BMC Systems Biology

A unified frame of predicting side effects of drugs by using linear neighborhood similarity

Wen Zhang[1], Xiang Yue[2], Feng Liu[2], Yanlin Chen[3], Shikui Tu[4] and Xining Zhang[1*]

Abstract

Background: Drug side effects are one of main concerns in the drug discovery, which gains wide attentions. Investigating drug side effects is of great importance, and the computational prediction can help to guide wet experiments. As far as we known, a great number of computational methods have been proposed for the side effect predictions. The assumption that similar drugs may induce same side effects is usually employed for modeling, and how to calculate the drug-drug similarity is critical in the side effect predictions.

Results: In this paper, we present a novel measure of drug-drug similarity named "linear neighborhood similarity", which is calculated in a drug feature space by exploring linear neighborhood relationship. Then, we transfer the similarity from the feature space into the side effect space, and predict drug side effects by propagating known side effect information through a similarity-based graph. Under a unified frame based on the linear neighborhood similarity, we propose method "LNSM" and its extension "LNSM-SMI" to predict side effects of new drugs, and propose the method "LNSM-MSE" to predict unobserved side effect of approved drugs.

Conclusions: We evaluate the performances of LNSM and LNSM-SMI in predicting side effects of new drugs, and evaluate the performances of LNSM-MSE in predicting missing side effects of approved drugs. The results demonstrate that the linear neighborhood similarity can improve the performances of side effect prediction, and the linear neighborhood similarity-based methods can outperform existing side effect prediction methods. More importantly, the proposed methods can predict side effects of new drugs as well as unobserved side effects of approved drugs under a unified frame.

Keywords: Drug side effects, Linear neighborhood similarity, Missing side effects

Background

A drug is a chemical substance which can treat, cure or prevent diseases, but all drugs can may have unexpected effects. In this paper, side effects refer to adverse effects of drugs. According to the reports of Food and Drug Administration (FDA), many drugs were withdrawn from markets because of fatal side effects. Identifying side effects of candidate drug molecules is critical for the success of drug discovery [1–6]. For drug safety, the investigation of side effects should be conducted before marketing new

drugs. Since wet methods are usually time-consuming and labor-intensive, researchers developed the computational methods to predict drug side effects.

For a long time, researchers defined preclinical drug-induced effect patterns to investigate the structure-response relationships or structure-property relationships [7–11], and then utilized them to identify drug side effects. However, these methods have to analyze data case by case, and are not suitable for complicated data. In recent years, the machine learning technique becomes more and more popular, and has been introduced to predict drug side effects. In general, machine learning-based methods are designed to complete two tasks. As demonstrated in Fig. 1,

* Correspondence: zhangxn@whu.edu.cn
[1]School of Computer, Wuhan University, Wuhan 430072, China
Full list of author information is available at the end of the article

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 24 of 88
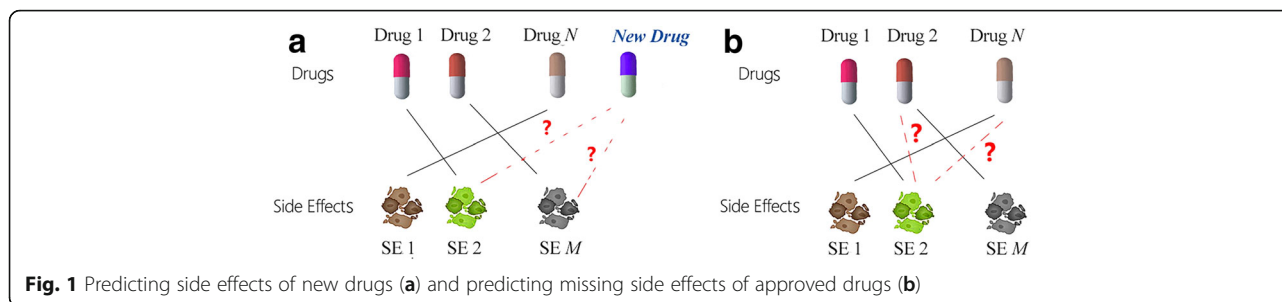
Fig. 1 Predicting side effects of new drugs (**a**) and predicting missing side effects of approved drugs (**b**)

one task is to predict side effects of new drugs (abbreviated "SEND"), and the other task is to predict missing side effects of approved drugs (abbreviated "SEAD").

As far as we know, many methods have been proposed for the SEND task, and they usually predict drug side effects from their structures or related features. Huang [12] considered drug targets, protein-protein interaction networks and gene ontology annotations, and adopted two types of classifiers: support vector machine (SVM) and logistic regression, and then built prediction models. Pauwels [13] explored chemical substructures of drugs, and utilized k-nearest neighbor classifier, support vector machine, ordinary canonical correlation analysis and sparse canonical correlation analysis to construct prediction models respectively. Yamanishi [14, 15] adopted the sparse canonical correlation analysis to build models based on drug substructures and drug targets. Liu [16] merged five types of drug feature vectors, and respectively utilized logistic regression, naive Bayes, k-nearest neighbor classifier, random forest and SVM to build prediction models. Huang [17] combined protein-protein interaction networks and drug substructures to build prediction models by using SVM. Zhang formulated the side effect prediction as the multi-label learning, and adopted the multi-label KNN to make predictions [18]. There are also several methods designed for the SEAD task. Cheng [19] utilized the resource allocation method to infer missing side effects from the known side effect-based network. Zhang formulated the original problem as the recommender systems, and utilized the resource allocation method, the restricted Boltzmann machine method and the collaborative filtering method to predict unobserved side effects [20]. In general, most existing methods were developed for either SEND task or SEAD task, but few methods can be used for both tasks.

In related studies, researchers usually assumed that similar drugs may induce same side effects, and then built side effect prediction models based on the assumption. The assumption is established on the biological common sense, and similarity-based models have good performances in the side effect prediction. Clearly, the drug-drug similarity is the key to the development of similarity-based models. In previous work [21], we considered a new measure named "linear neighborhood similarity" to calculate drug-drug similarity, and built prediction models to predict side effect of new drugs. In this paper, we present a unified frame based on linear neighborhood similarity to predict side effects of new drugs (SEND task) as well as unobserved side effects of approved drugs (SEAD task).

In this paper, we present the linear neighborhood similarity to calculate drug-drug similarity in a drug feature space, and then transfer the linear neighborhood similarity from the feature space into the side effect space. Therefore, we can predict drug side effects by propagating known side effect information through a similarity-based graph. We propose method "LNSM" and its extension "LNSM-SMI", which respectively make use of single features and multiple features to predict side effects of new drugs (SEND task); we propose the method "LNSM-MSE" which can predict unobserved side effect of approved drugs based on known side effects (SEAD task). The computational experiments show that the linear neighborhood similarity can produce better performances than other similarity measures in our models. When evaluated by cross validation, the proposed methods can produce high-accuracy performances for both SEND task and SEAD task, and outperform benchmark methods.

