UNIVERSITY OF NOTRE DAME

Brandeis UNIVERSITY

From Coarse to Fine: Enable Comprehensive Graph Self-supervised Learning with Multi-granular Semantic Ensemble

Author: Qianlong Wen, Mingxuan Ju, Zhongyu Ouyang,
Chuxu Zhang, Yanfang Ye

cerebras

ICML
International Conference
On Machine Learning

CONTENTS

1 Motivation

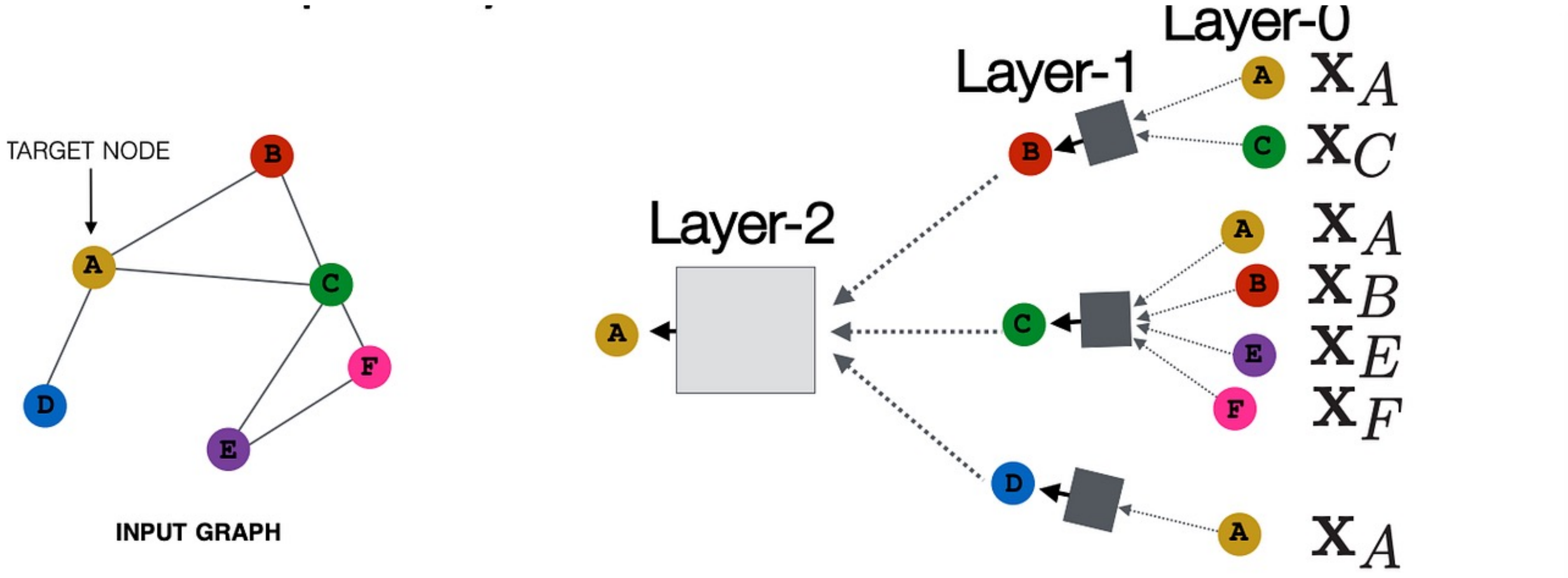
2 Challenge

3 Methodology

4 Experiment

Graph Neural Networks (GNNs)

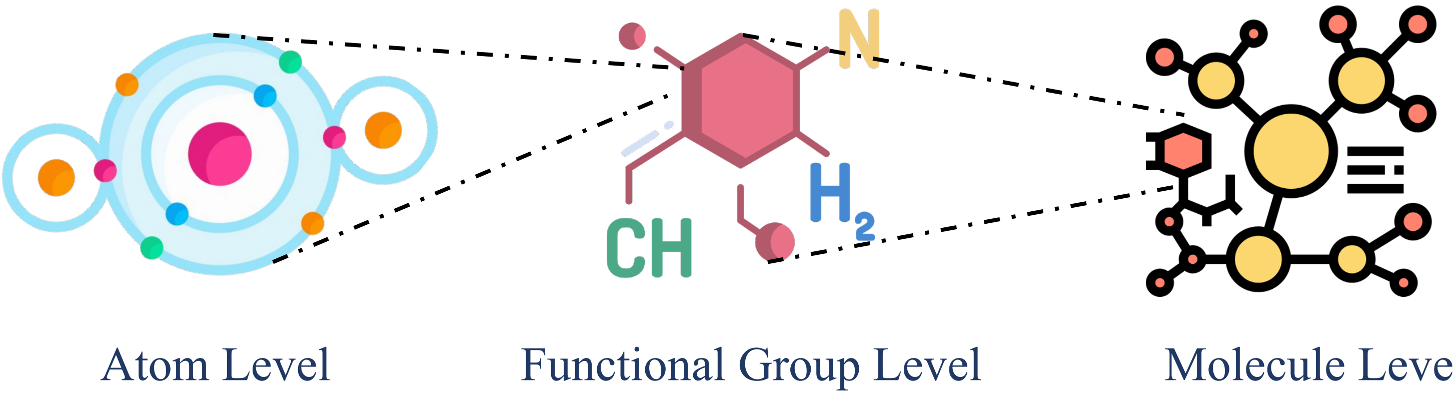
$$f_w(X, A) = \mathrm{Softmax}(\hat{A}\ \mathrm{ReLu}(\hat{A}\ X\ W_1)\ W_2),$$



The aggregated local neighborhood can usually represent the graph semantics.

Graph Semantics (Substructures) in Different Granularities

There could exist different underlying patterns within a single graph, which makes it challenging to train a single model to fits well on the data. And those knowledge could be in diversifying granularities, which further increase the difficulties.

