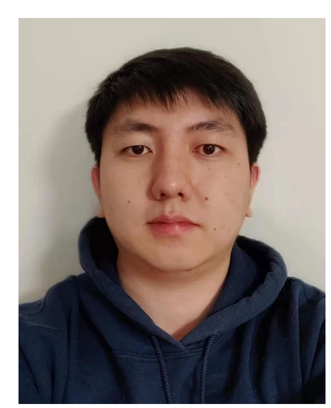
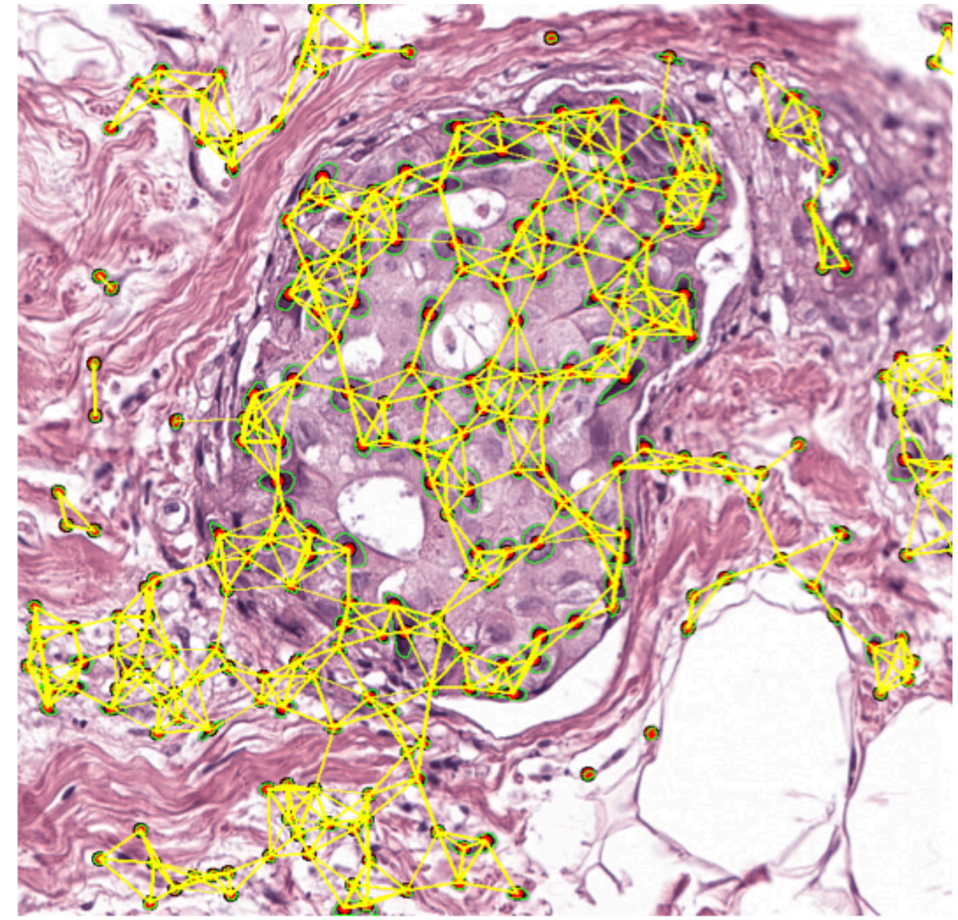
