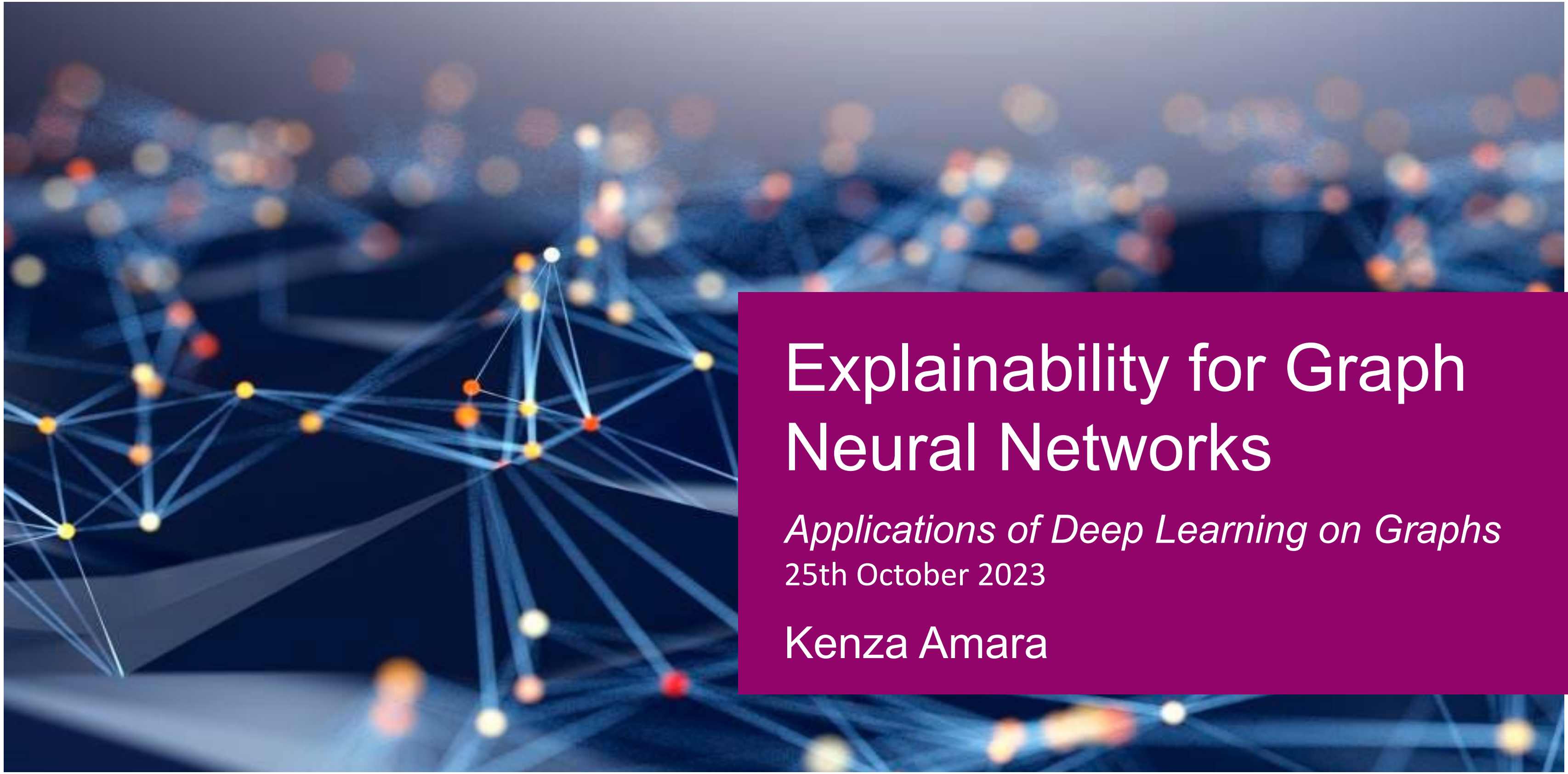


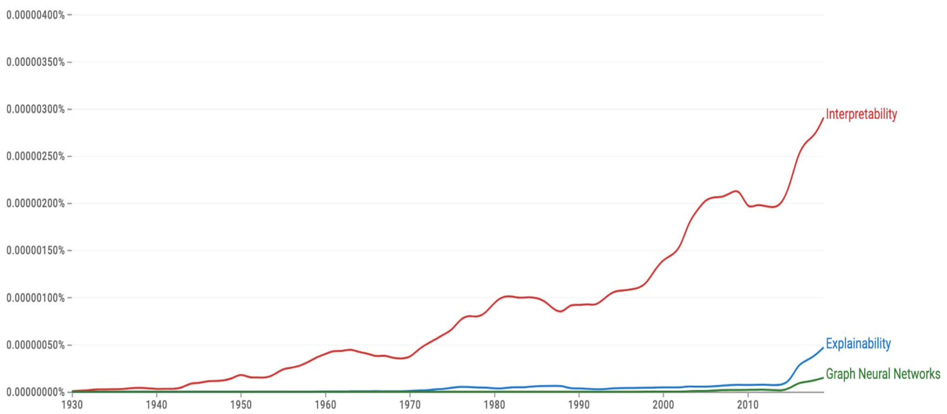
Explainability for Graph Neural Networks

Applications of Deep Learning on Graphs

25th October 2023

Kenza Amara

New research interests…



Google Ngram Viewer

Springer Link

Explainability

Discipline	see all
Computer Science	3,573
Engineering	956
Medicine & Public Health	364
Business and Management	262
Philosophy	215

Subdiscipline	see all
Artificial Intelligence	2,963
Information Systems Applications (incl. Internet)	728
Computational Intelligence	705
Computer Applications	646
Data Mining and Knowledge Discovery	623

Language	
English	6,037
German	67
Dutch	2

Graph Neural Networks

Discipline	see all
Computer Science	52,633
Engineering	31,726
Biomedicine	7,218
Life Sciences	4,803
Medicine & Public Health	4,770

Subdiscipline	see all
Artificial Intelligence	47,783
Computational Intelligence	16,208
Image Processing and Computer Vision	14,949
Computer Communication Networks	12,417
Data Mining and Knowledge Discovery	11,858

