Prototypical Graph Contrastive Learning

Shuai Lin, Chen Liu, and Xiaodan Liang
School of Intelligent Systems Engineering, Sun Y at-sen University

Pan Zhou
Sea AI Laboratory, Singapore
panzhou3@gmail.com

Zi-Yuan Hu and Liang Lin
School of Computer Science and Engineering, Sun Y at-sen University

Shuojia Wang, Ruihui Zhao, and Yefeng Zheng
Tencent Jarvis Lab, Shenzhen

Eric Xing
School of Computer Science, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Code: https://github.com/ha-lins/PGCL

(IEEE-2022)

Reported by Rongrong Ma

1. Introduction

2. Approach

3. Experiments

Introduction

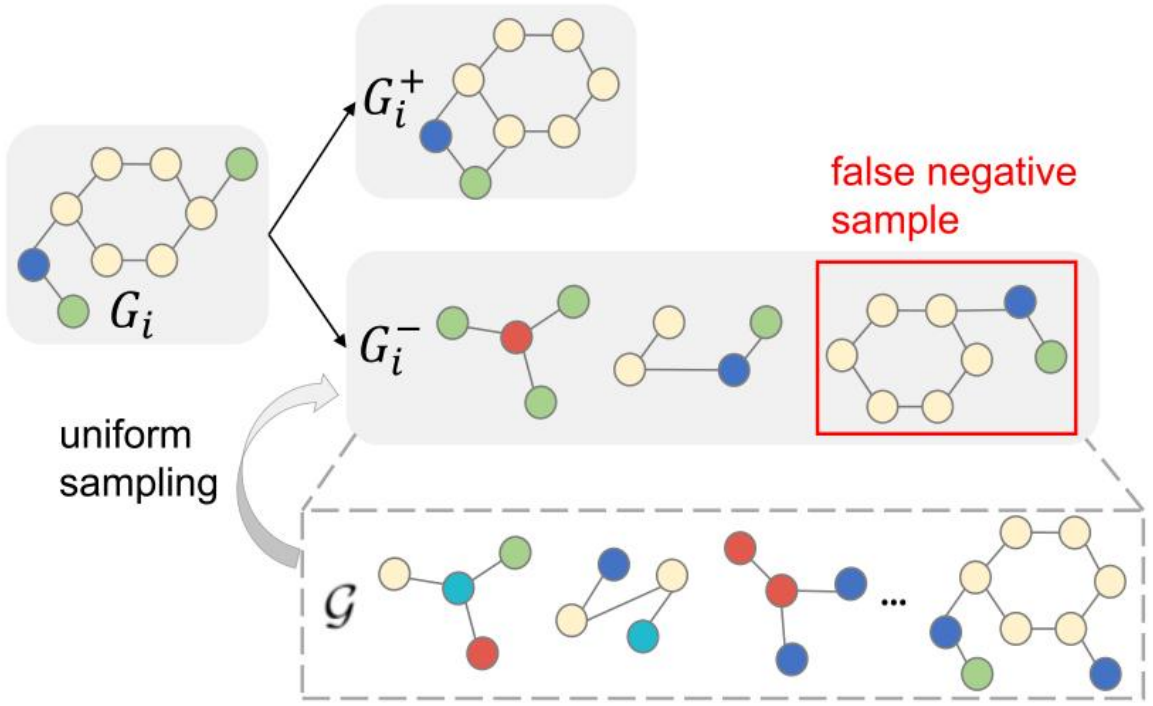


Fig. 1. "Sampling bias": the strategy of sampling negative examples uniformly from the data distribution $\mathcal{G}$ could result in that the sampled negatives G_i^- are semantically similar to the query G_i, e.g., they all contain the hexagonal structure that resembles a benzene ring.

For a query, its negatives are uniformly sampled from all graphs, existing methods suffer from the critical sampling bias issue, i.e., the negatives likely having the same semantic structure with the query, leading to performance degradation.

