Link Prediction Only With Interaction Data and Its Application on Drug Repositioning

Juan Liu, *Member, IEEE*, Zhiqun Zuo, and Guangsheng Wu

***Abstract*— To assist drug development, many computational methods have been proposed to identify potential drug-disease treatment associations before wet experiments. Based on the assumption that similar drugs may treat similar diseases, most methods calculate the similarities of drugs and diseases by using various chemical or biological features. However, since these features may be unknown or hard to collect, such methods will not work in the face of incomplete data. Besides, due to the lack of validated negative samples in the drug-disease associations data, most methods have no choice but to simply select some unlabeled samples as negative ones, which may introduce noises and decrease the reliability of prediction. Herein, we propose a new method (TS-SVD) which only uses those known drug-protein, disease-protein and drug-disease interactions to predict the potential drug-disease associations. In a constructed drug-protein-disease heterogeneous network, assuming that drugs/diseases relating to some common proteins or diseases/drugs may be similar, we get the common neighbors count matrix of drugs/diseases, then convert it to a topological similarity matrix. After that, we get low dimensional embedding representations of drug-disease pairs by using topological features and singular value decomposition. Finally, a Random Forest classifier is trained to do the prediction. To train a more reasonable model, we select out some reliable negative samples based on the *k*-step neighbors relationships between drugs and diseases. Compared with some state-of-the-art methods, we use less information but achieve better or comparable performance. Meanwhile, our strategy for selecting reliable negative samples can improve the performances of these methods. Case studies have further shown the practicality of our method in discovering novel drug-disease associations.**

***Index Terms*— link prediction, drug reposition, common neighbors, topological similarity, negative sample selection.**

I. Introduction

LINK prediction aims to predict new links in the network or to identify links which exist but are not represented in the data. Link prediction is commonly used in both industrial applications and scientific researches, such as proposing friendships, recommending products, predicting outbreak of a disease, controlling privacy in networks, detecting spam emails, and so on. Recently, link prediction has also shown its application in virtual screening of drugs. Now that de novo drug discovery is not only time consuming and expensive, but also low success rate, more and more attention has been paid to discovering new indications for approved drugs in the market, that is, drug repositioning.

In the past years, there were many successful examples in drug repositioning. For instance, Minoxidil which was used to treat hypertension, was found by chance to have the potential to treat seborrheic hair loss [1]; Sildenafil which was marketed to treat cardiovascular diseases and had poor treatment effect, was accidentally found to have the treatment efficacy for erectile dysfunction [2]. However, this occasional way of drug repositioning was lack of guidance and had poor productivity, thus can not meet the needs for drug development.

Thanks to the great progress of biotechnologies, a large amount of drugs and diseases related data has being accumulated, which makes it possible to predict novel drug-disease links by computational methods. Drug repositioning is based on the commonly accepted assumption that similar drugs may have similar indications so as to treat similar diseases. Accordingly, many researches focus on how to measure the similarities between drugs and diseases, and how to build link prediction models based on known association pairs of drugs and diseases. For instance, Gottlieb *et al.* proposed a drug-disease link prediction method (PREDICT) [3], in which they first measured the drug-drug similarities by integrating chemical based, side effect based, sequence based, protein-protein interaction (PPI) network based and gene ontology (GO) based similarities, and measured the disease-disease similarities by assembling either two phenotype based similarities or three signature-based similarities; then they exploited the similarity measures to construct features for representing drug-disease pairs. Whereafter, a logistic regression classifier could be used to predict new drug-disease associations. Wang *et al.* proposed a three-layer heterogeneous network model (TL-HGBI) for both drug repositioning and target prediction tasks [4]. The network contains diseases, drugs and drug targets nodes. They calculated the drug similarities based on their chemical structures, target similarities based on the sequence information, and directly got the disease similarities from MinMiner [5]. By formulating the drug repositioning

Manuscript received April 18, 2020; accepted April 21, 2020. Date of publication April 24, 2020; date of current version July 1, 2020. This work was supported in part by the National Key Research and Development Project of China under Grant 2019YFA0904300 and Grant 1502-211100026, in part by the Major Projects of Technological Innovation in Hubei Province under Grant 2019AEA170, and in part by the Frontier Projects of Wuhan for Application Foundation under Grant 2019010701011381. *(Corresponding author: Juan Liu.)*

The authors are with the School of Computer Science, Research Institute of Artificial Intelligence, Wuhan University, Wuhan 430072, China (e-mail: liujuan@whu.edu.cn).

Digital Object Identifier 10.1109/TNB.2020.2990291

as a missing edge prediction problem on the heterogeneous network, they developed an iterative updating algorithm that propagates information across the network to solve the missing edge prediction problem. Martínez *et al.* built a network of interconnected drugs, proteins and diseases, based on which they proposed the heterogeneous network prioritization method (DrugNet) to perform drug repositioning [6]. Wu *et al.* presented a method to first hierarchically integrate the heterogeneous data of drugs, proteins and diseases into three layers, then integrate the drug and disease similarities from different layers. Based on the fused similarities, they proposed a semi-supervised graph cut method (SSGC) to predict new drug-disease links [7], [8]. Moghadam *et al.* proposed a computational method named scored mean kernel fusion (SMKF) to make high-level features by systematically combining multiple features related to drugs or diseases at drug-drug and drug-disease levels. Based on the features they constructed the prediction model of drug-disease links [9]. Liang *et al.* integrated drug chemical structures, protein domains and gene ontology terms to compute the similarities, and then proposed a sparse subspace learning method (LRSSL) to build the prediction model of drug-disease links [10]. Zhang *et al.* proposed to integrate known drug-disease interactions and several drug and disease features to predict drug-disease associations [11], [12]. Luo *et al.* integrated several drug and disease similarities and proposed a Bi-Random Walk based method (MBiRW) to identify novel drug-disease links [13]. Cui *et al.* adopted the commonly used drug structure similarity and disease semantics similarity and proposed a dual-network L2,1-collaborative matrix factorization method to predict the drug-disease interactions [14].

