Partner-Specific Drug Repositioning Approach Based on Graph Convolutional Network

Xinliang Sun, Bei Wang, Jie Zhang, and Min Li, *Member, IEEE*

***Abstract*—Drug repositioning identifies novel therapeutic potentials for existing drugs and is considered an attractive approach due to the opportunity for reduced development timelines and overall costs. Prior computational methods usually learned a drug's representation from an entire graph of drug-disease associations. Therefore, the representation of learned drugs representation are static and agnostic to various diseases. However, for different diseases, a drug's mechanism of actions (MoAs) are different. The relevant context information should be differentiated for the same drug to target different diseases. Computational methods are thus required to learn different representations corresponding to different drug-disease associations for the given drug. In view of this, we propose an end-to-end partner-specific drug repositioning approach based on graph convolutional network, named PSGCN. PSGCN firstly extracts specific context information around drug-disease pairs from an entire graph of drug-disease associations. Then, it implements a graph convolutional network on the extracted graph to learn partner-specific graph representation. As the different layers of graph convolutional network contribute differently to the representation of the partner-specific graph, we design a layer self-attention mechanism to capture multi-scale layer information. Finally, PSGCN utilizes sortpool strategy to obtain the partner-specific graph embedding and formulates a drug-disease association prediction as a graph classification task. A fully-connected module is established to classify the partner-specific graph representations. The experiments on three benchmark datasets prove that the representation learning of partner-specific graph can lead to superior performances over state-of-the-art methods. In particular, case studies on small cell lung cancer and breast carcinoma confirmed that PSGCN is able to retrieve more actual drug-disease associations in the top prediction results. Moreover, in comparison with other static approaches, PSGCN can partly distinguish the different disease context information for the given drug.**

***Index Terms*—Drug repositioning, graph convolutional network, partner-specific graph, layer self-attention.**

Manuscript received 18 April 2022; revised 1 July 2022; accepted 24 July 2022. Date of publication 3 August 2022; date of current version 7 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61832019, in part by Hunan Provincial Science and Technology Program under Grant 2019CB1007, and in part by the Science and Technology innovation Program of Hunan Province under Grant 2021RC4008. *(Xinliang Sun and Bei Wang are contributed equally to this work.) (Corresponding authors: Jie Zhang; Min Li.)*

Xinliang Sun is with the Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China, and also with the SenseTime, Shanghai 200233, China (e-mail: xinliang-sun123456@csu.edu.cn).

Bei Wang is with the SenseTime, Shanghai 200233, China (e-mail: wangbei1@sensetime.com).

Jie Zhang is with the SenseTime, Shanghai 200233, China, and also with the Qing yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: stzhangjie@hotmail.com).

Min Li is with the Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: limin@mail.csu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2022.3194891, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2022.3194891

I. Introduction

DRUG repositioning, also known as drug repurposing, is a strategy aiming to investigate existing drugs for new therapeutic opportunities [1]. Compared to the laborious and expensive de novo drug discovery process, drug repositioning offers an effective and efficient way to facilitate potential drugs reaching the market, since pre-clinical and clinical information of the repurposed drugs are already available. Recently, the repositioned drugs, such as Remdesivir, Ritonavir, and Ocilizumab, have provided a rapid response to address the worldwide Coronavirus disease (COVID-19) [2], [3], suggesting that drug repositioning is a promising way to fight against diseases with no curative treatment.

For drug repositioning, computational methods are rising due to the explosion of biological data. Prior computational methods can be mainly divided into three categories: (1) matrix factorization and completion based methods, (2) two-stage machine learning based methods, and (3) deep learning methods, especially graph learning methods [4]. The basic idea of matrix factorization based methods is to map the drug-disease association matrix into a low-rank feature space and then minimize the relation reconstruction error by the latent embeddings, where the similarity matrices of drugs and diseases are taken as biological side information for additional constraints. For instance, SCMFDD [5] defined similarity measures based on biological context as additional constraints for the matrix factorization method. iDrug [6] constructed a cross-network between the drug-disease associations and drug-target associations, within which matrix factorization was adopted to predict potential drug-disease associations. MLMC [7] proposed a multi-view learning with matrix completion method to predict the potential associations between drugs and diseases. Although matrix factorization and completion based methods have shown progressive results, the high computational complexity poses scalability challenges for dealing with growing data.

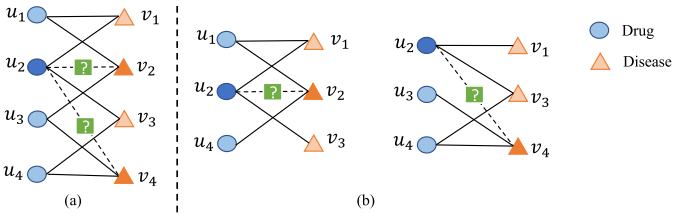

Fig. 1. An illustration of the difference between an entire bipartite graph of drug-disease associations and partner-specific graph. (a) A toy example of bipartite drug-disease graph, in which circles represent drugs and triangles represent diseases. (b) Describing the different context environment for two drug-disease pairs involving the same drug(node: u_2).

Machine learning based methods first pre-process side information as features and then predict whether a drug-disease pair is positive or negative based on these features. PREDICT [8] assembled multiple drug-drug and disease-disease similarity measures to construct features and fed them into a logistic regression classifier. HED [9] employed the metapath2vec [10] algorithm to generate feature vectors based on the constructed heterogeneous network with drug-drug similarity, disease-disease similarity and drug-disease association networks. HED then trains a support vector machine (SVM) model on the generated feature vectors. The two-stage machine learning methods heavily rely on arbitrary featurization, which requires domain knowledge and experience.

Deep learning algorithms are increasingly exploited for drug repositioning. CBPred [11] employed drug similarity and disease similarity information and considered the multiple paths information between drug-disease associations to improve the model performance. deepDR [12] integrated various heterogeneous networks information by a multi-modal deep autoencoder to infer candidates for approved drugs. LAGCN [13] developed a layer attention graph convolutional network (GCN) method. It constructed a heterogeneous network based on drug-disease associations, drug-drug similarities, and disease-disease similarities, then an attention-based GCN was utilized to encode the nodes, while a bilinear decoder was applied to reconstruct the drug-disease adjacency matrix. Such GCN-based methods have achieved promising performance for drug repositioning.