Method

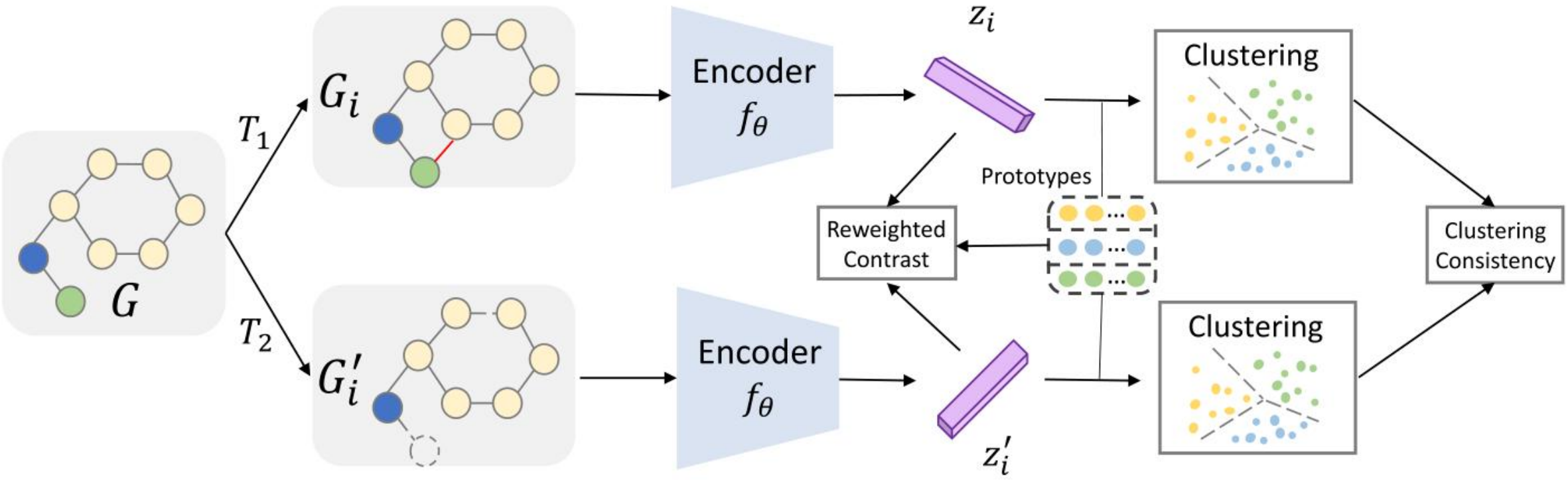


Fig. 2. Overview of PGCL. Two graph data augmentations T_1 and T_2 are applied to the input graph G. Then, two graph views G_i and G_i' are fed into the shared encoder f_θ (including GNNs and a projection head) to extract the graph representations z_i and z_i'. We perform the online clustering via assigning the representations of samples within a batch to prototype vectors (cluster centroids). The representations are learned via encouraging the clustering consistency between correlated views (Section IV-A) and a reweighted contrastive objective (Section IV-B), where prototype vectors are also updated along with the encoder parameters by backpropagation.

Method

Self-supervised Graph-level Representation Learning with Local and Global Structure

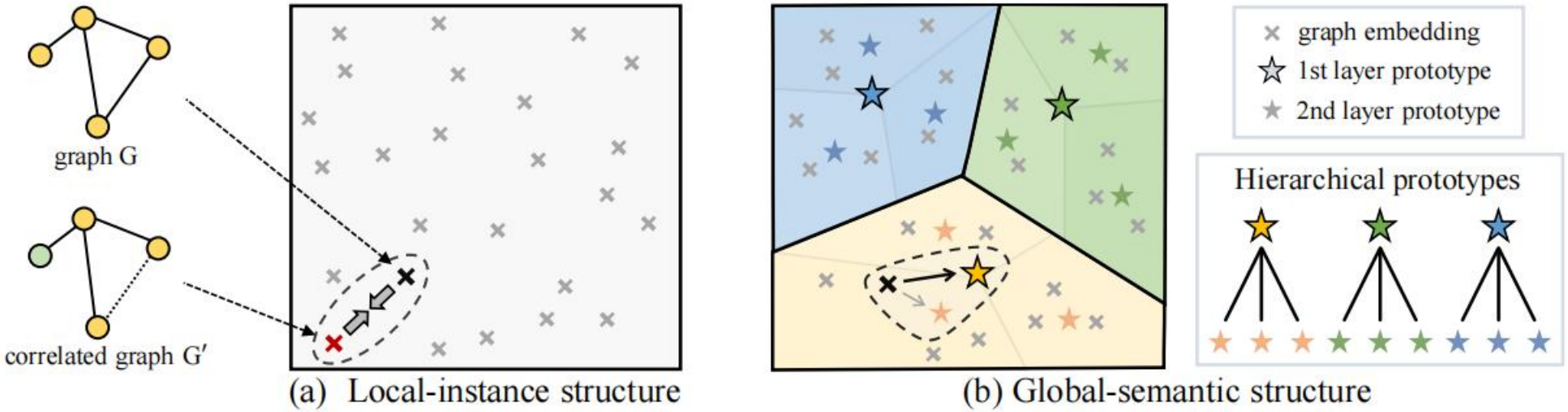


Figure 1. **Illustration of GraphLoG.** (a) Correlated graphs are constrained to be adjacently embedded to pursue the local-instance structure of the data. (b) Hierarchical prototypes are employed to discover and refine the global-semantic structure of the data.