Obviously, most of the computational methods follow the similar outlines: developing approaches to measure drug and disease similarities, and constructing link prediction models. In order to calculate the similarities, most of the existing methods collect comprehensive information from various sources, such as the chemical structures, the side effects, the target sequences, the GO items, and so on, which makes it tedious to integrate the heterogenous data, or even hard due to incomplete data. Furthermore, most of the current computational methods are based on supervised learning to build prediction models, which need reliable training data containing both positive and negative samples. Nevertheless, the biological wet experiments can only tell whether there is a therapeutic relationship between drugs and diseases (positive), but not whether there is no therapeutic relationship between them (negative). Consequently, one of the main difficulties that the supervised learning based methods are facing when used for drug repositioning is lacking validated negative samples. To alleviate the problem, many methods randomly select unlabeled samples as the negative ones. Obviously, such strategy will inevitably introduce some potential positive drug-disease pairs as negative samples, which results in the noisy training data and low prediction performance. Instead of random selection, we proposed a simple strategy to choose negative samples from unlabeled drug-disease pairs in our former work [15] and got good results. However, more sophisticate strategy should be considered to get more reliable negative samples.

To address above issues, we first present a method that use neither biological nor chemical features of drugs or diseases to measure the similarities. Instead, we build a drug-protein-disease heterogeneous network by integrating the drug-disease, drug-protein and disease-protein interaction data. Then we propose a new similarity measure of drugs and diseases that can be calculated via the topological information of the network. And then we develop a new strategy of selecting negative samples based on the network, which promises the reliability of the selected samples. Finally, we build the drug-disease link prediction model and evaluate its performance.

II. Data Collection and Preproessing

Different with most of other researches, we only need to collect the known drug-protein, disease-protein and drug-disease associations which were used to build the model in this work. We obtained the drugs and drug-protein interaction data from DrugBank [16], a database with drugs and targets information; diseases and disease-gene (protein) interaction data from OMIM [17]; and drug-disease interaction data from Gottlieb's golden data set [3]. After excluding the polypeptides and drugs whose targets are not in human cells, we obtained 4642 drug-protein interactions of 1186 small molecule drugs and 1467 proteins; 1365 disease-protein associations of 449 diseases and 1467 proteins; and 1827 drug-disease associations of 302 diseases and 551 drugs. Based on the drug-protein, disease-protein and drug-disease pairs, we can construct a drug-protein-disease heterogeneous network with drugs, proteins and diseases as the nodes.

Although no chemical or biological information related to drugs and diseases is needed in our method, we still collected the chemical structure data of drugs from DrugBank and the protein sequence data in FASTA format from UniProt [18] so that we could implement the representative methods for comparing with our method. Just the same as other methods, we calculated the drug-drug chemical similarities via Openbabel tool [19] based on the SMILES strings [20], the protein-protein similarities by using Smith-Waterman algorithm [21]. We also obtained the disease-disease phenotype similarities data directly from MimMiner [5].

III. Methods

A. Framework of our Method

The framework of our proposed method, which we call as TS-SVD, is shown in Fig. 1. First, we constructed a drug-protein-disease heterogeneous network according to the drug-protein, disease-protein, and drug-disease interaction data. Then we generated the matrix with the common neighbor numbers of the drugs as the elements, named as common neighbors count matrix. Obviously, the matrix is symmetric and each row or column represents a common neighbor profile of a drug with other drugs. It is very likely that two drugs with similar profiles have similar indications, and the more similar profiles the more similar indications. Therefore, we then generated the topological similarity matrix for drugs by calculating the cosine similarities of drug profiles. By the same way, we also generated the topological similarity matrix

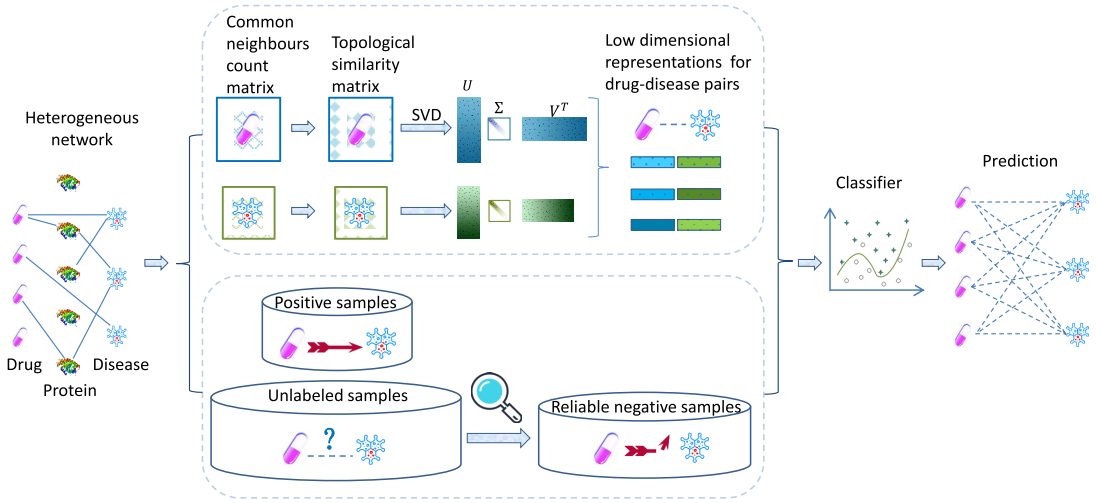

Fig. 1. The framework of our proposed TS-SVD.

for diseases. And then we used the similarities of drugs and diseases as the feature vectors to represent the drug-disease pairs. Since the dimension of the feature vector is high, we then used singular value decomposition (SVD) to do dimension reduction. Finally, we used the Random Forest (RF) algorithm to build the classifier for drug-disease link prediction.

Same with other biological data, all the drug-disease pairs consist of positive (validated treatment relations) and unlabeled (no evidence) instanes. On one hand, we need the classifier to identify whether the unlabeled pairs are positive or negative (having no treatment relations); on the other hand, we should use some of the unlabeled data as negative samples to build the classifier. That is a dilemma. In this work, in order to obtain more reliable classifier, we also presented a strategy to choose more likely negative samples from the unlabeled data.