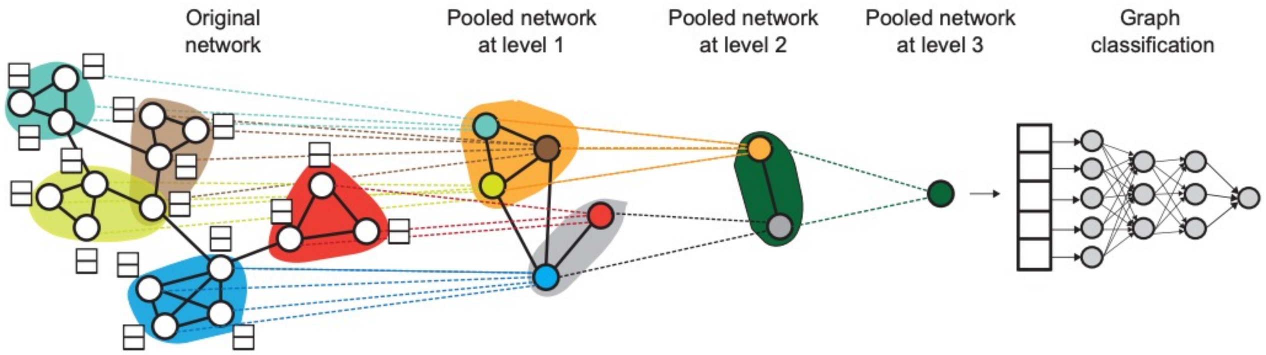


Each level can used for the prediction tasks of different properties. Therefore, we need a model with larger capacity and higher expressiveness power on graph substructure learning so that it can extract the knowledge in different levels for better generalization ability.

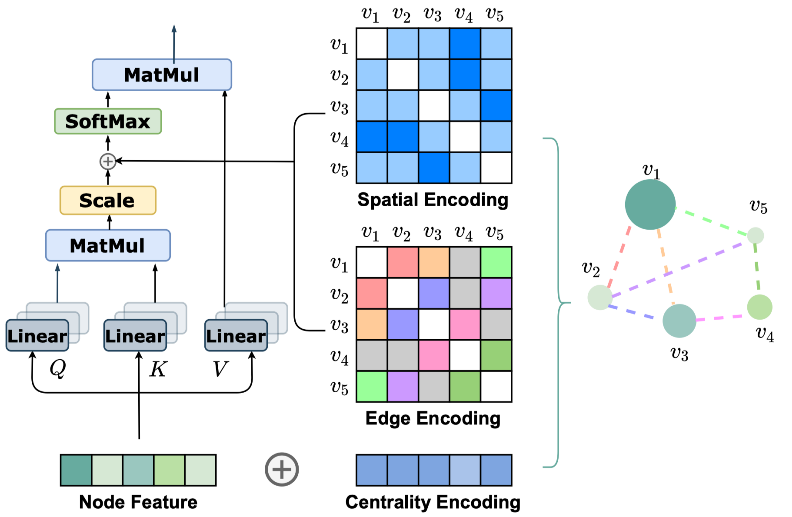
How to Improve the Expressiveness Power on Graph Substructures?

Basically, current research on improving the expressive power of deep graph models on graph substructure learning mainly focus on modifying model architectures to improve WL-Test,

- Hierarchical pooling operations: SubGNN, DiffPool and etc.
- Transformer backbone: Graphormer, SAN, GraphGPS and etc.


Original network
Pooled network at level 1
Pooled network at level 2
Pooled network at level 3
Graph classification

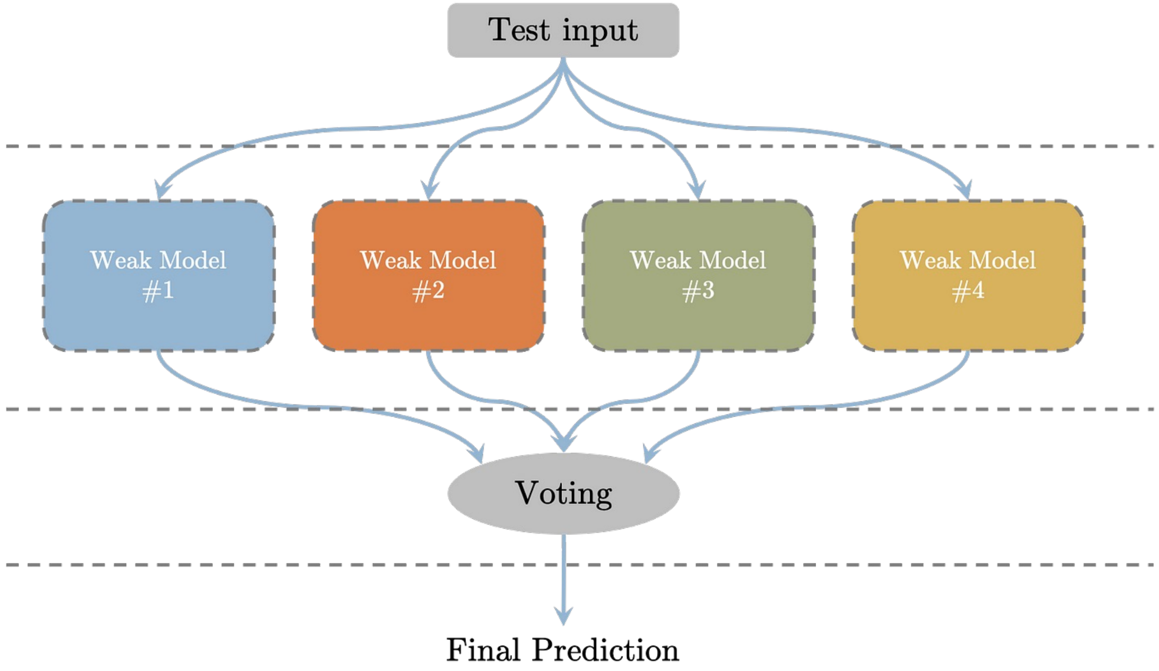
[NeurIPS 2018] Hierarchical Graph Representation Learning with Differentiable Pooling.



[NeurIPS 2021]Do Transformers Really Perform Bad for Graph Representation?

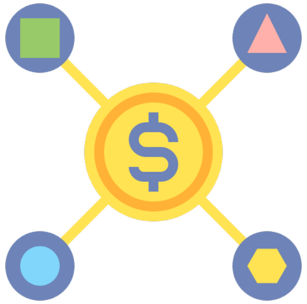
How to Improve the Expressiveness Power on Graph Substructures?

Meanwhile, another solution could be combining existing deep graph models into the ensemble learning and train each classifier to learn the graph semantics in different levels.