Nevertheless, in previous studies, GCN-based repositioning methods usually learn drug or disease representation from an entire graph of drug-disease associations. As shown in Fig. 1(a), it is a toy example of drug-disease association bipartite graph. The circles and triangles represent drugs and diseases, respectively. In this bipartite graph, we have four drugs (u_1, u_2, u_3, u_4) and four diseases (v_1, v_2, v_3, v_4). The solid and dashed lines represent known and unknown associations between drugs and diseases, respectively. The two unknown associations between u_2 and v_2 / v_4 need to be predicted. When predicting u_2-v_2 and u_2-v_4, prior GCN-based methods learn the same representative embedding for u_2 from the whole graph, which is static and agnostic to various diseases (such as v_2 and v_4). However, despite the same drug, its mechanism of actions (MoAs) to different diseases are different. Therefore, when targeting different diseases, a drug needs to differentiate the relevant context information. As Fig. 1(b) depicts, two different relation graphs are extracted for the two target drug-disease pairs (u_2, v_2) and (u_2, v_4), respectively.

To emphasize the difference of target partners, inspired by the subgraph-based studies [14], [15], we propose a partner-specific method based on GCN, termed PSGCN. For drug repositioning, PSGCN transforms a link prediction problem between a drug and a disease into a graph classification task. Each extracted graph collects the one-hop neighborhood information for the two drug-disease pairs. Implementing a GCN on such partner-specific graph can differentiate various context information as message propagation and integration, inducing more refined local structural features for inferring potential associations. The source code are at https://github.com/SenseTime-Knowledge-Mining/PSGCN.

Briefly, our contributions can be summarized as follows. (1) To the best of our knowledge, PSGCN is the first partner-specific method based on graph convolutional network for drug repositioning. (2) PSGCN can automatically learn suitable context information about partner-specific graph and utilize layer self-attention to capture multi-scale information. (3) The experiments demonstrate the effectiveness of PSGCN, and the case studies further present the potential of PSGCN for real-world application.

TABLE I
SUMMARY OF THE THREE BENCHMARK DATASETS

Dataset	Drugs	Diseases	Known associations
Gdataset	593	313	1,933
Cdataset	663	409	2,532
LRSSL	763	681	3,051

II. MATERIALS AND METHODS

In this study, we propose PSGCN, a novel learning framework based on graph convolutional network to capture latent relationships between drugs and diseases for effective drug repositioning. As Fig. 2 depicts, our framework mainly comprises three components: 1) construction of partner-specific graph around target drug-disease pairs; 2) structural information encoding of partner-specific graph with GCN; 3) prediction of the potential drug-disease associations.

A. Dataset

To evalute the performance of PSGCN, we use three public benchmarking datasets: Gdataset [8], Cdataset [16], and LRSSL [17]. Table I summarizes the detailed information of the three datasets. Gdataset contains 1,933 verified drug-disease associations between 593 drugs and 313 diseases, where the drugs are collected from DrugBank [18] and diseases are collected from Online Mendelian Inheritance in Man (OMIM) [19]. Cdataset contains 663 drugs and 409 diseases with 2,532 known drug-diseases associations derived from Comparative Toxicogenomics Database (CTD) [20]. LRSSL dataset includes 3,051 validated drug-disease associations involving 763 drugs and 681 diseases.

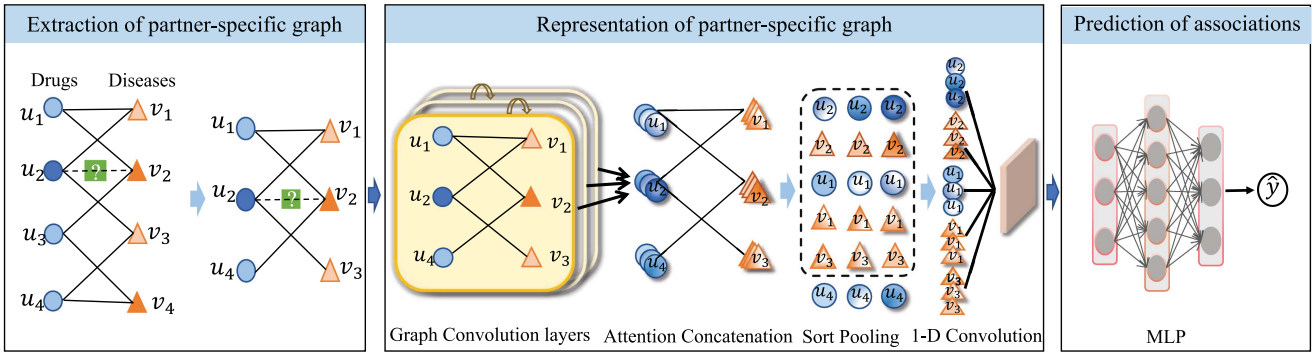

Fig. 2. The framework of PSGCN. PSGCN mainly consists of three modules: extraction of partner-specific graphs, representation learning of partner-specific graphs, and prediction of drug-disease associations. Firstly, target drug-disease associations are taken as centroids and their h-hop neighbors ($h = 1$ in this figure) are collected as context information, constituting partner-specific graphs. Then, representation of the graphs are obtained by graph convolution model and pooling operation. Finally, potential drug-disease associations are predicted as binary classification of the graphs.

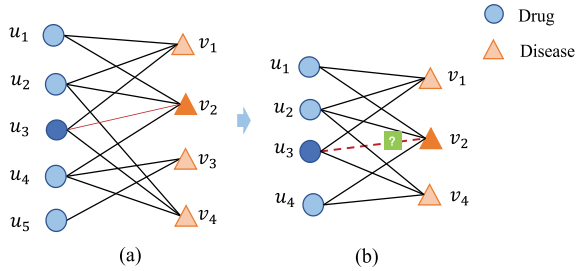

Fig. 3. An illustration of the partner-specific graph extraction. (a) Bipartite drug-disease network. (b) Extractd partner-specific graph around target node pair.