Language	
English	124,393
German	656
French	10
Italian	2

Explain module in PyTorch Geometric

torch_geometric.explain

Warning

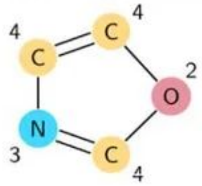
This module is in active development and may not be stable. Access requires installing PyG from master.

Contents

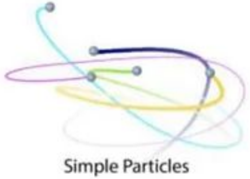
Philoshopy

This module provides a set of tools to explain the predictions of a PyG model or to explain the underlying phenomenon of a dataset (see the "GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks" paper for more details).

Graph Structure Data



Chemistry [1]
- Learn on molecules and predict chemical properties
- Use in drug repurposing

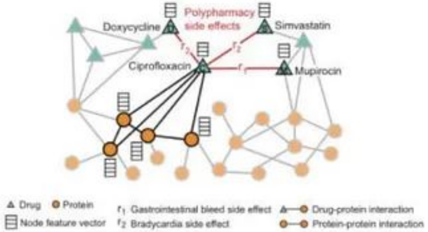


Physics [2]
- Learn from interactions of particles in systems
- Accelerate physics research

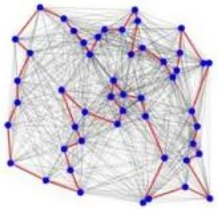
Neuroscience [5]
- Learn functions of brain regions through connectivity
- Accelerate brain-understanding and neuro-disease research

Numerous such examples of graph data.