Method

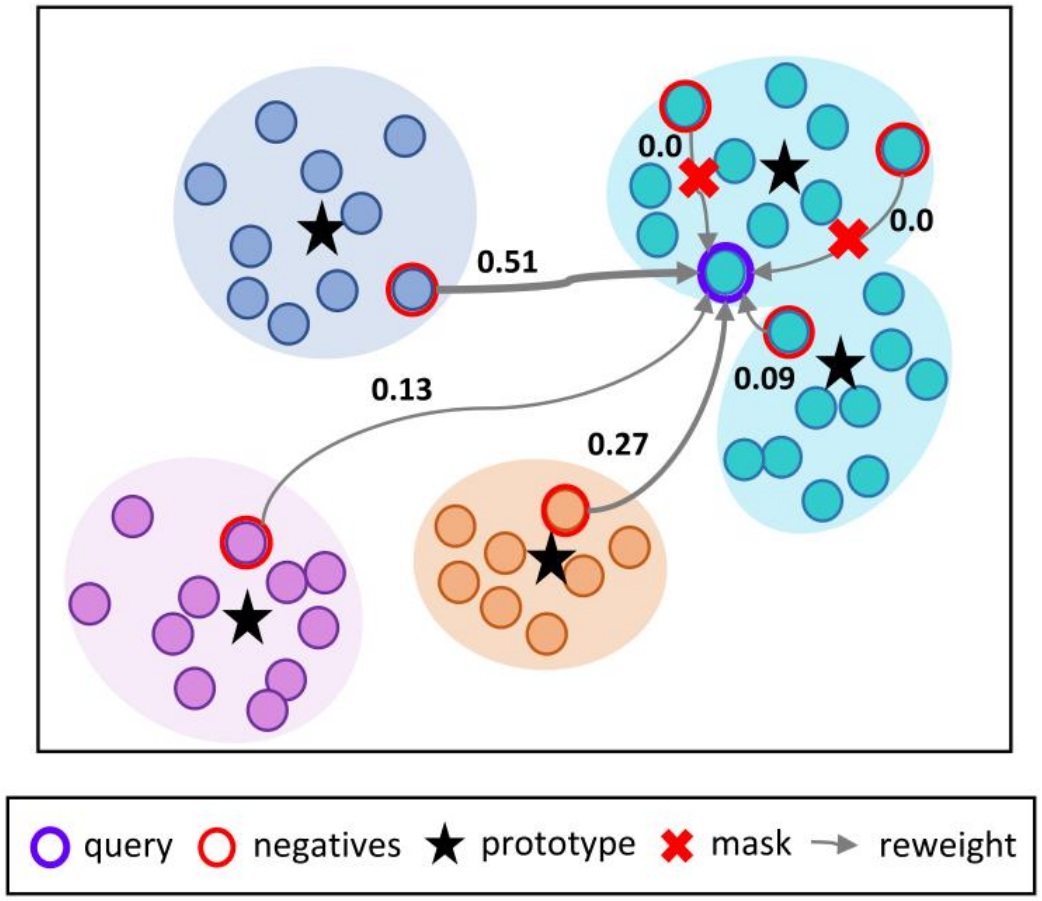
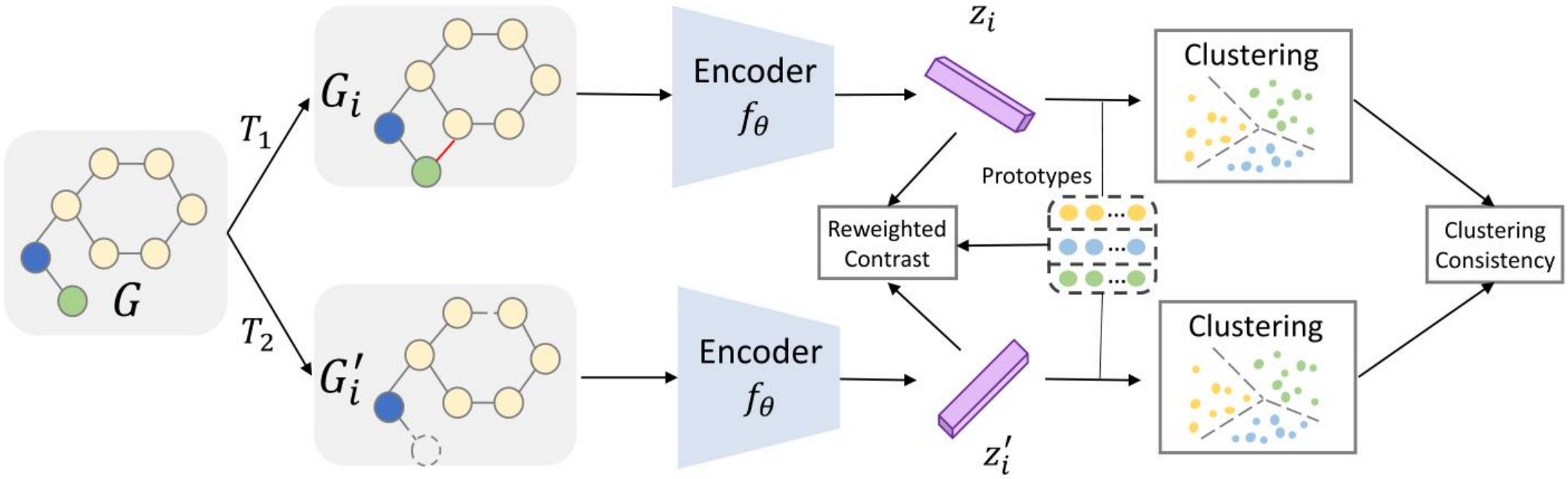


