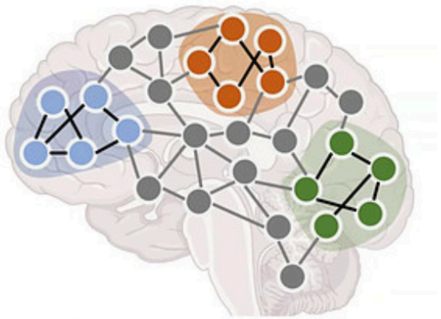
Structure-aware Graph Explainer via Myerson-Taylor Interaction Index

Ngoc Bui, Hieu Trung Nguyen, Viet Anh Nguyen, Rex Ying

Explainability in Graph Neural Network

- Graph Neural Networks (GNNs) are ubiquitous thanks to their predictive power in many applications. However, the transparency of GNN's predictions remains a major problem since they are usually considered as a black-box predictor.



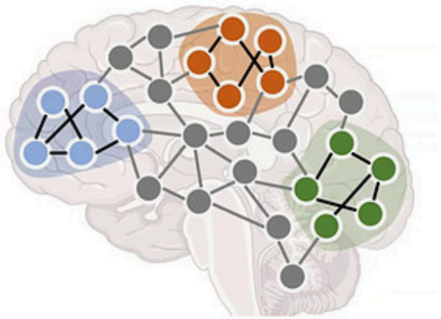
Brain cognitive prediction

Molecular Property Prediction

Terrorist Forcasting

Explainability in Graph Neural Network

- Graph Neural Networks (GNNs) are ubiquitous thanks to their predictive power in many applications. However, the transparency of GNN's predictions remains a major problem since they are usually considered as a black-box predictor.
- Explainability is crucial in domains where the impact of decisions is considerable.



Brain cognitive prediction

Which brain regions are associated with a specific cognitive task?

Molecular Property Prediction

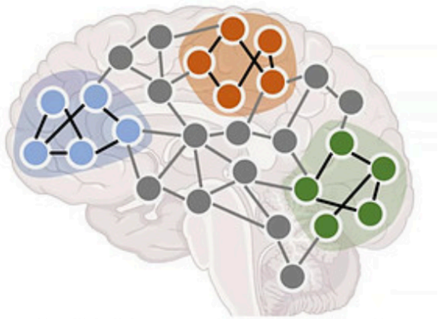
Which functional groups are associated with mutagenic?

Terrorist Forcasting

Why is this user likely to be a terrorist?

Explainability in Graph Neural Network

- Graph Neural Networks (GNNs) are ubiquitous thanks to their predictive power in many applications. However, the transparency of GNN's predictions remains a major problem since they are usually considered as a black-box predictor.
- Explainability is crucial in domains where the impact of decisions is considerable.
- **The goal of explainability**: Which **motifs/patterns** are **"most influential"** to the model prediction?



Brain cognitive prediction

Which brain regions are associated with a specific cognitive task?

Molecular Property Prediction

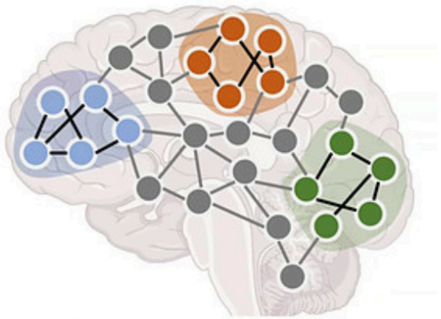
Which functional groups are associated with mutagenic?

Terrorist Forcasting

Why is this user likely to be a terrorist?

Explainability in Graph Neural Network

- Graph Neural Networks (GNNs) are ubiquitous thanks to their predictive power in many applications. However, the transparency of GNN's predictions remains a major problem since they are usually considered as a black-box predictor.
- Explainability is crucial in domains where the impact of decisions is considerable.
- **The goal of explainability**: Which **motifs/patterns** are **"most influential"** to the model prediction?
  - What is considered a **motif/pattern**? What should **an explanation** look like?
  - How do we measure the "**influence**" of a motif on the model prediction?



Brain cognitive prediction

Which brain regions are associated with a specific cognitive task?

Molecular Property Prediction

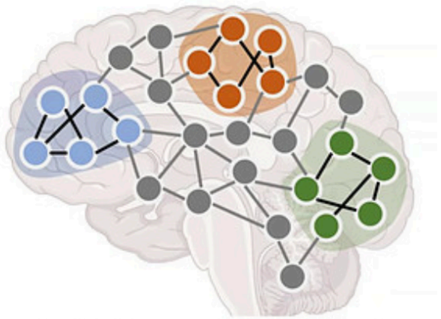
Which functional groups are associated with mutagenic?

Terrorist Forcasting

Why is this user likely to be a terrorist?

What should an explanation look like?

- What is considered a **motif/pattern**?
  - Motif is a connected subgraph of the input graph.



Brain cognitive prediction

Which brain regions are associated with a specific cognitive task?

Molecular Property Prediction

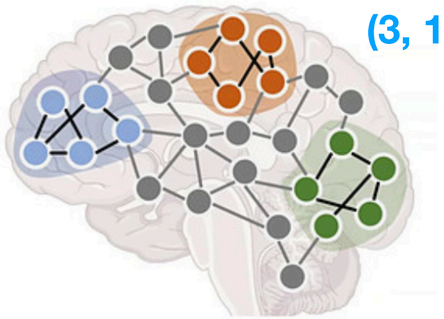
Which functional groups are associated with mutagenic?

Terrorist Forcasting

Why is this user likely to be a terrorist?

What should an explanation look like?

- What is considered a **motif/pattern**?
  - Motif is a connected subgraph of the input graph.
- What should **an explanation** look like?
  - An explanation may consist of one or multiple disjoint motifs (subgraph) of the input graph.
  - Formally, an (m, M)-explanation for the input graph G contains *at most* m disjoint subgraphs with the total *at most* M nodes.



(3, 15)-explanation

Brain cognitive prediction

Which brain regions are associated with a specific cognitive task?

(3, 9)-explanation

(3, 10)-explanation

(4, 10)-explanation

Molecular Property Prediction

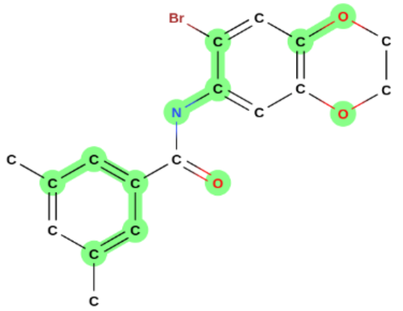
Which functional groups are associated with mutagenic?

Terrorist Forcasting

Why is this user likely to be a terrorist?

What should an explanation look like?

- What is considered a **motif/pattern**?
  - Motif is a connected subgraph of the input graph.
- What should **an explanation** look like?
  - An explanation may consist of one or multiple disjoint motifs (subgraph) of the input graph.
  - Formally, an (m, M)-explanation for the input graph G contains *at most* m disjoint subgraphs with the total *at most* M nodes.
- Why? This is a *general* definition covering every possible subgraph, and *users can choose* (m, M) for their interpretability budget. Previous work only considers M nodes as a constraint.
  - As m and M increase, the explanation becomes harder to understand for humans.
- **Our goal**: given a budget of explanation (m, M),
  we want to retain as "**much information**" about the model prediction as possible.

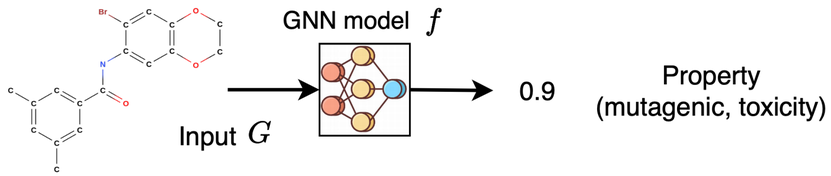


(5, 12)-explanation

not intuitive for human to understand

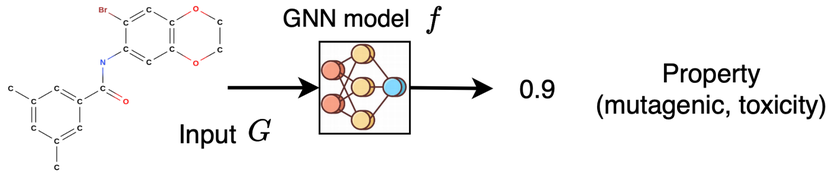
Settings

- **Given**:
  - A graph $G = (V, E)$ where V is the set of vertices and E is the set of edges.
  - A black-box model $f(\cdot)$ that takes a graph as the input and produces a prediction $f(G) \in \mathbb{R}$.