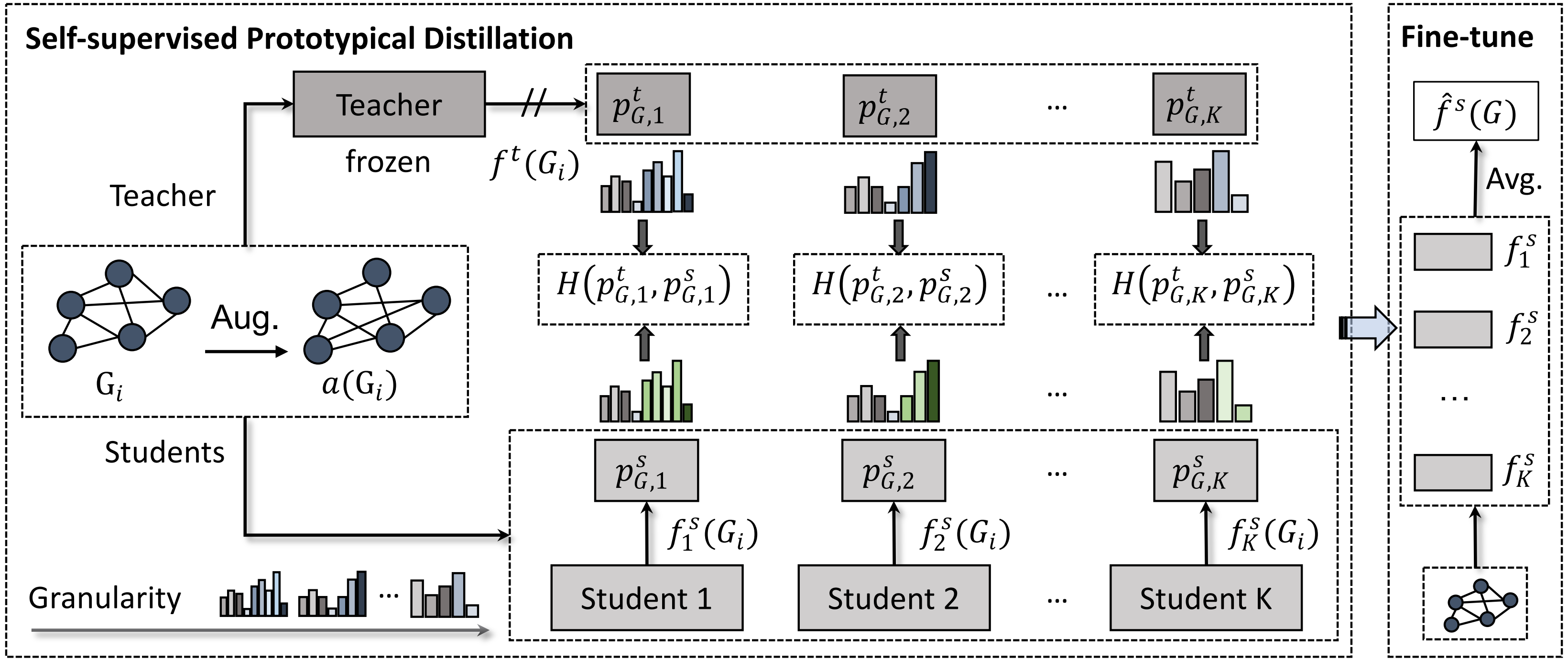


Challenges on Incorporating Ensemble Learning

- Higher Training and Inference Cost: Ensemble learning will naturally introduce extra computation cost, so the proposed framework should not be too expensive for training and inference.
- Loading Balance: Ensuring that all classifiers are utilized effectively can be challenging, as some of them may become over-utilized while others remain under-utilized.
- Regularization on Multi-granularity: Each classifier is included to learn knowledge in different granularities, so regularization should be applied to achieve this goal