Social networks [3]
- Learn from multi-faceted interactions among users
- Use for commercial and social applications



Medicine [4]
- Learn the effects of multiple drugs on body proteins
- Use for efficient multi-drug medical therapies



Combinatorial Optimization [6]
- Exploit the fact that most CO problems are rep. as graphs
- Develop better approximated solutions for NP-hard problems

Applications of xAI for GNN → EXPLAINABILITY

- Health sciences: explain the activity of a **molecule** with chemical groups, atoms and bonds.
- Climate sciences: explain the radiation level in the atmosphere with **spatial and temporal atmospheric graph**.
- Social networks: explain the voting behavior of an individual based on the **community** he belongs to.
- Finance: detect and explain fraudulent behaviour of users based on personal details (email address, bank account status,...) encoded as **an heterogeneous graph**.
- E-commerce: explain the purchases of users on an e-commerce platform.

GNN xAI in Computer Science → INTERPRETABILITY

- Understand the inner workings of a model
- Debug a model
- Explain out-of-distribution shifts

PLAN

Problematic: What is Explainability of Graph Neural Network?

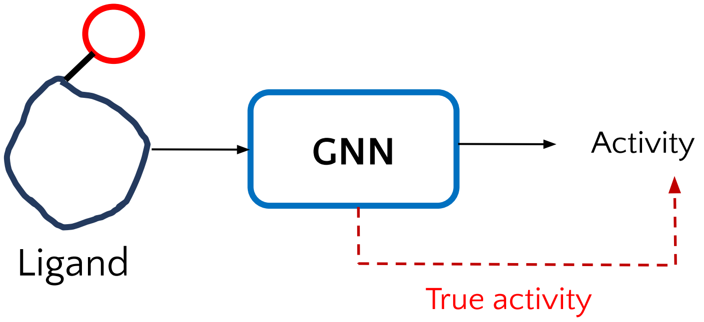
1. Definitions of xAI for Graphs
2. Taxonomy of Methods
3. Evaluation of Explanations
4. Challenges

PART 1: DEFINITIONS OF GNN EXPLAINABILITY

Example: Explain the toxicity of a molecule

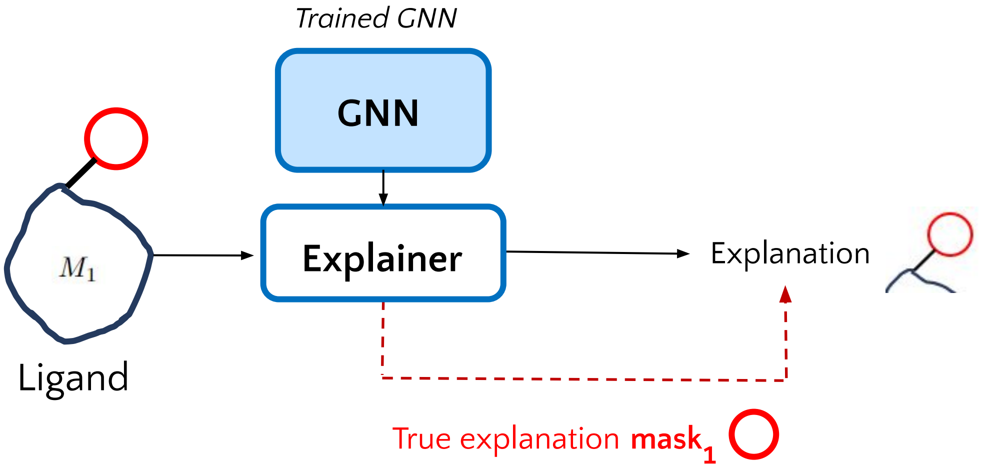
Step 1 – GNN training:

Predict compound activity



Step 2 - Explainability:

Color important atoms



Explainable AI for Graph Neural Networks (GNNs)

Graph neural networks = neural networks that take as input nodes, edges and node features.

