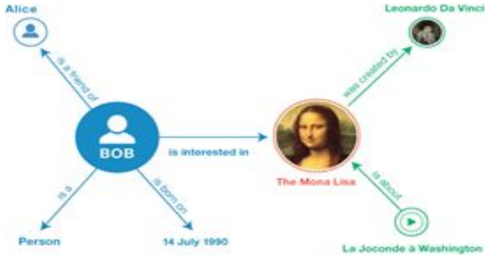
Bagel:A Benchmark for Assessing Graph Neural Network Explanations

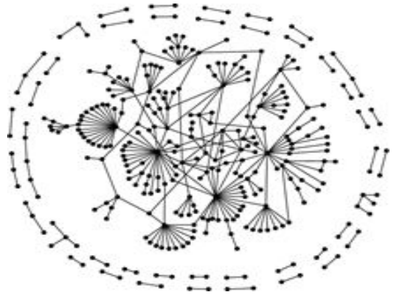
Mandeep Rathee, Thorben Funke, Avishek Anand, and Megha Khosla

Graphs Are Everywhere

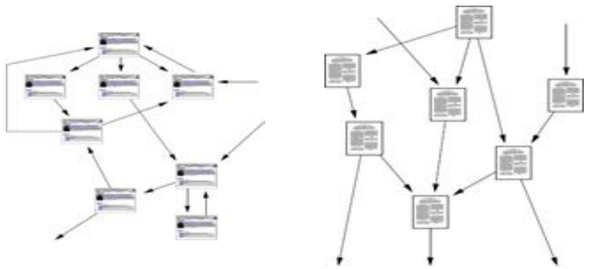
Social Networks



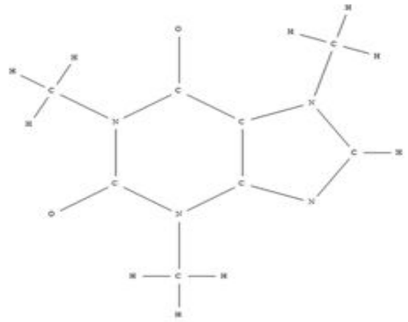
Knowledge Graphs



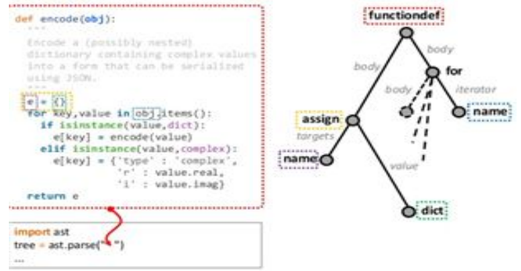
Biological Networks



Citation Networks

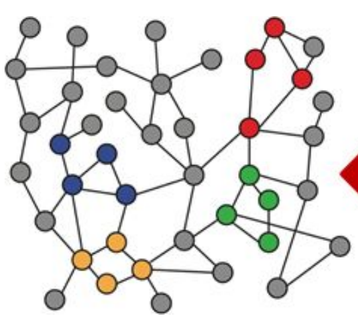


Molecules

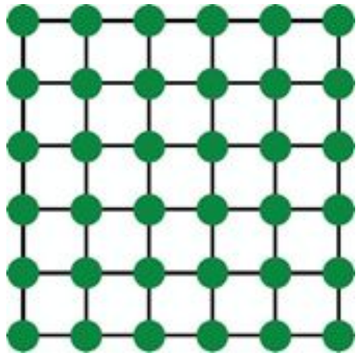


Code

Graphs Are Complex



VS.



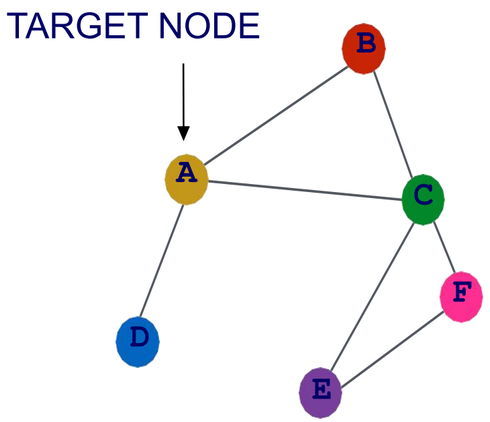
Random Network

Images

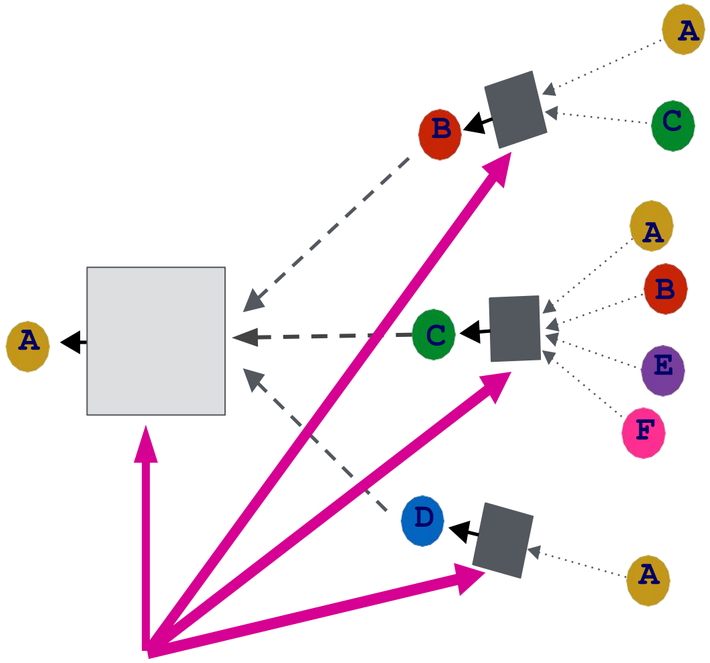
Text/Audio

Tabular

Graph Neural Networks (GNNs)



