# Prototype-Based Explanations for Graph Neural Networks

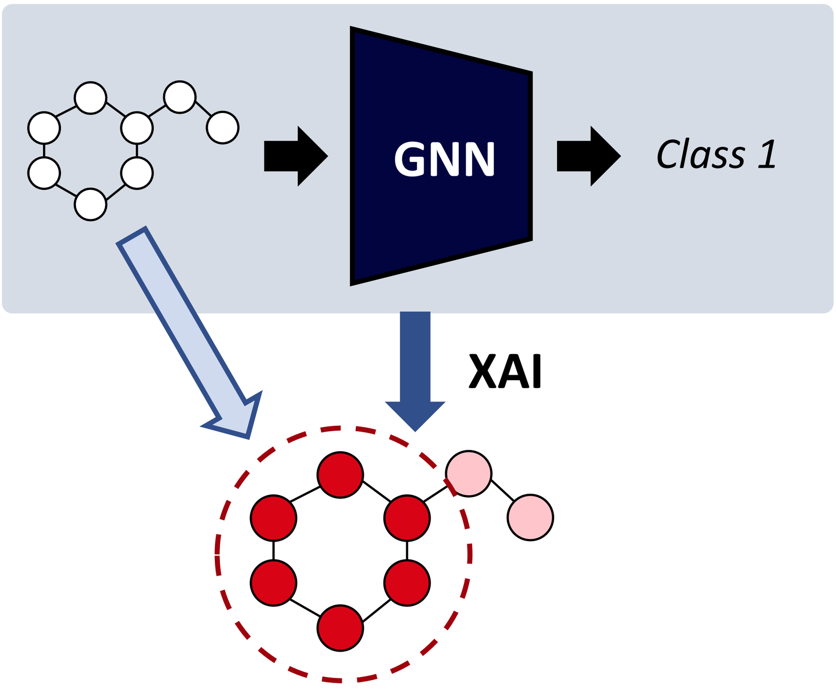
Yong-Min Shin, Sun-Woo Kim, Eun-Bi Yun, Won-Yong Shin
**Yonsei University**

## Research Problem

### Explainable AI for graph neural networks (GNNs)

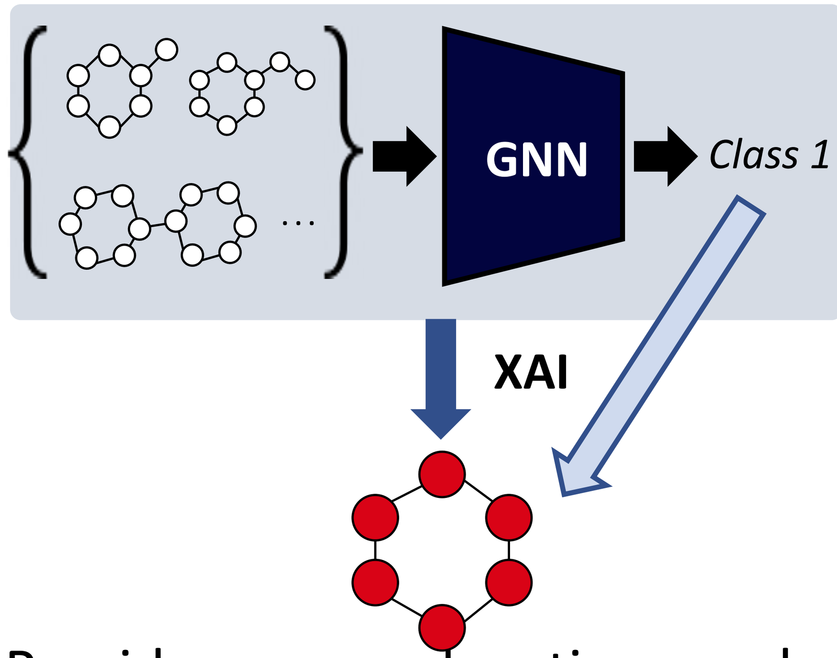
**Explainable AI (XAI)** is interested in explaining deep neural network models, which can provide model-user trust and avoid 'clever Hans' predictions [1].