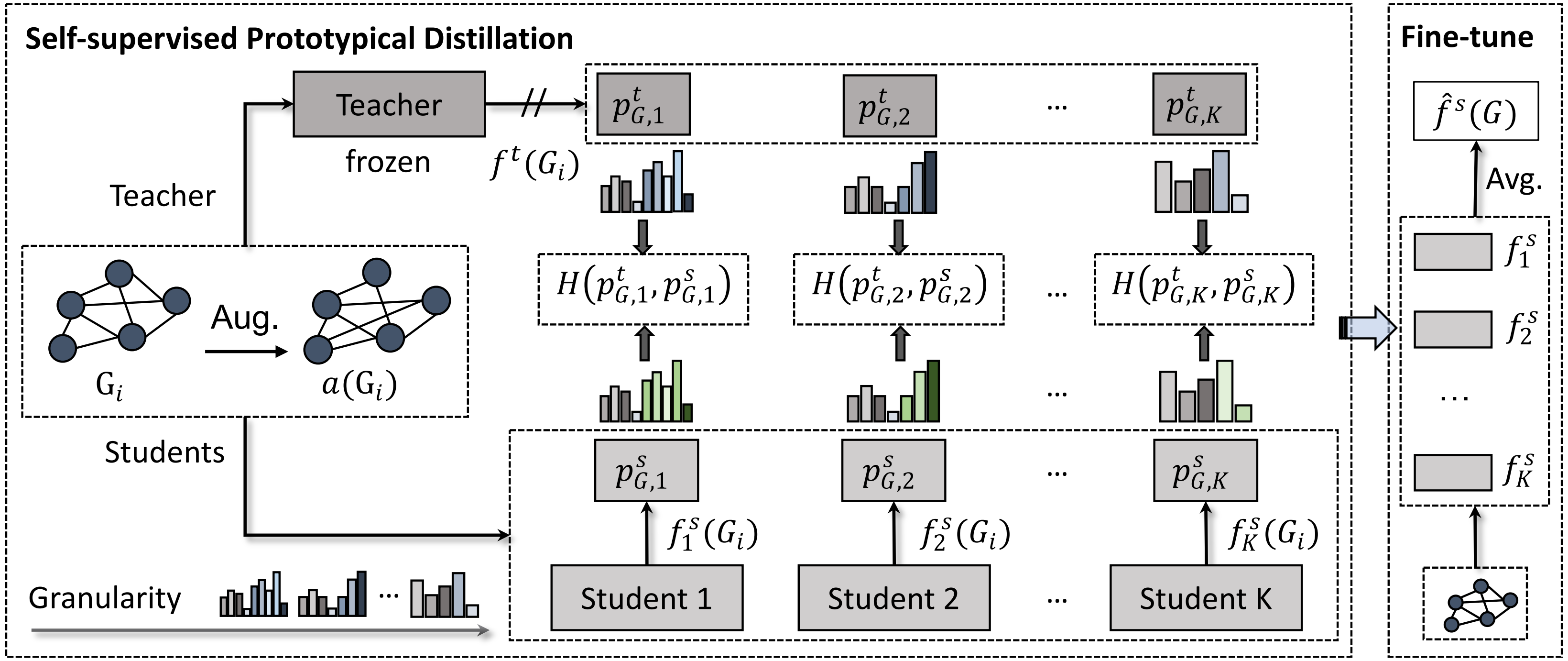
MGSE Framework



$$\mathcal{L} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\left[\boxed{\mathcal{D}(\mathbf{p}^t_{i,k}, \mathbf{p}^s_{i,k})} - \lambda\mathcal{H}(\overline{\mathbf{p}}^s_k)\right] \qquad \mathcal{D}(\mathbf{p}^t_k, \mathbf{p}^s_k) = -\mathcal{H}(\mathbf{p}^t_k, \mathbf{p}^s_k) = \frac{1}{ND_k}\sum_{i=1}^{N}\sum_{d=1}^{D_k} -\left[\mathbf{p}^t_{i,k}\right]_d \cdot \log\left[\mathbf{p}^s_{i,k}\right]_d.$$

Using knowledge distillation for quick knowledge transfer and model converge]

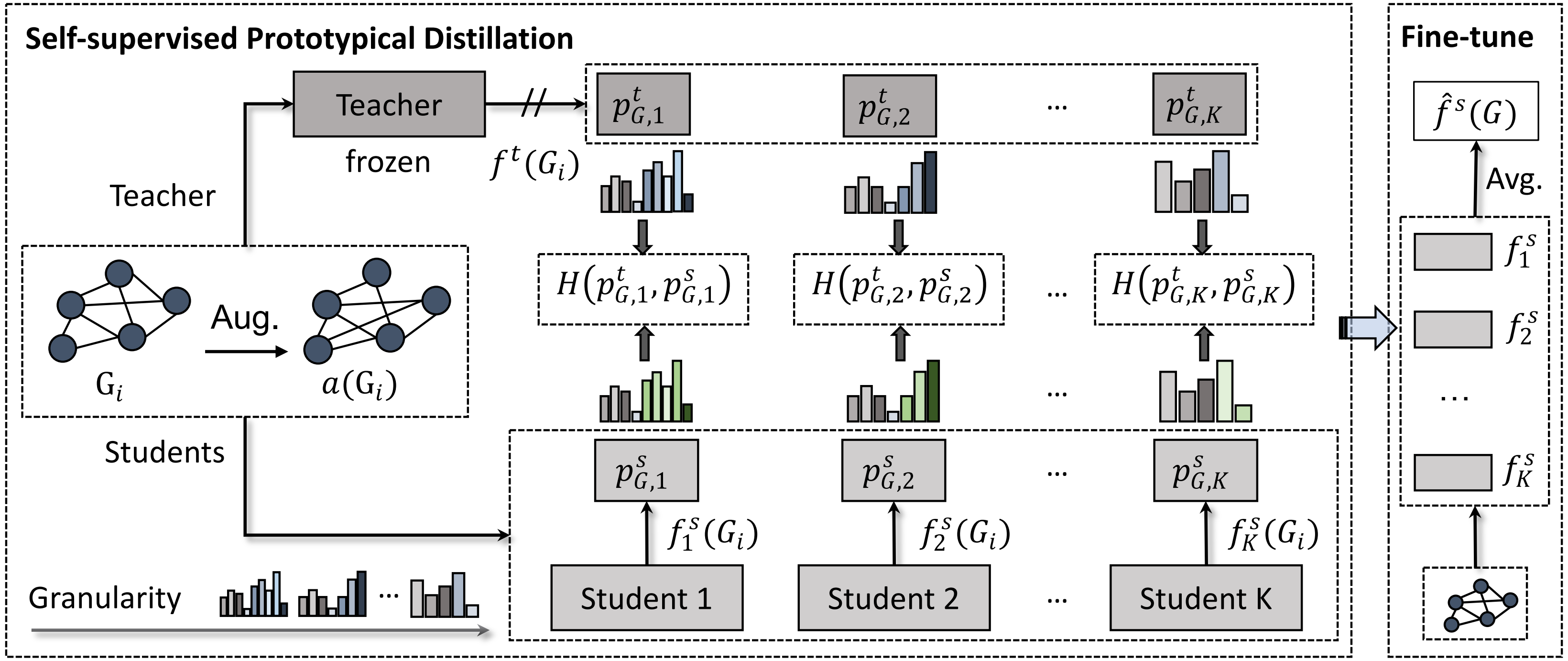
MGSE Framework



$$\mathcal{L} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} \left[\mathcal{D}(\mathbf{p}^t_{i,k}, \mathbf{p}^s_{i,k}) - \lambda \mathcal{H}(\overline{\mathbf{p}}^s_k) \right] \qquad \mathcal{D}(\mathbf{p}^t_k, \mathbf{p}^s_k) = -\mathcal{H}(\mathbf{p}^t_k, \mathbf{p}^s_k) = \frac{1}{ND_k} \sum_{i=1}^{N} \boxed{\sum_{d=1}^{D_k}} - \left[\mathbf{p}^t_{i,k}\right]_d \cdot \log \left[\mathbf{p}^s_{i,k}\right]_d.$$

Manually setting different prototype number for *K* student models to model the multi-granularity

MGSE Framework



$$\mathcal{L} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} \left[\mathcal{D}\left(\mathbf{p}^t_{i,k}, \mathbf{p}^s_{i,k}\right) - \boxed{\lambda \mathcal{H}\left(\overline{\mathbf{p}}^s_k\right)} \right] \qquad \mathcal{H}\left(\overline{\mathbf{p}}^s_k\right) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{p}^s_{i,k}\right] \cdot \log\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{p}^s_{i,k}\right]$$

