Communicative Subgraph Representation Learning for Multi-Relational Inductive Drug-Gene Interaction Prediction

Jiahua Rao[1,†], **Shuangjia Zheng**[1,4,*], **Sijie Mai**[2] and **Yuedong Yang**[1,3,*]
[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]School of Electronic and Information Technology, Sun Yat-sen University
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University
[4]Galixir Technologies Ltd, Beijing
{raojh6,zhengshj9,maisj}@mail2.sysu.edu.cn, yangyd25@mail.sysu.edu.cn

Code : https://github.com/biomed-ai/cosmig

IJCAI_2022

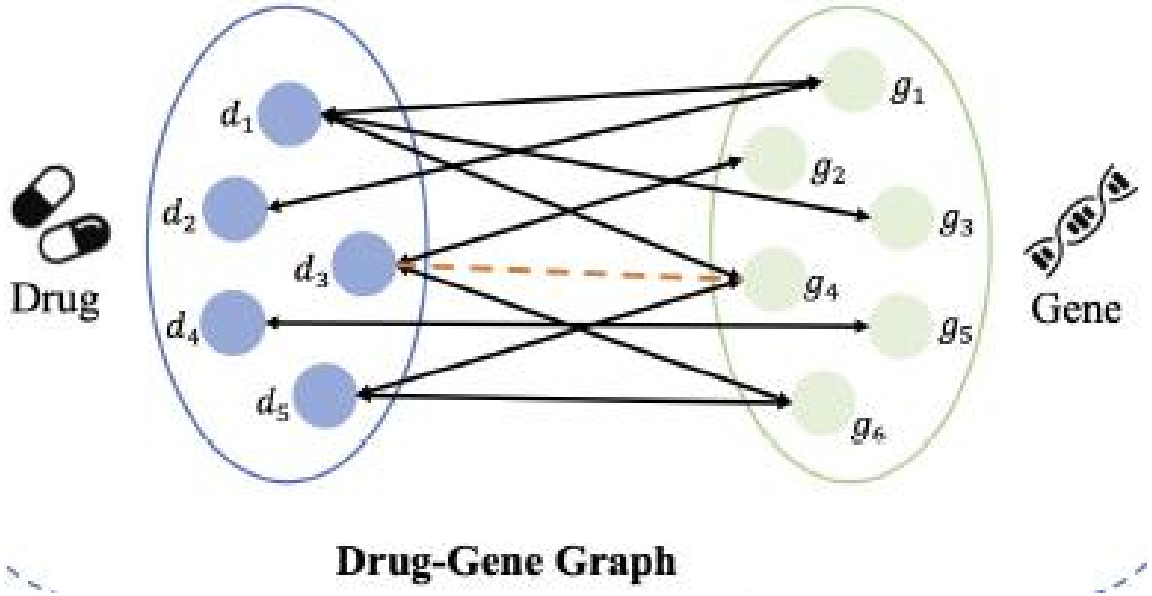
Reported by Xinsheng Wang

1.Introduction

2.Method

3.Experiments

Introduction



Drug-Gene Graph

Method

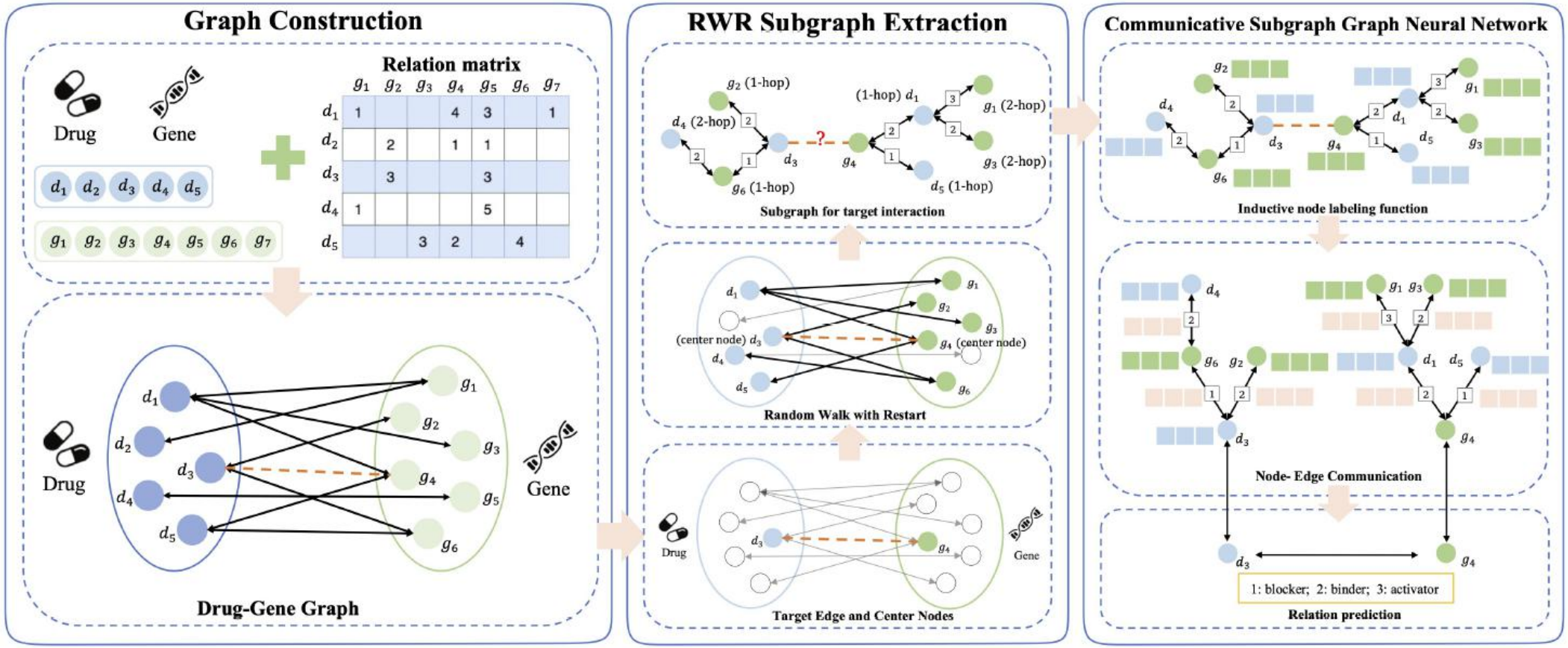
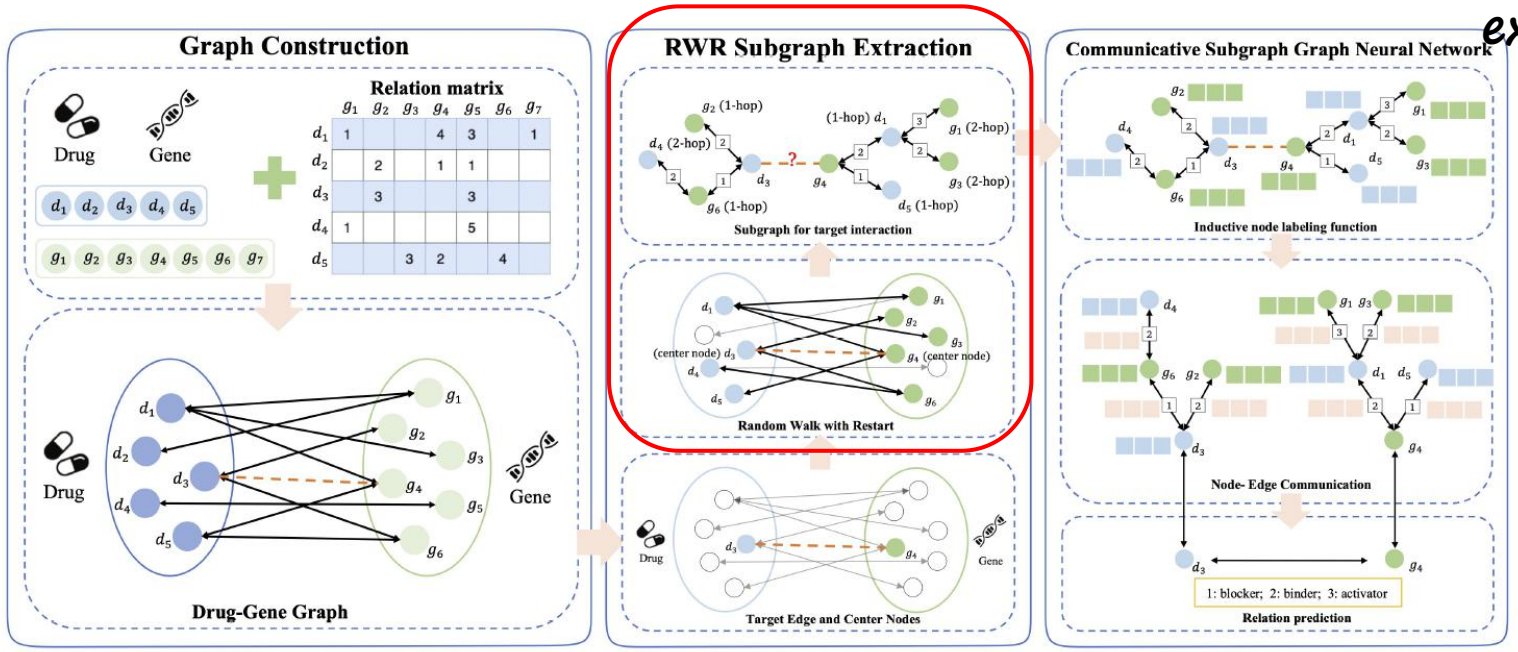