Explainability of GNNs consists in finding the entities in the graph that contribute the most to the GNN predictions.

Explanation of a GNN = subgraph from the computation graph, with subset of node features OR mask on nodes/edges/node features

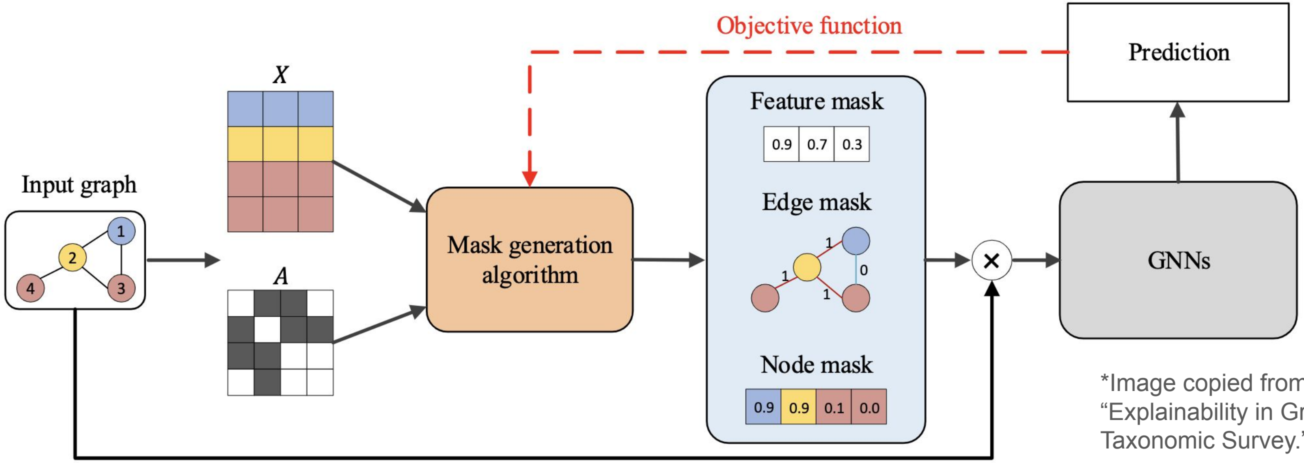
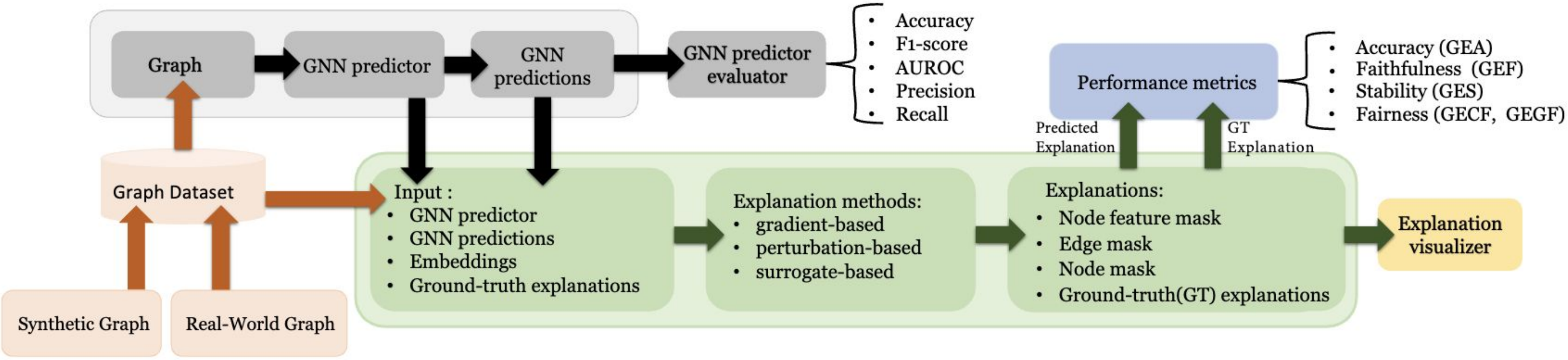