INPUT GRAPH



Neural networks

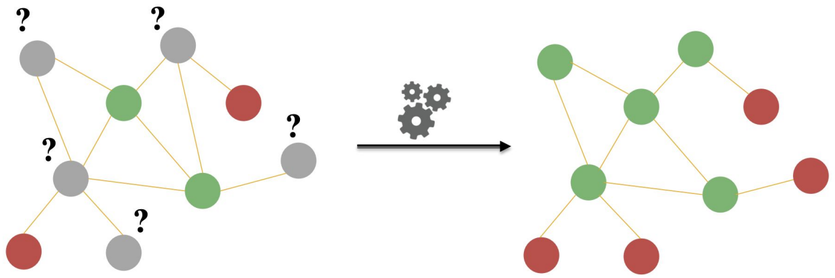
Each node defines a computation graph

- Each edge in this graph is a transformation/aggregation function

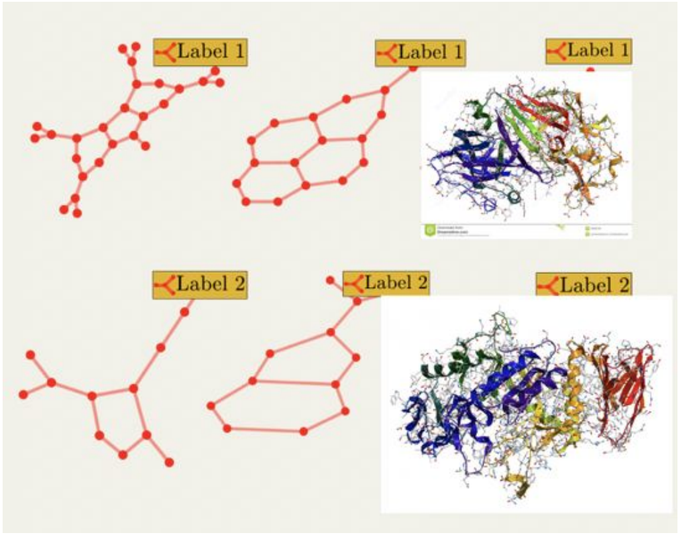
e.g., GCN and GAT

Scarselli et al. 2005. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*.

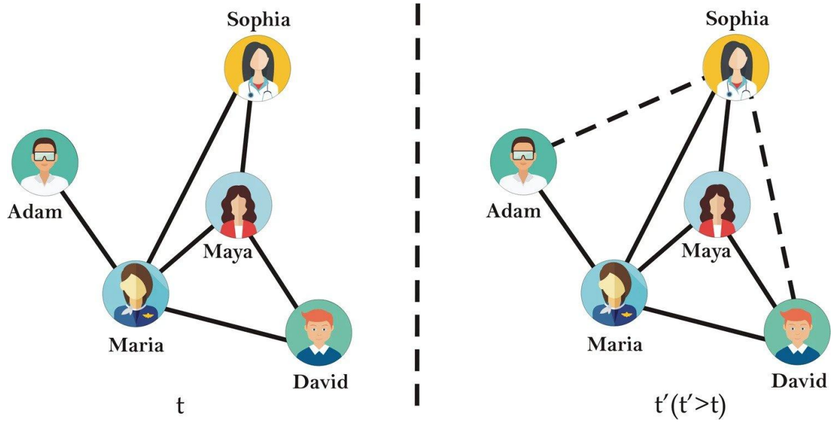
Tasks on Graph-Structured Data



Node Classification



Graph Classification



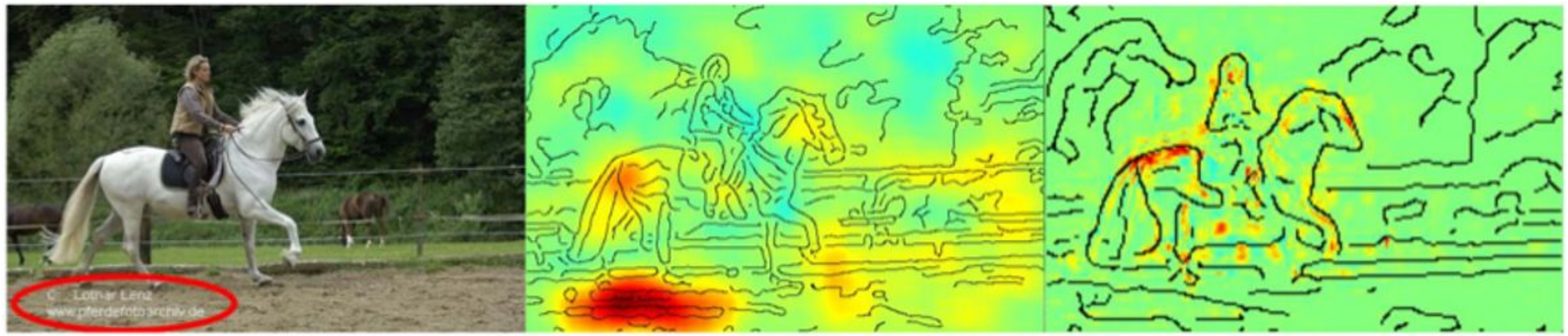
Link Prediction

Applications in

- Health
- Recommendation
- Finance

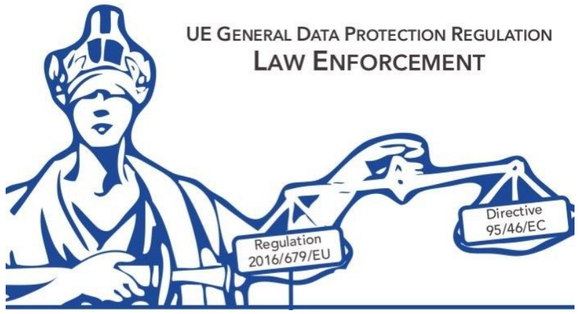
Why Explainability?

Given a machine learning model and machine learning task, we are interested in finding the rationales behind the prediction.