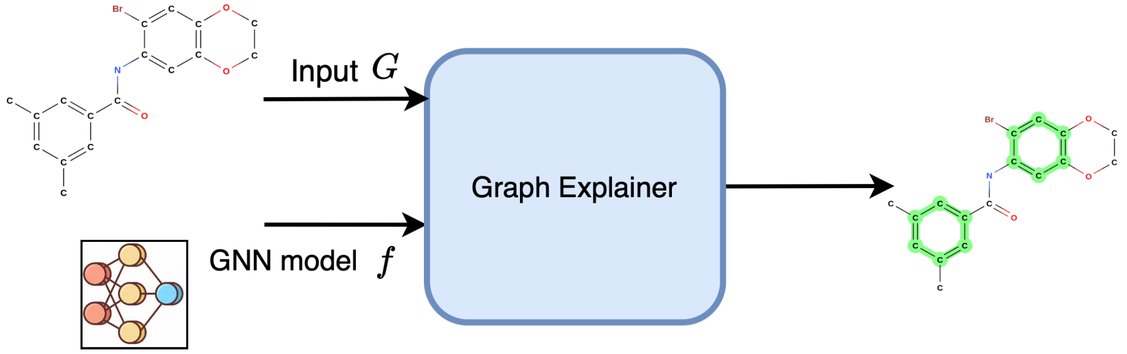


Settings

- **Given**:
  - A graph $G = (V, E)$ where V is the set of vertices and E is the set of edges.
  - A black-box model $f(\cdot)$ that takes a graph as the input and produces a prediction $f(G) \in \mathbb{R}$.



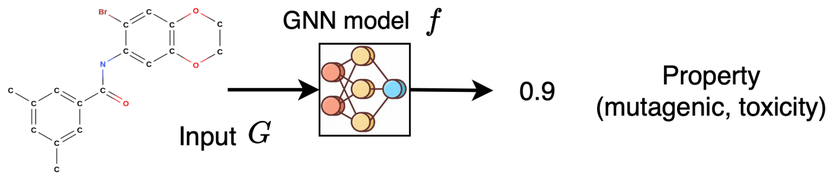
- **Goal of the Graph Explainer:**
  - Find a subgraph $H = (S, E_S)$ that is most influential to the model prediction such that $S = \{S_1, S_2, \ldots, S_m\} \subseteq V$ and E_S is the subset of edges induced by the set S



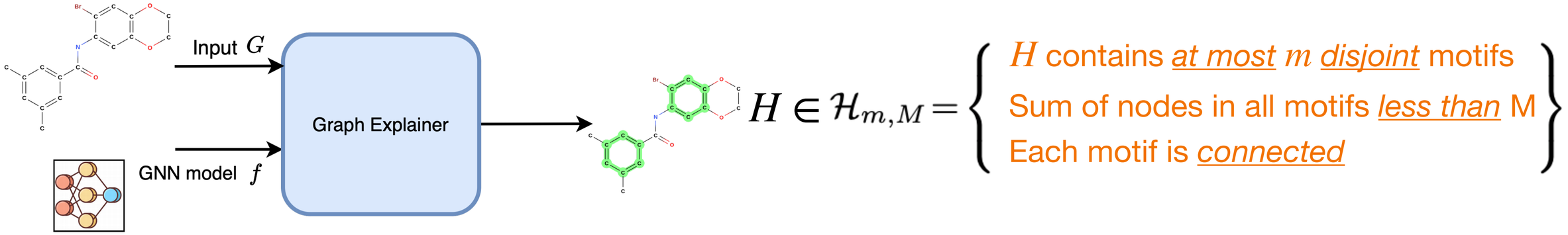
$$H \in \mathcal{H}_{m,M} = \left\{ \begin{array}{l} (S_1, \ldots, S_m) \subseteq V^m \text{ such that :} \\ S_l \cap S_h = \emptyset \; \forall l, h, \;\; |\cup_{l=1}^{m} S_l| \leq M \\ S_l \text{ induces a } \textit{connected} \text{ subgraph} \end{array} \right\}$$

Settings

- **Given**:
  - A graph $G = (V, E)$ where V is the set of vertices and E is the set of edges.
  - A black-box model $f(\cdot)$ that takes a graph as the input and produces a prediction $f(G) \in \mathbb{R}$.



- **Goal of the Graph Explainer:**
  - Find a subgraph $H = (S, E_S)$ that is most influential to the model prediction such that $S = \{S_1, S_2, \ldots, S_m\} \subseteq V$ and E_S is the subset of edges induced by the set S



How to measure the “importance” of a motif?

- Most of the **existing methods** for explaining black-box GNN models boil down to assigning node-wise importance scores.
  - assigning a node i a value w_i
  - the importance/attribution of a model is then defined as the sum of node-wise importance

$$\text{GroupAttr}(S) = \sum_{i \in S} w_i$$

We will use importance and attribution interchangeably.

How to measure the “importance” of a motif?

- Most of the **existing methods** for explaining black-box GNN models boil down to assigning node-wise importance scores.
  - assigning a node i a value w_i
  - the importance/attribution of a model is then defined as the sum of node-wise importance

$$\text{GroupAttr}(S) = \sum_{i \in S} w_i$$

- One prominent method to assign such importance scores is to use the Shapley values

$$\phi_i = \frac{1}{|V|} \sum_{T \subseteq V \setminus i} \frac{1}{\binom{|V|-1}{|T|}} (f(T \cup i) - f(T))$$

i.e., the contribution of a node i is computed by its weighted average marginal contribution over all possible subgraphs the node i could join.

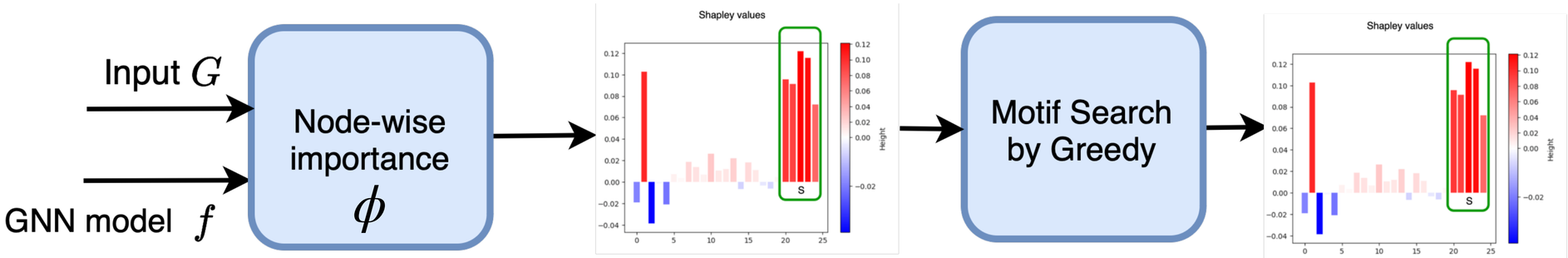
We will use importance and attribution interchangeably.

How to measure the “importance” of a motif?

