SemiGNN-PPI: Self-Ensembling Multi-Graph Neural Network for Efficient and Generalizable Protein-Protein Interaction Prediction

Ziyuan Zhao1,2,∗ , Peisheng Qian1,∗ , Xulei Yang1 , Zeng Zeng3 , Cuntai Guan2 ,
Wai Leong Tam4 and Xiaoli Li1,2
School of Computer Science and Engineering
Institute for Infocomm Research (I2R) , Genome Institute of Singapore (GIS),, A*STAR, Singapore
School of Microelectronics, Shanghai University, China
zeminliu@nus.edu.sg, {tknguyen, yfang}@smu.edu.sg

Code: https://github.com/jacobzhaoziyuan/jacobzhaoziyuan.github.io

—— IJCAI 2023

2023.6.18 • ChongQing

Reported by Jinyuan Zhang

Introduction

QUESTION:

Domainshift: existing methods are only developed and validated using in-distribution data 。

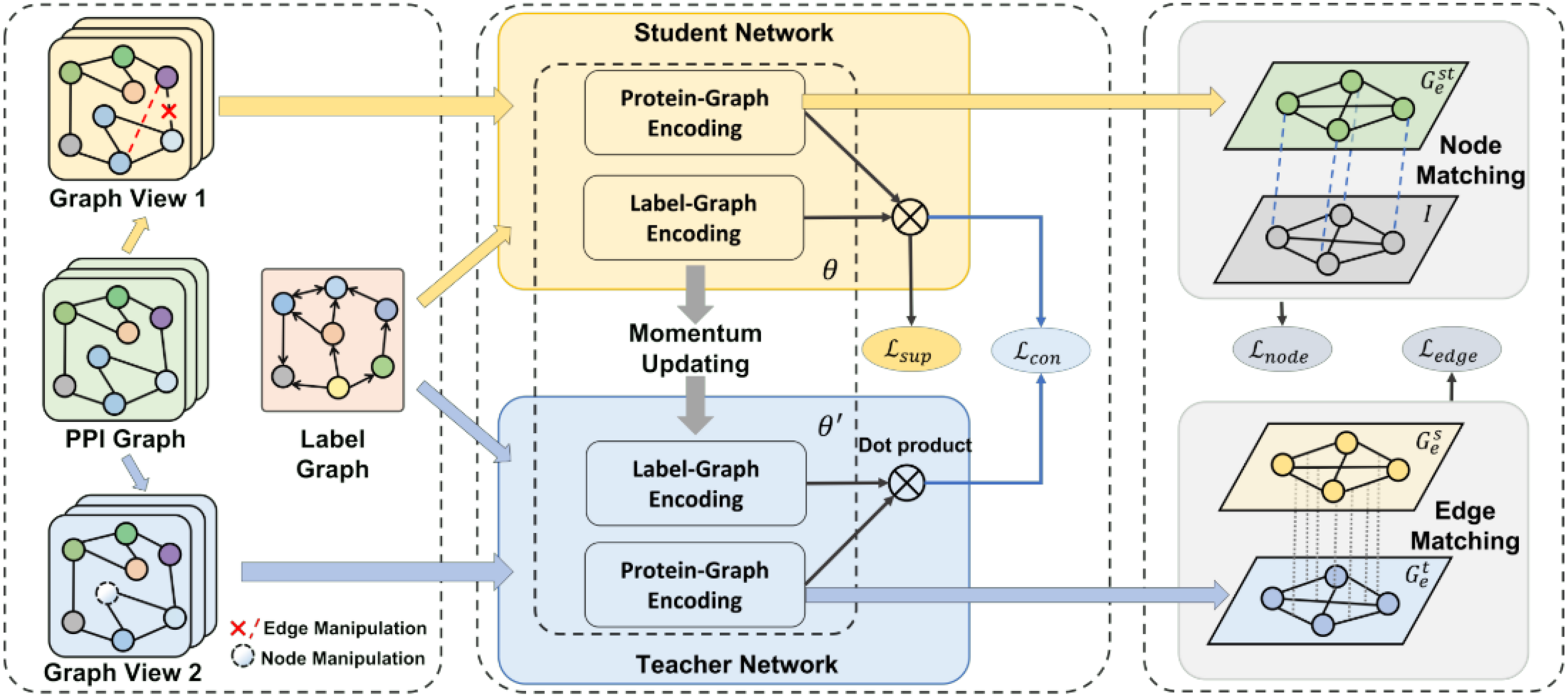
Label scarcity: many interactions still need to be annotated from experimental data， only a small portion of labeled samples can be used for model training

