GNNX-BENCH: Unravelling The Utility Of Perturbation-Based GNN Explainers Through In-Depth Benchmarking

Mert Kosan[1*], Samidha Verma[2*], Burouj Armgaan[2], Khushbu Pahwa[3],

Ambuj Singh[1], Sourav Medya[4], Sayan Ranu[2]

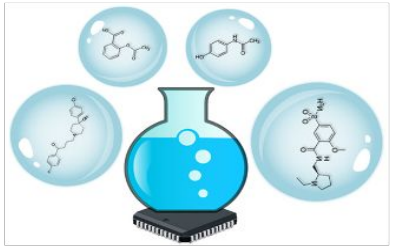
[1]University of California, Santa Barbara ; [2]Indian Institute of Technology, Delhi ;

[3]Rice University, Texas ; [4]University of Illinois, Chicago

ICLR 2024

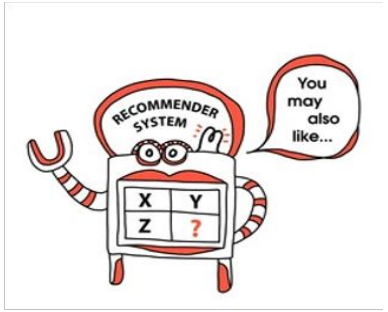
Need for explainability

GNNs can be used in research fields, industrial applications and high-stake use-cases.