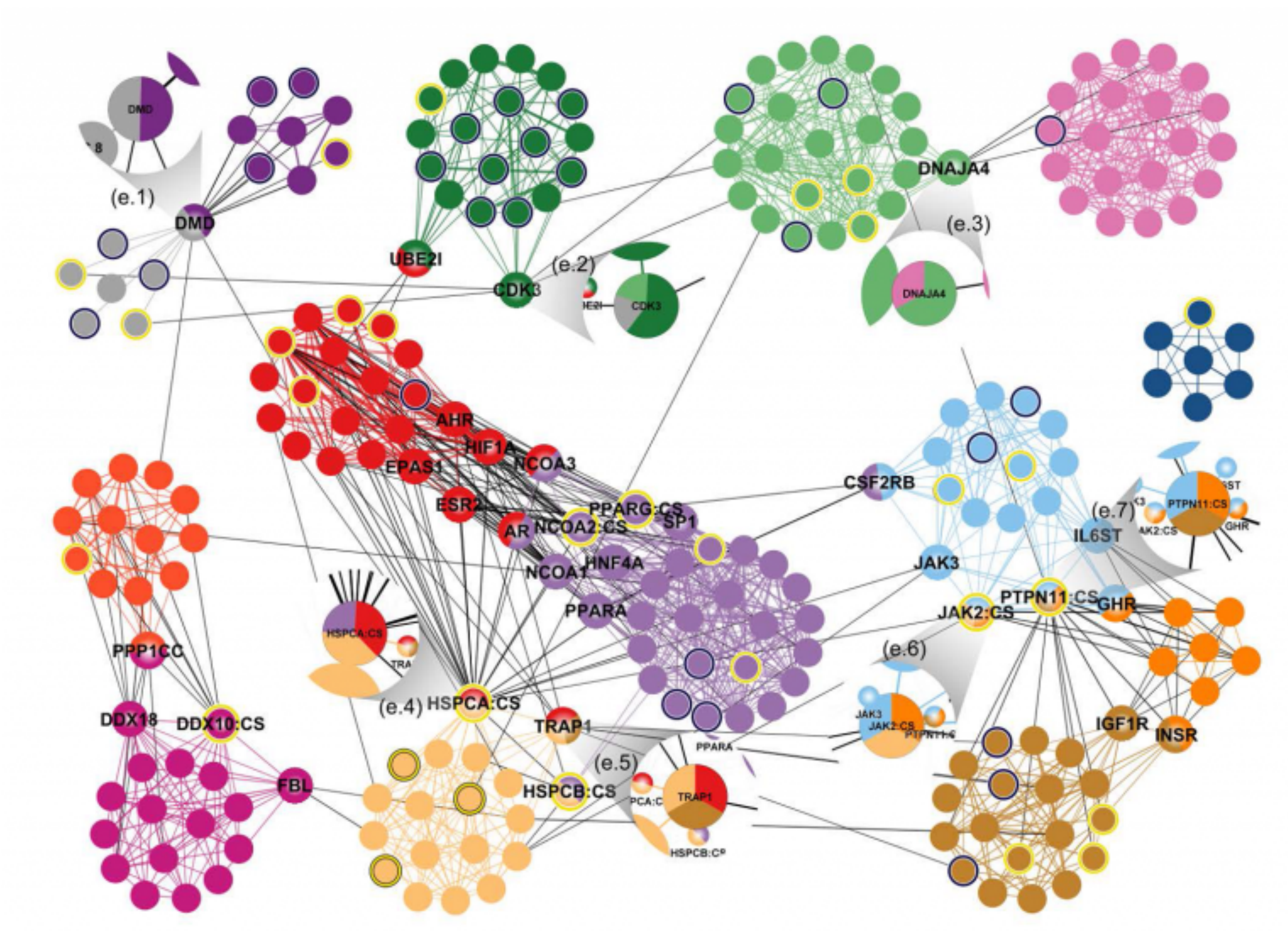
Generative Causal Explanations for Graph Neural Networks

