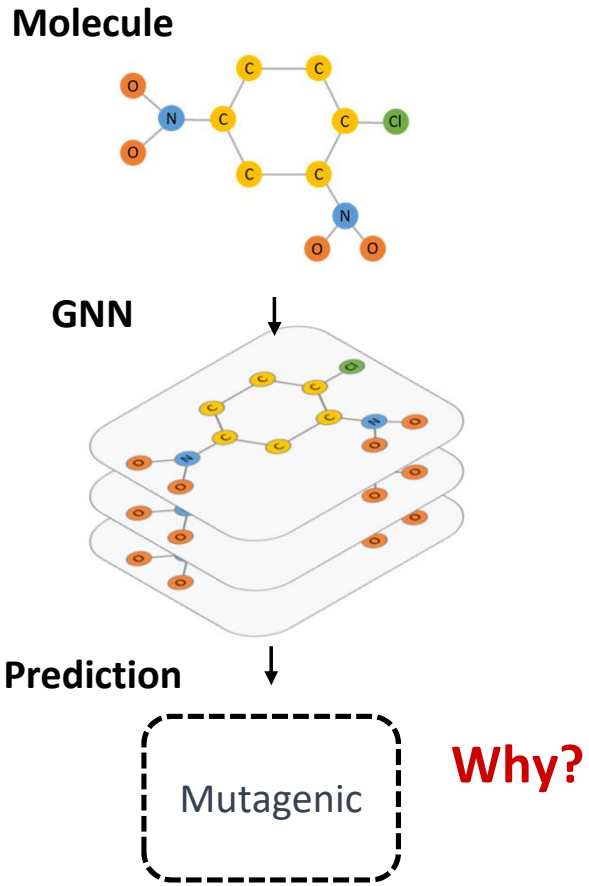
SAME: Uncovering GNN Black Box with Structure-aware Shapley-based Multipiece Explanation

Ziyuan Ye*, Rihan Huang*, Qilin Wu, Quanying Liu†

Ye, Z.*, Huang, R.*, Wu, Q., & Liu, Q. (2023). SAME: Uncovering GNN black box with structure-aware Shapley-based multipiece explanation. *Thirty-seventh Conference on Neural Information Processing Systems.*

* Equal contribution, co-first author
† Corresponding author.

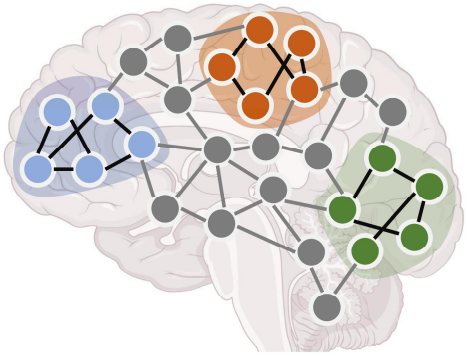
A brief intro: XAI in graph



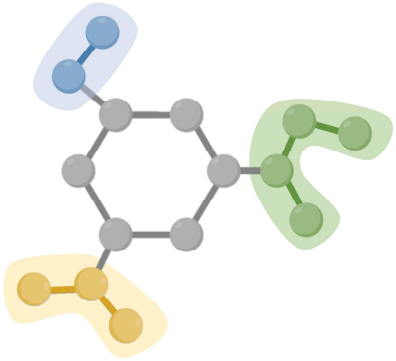
Why?

- Deep graph models becoming more widespread
- Black-box graph models are the mainstream
  - Graph Convolutional Network (GCN)
  - Graph Attention Network (GAT)
  - Graph Isomorphism Network (GIN)
  - …
- Various concerns about model transparency
- Analyzing the influence of a single node or edge is not enough.

