Can Graph Neural Networks Learn Language with Extremely Weak Text Supervision?

ACL 2025 VIENNA

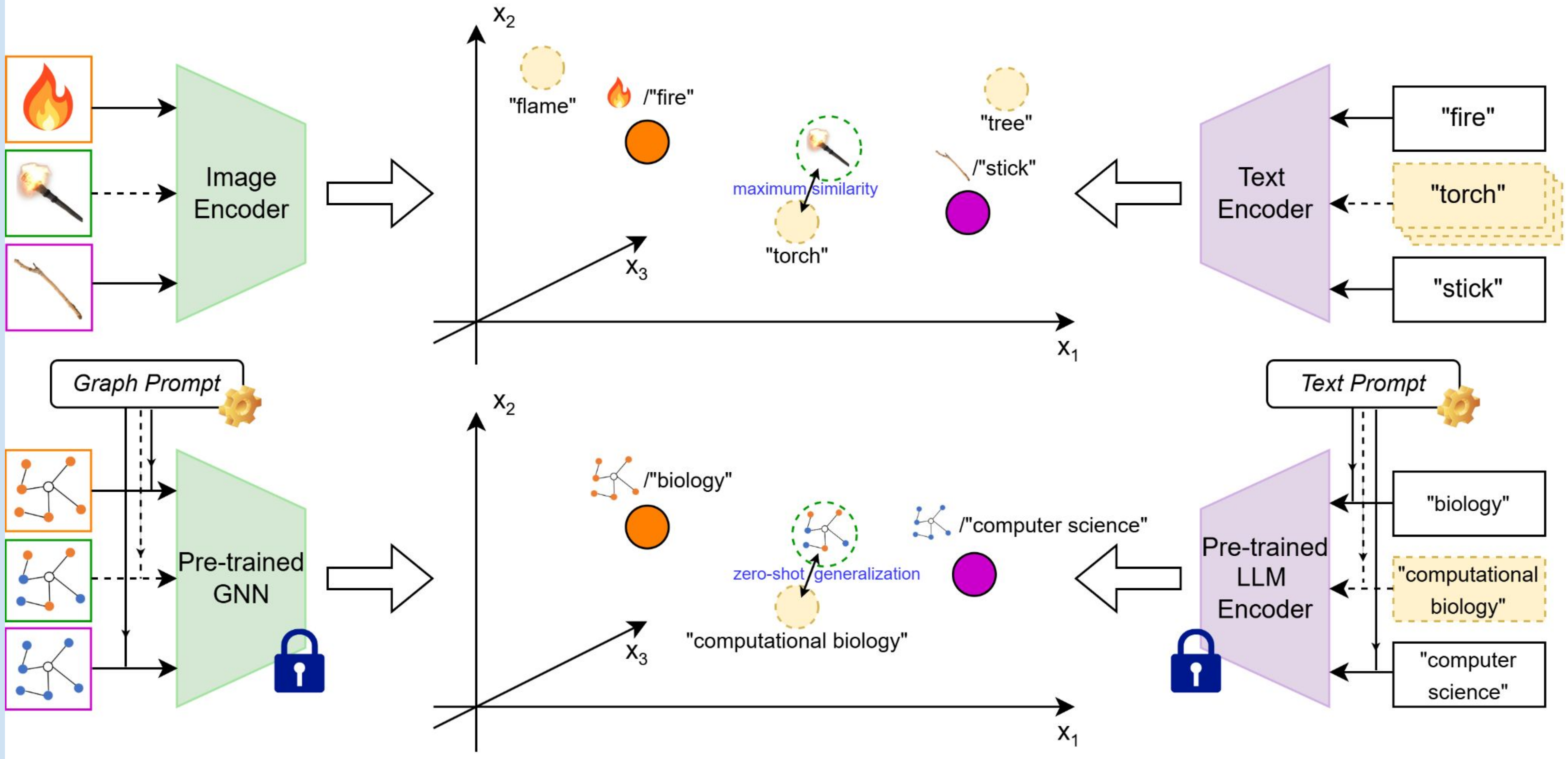
Zihao Li[1], Lecheng Zheng[1], Bowen Jin[1], Donqi Fu[2], Baoyu Jing[1], Yikun Ban[1], Jingrui He[1], Jiawei Han[1]

[1]University of Illinois Urbana-Champaign [2]Meta AI contact: zihaoli5@illinois.edu

paper

author (open for internship in USA)

Vision: CLIP in graph domain for foundational graph representation learning



Challenges:

- graph data is very scarce and the text supervision is extremely weak (v.s., million-scale image-text pairs in CLIP)
- Conceptual gap between different graph domains (v.s., language tokens and visual objects retain their concepts)
- Multiple task space: node-level, edge-level, graph-level

Observations:

- We don't need joint-pretraining. we already have pre-trained LLMs and many GNN pretraining methods
- When data is limited, prompt learning usually provides better option

Let's try *co-adapting GNN and LLM with prompt learning*!