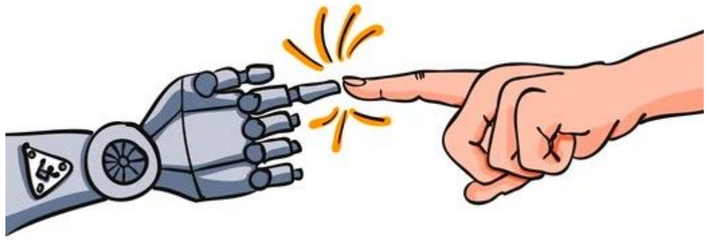


Right for the Right reasons

Utilize insights to improve models

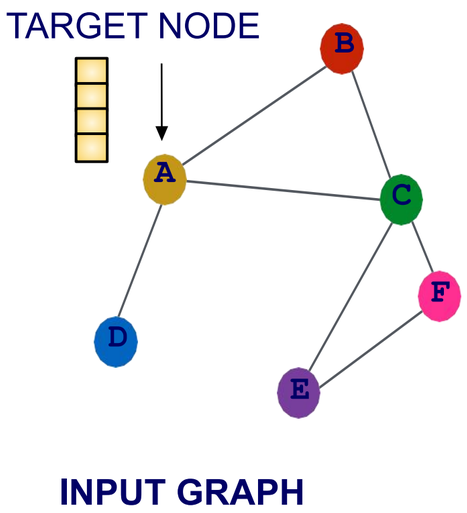


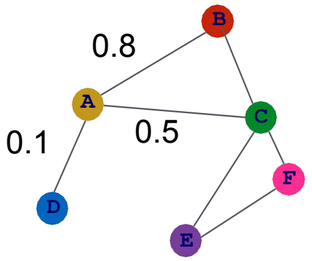
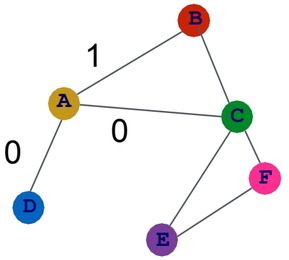
Legal recourse



Improve trust

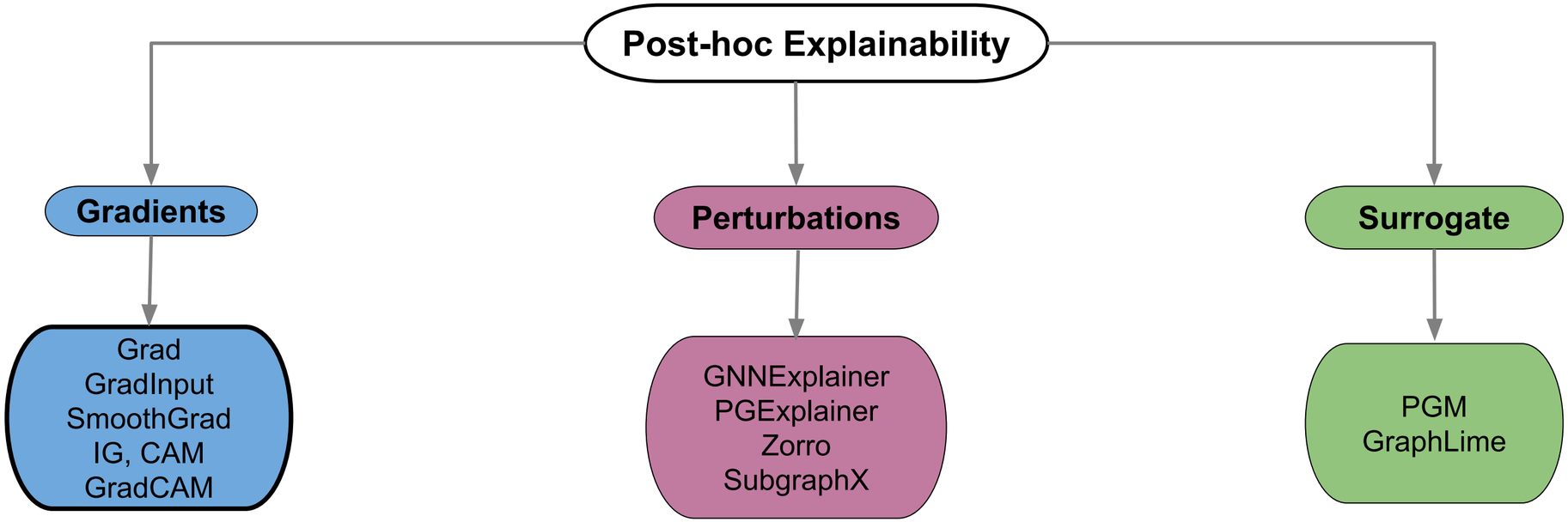
Explainability in GNNs



	Hard Mask	Soft Mask
Node Level	1 1 0 0 1 1	0.9 0.6 0.2 0.1 0.7 0.5
Feature Level	1 0 1 0	0.5 0.1 0.8 0
Edge Level		

GNN Explanation Methods

Given an input graph G(A,X), a trained GNN Φ, and its prediction



gradient of the prediction with respect to the input as the importance score (mask) for input nodes/edges/features

provide a minimal subgraph of the original graph that is deemed to be important of the prediction