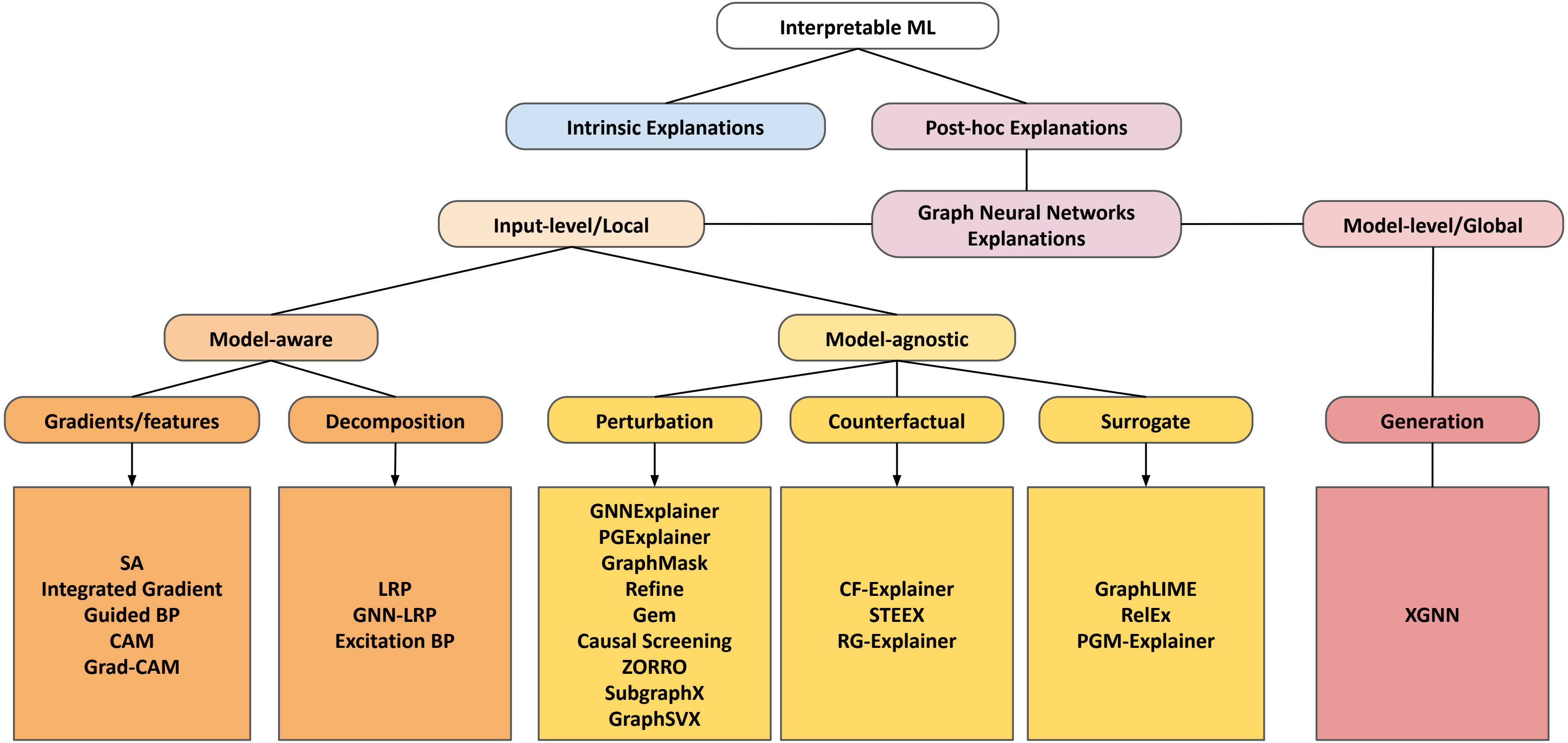
*Image copied from: Yuan et al. 2020. "Explainability in Graph Neural Networks: A Taxonomic Survey."