Fig. 3. Illustration of the negative sample reweighting. The linewidth of the arrow denotes the weight value.

Method



GNN

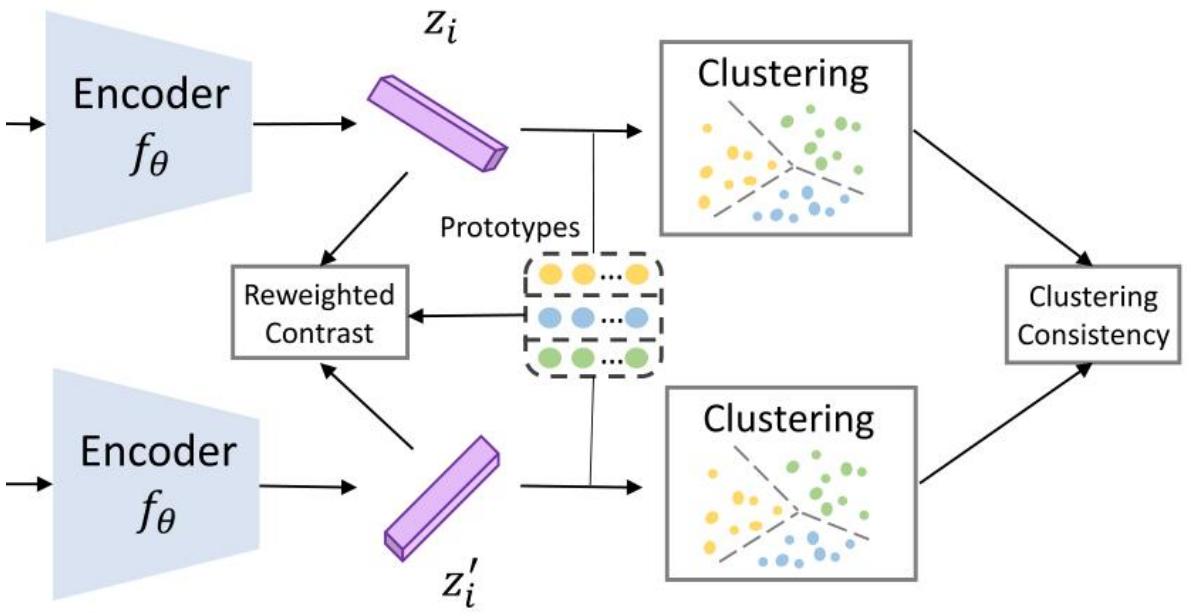
$$\mathbf{h}_v^k = \mathrm{COMBINE}^k\left(\mathbf{h}_v^{k-1}, \mathrm{AGGREGATE}_k\left(\left\{\mathbf{h}_u^{k-1} \forall u \in \mathcal{N}(v)\right\}\right)\right). \quad (1)$$

$$f_\theta(G_i) = \mathrm{READOUT}\left(\left\{\mathrm{CONCAT}\left(\left\{\mathbf{h}_j^k\right\}_{k=1}^K\right)\right\}_{j=1}^N\right) \quad (2)$$

GCL

$$\mathcal{L}_{\mathbf{InfoNCE}} = -\sum_{i=1}^{n} \log \frac{\exp\left(z_i \cdot z_i'/\tau\right)}{\exp\left(z_i \cdot z_i'/\tau\right) + \sum_{j=1, j\neq i}^{2N} \exp\left(z_i \cdot z_j/\tau\right)} \quad (3)$$

Method



C={c1, . . . , cK } a set of K trainable prototype vectors .