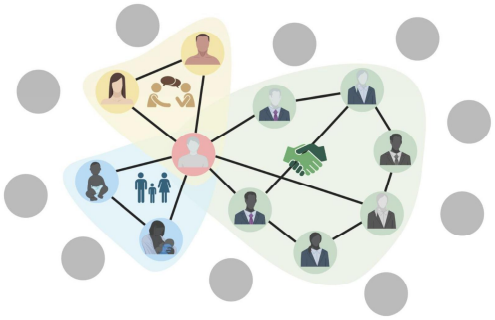
Motivation



Brain Networks Associated with Specific Cognitive Task



Functional Groups Associated with Molecular Property

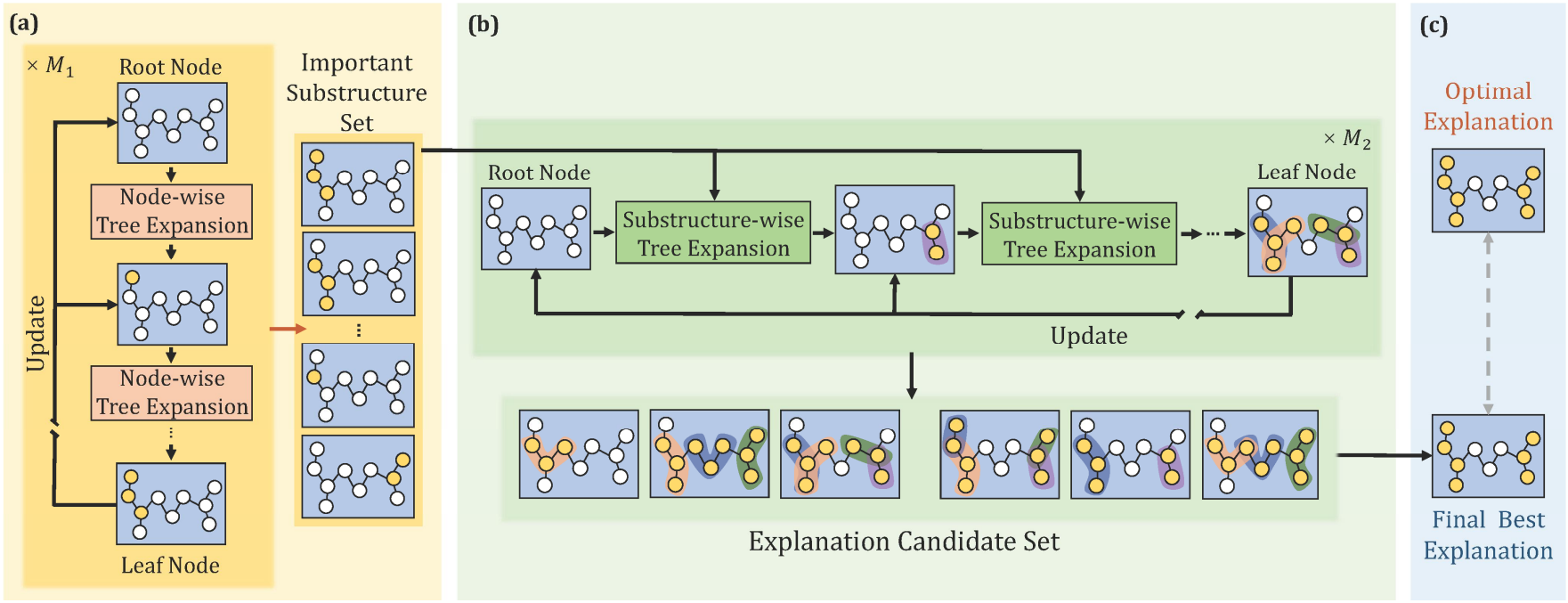


Social Groups Associated with Single Person

- The characteristics and properties within a graph or node tend to be jointly influenced by more than one high-order connected community of the graph.

➢ To design an explanation method: Retain important nodes while avoiding irrelevant nodes.

Methodology



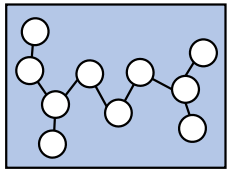
(a) **Important substructure initialization phase**

- Search the single connected important substructure.

(b) **Explanation exploration phase**

- Provide a candidate set of explanations
- Optimize the combination of different important substructures.

