



Microsoft

# Train Once and Explain Everywhere: Pre-training Interpretable Graph Neural Networks

Jun Yin*, Chaozhuo Li*, Hao Yan, Jianxun Lian, Senzhang Wang

Central South University, Changsha, China

Microsoft Research Asia, Beijing, China

NeurIPS 2023

# CONTENT


01 BACKGROUND

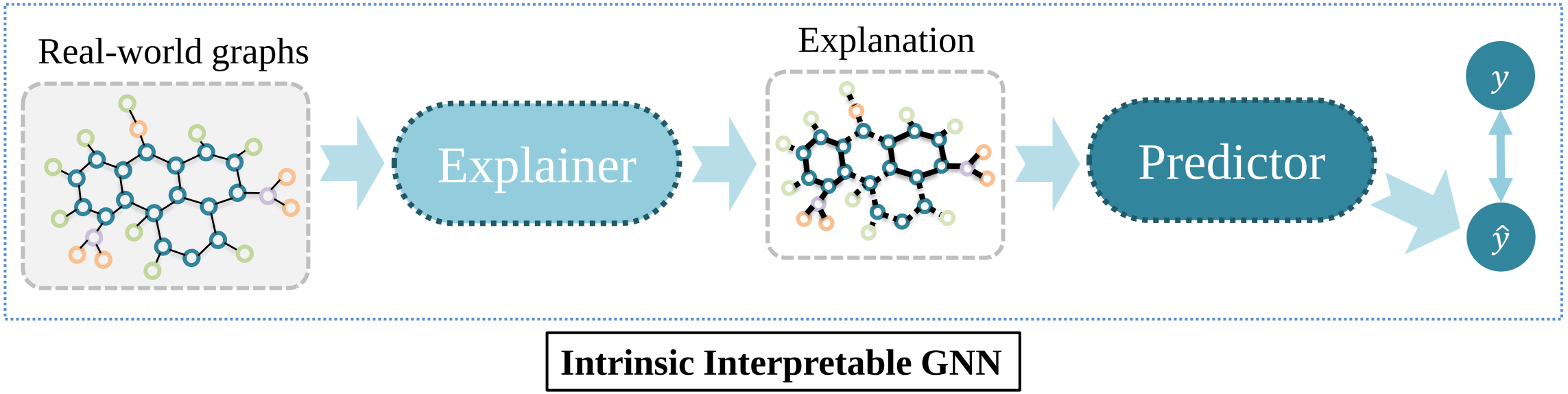
02 MODEL FRAMEWORK

03 EXPERIMENT

04 SUMMARY

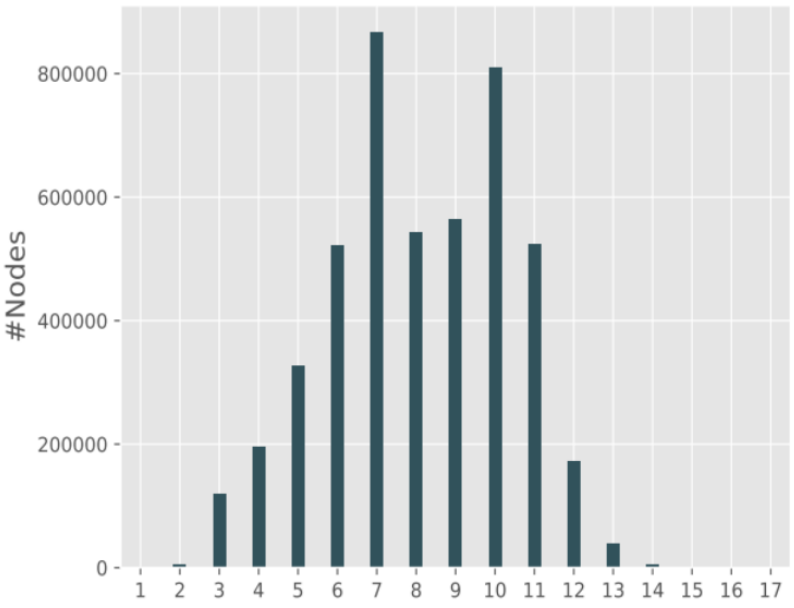
# BACKGROUND

- GNNs have achieved **remarkable success** in various applications.
- Black-box nature makes it hard to understand the **inner decision-making mechanism**.
- Intrinsic interpretable GNNs aim to provide **transparent predictions** by identifying the **influential fraction of the input graph** that guides the model prediction.