- Most of the **existing methods** for explaining black-box GNN models boil down to assigning node-wise importance scores.
  - assigning a node i a value ϕ_i
  - the importance/attribution of a model is then defined as the sum of node-wise importance

$$\text{GroupAttr}(S) = \sum_{i \in S} \phi_i$$

- Pipeline of most **existing methods**:



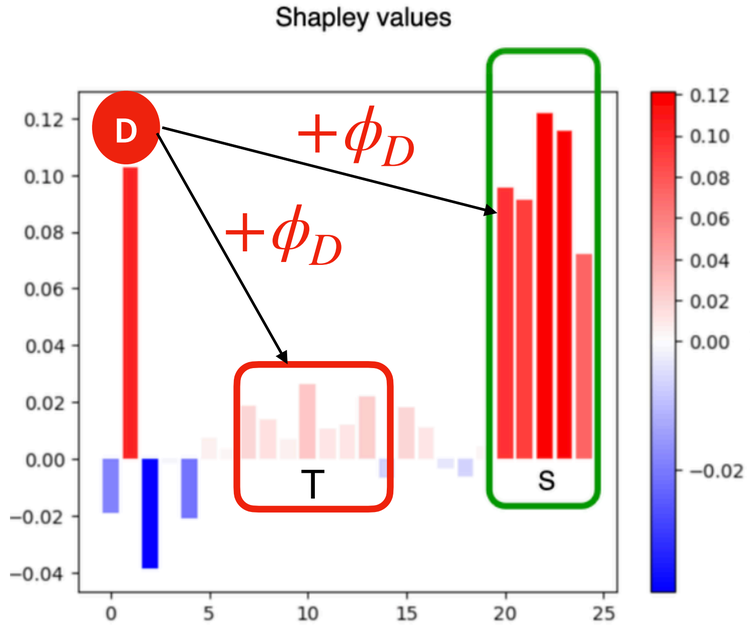
Compute node-wise importance for each node using Shapley value

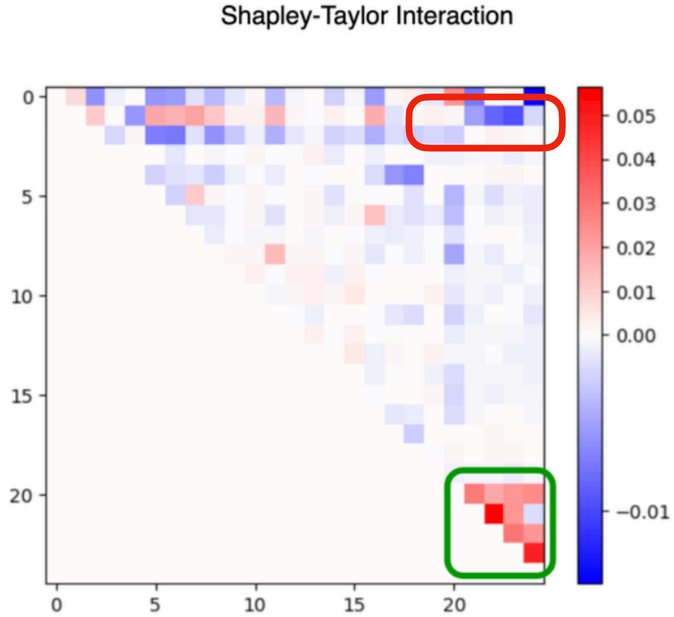
Greedily choose M nodes with highest importance score

Drawbacks

- Node-wise importance is not suitable to identify multiple motifs.

Should we highlight $\{D\} \cup S$ or $\{D\} \cup T$?





D has the same value when joining different groups
Interactions of D with members in S and T is neglected.
Using greedy algorithm, we can **highlight** $\{S \cup D\}$

However, using a higher-order interaction index, we can observe that
D has a negative contribution when interacting with members in S
Thus, we should not form $D \cup S$ as a motif.

Drawbacks

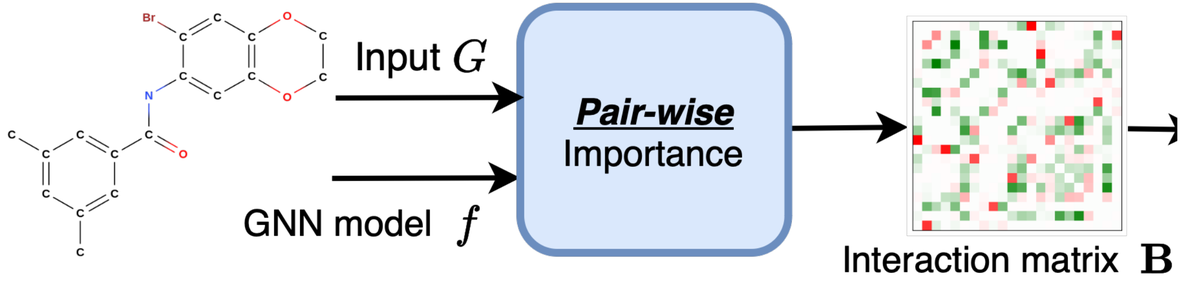
- Node-wise importance is not suitable to identify multiple motifs.
- Shapley value does not take the graph structure into account when computing importance score.

$$\phi_i = \frac{1}{|V|} \sum_{T \subseteq V \setminus i} \frac{1}{\binom{|V|-1}{|T|}} (f(T \cup i) - f(T))$$

  - Thus, Shapley value requires evaluating the black-box model at *all possible subgraphs.*
  - Some of them might be disconnected and pathological, which the model F does not see during the training —> Evaluating the model at these examples might inject bias into the importance score.

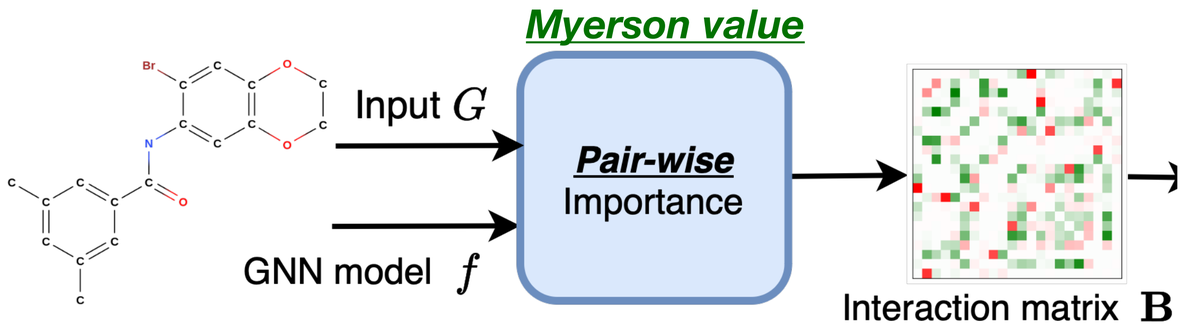
Our Key Idea

- Instead of using *node-wise* importance scores, we propose to use *pair-wise* importance scores.



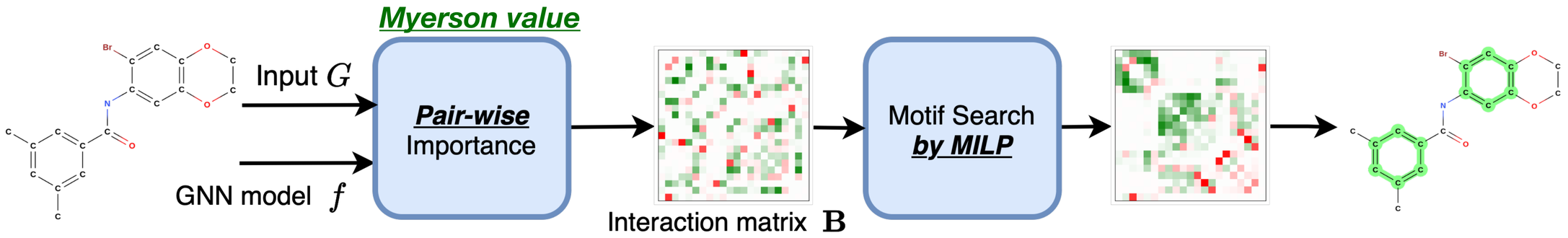
Our Key Idea

- Instead of using *node-wise* importance scores, we propose to use *pair-wise* importance scores.
- Generalize *the Shapley value* to *the Myerson value* to take the graph structure into account.



Our Key Idea

- Instead of using *node-wise* importance scores, we propose to use *pair-wise* importance scores.
- Generalize *the Shapley value* to *the Myerson value* to take the graph structure into account.
- Instead of using *greedy*, we use pair-wise importance to formulate a *mixed integer linear problem* to highlight multiple influential motifs to the model predictions, both positive and negative.