Adding ME-MAX regularization for loading balance

Experiment - Dataset

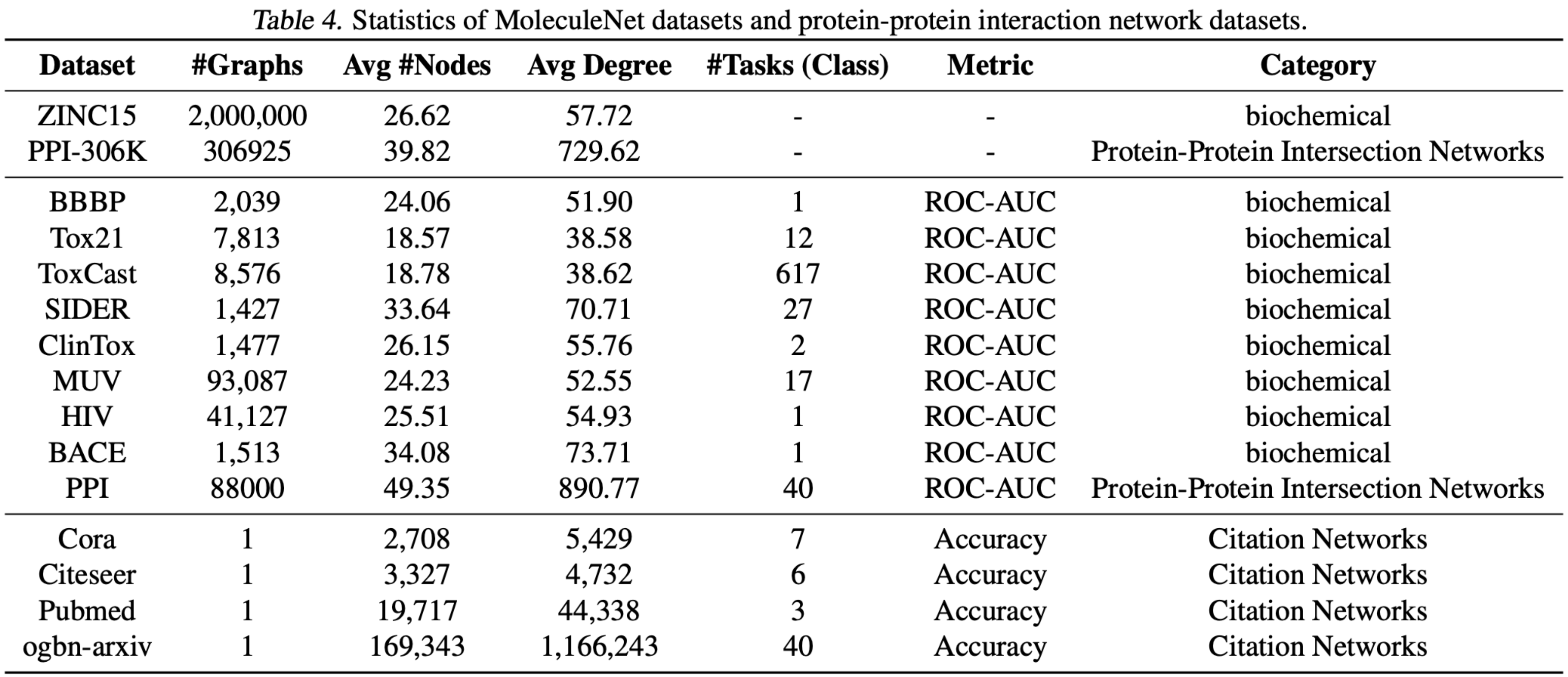
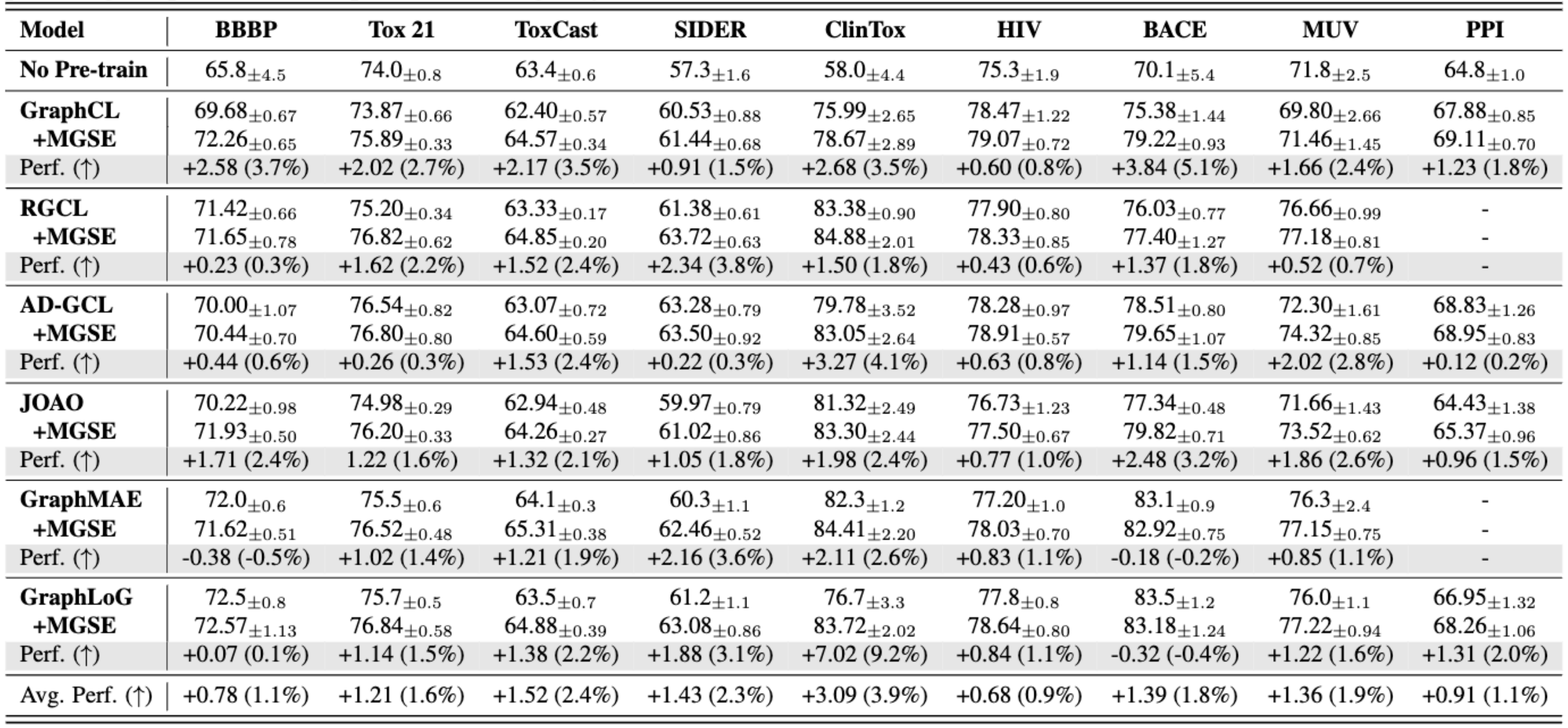


Table 4. Statistics of MoleculeNet datasets and protein-protein interaction network datasets.