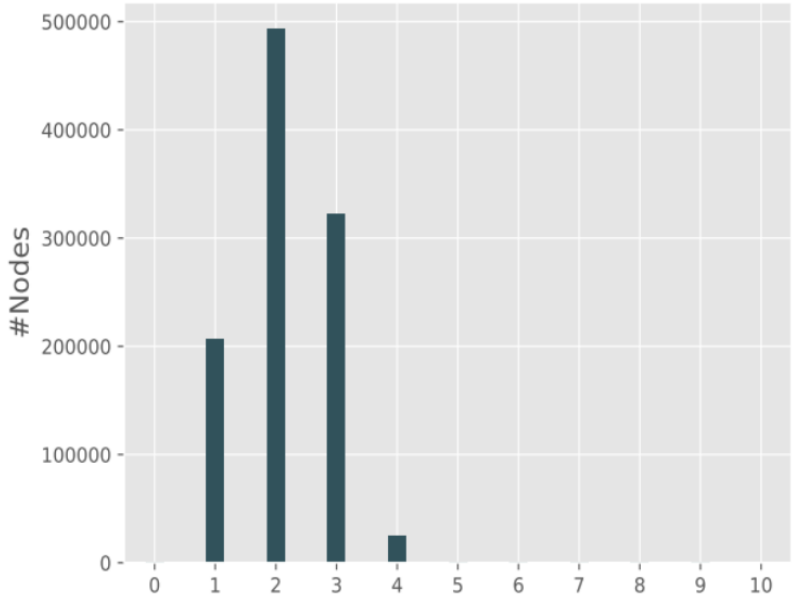


# BACKGROUND

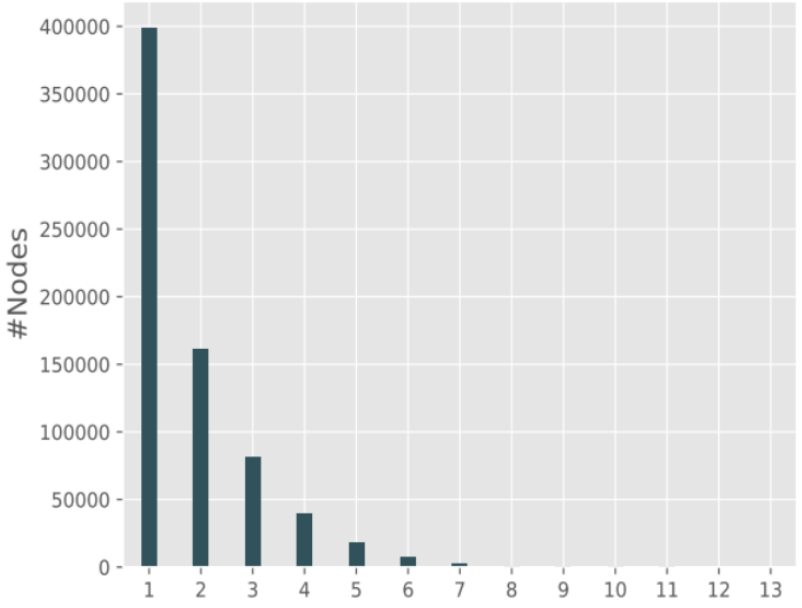
- Existing GNN explanation methods are dataset-specific.
- How to construct a interpretable GNN that can generalize to different datasets?