#### Instance-level explanation



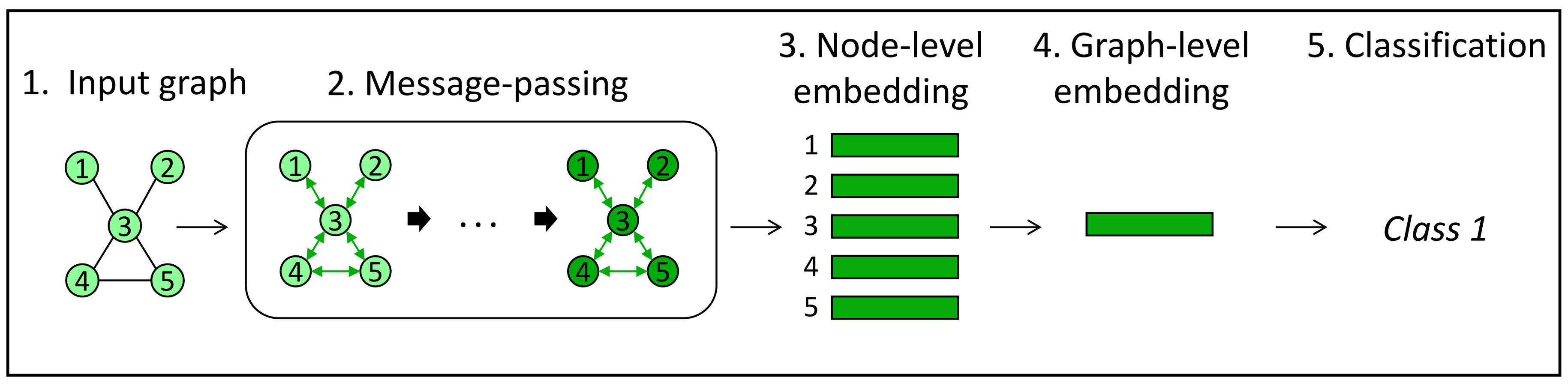

Provides explanation for **each instance**. Usually highlights important subgraph or features of the given instance.
Most XAI methods for GNN models consider this approach.

#### Model-level explanation




Provides explanation by describing the **general behavior** of a model without referring to a specific example.
Our work aims to design a XAI method for GNNs by adopting this approach.