Drug Discovery

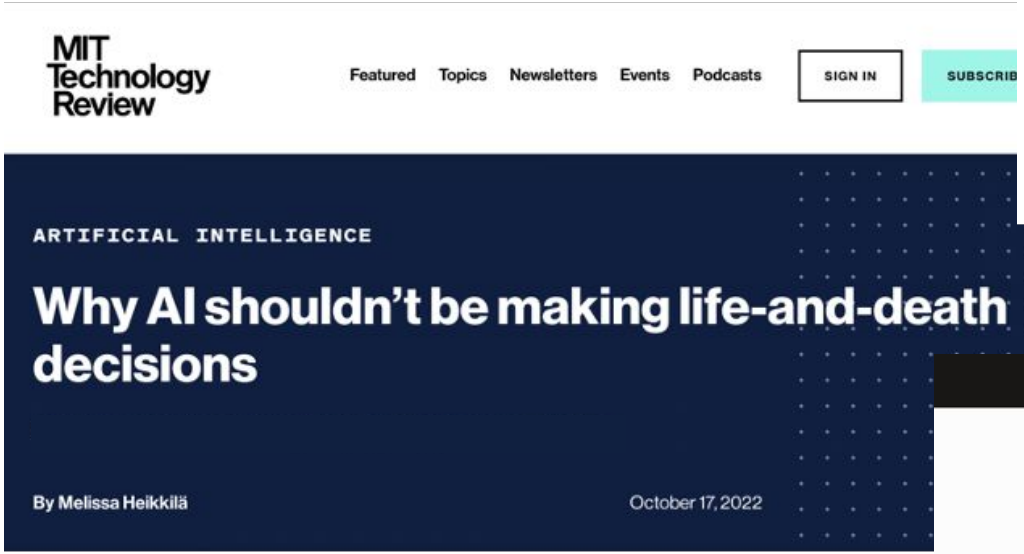
Social Networks

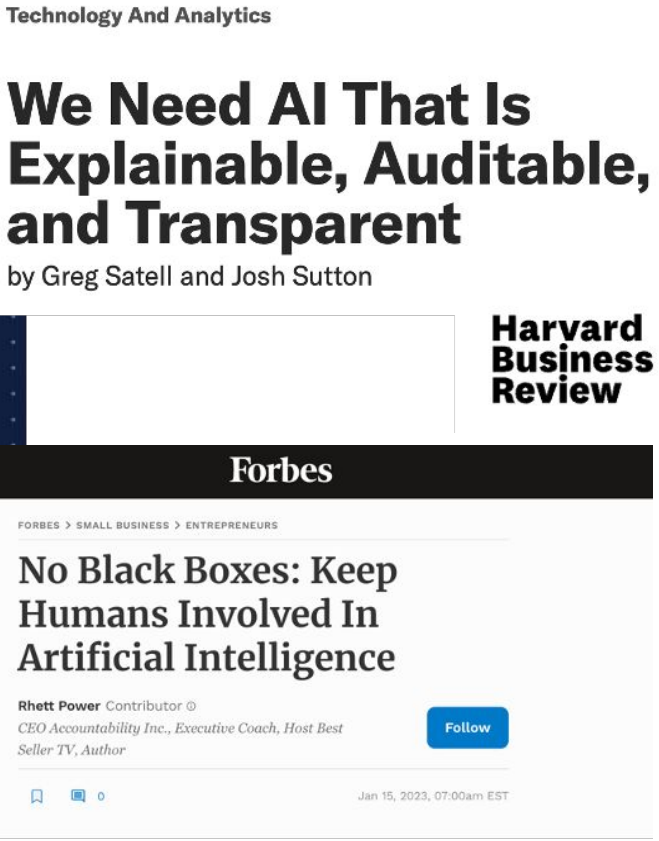


Recommendation Systems

Criminal Justice

“With great power comes great responsibility”



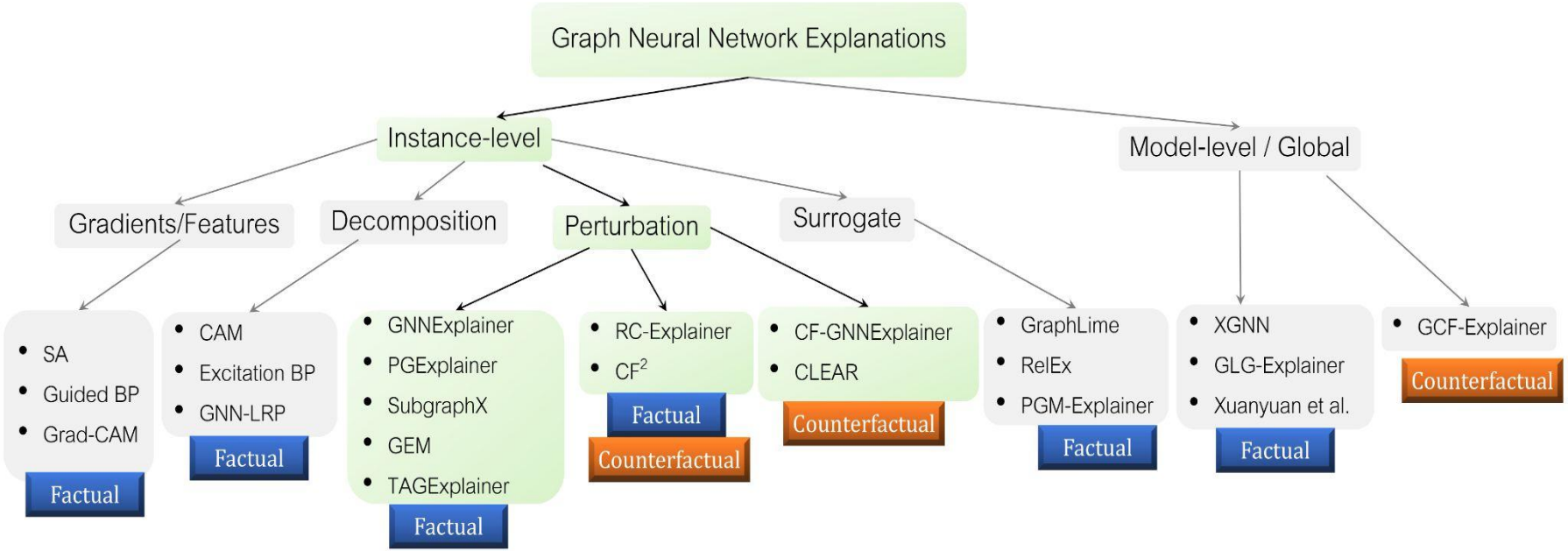


“With great power comes great responsibility”

Good Performance
(Accurate Results)

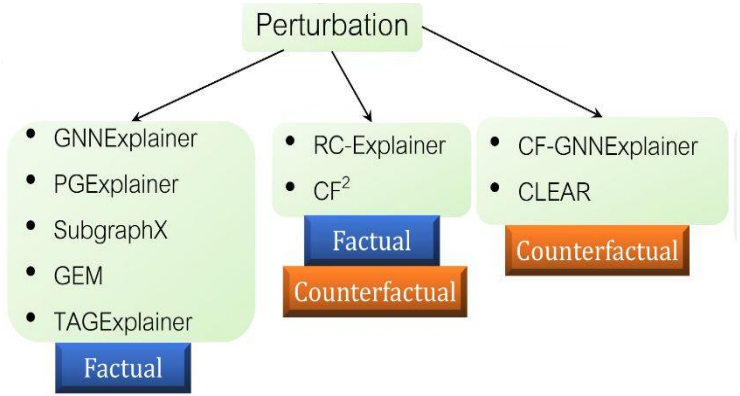
Explainable
(Trustworthy)

GNNX-BENCH : An In-Depth Benchmarking of Perturbation -based Explainers



Why GNNX-BENCH ?