Figure 1: The framework of CoSMIG. We first extract a subgraph around each interaction and train a communicative subgraph graph neural network to map subgraphs to interactions. Each subgraph is induced by the drug and gene associated with the target interaction as well as their h-hop neighbors (here $h = 3$). Finally, the learned subgraph embedding of each interaction is used to predict the various interactions between drug and gene.

Method

RWR subgraph ext



$$p = cAD^{-1}p + (1-c)e \quad (1)$$

where c is a number in the range $(0, 1)$ called restart probability and p is a column vector with p_i denoting the probability at node i. D is the degree matrix of adjacency matrix A with each diagonal value $D_{ii} = \sum_j A_{ij}$. The restart probability controls whether the next walk is jumping to a randomly selected neighbor (with probability c) or going to the starting node (with probability $1-c$). For the starting vector e, we set $e_i = 1$ if node i is the starting node else 0, and thus the starting vector e allows us to preserve the node's local topological structure and AD^{-1} allows us to further visit their neighborhoods.

Figure 1: The framework of CoSMIG. We first extract a subgraph around each interaction and train a communicative subgraph graph neural network to map subgraphs to interactions. Each subgraph is induced by the drug and gene associated with the target interaction as well as their h-hop neighbors (here $h = 3$). Finally, the learned subgraph embedding of each interaction is used to predict the various interactions between drug and gene.