WORK:

1. propose an effective Self-ensembling multi-Graph Neural Network-based PPI prediction (SemiGNN-PPI) framework

2. combining GNN with Mean Teacher（SSL model）, to explore unlabeled data for self-ensemble graph learning and effectively utilize unlabeled data by consistency regularization with multiple constraints..

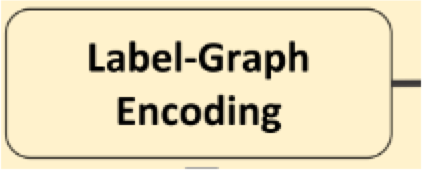
Overview


Graph View 1
PPI Graph
Label
Graph
Graph View 2
Edge Manipulation
Node Manipulation
Student Network
Protein-Graph
Encoding
Label-Graph
Encoding
θ
Momentum
Updating
θ′
Label-Graph
Encoding
Protein-Graph
Encoding
Teacher Network
Dot product
$\mathcal{L}_{sup}$
$\mathcal{L}_{con}$
G_e^{st}
Node
Matching
I
$\mathcal{L}_{node}$
$\mathcal{L}_{edge}$
G_e^s
Edge
Matching
G_e^t

Method

PPI graph Encoding(GNN)

$$h_p^{(l)} = \phi^{(l)}(h_p^{(l-1)}, f^{(l)}(\{h_p^{(l-1)} : u \in \mathcal{N}_k(p)\})), \quad (1)$$



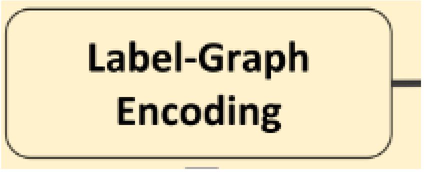
PPI graph Encoding(GNN+GIN+MLP)

$$h_p^{(l)} = g^l((1 + \epsilon^l) \cdot h_p^{(l-1)} + \sum\nolimits_{u \in \mathcal{N}_k(p)} h_u^{(l-1)}). \quad (2)$$