- ❑ Prior studies prefer breadth over depth.
- ❑ Significant increase in perturbation-based explainers.
- ❑ No benchmarks on counterfactual explainers.
- ❑ Clean codebase.



But … What are factual and counterfactual explainers ?

Factual

Finds smallest subgraph G_S of G, such that prediction on G and G_S is same.

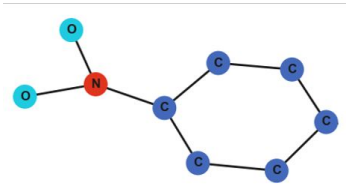
$$\mathcal{G}_S = \arg \min_{\mathcal{G}' \subseteq \mathcal{G},\ \Phi(\mathcal{G}) = \Phi(\mathcal{G}')} ||\mathcal{A}(\mathcal{G}_S)||$$

Counterfactual

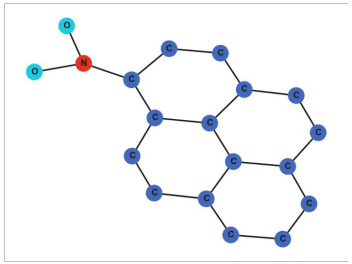
Finds minimally perturbed graph G′ for G, such that prediction on G and G′ is different.

$$\mathcal{G}^* = \arg \min_{\mathcal{G}' \in \mathbb{G},\ \Phi(\mathcal{G}) \neq \Phi(\mathcal{G}')} dist(\mathcal{G}, \mathcal{G}')$$

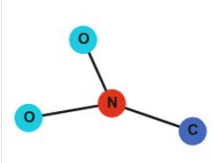
$$dist(\mathcal{G}, \mathcal{G}') = \|\mathcal{A}_{\mathcal{G}} - \mathcal{A}_{\mathcal{G}'}\|$$



(Sufficient for Mutagenicity)



Input graph
GNN Prediction: Mutagenic



(Necessary for Mutagenicity)

Key Empirical Investigations

Which is the best explainer?

Are these explainers stable in the face of optimization stochasticity, change in GNN architecture and topological noise?

How well do the explainers explain the model vs the underlying data?

Are the counterfactual recourses feasible?

Which is the best factual explainer?