B. Common Neighbors Count Matrices

Using the drug-protein, disease-protein and drug-disease pairs, we constructed a drug-protein-disease heterogeneous network with drugs, proteins and drugs as the nodes, and the known interactions as the edges. It should be noted that we did not consider edges between drug, protein or disease nodes in the network for simplicity. Assuming that in this network, if two drug or disease nodes show similar topological features, they may play similar roles. Based on this assumption, we used the common neighbors count information as the basis of the similarity measure.

Let $N(.)$ denote the set of all neighbors of a node in a network, we calculate the common neighbors count of node u and v as:

$$C(u, v) = |N(u) \cap N(v)| \tag{1}$$

where $|N|$ is the cardinality of the set N.

Obviously, $C(u, v)$ means the number of paths with length 2 bridging node u and node v. Suppose A is the adjacency matrix of a homogeneous network, it can be easily seen that $C(u, v) = A^2(u, v)$.

Now that the constructed drug-protein-disease network is a heterogeneous one with three different types of nodes, and we focus on finding the potential drug-disease interactions, we extend the calculation of common neighbors count for drug and disease nodes as following.

Let A_{ds}, A_{dp} and A_{sp} respectively denote the drug-disease, drug-protein and disease-protein interaction matrices. For drug nodes, only considering the neighbor disease nodes, we can get a common disease neighbors count matrix ${C_1}^{drug} = A_{ds} \times (A_{ds})^T$; and only considering the neighbor protein nodes, we can get a common protein neighbors count matrix ${C_2}^{drug} = A_{dp} \times (A_{dp})^T$. Accordingly, we can get a comprehensive common neighbors count matrix for drug nodes as $C^{drug} = {C_1}^{drug} + {C_2}^{drug}$. Similarly, we can get a comprehensive common neighbors count matrix for disease nodes as $C^{disease} = {C_1}^{disease} + {C_2}^{disease}$, where $C_1^{disease} = (A_{ds})^T \times A_{ds}$ and $C_2^{disease} = A_{sp} \times (A_{sp})^T$. It is obvious that element $C^{drug}(i, j)$ denotes the number of paths with length 2 bridging drug node i and j, and element $C^{disease}(i, j)$ denotes the number of paths with length 2 bridging disease node i and j. The procedure for calculating common neighbors count matrices in a drug-protein-disease heterogeneous network is shown in Algorithm 1. And it can be easily modified or extended to other application fields.

C. Topological Similarities

We have obtained the common neighbors count matrices C^{drug} for drugs and $C^{disease}$ for diseases, which imply some topological relationship among the drug or disease nodes. However, it is not intuitive enough to see how similar these drugs or diseases are. Therefore, we will transform the common neighbors count matrices to topological similarity matrices in the following.

According to the definition, each element $C(i, j)$ in a common neighbors count matrix C denotes the number of common neighbors of node i and j. Then we can define the

Algorithm 1 CommonHeterogeneousNeighbors

Input:
Drug-disease interaction matrix $A_{ds} \in \mathbb{R}^{m\times n}$;
Drug-protein interaction matrix $A_{dp} \in \mathbb{R}^{m\times p}$;
Disease-protein interaction matrix $A_{sp} \in \mathbb{R}^{n\times p}$.

Output:
Common neighbors count matrix of drugs $C^{drug} \in \mathbb{R}^{m\times m}$;
Common neighbors count matrix of diseases $C^{disease} \in \mathbb{R}^{n\times n}$.

$C_1^{drug} = A_{ds} \times (A_{ds})^T$;
$C_2^{drug} = A_{dp} \times (A_{dp})^T$;
$C^{drug} = C_1^{drug} + C_2^{drug}$;
$C_1^{disease} = (A_{ds})^T \times A_{ds}$;
$C_2^{disease} = A_{sp} \times (A_{sp})^T$;
$C^{disease} = C_1^{disease} + C_2^{disease}$;
return C^{drug}, $C^{disease}$;

i-th row as the topology profile of node i, or the j-th column as the topology profile of node j. Obviously, the topology profile represents the common neighbors count variation of a node with other nodes in the network. In this work, we adopt the row vectors of the matrix as profile denotations.

Let $x = (x_1, x_2, \ldots .x_n)$ and $y = (y_1, y_2, \ldots .y_n)$ be the topology profiles of two nodes, we can calculate their topological similarity according to the cosine measure:

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

Since x_i and y_i are the number of common neighbors which are integers greater than or equal to 0, the topological similarity value of them is in range [0,1]. The larger the more similar.

Therefore, we can obtain topological similarity matrices S^{drug} and $S^{disease}$ for drug nodes and disease nodes respectively based on common neighbors count matrices C^{drug} and $C^{disease}$. Each element of the matrices represents the topological similarity between two drug or disease nodes in the network, and each row represents the similarity profile of a drug or disease node.

D. SVD Based Dimension Reduction

After getting the similarity profile of each drug/disease node, we can follow the same methodology with some "state-of-the-art" methods to represent every drug-disease pair as the sample so that we can build a prediction model by using machine learning strategies. We use the similarity profile as the feature vector to represent every drug/disease, and for a drug-disease pair, we just represent it by concatenating the corresponding feature vectors of the drug and the disease. Let m be the number of drug nodes, and n be the number of disease nodes, then the dimension of the feature vector for a drug-disease pair would be $m + n$.

However, owing to the large amount of drugs and diseases in the data, the drug-disease representation would be high dimensional, which will increase the complexity of the model and result in poor generalization ability. Therefore, dimension reduction is necessary before building the classifier. Singular Value Decomposition (SVD) has been widely used for dimensionality reduction thanking to its simplicity and effectiveness. We also adopt SVD to do the task in this work.

By SVD, a matrix $M \in \mathbb{R}^{m\times n}$ will be decomposed into the product of three matrices:

$$M = U\Sigma V^T \tag{3}$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthonormal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix of singular values.

The singular values in Σ are ordered in descending order. In most cases, the sum of squares of the first 10% of the singular values is over 90% of the total sum of squares of all singular values. If we choose the first r singular values to form a new Σ, and use the first r columns of U and V, we can approximate the original matrix M as:

$$M_{m\times n} \approx U_{m\times r}\Sigma_{r\times r}{V^T}_{r\times n} \tag{4}$$

where $r \ll min(m, n)$.