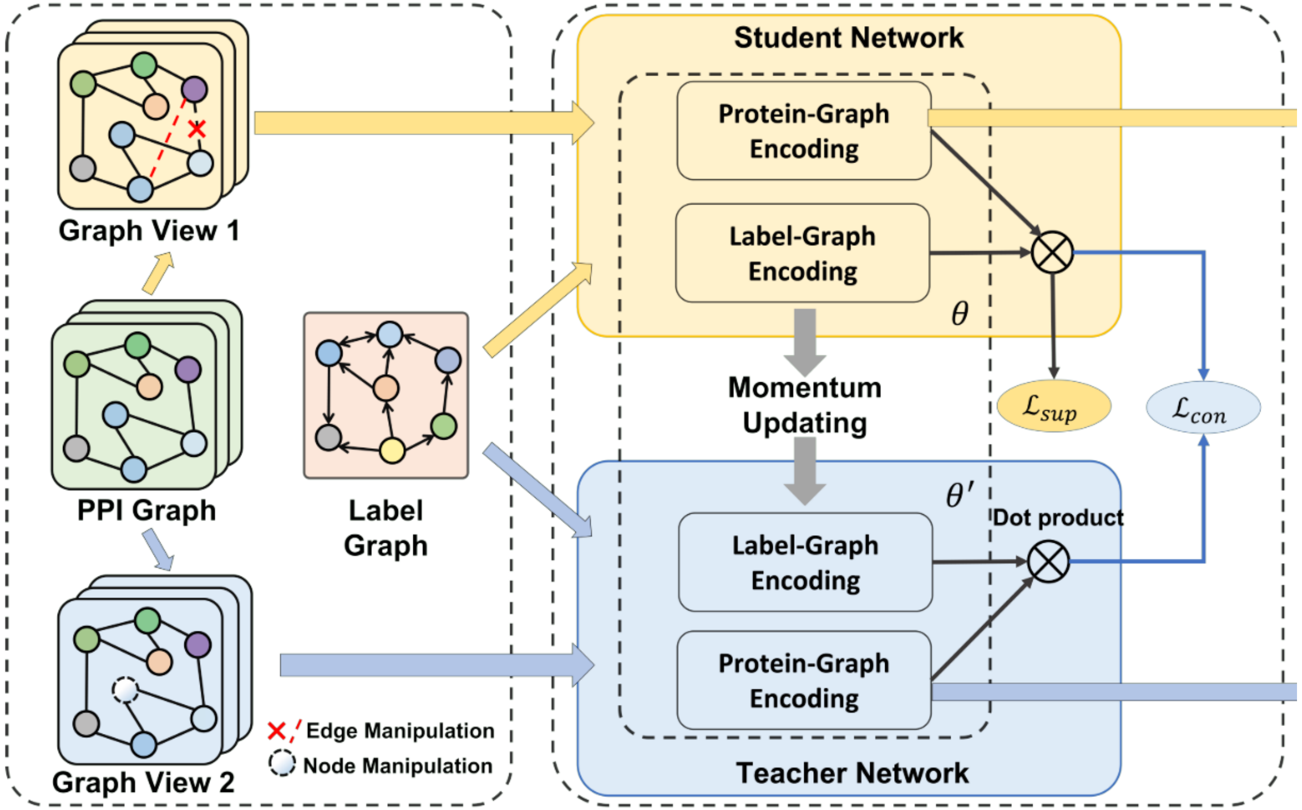
Label graph Encoding(GCN)

$$h_c^{(l+1)} = f(h_c^{(l)}, A), A \in \mathbb{R}^{t \times t}, \quad (3)$$

$$h_c^{(l+1)} = \delta\left(\widehat{A} h_c^{(l)} W^l\right), \quad (4)$$



Method



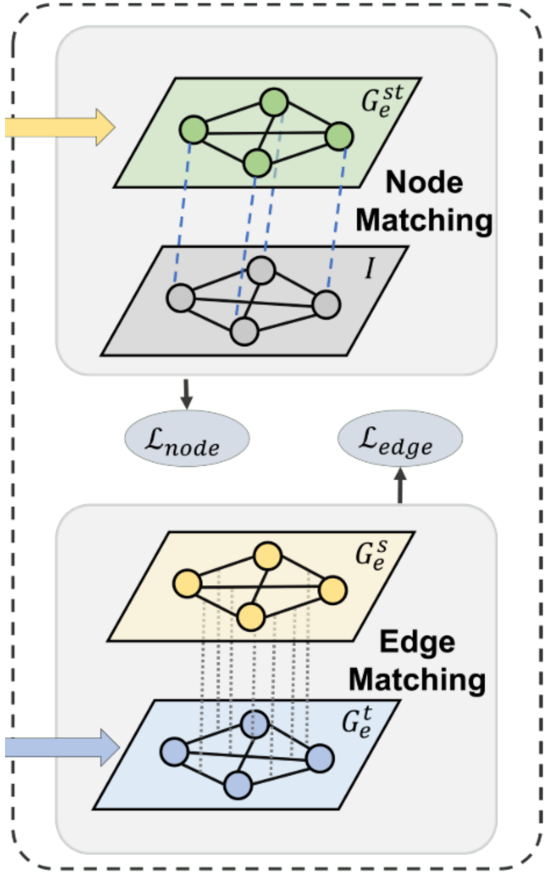
$$\hat{y}_{ij} = W(h_{p_i} \cdot h_{p_j}). \quad (5)$$

$$\mathcal{L}_{sup} = \sum_{c=1}^{t} \left(y^c \log\left(\sigma\left(\hat{y}^c\right)\right) + (1 - y^c)\log\left(1 - \sigma\left(\hat{y}^c\right)\right)\right)$$

$$\theta'_{\mathrm{k}} = m\theta'_{\mathrm{k}-1} + (1-m)\theta_{\mathrm{k}}, \quad (6)$$

$$\mathcal{L}_{con} = \|f_t(E_u|G, \theta'_k, \xi') - f_s(E_u|G, \theta_k, \xi)\|_2, \quad (7)$$