### GNNs for graph classification

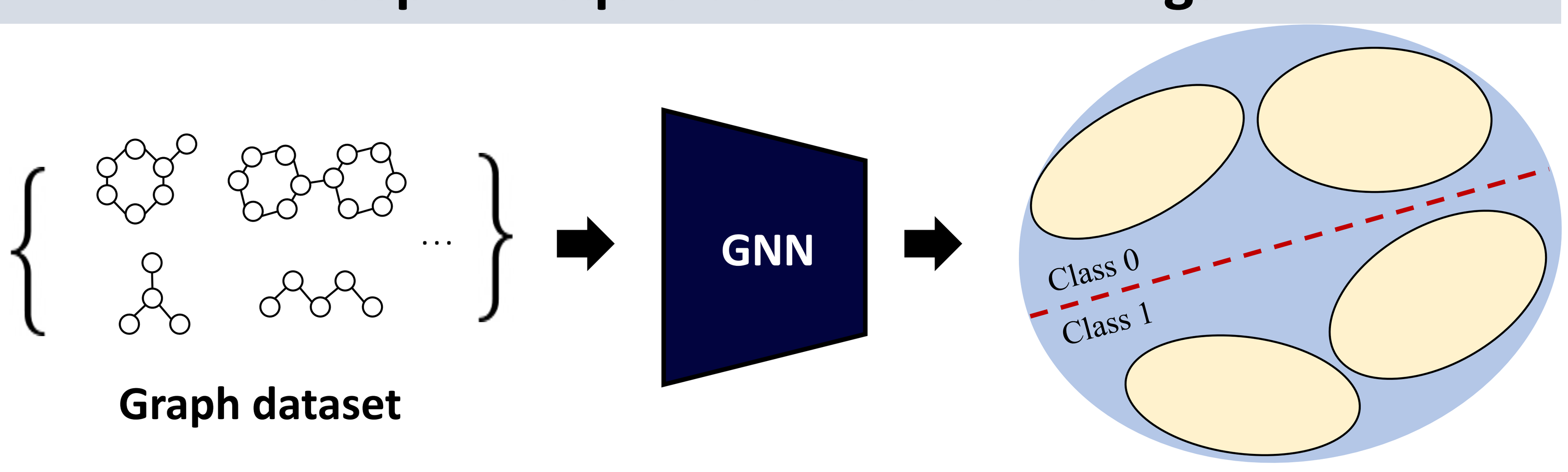



### Problem formulation

In our study, we aim at designing a **model-level explanation method** of GNNs for graph classification, which provides an abstract and concise explanation by capturing what the model has learned from the training data.