Degree distribution of MNIST-75sp



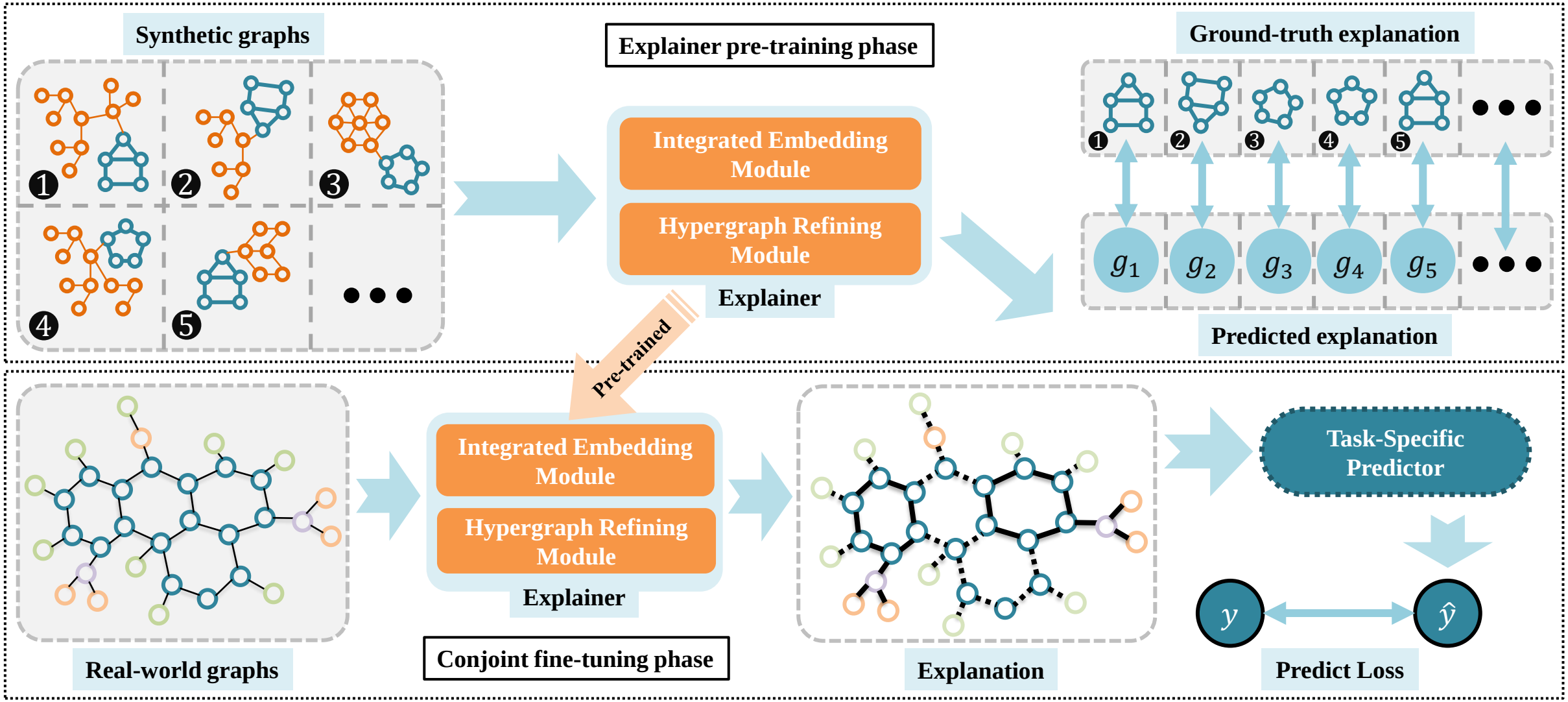

Degree distribution of Ogbg-Molhiv




Degree distribution of Graph-SST2

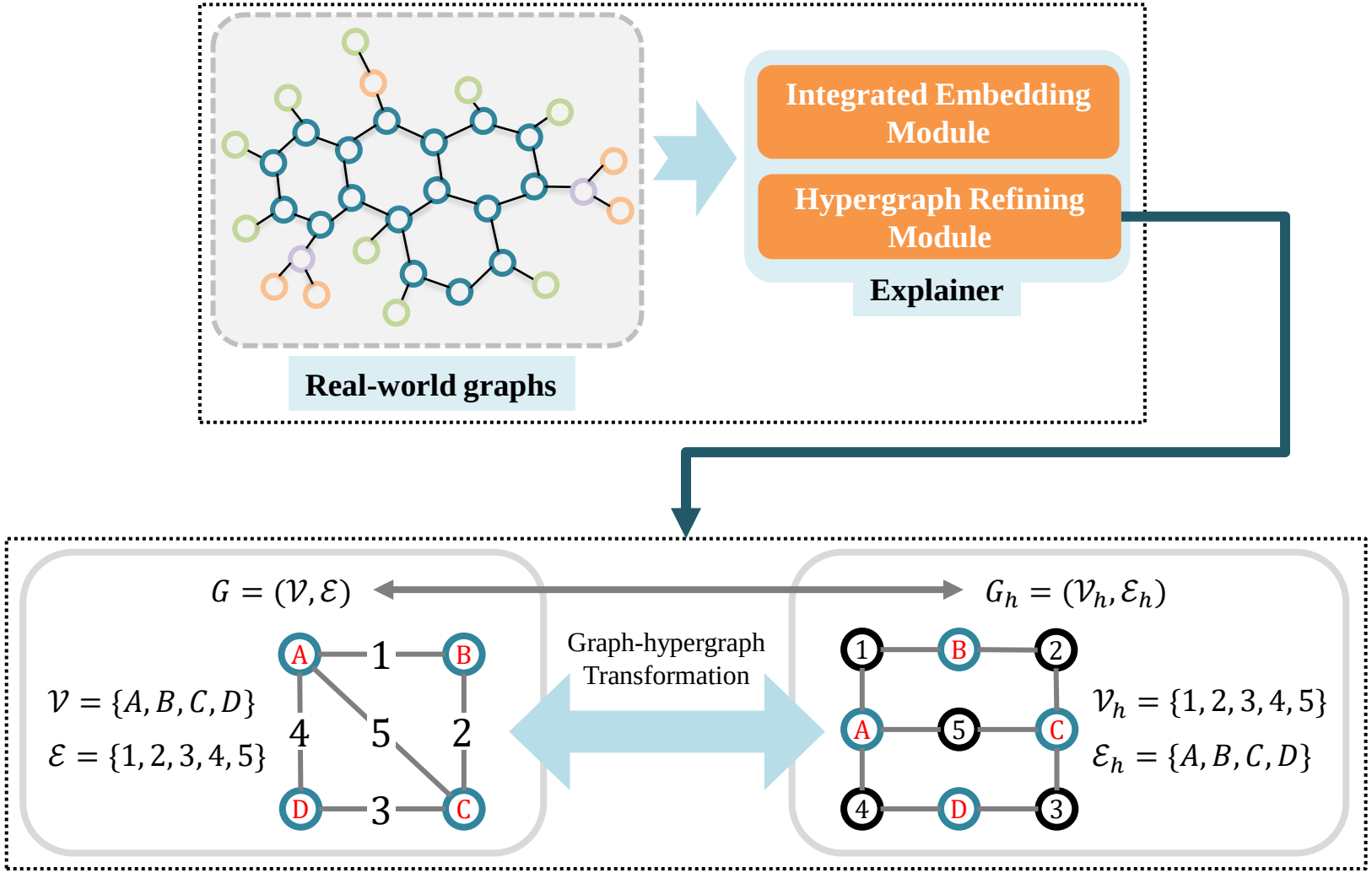
# MODEL FRAMEWORK

- Pre-training on synthetic graphs.



# MODEL FRAMEWORK

- Hyper-graph refining module.



# EXPERIMENT

- Interpretation Performance.