That is to say, we can factorize M as

$$M_{m\times n} \approx P_{m\times r}Q_{r\times n} \tag{5}$$

where

$$P = U_{m\times r}\sqrt{\Sigma}_{r\times r}, \quad Q = \sqrt{\Sigma}_{r\times r}{V^T}_{r\times n} \tag{6}$$

Thus, PQ is a matrix with rank r that best approximates the original matrix M. If M is a matrix related to features of a graph, in the view of graph representation, each row of $P = U\sqrt{\Sigma}$ or $Q^T = V\sqrt{\Sigma}$ can be used as the feature vector to represent the corresponding node in the graph [22], [23]. That means, if we use SVD to decompose S^{drug} and $S^{disease}$ respectively, we can use rows in either $U\sqrt{\Sigma}$ or $V\sqrt{\Sigma}$ to represent the drug or disease nodes.

Concretely, the topological similarity matrix of drugs, $S^{drug} \in \mathbb{R}^{m\times m}$, can be approximately factorized as

$$\begin{aligned} S^{drug} &\approx P^{drug}Q^{drug} \\ &= U^{drug}\sqrt{\Sigma^{drug}}\sqrt{\Sigma^{drug}}(V^{drug})^T \end{aligned} \tag{7}$$

where $P^{drug} \in \mathbb{R}^{m\times r_1}$, $Q^{drug} \in \mathbb{R}^{r_1\times m}$, $U^{drug} \in \mathbb{R}^{m\times r_1}$, $\Sigma^{drug} \in \mathbb{R}^{r_1\times r_1}$, $V^{drug} \in \mathbb{R}^{m\times r_1}$, $r_1 \ll m$.

At the same time, the topological similarity matrix of diseases, $S^{disease} \in \mathbb{R}^{n\times n}$, can be approximately factorized as

$$\begin{aligned} S^{disease} &\approx P^{disease}Q^{disease} \\ &= U^{disease}\sqrt{\Sigma^{disease}}\sqrt{\Sigma^{disease}}(V^{disease})^T \end{aligned} \tag{8}$$

where $P^{disease} \in \mathbb{R}^{n\times r_2}$, $Q^{disease} \in \mathbb{R}^{r_2\times n}$, $U^{disease} \in \mathbb{R}^{n\times r_2}$, $\Sigma^{disease} \in \mathbb{R}^{r_2\times r_2}$, $V^{disease} \in \mathbb{R}^{n\times r_2}$, $r_2 \ll n$.

Thus, we obtain the r_1-dimensional representations of drug nodes, $F^{drug} = U^{drug}\sqrt{\Sigma^{drug}}$ and the r_2-dimensional representations of disease nodes, $F^{disease} = U^{disease}\sqrt{\Sigma^{disease}}$. In the end, the dimension of the feature vector for each drug-disease pair is reduced from $m + n$ to $(r_1 + r_2)$.

To control the value of r_1 and r_2, we use a parameter $feature\_percent$ ranging from 0 to 1 in this work, such that $r_1 = feature\_percent \times m$ and $r_2 = feature\_percent \times n$.

E. Strategy of Selecting Reliable Negative Samples

In order to build the classification model via supervised learning, both positive and negative samples are needed. However, just as most of other biological data, only small part of the drug-disease pairs are confirmed by experiments to have treatment relations and can be labeled as positive, while the rest pairs remain unlabeled. Some of the unlabeled samples may be positive cases that have not yet been experimentally confirmed, while others may be negative cases that do not have a therapeutic relationship. So in this work we try to find the reliable negative samples from the unlabeled ones to build a reasonable prediction model.

We first introduce two definitions which are used in this paper.

Definition 1 (k-Step Neighbor): Given a graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. If there is a path with length k between vertex v_i and v_j, then v_j is a k-step neighbor of v_i, and v_i is also a k-step neighbor of v_j, where $v_i \in V$ and $v_j \in V$.

Definition 2 (k-Step Commuting Matrix): Given a heterogeneous graph $G = (V, E)$, where V is the set of vertices respectively belong to type T_x, $T_y, \ldots$, and E is the set of edges, a k-step commuting matrix D between type T_x and T_y is a matrix whose element $D(i, j)$ is the number of paths with length k between vertex $x_i \in T_x$ and $y_j \in T_y$.

Specifically, in a drug-protein-disease heterogeneous network, a k-step commuting matrix D between type $Drug$ and type $Disease$ is a matrix whose element $D(i, j)$ is the number of paths with length k between vertex $d_i \in Drug$ and vertex $s_j \in Disease$. And the element $D(i, j)$ can be obtained from the product of k association matrices.

Let A_{ds} denote the drug-disease interaction matrix, A_{dp} denote the drug-protein interaction matrix, and A_{sp} denote the disease-protein interaction matrix. We will discuss the k-step commuting matrix between drugs and diseases with $k = 1$, 2 and 3 as follows:

If $k = 1$, we consider all the path instances following the path schema "$Drug - Disease$". For example, a path instance "$d_4 - s_3$" (d_4 is a drug and s_3 is a disease) means d_4 and s_3 are 1-step neighbors to each other. The 1-step commuting matrix between drugs and diseases, denoted as $D1$, is just the existing drug-disease interaction matrix. We have

$$D1 = A_{ds} \tag{9}$$

If $k = 2$, we consider all the path instances following the path schema "$Drug - Protein - Disease$". For example, a path instance "$d_2 - p_3 - s_1$" (d_2 is a drug, s_1 is a disease, and p_3 is a protein) means d_2 and s_1 are 2-step neighbors to each other. The 2-step commuting matrix between drugs and diseases, named $D2$, is calculated by

$$D2 = A_{dp} \times A_{sp}^T \tag{10}$$

It means that if a drug and a disease share some proteins, they may have potential treatment relations.