Important substructure initialization phase

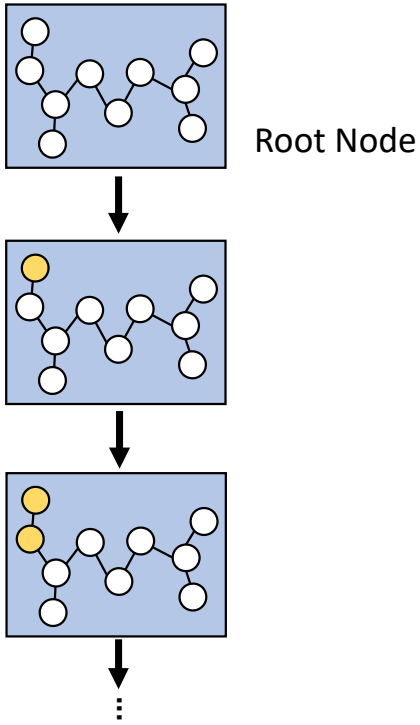


Root Node

Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph

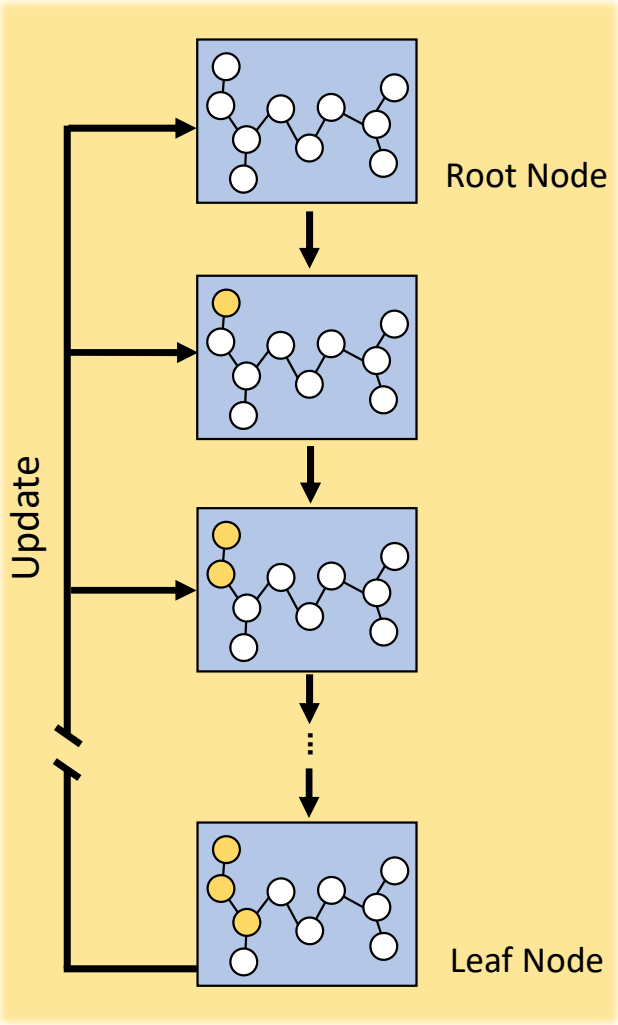
Important substructure initialization phase



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph
- Expand within 1-hop neighbors of the associate substructure
- Choose the best children according to the Shapley value

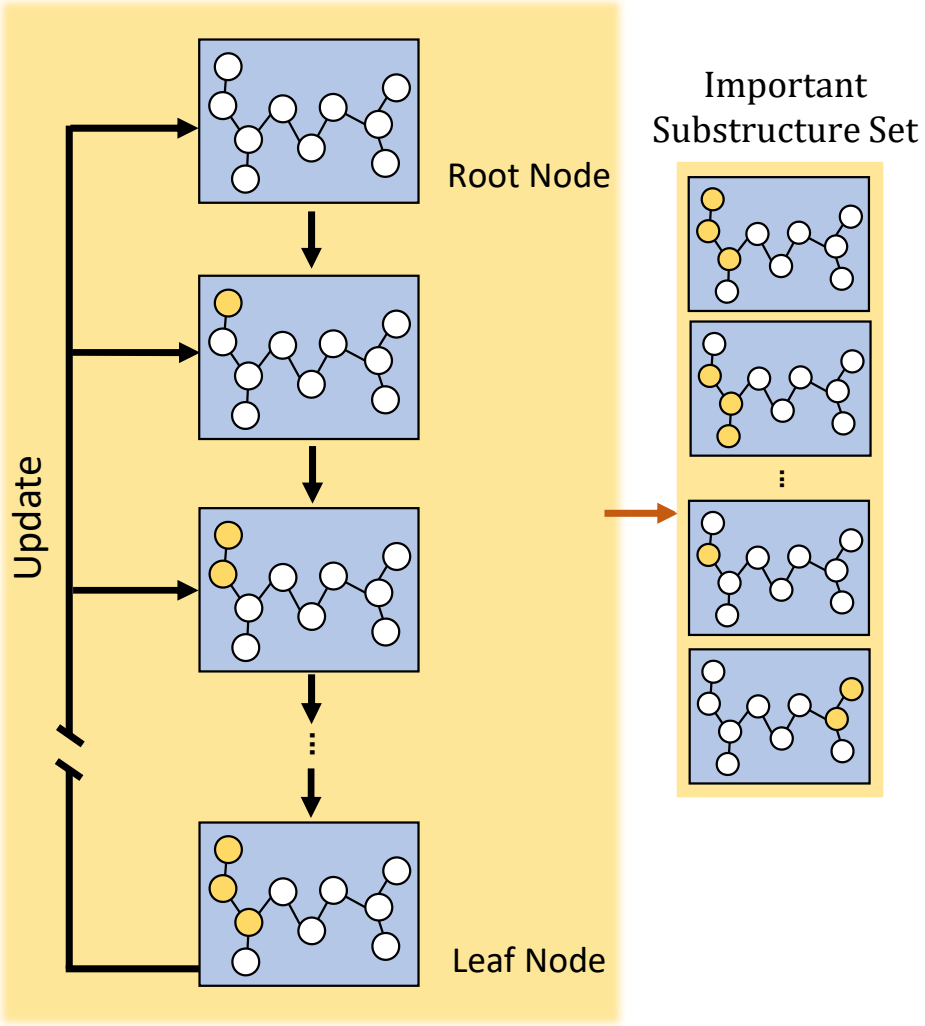
Important substructure initialization phase