Besides, since baseline models need additional information, for the three datasets, we also collect SMILES [21] of drugs from DrugBank database, and utilize the Chemical Development Kit [22] to measure the similarity of two drugs based on the Tanimoto score of their 2D chemical fingerprints. Disease-disease similarity is obtained from MimMiner [23], where the similarity score has been normalized into [0,1]. Please note that Gdataset is regarded as the gold standard dataset, due to its clinically-validated drug-disease associations. Therefore, we take Gdataset as the main dataset and conduct comprehensive experiments on it for evaluation.

B. Partner-Specific Graph Construction

Given the known drug-disease associations, we construct a bipartite drug-disease network, $G = \{U, V, E\}$, where $U = \{u_1, u_2, \ldots, u_m\}$ represents drug nodes, m is the number of drugs; and $V = \{v_1, v_2, \ldots, v_n\}$ represents disease nodes, n is the number of diseases. $E = \{(u_i, v_j)|u_i \in U, v_j \in V\}$ is the edge set representing known associations between drugs and diseases. The adjacency matrix of the network is $A \in R^{m \times n}$, where $a_{ij} = 1$ ($a_{ij} \in A$) when an edge between drug u_i and disease v_j exists, otherwise, $a_{ij} = 0$. Based on the constructed drug-disease bipartite network, the partner-specific graphs are extracted. For a target drug-disease pair to be predicted, e.g. (u_3, v_2) in Fig. 3, we take the two nodes as centroid and extract h-hop neighbors ($h = 2$) around them, i.e., u_1, u_2, u_4, and v_1, v_4 in Fig. 3(b). The extracted partner-specific sub-graph around drug u_i and disease v_j is denoted as P_{ij} ($P_{ij} \subseteq G$). The two nodes u_i and v_j are called *center nodes*, and the other nodes (e.g. u_1, u_2, u_4, v_1, v_4) are called *context nodes*.

A node labeling strategy [24] is then adopted on the extracted partner-specific sub-graphs, to (1) distinguish the center nodes from context nodes and (2) differentiate the types of nodes, i.e., drug or disease. Specifically, the node labels of the drug and disease at center are initialized as 0 and 1, respectively. The context nodes are then labeled according to their hops. A drug node at the h-th hop is labeled as $2h$. A disease node at the h-th hop is labeled as $2h + 1$. The labels are based on the roles (target or context) and types (drug or disease) of the nodes in a partner-specific graph (P_{ij}), and independent of the entire drug-disease association graph G. Thus, drug and disease nodes have distinguishable labels for graph representation learning. It is worth noting that, as shown in Fig. 3(b), the target drug-disease association in a partner-specific graph is removed to avoid information leakage.

C. Representation Learning of Partner-Specific Graph

After extracting the partner-specific graphs, we leverage a graph neural network to learn the partner-specific graph representation for association prediction. Most of the existing GCN based drug repositioning methods, such as [13], [25], apply a node-level GCN on the whole drug and disease related networks to learn the node embeddings, then directly predict the association probability with an inner-product or bilinear operator. Differently, in this work, we learn *graph-level* embeddings by implementing a *graph-level* GCN model to learn the extracted partner-specific graph representation. The graph-level GCN consists of two components: 1) message propagation layers to produce node representations with context information in partner-specific graph, and 2) a pooling layer to aggregate the node embeddings into a comprehensive partner-specific graph-level representation.

Message propagation on partner-specific graphs: To capture the context information of the center nodes in the extracted partner-specific graph, a graph convolutional network [26] is

applied to learn the node representations. Moreover, to fully acquire multi-scale structure information of the extracted graph, we stacked L message passing layers. The convolutional operation for a partner-specific graph P_{ij} at the l-th layer is formed as (1).

$$Z^l = f(\widetilde{D}^{-\frac{1}{2}}\widetilde{A^p}\widetilde{D}^{-\frac{1}{2}}Z^{l-1}W^l) \tag{1}$$

where $Z^l \in \mathcal{R}^{N\times d_l}$ denotes the output node embeddings at layer l, N is the number of nodes in the partner-specific graph, and d_l is the number of output channels of layer l. For $l = 1$, the initial node features $Z^0 = X$ is a one-hot encoding of the node labels in the partner-specific graph. $\widetilde{A^p} = A^p + I$, is the adjacency matrix of partner-specific graph P_{ij} with added self-connections. I is the identity matrix, and $\widetilde{D}$ is a diagonal degree matrix with $\widetilde{D}_{ii} = \sum_j \widetilde{A^p}_{ij}$. W^l is trainable parameter matrix at layer l. f is a nonlinear activation function. After obtaining the nodes embeddings from different convolutional layers, inspired by recent research [27], a learnable layer self-attention vector $\overrightarrow{\alpha}$ is applied to retain more important information from different layers based on both graph features and topology. The final node representation $Z^{1:L}$ is defined as follows:

$$\begin{aligned} Z^{1:L} &= concat(\alpha_1 Z^1, \alpha_2 Z^2, \ldots, \alpha_L Z^L) \\ \overrightarrow{\alpha} &= (\alpha_1, \alpha_2, \ldots, \alpha_L) \end{aligned} \tag{2}$$

where $Z^{1:L} \in \mathcal{R}^{N\times\sum_1^L d_l}$ is the concatenated output of L graph convolutional layers, and each row is a node feature embedding and each column is a feature channel.

Graph-level representation learning: When acquiring the final node representations, we integrate them into the graph feature vector by a sorted pooling layer [28]. Specifically, we sort all the nodes based on the values of $Z^{1:L}$ in a descending order. That is, we first compare the value of two nodes from the last channel of Z^L to the first channel of Z^1 until the value is not equal. After sorting the node features in a consistent order, the number of nodes is unified into a fixed size, truncating the output node representations from N to K by deleting the last $N - K$ rows if $N > K$, or extending the output by adding $K - N$ zero rows if $N < K$. The sorted pooling layer is formed as (3).

$$Z^p = \Gamma(Z^{1:L}) = \Gamma_{l:L\to 1}\Gamma_{d:d_l\to 1}(Z^{ld}) \tag{3}$$

where $Z^p \in K \times \sum_1^L d_l$ is the output of the sort pooling layer. Γ is descending sort operation and Z^{ld} denotes the value of the dth channel in l layer. Detailedly, $\Gamma_{l:L\to 1}$ means to sort the final node embedding from the output of layer L to layer 1, and for the output of each layer l, $\Gamma_{d:d_l\to 1}(Z^{ld})$ means to sort node vector Z^l from its last channel d_l to the first channel. Next, a reshape operation is utilized to map the partner-specific graph into a $K\sum_1^L d_l \times 1$ vector. After that, MaxPooling layers and 1-D convolutionalal layers are utilized to learn local partner-specific graph patterns on the node sequence.