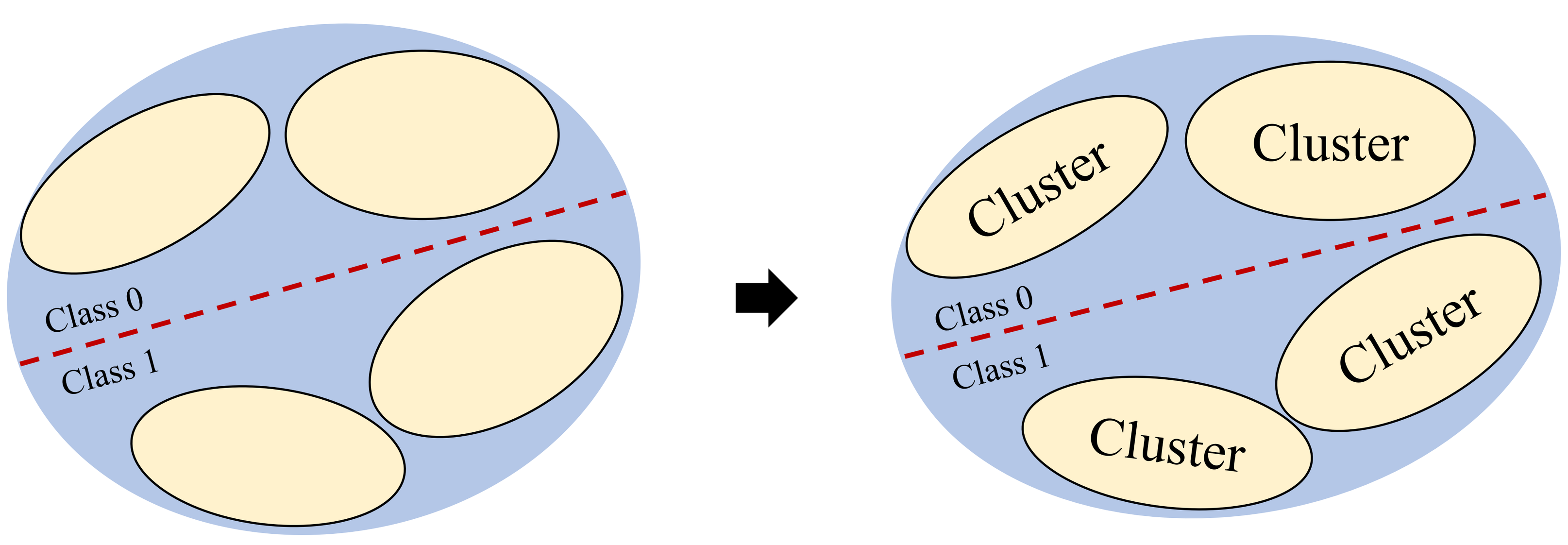
## Proposed methodology: PAGE

### Step 1: Acquisition of embeddings



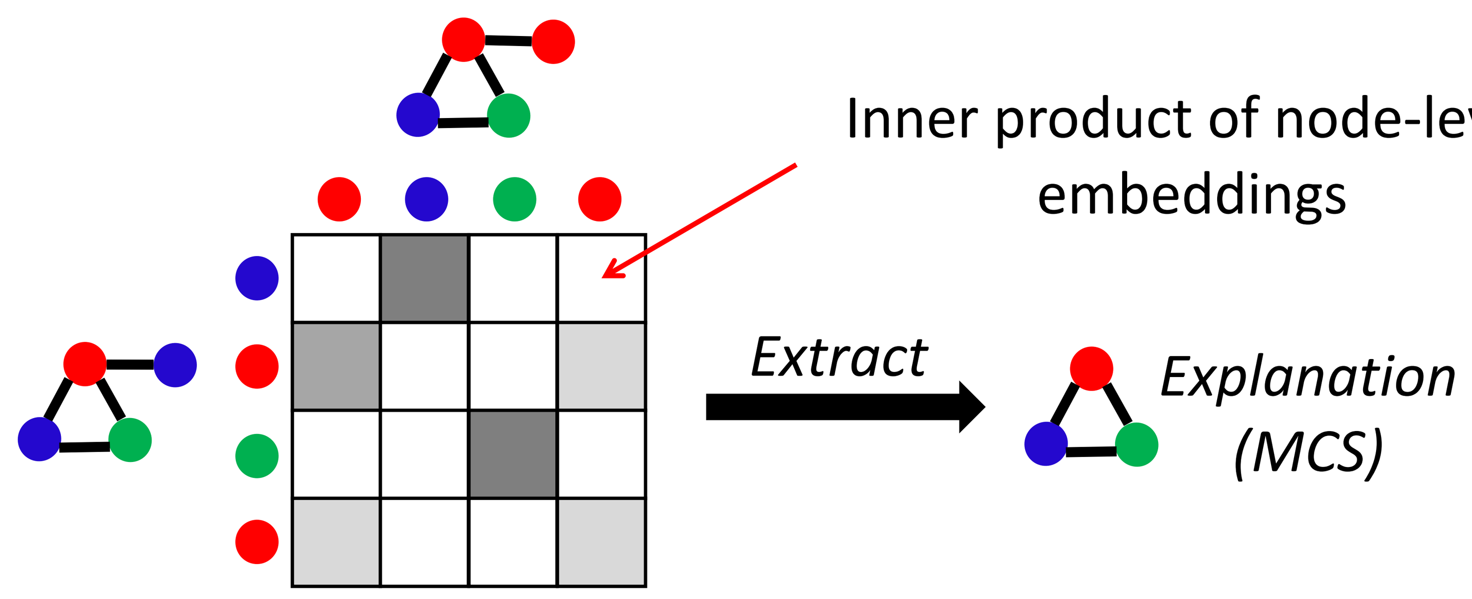

First, we **acquire graph-level embeddings** that are generated during the feed-forward process of GNN.

### Step 2: Clustering on the embedding space




Second, we fit a **Gaussian Mixture Model** to find clusters on the embedding space. The graph-level embedding vectors closest to each centroid is then **selected**.