Dataset	#Graphs	Avg #Nodes	Avg Degree	#Tasks (Class)	Metric	Category
ZINC15	2,000,000	26.62	57.72	-	-	biochemical
PPI-306K	306925	39.82	729.62	-	-	Protein-Protein Intersection Networks
BBBP	2,039	24.06	51.90	1	ROC-AUC	biochemical
Tox21	7,813	18.57	38.58	12	ROC-AUC	biochemical
ToxCast	8,576	18.78	38.62	617	ROC-AUC	biochemical
SIDER	1,427	33.64	70.71	27	ROC-AUC	biochemical
ClinTox	1,477	26.15	55.76	2	ROC-AUC	biochemical
MUV	93,087	24.23	52.55	17	ROC-AUC	biochemical
HIV	41,127	25.51	54.93	1	ROC-AUC	biochemical
BACE	1,513	34.08	73.71	1	ROC-AUC	biochemical
PPI	88000	49.35	890.77	40	ROC-AUC	Protein-Protein Intersection Networks
Cora	1	2,708	5,429	7	Accuracy	Citation Networks
Citeseer	1	3,327	4,732	6	Accuracy	Citation Networks
Pubmed	1	19,717	44,338	3	Accuracy	Citation Networks
ogbn-arxiv	1	169,343	1,166,243	40	Accuracy	Citation Networks

Experiment - Main Results

Table 1. Performance (i.e., AUC) of state-of-the-art SSL-based GNN frameworks in the transfer learning setting, and improvements after MGSE is applied. "-" means baseline results are not reported in the original papers. The percentage in the parentheses refers to the percentage of performance improvement brought by MGSE.



Model	BBBP	Tox 21	ToxCast	SIDER	ClinTox	HIV	BACE	MUV	PPI
No Pre-train	$65.8_{\pm 4.5}$	$74.0_{\pm 0.8}$	$63.4_{\pm 0.6}$	$57.3_{\pm 1.6}$	$58.0_{\pm 4.4}$	$75.3_{\pm 1.9}$	$70.1_{\pm 5.4}$	$71.8_{\pm 2.5}$	$64.8_{\pm 1.0}$
GraphCL	$69.68_{\pm 0.67}$	$73.87_{\pm 0.66}$	$62.40_{\pm 0.57}$	$60.53_{\pm 0.88}$	$75.99_{\pm 2.65}$	$78.47_{\pm 1.22}$	$75.38_{\pm 1.44}$	$69.80_{\pm 2.66}$	$67.88_{\pm 0.85}$
+MGSE	$72.26_{\pm 0.65}$	$75.89_{\pm 0.33}$	$64.57_{\pm 0.34}$	$61.44_{\pm 0.68}$	$78.67_{\pm 2.89}$	$79.07_{\pm 0.72}$	$79.22_{\pm 0.93}$	$71.46_{\pm 1.45}$	$69.11_{\pm 0.70}$
Perf. (↑)	+2.58 (3.7%)	+2.02 (2.7%)	+2.17 (3.5%)	+0.91 (1.5%)	+2.68 (3.5%)	+0.60 (0.8%)	+3.84 (5.1%)	+1.66 (2.4%)	+1.23 (1.8%)
RGCL	$71.42_{\pm 0.66}$	$75.20_{\pm 0.34}$	$63.33_{\pm 0.17}$	$61.38_{\pm 0.61}$	$83.38_{\pm 0.90}$	$77.90_{\pm 0.80}$	$76.03_{\pm 0.77}$	$76.66_{\pm 0.99}$	-
+MGSE	$71.65_{\pm 0.78}$	$76.82_{\pm 0.62}$	$64.85_{\pm 0.20}$	$63.72_{\pm 0.63}$	$84.88_{\pm 2.01}$	$78.33_{\pm 0.85}$	$77.40_{\pm 1.27}$	$77.18_{\pm 0.81}$	-
Perf. (↑)	+0.23 (0.3%)	+1.62 (2.2%)	+1.52 (2.4%)	+2.34 (3.8%)	+1.50 (1.8%)	+0.43 (0.6%)	+1.37 (1.8%)	+0.52 (0.7%)	-
AD-GCL	$70.00_{\pm 1.07}$	$76.54_{\pm 0.82}$	$63.07_{\pm 0.72}$	$63.28_{\pm 0.79}$	$79.78_{\pm 3.52}$	$78.28_{\pm 0.97}$	$78.51_{\pm 0.80}$	$72.30_{\pm 1.61}$	$68.83_{\pm 1.26}$
+MGSE	$70.44_{\pm 0.70}$	$76.80_{\pm 0.80}$	$64.60_{\pm 0.59}$	$63.50_{\pm 0.92}$	$83.05_{\pm 2.64}$	$78.91_{\pm 0.57}$	$79.65_{\pm 1.07}$	$74.32_{\pm 0.85}$	$68.95_{\pm 0.83}$
Perf. (↑)	+0.44 (0.6%)	+0.26 (0.3%)	+1.53 (2.4%)	+0.22 (0.3%)	+3.27 (4.1%)	+0.63 (0.8%)	+1.14 (1.5%)	+2.02 (2.8%)	+0.12 (0.2%)
JOAO	$70.22_{\pm 0.98}$	$74.98_{\pm 0.29}$	$62.94_{\pm 0.48}$	$59.97_{\pm 0.79}$	$81.32_{\pm 2.49}$	$76.73_{\pm 1.23}$	$77.34_{\pm 0.48}$	$71.66_{\pm 1.43}$	$64.43_{\pm 1.38}$
+MGSE	$71.93_{\pm 0.50}$	$76.20_{\pm 0.33}$	$64.26_{\pm 0.27}$	$61.02_{\pm 0.86}$	$83.30_{\pm 2.44}$	$77.50_{\pm 0.67}$	$79.82_{\pm 0.71}$	$73.52_{\pm 0.62}$	$65.37_{\pm 0.96}$
Perf. (↑)	+1.71 (2.4%)	1.22 (1.6%)	+1.32 (2.1%)	+1.05 (1.8%)	+1.98 (2.4%)	+0.77 (1.0%)	+2.48 (3.2%)	+1.86 (2.6%)	+0.96 (1.5%)
GraphMAE	$72.0_{\pm 0.6}$	$75.5_{\pm 0.6}$	$64.1_{\pm 0.3}$	$60.3_{\pm 1.1}$	$82.3_{\pm 1.2}$	$77.20_{\pm 1.0}$	$83.1_{\pm 0.9}$	$76.3_{\pm 2.4}$	-
+MGSE	$71.62_{\pm 0.51}$	$76.52_{\pm 0.48}$	$65.31_{\pm 0.38}$	$62.46_{\pm 0.52}$	$84.41_{\pm 2.20}$	$78.03_{\pm 0.70}$	$82.92_{\pm 0.75}$	$77.15_{\pm 0.75}$	-
Perf. (↑)	-0.38 (-0.5%)	+1.02 (1.4%)	+1.21 (1.9%)	+2.16 (3.6%)	+2.11 (2.6%)	+0.83 (1.1%)	-0.18 (-0.2%)	+0.85 (1.1%)	-
GraphLoG	$72.5_{\pm 0.8}$	$75.7_{\pm 0.5}$	$63.5_{\pm 0.7}$	$61.2_{\pm 1.1}$	$76.7_{\pm 3.3}$	$77.8_{\pm 0.8}$	$83.5_{\pm 1.2}$	$76.0_{\pm 1.1}$	$66.95_{\pm 1.32}$
+MGSE	$72.57_{\pm 1.13}$	$76.84_{\pm 0.58}$	$64.88_{\pm 0.39}$	$63.08_{\pm 0.86}$	$83.72_{\pm 2.02}$	$78.64_{\pm 0.80}$	$83.18_{\pm 1.24}$	$77.22_{\pm 0.94}$	$68.26_{\pm 1.06}$
Perf. (↑)	+0.07 (0.1%)	+1.14 (1.5%)	+1.38 (2.2%)	+1.88 (3.1%)	+7.02 (9.2%)	+0.84 (1.1%)	-0.32 (-0.4%)	+1.22 (1.6%)	+1.31 (2.0%)
Avg. Perf. (↑)	+0.78 (1.1%)	+1.21 (1.6%)	+1.52 (2.4%)	+1.43 (2.3%)	+3.09 (3.9%)	+0.68 (0.9%)	+1.39 (1.8%)	+1.36 (1.9%)	+0.91 (1.1%)