Method



$$\mathcal{L}_{node} = ||\text{diag}(\text{Adj}(G_e^{st})) - \text{diag}(I)||_2, \quad (9)$$

$$\mathcal{L}_{edge} = ||\text{Adj}(G_e^s) - \text{Adj}(G_e^t)||_2, \quad (8)$$

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{con}\mathcal{L}_{con} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{node}\mathcal{L}_{node}, \quad (10)$$

Experiments

Method		SHS27k			SHS148k			STRING		
		Random	DFS	BFS	Random	DFS	BFS	Random	DFS	BFS
ML	RF	$78.45_{0.88}$	$35.55_{2.22}$	$37.67_{1.57}$	$82.10_{0.20}$	$43.26_{3.43}$	$38.96_{1.94}$	$88.91_{0.08}$	$70.80_{0.45}$	$55.31_{1.02}$
	LR	$71.55_{0.93}$	$48.51_{1.87}$	$43.06_{5.05}$	$67.00_{0.07}$	$51.09_{2.09}$	$47.45_{1.42}$	$67.74_{0.16}$	$61.28_{0.53}$	$50.54_{2.00}$
DL	DPPI	$73.99_{5.04}$	$46.12_{3.02}$	$41.43_{0.56}$	$77.48_{1.39}$	$52.03_{1.18}$	$52.12_{8.70}$	$94.85_{0.13}$	$66.82_{0.29}$	$56.68_{1.04}$
	DNN-PPI	$77.89_{4.97}$	$54.34_{1.30}$	$48.90_{7.24}$	$88.49_{0.48}$	$58.42_{2.05}$	$57.40_{9.10}$	$83.08_{0.11}$	$64.94_{0.93}$	$53.05_{0.82}$
	PIPR	$83.31_{0.75}$	$57.80_{3.24}$	$44.48_{4.44}$	$90.05_{2.59}$	$63.98_{0.76}$	$61.83_{10.23}$	$94.43_{0.10}$	$67.45_{0.34}$	$55.65_{1.60}$
Graph	GNN-PPI	$87.91_{0.39}$	$74.72_{5.26}$	$63.81_{1.79}$	$92.26_{0.10}$	$82.67_{0.85}$	$71.37_{5.33}$	$95.43_{0.10}$	$91.07_{0.58}$	$78.37_{5.40}$
	GNN-PPI*	$88.87_{0.23}$	$75.68_{3.95}$	$68.84_{3.16}$	$92.13_{0.10}$	$83.77_{1.34}$	$69.02_{3.07}$	$94.94_{0.17}$	$90.62_{0.23}$	$79.76_{2.43}$
M-Graph	SemiGNN-PPI	$\mathbf{89.51_{0.46}}$	$\mathbf{78.32_{3.15}}$	$\mathbf{72.15_{2.87}}$	$\mathbf{92.40_{0.22}}$	$\mathbf{85.45_{1.17}}$	$\mathbf{71.78_{3.56}}$	$\mathbf{95.57_{0.08}}$	$\mathbf{91.23_{0.26}}$	$\mathbf{80.84_{2.05}}$

Table 1: Performance of SemiGNN-PPI and baseline methods over different datasets and data partition schemes. GNN-PPI: reported results in the original paper. GNN-PPI*: reproduced GNN-PPI results. The scores are presented in the format of mean_{std}.