Methods
Datasets
Motivated by studies on big data, researchers have constructed several databases to facilitate the computational works about drugs. SIDER database [22] contains approved drugs and their reported side effects, which were extracted from public documents and package inserts. PubChem Compound Database [23, 24] contains experimentally validated information about substances, especially their structures. DrugBank database [25–28] contains FDA-approved small molecule drugs, biotech drugs, nutraceuticals, experimental drugs and their related non-redundant protein (drug target, enzyme, transporter, carrier) sequences. KEGG DRUG database [29] is a comprehensive database for approved drugs in Japan,

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 25 of 88

USA, and Europe, providing chemical structures, targets, metabolizing enzymes and etc.

Various features about drugs can be extracted from above databases. The drug chemical substructures provide direct information related with side effects, and are available in PubChem Compound Database. Drug targets may play roles in the particular metabolic or signaling pathway, and thus incur side effects; transporters are responsible for drug absorption, distribution and excretion in tissues; enzymes affect the metabolism to activate drugs, and may be associated with side effects. The pathways and indications are usually considered as the direct factors that induce drug side effects. The information about targets, transporters, enzymes and pathways are available in DrugBank database. Drug indications are provided in SIDER database.

From above data sources, Pauwels et al. [13], Mizutani et al. [14] and Liu et al. [16] compiled several benchmark datasets, and used them for the drug side effect prediction. In our previous work [18], we also compiled a dataset, and we named it "SIDER 4 dataset" [18]. Table 1 detailedly describe above mentioned datasets. The datasets contain drugs and their side effects, and include drug-related features as well. The features in different datasets are introduced. Pauwels's dataset has only one drug feature: substructure, and Mizutani's dataset has two features: substructures and targets; both Liu's dataset and SIDER 4 dataset has six drug-related features. Numbers in Table 1 represent the number of corresponding descriptors for a feature. For example, 881 types of substructures are defined in PubChem, and the feature "substructure" has 881 descriptors because of 881 types of substructures.

Linear neighborhood similarity

As introduced above, we usually have different features to describe the chemical or biological characteristics of drugs. Since one feature is actually a set of descriptors, a drug can be described by a subset of descriptors in the feature, and thus represented as a binary feature vector, whose dimensions means the presence or absence of descriptors by using the value 1 or 0. When we have different features, we can represent a drug as feature vectors in different feature spaces.

A drug can be considered as a data point in the feature space. How to calculate drug-drug similarity in a feature space is of the most importance for the drug side effect prediction. As far as we know, researcher have proposed several measures to calculate the similarity between data points in the feature space, and popular similarity measures are Jaccard similarity, Cosine similarity and Gauss similarity. Here, we present a novel similarity measure "linear neighborhood similarity" for the side effect prediction, and introduce them as below.

Roweis et al. [30] revealed that the locally linear patch of the manifold in a feature space can be described by data points and neighbor data points; Wang et al. [31] discovered that each point in the high-dimension space may be reconstructed by its neighbors.

Let X_i denote the p-dimensional feature vector of drugs d_i in a feature space, $i = 1, 2, \cdots N$. By considering feature vectors as data points in the feature space, we assume that a data point X_i approximate to the linear combination of neighbor data points, and write the objective function, which minimizes the reconstruction error,

$$
\begin{aligned}
\varepsilon_i &= \left\| X_i - \sum_{i_j,\, X_{i_j} \in N(X_i)} w_{i,i_j} X_{i_j} \right\|^2 + \lambda \| w_i \|^2 \\
&= \sum_{X_{i_k},\, X_{i_j} \in N(X_i)} w_{i,i_j} G_{i_j,i_k} w_{i,i_k} + \lambda \sum_{X_{i_j}} \in N(X_i) (w_{i,i_j})^2 \\
&= w_i^T (G^i + \lambda \mathrm{I}) w_i \\
&\quad \text{s.t.} \sum_{X_{i_j}} \in N(X_i) w_{i,i_j} = 1, \quad w_{i,i_j} \geq 0
\end{aligned}
$$

(1)

where $N(X_i)$ are the set of K nearest neighbors of X_i. I is the identity matrix of order N. $w_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,K})^T$. $G^i = (G_{i_j,i_k})$. If $X_{i_j}, X_{i_k} \in N(X_i)$, $G_{i_j,i_k} = (X_i - X_{i_j})^T (X_i - X_{i_j})$; otherwise, $G_{i_j,i_k} = 0$; $i_j = 1, 2, \cdots, K$, $i_k = 1, 2, \cdots, K$. w_{i,i_j} describe how to construct X_i from X_{i_j}, and be approximately taken as the similarity between two drugs. The first term of (1) is the reconstruction error; the second term of (1) is for regularization, and λ is the hyper parameter.

The parameter λ is very important for the regularization form of (1). Here, we discuss how to set the parameter. Since $\sum_{X_{i_j}} \in N(X_i) w_{i,i_j} = 1$, $0 \leq \| w_i \|^2 \leq 1$,

Table 1 Details about benchmark datasets

Dataset	#drug	#side effect	#substructure	#target	#transporter	#enzyme	#pathway	# indication
Pauwels's dataset	888	1385	881	N.A	N.A	N.A	N.A	N.A
Mizutani's dataset	658	1339	881	1368	N.A	N.A	N.A	N.A
Liu's dataset	832	1385	881	786	72	111	173	869
SIDER 4 dataset	1080	2260	881	1050	96	160	268	2537

N.A means unavailable information

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 26 of 88

and then $\|X_i - \sum_{i_j,} X_{i_j} \in N(X_i) w_{i,i_j} X_{i_j}\|^2 = \|\sum_{i_j,} X_{i_j} \in N(X_i) w_{i,i_j}(X_i - X_{i_j})\|^2 \leq p$. p is the dimension of feature vectors in the feature space. Clearly, $p \gg 1$, and we can let $\lambda = 1$ to make sure that the error term is greater than the regularization term in (1).

We can adopt the standard quadratic programming technique to solve (1) for each data point X_i, $i = 1, 2, \cdots, N$. The pairwise similarities between N drugs can be written as a $N \times N$ similarity matrix $W = (w_1, w_2, \cdots, w_N)^T$. We notice that the regularization term is not used if $\lambda = 0$. Therefore, we can calculate the linear neighborhood similarity which we name "LN similarity" if $\lambda = 0$, and calculate the regularization form of linear neighborhood similarity which we name "RLN similarity" if $\lambda = 1$.

By using linear neighborhood similarity, we can develop prediction methods for the SEND task and SEAD task, which are described in Fig. 2. Methods for SEND task are introduced in section 2.3, and the method for SEAD Task is introduced in section 2.4.

Linear neighborhood similarity-based methods for SEND task

In this section, we propose methods for the SEND task by using the linear neighborhood similarity. One method named "LNSM" is to make predictions based on single features about drugs; the other named "LNSM-SMI" is the extension of LNSM, which can make predictions by integrating multiple features about drugs.