Method

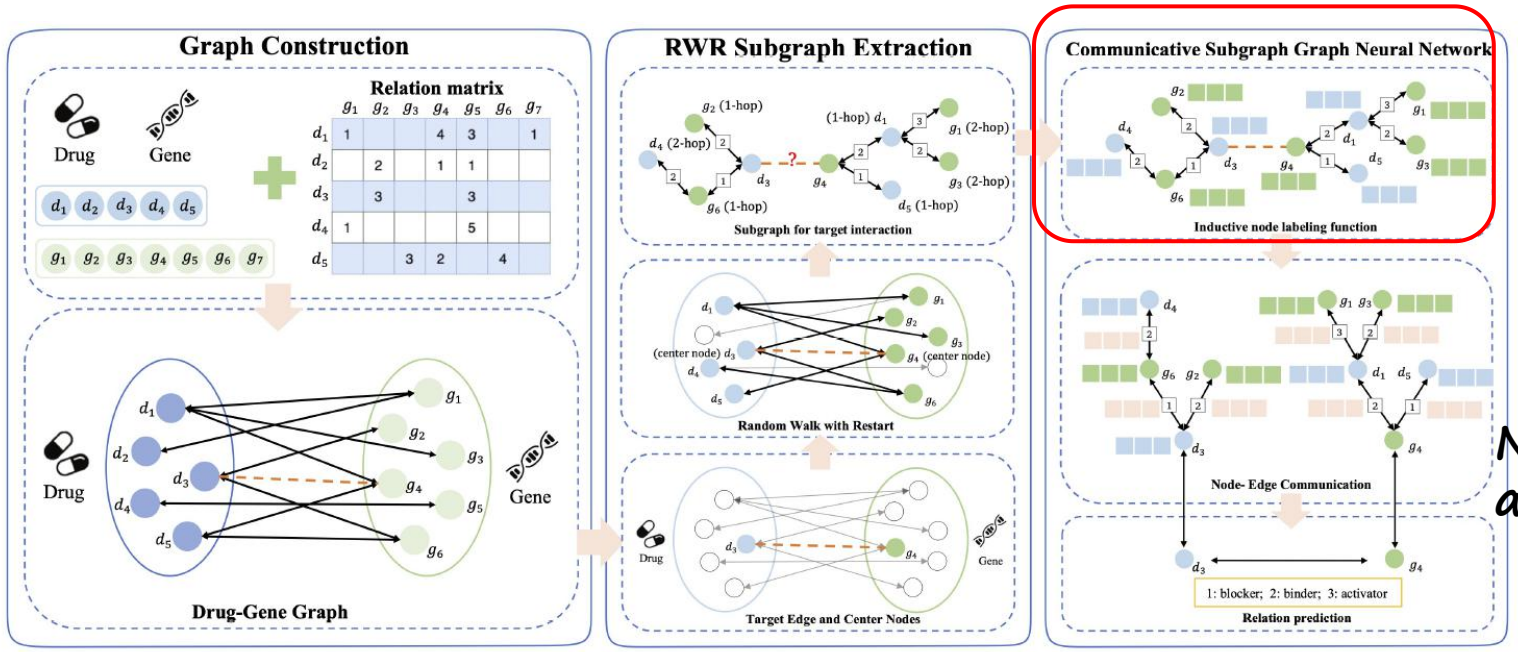


Figure 1: The framework of CoSMIG. We first extract a subgraph around each interaction and train a communicative subgraph graph neural network to map subgraphs to interactions. Each subgraph is induced by the drug and gene associated with the target interaction as well as their h-hop neighbors (here $h = 3$). Finally, the learned subgraph embedding of each interaction is used to predict the various interactions between drug and gene.

Communicative subgraph Neural Netwo

Before feeding to CoSMIG, we first apply an inductive node labeling function to it, which uses different labels to mark nodes' different roles in the subgraph without leveraging any external domain features and global information. Our node labeling function is defined as $(2i+j)$ where i is the hop number of the node and j is the node type with 0 representing the drug nodes and 1 representing the gene nodes. The one-hot encoding of these node labels will be treated as the initial node features of the subgraph, denoted as N_0. The initial relational feature R_0 is the one-hot encoding of interaction types and the initial edge features E_0 is the initial relation feature of each edge.

$$N^{(0)} = \sigma(N_0 W_n^{(0)}), E^{(0)} = \sigma(E_0 W_e^{(0)}) \quad (2)$$

Node embedding aggregation

$$\alpha_{i,j}^{(k)} = \sigma(\sigma([N_i^{(k)} || N_j^{(k)} || E_{i,j}^{(k)}] W_{a_0}^{(k-1)}) W_{a_1}^{(k)}) \quad (3)$$

$$E_{i,j}^{(k-1)} = \alpha_{i,j}^{(k-1)} E_{i,j}^{(k-1)} \quad (4)$$

$$N_{agg}^{(k)} = A^{ne} E^{(k-1)} \quad (5)$$

$$N^{(k)} = \sigma((N_{agg}^{(k)} + N^{(k-1)}) W_n^{(k)}) \quad (6)$$

A^{ne} and A^{re}

which represents the node-to-edge and relation type-to-edge adjacency matrix, respectively.