Experiments

Method	STRING				SHS148k				SHS27k			
	5%	10%	20%	100%	5%	10%	20%	100%	5%	10%	20%	100%
Partition Scheme = Random												
GNN-PPI	$89.94_{0.29}$	$92.38_{0.51}$	$93.30_{0.56}$	$94.94_{0.17}$	$79.19_{0.67}$	$82.86_{0.49}$	$86.67_{0.22}$	$92.13_{0.10}$	$52.04_{3.32}$	$60.28_{12.26}$	$79.44_{1.19}$	$88.87_{0.23}$
Ours	**$90.55_{0.10}$**	**$92.66_{0.59}$**	**$93.90_{0.41}$**	**$95.57_{0.08}$**	**$79.50_{0.31}$**	**$83.48_{0.30}$**	**$87.38_{0.24}$**	**$92.40_{0.22}$**	**$57.97_{1.13}$**	**$62.67_{11.26}$**	**$81.01_{0.47}$**	**$89.51_{0.46}$**
Partition Scheme = DFS												
GNN-PPI	$86.60_{0.37}$	$87.91_{0.30}$	$89.42_{0.46}$	$90.62_{0.23}$	$68.77_{11.20}$	$78.36_{2.23}$	$80.96_{1.61}$	$83.77_{1.34}$	$53.41_{1.64}$	$58.43_{2.27}$	$65.73_{4.18}$	$75.68_{3.95}$
Ours	**$87.54_{0.06}$**	**$88.98_{0.26}$**	**$90.23_{0.12}$**	**$91.23_{0.26}$**	**$69.94_{9.57}$**	**$81.12_{0.98}$**	**$83.63_{0.86}$**	**$85.45_{1.17}$**	**$58.48_{1.11}$**	**$61.18_{1.98}$**	**$70.31_{2.38}$**	**$78.32_{3.15}$**
Partition Scheme = BFS												
GNN-PPI	$71.35_{4.67}$	$74.94_{2.35}$	$79.99_{2.75}$	$79.76_{2.43}$	$61.42_{3.29}$	$62.51_{3.07}$	$67.10_{3.48}$	$69.02_{3.07}$	$57.93_{4.11}$	$56.84_{12.19}$	$61.18_{6.58}$	$68.84_{3.16}$
Ours	**$73.35_{4.90}$**	**$76.94_{2.53}$**	**$81.39_{2.44}$**	**$80.84_{2.05}$**	**$64.86_{2.97}$**	**$68.76_{1.62}$**	**$71.06_{3.35}$**	**$71.78_{3.56}$**	**$60.15_{2.09}$**	**$66.13_{2.01}$**	**$67.69_{8.47}$**	**$72.15_{2.87}$**

Table 2: Performance comparison of different methods under different label ratios. The scores are presented in the format of $\mathrm{mean}_{\mathrm{std}}$.

Experiments

Method	% Labels	Random Partition			DFS Partition		BFS Partition	
	100	BS (92.66%)	ES (6.95%)	NS(0.39%)	ES (75.95%)	NS(24.05%)	ES (85.70%)	NS(14.30%)
GNN-PPI		89.17	72.44	50.00	77.81	63.44	71.03	44.80
SemiGNN-PPI		**89.68**	**72.93**	50.00	**81.75**	**66.32**	**75.14**	**57.00**
	20	BS (73.18%)	ES (24.98%)	NS (1.84%)	ES (72.87%)	NS (27.13%)	ES (47.71%)	NS (52.29%)
GNN-PPI		83.46	70.10	43.68	64.40	54.21	**59.04**	66.33
SemiGNN-PPI		**84.09**	**71.95**	**45.78**	**73.30**	**55.46**	58.10	**73.82**
	10	BS (55.80%)	ES (38.03%)	NS (6.16%)	ES (63.36%)	NS (36.64%)	ES (41.14%)	NS (58.86%)
GNN-PPI		79.64	69.64	38.41	56.13	53.85	36.02	47.89
SemiGNN-PPI		**80.22**	**70.33**	**41.67**	**61.07**	**57.90**	**57.39**	**72.73**
	5	BS (38.16%)	ES (47.61%)	NS (14.23%)	ES (46.63%)	NS (53.37%)	ES (43.18%)	NS (56.82%)
GNN-PPI		53.43	44.33	40.64	53.85	49.62	56.10	51.95
SemiGNN-PPI		**59.76**	**57.82**	**42.71**	**58.25**	**56.25**	**58.18**	**58.60**

Table 3: Analysis on performance between GNN-PPI and SemiGNN-PPI over BS, ES, and NS subsets in the SHS27k dataset. The ratios of the subsets are annotated in brackets. The BS subsets are empty under DFS and BFS partitions and are omitted for brevity.