If $k = 3$, there are three cases:

Case 1: We consider all the path instances following the path schema "$Drug - Protein - Drug - Disease$". For example, a path instance "$d_1 - p_3 - d_2 - s_2$" (d_1 and d_2 are drugs, s_2 is a disease, and p_3 is a protein) indicates d_1 and s_2 are 3-step neighbors to each other. The 3-step commuting matrix between drugs and diseases, named $D31$, is calculated by

$$D31 = A_{dp} \times A_{dp}^T \times A_{ds} \tag{11}$$

It means that if drug d_1 and d_2 target to some common proteins, indications of drug d_2 might also be indications of drug d_1.

Case 2: We consider all the path instances following the path schema "$Drug - Disease - Drug - Disease$". For example, a path instance "$d_1 - s_1 - d_2 - s_2$" (d_1 and d_2 are drugs, s_1 and s_2 are diseases) shows d_1 and s_2 are 3-step neighbors to each other. The 3-step commuting matrix between drugs and diseases, named $D32$, is calculated by

$$D32 = A_{ds} \times A_{ds}^T \times A_{ds} \tag{12}$$

It means that if drug d_1 and d_2 have some common known indications, other indications of drug d_2 might also be potential indications of drug d_1.

Case 3: We consider all the path instances following the path schema "$Drug - Disease - Protein - Disease$". For example, a path instance "$d_1 - s_1 - p_3 - s_2$" (d_1 is a drug, s_1 and s_2 are diseases, and p_3 is a protein) means d_1 and s_2 are 3-step neighbors to each other. The 3-step commuting matrix between drugs and diseases, named $D33$, is calculated by

$$D33 = A_{ds} \times A_{sp} \times A_{sp}^T \tag{13}$$

It means that if disease s_1 and s_2 share some common target proteins, drugs used to treat disease s_1 might also be used to treat disease s_2.

The sum of all the k-step commuting matrices ($k = 1, 2, 3$) between drugs and diseases, denoted as D, can be obtained by

$$D = D1 + D2 + D31 + D32 + D33 \tag{14}$$

Elements in $D1$ are the direct true drug-disease associations. Elements in $D2$, $D31$, $D32$ and $D33$ are numbers of indirect associations between drugs and diseases meaning that these drug-disease pairs have a certain probability p to be treatment relations, where $0 \leqslant p \leqslant 1$.

Since matrix D reflects all the 1-step, 2-step and 3-step links from drugs to diseases, a zero element in matrix D means that the corresponding drug-disease pair either have links longer than 3 steps or have no link and the corresponding drug-disease pair is more likely to be negative.

In this way, we can pick out some more reliable negative ones from the unlabeled samples. The idea can be summarized as Algorithm 2.

With the positive samples confirmed as true treatment interactions and the selected reliable negative samples, we then train the Random Forest classifier which is implemented by the $RandomForestClassifier$ function in the scikit-learn package, and can be used for the prediction of the drug-disease treatment relations later.

Algorithm 2 selectReliableNegativeSamples

Input:
Drug-disease interaction matrix $A_{ds} \in \mathbb{R}^{m\times n}$;
Drug-protein interaction matrix $A_{dp} \in \mathbb{R}^{m\times p}$;
Disease-protein interaction matrix $A_{sp} \in \mathbb{R}^{n\times p}$.
Output:
Selected negative samples set $reliableNeg$.

$D1 = A_{ds}$;
$D2 = A_{dp} \times A_{sp}^T$;
$D31 = A_{dp} \times A_{dp}^T \times A_{ds}$;
$D32 = A_{ds} \times A_{ds}^T \times A_{ds}$;
$D33 = A_{ds} \times A_{sp} \times A_{sp}^T$;
$D = D1 + D2 + D31 + D32 + D33$;
$reliableNeg \leftarrow \emptyset$;
for $i = 1$ to m **do**
  for $j = 1$ to n **do**
    if $D(i, j) = 0$ **then**
      $reliableNeg \leftarrow reliableNeg \cup (i, j)\ pair$;
    end if
  end for
end for
return $reliableNeg$;

IV. Results and Discussions

In this section, first we will introduce the evaluation metrics used in the experiments. Then we show the results of experiments carried out in three aspects: (1) We examined how the performance of TS-SVD varied with the parameter $feature\_percent$ and determined the proper parameter setting. (2) We applied different negative samples selection strategies on state-of-the-art methods to investigate whether our proposed negative samples selection approach could benefit the performance of classifiers. (3) We tested the practicality of our method in discovering novel drug-disease associations by literature investigation.

A. Evaluation Metrics

In traditional classification problems, these metrics such as Precision (PRE), Recall (REC), Accuracy (ACC), Matthews Correlation Coefficient (MCC) and F_1 score (F_1) are often used. They can be calculated by:

$$PRE = \frac{TP}{TP + FP} \tag{15}$$

$$REC = \frac{TP}{TP + FN} \tag{16}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{18}$$

$$F_1 = \frac{2 \times PRE \times REC}{PRE + REC} \tag{19}$$

where TP, FP, TN and FN are the number of true positive samples, false positive samples, true negative samples and false negative samples respectively.

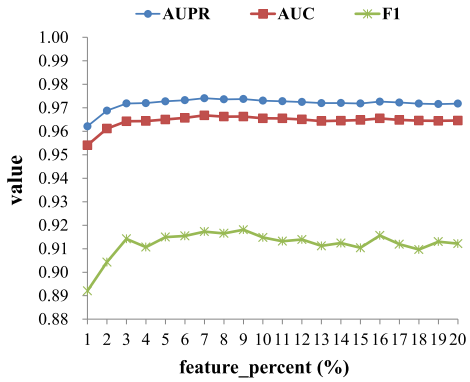

Fig. 2. Influence of different *feature_percent* values on *AUPR*, *AUC* and F_1.

F_1 is a comprehensive metric, as it is the harmonic average of PRE and REC. Another two metrics, $AUPR$ (Area under Precision-Recall curve) and AUC (Area under ROC curve) are also often used. $AUPR$ reflects both Precision and Recall, AUC reveals both the True Positive Rate(TPR, the same as REC) and the False Positive Rate ($FPR = \frac{FP}{FP+TN}$), so they are also comprehensive metrics. Therefore, in this work, we adopted the comprehensive metrics $AUPR$, AUC and F_1 as the main metrics.