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph
- Expand within 1-hop neighbors of the associate substructure
- Leaf Node: The substructure reaches the maximum size predefined
- Backpropagation to update the previous nodes

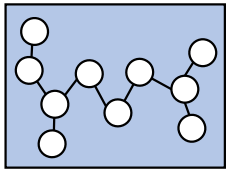
Important substructure initialization phase



Expansion-based Monte Carlo Tree Search (MCTS)

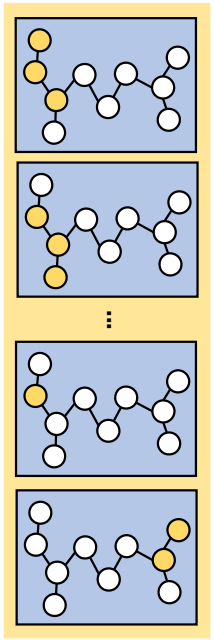
- Root Node: Empty graph
- Expand within 1-hop neighbors of the associate substructure
- Leaf Node: The substructure reaches the predefined maximum size
- Backpropagation to update the previous nodes
- Important substructure set: All the substructures in the MCTS

Explanation exploration phase



Root Node

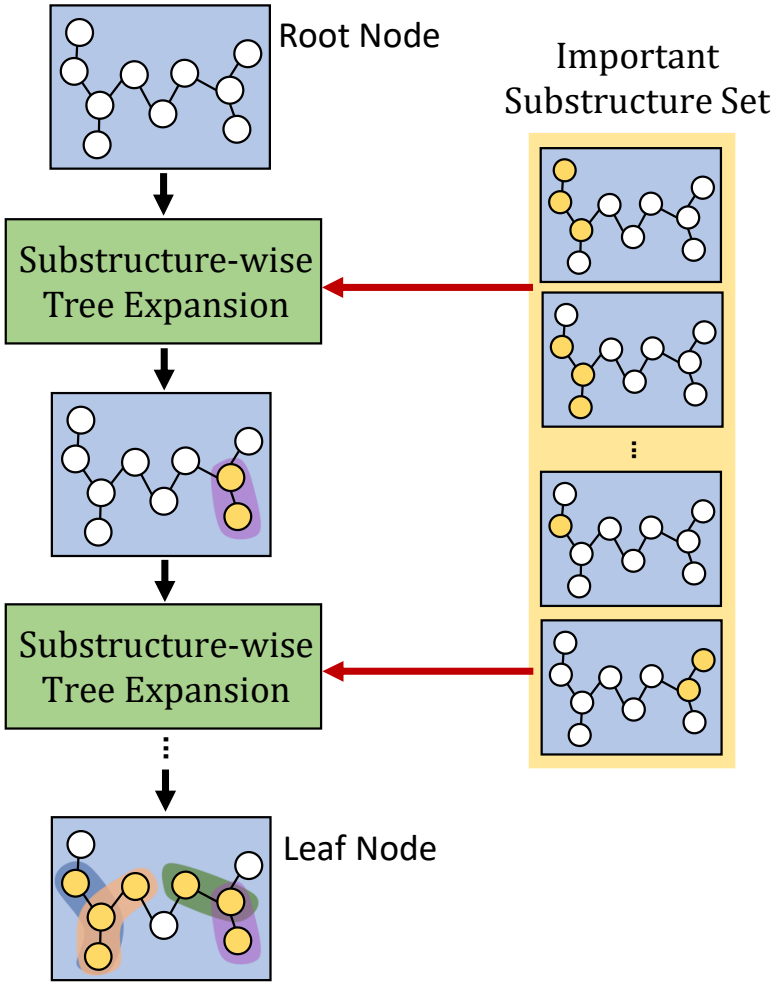
Important
Substructure Set



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph

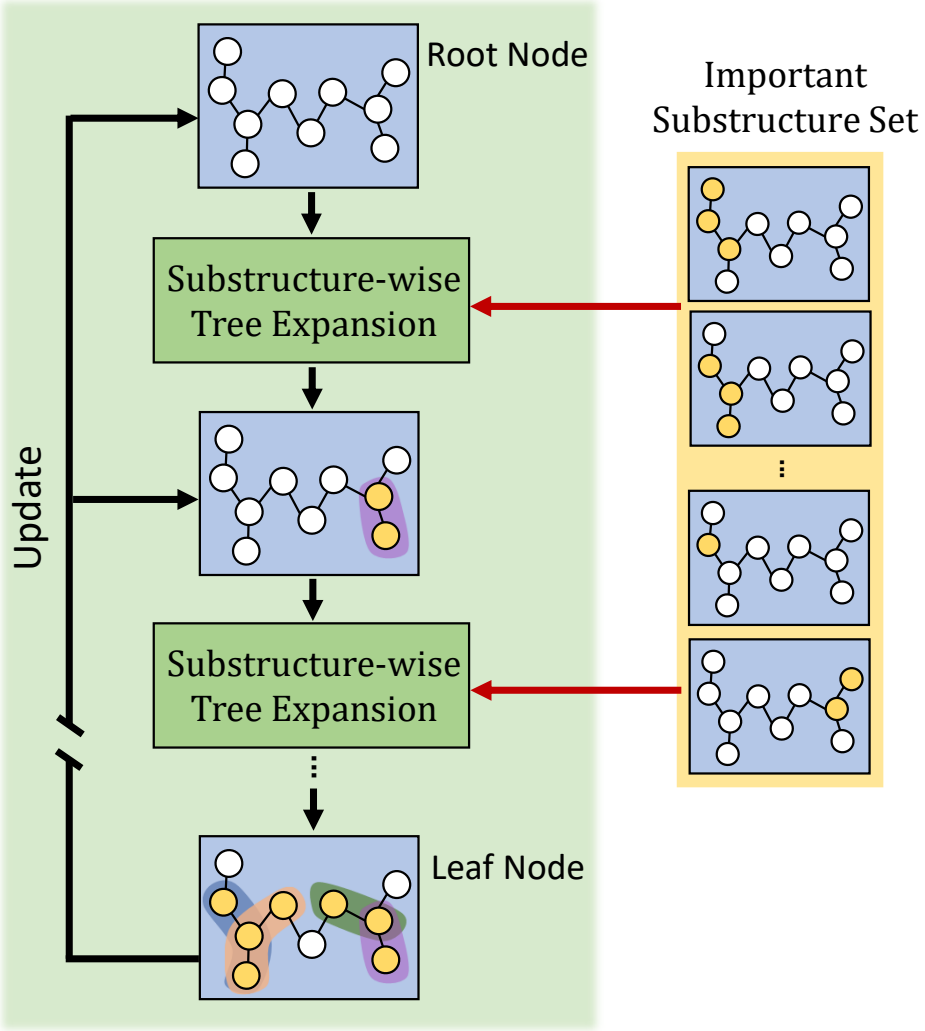
Explanation exploration phase



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph
- Expand an important substructure
- Leaf Node: The size reaches the threshold