Our proposed method can generally further improve the performance of existing graph SSL methods.

Experiment – Multi-Granularity Design

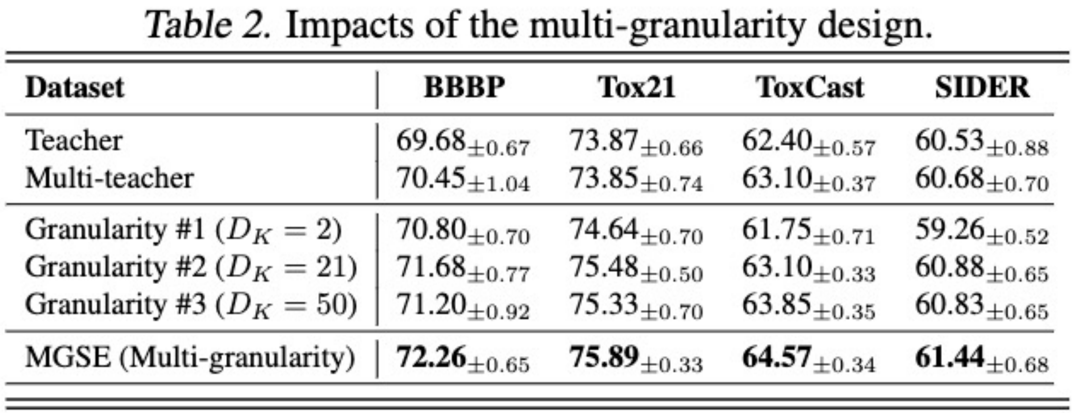
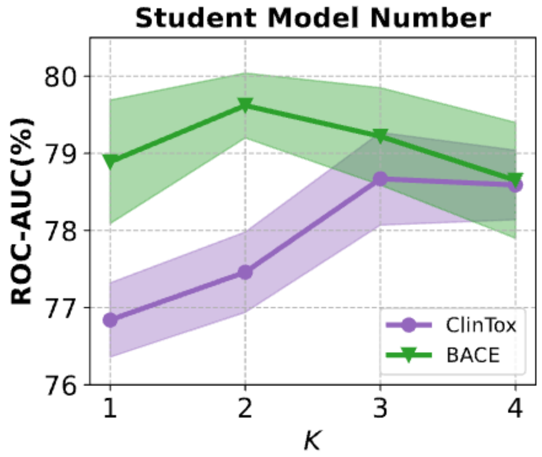