Multi-modal Prompt Learning for Graph Neural Networks

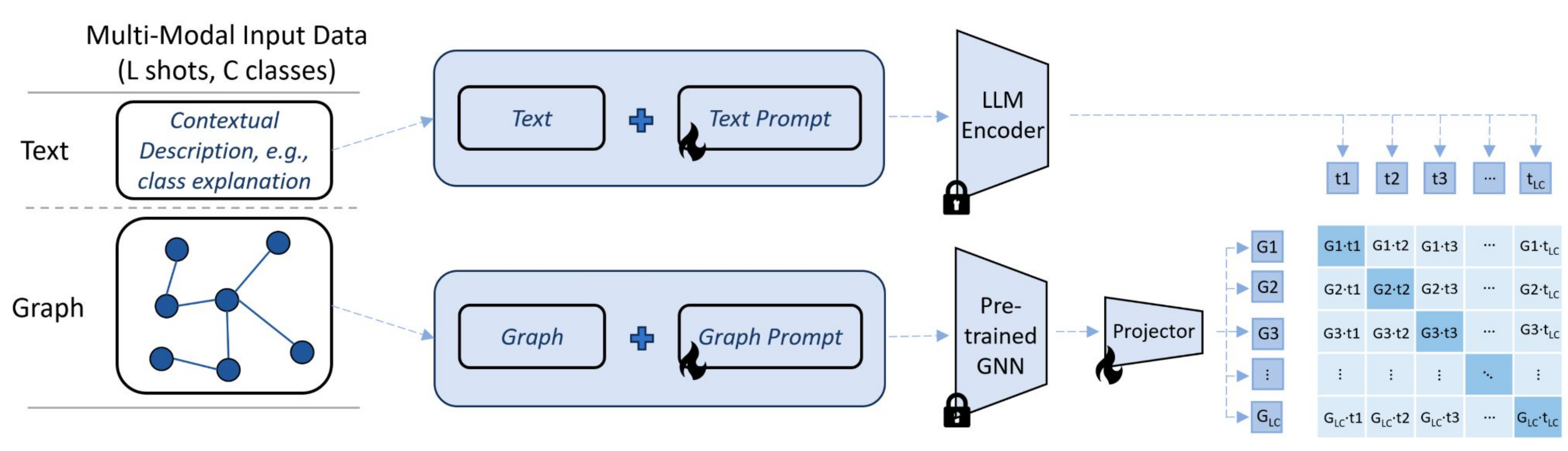


Figure 2: Similar to CLIP backbone, Morpher adapts the graph representations to semantic space through multi-modal prompt learning, even if the GNN and LLM are not jointly trained and are kept frozen.