B. Influence of Parameter Value on Performance

In TS-SVD, we used a parameter $feature\_percent$ to control the value of r_1 and r_2, so the number of reduced features representing drug-disease pair can be determined. In order to examine how different parameter value settings influenced the performance of our method, we separately set the value of $feature\_percent$ from 1% to 20% by step 1% and the performance is shown in Fig. 2. When the parameter value is very small (say no greater than 3%), the performance improves with the increase of parameter value, meaning that too few features may lose a lot of useful information for classification which can be included by adding additional features. As the parameter value continues to increase, the classification performance does not increase or decrease significantly with the parameter value, which illustrates the robustness of our method. Now that after SVD the sum of squares of the first 10% of the singular values is in most cases over 90% of the total sum of squares of all singular values, we set $feature\_percent$ value as 9% in this work according to the curves in Fig. 2.

C. Comparison of Different Negative Samples Selection Strategies

To investigate the effectiveness of our proposed strategy of selecting negative samples described in Algorithm 2, we adopted three kinds of approaches to select negative samples then together with the same positive samples to train and test TS-SVD and the following representative drug-disease classification models: PREDICT [3], TL-HGBI [4], MBiRW [13], LRSSL [10], SCMFDD [11] and EMP-SVD [15].

TABLE I
PERFORMANCES COMPARISON WITH DIFFERENT NEGATIVE SAMPLES SELECTING STRATEGIES

Negative Samples	Methods	AUPR	AUC	PRE	REC	ACC	MCC	F1
randomly selected unlabeled samples	TS-SVD	**0.957**	***0.949***	0.871	***0.899***	0.884	**0.769**	0.884
	PREDICT	0.902	0.891	0.821	0.825	0.820	0.642	0.822
	TL-HGBI	0.849	0.843	0.842	0.735	0.766	0.540	0.783
	LRSSL	0.879	0.860	0.880	0.718	0.766	0.546	0.790
	SCMFDD	0.849	0.864	0.916	0.724	0.782	0.585	0.809
	MBiRW	0.952	0.943	0.876	***0.899***	**0.887**	0.775	**0.887**
	EMP-SVD	0.954	***0.949***	**0.924**	0.837	0.871	0.745	0.878
reliable negative samples in EMP-SVD	TS-SVD	**0.959**	***0.951***	0.890	0.892	**0.887**	**0.777**	**0.889**
	PREDICT	0.908	0.895	0.809	0.850	0.830	0.662	0.828
	TL-HGBI	0.852	0.846	0.829	0.750	0.774	0.552	0.787
	LRSSL	0.881	0.861	0.864	0.732	0.770	0.553	0.790
	SCMFDD	0.836	0.854	**0.926**	0.713	0.774	0.575	0.805
	MBiRW	0.952	0.942	0.867	**0.901**	0.884	0.769	0.884
	EMP-SVD	0.956	***0.951***	0.913	0.854	0.876	0.755	0.882
reliable negative samples in TS-SVD	TS-SVD	**0.974**	***0.966***	0.899	***0.939***	**0.919**	**0.840**	**0.918**
	PREDICT	0.923	0.907	0.836	0.840	0.836	0.673	0.838
	TL-HGBI	0.882	0.870	0.839	0.776	0.797	0.597	0.806
	LRSSL	0.900	0.879	0.841	0.782	0.796	0.603	0.805
	SCMFDD	0.858	0.874	0.930	0.733	0.794	0.611	0.819
	MBiRW	0.964	0.954	0.873	***0.939***	0.906	0.815	0.904
	EMP-SVD	0.970	***0.966***	**0.935**	0.874	0.899	0.800	0.903

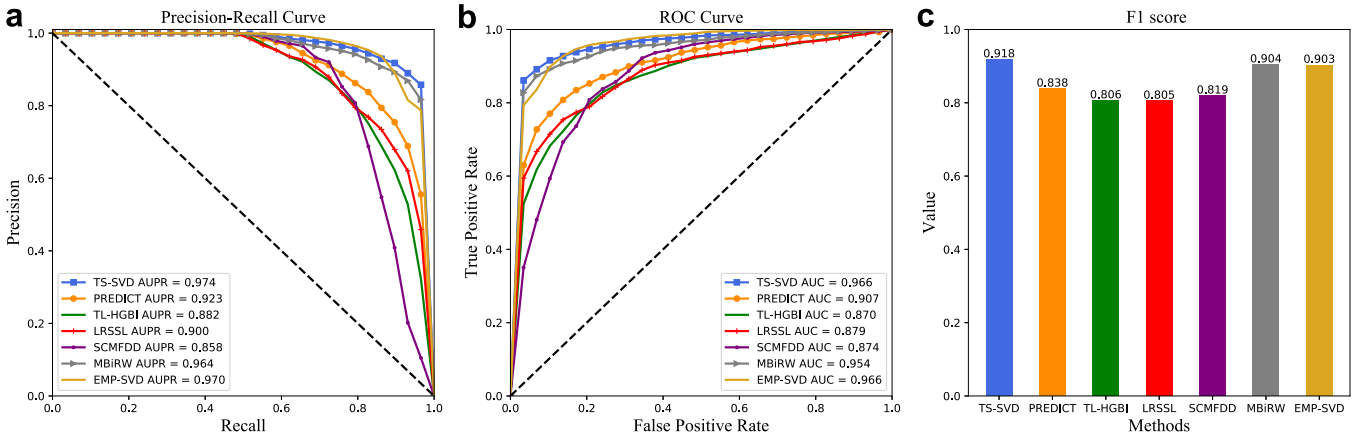

Fig. 3. (a) Precision-recall curve (b) ROC curve (c) F_1 score of TS-SVD and compared methods (all methods use reliable negative samples in TS-SVD).

Specifically, we respectively selected negative samples by (1) choosing randomly from unlabeled ones; (2) using method proposed in EMP-SVD and (3) using method in this work. Each of them together with the same set of positive samples formed the training data to perform 5-fold cross-validation on above classifiers. To be fair, five parts of data were kept the same partition in each method when performing the cross-validation.