Method

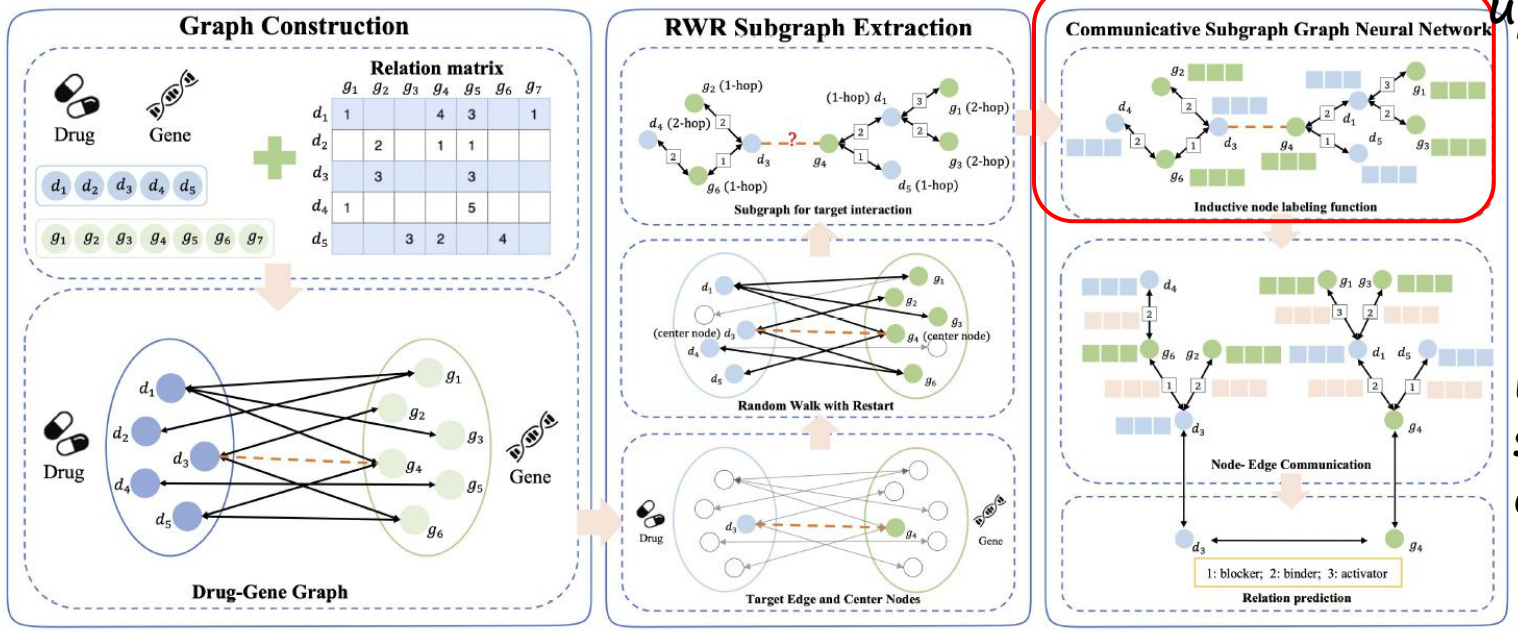


Figure 1: The framework of CoSMIG. We first extract a subgraph around each interaction and train a communicative subgraph graph neural network to map subgraphs to interactions. Each subgraph is induced by the drug and gene associated with the target interaction as well as their h-hop neighbors (here $h = 3$). Finally, the learned subgraph embedding of each interaction is used to predict the various interactions between drug and gene.

Relation embedding

$$E_{agg}^{(k)} = (A^{ne})^T N^{(k)} + (A^{re})^T R^{(k)} \quad (7)$$

R for the relation embeddings

$$E^{(k)'} = \sigma(E^{(k-1)} + \sigma(E_{agg}^{(k)})) \quad (8)$$

$$E^{(k)} = \sigma(E^{(k)'} W_e^{(k)} + E^{(0)}) \quad (9)$$

Relation prediction based on subgraph embedding

$$h = concat(N_d, N_g) \quad (10)$$

N_d and N_g denote the final representations

$$\widehat{r} = w_y^T \sigma(W_h h) \quad (11)$$

$$\mathcal{L} = \left(\frac{1}{|(d,g)|\Omega_{d,g} = 1|} \right) \sum_{(d,g):\Omega_{d,g}=1} (r_{d,g} - \widehat{r}_{d,g})^2 \quad (12)$$

Experiments

Dataset	DrugBank	DGIdb
Number of Drug	425	1185
Number of Gene	11284	1664
Interactions	80924	11366
Interaction types	2	14