$$p(y|z_i) = \text{softmax}(C \cdot f_\theta(G_i)). \tag{4}$$

$$\ell(p_i, q_{i'}) = -\sum_{y=1}^{K} q\left(y|z_i'\right) \log p(y|z_i) \tag{5}$$

$$\mathcal{L}_{\text{consistency}} = \sum_{i=1}^{n} [\ell(p_i, q_{i'}) + \ell(p_{i'}, q_i)]. \tag{6}$$

$$\min_{p,q} \mathcal{L}_{\text{consistency}}$$

$$\text{s.t. } \forall y : q(y|z_i) \in [0, 1] \text{ and } \sum_{i=1}^{N} q(y|z_i) = \frac{N}{K}. \tag{7}$$

$$P = \frac{1}{N} p(y|z_i); \quad Q = \frac{1}{N} q(y|z_i). \tag{8}$$

$$T = \left\{ Q \in \mathbb{R}_+^{K \times N} \mid Q\mathbb{1}_N = \frac{1}{K}\mathbb{1}_K, Q^\top \mathbb{1}_K = \frac{1}{N}\mathbb{1}_N \right\} \tag{9}$$

$$\min_{p,q} \mathcal{L}_{\text{consistency}} = \min_{Q \in \mathbf{T}} \langle Q, -\log P \rangle - \log N \tag{10}$$

$$Q = \text{Diag}(\alpha) P^\eta \text{Diag}(\beta) \tag{11}$$

Method

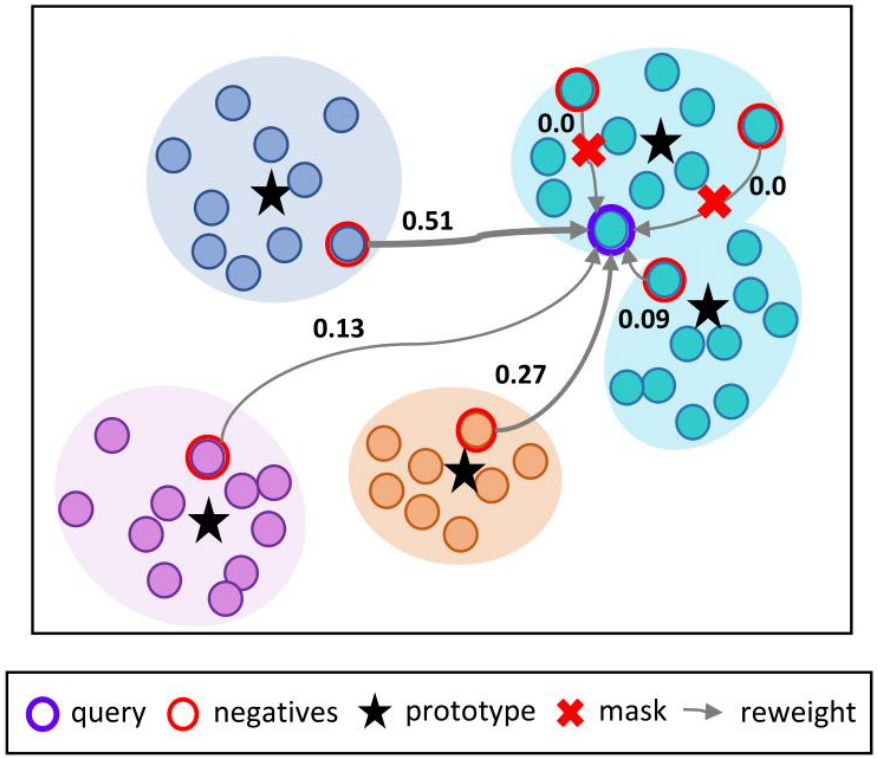


Fig. 3. Illustration of the negative sample reweighting. The linewidth of the arrow denotes the weight value.

Reweighted Contrastive Objective

$$\mathcal{L} = -\sum_{i=1}^{n} \log \frac{\exp\left(z_i \cdot z_i'/\tau\right)}{\exp\left(z_i \cdot z_i'/\tau\right) + \sum_{j=1, j\neq i}^{2N} \mathbb{1}_{c_i \neq c_j} \cdot \exp\left(z_i \cdot z_j'/\tau\right)} \tag{12}$$

$$\mathcal{L}_{\text{Reweighted}} = -\sum_{i=1}^{n} \times \log \frac{\exp\left(z_i \cdot z_i'/\tau\right)}{\exp\left(z_i \cdot z_i'/\tau\right) + M_i \sum_{j=1, j\neq i}^{2N} \mathbb{1}_{c_i \neq c_j} \cdot \boldsymbol{w}_{ij} \cdot \exp\left(z_i \cdot z_j'/\tau\right)} \tag{13}$$

$$\boldsymbol{w}_{ij} = \exp\left\{-\frac{\left[\mathcal{D}(c_i, c_j) - \mu_i\right]^2}{2\sigma_i^2}\right\} \tag{14}$$

$$\mathcal{D}(c_i, c_j) = 1 - (c_i \cdot c_j)/(\|c_i\|_2 \|c_j\|_2).$$

$M_i = (2N/(\sum_{j=1}^{2N} \boldsymbol{w}_{ij}))$ is the normalization factor.