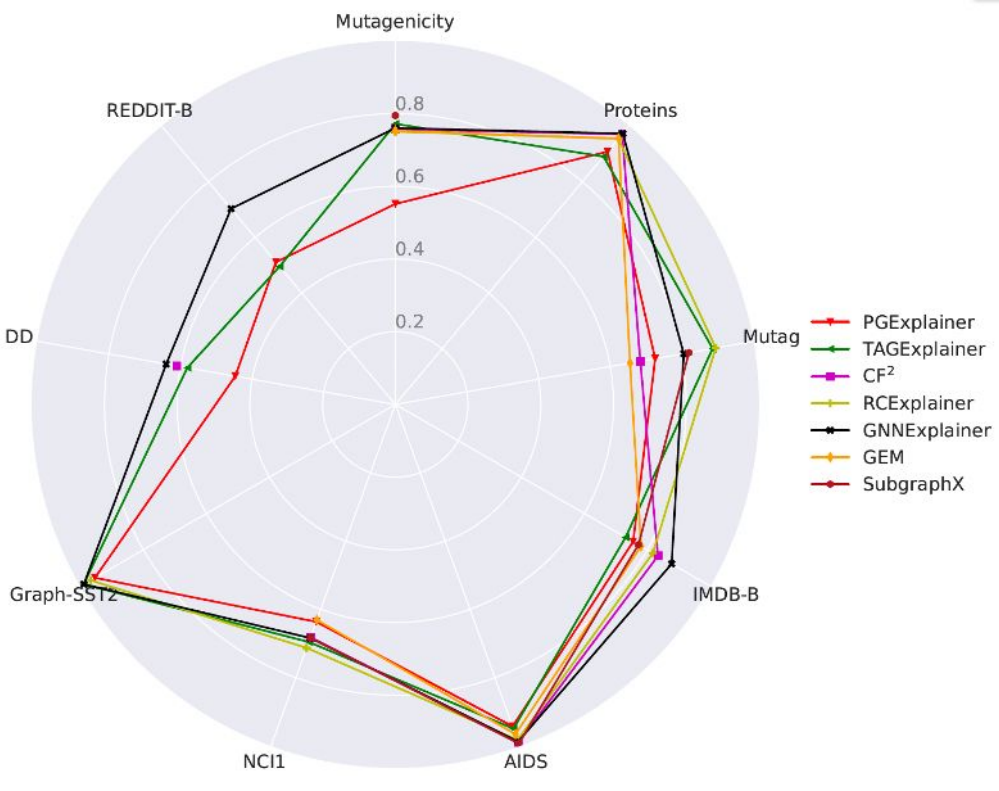
Sufficiency

G : Graph, G_S: explanation, Φ : GNN
P : #graphs for which $L_\Phi(G_S) = L_\Phi(G)$
N : Total #graphs

Sufficiency = P/N

Factual : Higher is better

GNNExplainer and RCExplainer outperform all other explainers.



Which is the best counterfactual explainer?

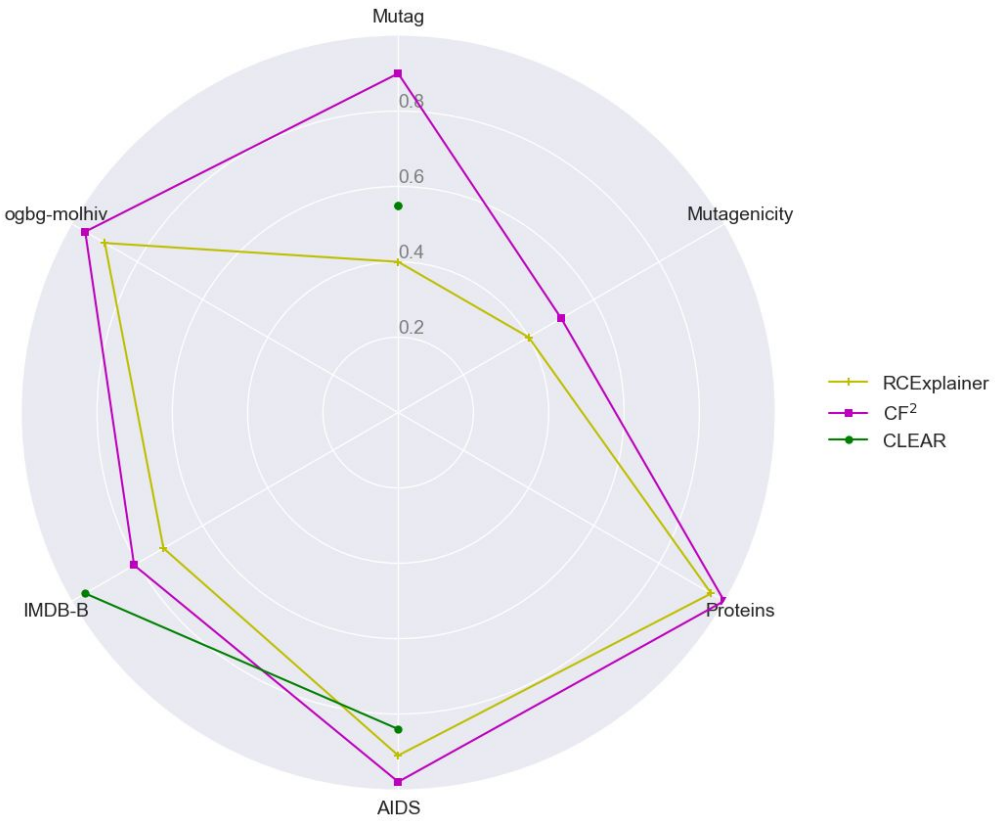
Sufficiency

G : Graph, G_C: Counterfactual of G,
Φ : GNN
P : #graphs for which $L_\Phi(G_C) = L_\Phi(G)$
N : Total #graphs

Sufficiency = P/N

Counterfactual : Lower is better

RCExplainer outperforms other counterfactual explainers.



How stable is this explainer ?