Linear neighborhood similarity method (LNSM)

Given N drugs, these drugs represented as feature vectors $X_1, X_2, \cdots, X_N$ in a p-dimensional feature space, where $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$. Suppose we want to predict M types of side effects for drugs, the presence or absence of side effects for N drugs can be represented as M-dimensional vectors named side effect profiles $Y_1, Y_2, \cdots, Y_N$. $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iM})$, where $Y_{ij} = 1$, if the ith drug has the jth side effect; else, $Y_{ij} = 0$, $i = 1, 2, \cdots N, j = 1, 2, \cdots M$. Therefore, $\{(X_i, , Y_i)\}_{i=1}^N$ are annotated

dataset for training models. We respectively concentrate $X_1, X_2, \cdots, X_N$ and $Y_1, Y_2, \cdots, Y_N$ row by row, and obtain two matrices and Y. In the feature space, we can easily calculate linear neighborhood similarities between N drugs, which are denoted by a similarity matrix W. Then, we describe how to build LNSM models.

First of all, we construct a directed graph, which uses N given drugs as nodes and drug-drug similarities as edge weights. We consider a side effect term as a type of label, and a node has the label if the drug has the side effect. The ith column of Y response to the labels for N nodes in terms of ith side effect term. Label information is propagated on the graph, by following the rule that a node absorbs labels of neighbors with the probability α and retain the initial labels with the probability $1 - \alpha$. Considering all side effect terms simultaneously, we can formulate the update equation in the matrix from,

$$Y^{p+1} = \alpha W Y^p + (1-\alpha) Y^0 \qquad (2)$$

Where Y^0 is the matrix for initial label information, and $Y = Y^0$. Y is matrix representing the updated labels for N nodes. The iteration will converge to

$$Y' = (1-\alpha)(I - \alpha W)^{-1} Y^0 \qquad (3)$$

Where I is the identity matrix of order N. Y' is final labels for N nodes.

When we have a new drug X_{new} for prediction, we take the drug as out-of-sample data, and calculate the similarities between X_{new} and N known drugs in the feature space. The similarities are represented by a vector $W_{new} = (w_{new, 1}, w_{new, 2}, \cdots, w_{new, N})$. Thus, we can predict the side effects of X_{new},

$$Y_{new=} W_{new} \times Y$$

According to the above discussion, LNSM predicts side effect of new drugs from single drug features.
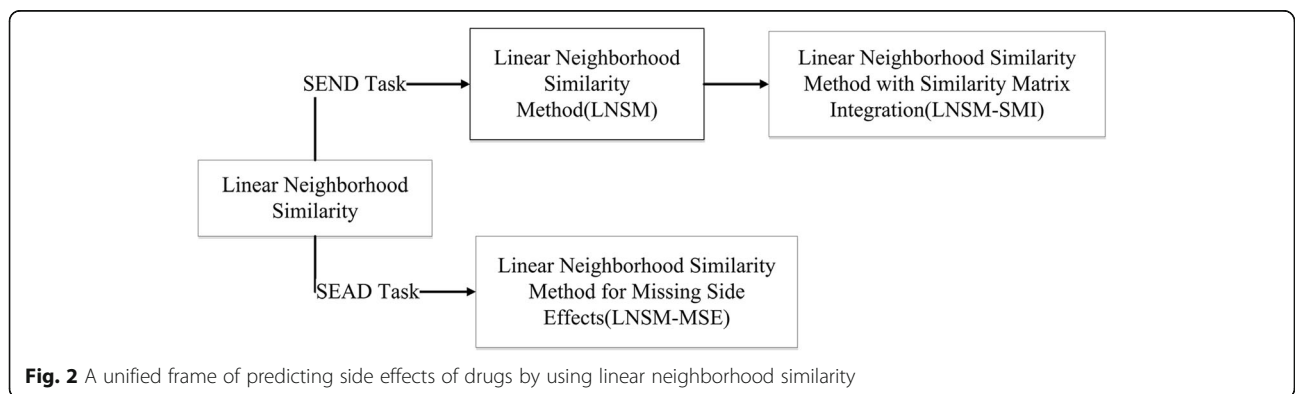
Fig. 2 A unified frame of predicting side effects of drugs by using linear neighborhood similarity

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 27 of 88

Linear neighborhood similarity method with similarity matrix integration

In order to predict side effects of new drugs, researchers usually collect various drug features, and construct the relationship between features and their side effects. When we have multiple drug features, we have to face the challenges of integrating features to make predictions. For the purpose, we propose the linear neighborhood similarity method with similarity matrix integration (LNSM-SMI) by extending LNSM.

Given N drugs, we have K features to describe characteristics of drugs. Let X_i^k denote the feature vector based on kth feature for the ith drug, and Y_i denotes the side effect vector for the ith drug. In K feature spaces, we calculate similarities between N drugs, and represent them as similarity matrices. K features can produce K similarity matrices $W_1, W_2, \cdots, W_K$. Then, we describe how to build models based on multiple features.

First of all, the study in [31] proved the label propagation on the graph shown in (2) is equivalent to a convex optimization problem,

$$\min_Y \alpha\left(tr\left(Y^T\left(I-W\right)Y\right)\right) + \left(1-\alpha\right)\left\|Y-Y^0\right\|_F^2 \quad (4)$$

When we have similarity matrices $W_1, W_2, \cdots, W_K$ based on K features, we consider the linear sum of these matrices $\sum_{i=1}^{K}\theta_i W_i$. By replacing W in (4) with $\sum_{i=1}^{K}\theta_i W_i$, we can obtain the optimization problem,

$$\min_{Y,\theta} \alpha\left(tr\left(Y^T\sum_{i=1}^{K}\theta_i(I-W_i)Y\right)\right) + (1-\alpha)\left\|Y-Y^0\right\|_F^2 + \|\theta\|^2 \quad (5)$$

$$s.t. \sum_{i=1}^{K}\theta_i = 1; \forall i, \theta_i \geq 0$$

where $\delta(>0)$ is hyper parameter for the regularization term $\|\theta\|^2$.

The matrix $Y^0 = [Y_1, Y_2, \cdots, Y_N]^T$ represents observed side effects for N known drugs, and we can set $Y = Y^0$ and rewrite (5) as

$$\min\theta \, \alpha\left(tr\left((Y^0)^T\sum_{i=1}^{K}\theta_i(I-W_i)Y^0\right)\right) + \delta\|\theta\|^2 \quad (6)$$

$$s.t. \sum_{i=1}^{K}\theta_i = 1; \forall i, \, _i \geq 0$$

We introduce the Lagrange Multiplier terms λ and $\eta = (\eta_1, \eta_2, \cdots, \eta_K)^T$ to solve the optimization problem,

$$L(\theta, \lambda, \eta) = \delta\theta^T\theta + \alpha C^T\theta - \lambda(e^T\theta - 1) - \eta^T \quad (7)$$

Where $c_i = trace((Y^0)^T(I - W_i)Y^0)$, $C = (c_1, c_2, \cdots, c_K)^T$, and $e = (1, 1, \cdots, 1)^T$. The KKT condition is,

$$\begin{cases} \nabla_\theta L(\theta, \lambda, \eta) = 2\delta\theta + \alpha C - \lambda e - \eta = 0 \\ e^T\theta - 1 = 0 \\ \eta_i \geq 0, \theta_i \geq 0, \eta_i\theta_i = 0, i = 1, \cdots, K \end{cases} \quad (8)$$