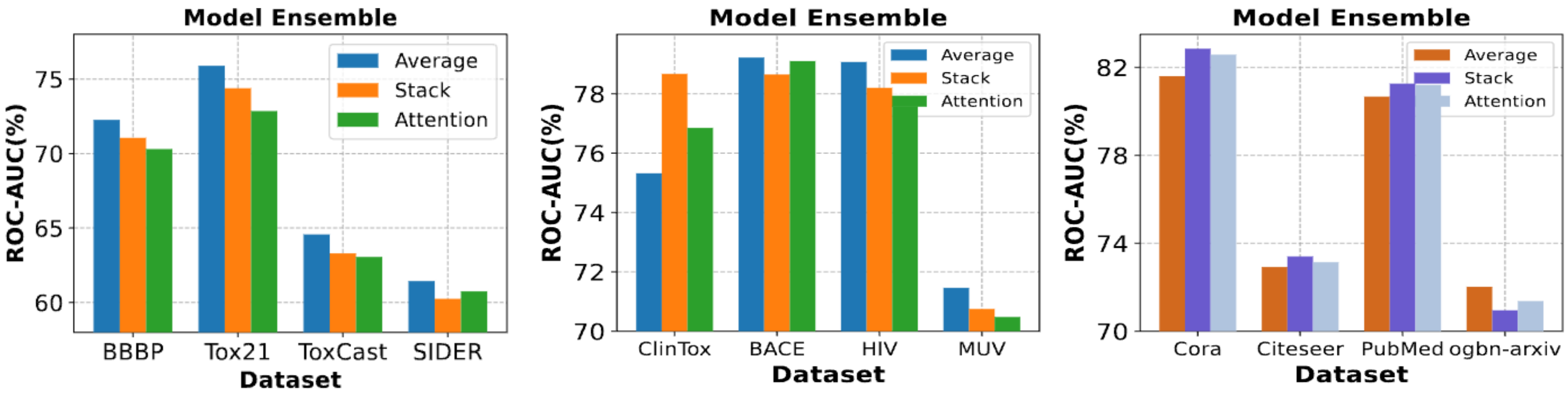
Table 2. Impacts of the multi-granularity design.

Dataset	BBBP	Tox21	ToxCast	SIDER
Teacher	$69.68_{\pm 0.67}$	$73.87_{\pm 0.66}$	$62.40_{\pm 0.57}$	$60.53_{\pm 0.88}$
Multi-teacher	$70.45_{\pm 1.04}$	$73.85_{\pm 0.74}$	$63.10_{\pm 0.37}$	$60.68_{\pm 0.70}$
Granularity #1 ($D_K = 2$)	$70.80_{\pm 0.70}$	$74.64_{\pm 0.70}$	$61.75_{\pm 0.71}$	$59.26_{\pm 0.52}$
Granularity #2 ($D_K = 21$)	$71.68_{\pm 0.77}$	$75.48_{\pm 0.50}$	$63.10_{\pm 0.33}$	$60.88_{\pm 0.65}$
Granularity #3 ($D_K = 50$)	$71.20_{\pm 0.92}$	$75.33_{\pm 0.70}$	$63.85_{\pm 0.35}$	$60.83_{\pm 0.65}$
MGSE (Multi-granularity)	**72.26**$_{\pm 0.65}$	**75.89**$_{\pm 0.33}$	**64.57**$_{\pm 0.34}$	**61.44**$_{\pm 0.68}$



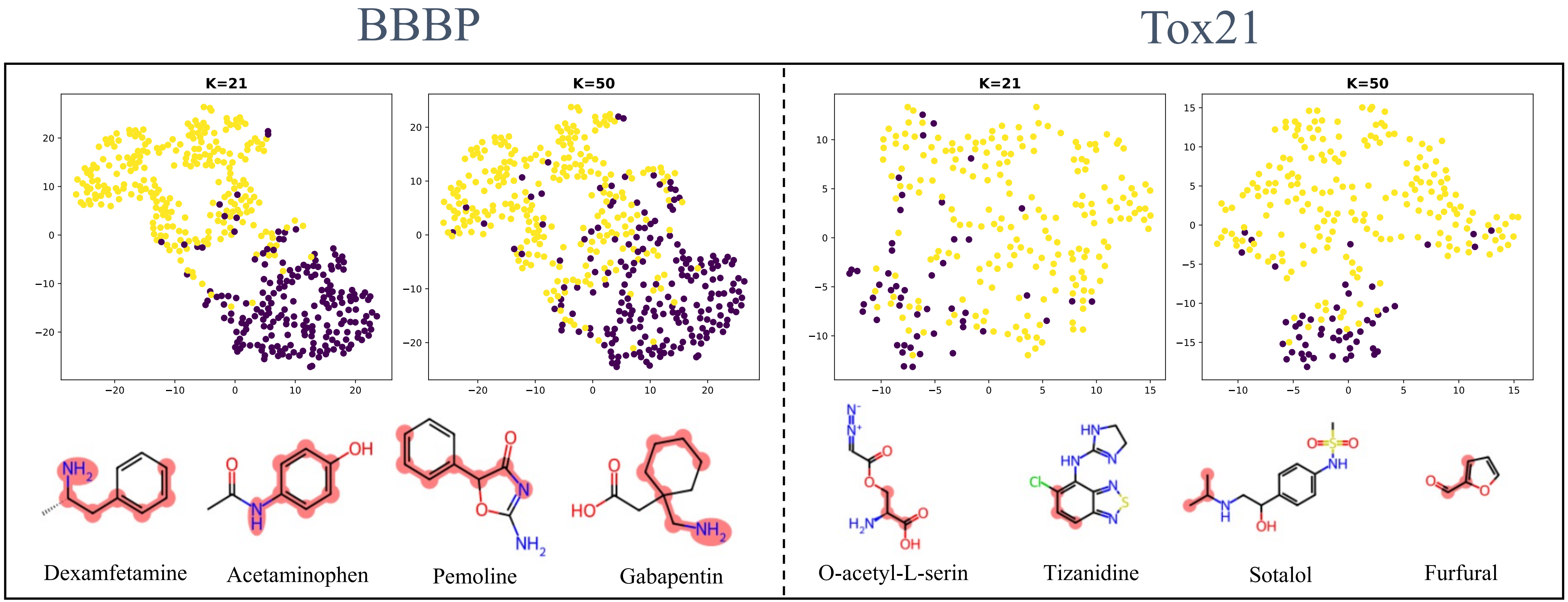
- Multi-granularity has more significant advantage on multi-label prediction tasks which require more complex knowledge.
- Our design provides more flexibility to combine different granularities for various tasks.

Experiment – Model Ensemble Strategy



The distribution shift between training and testing set may have an influence on the best ensemble strategy.

Experiment – Prototype Visualization



The key substructure size of positive samples for discrimination of the BBBP dataset is larger than that from the Tox21 dataset. This implies that the classification of BBBP relies on high-level abstract features, whereas fine-grained substructure information is more helpful to the classification of the Tox21 dataset.

Thanks!

Q&A

Paper

Code