A. Optimization Stochasticity

B. Change in GNN Architecture

C. Perturbations: Features / Topology

Stability

$G(V, E)$: Graph, V : Vertex set, E : Edge set

Let, $E_X \subset E$ = Set of edges in original explanations

$E'_X \subseteq E$ = Set of edges in explanation after variation.

$$\text{Stability} = \frac{|E_X \cap E'_X|}{|E_X \cup E'_X|}$$

Higher is better

How stable is this explainer ?

A. Optimization Stochasticity : RCExplainer is the most stable.

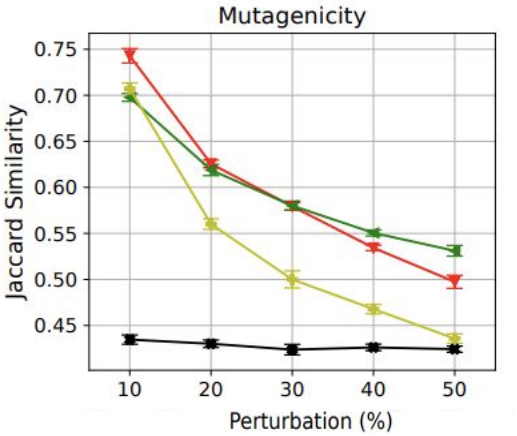
	PGExplainer			TAGExplainer			CF^2			RCExplainer			GNNExplainer		
Dataset / Seeds	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$
Mutagenicity	0.69	0.75	0.62	0.76	0.78	0.74	0.77	0.77	0.77	0.75	0.71	0.71	0.46	0.47	0.47
Proteins	0.38	0.51	0.38	0.55	0.48	0.46	0.34	0.34	0.35	0.88	0.85	0.91	0.28	0.28	0.28
Mutag	0.5	0.54	0.51	0.36	0.43	0.72	0.78	0.79	0.79	0.86	0.92	0.87	0.57	0.57	0.58
IMDB-B	0.67	0.76	0.67	0.67	0.60	0.56	0.32	0.32	0.32	0.75	0.73	0.70	0.18	0.19	0.18
AIDS	0.88	0.87	0.82	0.81	0.83	0.87	0.85	0.85	0.85	0.95	0.96	0.97	0.80	0.80	0.80
NCI1	0.58	0.55	0.64	0.69	0.81	0.65	0.60	0.60	0.60	0.71	0.71	0.94	0.44	0.44	0.44

B. GNN Architecture : PGExplainer and RCExplainer are the most stable.

	PGExplainer			TAGExplainer			CF^2			RCExplainer			GNNExplainer		
Dataset / Architecture	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE
Mutagenicity	0.63	0.65	0.60	0.24	0.25	0.32	0.52	0.47	0.54	0.56	0.52	0.46	0.43	0.42	0.43
Proteins	0.22	0.47	0.38	0.45	0.41	0.18	0.28	0.28	0.28	0.37	0.41	0.42	0.28	0.28	0.28
Mutag	0.57	0.58	0.69	0.60	0.65	0.64	0.58	0.56	0.62	0.47	0.76	0.54	0.55	0.57	0.55
IMDB-B	0.48	0.45	0.56	0.44	0.35	0.47	0.17	0.23	0.17	0.30	0.33	0.26	0.17	0.17	0.17
AIDS	0.81	0.85	0.87	0.83	0.83	0.84	0.80	0.80	0.80	0.81	0.85	0.81	0.8	0.8	0.8
NCI1	0.39	0.41	0.37	0.45	0.17	0.58	0.37	0.38	0.38	0.49	0.53	0.52	0.37	0.38	0.39

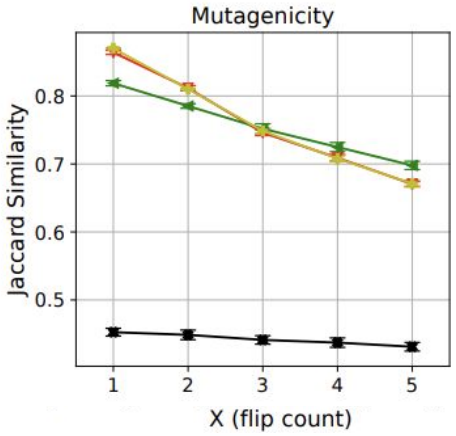
How stable is this explainer ?