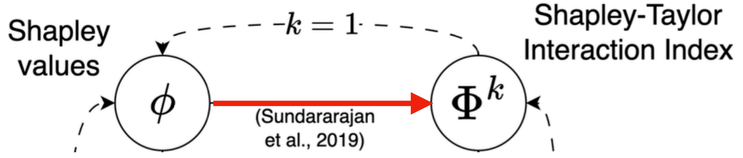
The Shapley Value to Higher-order Interactions

- **Recall**: Shapley value assigns for *each node* i an importance score ϕ_i.
- We want to generalize it to assign an importance score for *each subset of nodes* S.

Definition (Shapley-Taylor index (Sundararajan et al. '19)): Given a set of nodes V and a black-box model f, a k-order Shapley-Taylor interaction index for a subset S, $|S| \leq k$ is defined as:

$$\Phi_S^k = \begin{cases} \delta_S(\emptyset) & \text{if } |S| < k, \\ \frac{k}{|V|} \sum_{T \subseteq V \setminus S} \frac{1}{\binom{|V|-1}{|T|}} \delta_S f(T) & \text{if } |S| = k. \end{cases}$$

- Φ_S^k represents the importance of the subset S to the model prediction.
- If $k = 1$, the Shapley-Taylor index recovers the Shapley value.



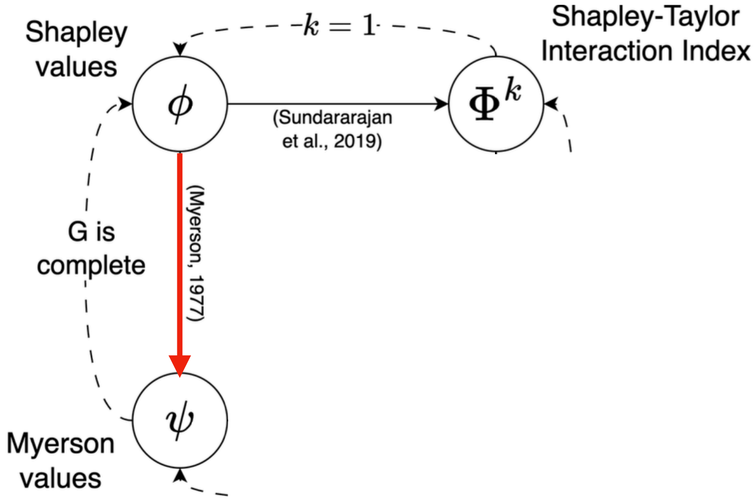
The Shapley Value to Graph Structure

- **Recall**: Shapley value assigns for *each node* i an importance score ϕ_i *without considering the connectivity* of subgraph coalitions.
- We want to generalize it to *take the graph structure* into account.

Definition (Myerson value (Myerson '77)): Given a set of nodes $G = (V, E)$ and a black-box model f, a Myerson value of a node i is computed by:

$$\psi_i = \frac{1}{|V|} \sum_{T \subseteq V \setminus i} \frac{1}{\binom{|V|-1}{|T|}} (f|_E(T \cup i) - f|_E(T)).$$

- where $f|_E(T) = \sum_{R \in \text{set of components of } T} f(R)$
  This is a communication-restricted function.
- The Myerson value is the Shapley value of the communication-restricted function.
- If the graph G is complete, then the Myerson value coincides with the Shapley value.

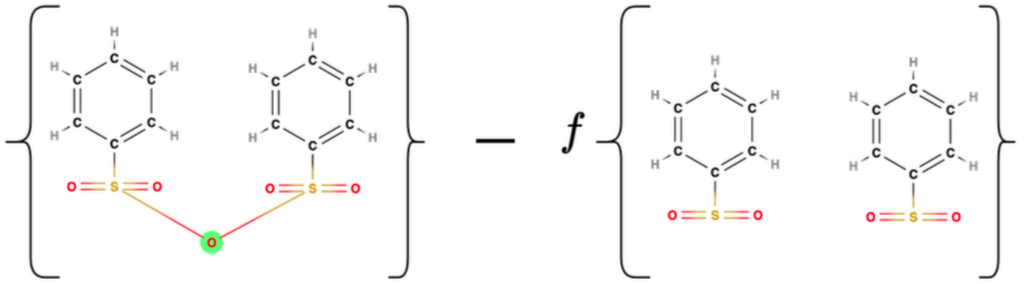


The Shapley Value to Graph Structure

- The marginal contribution of a node in the Shapley values compared to the Myerson value.
  - Therefore, the contribution of the node i in connecting components in T is also considered.

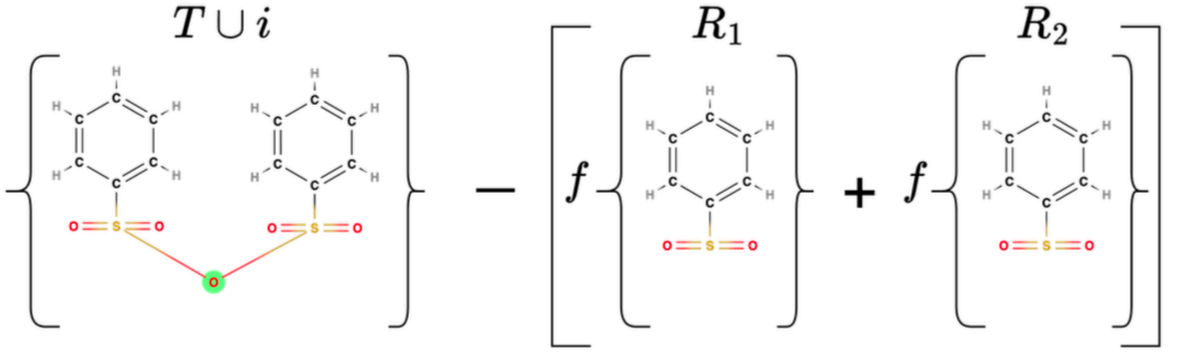
Shapley Values

$$\delta_i(T) = f\{T \cup i\} - f\{T\}$$



Myerson Values

$$\delta_i f|_E(T) = \underbrace{f\{T \cup i\}}_{f|_E(T \cup i)} - \underbrace{[f\{R_1\} + f\{R_2\}]}_{f|_E(T)}$$



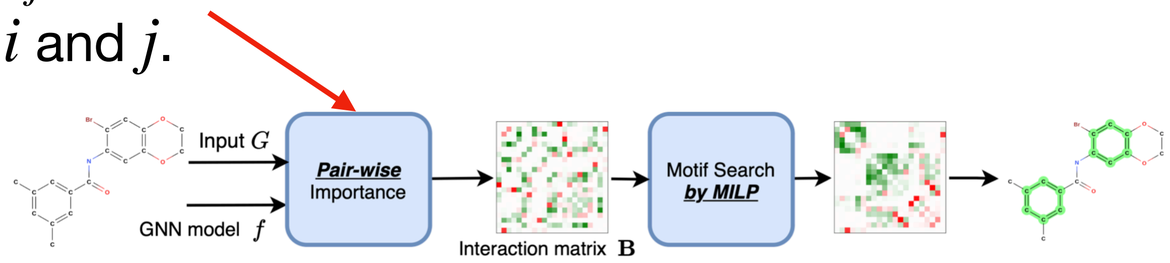
The Myerson-Taylor Interaction Index

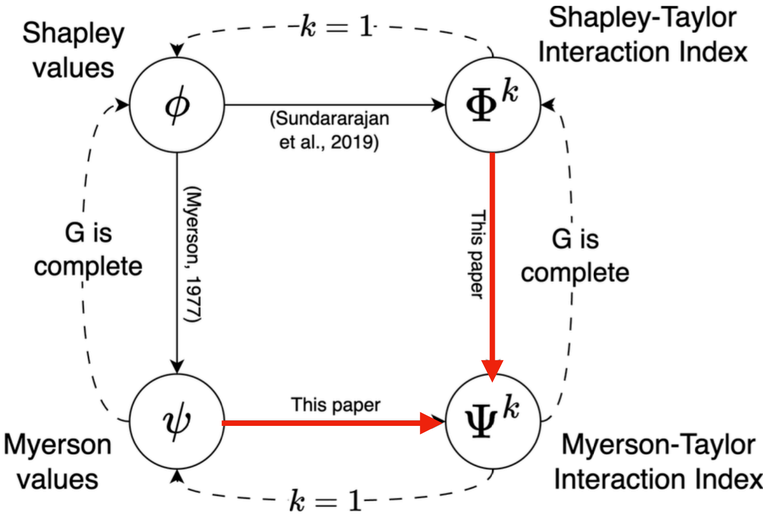
- **Recall**: Shapley value assigns for *each node* i an importance score ϕ_i *without considering the connectivity* of subgraph coalitions.
- We want to generalize it to take both *higher-order interaction* and *the graph structure* into account.