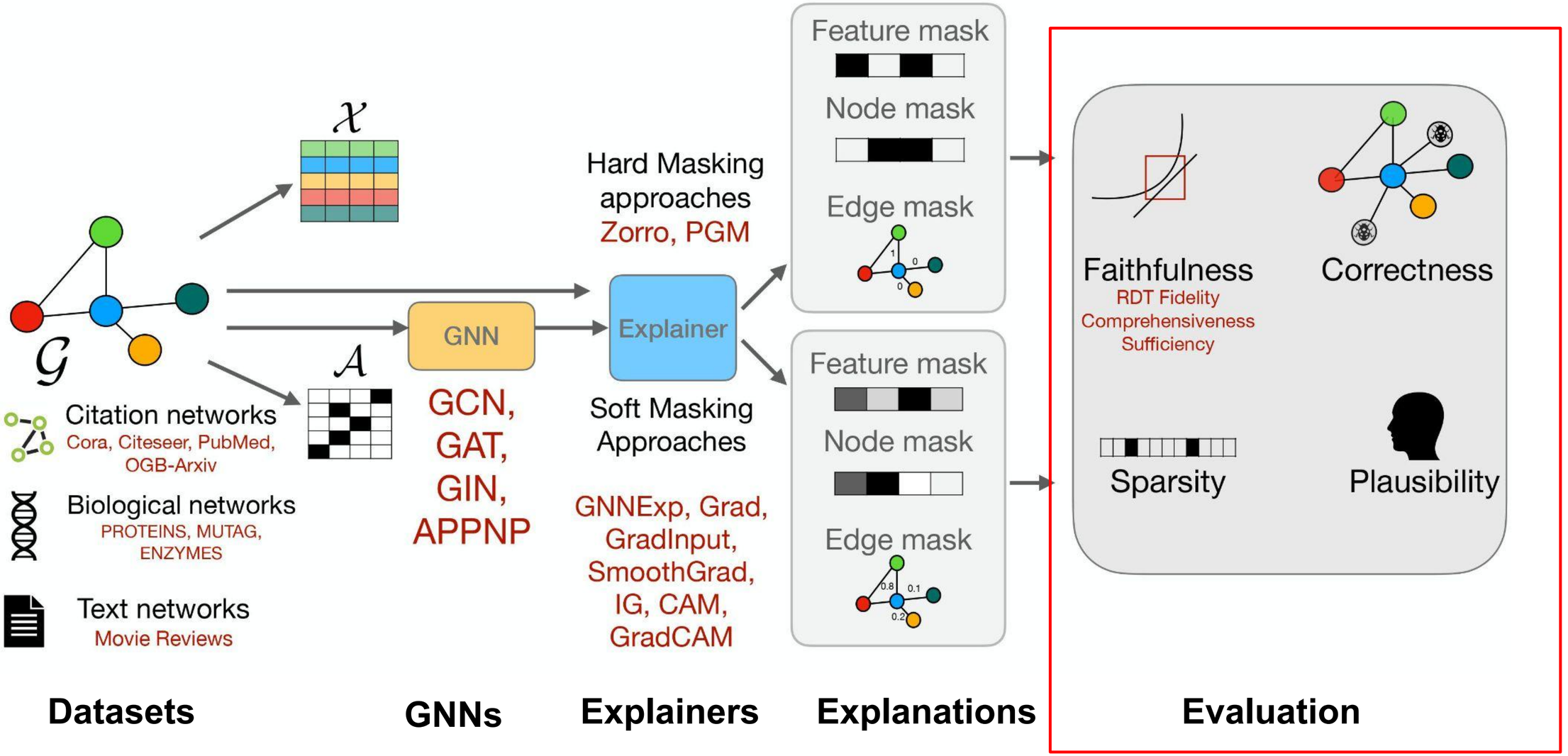
sample a local dataset to represent the relationships around the target node. Usually interpretable by design are applied to fit the sampled local dataset.

But How to judge if an explanation is good or not?

“a good explanation should be relatively faithful to how the model works, understandable to the receiver, and useful for the receiver’s end-goals.”

- AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap, (Vera Liao et al.,2023)

Bagel: A Benchmark for Assessing Graph Neural Network Explanations



Faithfulness

How well does the explanation approximate the model's behavior?

The **faithfulness** or **fidelity** of an explanation ***S*** can be defined as prediction or accuracy change by removing important nodes/edges/node features.

$$F(S) = \Phi(\quad) - \Phi(\quad \odot \begin{matrix}1\\0\\1\\0\end{matrix})$$

Drawbacks:

- We do not want to replace the unimportant features with 0, rather its value should not matter.
- Out of distribution features.
- Special case of dense features.

Explainability methods for graph convolutional neural networks (Pope et al., 2019).

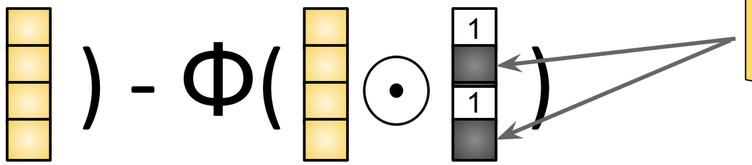
Faithfulness

How well does the explanation approximate the model's behavior?

The **faithfulness** or **fidelity** of an explanation ***S*** can be defined as prediction or accuracy change by removing important nodes/edges/node features.

RDT-Fidelity

$$F(S) = \Phi(\;) - \Phi(\; \odot \;)$$



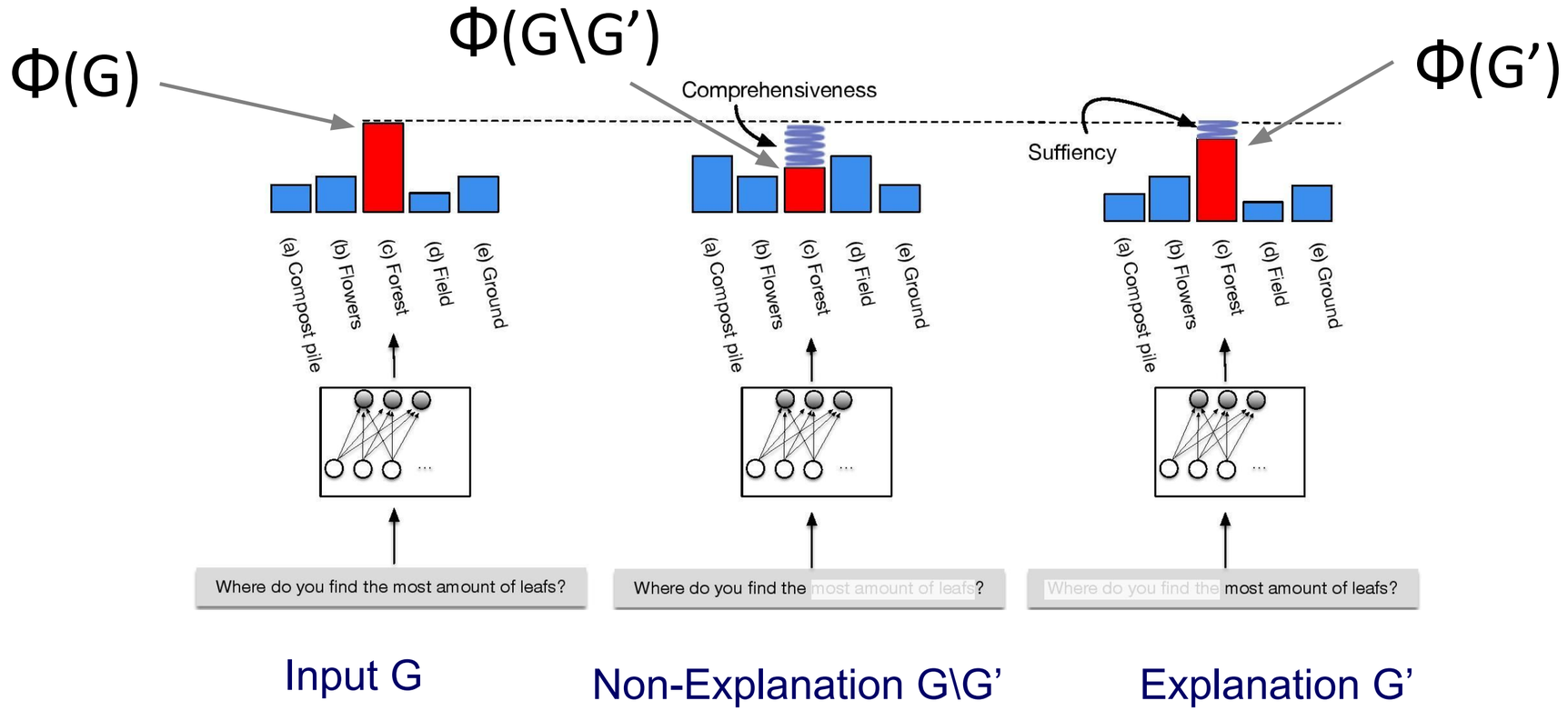
sample from the underline training data distribution