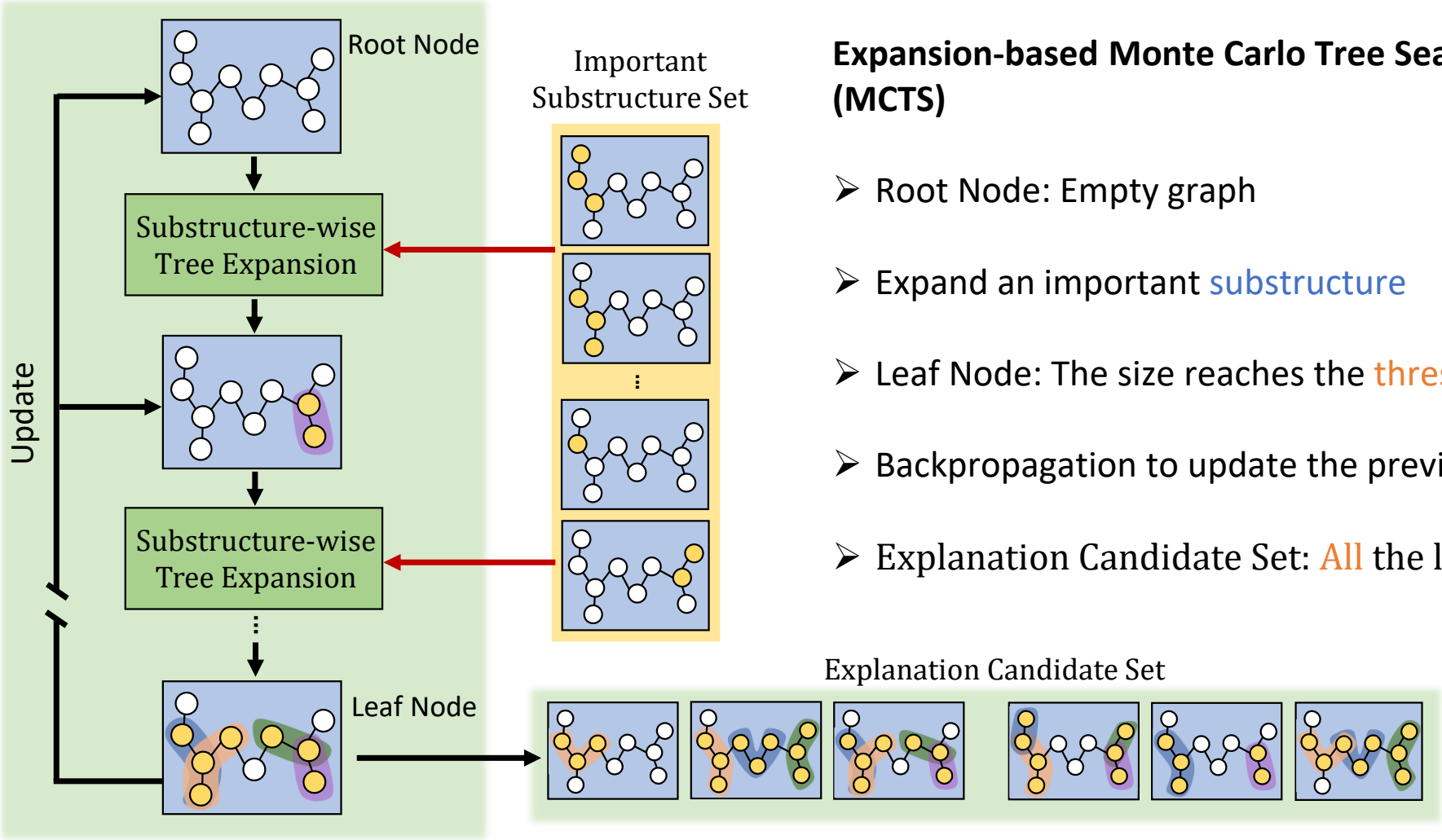
Explanation exploration phase



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph
- Expand an important substructure
- Leaf Node: The size reaches the threshold
- Backpropagation to update the previous nodes

Explanation exploration phase



Expansion-based Monte Carlo Tree Search (MCTS)

- Root Node: Empty graph
- Expand an important substructure
- Leaf Node: The size reaches the threshold
- Backpropagation to update the previous nodes
- Explanation Candidate Set: All the leaf nodes

Experiments

Table 2: Comparison of our SAME and other baseline using fidelity.

Methods \ Dataset	Graph classification					Node classif.
	Molecular graph		Semantic graph		Synthetic graph	
	BBBP	MUTAG	Graph-SST2	Graph-SST5	BA-2Motifs	BA-Shapes
Grad-CAM [22]	0.226±0.036	0.261±0.018	0.257±0.056	0.229±0.042	0.472±0.010	-
GNNExplainer [35]	0.148±0.041	0.188±0.031	0.143±0.041	0.170±0.046	0.442±0.026	0.154±0.000
PGExplainer [19]	0.197±0.043	0.156±0.004	0.219±0.040	0.207±0.036	0.431±0.011	0.135±0.020
GNN-LRP [25]	0.111±0.040	0.253±0.030	0.103±0.042	0.131±0.057	0.146±0.010	0.155±0.000
SubgraphX [38]	0.433±0.073	0.379±0.030	0.262±0.027	0.283±0.042	0.493±0.003	0.181±0.005
GStarX [40]	0.117±0.043	0.656±0.096	0.183±0.050	0.186±0.050	0.476±0.014	-
SAME	**0.489±0.034**	**0.702±0.125**	**0.373±0.042**	**0.393±0.022**	**0.549±0.004**	**0.214±0.000**
Relative Improve	**12.9%↑**	**7.01%↑**	**42.3%↑**	**38.9%↑**	**11.3%↑**	**18.2%↑**

Note: The fidelity results are averaged across different sparsity from 0.5 to 0.8. The quantitative results are presented in the form of mean ± std. The previous SOTA results on different datasets are marked with an underline. *Relative Improve* denotes the relative improvement of our SAME method over the SOTA methods.