Drug-disease association prediction: Finally, fully connected layers perform classification for each partner-specific graph P_{ij} with the pooled feature vector Z^p (4).

$$\hat{y} = W_2 \cdot relu(W_1 Z^p + b_1) + b_2 \tag{4}$$

where $\hat{y}$ indicates the probability of the association between drug u_i and disease v_j. W_1 and W_2 are trainable weights, and b_1 and b_2 are bias.

D. Model Training

We transform the drug-disease association prediciton problem to graph classification task, and adopt the cross-entropy loss function (5) to train the model. The known drug-disease associations in the dataset are considered as positive samples and others as negative samples. Accordingly, the classification labels of extracted partner-specific graphs around positive samples are 1, and 0 otherwise.

$$\mathcal{L} = \sum_{(i,j)} -y_{ij} \cdot log(\hat{y}_{ij}) + (1 - y_{ij}) \cdot log(1 - \hat{y}_{ij}) \tag{5}$$

where (i, j) denotes the pair for drug u_i and disease v_j. y_{ij} is the truth label, while $\hat{y}_{ij}$ is the predicted association probability of (u_i, v_j). To optimize the model, we use the Adam optimizer [29] and train the model in a denoising setup by randomly dropping out all outgoing messages of a particular edge with a fixed probability. We also apply regular dropout [30] to prediction layers.

III. Results and Discussion

In this section, we evaluate our proposed PSGCN on three benchmark datasets against competitive drug repositioning methods. And we analyse the impact of hop depth on our model. Furthermore, we conduct a *de novo* experiment to verify the performance of our model for identifying potential indications. Finally, we show specific examples of the drug repositioning results to illustrate the ability of PSGCN in practical application.

A. Evaluation Metrics and Baseline Models

We conduct 10-fold cross validation to evaluate the performance of our approach. Specifically, all the known drug-disease associations are considered as positive samples and are randomly splited into ten subsets with the same size. In each fold, nine subsets are combined as the positive training set, while the remaining subset is treated as the positive testing set. Besides, we randomly select the same number of negative instances as the positive ones in training and testing sets. We utilize Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision Recall curve (AUPR) as the metrics to assess the performance of models, which are widely used for drug repositioning prediction tasks. To provide robust estimation of performance, we repeated the 10-fold cross validation procedure ten times, and reported the mean results.

We compared the performance of our model against several competitive drug repositioning approaches. For SCMFDD, iDrug, NRLMF and DRWBNCF, we adopt the same experimental setting as their paper recommended. For GRMF, following [6], the parameters $\lambda_l = 0.5, \lambda_d = \lambda_t = 10^{-3}$ are chosen. For NIMCGCN, we set the parameters following that in [31].

- SCMFDD [5] is a matrix factorization method, which maps the high dimensional drug and disease latent vectors

TABLE II
PERFORMANCE COMPARISON OF 10 TIMES 10-FOLD CROSS VALIDATION PREDICTION RESULTS BETWEEN OUR METHOD AND BASELINES OVER GDATASET, CDATASET AND LRSSL DATASET

	Dataset	SCMFDD	iDrug	GRMF	NRLMF	NIMCGCN	DRWBNCF	PSGCN
AUROC	Gdataset	0.7731±0.0196	0.9078±0.0158	0.7476±0.0299	0.9097±0.0151	0.8234±0.0157	0.9061±0.0149	**0.9485±0.0097**
	Cdataset	0.7896±0.0157	0.9294±0.0137	0.7469±0.0240	0.9257±0.0129	0.8393±0.0087	0.9277±0.0157	**0.9566±0.0114**
	LRSSL	0.7698±0.0211	0.8993±0.0104	0.6924±0.0369	0.8854±0.0147	0.7581±0.0123	0.9232±0.0128	**0.9395±0.0104**
AUPR	Gdataset	0.7749±0.0210	0.9265±0.0127	0.7978±0.0264	0.9302±0.0124	0.8590±0.0156	0.9307±0.0109	**0.9558±0.0087**
	Cdataset	0.7878±0.0180	0.9454±0.0098	0.8007±0.0194	0.9441±0.0097	0.8728±0.0092	0.9476±0.0099	**0.9627±0.0101**
	LRSSL	0.7860±0.0203	0.9212±0.0080	0.7689±0.0250	0.9102±0.0113	0.7962±0.0096	0.9339±0.0100	**0.9462±0.0101**

The best reported result is bolded and the second best result is underlined.

to low dimensional space for prediction, and incorporates drug and disease similarities as constraints.

- iDrug [6] is a cross-network framework that integrates drug repositioning and drug-target prediction into a unified networks with the overlapped drugs as the anchor nodes.
- GRMF [32] uses graph regularization to learn low-dimensional non-linear manifolds. In addition, the method considers that many of the non-occurring edges in the network are actually unknown or missing cases, and developed a preprocessing step to enhance prediction.
- NRLMF [33] utilizes a logistic matrix factorization based method with a nearest neighborhood regularization for drug-target interaction prediction.
- NIMCGCN [34], a graph convolutional network based method for miRNA-disease association prediction, which are demostrated to have great potential for drug repositioning.
- DRWBNCF [35] adopts weighted bilinear graph convolution operation to predict potential drug-disease associations by integrating the prior information of drugs and diseases.

B. Performance of PSGCN in the Cross-Validation

For a fair comparison, we reported the average results of ten times 10-fold cross validation on three datasets, with variance to show the stability of the results. As shown in Table II, we find that PSGCN consistently attains the best AUROC and AUPR values over three datasets. The AUROC values achieved by PSGCN on Gdataset is 0.9485, which is higher than the second best method NRLMF 3.88%. It is also observed that PSGCN achieves 0.9566 and 0.9395 for AUROC values on Cdataset and LRSSL, respectively, which is an improvement of 2.72% and 1.63% compared to the iDrug and DRWBNCF. For the AUPR metric, PSGCN yields the best performance among all methods, leading to an average improvement of 2.51% compared to the DRWBNCF method on Gdataset, and about 1.50% relative improvement over previous state-of-the-art method DRWBNCF on both Cdataset and LRSSL datasets. This demonstrates the superiority of our proposed method in drug repositioning task.