Wanyu Lin, Hao Lan, Baochun Li
The Hong Kong Polytechnic University, University of Toronto


Cellular interaction graphs





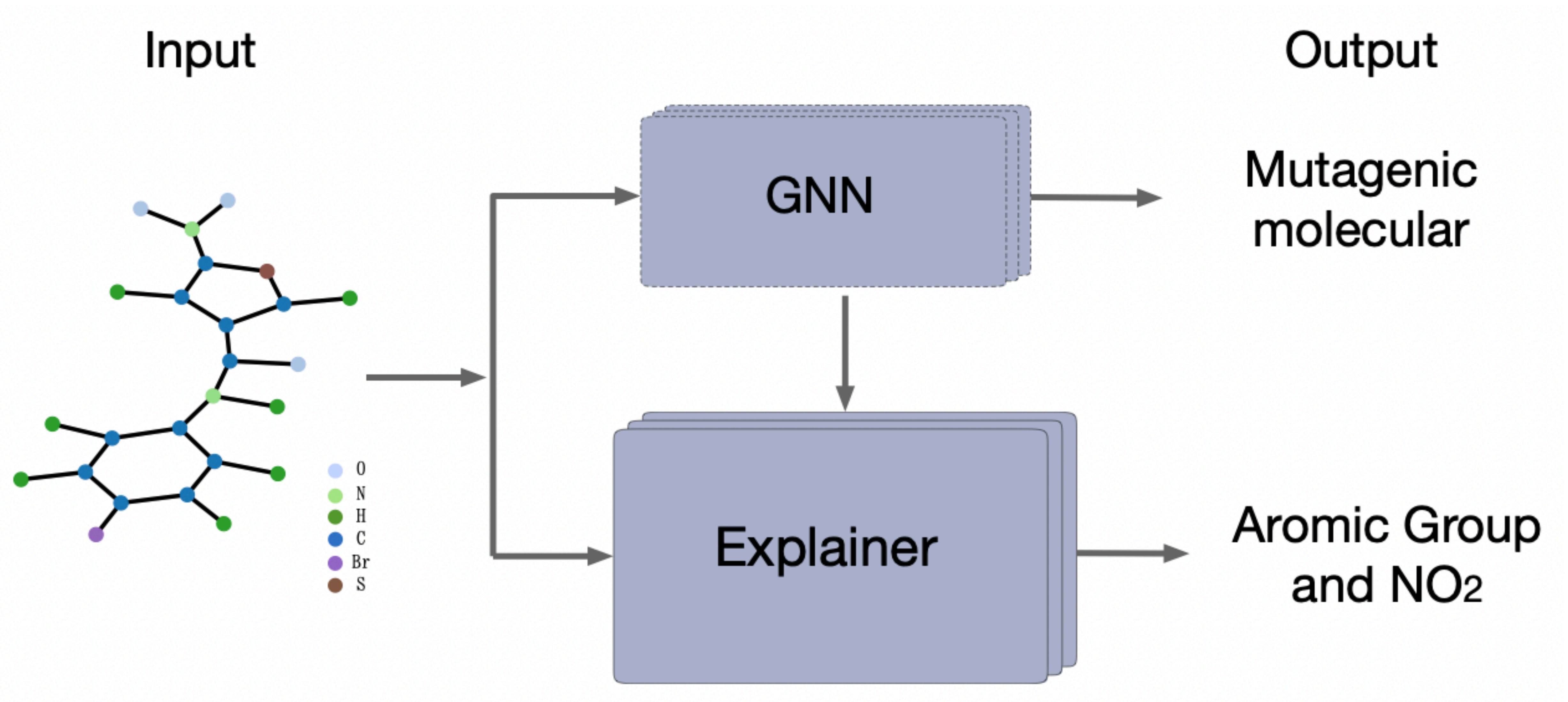
Protein-protein interaction graph

We aim to uncover the veil of the GNN by interpreting its predictions.

Problem

- Given: a pre-trained GNN for classification, an instance (an input graph) from the data distribution.

- Objective: to obtain an explanation mechanism that can identify the most relevant part of the input (a compact subgraph), causing the prediction of the GNN.



Mutag [1]

[1] Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., and Hansch, C. Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital En- ergies and Hydrophobicity. Journal of Medicinal Chem- istry, 34(2):786–797, 1991

Related work

- GNNExplainer (NeurIPS 2019): explains each instance separately.

- PGExplainer (NeurIPS 2020): trains a multilayer-perceptron (MLP) to provide explanations.

Our solution

- We propose to train a graph generator as an explainer — the input is a graph, and the output is the explanatory subgraph structure.
  - Once trained, it can be used to explain any input graph with little time.
  - Our explainer is model-agnostic — does not need to know the internal structure of the target GNN.

What is the supervision signal for training our explanation model?

Our solution (*Cont.*)

- We propose a graph distillation mechanism that can extract the most relevant part of the graph leading to the predictions of the target GNN.

  - We quantify the edge importance with the notion of *Granger causality* — In the graph domain, if the absence of an edge decreases the ability to predict Y, then there is a causal relationship between this edge and its corresponding prediction.