Definition (Myerson-Taylor index (ours)): Given a set of nodes $G = (V, E)$ and a black-box model f, a Myerson value of a subset S is computed by:

$$\Psi_S^k = \begin{cases} \delta_S f|_E(\emptyset) & \text{if } |S| < k, \\ \dfrac{k}{|V|} \displaystyle\sum_{T \subseteq V \setminus S} \dfrac{1}{\binom{|V|-1}{|T|}} \delta_S f|_E(T) & \text{if } |S| = k. \end{cases}$$

- Idea: Evaluate the Shapley-Taylor interaction index on the communication-restricted function.
- We can use Ψ_{ij}^2 to compute the pair-wise importance of two nodes i and j.

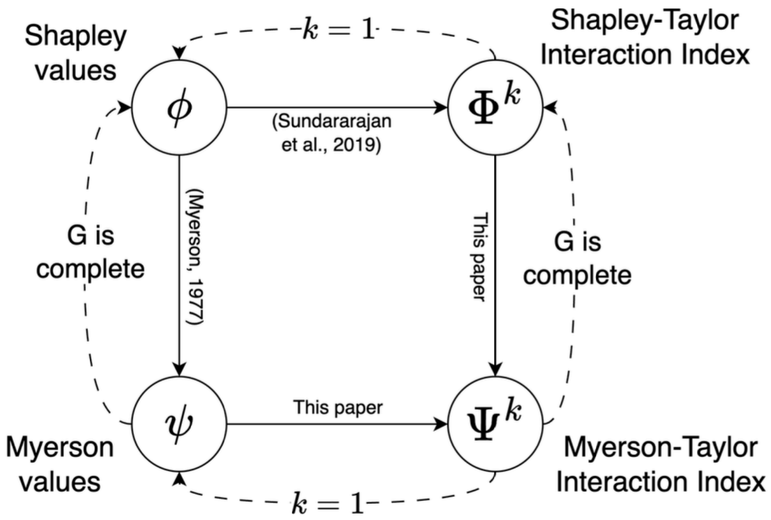




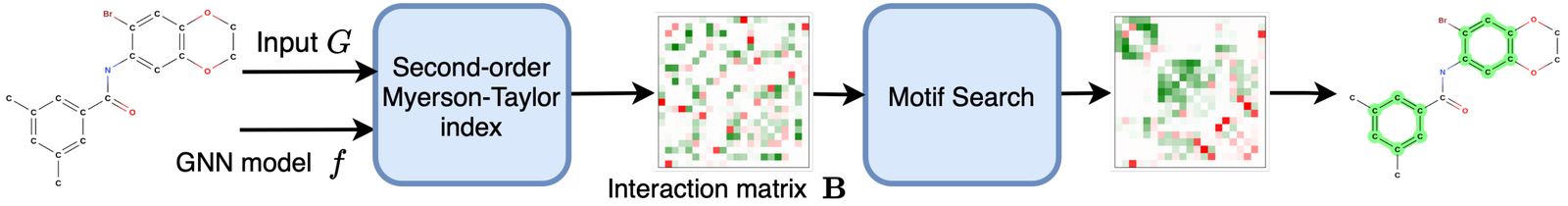
The Myerson-Taylor Interaction Index

- As the Shapley value, Myerson value, and Shapley-Taylor index are the unique allocation rules that satisfy a set of their own axioms.
- A similar result is also required for Myerson-Taylor Interaction Index

Theorem (Uniqueness): Myerson-Taylor interaction index is the unique interaction allocation index that satisfies a set of five axioms: *Linear, Restricted Null Player, Coalitional Fairness, Interaction Distribution and Component Efficiency.*



Motif Search



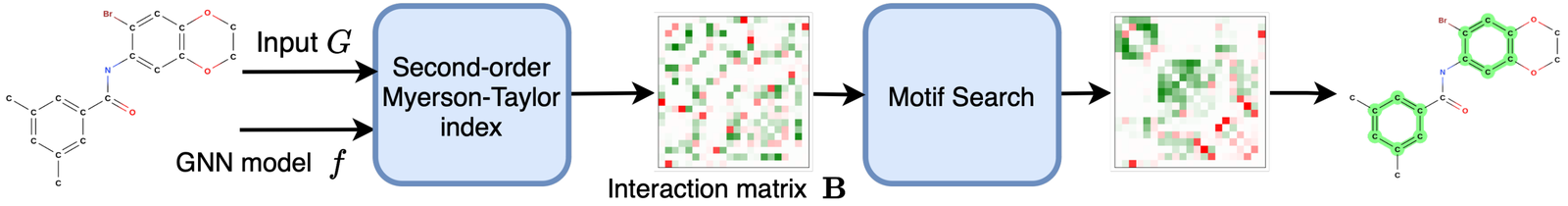
- Given an interaction matrix $\mathbf{B}$ where $\mathbf{B}_{ij}$ is the pair-wise importance score computed by the second order Myerson-Taylor interaction index.
  - $\mathbf{B}_{ij}^{+} = \max(0, \mathbf{B}_{ij})$: Only positive interaction.
  - $\mathbf{B}_{ij}^{-} = \min(0, \mathbf{B}_{ij})$: Only negative interaction.

$$H = \mathscr{H}_{m,M} = \left\{ \begin{array}{l} H \text{ contains } \underline{\textit{at most}}\ m\ \underline{\textit{disjoint}} \text{ motifs} \\ \text{Sum of nodes in all motifs } \underline{\textit{less than}}\ \mathrm{M} \\ \text{Each motif is } \underline{\textit{connected}} \end{array} \right\}$$

- The *importance score* of a group S_i

$$\mathrm{GroupAttr}(S_i) = \sum_{i,j \in S_i} \tau \mathbf{B}_{ij}^{+} + (1-\tau)\mathbf{B}_{ij}^{-}$$

Motif Search



- Given an interaction matrix $\mathbf{B}$ where $\mathbf{B}_{ij}$ is the pair-wise importance score computed by the second order Myerson-Taylor interaction index.
  - $\mathbf{B}_{ij}^{+} = \max(0, \mathbf{B}_{ij})$: Only positive interaction.
  - $\mathbf{B}_{ij}^{-} = \min(0, \mathbf{B}_{ij})$: Only negative interaction.

$$H = \mathcal{H}_{m,M} = \left\{ \begin{array}{l} H \text{ contains } \underline{\textit{at most}}\ m\ \underline{\textit{disjoint}} \text{ motifs} \\ \text{Sum of nodes in all motifs } \underline{\textit{less than}}\ \text{M} \\ \text{Each motif is } \underline{\textit{connected}} \end{array} \right\}$$

- The *importance score* of a group S_i

$$\text{GroupAttr}(S_i) = \sum_{i,j \in S_i} \tau \mathbf{B}_{ij}^{+} + (1-\tau)\mathbf{B}_{ij}^{-}$$

- We want to find a subgraph $H = (S, E_S) \in \mathcal{H}_{m,M}$, $S = \{S_1, S_2, \ldots, S_m\}$ that is ***most influential*** to the model prediction by solving the following optimization problem:

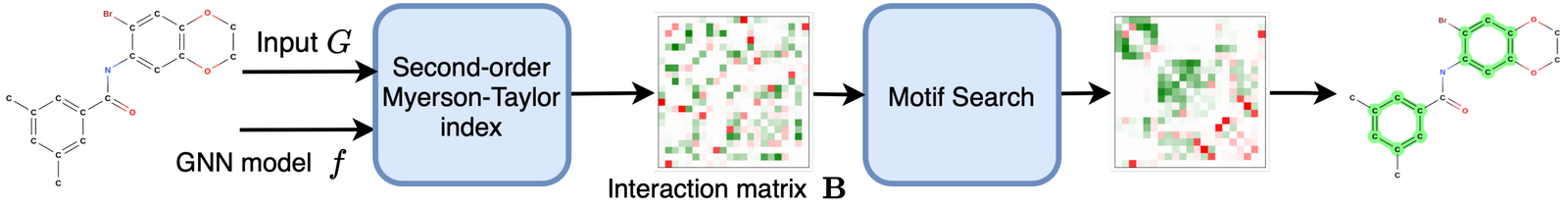
Motif search

$$\max_{(S_1, S_2, \ldots, S_m) \in \mathcal{H}_{m,M}} \sum_{l=1}^{m} \left| \text{GroupAttr}(S_l) \right|.$$

The absolute operator is to ensure both positive and negative interactions are considered.

- *Ideally, nodes in the same motifs should strongly interact with each other, while interactions of nodes from different motifs should be negligible*

Motif Search



- Given an interaction matrix $\mathbf{B}$ where $\mathbf{B}_{ij}$ is the pair-wise importance score computed by the second order Myerson-Taylor interaction index.
  - $\mathbf{B}_{ij}^{+} = \max(0, \mathbf{B}_{ij})$: Only positive interaction.
  - $\mathbf{B}_{ij}^{-} = \min(0, \mathbf{B}_{ij})$: Only negative interaction.

$$H = \mathcal{H}_{m,M} = \left\{ \begin{array}{l} H \text{ contains } \underline{\textit{at most}}\ m\ \underline{\textit{disjoint}} \text{ motifs} \\ \text{Sum of nodes in all motifs } \underline{\textit{less than}}\ \text{M} \\ \text{Each motif is } \underline{\textit{connected}} \end{array} \right\}$$

- The *importance score* of a group S_i

$$\text{GroupAttr}(S_i) = \sum_{i,j \in S_i} \tau \mathbf{B}_{ij}^{+} + (1-\tau)\mathbf{B}_{ij}^{-}$$

- We want to find a subgraph $H = (S, E_S) \in \mathcal{H}_{m,M}$, $S = \{S_1, S_2, \ldots, S_m\}$ that is ***most influential*** to the model prediction by solving the following optimization problem:

Motif search

$$\max_{(S_1, S_2, \ldots, S_m) \in \mathcal{H}_{m,M}} \sum_{l=1}^{m} \left| \text{GroupAttr}(S_l) \right|.$$

The absolute operator is to ensure both positive and negative interactions are considered.

- ***This problem can be solved by an off-the-shelf Mixed Integer Linear Programming (MILP) solvers.***

Experiments

- **Datasets:** ten datasets from
  - *Synthetic datasets (4)*:
    - BA-2motifs, BA-HouseGrid, SPMotif, BA-HouseAndGrid, BA-HouseOrGrid
  - *Molecular property prediction (2)*:
    - Mutagenic, Benzene
  - *Image classification (1)*:
    - MNIST75SP
  - *Sentiment classification (2):*
    - GraphSST2, Twitter

Has ground truth explanation

No ground truth explanation

- **Model:**
  - GCN, GIN, GAT
- **Baselines:**
  - White-box explainer: Grad-CAM
  - Perturbation-based Explainers: GNNExplainer, PGExplainer, Refine, MatchExplainer
    - Cooperative game-based Explainers: SubgraphX, GStarX, SAME
  - **Ours: MAGE**

Experiments

- **Metrics:**
  - For datasets *with* ground truth explanations:
    - *F1 score:* formulate as binary classification and evaluate the performance by measuring the overlap in highlighted nodes.
    - *Area under the curve (AUC):* Use for datasets with single motif only.
    - *Adjusted Mutual Information (AMI)*: usually used in clustering tasks. we use to measure the explainer's ability to identify multiple motifs.
  - For datasets *without* ground truth explanations:
    - *Fidelity (Fid)*: faithfulness of the explanations to the model's prediction by accessing the drop when including or excluding the explanation graph.

$$\begin{aligned}\mathrm{Fid}^{+}(S) &= f(V) - f(V \setminus S),\\ \mathrm{Fid}^{-}(S) &= f(V) - f(S),\\ \mathrm{Fid}(S) &= \mathrm{Fid}^{+}(S) - \mathrm{Fid}^{-}(S).\end{aligned}$$

    - *Robust Fidelity (Fid_{α})*: a metric proposed to reduce the OOD problem of Fid

Experiments - Main Result

Datasets with ground truth explanations

Table 1. Single Motif GCN & Multiple Motifs GIN Experiment. On average, MAGE achieves a 59.29% improvement in F1 score on single motif datasets, a 28.11% improvement in AMI score on multi-motif datasets, and a 12.61% improvement in AUC score across all datasets.

Method	Single Motif - GCN								Multiple Motifs - GIN							
	BA-2Motifs		BA-HouseGrid		SPMotif		MNIST75SP		BA-HouseAndGrid		BA-HouseOrGrid		Mutagenic		Benzene	
	F1↑	*AUC*↑	*F1*↑	*AUC*↑	*F1*↑	*AUC*↑	*F1*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑
GradCAM	0.634	0.753	0.459	0.485	0.538	0.675	0.193	0.492	0.825	**0.994**	0.931	**0.997**	-0.001	0.514	0.789	**0.964**
GNNExplainer	0.222	0.44	0.297	0.546	0.185	0.465	0.22	0.531	0.275	0.533	0.148	0.532	0.228	0.679	0.178	0.487
PGExplainer	0.042	0.498	0.057	0.434	0.066	0.097	0.236	0.607	0.1	0.088	0.17	0.002	0.099	0.573	0.186	0.042
Refine	0.144	0.474	0.191	0.398	0.164	0.508	0.153	0.459	0.254	0.429	0.123	0.488	0.21	0.623	0.207	0.529
MatchExplainer	0.586	0.706	0.587	0.712	0.19	0.513	0.162	0.483	0.537	0.810	0.521	0.788	0.216	0.576	0.318	0.545
SubgraphX	0.606	0.7	0.168	0.51	0.169	0.523	0.08	0.501	0.494	0.697	0.526	0.767	0.595	0.784	0.731	0.832
GStarX	0.153	0.491	0.267	0.544	0.203	0.498	0.280	0.517	0.203	0.494	0.13	0.484	-0.018	0.462	0.122	0.505
SAME	0.587	0.692	0.474	0.693	0.37	0.582	0.272	0.531	0.497	0.681	0.606	0.796	0.48	0.709	0.617	0.791
MAGE (ours)	**0.915**	**0.934**	**0.832**	**0.849**	**0.635**	**0.757**	**0.634**	**0.716**	**0.998**	**0.999**	**0.998**	**0.999**	**1.000**	**1.000**	**0.917**	**0.959**
Improvement (%)	44.32	24.03	41.73	19.24	18.02	12.14	133.09	17.96	20.96	0	7.20	0	68.07	27.55	16.22	0

MAGE outperforms most of the current explainers in most of the settings.

Experiments - Main Result

Datasets with*out* ground truth explanations

Table 2. Fidelity evaluation on sentiment classification and GCN.

Method	GraphSST2		Twitter	
	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$
GradCAM	<u>0.169</u>	0.253	0.268	0.363
SubgraphX	0.141	0.225	0.238	0.32
GStarX	<u>0.161</u>	<u>0.273</u>	0.276	<u>0.425</u>
SAME	0.141	0.216	<u>0.293</u>	0.397
MAGE (ours)	**0.200**	**0.337**	**0.317**	**0.471**

MAGE outperforms most of the current explainers in most of the settings.