Notice that the baseline methods achieve the results in Table II by using additional information, while our PSGCN is only based on learning from the bipartite drug-disease network. For instance, iDrug utilizes drug-target interaction information as cross-domain bridge. Furthermore, when compared to GCN-based methods, such as NIMCGCN and DRWBNCF, PSGCN still outperforms them by a large margin on three datasets, which demonstrates the significance of the partner-specific representation learning.

C. Parameter Analysis

To investigate the effect of hop depth when extracting partner-specific graph on model performance, we searched the number of hop in the range of $\{1, 2, 3, 4\}$ and performed 10-fold cross validation on Gdataset. The mean results of AUROC and AUPRC are illustrated in Fig. 4. Clearly, extracting only 1-hop induces the lowest performance, and 2-hop depth is capable of endowing the model with better representation ability. While continuing increasing the hop depth, the improvements are not obvious. This suggests that the 2-hop information is sufficient to differentiate partner-specific context information. Thus, we choose the hop as 2 in this paper. The other two dataset h-hop experiment results are shown in Supplementary Figure S1 and Figure S2.

D. Test of PSGCN on Sparse Data

In this section, we investigate the model performance on sparsity data, we test the model on Gdataset under different sparsity levels of the drug-disease associations. Here we randomly retain a part of known associations in the Gdataset at a ratio $\lambda \in \{80\%, 85\%, 90\%, 95\%, 100\%\}$ and implement 10-fold cross validation to evaluate the method. As shown in Fig. 5, we found that the number of drug–disease associations is an important factor for the drug-disease association prediction, and more associations can result in better prediction models. Besides, we also found that PSGCN can produce the robust and high performances across different sparsity data and the results still outperform the baseline models under low associations scenario.

E. Discovering Candidates for New Drugs

To validate the ability of PSGCN on prediction for new drugs, we performed de novo test on the Gdataset. For a test drug, we removed all of its known disease associations as the testing set, and used all the remaining associations as the training samples. Such setting makes the test drug isolated in the known drug-disease association graph. Therefore, to relieve the cold-start problem, we conduct a K-Nearest Neighbor preprocessing step for these new drugs. Specifically, for each novel drug, K

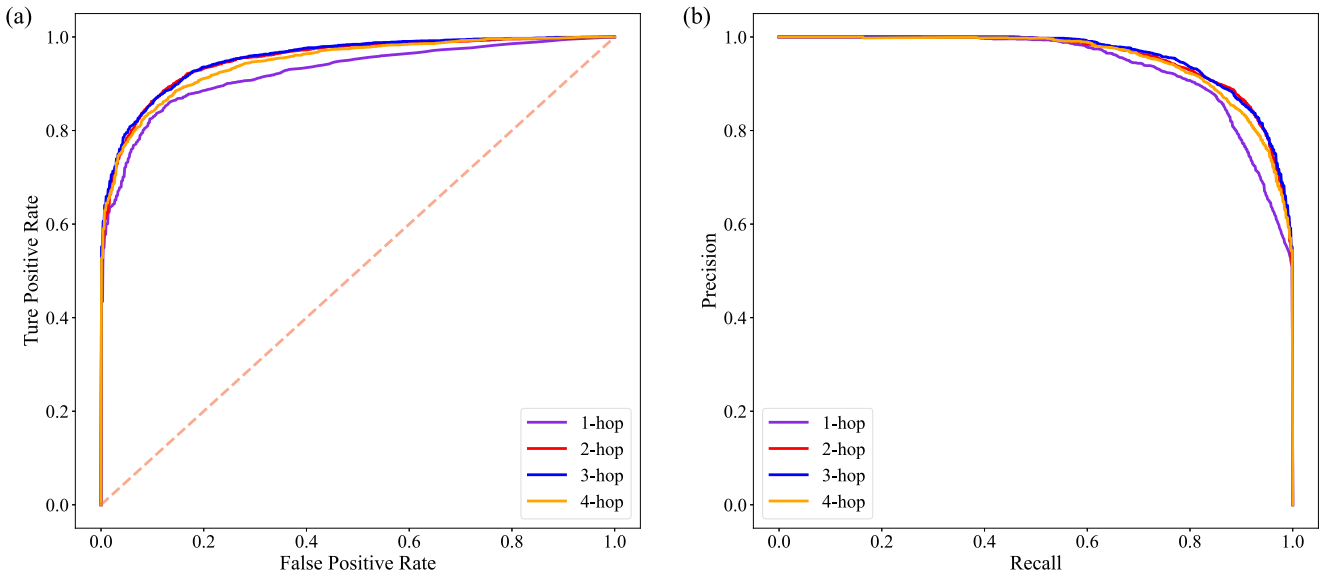

Fig. 4. Impact of the number of hop on PSGCN model performance on Gdataset. (a) The Area Under the Receiver operating characteristic (AUROC) curves of 10-fold cross validation results obtained by searching different hops. (b) The Area Under the Precision Recall (AUPR) curves of 10-fold cross validation results obtained by searching different hops.

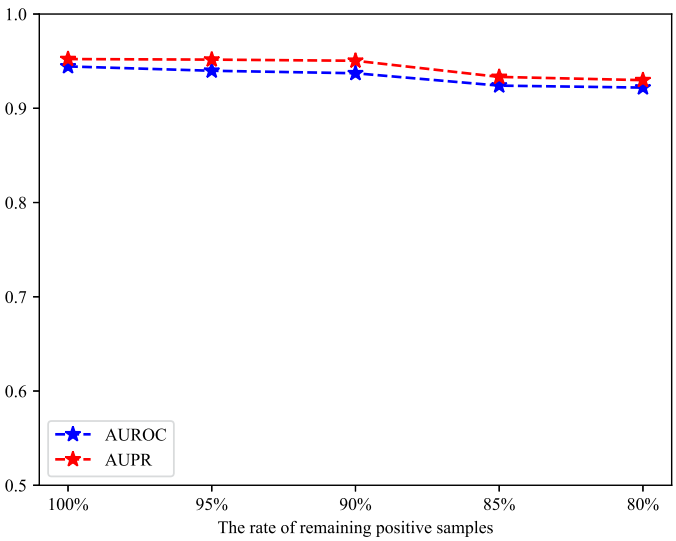

Fig. 5. PSGCN performances for different sparsity ratios on Gdataset.

nearest neighbor drugs of the given drug are picked based on their drug similarities in descending order. Then we update its associations in the bipartite drug-disease graph with a part of its nearest neighbor drugs' association information. As shown in Table III, we can find that PSGCN achieves the best results on both evaluation metrics (AUROC is 0.8970, AURP is 0.3484), which demonstrates the superiority of PSGCN for indicating indications for novel drugs.