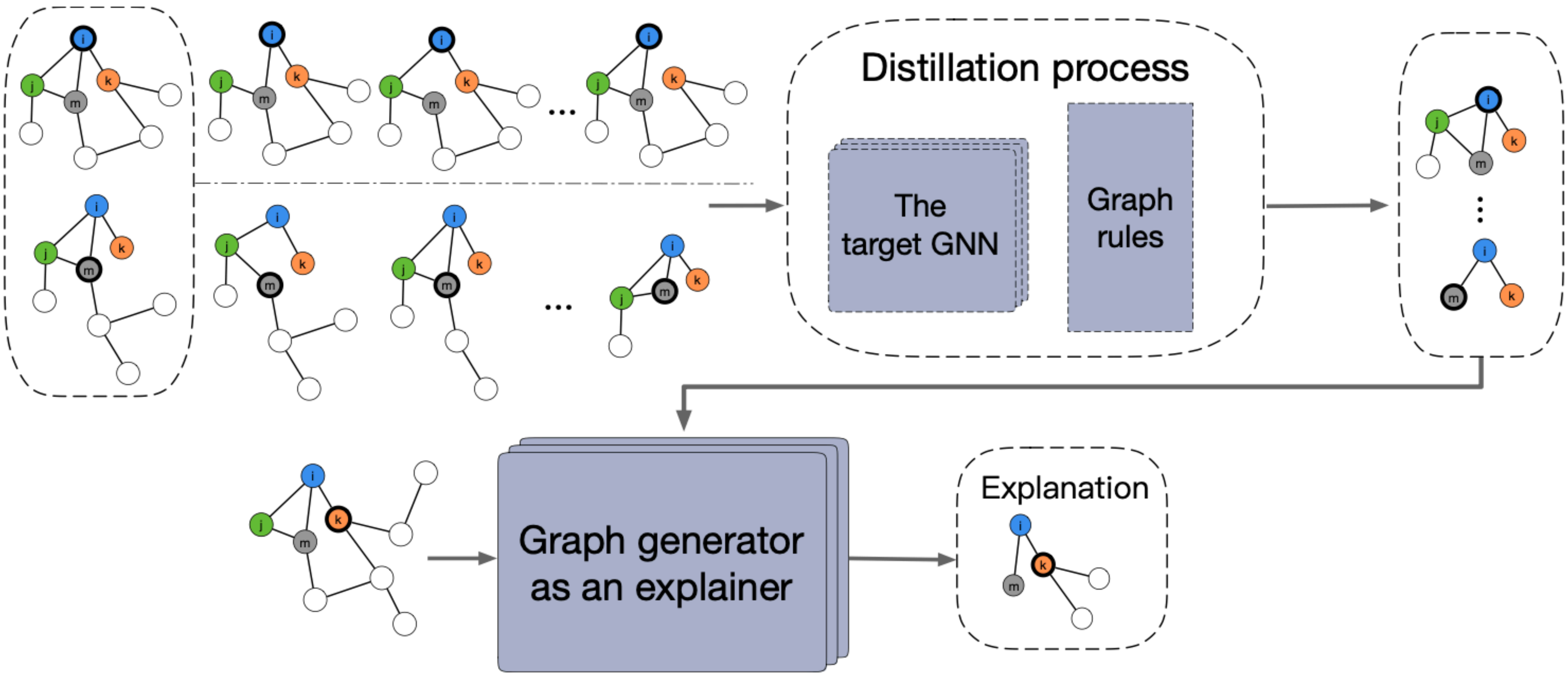
  - With the importance quantification, we can extract the top-K most important edges as the explanatory subgraph.

Our solution (*Cont.*)

- The causal contribution of the edge e_j is defined as the decrease in model error, formulated as:

$$\Delta_{\delta, e_j} = \delta_{G^c \setminus \{e_j\}} - \delta_{G^c}$$

- Incorporate graph rules: such connectivity checking, etc.



Our framework | *Gem*

Experiments

- Baselines: GNNExplainer (NeurIPS 2019) and PGExplainer (NeurIPS 2020)
- Benchmarking datasets:
  - Graph classification tasks: MUTAG and NCI1
  - Node classification tasks: BA-shapes and Tree-cycles