Experiments - Qualitative Comparison

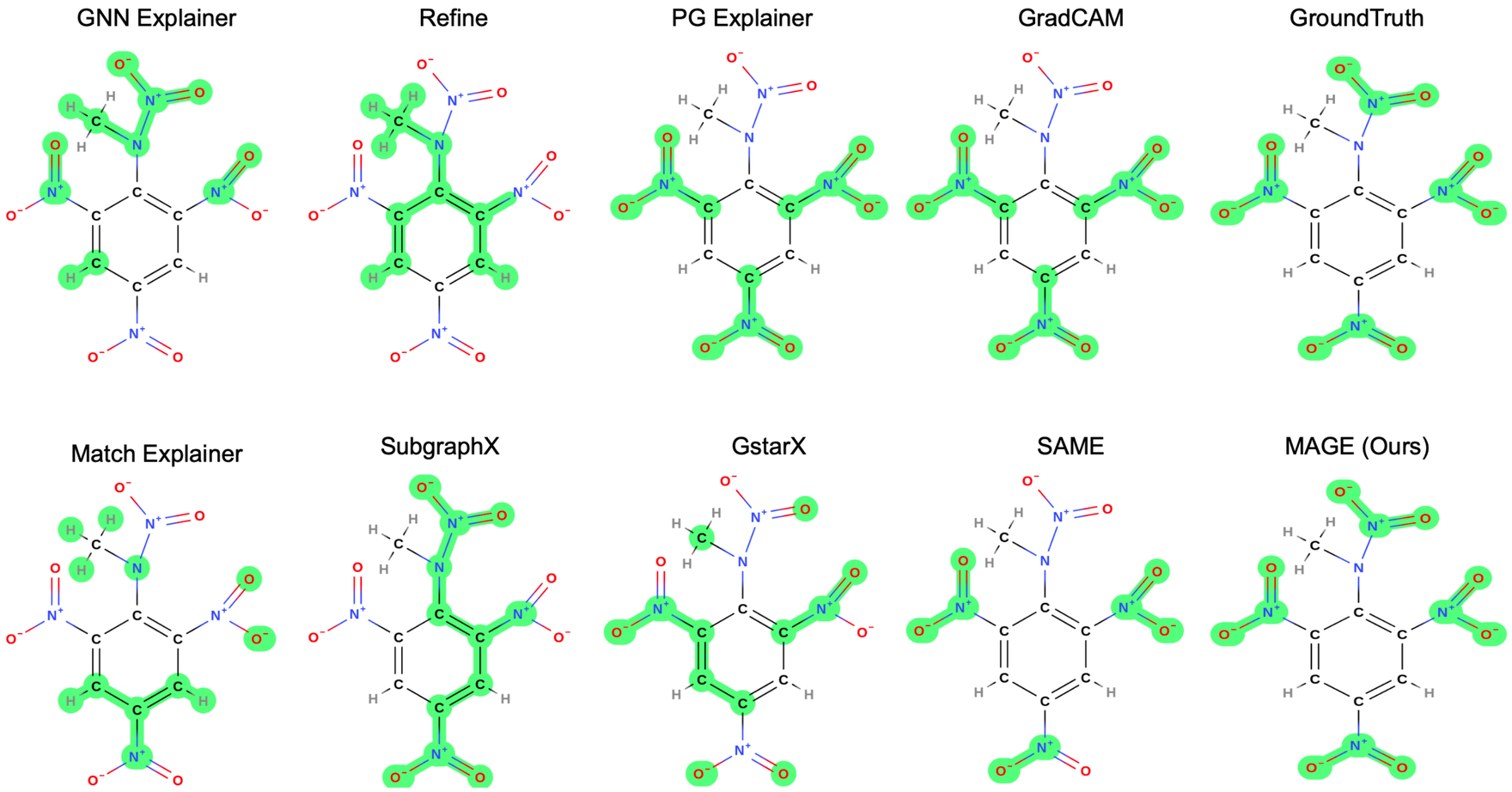


Figure 8. Explanations of competing methods on a molecular graph from Mutagenic dataset.

Experiments - Qualitative Comparison

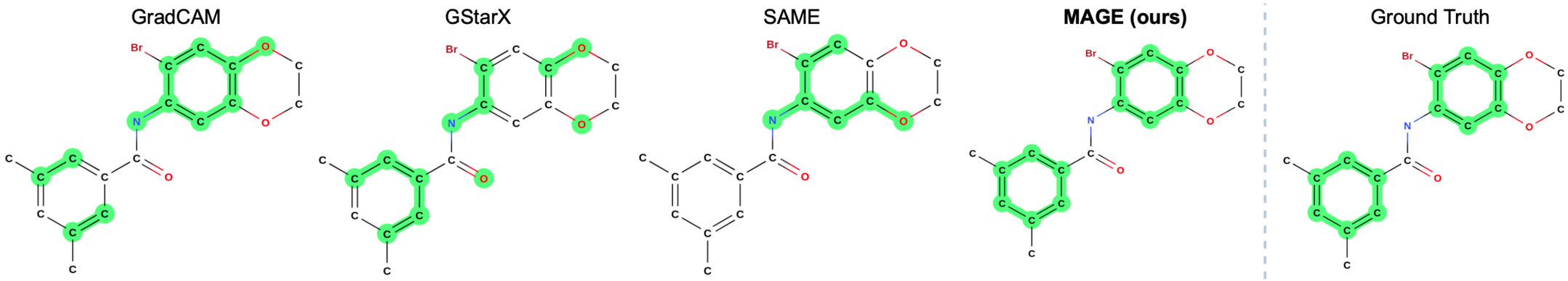


Figure 1. Molecule C17H16BrNO3 input is predicted in class ‘have benzene ring’ by GNN. Our MAGE multi-motif explanations correctly identify the two benzene rings; while competing methods such as GradCAM, GStarX, and SAME fail.

Experiments - Qualitative Comparison

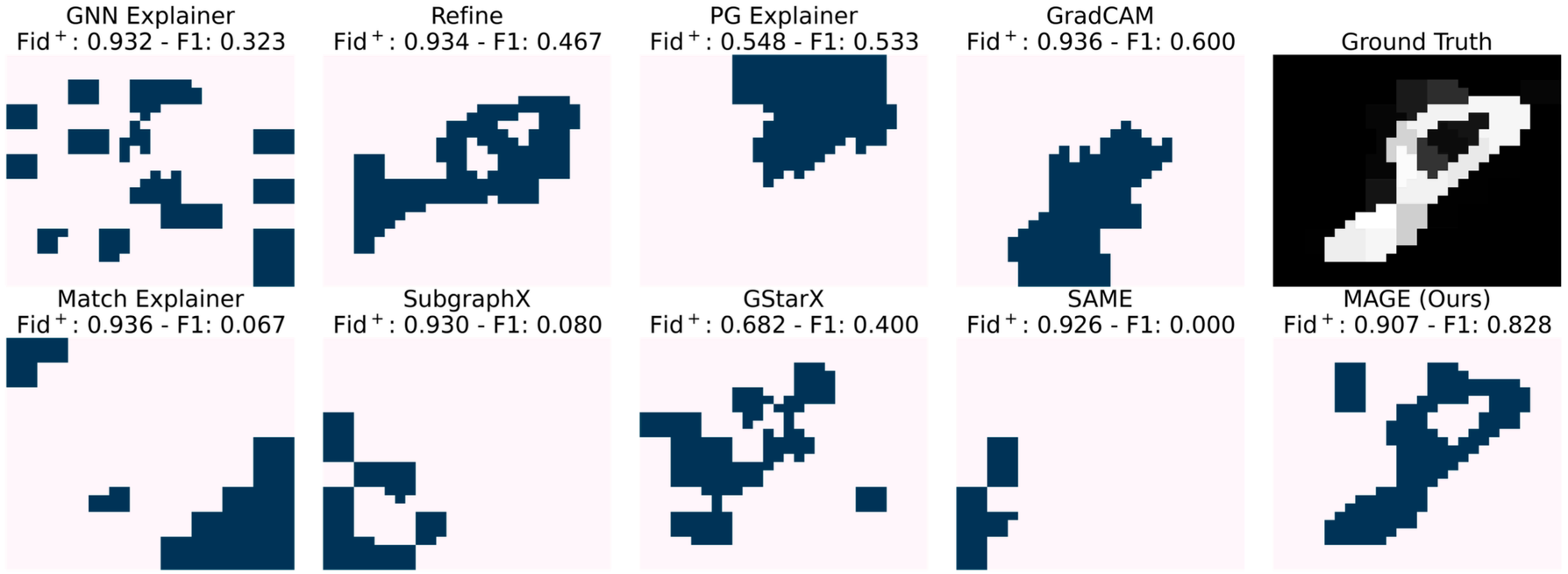
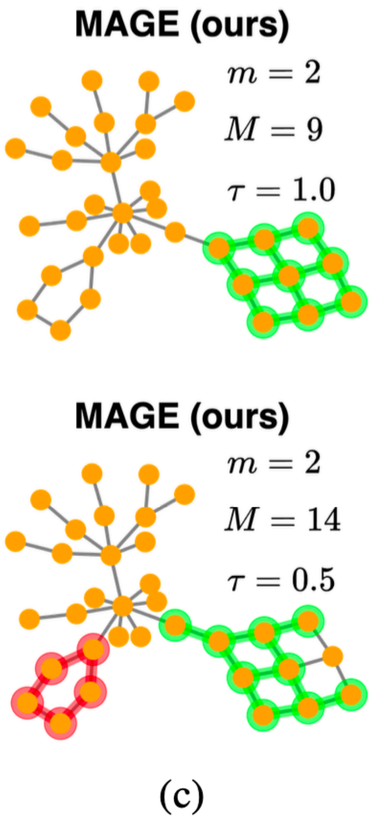