As shown in Table I, when using method in EMP-SVD, most of these methods achieved slightly better performances in comprehensive metrics ($AUPR$, AUC and F_1) than randomly selected negative samples. When using method in TS-SVD, the performance are much more better. It means that the strategy proposed in this work is more effective, and these reliable negative samples are more beneficial to building discriminative models.

D. Comparison With Other Methods

Table I also illustrates that the proposed TS-SVD achieves better or comparable performance than other methods in most metrics, especially in comprehensive metrics ($AUPR$, AUC and F_1) when using the same kinds of training data. To be more intuitive, we respectively plot the Precision-Recall Curve, ROC Curve and F_1 score (all methods use reliable negative samples in TS-SVD) in Fig. 3a, b and c. Obviously, TS-SVD performs better than the compared methods. It should be noted that the comparisons are performed by using the same negative samples and the same experiment settings. If TS-SVD compares with other methods with their original strategies to select negative samples, the superiority of TS-SVD is more obvious.

What is more, TS-SVD only makes use of drug-protein, disease-protein and drug-disease information to build the model, which is very different than most of the other methods that usually need additional biological or chemical information related to drugs or diseases. The comparison results illustrate the advantages of the simple yet powerful topological similarity metric, the SVD based feature reduction and representation method, and the effective negative sample selecting strategy proposed in this work.

TABLE II
THE PREDICTED DRUG-DISEASE ASSOCIATIONS (TOP 10)

Rank	Score	DrugBank ID	Drug Name	OMIM ID	Disease Name	Literature
1	0.996	DB00509	Dextrothyroxine	138800	Goiter, multinodular 1, with or without Sertoli-Leydig cell tumors; MNG1	[24]
2	0.992	DB00509	Dextrothyroxine	255900	Myxedema	[25]
3	0.992	DB00880	Chlorothiazide	208300	Ascites, chylous	
4	0.992	DB01202	Levetiracetam	267740	Retinal degeneration and epilepsy	
5	0.992	DB01202	Levetiracetam	208700	Ataxia with myoclonic epilepsy and presenile dementia	[26]
6	0.992	DB00214	Torasemide	256370	Nephrotic syndrome, type 4; NPHS4	
7	0.992	DB01210	Levobunolol	137760	Glaucoma, primary open angle; POAG	[27]
8	0.992	DB01210	Levobunolol	601682	Glaucoma 1, primary open angle, C; GLC1C	
9	0.989	DB00763	Methimazole	138800	Goiter, multinodular 1, with or without Sertoli-Leydig cell tumors; MNG1	[28]
10	0.989	DB00550	Propylthiouracil	138800	Goiter, multinodular 1, with or without Sertoli-Leydig cell tumors; MNG1	[29]

TABLE III
TOP 10 PREDICTIONS FOR THE DISEASE "NON-SMALL CELL LUNG CANCER"

Rank	Score	DrugBank ID	Drug Name	Literature
1	0.801	DB00563	Methotrexate	[30]
2	0.781	DB00515	Cisplatin	[31]
3	0.707	DB00541	Vincristine	[32]
4	0.707	DB00773	Etoposide	[33]
5	0.703	DB00762	Irinotecan	[34]
6	0.691	DB00444	Teniposide	[35]
7	0.688	DB00262	Carmustine	
8	0.680	DB00958	Carboplatin	[36]
9	0.676	DB00987	Cytarabine	
10	0.660	DB00851	Dacarbazine	

It is mentionable that our previously proposed EMP-SVD [15] also only needs drug-protein, disease-protein and drug-disease associations to construct a heterogenous network so as to build the prediction model, which is somewhat similar to TS-SVD. However, these two methods are actually very different in the way of utilizing the network information to construct the final classifier, the drug-disease pair representation and the selecting strategy of negative samples. Compared to EMP-SVD, the idea of TS-SVD is much more intuitive and easier to understand, and the condition to select negative samples is stricter which promises the selected negative samples being more reliable.

E. Literature Investigation of Predicted Interactions

To test the usefulness of our TS-SVD in predicting unknown drug-disease associations, we first used all 1827 known drug-disease associations in the gold standard data set as positive samples and equal number of reliable negative samples to compose a training set to train the prediction model; and then we used the trained model to make predictions on other unknown drug-disease pairs; finally we validated the new predicted results by literature investigation.

The top 10 predicted drug-disease associations are listed in Table II, and 6 of them were validated in literatures. More over, we also checked the predicted drugs associated with Non-small Cell Lung Cancer (OMIM ID: 211980) by our work, shown in Table III. After literature investigation, we found that in the top 10 predicted drugs for treating Non-small Cell Lung Cancer, 7 of them were confirmed to be used in the treatment or clinical trials. Above drug-disease associations are not in the original data set, but can be successfully predicted by our method, demonstrating that TS-SVD is a practical tool in discovering potential unknown drug-disease associations.

V. CONCLUSIONS

In this work, we proposed a method named TS-SVD to predict the potential drug-disease associations. Different with most existing methods, we only made use of known interaction information instead of various chemical or biological data which may be unknown or hard to collect. We first built a heterogeneous network composed of drugs, proteins and diseases. Based on the assumption that drugs/diseases related to some common proteins or diseases/drugs may be more similar, we got the common neighbors count matrix of drugs/diseases, then converted to topological similarity matrices. After that, we decomposed the matrices by SVD and got low dimensional embedding representations of drug-disease pairs. Finally, we selected reliable negative samples and constructed the Random Forest classifier to predict the potential drug-disease associations. Compared with some state-of-the-art methods, we used less information but achieved better or comparable performance. Compared with our previous work EMP-SVD, TS-SVD is more intuitive and easier to be understood and applied. Literature investigation has further shown the usefulness of our method.

ACKNOWLEDGMENT

The authors thank F. Huang of Wuhan University for his help in running a part of the comparison experiments.

REFERENCES

[1] S. Varothai and W. F. Bergfeld, "Androgenetic alopecia: An evidence-based treatment update," *Amer. J. Clin. Dermatol.*, vol. 15, no. 3, pp. 217–230, Jul. 2014.
[2] N. Novac, "Challenges and opportunities of drug repositioning," *Trends Pharmacological Sci.*, vol. 34, no. 5, pp. 267–272, May 2013.
[3] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 496, Jan. 2011.
[4] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, Oct. 2014.
[5] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Human Genet.*, vol. 14, no. 5, pp. 535–542, May 2006.