In (8), $L(\alpha, \lambda, \eta) = 2\delta\theta_i + \alpha c_i - \lambda - \eta_i = 0$ and $\eta_i = 2\delta\theta_i + \alpha c_i - \lambda$, and thus we know that $\theta_i(2\delta\theta_i + \alpha c_i - \lambda) = 0$. Since $0 \leq \theta_i \leq 1$, we can know that $\theta_i = 0$ if $\lambda - \alpha c_i \leq 0$; otherwise, $\theta_i = (\lambda - \alpha c_i)/(2\delta)$. We reorder $c_1, c_2, \cdots, c_K$ as $c_1 \leq c_2 \leq \cdots \leq c_K$, and then the corresponding weights $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_l > \theta_{l+1} = \cdots = \theta_K = 0$. Therefore, we can obtain the solution for the optimization problem in (5),

$$\begin{cases} l = \max\{n \,|\delta \geq \frac{\alpha}{2}\sum_{k=1}^{n}(c_k - c_i), n = 1, 2\cdots, K\} \\ \lambda = \frac{1}{l}\left(2\delta + \alpha\sum_{k=1}^{l}c_i\right) \\ \theta_i = \frac{2\delta + \alpha\sum_{k=1}^{l}(c_k - c_i)}{2l\delta}, \eta_i = 0, i = 1, \cdots, l \\ \theta_i = 0, \eta_i = \frac{1}{l}\left(\alpha\sum_{k=1}^{l}(c_i - c_k) - 2\delta\right), i = l+1, \cdots, K \end{cases}$$

$$(9)$$

Let $c_{max} = \max\{c_1, c_2, \cdots, c_K\}$. Clearly, the free parameter δ determine the number of nonzero weights. In order to guarantee $\delta \geq \frac{\alpha}{2}\sum_{k=1}^{n}(c_k - c_i)$, we can set $\delta = \frac{\alpha}{2}\sum_{k=1}^{K}(c_{max} - c_k)$. Therefore, we can estimate weights in a simple form,

$$\theta_i = \frac{c_{max} - c_i}{\sum_{k=1}^{K}(c_{max} - c_k)}, i = 1, 2, \cdots, K \quad (10)$$

When we have a new drug X_{new} described by K features, we can calculate similarities between the new drug X_{new} and known drugs, represented by K vectors W_{new}^i, $i = 1, 2, \cdots, K$. Thus, we can predict the side effects of X_{new} based on K features,

$$Y_{new} = \left(\sum_{i=1}^{K}\theta_i W_{new}^i\right)Y^0$$

Clearly, LNSM-SMI is the extension of LNSM to make use of multiple features for prediction.

Linear neighborhood similarity method for SEAD task

In this section, we propose the method "LNSM-MSE" to predict missing or unobserved side effects of approved drugs by using the linear neighborhood similarity.

Given N Drugs and M side effect terms, we known that these drugs have observed side effects. By linking drugs and induced side effects, relations between drugs

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 28 of 88

and induced side effects can be formulated as a bipartite network. The bipartite network can be described by an $N \times M$ association matrix A, where $A_{ij} = 1$ if the drug i induces side effect j and $A_{ij} = 0$ otherwise. For each drug d_i, i = 1, 2, $\cdots$, N, the associate profile of d_i is the vector $A(i, :)$ = $(A_{i1}, A_{i2}, \cdots, A_{iM})$, which represents the known side effects of the drug. The drug-side effect bipartite network and the association matrix are demonstrated in Fig. 3.

Then, we calculate linear neighborhood similarities W between drugs based on their association profiles, and construct the directed graph which uses drugs as nodes and use similarities as edge weights. The known side effect information is propagated on the graph as described in section 2.2.2, and the update will converge. Thus, we can predict missing side effects of N approved drugs,

$$Y = (1-\alpha)(I-\alpha W)^{-1} A$$

If $A_{ij} = 0$, the entry Y_{ij} indicates the probability of drug d_i inducing the jth side effect. Therefore, LNSM-MSE predict missing side effects of approved drugs based on their known side effects.

Results and discussion
Evaluation metrics
In the paper, we evaluate prediction models by using five-fold cross validation (5-CV). The five-fold cross validation in the SEND task randomly splits all drugs into equal-sized subsets. In each fold, four subsets of drugs are used as the training set, and other drugs are used as the testing set. The models are constructed on training set with annotated features and side effects, and then predict side effects of drugs in the testing test from features. In the SEAD task, the five-fold cross validation splits all known side effects into equal-sized subsets. We construct the prediction models based on all drugs and known side effects in the training set, and apply the model to predict unobserved side effects for all drugs.

In the SEND task, the side effect prediction is a multi-label learning task [18]. Therefore, we adopt several evaluation metrics for the multi-label classification to evaluate models, i.e. Hamming loss, one-error, coverage, ranking loss and average precision. In addition, we use the area under ROC curve (AUC) and the area under the precision-recall curve (AUPR). The smaller scores of one-error, coverage, ranking loss and hamming loss indicate better results, and the smaller scores of AUC and AUPR mean better results.

For the SEAD task, we adopt several binary classification metrics to evaluate the performances of models, including specificity (SP), sensitivity (SN), accuracy (ACC), F-measure (F), recall, precision, AUC and AUPR,