Table 3: Comparison of inference time (in seconds) on different datasets.

Methods \ Dataset	BBBP	MUTAG	Graph-SST2	Graph-SST5	BA-2Motifs	BA-Shapes
Grad-CAM [22]	0.16	0.23	0.39	0.44	0.14	-
GNNExplainer [35]	7.56	1.96	7.64	19.39	1.89	2.72
PGExplainer [19]	0.15	0.21	0.35	0.43	0.12	0.13
GNN-LRP [25]	2.37	1.97	5.84	5.47	3.30	51.77
SubgraphX [38]	26.72	151.75	36.48	71.32	85.50	162.80
GStarX [40]	84.54	25.24	30.64	54.49	77.99	-
SAME	7.86	5.67	6.06	8.83	8.19	14.08

Note: The PGExplainer needs training before inferring the explanation.

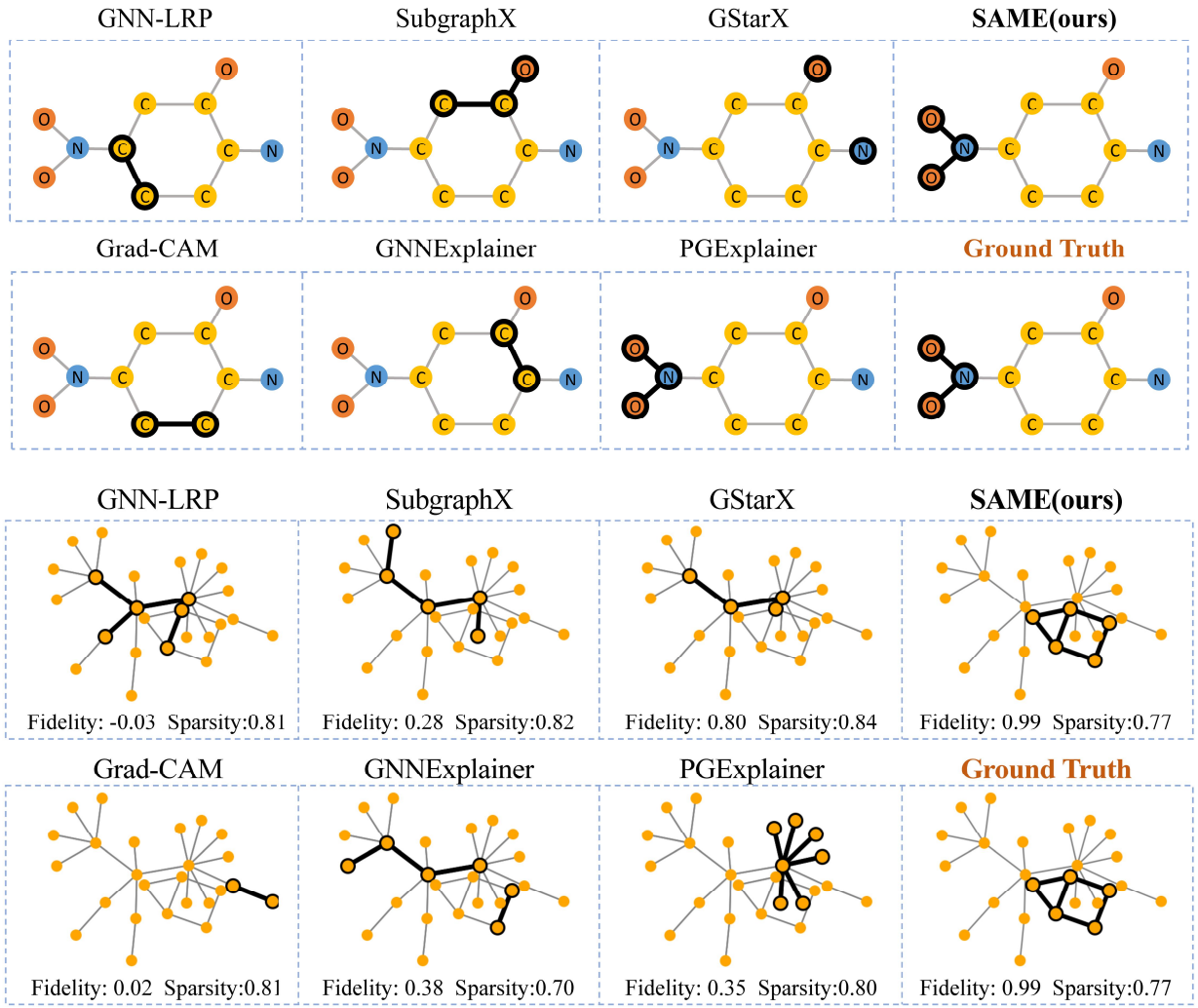
Fidelity

- A higher fidelity demonstrates a better explainability.
- SAME outperforms the SOTA baselines among different tasks and datasets.