Key Idea: Perturb the unimportant features.

Special case of dense features.

Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks (Funke et al., 2022)

Faithfulness



Sufficiency = Φ(G) - Φ(G'), if the extracted nodes/edges are sufficient to come up the original prediction.

Comprehensiveness = Φ(G) - Φ(G\G'), if all nodes/edges in the graph needed to make a prediction were selected.

ERASER: DeYoung et al. 2020

Sparsity

But the full input is also a faithful explanation. Are the explanations non-trivial?

Sparsity for hard masks (binary explanations) = Selection size/total

Drawbacks:

- does not work for soft masks.

Explainability methods for graph convolutional neural networks (Pope et al., 2019).

Sparsity

But the full input is also a faithful explanation. Are the explanations non-trivial?

Entropy based Sparsity

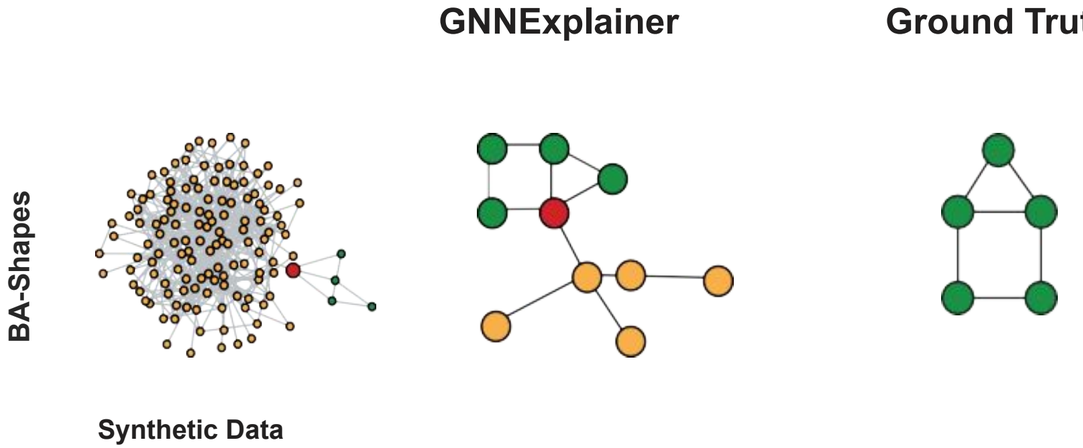
$$H(p) = -\sum_{x} p(x) \log p(x)$$

Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks (Funke et al., 2022)

Correctness

Explanation Accuracy

How well does the explanation agree with the ground truth?



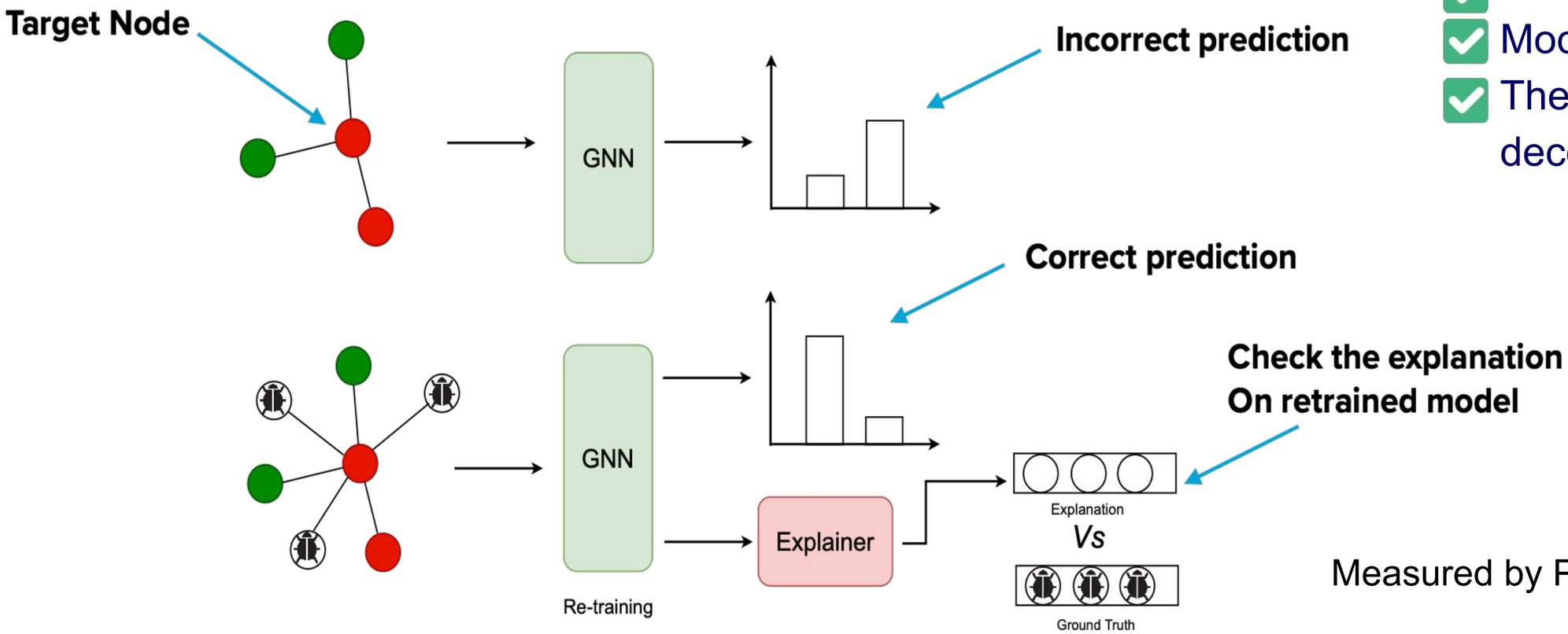
Drawbacks:

- We do not know if the model really used the ground truth for prediction?
- Ground truth is not available for most of the graph datasets.

GNNExplainer, Ying et al., 2019

Correctness

Introduce correlations (decoys) in the training data which can change the decision on a node/graph. Then check if explanation discovers the added correlations.



- Model uses the decoys.
- Model learns the decoys.
- The explainer picks the decoys.

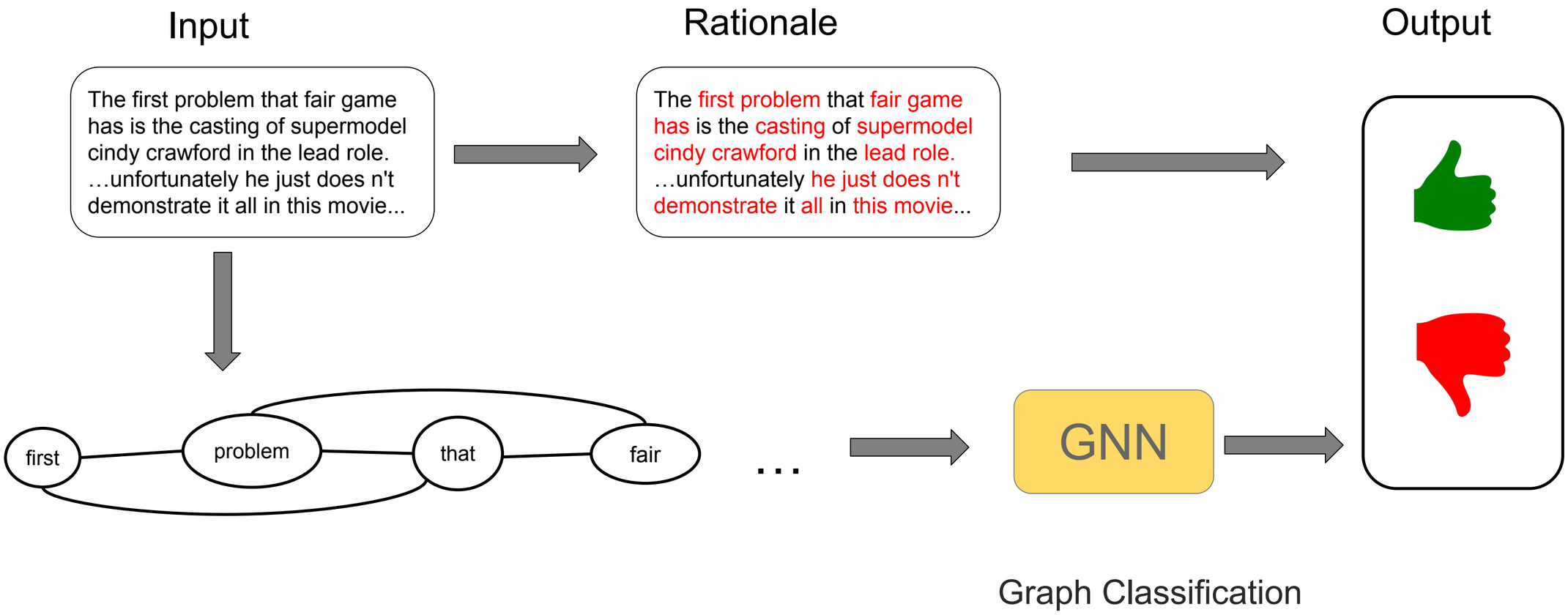
Measured by Precision and Recall.

Plausibility

An explanation is considered plausible if it is coherent with human reasoning and understanding.

Graph neural network explanations often disregard plausibility.

Dataset for Plausibility

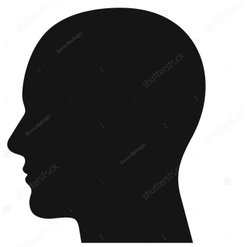


ERASER: DeYoung et al. 2020

Plausibility

Human Rationales	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
GNNExp	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
Grad	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
CAM	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

Explainer **Vs.**



Measured by
- Token level F1.
- AUPRC(only for soft masks).