Improved graph prompts by cross-connection pruning

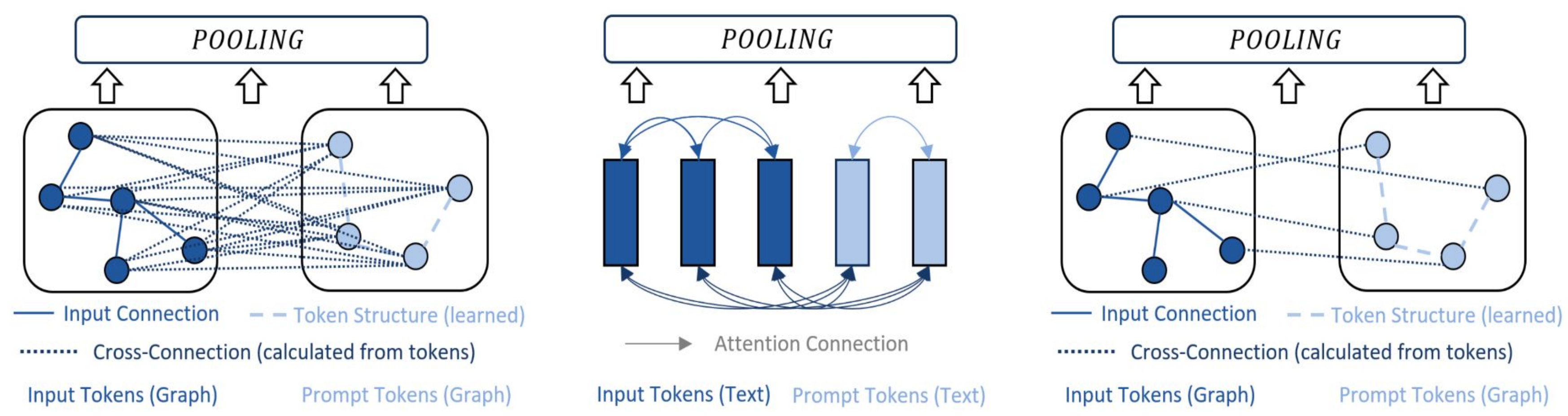


Figure 1: Cross-connections overwhelm inner-connections in current graph prompt design, which may be unstable during training (left); attention in NLP where $3 \times 2 = 6$ cross-connections and $3 + 1 = 4$ inner-connections are balanced (middle); and our balanced graph prompt design (right). **The cross-connections between input and prompt should have a consistent scale with the input connections.**

Main Results: few-shot classification

Training schemes	GNN pretraining	MUTAG Acc	MUTAG F1	ENZYMES Acc	ENZYMES F1	PROTEINS Acc	PROTEINS F1	MSRC_21C Acc	MSRC_21C F1
Supervised	N/A + GCN	66.00	66.67	16.67	8.68	65.89	60.77	38.85	35.32
	N/A + GAT	66.00	65.69	16.45	4.65	64.75	64.08	41.14	39.86
	N/A + GT	66.66	66.26	15.62	4.22	62.81	57.12	38.28	41.62
Pre-train + Fine-tune	GraphCL+GCN	70.00	70.23	17.91	11.82	65.89	61.23	40.00	43.89
	GraphCL+GAT	70.00	69.73	17.91	10.46	65.16	63.92	44.57	45.74
	GraphCL+GT	68.00	67.81	17.70	8.99	63.28	56.41	41.71	43.73
	SimGRACE+GCN	66.67	67.27	17.29	8.78	66.82	64.70	40.57	43.84
	SimGRACE+GAT	70.67	69.10	16.87	7.18	65.42	63.65	42.85	42.37
	SimGRACE+GT	69.33	69.77	16.24	6.08	65.98	62.31	39.42	40.78
AIO (Sun et al., 2023a)	GraphCL+GCN	64.67	39.27	17.50	4.97	61.35	44.93	3.59	10.09
	GraphCL+GAT	64.67	39.27	17.50	4.97	59.21	37.19	14.37	3.11
	GraphCL+GT	73.33	72.06	18.33	9.09	40.79	28.97	17.96	8.30
	SimGRACE+GCN	64.67	39.27	16.04	4.61	67.42	60.87	34.73	18.16
	SimGRACE+GAT	64.67	39.27	16.04	4.61	59.21	37.19	7.78	1.79
	SimGRACE+GT	36.00	27.26	17.50	8.15	50.56	49.34	32.34	15.13
GPF-plus (Fang et al., 2023)	GraphCL+GCN	68.67	67.27	16.88	15.48	64.75	61.45	47.42	29.02
	GraphCL+GAT	68.67	62.84	16.45	13.23	65.89	60.07	47.42	26.28
	GraphCL+GT	69.33	67.87	18.12	15.56	59.66	37.37	41.71	21.35
	SimGRACE+GCN	65.33	39.52	18.96	15.83	65.16	58.80	45.71	23.32
	SimGRACE+GAT	69.33	66.72	18.54	12.58	63.28	53.50	42.85	21.40
	SimGRACE+GT	70.00	67.31	17.91	14.69	64.83	52.97	34.13	20.13
Gprompt (Liu et al., 2023d)	GraphCL+GCN	73.33	66.93	17.91	8.44	61.01	60.01	1.80	0.21
	GraphCL+GAT	64.67	62.63	17.08	14.18	50.56	50.55	1.80	0.22
	GraphCL+GT	70.67	70.02	17.91	9.64	63.28	58.65	1.80	0.21
	SimGRACE+GCN	65.33	39.52	17.29	14.48	52.70	52.68	1.80	0.21
	SimGRACE+GAT	67.33	65.88	16.25	11.31	59.10	58.72	1.80	0.21
	SimGRACE+GT	73.33	67.84	16.87	13.54	64.75	62.37	1.80	0.223
Improved AIO (Ours)	GraphCL+GCN	77.33	77.74	18.13	11.98	65.89	65.97	42.85	45.91
	GraphCL+GAT	74.67	75.51	18.33	11.26	65.76	66.05	46.85	51.39
	GraphCL+GT	74.67	74.67	19.16	9.04	68.12	68.18	42.85	43.54
	SimGRACE+GCN	68.00	69.01	17.91	9.02	66.82	66.40	44.57	49.24
	SimGRACE+GAT	77.33	77.20	18.75	9.39	66.91	65.49	45.14	42.31
	SimGRACE+GT	71.33	72.06	18.95	11.25	68.59	68.84	40.57	42.82
Morpher (Ours)	GraphCL+GCN	78.67	78.09	20.41	15.20	67.47	66.40	45.14	49.62
	GraphCL+GAT	79.33	79.15	23.12	18.01	70.89	70.30	50.85	54.48
	GraphCL+GT	76.00	76.51	19.58	13.28	73.53	72.48	45.71	48.41
	SimGRACE+GCN	69.33	70.27	19.79	14.94	67.10	66.15	45.71	51.24
	SimGRACE+GAT	78.00	77.65	20.21	16.27	68.12	67.26	45.71	51.13
	SimGRACE+GT	74.00	74.84	19.16	14.29	71.76	71.75	44.00	48.16
IMP of ImprovedAIO		2.00 ↑	5.01 ↑	0.52 ↑	4.41 ↓	2.01 ↑	4.37 ↑	0.28 ↓	2.50 ↑
IMP of Morpher		4.00 ↑	6.73 ↑	2.36 ↑	0.60 ↑	4.81 ↑	6.61 ↑	2.66 ↑	7.14 ↑