$$\mathcal{L} = \mathcal{L}_{\text{Reweighted}} + \lambda \mathcal{L}_{\text{Consistency}} \tag{15}$$

Experiments

GRAPH CLASSIFICATION ACCURACIES (%) OF KERNEL, SUPERVISED, AND UNSUPERVISED METHODS. WE REPORT THE MEAN TENFOLD CROSS-VALIDATION ACCURACY WITH FIVE RUNS. "> 1 DAY" REPRESENTS THAT THE COMPUTATION EXCEEDS 24 h

	Datasets	MUTAG	PTC	PROTEINS	NCI1	COLLAB	RDT-B	RDT-M5K
Datasets	# graphs	188	344	1113	4110	5000	2000	2000
	Avg # nodes	17.9	14.3	39.1	29.9	74.5	429.6	429.6
Supervised	GRAPHSAGE [39]	85.1±7.6	63.9±7.7	75.9±3.2	77.7±1.5	-	-	-
	GCN [11]	85.6±5.8	64.2±4.3	76.0±3.2	80.2±2.0	79.0±1.8	50.0±0.0	20.0 ± 0.0
	GIN-0 [12]	**89.4±5.6**	**64.6±7.0**	**76.2±2.8**	**82.7±1.7**	**80.2±1.9**	**92.4±2.5**	**57.5±1.5**
	GIN-ϵ [12]	89.0±6.0	63.7±8.2	75.9±3.8	**82.7±1.6**	80.1±1.9	92.2±2.3	57.0±1.7
Kernel	GL [52]	81.7±2.1	57.3±1.4	-	53.9±0.4	**56.3±0.6**	77.3±0.2	41.0±0.2
	WL [21]	80.7±3.0	58.0±0.5	72.9±0.6	80.0±0.5	-	68.8±0.4	46.1±0.2
	DGK [53]	87.4±2.7	60.1±2.6	**73.3±0.8**	**80.3±0.5**	-	**78.0±0.4**	41.3±0.2
	MLG [23]	87.9±1.6	**63.3±1.5**	41.2±0.0	>1 Day	>1 Day	63.3±1.5	**57.3±1.4**
	GCKN [54]	**87.2±6.8**	-	50.8±0.8	70.6±2.0	54.3±1.0	58.4±7.6	**57.3±1.4**
Unsupervised	GRAPH2VEC [55]	83.2±9.3	60.2±6.9	73.3±2.1	73.2±1.8	-	75.8±1.0	47.9±0.3
	INFOGRAPH [13]	89.0±1.1	61.7±1.7	74.4±0.3	73.8±0.7	67.6±1.2	82.5 ±1.4	53.5±1.0
	MVGRL [16]	89.7±1.1	62.5±1.7	-	75.0±0.7	68.9±1.9	84.5±0.6	-
	GCC [25]	86.4±0.5	58.4±1.2	72.9±0.5	66.9±0.2	75.2±0.3	88.4±0.3	52.6±0.2
	GRAPHCL [18]	86.8±1.3	58.4±1.7	74.4±0.5	77.9±0.4	71.4±1.2	89.5±0.8	56.0±0.3
	PGCL (ours)	**91.1±1.2**	**63.3±1.3**	**75.7±0.2**	**78.8±0.8**	**76.0±0.3**	**91.5±0.7**	**56.3±0.2**

Experiments

Transfer Learning

TRANSFER LEARNING PERFORMANCE FOR CHEMICAL MOLECULES PROPERTY PREDICTION (MEAN ROC-AUC ± STD. OVER TEN RUNS). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Downstream Dataset	BBBP	Tox21	SIDER	ToxCast	ClinTox	BACE	MUV	Average
#Molecules	2039	7831	1427	8575	1478	1513	93087	Rank
#Tasks	1	12	27	617	2	1	17	(↓)
No Pre-Train	65.8 ± 4.5	74.0 ± 0.8	57.3 ± 1.6	63.4 ± 0.6	58.0 ± 4.4	70.1 ± 5.4	71.8 ± 2.5	6.1
EdgePred [68]	67.3 ± 2.4	76.0 ± 0.6	60.4 ± 0.7	64.1 ± 0.6	64.1 ± 3.7	**79.9 ± 0.9**	74.1 ± 2.1	3.7
AttrMasking [68]	64.3 ± 2.8	**76.7 ± 0.4**	61.0 ± 0.7	64.2 ± 0.5	71.8 ± 4.1	79.3 ± 1.6	74.7 ± 1.4	2.9
ContextPred [68]	68.0 ± 2.0	75.7 ± 0.7	60.9 ± 0.6	63.9 ± 0.6	65.9 ± 3.8	79.6 ± 1.2	**75.8 ± 1.7**	3.1
InfoGraph [13]	68.8 ± 0.8	75.3 ± 0.5	58.4 ± 0.8	62.7 ± 0.4	69.9 ± 3.0	75.9 ± 1.6	75.3 ± 2.5	4.4
GraphCL [18]	69.7 ± 0.7	73.9 ± 0.7	60.5 ± 0.9	62.4 ± 0.6	**75.9 ± 2.7**	75.4 ± 1.4	69.8 ± 2.7	5.0
PGCL (Ours)	**69.8 ± 1.3**	75.6 ± 0.5	**61.6 ± 1.1**	**66.4 ± 0.2**	69.4 ± 1.4	79.3 ± 1.5	71.2 ± 1.3	**2.9**