F. Case Studies

To assess the practical use of PSGCN, we take small cell lung cancer (SCLC) and breast carcinoma as case studies. Specifically, we employ all the known drug-disease associations in the Gdataset as training set and take the missing drug-disease associations as candidate pairs for SCLC and breast carcinoma. We subsequently rank all the candidate drugs by the computed prediction scores for each disease, and verify the predicted top 10 potential drug-disease associations in acknowledged CTD [36], PubChem,[1] and DrugCentral [37] databases.

TABLE III
PERFORMANCE OF ALL METHODS IN PREDICTING POTENTIAL INDICATIONS FOR NEW DRUGS ON GDATASET

Methods	AUROC	AUPR
SCMFDD	0.7625	0.1228
iDrug	0.8260	0.2094
GRMF	0.6299	0.1322
NRLMF	0.8062	0.3287
NIMCGCN	0.8652	0.1857
DRWBNCF	0.8343	0.3196
PSGCN	**0.8970**	**0.3484**

SCLC: SCLC is the most malignant type of lung cancer, with the characteristics of rapid progression, high metastatic tendenc and easy recurrence. As shown in Table IV, among the top 10 predicted candicate drugs by PSGCN, eight drugs out of the top predicted ten candidate drugs can be confirmed by authoritative public databases (80% hit rate). For example, Doxorubicin is the top predicted candidate, which has been proved that the combination of NGR-hTNF(a vascular-targeting agent) and Doxorubicin shows manageable toxicity and promising activity in patients with relapsed SCLC [38]. The second predicted candidate drug is ifosfamide, which is an alkylating and immunosuppressive agent used in chemotherapy for the treatment of cancers. The therapeutic effect of ifosfamide on SCLC has been demonstrated on CTD database [39], [40].

[1][Online]. Available: https://pubchem.ncbi.nlm.nih.gov

TABLE IV
TOP 10 PREDICTED DRUGS FOR POTENTIALLY TREATING SMALL CELL LUNG CANCER

Rank	DrugBank ID	Candidate drug	Evidence
1	DB00997	Doxorubicin	[43] [38]
2	DB01181	Ifosfamide	[39] and [40]
3	DB01254	Dasatinib	[44]
4	DB00958	Carboplatin	[45] [46]
5	DB00570	Vinblastine	[47]
6	DB00619	Imatinib	[48]
7	DB01073	Fludarabine	[49]
8	DB01234	Dexamethasone	[50]
9	DB00262	Carmustine	NA
10	DB00851	Dacarbazine	NA

TABLE V
TOP 10 PREDICTED DRUGS FOR POTENTIALLY TREATING BREAST CANCER

Rank	DrugBank ID	Candidate drug	Evidence
1	DB00541	Vincristine	[41] [51]
2	DB00977	Ethinylestradiol	[42]
3	DB00399	Zoledronic acid	[52]
4	DB01005	Hydroxyurea	[53]
5	DB00851	Dacarbazine	[54]
6	DB01204	Mitoxantrone	[55] [56]
7	DB00515	Cisplatin	[57]
8	DB00694	Daunorubicin	[58]
9	DB00262	Carmustine	[59]
10	DB00762	Irinotecan	[60]

Breast cancer: Similar to SCLC, we also focus on analyzing the top 10 drug candidates for Breast cancer predicted by PSGCN. Table V shows the Top 10 potential drugs recommended by PSGCN. These 10 potential drugs have been verified by the reliable evidence with 100% hit rate. [41] find that simultaneous liposomal delivery of Vincristine and Quercetin has the ability to enhance estrogen-receptor-negative breast cancer treatment. Ethinylestradiol (a synthetic compound) has been used primarily as contraceptive. However, the recent research [42] finds that Ethinylestradiol can cure a patient with metastatic breast cancer.

To sum up, such case studies demonstrate the promising ability of PSGCN for discovering potential drugs for specific diseases. We expect that the predicted candidate drugs by PSGCN will provide a meaningful reference for clinicians in practical application.

G. Visualization

To intuitively present the characteristic of our PSGCN that extracts specific contextual information for association prediction, we visualize the embeddings of partner-specific graphs corresponding to Doxorubicin (a drug) with different targeting diseases. Specifically, we employ principal component analysis (PCA) to transform the high dimensional embeddings to two primary dimensions for visualization, and we set the same color for nodes close to each other.

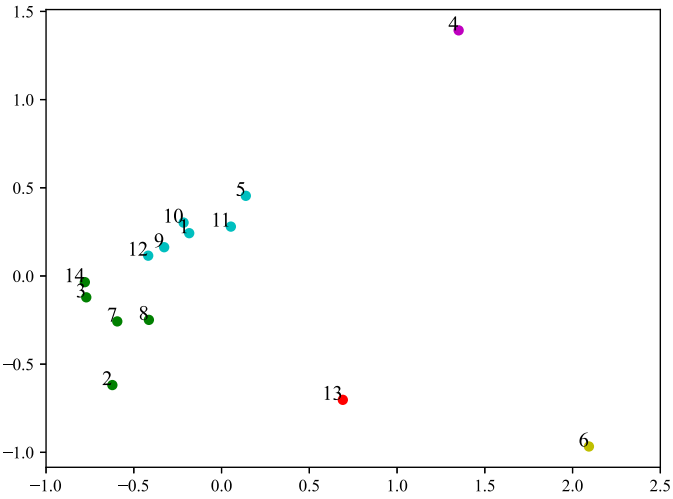

Fig. 6. Visualization of partner-specific graph representation learned from PSGCN model. The X- and Y- axes represent the two primary dimensions after performing PCA, respectively. The nodes represent the learned embeddings of partner-specific graphs when Doxorubicin targets different diseases.