Inference Time

- SAME consistently achieves much lower computational cost compared to GStarX and SubgraphX, reflecting its efficiency and robustness.

Experiments



MUTAG dataset

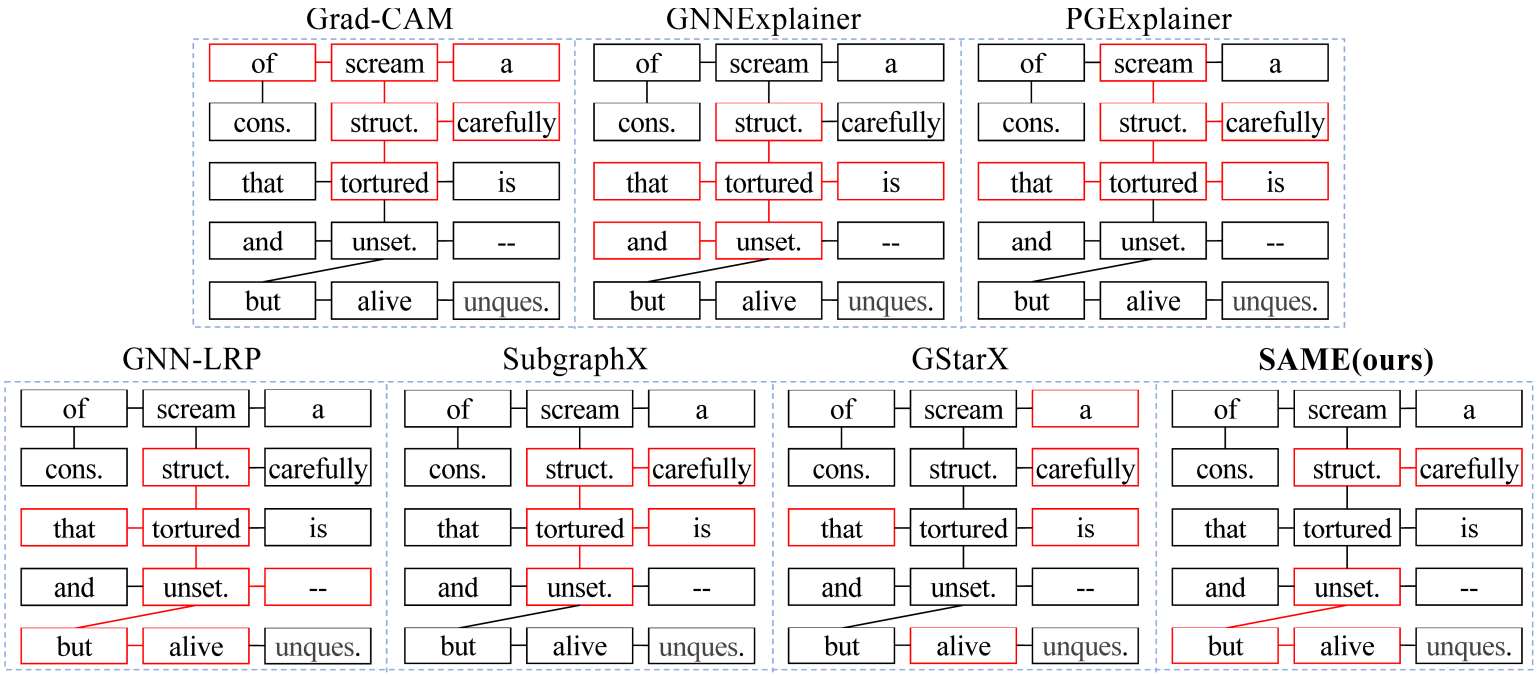
- SAME achieves to provide the explanations the same as the ground truth (-NO_2) which are labeled by human experts.

BA-2Motifs dataset

- SAME exactly finds the ground-truth explanation (a 5-node-house-structure motif) compared to other baselines.

Experiments

Sentence: “a carefully structured scream of consciousness that is tortured and unsettling -- but unquestionably alive.”



Graph-SST2 dataset

- SAME can well capture the adjectives-or-adverbs-like graph structures than other baselines.

Thanks for your attention!