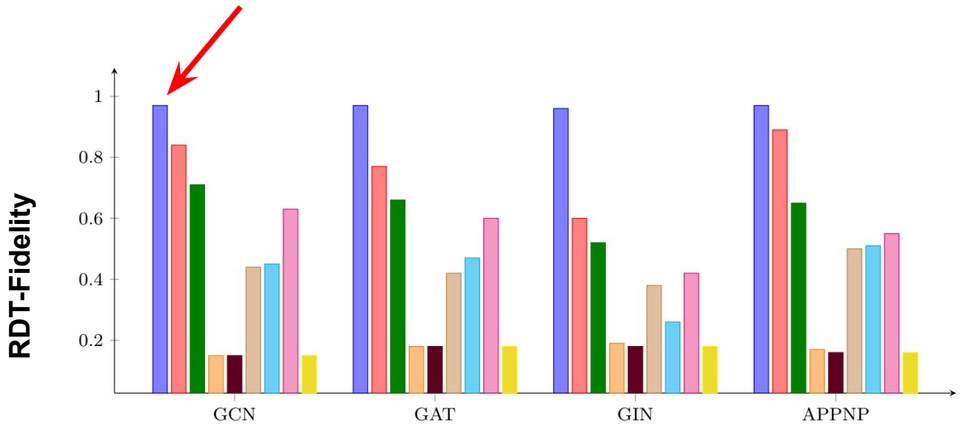
Results

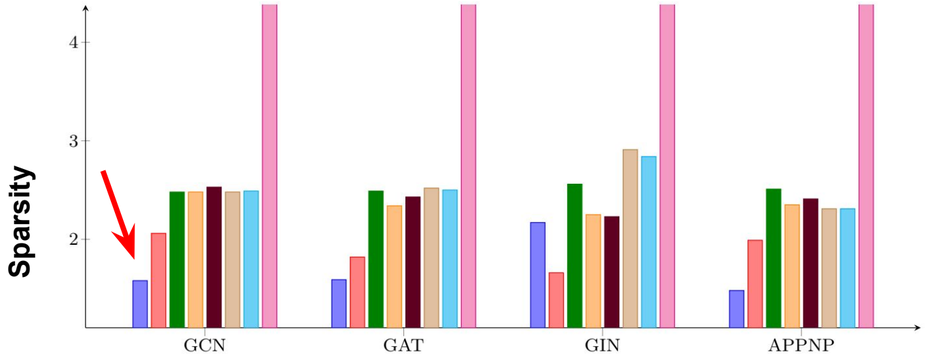
We did ton of experiments in our Bagel benchmark and can be extended easily.

Task	Dataset	Metric				
		Faithfulness		Sparsity	Correctness	Plausibility
		RDT-Fidelity	Suff. & Comp.			
Node	CORA	✓	✗	✓	✓	✗
	CITESEER	✓	✗	✓	✓	✗
	PUBMED	✓	✗	✓	✗	✗
	OGBN-ARXIV	✓	✗	✓	✗	✗
	Synthetic	✓	✗	✓	✓	✓
Graph	Movie Reviews	✗	✓	✗	✗	✓
	MUTAG	✓	✓	✓	✗	✗
	PROTEINS	✓	✓	✓	✗	✗
	ENZYMES	✓	✓	✓	✗	✗