For all drugs and all side effect terms, the associated drug-side effect pairs which indicate that drug induces the side effect are much more than other pairs. Since data is imbalanced, we adopt the AUPR as the primary metric to evaluate the models in both SEND task and SEAD task.
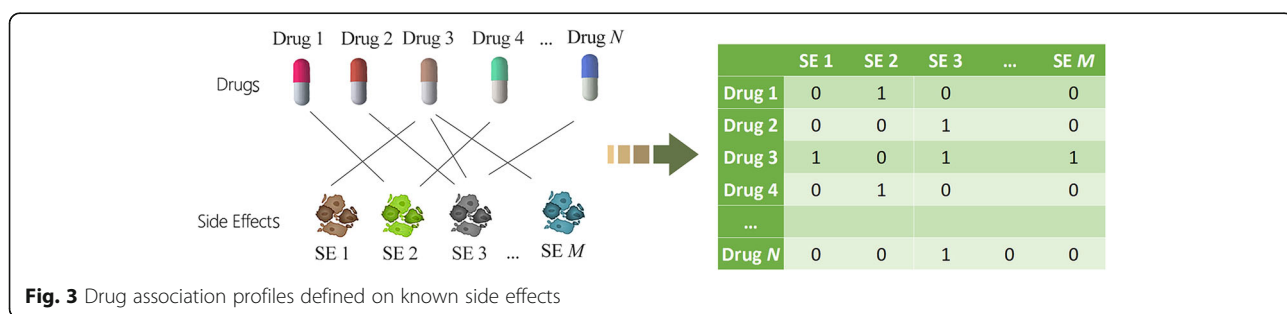
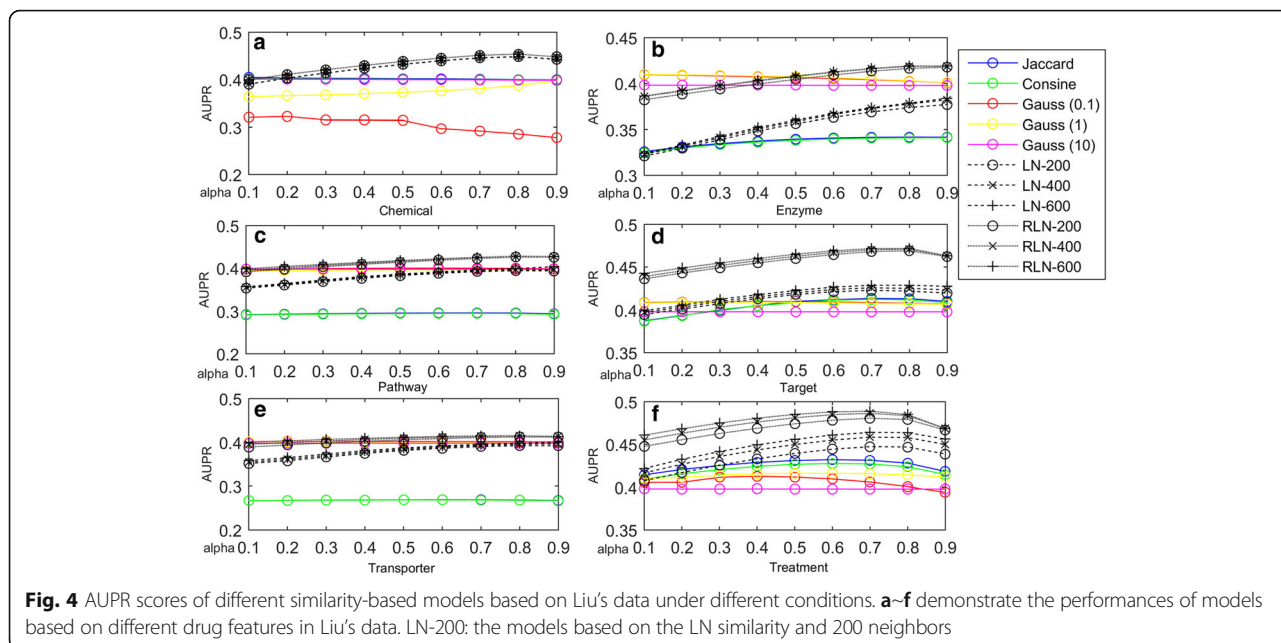### Performances of linear neighborhood similarity methods for SEND task
By using the linear neighborhood similarity, we present the linear neighborhood similarity method (LNSM) and the linear neighborhood similarity method with similarity matrix integration (LNSM-SMI). LNSM uses single drug features to make predictions; as the extension of LNSM, LNSM-SMI integrates multiple features for predictions. In this section, we evaluate LNSM and LNSM-SMI based on Liu's dataset.

Performances of LNSM
LNSM can build the prediction models based on the single features. Liu's dataset has a variety of features, and we respectively construct prediction models based on the different features, and evaluate their usefulness.

LNSM calculates drug-drug similarity in a feature space, and then predict side effects of new drugs. There are two parameters in LNSM: the absorbing probability α and the neighbor number K. Liu's dataset has 832 drugs, and thus the five-fold cross validation has about 665 training drugs in each fold. Therefore, the neighbor number K should be less than 665 in our study. To test the impact of parameters on LNSM, we consider α in $\{0.1, 0.2, \cdots 0.9\}$ and K in $\{200, 400, 600\}$ to build prediction models. In addition, we consider different similarities: Jaccard similarity, Cosine similarity and Gauss similarity

	SE 1	SE 2	SE 3	...	SE *M*
Drug 1	0	1	0		0
Drug 2	0	0	1		0
Drug 3	1	0	1		1
Drug 4	0	1	0		0
...					
Drug *N*	0	0	1	0	0

Fig. 3 Drug association profiles defined on known side effects

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 29 of 88

Fig. 4 AUPR scores of different similarity-based models based on Liu's data under different conditions. **a~f** demonstrate the performances of models based on different drug features in Liu's data. LN-200: the models based on the LN similarity and 200 neighbors

to compare with the linear neighborhood similarity (LN) and regularized linear neighborhood similarity (RLN). Figure 4 demonstrates AUPR scores of all prediction models evaluated by five-fold cross validation.

According to the results in Fig. 4, LNSM prediction models which use LN similarity and RLN similarity produce robust results for the parameters: the neighbor number K and absorbing probability α. RLN similarity is the LN similarity with the regularization term. The introduction of the regularization term usually enhances generalization capability of prediction models. One drawback of LN is that the G^i in the Eq. (1) may be a singular matrix, and the introduction of the regularization term can alleviate the singular matrix problem in solving quadratic programming. Therefore, we have observed that LNSM models based on RLN similarity can lead to better experimental results than LNSM models based on LN similarity under all conditions. In general, the LNSM models produce the best results when using 400 neighbors and α of 0.8.

Figure 4 also demonstrates the results of prediction models based on different similarities. In fact, the linear neighborhood similarity and its regularized form calculate the similarity in a feature space by considering linear relationship of data points, and the similarity can be transferred into the side effect space and be used by the label propagation, which is also in a linear form. In contrast, other similarities (Jaccard similarity, Cosine similarity and Gauss similarity) calculates the drug-dug similarity in a nonlinear from. Therefore, the models based on LN similarity and RLN similarity yield better AUPR scores than models based on other similarities.
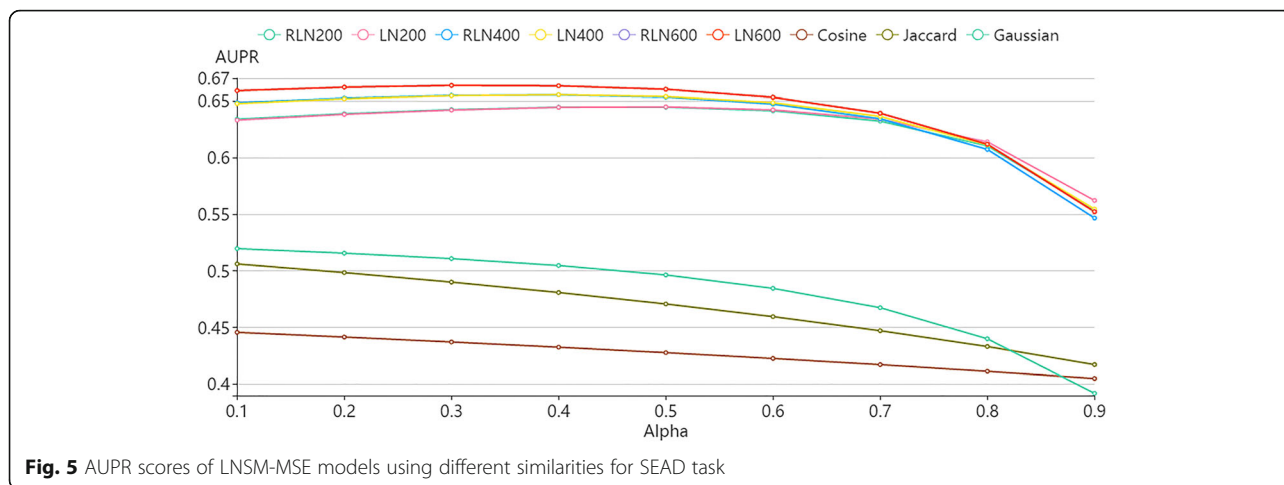
Superiority of LNSM is demonstrated in this section. The parameters: the neighbor number of 400 and α of 0.8 are used for LNSM in the following experiments.