### Step 3: Prototype discovery by calculating the MCS

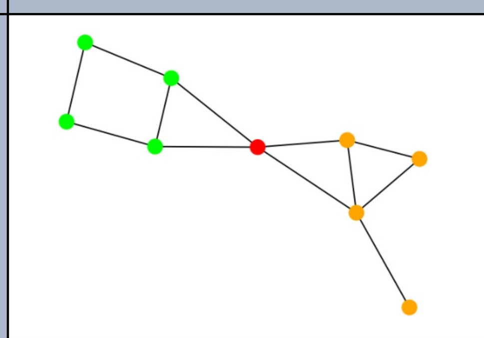
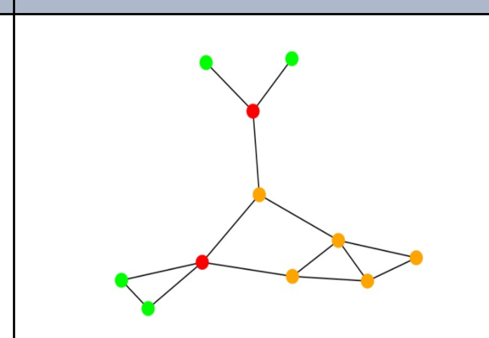
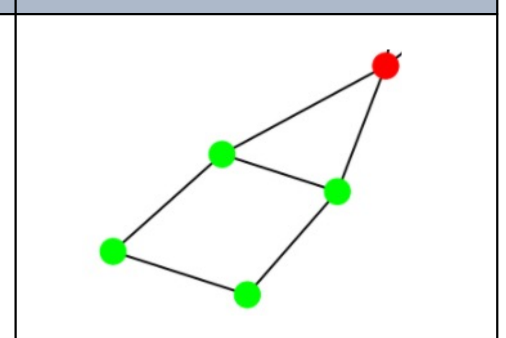
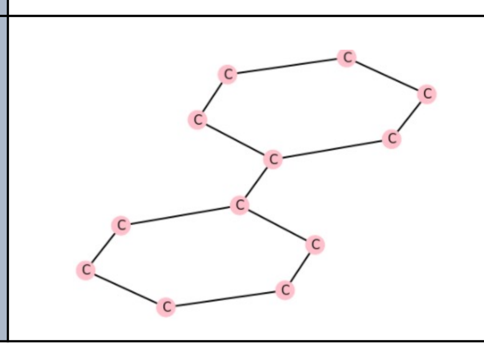
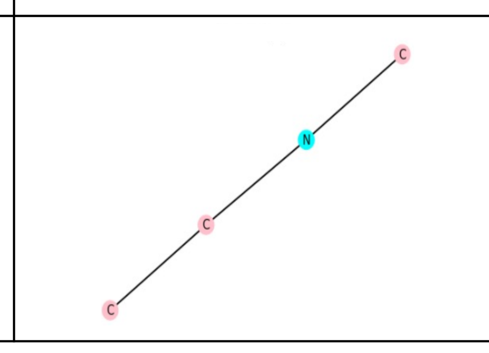
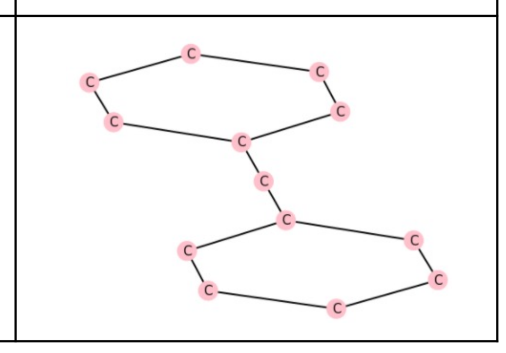



As the final step, we extract the **maximum common subgraph (MCS)** from the selected input graphs to acquire the **most important subgraph pattern** based on NeuralMCS [2].

## Experimental evaluation

We employ GCN [3] as the benchmark GNN model for our experiments.

### Qualitative evaluation

| | PAGE (Ours) | XGNN [4] | Ground-truth explanation |
|---|---|---|---|
| BA-house |  |  |  |
| Solubility |  |  |  |

### Quantitative evaluation

| | Consistency | | Faithfulness | |
|---|---|---|---|---|
| Dataset | **PAGE (Ours)** | XGNN [4] | **PAGE (Ours)** | XGNN [4] |
| BA-house | 0.048 | 0.312 | 0.733 | 0.328 |
| Solubility | 0.109 | 0.348 | 0.591 | 0.085 |

- **Consistency** measures the robustness of explanations across different GNN hyperparameters (the lower the better).
- **Faithfulness** measures the Kendall's tau coefficient between the performance of the GNN model and its explanation accuracy (the higher the better).

## Discussion & Conclusion

- In our work, we propose PAGE, a novel **model-level explanation of a GNN model** that performs graph classification.
- In contrast to XGNN, which relies on reinforcement learning that requires carefully designed reward functions along with domain knowledge, our method **discovers explanations within the dataset**.

## Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345), and by the Yonsei University Research Fund of 2021 (2021-22-0083).

## Reference

[1] Wojciech Samek and Klaus-Robert Muller. Towards explainable artificial intelligence. In ¨ Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pages 5–22. Springer, Cham, Switzerland, 2019. doi: 10.1007/978-3-030-28954-6\ 1.
[2] Ma, G.; Ahmed, N. K.; Willke, T. L.; and Yu, P. S. 2021. Deep graph similarity learning: A survey. Data Min. Knowl. Discov.,35(3): 688–725.
[3] Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In ICLR.
[4] Yuan, H.; Tang, J.; Hu, X.; and Ji, S. 2020. XGNN: Towards model-level explanations of graph neural networks. In KDD.