Table 1: Interpretation Performance (ROC-AUC) Comparison. The underlined results highlight the best baselines. The **bold** results mean the $\pi$-GNN or $\pi$-GNN$_{\text{DFT}}$ outperform the best baselines.

| Model | BA-2Motifs | Mutag | MNIST-75sp | Spurious-Motif $b = 0.5$ | Spurious-Motif $b = 0.7$ | Spurious-Motif $b = 0.9$ |
|---|---|---|---|---|---|---|
| GNNExplainer | 67.35 ± 3.29 | 61.98 ± 5.45 | 59.01 ± 2.04 | 62.62 ± 1.35 | 62.25 ± 3.61 | 58.86 ± 1.93 |
| PGExplainer | 84.59 ± 9.09 | 60.91 ± 17.10 | 69.34 ± 4.32 | 69.54 ± 5.64 | 72.33 ± 9.18 | 72.34 ± 2.91 |
| GraphMask | 92.54 ± 8.07 | 62.23 ± 9.01 | 73.10 ± 6.41 | 72.06 ±5.58 | 73.06 ± 4.91 | 66.68 ± 6.96 |
| IB-Subgraph | 86.06 ± 28.37 | 91.04 ± 6.59 | 51.20 ± 5.12 | 57.29 ± 14.35 | 62.89 ± 15.59 | 47.29 ± 13.39 |
| DIR | 82.78 ± 10.97 | 64.44 ± 28.81 | 32.35 ± 9.39 | 78.15 ± 1.32 | 77.68 ± 1.22 | 49.08 ± 3.66 |
| GIN-GSAT | 98.74 ± 0.55 | 99.60 ± 0.51 | 83.36 ± 1.02 | 78.45 ± 3.12 | 74.07 ± 5.28 | 71.97 ± 4.41 |
| PNA-GSAT | 93.77 ± 3.90 | 99.07 ± 0.50 | 84.68 ± 1.06 | 83.34 ± 2.17 | 86.94 ± 4.05 | 88.66 ± 2.44 |
| $\pi$-GNN | **99.33** ± 0.63 | **99.81** ± 0.17 | **92.77** ± 0.80 | **93.24** ± 0.72 | **96.92** ± 0.85 | **96.39** ± 0.92 |
| $\pi$-GNN$_{\text{DFT}}$ | 93.19 ± 1.48 | 95.29 ± 0.67 | **85.18** ± 1.08 | **86.29** ± 2.22 | **87.43** ± 2.47 | **89.64** ± 2.26 |

# EXPERIMENT

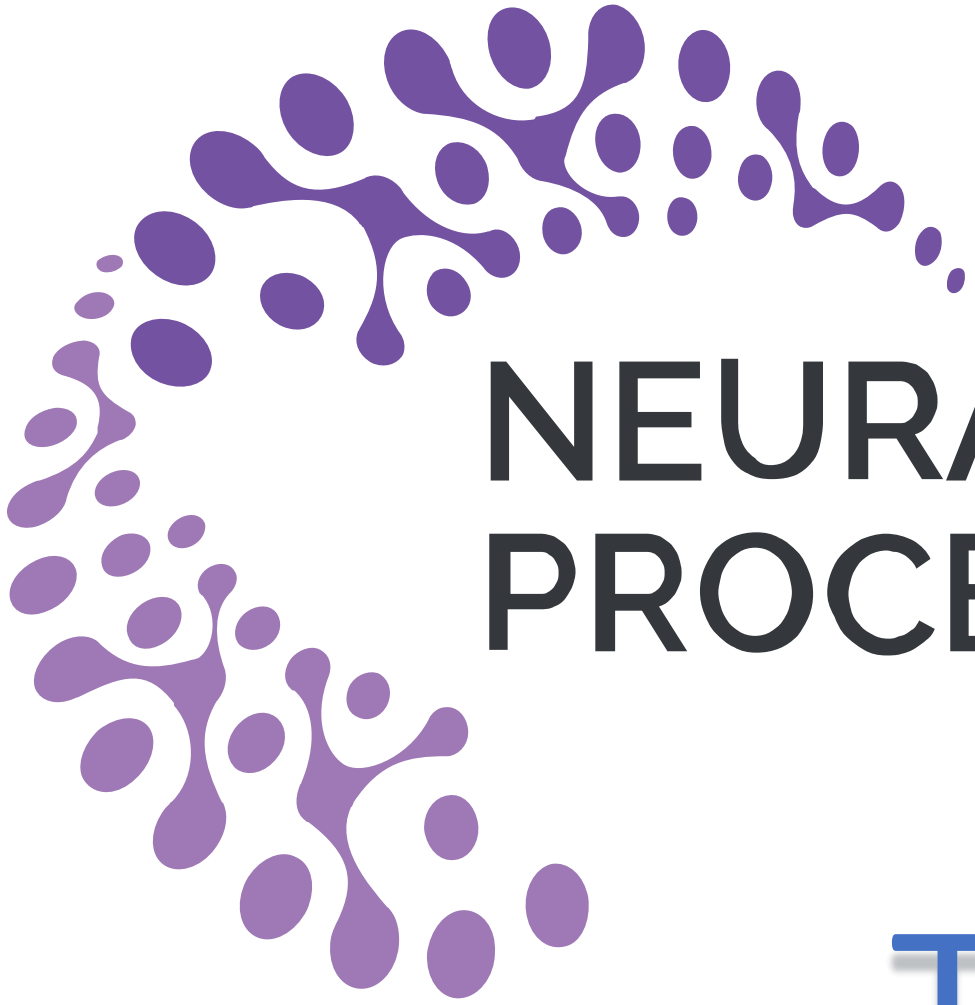
- Prediction Performance.