As shown in Fig. 6, these nodes are roughly clustered into two clusters (green nodes and light blue nodes). As we known, subject headings in the Mesh Tree [61] reveal the degree of category similarity between diseases. Thus it is reasonable to expect that diseases with closer cataloging in MeSH will induce more similar partner-specific embeddings. In the green cluster, the node *7: Hodgkin Disease* and node *8: Lymphoblastic Leukemia* are both associated with lymphoid tissue. Concretely, *Hodgkin Disease* belongs to Lymphoma (with subject heading [C04.557.386]) in Mesh Tree, and *Lymphoblastic Leukemia* under the category of Leukemia, Lymphoid (with subject heading [C04.557.337.428]). In the light blue cluster, node *10: Neuroblastoma* and *11: Osteosarcoma*, the subject headings of which in the Mesh Tree are under [C04.557.465] and [C04.557.450], respectively, both occur most often in young children.

Such examples indicate the potential of PSGCN on capturing distinguishable representations for different target pairs. Since these diseases are not standard for strict clustering, it is inevitable to appear a slight bias, (e.g. *13: Turcot syndrome* and *3: Stomach Neoplasms* are all under Digestive System Neoplasms (with subject headings [C04.588.274]) of the Mesh Tree. However, compared to other methods that learn a static embedding, our method provides a more differentiated representation for effective prediction. Additionally, such visual descriptions can be regarded as a reference to assist researchers to find the association among diseases. The more details about visualization are shown in Supplementary Table S1.

IV. CONCLUSION

In this paper, we present a novel partner-specific drug repositioning approach based on graph neural network, PSGCN. Instead of learning general feature for each node, PSGCN emphasizes to learn a summary representation for the graph of a

specific drug-disease association, which considers the different roles of an object corresponding to different cases.

Compared to previous drug repositioning methods, our PSGCN method applies graph convolutional network to automatically capture partner-specific context information for expressive feature learning. Extensive experiments have demonstrated the superior performance of PSGCN on the task of drug repositioning. The partner-graph representation visualization of a drug with different diseases indicates that our partner-specific strategy prompts the model to better capture target specific structural information, which provides high-quality representations for drug repositioning task. Furthermore, case studies suggest the ability of PSGCN to predict unknown drug-disease associations in terms of the concrete diseases.

Although PSGCN obtains satisfactory results, compared to the whole drug space, the known drug–disease associations are sometimes too sparse to provide abundant context information, resulting in the performance of PSGCN still having room for improvement. In the future, we will consider to enrich the partner-specific graph with more biological entities. That is, not only to capture the direct association context information, but also to incorporate the biological knowledge based context information for more robust and effective representation learning.

References

[1] S. Pushpakom et al., "Drug repurposing: Progress, challenges and recommendations," *Nature Rev. Drug Discov.*, vol. 18, no. 1, pp. 41–58, 2019.

[2] C. Harrison, "Coronavirus puts drug repurposing on the fast track," *Nature Biotechnol.*, vol. 38, no. 4, pp. 379–381, 2020.

[3] Y. Zhou, F. Wang, J. Tang, R. Nussinov, and F. Cheng, "Artificial intelligence in COVID-19 drug repurposing," *Lancet Digit. Health*, vol. 2, no. 12, pp. e667–e676, 2020.

[4] H. Luo, M. Li, M. Yang, F.-X. Wu, Y. Li, and J. Wang, "Biomedical data and computational models for drug repositioning: A comprehensive review," *Brief. Bioinf.*, vol. 22, no. 2, pp. 1604–1619, 2021.

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

[6] H. Chen, F. Cheng, and J. Li, "iDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding," *PLoS Comput. Biol.*, vol. 16, no. 7, 2020, Art. no. e1008040.

[7] Y. Yan, M. Yang, H. Zhao, G. Duan, X. Peng, and J. Wang, "Drug repositioning based on multi-view learning with matrix completion," *Brief. Bioinf.*, vol. 23, no. 3, 2022, Art. no. bbac054.

[8] 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, 2011, Art. no. 496.

[9] K. Yang, X. Zhao, D. Waxman, and X.-M. Zhao, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos: An Interdiscipl. J. Nonlinear Sci.*, vol. 29, no. 12, 2019, Art. no. 123109.

[10] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 135–144.

[11] P. Xuan, Y. Ye, T. Zhang, L. Zhao, and C. Sun, "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations," *Cells*, vol. 8, no. 7, 2019, Art. no. 705.

[12] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "deepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[13] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Brief. Bioinf.*, vol. 22, no. 4, 2021, Art. no. bbaa243.

[14] Z. Cao, L. Wang, and G. De Melo, "Link prediction via subgraph embedding-based convex matrix completion," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2803–2810.

[15] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5171–5181.

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

[17] 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, 2017.

[18] D. S. Wishart et al., "Drugbank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

[19] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[20] A. P. Davis et al., "The comparative toxicogenomics database: Update 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D972–D978, 2017.

[21] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[22] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The chemistry development kit (CDK): An open-source java library for chemo-and bioinformatics," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 493–500, 2003.

[23] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Hum. Genet.*, vol. 14, no. 5, pp. 535–542, 2006.

[24] M. Zhang and Y. Chen, "Inductive matrix completion based on graph neural networks," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–25.

[25] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "Neodti: Neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2019.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.

[27] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3734–3743.

[28] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4438–4445.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–41.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] L. Cai et al., "Drug repositioning based on the heterogeneous information fusion graph convolutional network," *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab319.

[32] A. Ezzat, P. Zhao, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 646–656, 2016.

[33] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS Comput. Biol.*, vol. 12, no. 2, 2016, Art. no. e1004760.

[34] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2538–2546, 2020.

[35] Y. Meng, C. Lu, M. Jin, J. Xu, X. Zeng, and J. Yang, "A weighted bilinear neural collaborative filtering approach for drug repositioning," *Brief. Bioinf.*, vol. 23, no. 2, 2022, Art. no. bbab581.

[36] A. P. Davis et al., "Comparative toxicogenomics database (CTD): Update 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, 2020.