Table 1: Statistics of two Drug-Gene Interaction datasets

Experiments

	Methods	Features	DrugBank		DGIdb	
			Validation ACC	Ind. Test ACC	Validation ACC	Ind. Test ACC
MF-based	MC	no	-	0.518 ± 0.013	-	0.559 ± 0.009
	GRALS	yes	-	0.532 ± 0.021	-	0.578 ± 0.016
	F-EAE	no	-	0.566 ± 0.004	-	0.623 ± 0.003
GNN-based	GC-MC	yes	-	0.586 ± 0.008	-	0.601 ± 0.005
	sRGCNN	yes	-	0.602 ± 0.010	-	0.689 ± 0.007
	PinSage	yes	-	0.629 ± 0.004	-	0.713 ± 0.005
	IGMC	no	-	0.634 ± 0.003	-	0.803 ± 0.006
Proposed	CoSMIG-w/GCN	no	0.562 ± 0.004	0.581 ± 0.004	0.778 ± 0.023	0.803 ± 0.009
	CoSMIG-w/GraphSAGE	no	0.584 ± 0.003	0.602 ± 0.008	0.807 ± 0.014	0.814 ± 0.010
	CoSMIG-w/RGCN	no	0.614 ± 0.004	0.637 ± 0.005	0.821 ± 0.013	0.832 ± 0.002
	CoSMIG-w/AvgPooling	no	0.619 ± 0.003	0.643 ± 0.006	0.822 ± 0.006	0.835 ± 0.003
	CoSMIG-w/SumPooling	no	0.625 ± 0.004	0.655 ± 0.003	0.824 ± 0.007	0.839 ± 0.004
	CoSMIG-w/SortPooling	no	0.639 ± 0.002	0.667 ± 0.004	0.833 ± 0.003	0.841 ± 0.005
	CoSMIG	no	$\mathbf{0.658 \pm 0.008}$	$\mathbf{0.678 \pm 0.003}$	$\mathbf{0.840 \pm 0.011}$	$\mathbf{0.852 \pm 0.012}$

Table 2: The comparison of different methods by the overall accuracy on the DrugBank and DGIdb datasets in transductive scenario.

Experiments

Methods		DrugBank		DGIdb	
		ACC	Δ	ACC	Δ
MF-based	IMC	0.441	14.8%	0.424	24.2%
	F-EAE	0.474	16.3%	0.532	14.6%
GNN-based	GC-MC	0.513	12.4%	0.553	7.99%
	PinSage	0.567	9.86%	0.654	8.27%
	IGMC	0.612	3.47%	0.778	3.11%
Proposed	CoSMIG	**0.672**	**0.88%**	**0.842**	**1.17%**

Table 3: Performance on the inductive scenario. Δ represents the decline rate between the transductive scenario and the inductive scenario of each model.