Performances of LNSM-SMI
When diverse features are available, researchers usually combine or integrate multiple features in order to achieve high-accuracy prediction models [18, 20, 32–38].

Table 2 5-CV performances of prediction models on Liu's dataset

Data	Methods	AUC	AUPR	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision
Enzyme	LNSM	0.8898	0.4187	0.0473	0.0821	0.1659	846.3846	0.4696
Pathway	LNSM	0.8886	0.4273	0.0470	0.0776	0.1647	814.6298	0.4932
Target	LNSM	0.8991	0.4708	0.0452	0.0690	0.1538	792.3726	0.5216
Transporter	LNSM	0.8896	0.4147	0.0477	0.0817	0.1611	849.3161	0.4762
Treatment	LNSM	0.9013	0.4836	0.0446	0.0710	0.1262	806.8558	0.5232
Substructure	LNSM	0.8944	0.4538	0.0459	0.0714	0.1490	803.5228	0.5184
All data	LNSM-SMI	0.8986	0.5053	0.0435	0.0670	0.1154	789.8486	0.5476

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 30 of 88

Fig. 5 AUPR scores of LNSM-MSE models using different similarities for SEAD task

As discussed above, we have multiple features to describe chemical and biological characteristics of drugs. Here, we test the performances of the integration method: the linear neighborhood similarity method with similarity matrix integration (LNSM-SMI), which integrate diverse and multiple features.
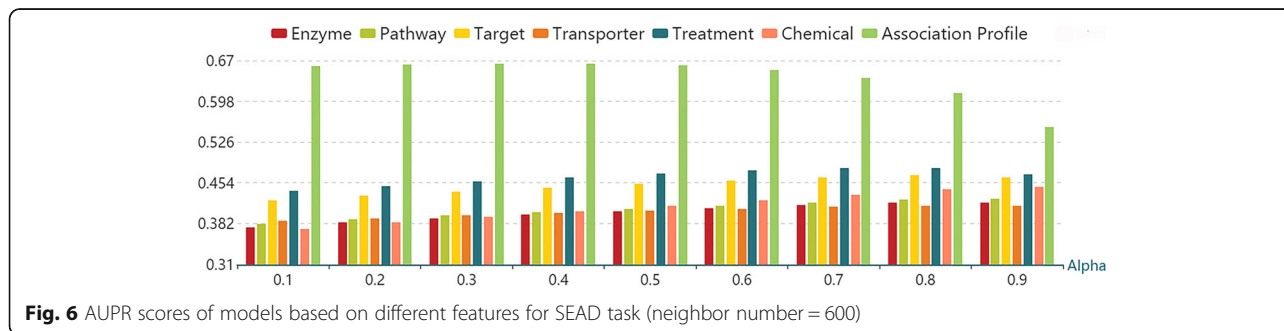
All prediction models are evaluated based on Liu's dataset by using 5-fold cross validation. Table 2 shows the performances of integration models LNSM-SMI which use multiple features and LNSM models based on single features. We respectively build six LNSM models by using six features, and build a LNSM-SMI model by integrating six features. As shown in Table 2, the feature "indication" can produce the LNSM model with best performances, and the performances of targets, substructures, pathways, enzymes and transporters are sorted descendingly. Clearly, the data integration model LNSM-SMI can greatly improve the performances of LNSM based on indications, achieving the AUPR scores of 0.5053. The improvements in terms of other evaluation metrics can be observed as well. Therefore, LNSM-SMI can effectively combine multiple features to predict side effects of new drugs.

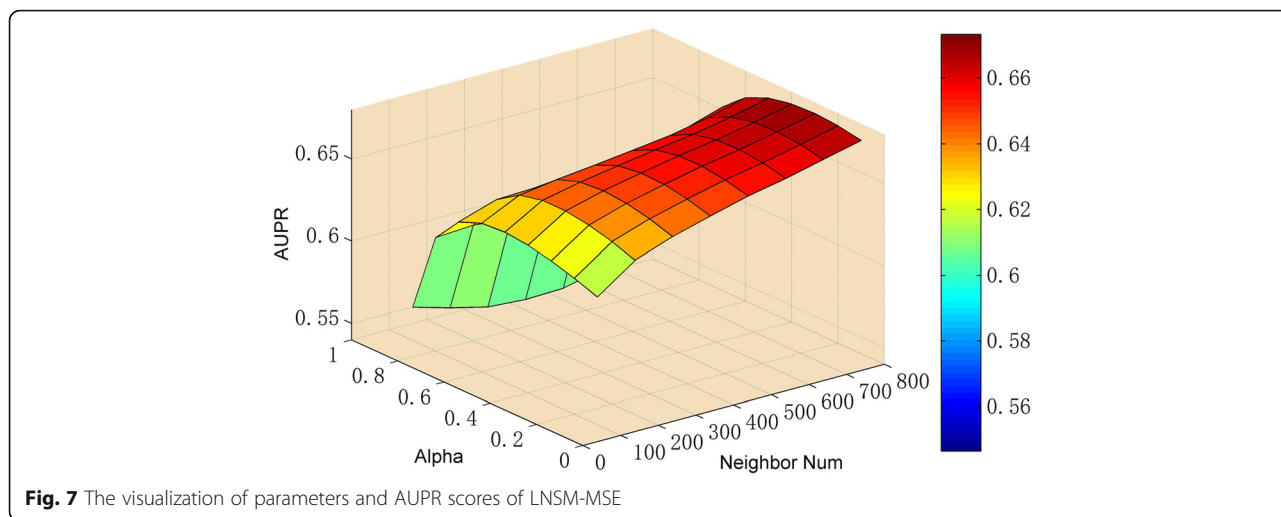LNSM-SMI has the weights $\alpha_1, \alpha_2, \cdots, \alpha_K$ for similarity matrices, which are calculated from K different features. We analyzed how to estimate weights in LNSM-SMI, and give out the analytical solutions in (10). Thus, we investigate weights $\alpha_1, \alpha_2, \cdots, \alpha_K$ in LNSM-SMI models. The weights $\alpha_1, \alpha_2, \cdots, \alpha_K$ directly indicate the features' contributions to the data integration models, and we can observe that features which have better performances in LNSM can usually gain greater weights in LNSM-SMI. We further conduct simulation experiments to demonstrate the importance of weights in LNSM-SMI. Here, we randomly generate 100 sets of weights, and use them to construct LNSM-SMI models. We analyze the AUPR scores of these LNSM-SMI models evaluated by 5-CV, and our statistics is 0.4912 ± 0.0104. The results show that the optimal weights are very important for LNSM-SMI, and arbitrary weights cannot yield the superior performances. Clearly, our estimation in (10) can effectively determine the optimal weights, and produce the satisfying results in the computational experiments,

Performances of LNSM-MSE for SEAD task

By using the linear neighborhood similarity, we develop LNSM-MSE to predict missing side effect of approved drugs.