Experiments

Ablation Studies

ABLATION STUDY FOR DIFFERENT OBJECTIVE FUNCTIONS ON DOWNSTREAM GRAPH CLASSIFICATION DATASETS. AS TWO VARIANTS OF THE VANILLA INFONCE LOSS, $\mathcal{L}_{\mathbf{S.R.}}$ DENOTES CALCULATING THE WEIGHT IN (14) WITH THE SAMPLE DISTANCE, WHILE $\mathcal{L}_{\mathbf{P.R.}}$ CORRESPONDS TO THE PROTOTYPE DISTANCE

$\mathcal{L}_{\text{Inf.}}$	$\mathcal{L}_{\text{Con.}}$	$\mathcal{L}_{\text{S.R.}}$	$\mathcal{L}_{\text{P.R.}}$	MUTAG	PTC	PRO.	COLLAB
✓				86.8±1.3	58.4±1.7	74.4±0.5	71.4±1.2
	✓			89.7±1.0	61.1±1.7	75.4±0.4	71.5±1.4
		✓		89.9±1.1	61.9±0.9	73.4±0.6	72.6±0.5
			✓	90.1±0.9	62.5±0.7	75.2±0.4	73.3±0.7
✓	✓			89.9±1.0	62.4±2.1	75.4±0.3	73.3±1.2
	✓	✓		91.0±1.4	**63.4±1.5**	73.6±1.1	74.6±0.6
	✓		✓	**91.1±1.2**	63.3±1.3	**75.7±0.2**	**76.0±0.3**

Experiments

Sensitivity Analysis

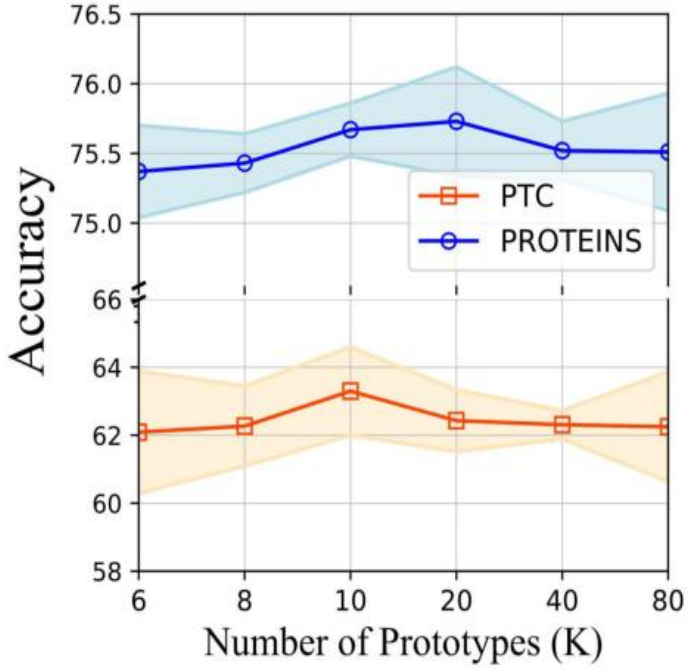


Fig. 4. Sensitivity analysis for the number of prototypes K.

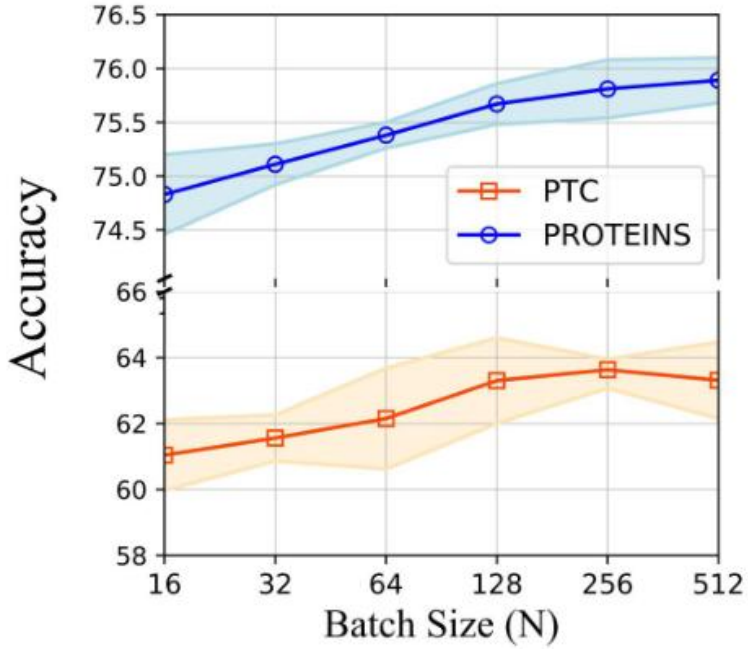


Fig. 5. Sensitivity analysis for batch size N.