Table 14: Few-shot graph classification performance (%). IMP (%): the average improvement (absolute value) compared to the **best result** among all the baseline methods.

Multi-task and domain-transfer performance

Tasks	Methods	Cora Acc	Cora F1	CiteSeer Acc	CiteSeer F1
Node Level	Supervised	52.83	47.73	63.91	64.82
	Fine-tune	56.37	55.04	64.87	66.42
	AIO (Sun et al., 2023a)	14.69	7.10	18.93	6.92
	ImprovedAIO	<u>58.46</u>	<u>55.10</u>	<u>66.44</u>	<u>66.53</u>
	Morpher	**61.26**	**62.36**	**68.20**	**68.56**
Edge Level	Supervised	51.78	50.62	52.14	50.81
	Fine-tune	52.50	51.00	52.50	51.12
	AIO (Sun et al., 2023a)	50.00	33.33	50.00	33.33
	ImprovedAIO	<u>54.64</u>	<u>54.57</u>	<u>53.92</u>	<u>53.55</u>
	Morpher	**55.71**	**55.05**	**55.35**	**55.05**

Table 2: Node-level, edge-level performance. Best results are bolded and second-best results are underlined.

Source	Methods	MUTAG (graph-level) Acc	MUTAG (graph-level) F1	PubMed (node-level) Acc	PubMed (node-level) F1
ENZYMES (graph-level)	Supervised	66.00	56.67	47.57	36.07
	Fine-tune	68.00	55.04	47.57	36.07
	AIO	64.00	54.50	44.85	34.13
	ImprovedAIO	<u>70.67</u>	<u>64.07</u>	<u>50.28</u>	<u>50.51</u>
	Morpher	**72.67**	**73.29**	**54.42**	**53.96**
CiteSeer (node-level)	Supervised	66.00	56.67	47.57	36.07
	Fine-tune	71.33	62.19	48.71	40.66
	AIO	65.33	57.20	45.71	34.39
	ImprovedAIO	<u>74.00</u>	<u>73.76</u>	<u>52.57</u>	<u>51.29</u>
	Morpher	**76.67**	**77.04**	**58.29**	**57.54**

Table 13: Domain Transfer Performance. Best results are bolded and second-best results are underlined.

Compatibility

Table 7: Few-shot graph classification performance (%) of Morpher with ELECTRA (Clark et al., 2020) as language encoder. Other experiment settings are identical to the main experiment.

GNN pretraining	MUTAG Acc	MUTAG F1	ENZYMES Acc	ENZYMES F1	PROTEINS Acc	PROTEINS F1	MSRC_21C Acc	MSRC_21C F1
GraphCL + GCN	78.00	78.17	20.41	15.79	67.38	65.66	43.42	47.19
GraphCL + GAT	76.67	75.75	20.41	11.37	66.26	65.66	44.57	49.01
GraphCL + GT	76.67	77.04	19.16	14.68	73.06	72.70	42.28	44.09
SimGRACE + GCN	70.00	70.99	19.79	12.41	68.96	67.77	45.71	48.44
SimGRACE + GAT	77.33	77.51	18.12	13.31	68.96	67.78	44.00	49.43
SimGRACE + GT	72.67	73.55	18.33	15.76	70.18	70.28	41.14	44.50

Table 9: Few-shot graph classification performance (%) of Morpher with the GNN pre-trained by GraphMAE (Hou et al., 2022). Other experiment settings are identical to the main experiment.