Fig. 6 AUPR scores of models based on different features for SEAD task (neighbor number = 600)

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 31 of 88

Fig. 7 The visualization of parameters and AUPR scores of LNSM-MSE

LNSM-MSE calculates drug-drug similarity based on the drug side effect association profiles, which are defined on the known side effects of approved drugs, and then build models. First of all, we consider different similarity measures, including Jaccard similarity, Cosine similarity, Gauss similarity, LN similarity and RLN similarity for the purpose of comparison. We consider the neighbor number K 200,400 and 600 for LN similarity and RLN similarity. The probability α 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 are considered for the label propagation. Figure 5 demonstrates AUPR scores of different models. The results indicate that LN similarity and RLN similarity also outperform other similarities in predicting missing side effects of approved drugs (SEAD task). Since LN similarity has similar performances as RLN similarity in the SEAD task, we use RLN similarity to construct LNSM-MSE in the following study.

Since we have multiple features about drugs in Liu's dataset, we can calculate drug-drug RLN similarities based on different features, and build the prediction models which are similar to LNSM-MSE. Here, we respectively use different features to construct LNSM-MSE models, and compare different features. As shown in Fig. 6, the results demonstrate that the association profile can have significantly better performances than other features. Clearly, association profiles of drugs can bring critical information for modelling, and LNSM-MSE can produce the AUPR score greater than 0.65 by only using the association profile.

Finally, we consider greater ranges for parameters neighbor number K and the absorbing probability α, and determine the optimal parameters for the LNSM-MSE, which utilizes the association profile and RLN similarity. For neighbor number K, we consider 100, 200, …., 800; we consider the absorbing probability α 0.1,0.2,…0.9. We try different parameter combinations, and AUPR scores of LNSM-MSE models based on different parameter values are visualized in Fig. 7. LNSM-MSE can produce the best results when K = 800 and $α$ = 0.3, and these parameter values are used for final LNSM-MSE models in following experiments.

Comparison with benchmark methods

As we mentioned, lots of methods have been proposed to predict drug side effects, and some methods which provided source codes and datasets are usually used as benchmark methods for comparison. These benchmark methods we consider are Pauwels's method [13], Mizutani's method [14], Cheng's method [19], Liu's method [16], RBMBM [20], INBM [20] and

Table 3 Performances of our methods and other state-of-the-art methods

Dataset	Method	AUC	AUPR	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision
Pauwels's dataset	Pauwels's method	0.8827	0.3883	0.0577	0.0827	0.1779	832.7827	0.4616
	LNSM	0.8941	0.4491	0.0444	0.0713	0.1633	790.9471	0.5126
Mizutani's dataset	Mizutani's method	0.8665	0.4107	0.0557	0.0888	0.1854	862.9757	0.4795
	LNSM	0.8946	0.4624	0.0499	0.0746	0.1581	805.8875	0.5170
Liu's dataset	Liu's method	0.8850	0.2514	0.0721	0.0927	0.9291	837.4579	0.2610
	FS-MLKNN	0.9034	0.4802	0.0524	0.0703	0.1202	795.9435	0.5134
	LNSM-SMI	0.8986	0.5053	0.0435	0.0670	0.1154	789.8486	0.5476

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 32 of 88

Table 4 Performances of different methods in the independent test

Method	AUC	AUPR	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision
Liu's method	0.8772	0.1766	0.0421	0.1150	0.9870	1587.5663	0.1816
FS-MLKNN	0.8722	0.3109	0.0373	0.1038	0.1851	1535.9223	0.3649
LNSM-SMI	0.8786	0.3465	0.0291	0.0969	0.2013	1488.2977	0.3906

FS-MLKNN [18]. In this paper, we present a unified frame to handle two side effect prediction tasks by using linear neighborhood similarity. However, these benchmark methods are usually designed for either SEND task [13, 14, 18] or SEAD task [19, 20]; only on method: Liu's method is suitable for both tasks. Therefore, we compare our proposed methods with benchmark methods respectively in two tasks.

Comparison with benchmark methods for SEND task

For the SEND task, we adopt Pauwels's method, Liu's method, Mizutani's method and FS-MLKNN as benchmark methods for comparison. We replicate these methods by using their publicly available source codes or following details in publications. We respectively construct our prediction models by using the same datasets which were ever used for benchmark methods. Since only one feature "substructure" in Pauwels's dataset and Mizutani's dataset was usually used for modeling, we build LNSM models on these datasets to compare with corresponding methods. Liu's dataset has multiple features, and were ever used by Liu's method and FS-MLKNN, and thus we build LNSM-SMI models based on multiple features to make the comparison. Table 3 shows results of all methods evaluated by 5-fold cross validation. Clearly, the proposed methods outperform

benchmark methods under the same experimental conditions.

We further implement the independent experiments to evaluate the practical capability of our methods. Here, we adopt Liu's method and FS-MLKNN for comparison, for they usually have good performances on different datasets. The SIDER 4 dataset covers 1080 drugs, which have 771 drugs overlapped with Liu's dataset and 309 newly added drugs. In independent experiments, we train prediction models based on 771 drugs, and then make prediction for 309 new drugs. Table 4 demonstrates results of all models, and LNSM-SMI has significant advantages on the AUPR scores.

For each testing drug, we respectively consider top 100 and top 200 predicted side effect terms, and investigate how much known side effects can be found out. We calculate recall scores for drugs one by one, and conduct statistics on the results. By evaluating top 100 predictions, the statistics on AUPR scores of Liu's method, FS-MLKNN and LNSM-SMI are 0.4161 ± 0.0239, 0.5157 ± 0.0293, 0.5421 ± 0.0334; the statistics in evaluating top 200 predictions are 0.6261 ± 0.0262, 0.6605 ± 0.0263, 0.6840 ± 0.0285. LNSM-SMI has identified about 54% known side effects on average when checking up top 100 predicted side effects out of 2260 side effect terms, and has identified about 68% known side effects on average

Table 5 Performances of LNSM-MSE and benchmark methods evaluated by 5-CV

Dataset	Methods	AUPR	AUC	SN	SP	Precision	Accuracy	F
Pauwels's dataset	Liu's method	0.345	0.920	0.643	0.950	0.400	0.934	0.493
	Cheng's method	0.588	0.922	0.587	0.975	0.547	0.955	0.566
	RBMBM	0.612	0.941	0.605	0.977	0.579	0.958	0.592
	INBM	0.641	0.934	0.608	0.979	0.605	0.961	0.607
	LNSM-MSE	0.671	0.948	0.629	0.980	0.625	0.963	0.627
Mizutani's dataset	Liu's method	0.366	0.918	0.637	0.948	0.418	0.930	0.505
	Cheng's method	0.599	0.923	0.593	0.973	0.560	0.951	0.576
	RBMBM	0.619	0.939	0.614	0.974	0.581	0.954	0.597
	INBM	0.646	0.932	0.616	0.976	0.605	0.956	0.611
	LNSM-MSE	0.676	0.944	0.627	0.979	0.635	0.959	0.631
Liu's dataset	Liu's method	0.278	0.907	0.669	0.930	0.341	0.917	0.452
	Cheng's method	0.592	0.922	0.589	0.974	0.550	0.954	0.569
	RBMBM	0.616	0.941	0.608	0.976	0.581	0.957	0.594
	INBM	0.641	0.934	0.607	0.979	0.606	0.959	0.606
	LNSM-MSE	0.673	0.948	0.631	0.979	0.624	0.962	0.628

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 33 of 88

when checking up top 200 predictions. Therefore, LNSM-SMI is effective for predicting side effect of new drugs.