Experiments

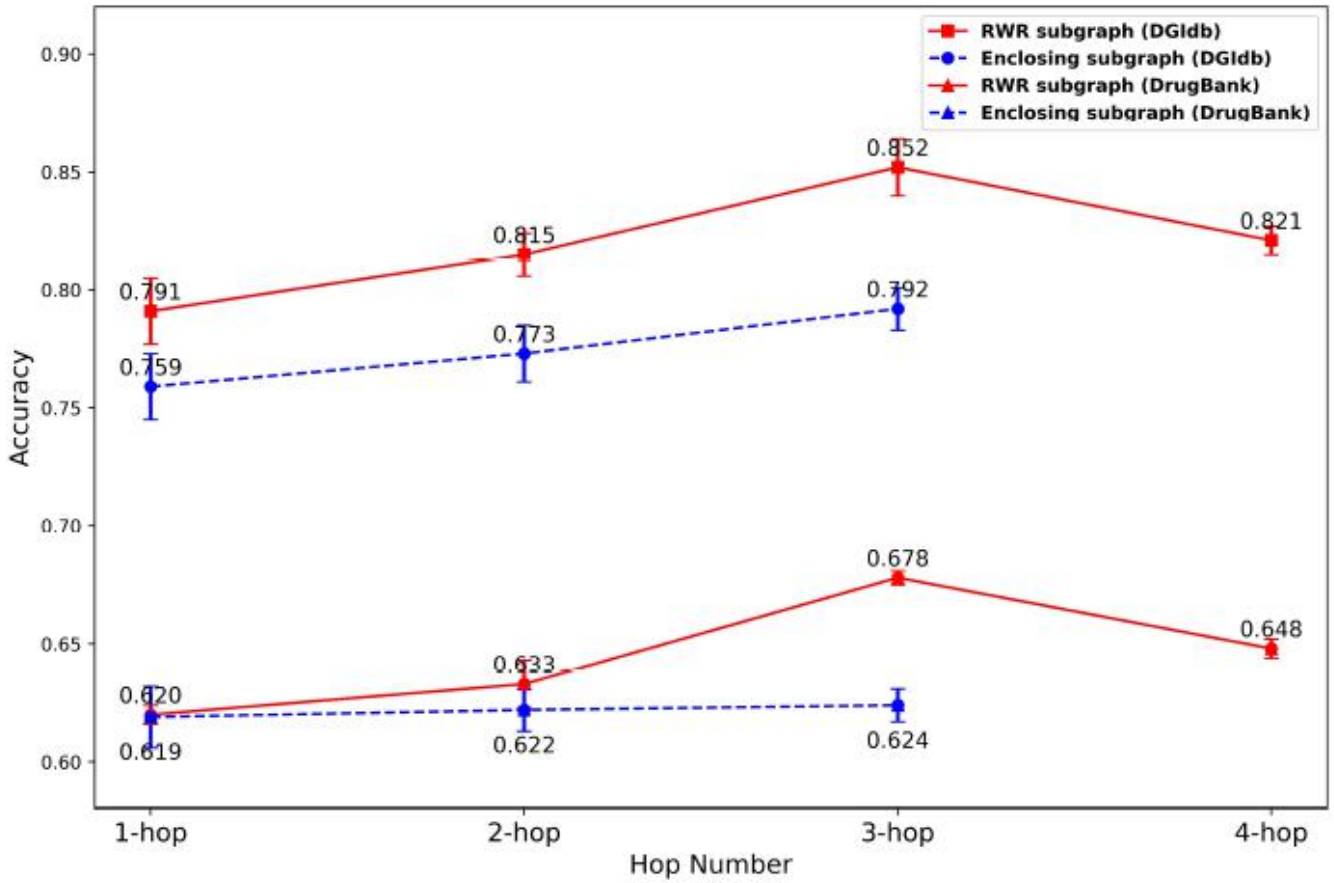


Figure 2: Evaluating the effect of Subgraph Extraction on DrugBank and DGIdb.

Experiments

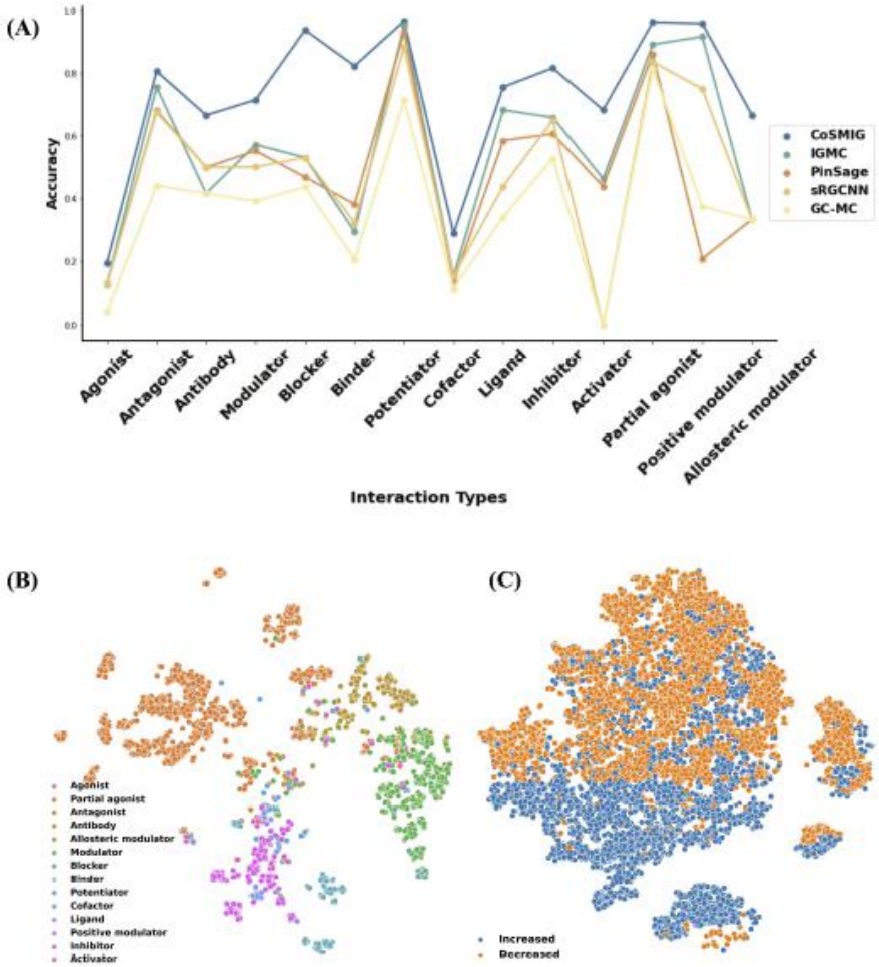


Figure 3: (A) The accuracy of each interaction type for DGIdb, and the pairs projected by t-SNE for (B) DGIdb and (C) DrugBank.

Thank you!