C.(i) Feature Perturbation

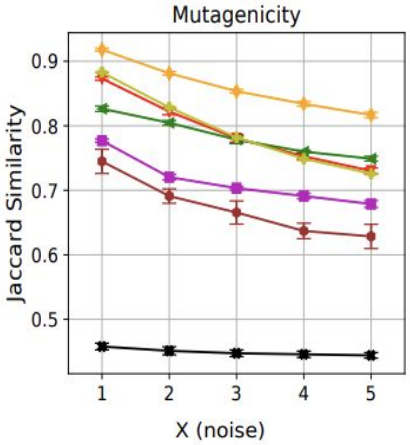


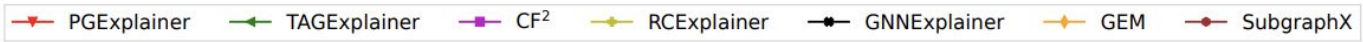
C.(ii) Topological Perturbation

Adversarial Attack



Random



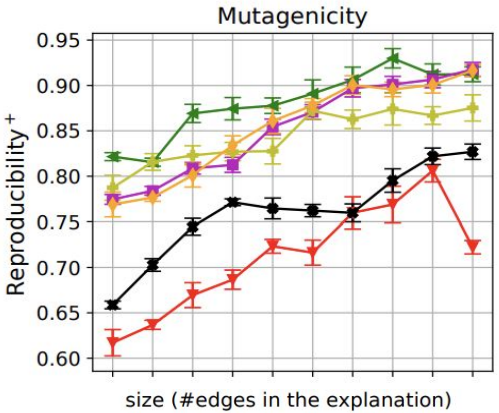


GEM, PGExplainer and RCExplainer are the most stable. But, significant stability issues exist in all explainers.

How well does the explainer explain the model vs the underlying data ?

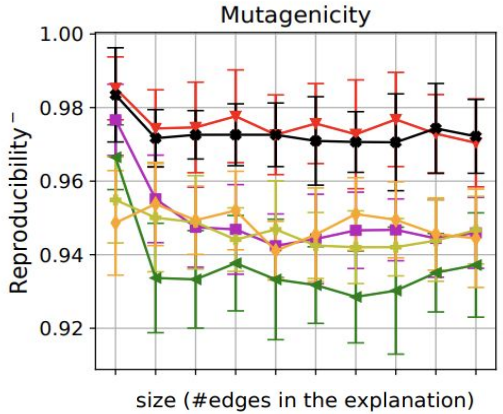
Reproducibility^{+}

Ratio of GNN's accuracy after retraining on only **explanation** graphs to the original accuracy.



Reproducibility^{-}

Ratio of GNN's accuracy after retraining on **residual graphs** (original graph minus explanation) to the original accuracy.

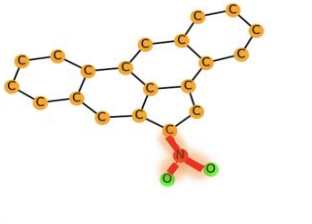


High reproducibility demonstrates that the explainers hardly capture the real cause of the GNN predictions.

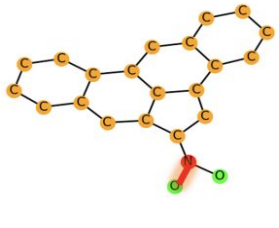
Are these counterfactual recourses feasible ?

Dataset	RCEXPLAINER			CF^2		
	Expected Count	Observed Count	p-value	Expected Count	Observed Count	p-value
Mutagenicity	233.05	70	< 0.00001	206.65	0	< 0.00001
Mutag	11	9	0.55	4	1	0.13
AIDS	17.6	8	< 0.00001	1.76	0	0.0001

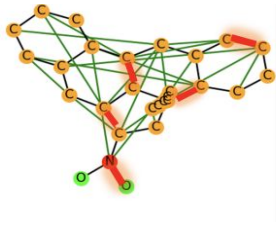
Statistical significance of deviations in the number of connected graphs between the test set and their corresponding counterfactual explanations on molecular datasets.



CF^2



RCExplainer



CLEAR

Predicted counterfactuals are not valid molecules with high probability !

Key Findings and Future Directions

RCExplainer shows superior performance in most cases.

- RCExplainer is consistently the most stable explainer.
- Most explainers suffer from significant deviations in the face of variational factors.

- Explainers only capture specific signals learned by the GNNs.
- They do not encompass all underlying data signals.

Counterfactual recourses showed statistically significant deviations in topological distribution from the original graphs.

Thank You !

Acknowledgements to our sponsors

Microsoft Research India
The National Science Foundation under grant #IIS-2229876
CSE Research Acceleration Fund of IIT Delhi.

Paper

Code