Experiments

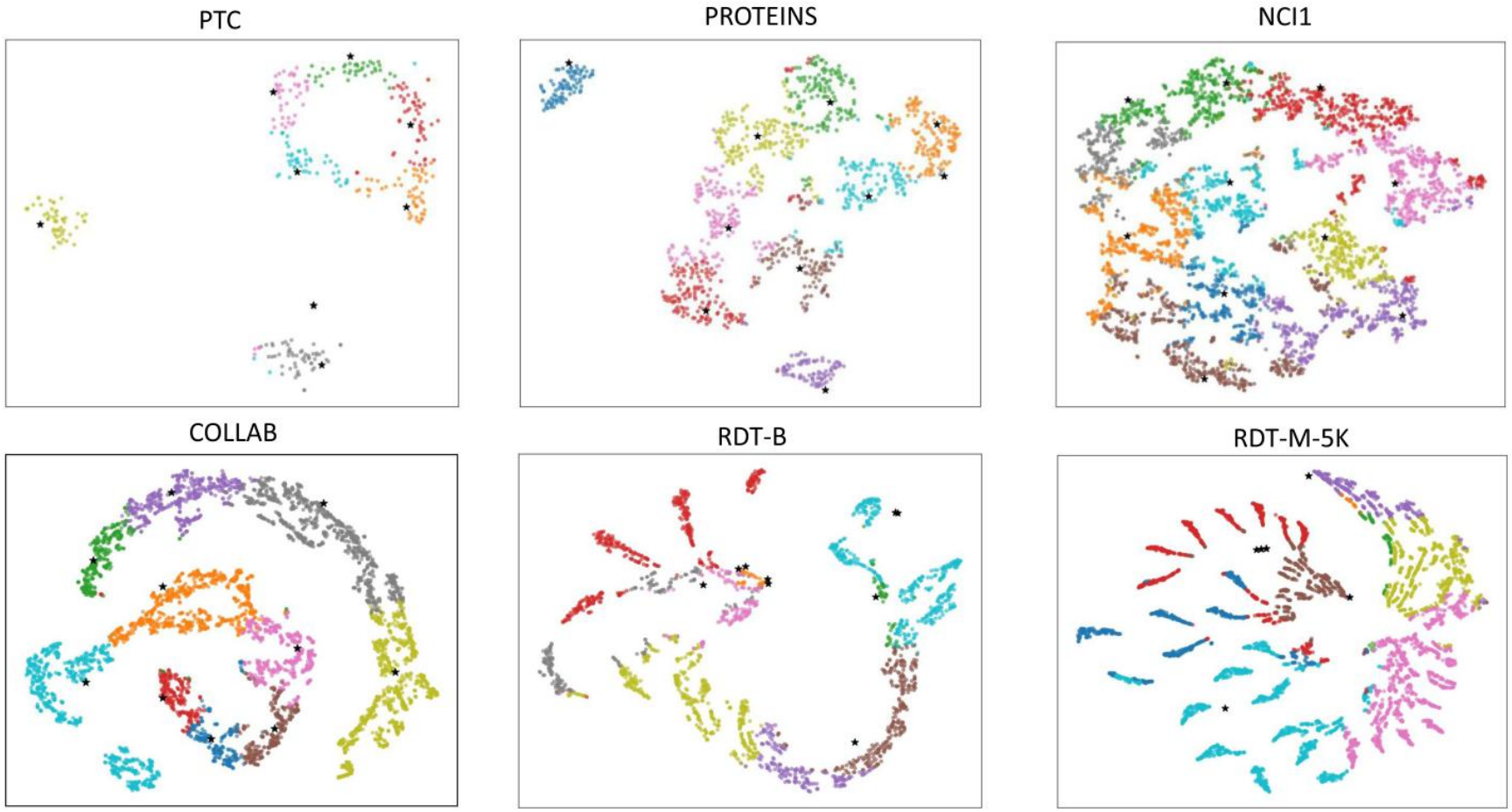


Fig. 6. T-SNE visualization of the learned representation on six datasets. "⋆" means the prototype vectors. Colors represent underlying classes that PGCL discovers.

Thank you !