Comparison with benchmark methods for SEAD task

We propose LNSM-MSE to predict missing side effects of approved drugs from known side effects. In predicting missing side effects of approved drugs, we adopt Cheng's method [19], Liu's method [16], INBM [20] and RBMBM [20] for comparison. Liu's method makes use of multiple features for predictions, and other methods only use the known side effects to predict missing ones. Therefore, we construct LNSM-MSE and benchmark methods on benchmark datasets, and use 5-fold cross validation to evaluate models.

The performances of all methods are shown in Table 5. Clearly, LNSM-MSE can outperform the benchmark methods on the benchmark datasets, and significantly improve the AUPR score from 0.64 to 0.67. Moreover, LNSM-MSE has the better performances in terms of other evaluation metrics. Therefore, LNSM-MSE is useful and suitable for the SEAD Task.

Conclusions

This paper presents a novel similarity measure named "linear neighborhood similarity" to calculate drug-drug similarity, and develop a unified frame of predicting side effects of new drugs (SEAD task) as well as missing side effects of approved drugs (SEND task). Therefore, we propose the method "LNSM" and its extension "LNSM-SMI" to predict the side effects of new drugs; we propose the method "LNSM-MSE" to predict missing side effects of approved drugs. In computational experiments, proposed methods can produce good results, and outperform benchmark methods in two tasks. The proposed methods have great potential in predicting drug side effects.

Abbreviations
5-CV: 5-fold cross validation; AUC: area under ROC curve; AUPR: area under precision-recall curve; SEAD: predicting missing side effect of approved drug; SEND: predicting side effects of new drugs

Availability of data and materials
Not applicable

About this supplement
This article has been published as part of *BMC Systems Biology* Volume 11 Supplement 6, 2017: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2016: systems biology. The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-11-supplement-6.

Authors' contributions
WZ, XZ designed the study, implemented the algorithm and drafted the manuscript. XY, YC, ST and FL helped prepare the data and draft the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

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

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]School of Computer, Wuhan University, Wuhan 430072, China. [2]International School of Software, Wuhan University, Wuhan 430072, China. [3]School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China. [4]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Published: 14 December 2017

References
1. Whitebread S, Hamon J, Bojanic D, Urban L. Keynote review:safety pharmacology profiling: an essential tool for successful drug development. Drug Discov Today. 2005;10(21):1421–33.
2. Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y. When good drugs go bad. Nature. 2007;446(7139):975–7.
3. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321(5886):263–6.
4. Scheiber J, Chen B, Milik M, Sukuru SCK, Bender A, Mikhailov D, Whitebread S, Hamon J, Azzaoui K, Urban L. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. J Chem Inf Model. 2009;49(2):308–17.
5. Tatonetti NP, Liu T, Altman RB. Predicting drug side-effects by chemical systems biology. Genome Biol. 2009;10(9):238.
6. Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLoS Comput Biol. 2009;5(5):e1000387.
7. Mizuno N, Niwa T, Yotsumoto Y, Sugiyama Y. Impact of drug transporter studies on drug discovery and development. Pharmacol Rev. 2003;55(3):425–61.
8. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Analysis of drug-induced effect patterns to link structure and side effects of medicines. Nat Chem Biol. 2005;1(7):389–97.
9. Merle L, Laroche ML, Dantoine T, Charmes JP. Predicting and preventing adverse drug reactions in the very old. Drugs Aging. 2005;22(5):375–92.
10. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem. 2007;2(6):861–73.
11. Fukuzaki M, Seki M, Kashima H, Sese J. Side effect prediction using cooperative pathways. In: Bioinformatics and biomedicine, 2009 BIBM'09 IEEE international conference on: Washington, DC: IEEE; 2009. p. 142–7.
12. Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC genomics. 2011;12(Suppl 5):S11.
13. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC bioinformatics. 2011;12:169.
14. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y. Relating drug-protein interaction network with drug side effects. Bioinformatics. 2012;28(18):i522–8.
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. Liu M, Wu YH, Chen YK, Sun JC, Zhao ZM, Chen XW, Matheny ME, Xu H. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc. 2012;19(E1):E28–35.

Zhang *et al. BMC Systems Biology* 2017, **11**(Suppl 6):101

Page 34 of 88

17. Huang LC, Wu X, Chen JY. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. Proteomics. 2013;13(2):313–24.

18. Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC bioinformatics. 2015;16:365.

19. Cheng F, Li W, Wang X, Zhou Y, Wu Z, Shen J, Tang Y. Adverse drug events: database construction and in silico prediction. J Chem Inf Model. 2013;53(4):744–52.

20. Zhang W, Zou H, Luo L, Liu Q, Wu W, Xiao W. Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing. 2016;173:979–87.

21. Zhang W, Chen Y, Tu S, Liu F, Qu Q. Drug side effect prediction through linear neighborhoods and multiple data source integration. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM); 2016. p. 427–34.

22. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol. 2010;6:343.

23. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009, 37(Web Server issue):W623-W633.

24. Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discov Today. 2010;15(23–24):1052–7.

25. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–72.

26. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(Database issue):D901–6.

27. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011;39(Database issue):D1035–41.

28. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42(Database issue):D1091–7.

29. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355–60.

30. Roweis S, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–6.

31. Wang F, Zhang C. Label propagation through linear neighborhoods. Knowledge and Data Engineering, IEEE Transactions on. 2008;20(1):55–67.

32. Hu X, Mamitsuka H, Zhu S. Ensemble approaches for improving HLA class I-peptide binding prediction. J Immunol Methods. 2011;374(1–2):47–52.

33. Dehzangi A, Paliwal K, Sharma A, Dehzangi O, Sattar A. A combination of feature extraction methods with an Ensemble of Different Classifiers for protein structural class prediction problem. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 2013;10(3):564–75.

34. Yang R, Zhang C, Gao R, Zhang L. An ensemble method with hybrid features to identify extracellular matrix proteins. PLoS One. 2015;10(2):e0117804.

35. Zhang W, Niu Y, Xiong Y, Zhao M, Yu R, Liu J. Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. PLoS One. 2012;7(8):e43575.

36. Li D, Luo L, Zhang W, Liu F, Luo F. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. BMC bioinformatics. 2016;17(1):329.

37. Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. Accurate prediction of Transposon-derived piRNAs by integrating various sequential and physicochemical features. PLoS One. 2016;11(4):e0153268.

38. Zhang W, Chen Y, Liu F, Luo F, Tian G, Li X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC bioinformatics. 2017;18(1):18.