Figure 6. This example visualizes the explanation for GCN model of MAGE against competing baselines on MNIST75SP. Despite achieving high fidelity (Fid^+) scores, the explanations of baselines are not meaningful. Meanwhile, only MAGE can generate an explanation that aligns with pixels that describe number '8'

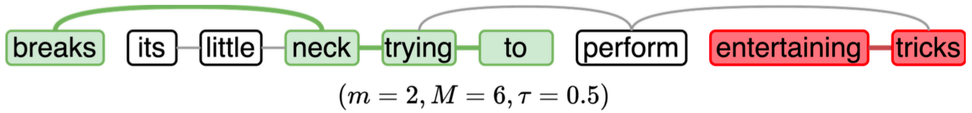
Only our method can provide a reasonable explanation for GNN on MNIST75SP datasets.

Note that other methods can provide explanations with a very *high fidelity*. However, we argue that these explanations are just o*ut-of-distribution samples, not representative* of the class.

Experiments - Qualitative Comparison



(c)



$(m = 2, M = 6, \tau = 0.5)$

(a) Model prediction: Negative sentiment. MAGE highlights the main negative verb phrase 'breaks neck', which contributes to the overall negative sentiment, and the phrase 'entertaining tricks', which shows a slightly positive sentiment.

$(m = 2, M = 10, \tau = 0.5)$

(b) Model prediction: positive sentiment. MAGE highlights the main adjective 'heart-warming' which contributes to the overall positive sentiment, and two minor adjectives 'corny' and 'schmaltzy' which display some negative sentiment.

Figure 7. MAGE can highlight text subgraphs with contradicting sentiments in GraphSST2 Dataset.

By calibrating the hyperparameters, MAGE can highlight both motifs that negatively and positively impact the model prediction.

Ablation study

We ablate the Myerson-Taylor and the Shapley-Taylor to see if we can gain the performance from the graph structure.

We also ablate the connectivity constraint in the optimization problem in Motif Search.

Table 3. Ablating Shapley-Taylor (Φ^2) and Myerson-Taylor (Ψ^2) indices and connectivity constraints in the problem (3).

Method	BA-2Motifs		BA-HouseGrid	
	F1 ↑	*AUC*↑	*F1* ↑	*AUC*↑
MAGE (Φ^2) w/o connectivity	0.699	0.773	0.634	0.735
MAGE (Φ^2) w/ connectivity	0.709	0.787	0.636	0.734
MAGE (Ψ^2) w/o connectivity	**0.854**	**0.885**	0.819	0.838
MAGE (Ψ^2) w/ connectivity	**0.858**	**0.890**	**0.832**	**0.849**

- Using Myerson-Taylor improves the results significantly compared to the Shapley-Taylor.
- The connectivity constraint in the Motif Search module also helps improve the performance and helps humans understand the explanations more easily.

Running time

Table 10. The average running time of competing methods on evaluated datasets (second/sample).

Method	BA-2Motifs	BA-HouseGrid	SPMotif	MNIST75SP	BA-HouseAndGrid	BA-HouseOrGrid	Mutagenic	Benzene
GNNExplainer	2.27	2.30	3.50	15.83	2.62	3.41	3.78	4.91
SubgraphX	66.60	54.40	74.14	823.55	70.18	81.41	63.63	4.40
GStarX	19.63	15.53	21.51	35.75	29.47	17.06	16.43	20.02
SAME	11.61	44.42	19.03	142.11	90.73	41.27	6.12	7.14
MAGE (ours)	5.16	6.18	14.61	44.28	17.74	9.38	4.89	3.63

- It’s faster than other cooperative game-based methods.

Conclusion

- We proposed a novel interaction index, namely the Myerson-Taylor interaction index, that takes both high-order interactions and structural information in the input into account when computing the importance scores for subgraphs.
  - We proved that this interaction index is a unique allocation rule that satisfies a set of five axioms.

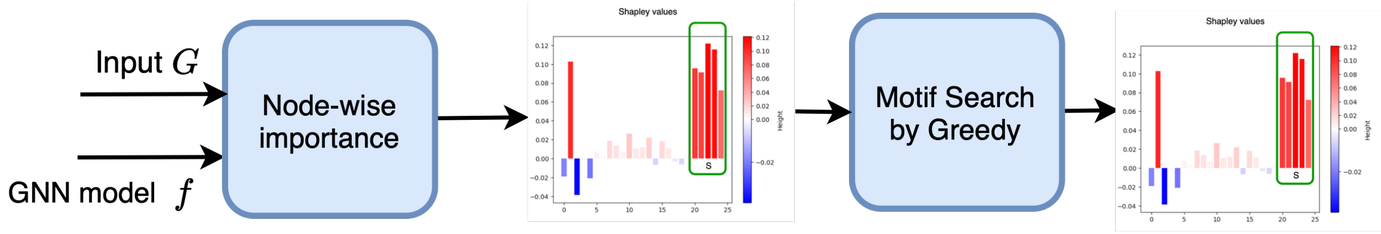
- We proposed a framework leveraging the second-order Myerson-Taylor interaction index to compute the pair-wise importance of nodes and then extract explanations by solving a MILP.

- Experiments show the effectiveness of the proposed methods.

Comparing to related works

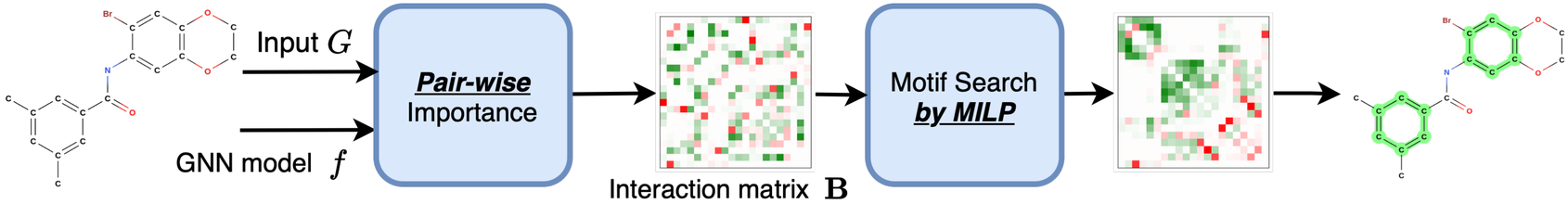
Graphsvx
GStarX

- Fast, but not accurate



MAGE

- A balance between efficiency and accuracy.



SubgraphX

SAME

- Accurate but slow