Experiments

PPI Type	Type Ratio	Random Partition		DFS Partition		BFS Partition	
		GNN-PPI	SemiGNN-PPI	GNN-PPI	SemiGNN-PPI	GNN-PPI	SemiGNN-PPI
Reaction	40.61%	$89.58_{0.15}$	$\mathbf{90.16_{0.43}}$	$81.90_{1.65}$	$\mathbf{85.86_{0.71}}$	$61.62_{1.29}$	$\mathbf{64.92_{5.73}}$
Binding	52.71%	$88.28_{0.48}$	$\mathbf{89.46_{0.57}}$	$83.52_{1.41}$	$\mathbf{86.39_{0.67}}$	$70.00_{4.10}$	$\mathbf{72.43_{6.33}}$
Ptmod	20.99%	$87.04_{0.29}$	$\mathbf{87.42_{0.33}}$	$77.94_{1.67}$	$\mathbf{82.99_{1.44}}$	$65.92_{5.52}$	$\mathbf{71.32_{5.04}}$
Activation	42.51%	$85.15_{0.38}$	$\mathbf{85.26_{0.46}}$	$73.48_{2.74}$	$\mathbf{77.95_{1.19}}$	$67.44_{8.43}$	$\mathbf{68.04_{8.06}}$
Inhibition	20.20%	$87.21_{0.18}$	$\mathbf{88.09_{0.31}}$	$72.46_{1.11}$	$\mathbf{78.12_{2.62}}$	$60.20_{4.62}$	$\mathbf{67.71_{7.21}}$
Catalysis	44.67%	$89.36_{0.44}$	$\mathbf{90.35_{0.31}}$	$82.30_{0.80}$	$\mathbf{85.77_{1.29}}$	$65.70_{4.42}$	$\mathbf{73.39_{6.33}}$
Expression	7.69%	$\mathbf{47.85_{0.79}}$	$46.99_{0.22}$	$\mathbf{34.96_{3.74}}$	$32.45_{5.96}$	$\mathbf{31.81_{6.87}}$	$28.99_{4.90}$
Macro-Average	-	$82.07_{0.39}$	$\mathbf{82.53_{0.38}}$	$72.37_{1.87}$	$\mathbf{74.16_{2.09}}$	$60.38_{5.03}$	$\mathbf{63.29_{5.29}}$
Micro-Average	-	$86.67_{0.22}$	$\mathbf{87.38_{0.24}}$	$80.96_{1.61}$	$\mathbf{83.63_{0.86}}$	$67.10_{3.48}$	$\mathbf{71.06_{3.35}}$

Table 4: Per-class results in the SHS148k dataset with 20% training labels. The type ratios are calculated over the whole dataset.