Table 2: Prediction Performance (Acc) Comparison. The underlined results highlight the best baselines. The **bold** results mean the $\pi$-GNN or $\pi$-GNN$_{\mathrm{DFT}}$ outperform the best baselines.

| Model | Molhiv(AUC) | Graph-SST2 | MNIST-75sp | Spurious-Motif $b = 0.5$ | Spurious-Motif $b = 0.7$ | Spurious-Motif $b = 0.9$ |
|---|---|---|---|---|---|---|
| GIN | $76.69 \pm 1.25$ | $82.73 \pm 0.77$ | $95.74 \pm 0.36$ | $39.87 \pm 1.30$ | $39.04 \pm 1.62$ | $38.57 \pm 2.31$ |
| PNA | $78.91 \pm 1.04$ | $79.87 \pm 1.02$ | $87.20 \pm 5.61$ | $68.15 \pm 2.39$ | $66.35 \pm 3.34$ | $61.40 \pm 3.56$ |
| IB-Subgraph | $76.43 \pm 2.65$ | $82.99 \pm 0.67$ | $93.10 \pm 1.32$ | $54.36 \pm 7.09$ | $48.51 \pm 5.76$ | $46.19 \pm 5.63$ |
| DIR | $76.34 \pm 1.01$ | $82.32 \pm 0.85$ | $88.51 \pm 2.57$ | $45.49 \pm 3.81$ | $41.13 \pm 2.62$ | $37.61 \pm 2.02$ |
| GIN-GSAT | $76.47 \pm 1.53$ | $82.95 \pm 0.58$ | $96.24 \pm 0.17$ | $52.74 \pm 4.08$ | $49.12 \pm 3.29$ | $44.22 \pm 5.57$ |
| PNA-GSAT | $80.24 \pm 0.73$ | $80.92 \pm 0.66$ | $93.96 \pm 0.92$ | $68.74 \pm 2.24$ | $64.38 \pm 3.20$ | $57.01 \pm 2.95$ |
| $\pi$-GNN | $\mathbf{80.86} \pm 0.61$ | $\mathbf{88.05} \pm 0.43$ | $\mathbf{96.89} \pm 0.20$ | $\mathbf{74.67} \pm 0.63$ | $\mathbf{77.52} \pm 0.77$ | $\mathbf{77.46} \pm 0.96$ |
| $\pi$-GNN$_{\mathrm{DFT}}$ | $79.71 \pm 1.08$ | $\mathbf{83.48} \pm 1.20$ | $92.89 \pm 0.95$ | $\mathbf{70.78} \pm 1.63$ | $\mathbf{71.02} \pm 1.43$ | $\mathbf{72.61} \pm 1.75$ |

# SUMMARY

- An interpretable GNN that can generalize to different graph datasets.
- Synthetic pre-training process.
- Hyper-graph transformation based edge representation.




Microsoft

# NEURAL INFORMATION PROCESSING SYSTEMS

# THANKS