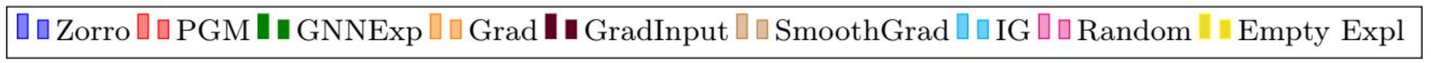
9 explainers
9 Datasets
4 GNNs

Results

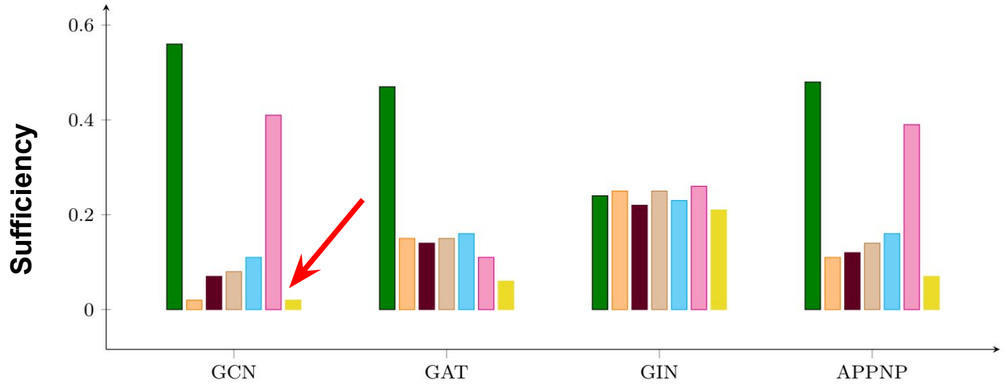


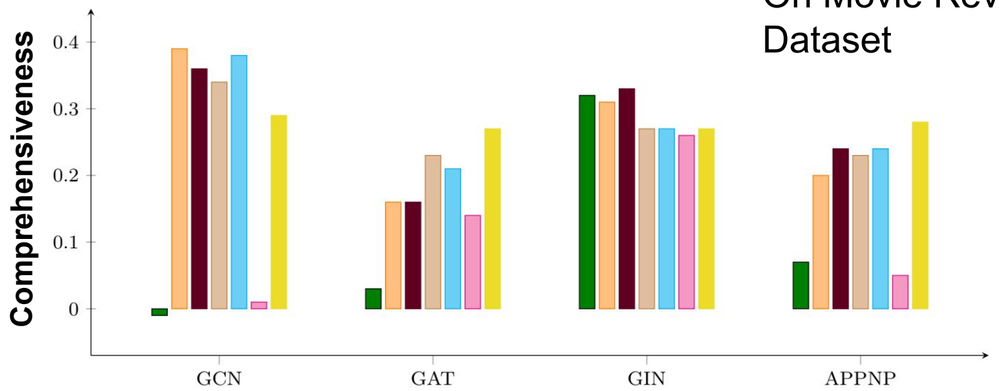


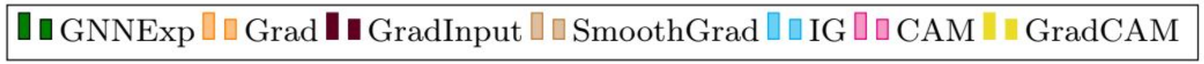
On Cora Dataset



On Movie Reviews Dataset







Results

Correctness

Mask	Methods	Cora															
		GCN				GAT				GIN				APPNP			
		P@k	R@k	F1	$\|\mathcal{S}\|$	P@k	R@k	F1	$\|\mathcal{S}\|$	P@k	R@k	F1	$\|\mathcal{S}\|$	P@k	R@k	F1	$\|\mathcal{S}\|$
Hard	Zorro	0.19	0.80	0.30	45	0.25	0.83	0.37	40	0.26	0.45	0.27	33	0.22	0.79	0.33	38
	PGM	0.11	0.22	0.15	20	0.18	0.36	0.24	20	0.18	0.36	0.25	20	0.19	0.38	0.25	20
Soft	GNNExp	**0.42**	**0.84**	**0.56**	20	**0.44**	**0.88**	**0.59**	20	**0.50**	**1.00**	**0.67**	20	**0.34**	**0.67**	**0.58**	20
	Grad	0.23	0.46	0.31	20	0.29	0.58	0.39	20	0.30	0.60	0.40	20	0.33	0.67	0.45	20
	GradInput	0.16	0.32	0.21	20	0.28	0.56	0.34	20	0.30	0.60	0.40	20	0.28	0.56	0.38	20
	SmoothGrad	0.12	0.25	0.16	20	0.24	0.48	0.32	20	**0.50**	**1.00**	**0.67**	20	0.22	0.43	0.29	20
	IG	0.16	0.32	0.22	20	0.24	0.49	0.33	20	**0.50**	**1.00**	**0.67**	20	0.28	0.55	0.37	20

We added 10 decoys and take k=20.

GNNExp is the best in detecting the injected decoys.

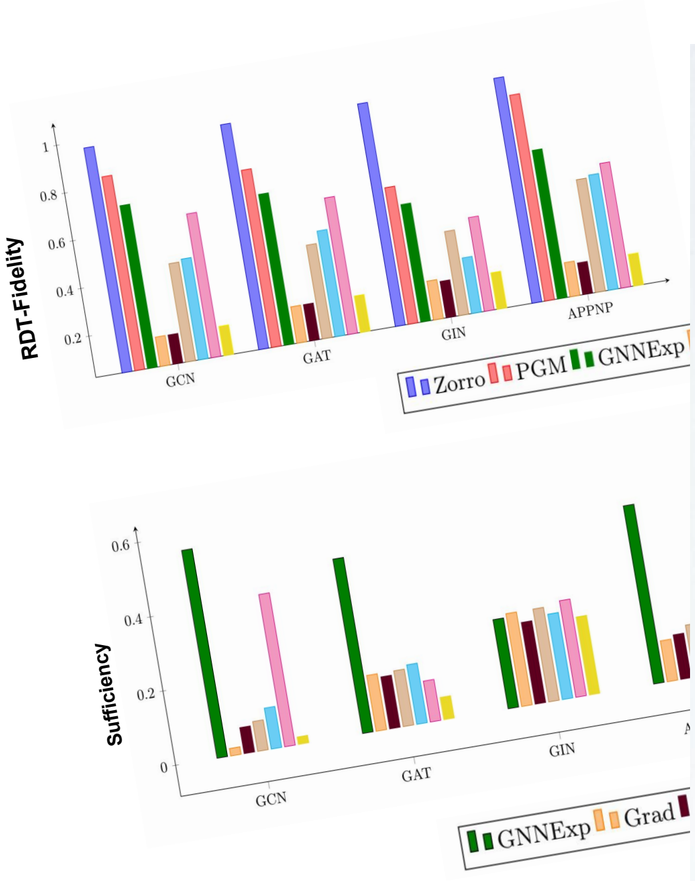
Results

Human Rationales	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
GNNExp	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
Grad	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...
CAM	The first problem that fair game has is the casting of supermodel cindy crawford in the lead role. not that cindy does that bad... sure william is n't a bad actor. unfortunately he just does n't demonstrate it all in this movie...

Mask	Methods	GCN auprc	GCN F1	GCN $\|\mathcal{S}\|$
Hard	PGM	—	0.42	25
Soft	GNNExp	0.46	0.54	168
	Grad	0.44	0.52	265
	GradInput	0.39	0.51	221
	SmoothGrad	0.40	0.52	219
	IG	0.37	0.49	225
	CAM	0.54	**0.61**	224
	GradCAM	0.67	0.34	175

Plausibility

Results



Mask	Methods	GCN				GAT				GIN				APPNP			
										CORA							
					F1	$\lvert\mathcal{S}\rvert$	P@k	R@k	F1	$\lvert\mathcal{S}\rvert$	P@k	R@k	F1	$\lvert\mathcal{S}\rvert$
										37	40	0.26	0.45	0.27	33	0.22	0.79	0.33	38
										24	20	0.18	0.36	0.25	20	0.19	0.38	0.25	20
										59	20	**0.50**	**1.00**	**0.67**	20	**0.34**	**0.67**	**0.58**	20
										39	20	0.30	0.60	0.40	20	0.33	0.67	0.45	20
										34	20	0.30	0.60	0.40	20	0.28	0.56	0.38	20
										32	20	**0.50**	**1.00**	**0.67**	20	0.22	0.43	0.29	20
													1.00	**0.67**	20	0.28	0.55	0.37	20

"A truly lively debate, yet no clear winner"
- Bill Abbott

e casting of supermodel cindy crawford in the lead role. not
n't a bad actor. unfortunately he just does n't demonstrate it

casting of supermodel cindy crawford in the lead role. not
't a bad actor. unfortunately he just does n't demonstrate it

asting of supermodel cindy crawford in the lead role. not
a bad actor. unfortunately he just does n't demonstrate it

sting of supermodel cindy crawford in the lead role. not
bad actor. unfortunately he just does n't demonstrate it

Conclusion

- Explainability in Graph
- Four dimensions of measuring the goodness of the explanation
    - Faithfulness
    - Sparsity
    - Correctness
    - Plausibility
- New dataset for Plausibility.
- Study different explainers.
- Lot more to explore

Thank You

rathee@l3s.de

Please scan for the
Github link

Leibniz Universität Hannover

L3S

TU Delft

Delft University of Technology