[6] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 41–49, Jan. 2015.

[7] G. Wu, J. Liu, and C. Wang, "Semi-supervised graph cut algorithm for drug repositioning by integrating drug, disease and genomic associations," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 223–228.

[8] G. Wu, J. Liu, and C. Wang, "Predicting drug-disease interactions by semi-supervised graph cut algorithm and three-layer data integration," *BMC Med. Genomics*, vol. 10, no. S5, p. 79, Dec. 2017.

[9] H. Moghadam, M. Rahgozar, and S. Gharaghani, "Scoring multiple features to predict drug disease associations using information fusion and aggregation," *SAR QSAR Environ. Res.*, vol. 27, no. 8, pp. 609–628, Aug. 2016.

[10] X. Liang *et al.*, "LRSSL: Predict and interpret drug–disease associations based on data integration using sparse subspace learning," *Bioinformatics*, vol. 33, no. 8, pp. 1187–1196, Apr. 2017.

[11] W. Zhang *et al.*, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, no. 1, p. 233, Dec. 2018.

[12] W. Zhang, X. Yue, F. Huang, R. Liu, Y. Chen, and C. Ruan, "Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network," *Methods*, vol. 145, pp. 51–59, Aug. 2018.

[13] H. Luo *et al.*, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, Sep. 2016.

[14] Z. Cui, Y.-L. Gao, J.-X. Liu, J. Wang, J. Shang, and L.-Y. Dai, "The computational prediction of drug-disease interactions using the dual-network l2,1-CMF method," *BMC Bioinf.*, vol. 20, no. 1, p. 5, Dec. 2019.

[15] G. Wu, J. Liu, and X. Yue, "Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition," *BMC Bioinf.*, vol. 20, no. S3, p. 134, Mar. 2019.

[16] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018.

[17] A. Hamosh, "Online Mendelian Inheritance In Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 52–55, Jan. 2002.

[18] T. UniProt Consortium, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019.

[19] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J. Cheminform.*, vol. 3, no. 1, p. 33, Dec. 2011.

[20] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Model.*, vol. 28, no. 1, pp. 31–36, Feb. 1988.

[21] T. F. Smith, M. S. Waterman, and C. Burks, "The statistical distribution of nucleic acid similarities," *Nucleic Acids Res.*, vol. 13, no. 2, pp. 645–656, 1985.

[22] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD," *Behav. Res. Methods*, vol. 44, no. 3, pp. 890–907, Sep. 2012.

[23] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM, 2015, pp. 891–900.

[24] H. Krawczynska, K. Wojcik-Musialek, and R. Illig, "10 years of successful treatment with dextrothyroxine in a girl with TSH-induced hyperthyroidism," *Hormone Res.*, vol. 35, no. 5, pp. 213–216, 1991.

[25] P. Starr, "Depression of the serum cholesterol level in myxedematous patients by an oral dosage of sodium dextrothyroxine which has no effect on the basal metabolic rate or electrocardiogram," *J. Clin. Endocrinology Metabolism*, vol. 20, no. 1, pp. 116–119, Jan. 1960.

[26] C. Hommet, K. Mondon, V. Camus, B. De Toffol, and T. Constans, "Epilepsy and dementia in the elderly," *Dementia Geriatric Cognit. Disorders*, vol. 25, no. 4, pp. 293–300, 2008.

[27] T. Li *et al.*, "Comparative effectiveness of first-line medications for primary open-angle glaucoma: A systematic review and network meta-analysis," *Ophthalmology*, vol. 123, no. 1, pp. 129–140, 2016.

[28] J. Bednarek, H. Wysocki, and J. Sowinski, "Oxidation products and antioxidant markers in plasma of patients with Graves' disease and toxic multinodular goiter: Effect of methimazole treatment," *Free Radical Res.*, vol. 38, no. 6, pp. 659–664, Jun. 2004.

[29] K. Hughes and C. Eastman, "Goitre: Causes, investigation and management," *Austral. Family Physician*, vol. 41, no. 8, p. 572, 2012.

[30] L. H. Einhorn, S. D. Williams, E. E. Stevens, W. H. Bond, and L. Chenoweth, "Random prospective study cyclophosphamide, doxorubicin, and methotrexate (CAM) combination chemotherapy versus single-agent sequential chemotherapy in non small cell lung cancer," *Cancer Treat. Rep.*, vol. 66, no. 12, pp. 2005–2011, 1982.

[31] D. A. Fennell *et al.*, "Cisplatin in the modern era: The backbone of first-line chemotherapy for non-small cell lung cancer," *Cancer Treatment Rev.*, vol. 44, pp. 42–50, Mar. 2016.

[32] A. Y.-C. Chang *et al.*, "Phase II evaluation of a combination of mitomycin C, vincristine, and cisplatin in advanced non-small cell lung cancer," *Cancer*, vol. 57, no. 1, pp. 54–59, 1986.

[33] R. Zhu *et al.*, "PH sensitive nano layered double hydroxides reduce the hematotoxicity and enhance the anticancer efficacy of etoposide on non-small cell lung cancer," *Acta Biomaterialia*, vol. 29, pp. 320–332, Jan. 2016.

[34] S. Negoro *et al.*, "Randomised phase III trial of irinotecan combined with cisplatin for advanced non-small-cell lung cancer," *Brit. J. Cancer*, vol. 88, no. 3, pp. 335–341, Feb. 2003.

[35] G. Giaccone, M. Donadio, G. Bonardi, F. Testore, and A. Calciati, "Teniposide in the treatment of small-cell lung cancer: The influence of prior chemotherapy," *J. Clin. Oncol.*, vol. 6, no. 8, pp. 1264–1270, Aug. 1988.

[36] R. S. Herbst *et al.*, "Gefitinib in combination with paclitaxel and carboplatin in advanced non–small-cell lung cancer: A phase III trial—Intact 2," *J. Clin. Oncol.*, vol. 22, no. 5, pp. 785–794, 2004.