Explanation accuracy

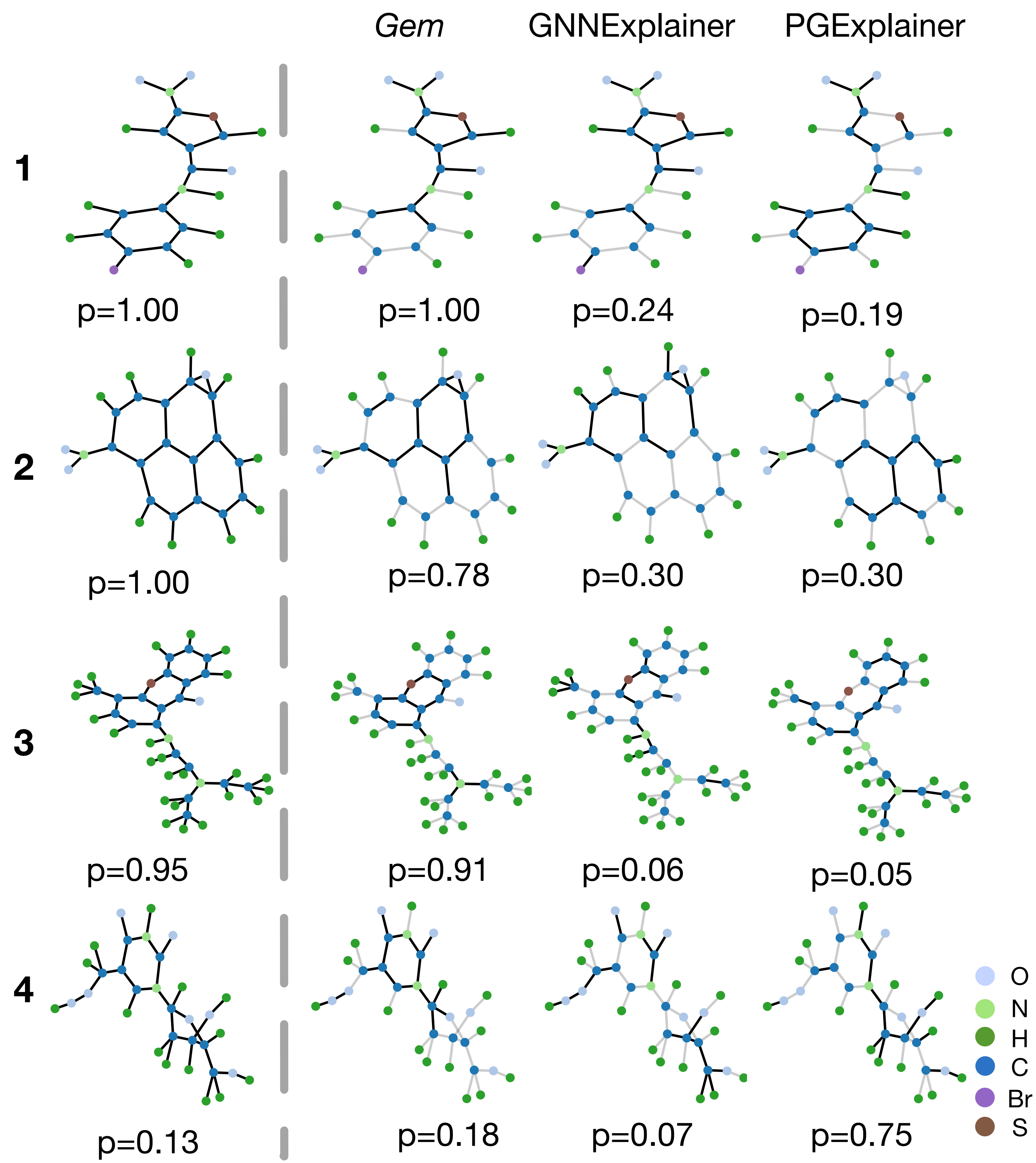
Table 1. Explanation Accuracy on Synthetic Datasets (%).

	BA-SHAPES					TREE-CYCLES				
K	5	6	7	8	9	6	7	8	9	10
Gem	**93.4**	**97.1**	**97.1**	**97.1**	**99.3**	86.1	**87.5**	**92.5**	**93.9**	95.4
GNNExplainer	82.4	88.2	91.2	91.2	94.1	14.3	46.8	74.6	91.4	**96.1**
PGExplainer	71.9	90.7	92.0	93.3	94.1	**94.4**	80.6	77.0	82.4	89.4

Table 2. Explanation Accuracy on Real-World Datasets (%).

	MUTAG				NCI1			
K	15	20	25	30	15	20	25	30
Gem-0	**64.0**	**78.1**	**81.0**	**85.0**	–	–	–	–
GNNExplainer-0	60.0	67.6	68.9	75.8	–	–	–	–
PGExplainer-0	22.5	38.5	57.6	72.3	–	–	–	–
Gem	66.3	**78.0**	**82.1**	**83.4**	56.9	**65.3**	68.9	**72.8**
GNNExplainer	**67.1**	74.9	75.8	80.9	**59.3**	61.8	**69.6**	72.0

Visualization



Explanation time

Table 3. Inference Time per Instance (ms).

DATASETS	BA-SHAPES	TREE-CYCLES	MUTAG	NCI1
GNNEXPLAINER	265.2	204.5	257.6	259.8
PGEXPLAINER	6.7	6.5	5.5	–
GEM	0.5	0.5	0.05	0.02

https://github.com/wanyu-lin/ICML2021-Gem

The Hong Kong Polytechnic University
香港理工大學
Department of Computing
電子計算學系