Experiments

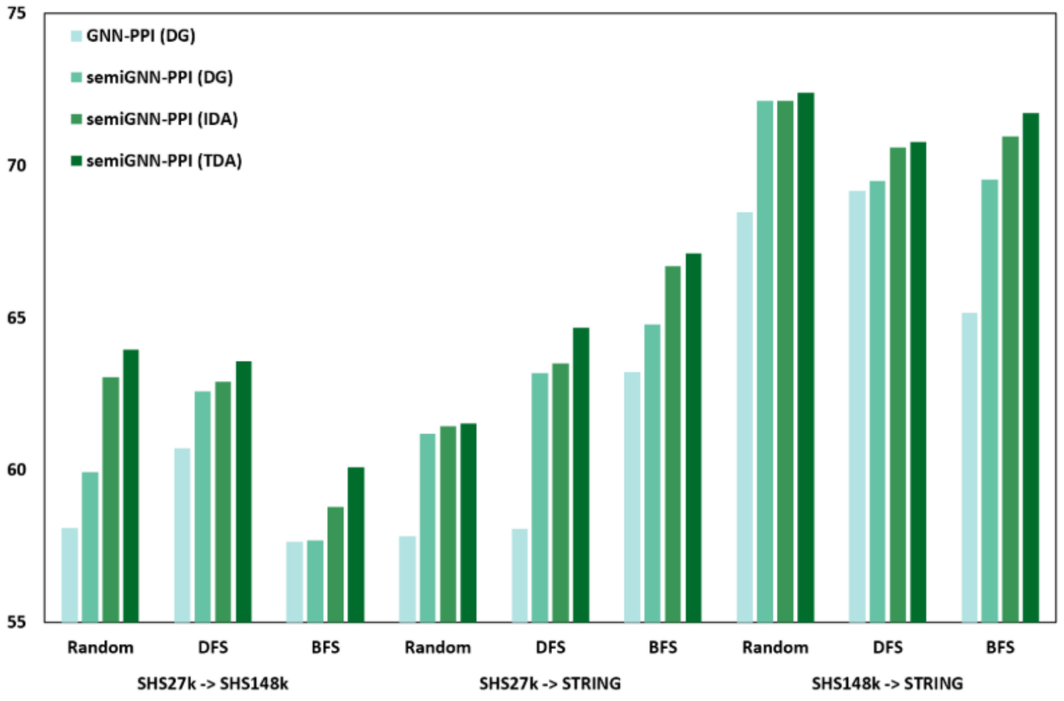


Figure 2: Performance comparison on trainset-heterologous test-sets. DG: domain generalization. IDA: inductive domain adaptation. TDA: transductive domain adaptation.

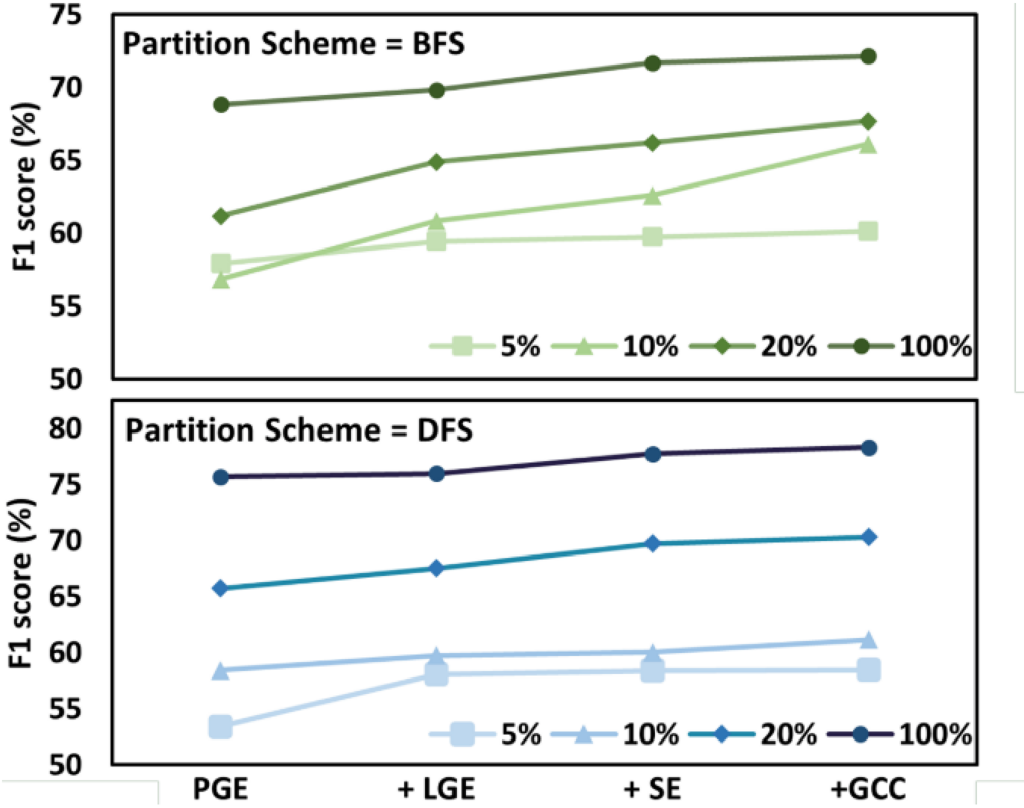


Figure 3: Results of ablation studies on different components of SemiGNN-PPI using the SHS27k dataset.

THANKS