[37] O. Ursu et al., "DrugCentral: Online drug compendium," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D932–D939,2016.

[38] V. Gregorc et al., "NGR-hTNF and doxorubicin as second-line treatment of patients with small cell lung cancer," *Oncologist*, vol. 23, no. 10, pp. 1133–e112, 2018 .

[39] D. Decaudin et al., "In vivo efficacy of STI571 in xenografted human small cell lung cancer alone or combined with chemotherapy," *Int. J. Cancer*, vol. 113, no. 5, pp. 849–856, 2005.

[40] I. Tanaka et al., "A phase II trial of ifosfamide combination with recommended supportive therapy for recurrent SCLC in second-line and heavily treated setting," *Cancer Chemotherapy Pharmacol.*, vol. 81, no. 2, pp. 339–345, 2018.

[41] M.-Y. Wong and G. N. Chiu, "Simultaneous liposomal delivery of quercetin and vincristine for enhanced estrogen-receptor-negative breast cancer treatment," *Anti-Cancer Drugs*, vol. 21, no. 4, pp. 401–410, 2010.
[42] A. Sueta et al., "Successful ethinylestradiol therapy for a metastatic breast cancer patient with heavily pre-treated with endocrine therapies," in *Proc. Int. Cancer Conf. J.*, 2016, vol. 5, pp. 126–130.
[43] N. B. Leighl et al., "A phase I study of pegylated liposomal doxorubicin hydrochloride (Caelyx) in combination with cyclophosphamide and vincristine as second-line treatment of patients with small-cell lung cancer," *Clin. Lung Cancer*, vol. 5, no. 2, pp. 107–112, 2003.
[44] H. Yang et al., "Pharmaco-transcriptomic correlation analysis reveals novel responsive signatures to HDAC inhibitors and identifies dasatinib as a synergistic interactor in small-cell lung cancer," *EBioMedicine*, vol. 69, 2021, Art. no. 103457.
[45] G. Rustin et al., "A phase Ib trial of CA4P (combretastatin A-4 phosphate), carboplatin, and paclitaxel in patients with advanced cancer," *Brit. J. Cancer*, vol. 102, no. 9, pp. 1355–1360, 2010.
[46] A. Mouri et al., "Combination therapy with carboplatin and paclitaxel for small cell lung cancer," *Respir. Investigation*, vol. 57, no. 1, pp. 34–39, 2019.
[47] J. Hardy, T. Noble, and I. Smith, "Symptom relief with moderate dose chemotherapy (mitomycin-C, vinblastine and cisplatin) in advanced non-small cell lung cancer," *Brit. J. Cancer*, vol. 60, no. 5, pp. 764–766, 1989.
[48] T. Yokoyama, K. Miyazawa, T. Yoshida, and K. Ohyashiki, "Combination of vitamin K2 plus imatinib mesylate enhances induction of apoptosis in small cell lung cancer cell lines," *Int. J. Oncol.*, vol. 26, no. 1, pp. 33–40, 2005.
[49] C. M. Rudin et al., "Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer," *Nature Genet.*, vol. 44, no. 10, pp. 1111–1116, 2012.
[50] M. Peifer et al., "Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer," *Nature Genet.*, vol. 44, no. 10, pp. 1104–1110, 2012.
[51] S. Esmaeili-Mahani, F. Falahi, and M. M. Yaghoobi, "Proapoptotic and antiproliferative effects of Thymus caramanicus on human breast cancer cell line (MCF-7) and its interaction with anticancer drug vincristine," *Evidence-Based Complement. Altern. Med.*, vol. 2014, 2014, Art. no. 893247.
[52] J. K. Woodward, H. L. Neville-Webbe, R. E. Coleman, and I. Holen, "Combined effects of zoledronic acid and doxorubicin on breast cancer cell invasion in vitro," *Anti-Cancer Drugs*, vol. 16, no. 8, pp. 845–854, 2005.
[53] Y. Tian et al., "Valproic acid sensitizes breast cancer cells to hydroxyurea through inhibiting RPA2 hyperphosphorylation-mediated DNA repair pathway," *DNA Repair*, vol. 58, pp. 1–12, 2017.
[54] F. Morales-Vásquez et al., "Adjuvant chemotherapy with doxorubicin and dacarbazine has no effect in recurrence-free survival of malignant phyllodes tumors of the breast," *Breast J.*, vol. 13, no. 6, pp. 551–556, 2007.
[55] F. Di Costanzo et al., "Paclitaxel and mitoxantrone in metastatic breast cancer: A phase II trial of the italian oncology group for cancer research," *Cancer Investigation*, vol. 22, no. 3, pp. 331–337, 2004.
[56] H. Qiao et al., "Redox-triggered mitoxantrone prodrug micelles for overcoming multidrug-resistant breast cancer," *J. Drug Targeting*, vol. 26, no. 1, pp. 75–85, 2018.
[57] H. E. Daaboul et al., "β-2-himachalen-6-ol inhibits 4t1 cells-induced metastatic triple negative breast carcinoma in murine model," *Chemico-Biol. Interact.*, vol. 309, 2019, Art. no. 108703.
[58] R.-J. Ju et al., "Octreotide-modified liposomes containing daunorubicin and dihydroartemisinin for treatment of invasive breast cancer," *Artif. Cells, Nanomedicine, Biotechnol.*, vol. 46, no. sup1, pp. 616–628, 2018.
[59] L. B. Marks et al., "Impact of high-dose chemotherapy on the ability to deliver subsequent local–regional radiotherapy for breast cancer: Analysis of cancer and leukemia group B protocol 9082," *Int. J. Radiat. Oncol.* Biol.* Phys.*, vol. 76, no. 5, pp. 1305–1313, 2010.
[60] M. E. Melisko, M. Assefa, J. Hwang, A. DeLuca, J. W. Park, and H. S. Rugo, "Phase II study of irinotecan and temozolomide in breast cancer patients with progressing central nervous system disease," *Breast Cancer Res. Treat.*, vol. 177, no. 2, pp. 401–408, 2019.
[61] I. K. Dhammi and S. Kumar, "Medical subject headings (MeSH) terms," *Indian J. Orthopaedics*, vol. 48, no. 5, 2014, Art. no. 443.