GNN pretraining	MUTAG Acc	MUTAG F1	ENZYMES Acc	ENZYMES F1	PROTEINS Acc	PROTEINS F1	MSRC_21C Acc	MSRC_21C F1
Pre-train + Fine-tune	71.33	71.41	16.04	12.14	65.86	65.22	39.42	40.20
ImprovedAIO	76.67	76.95	19.58	12.59	66.36	65.30	42.28	46.81
Morpher	78.67	78.67	20.20	16.95	67.38	65.66	45.71	48.49

On AI4Science tasks (with more graph-language pairs)

Table 11: AUC-ROC (↑) on MolecureNet (bace, tox21, hiv). Morpher-K denotes K shots.

Dataset	KVPLM	MoMu	Galactica-1.3B	GIMLET-64M-50-shots	GAT-1M-supervised	Morpher-10	Morpher-20	Morpher-50
bace	0.5126	0.6656	0.5648	0.729	0.697	0.6231	0.6513	0.6858
tox21	0.4917	0.5757	0.4946	0.652	0.754	0.6769	0.7275	0.7459
hiv	0.6120	0.5026	0.3385	0.721	0.729	0.5742	0.7034	0.7283

The performance of Our Morpher paradigm is comparable to much larger models

GNN zero-shot prototype ("reasoning over graph representations")

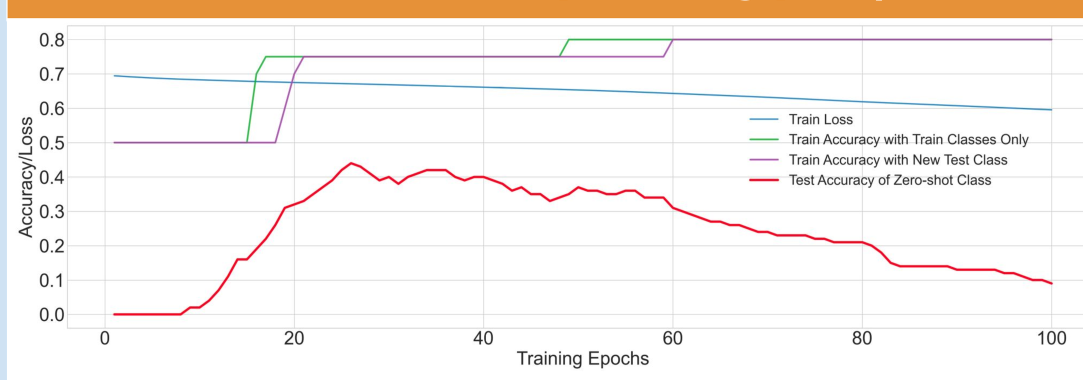


Figure 7: Novel class generalization result for our ZERO-CiteSeer dataset.

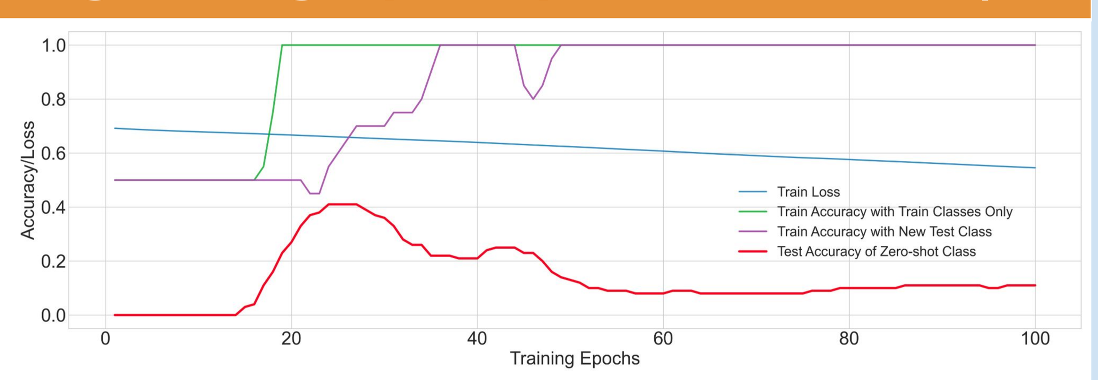


Figure 8: Novel class generalization result for our ZERO-PubMed dataset.