Overall xAI pipeline



PART 2:
EXPLAINABILITY METHODS FOR GNN

Non-generative explainability methods

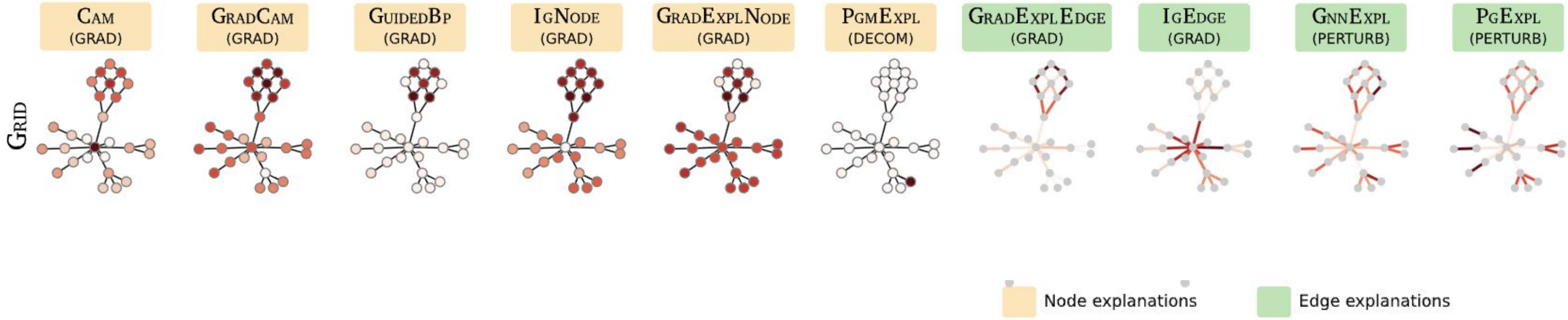


Example: Explanation of diverse explainability methods

BA-Grid dataset:

- Binary classification based on the presence of a grid motif attached to the Barabasi base graph.
- Human-intelligible explanation = the grid motif

Do explainers highlight the expected explanation?



Focus on GNNExplainer

… one of the most popular method in GNN xAi.

Main principle: reducing redundant information in a graph which does not directly impact the decisions.

Properties:

- Post-hoc explainability method
- Input-level explainer
- Perturbation-based method
- Discover the subgraph that preserves the best the model prediction

Ying, Zhitao, et al. "Gnnexplainer: Generating explanations for graph neural networks." *Advances in neural information processing systems* 32 (2019).

Mutual Information

Goal: Maximize the mutual information = the change in the probability of the initial prediction and the prediction when the graph is limited to the subgraph G_s and the node features limited to X_s.

$$\max_{G_S} MI\left(Y,(G_S,X_S)\right) = H(Y) - H(Y|G=G_S, X=X_S)$$

$\Rightarrow$ Minimize the conditional entropy of returning the initial predictions.

$$\min_{\mathcal{G}} \mathbb{E}_{G_S \sim \mathcal{G}} H(Y|G=G_S, X=X_S)$$

GNNExplainer Loss

<u>The loss has 3 terms:</u>

- **Mutual information**: The GNN prediction taking as input the masked graph should be as close as possible to the GNN prediction on the whole graph
  → constraint on the mutual information
- **Mask size**: The subgraph must be smaller than the initial graph
  → constraint on the mask size
- **Entropy**: The explanation should be discriminative
  → constraint on the entropy of the mask

$$\boxed{\mathcal{L} = -MI(Y, (G_S, X_S)) + \mu_e \cdot |\mathcal{V}_S| + \lambda_e \cdot \mathcal{H}_m}$$

Generative VS Non-generative explainers

Non-generative methods: optimize an explanation for individual instances.

Generative methods: learn a strategy to generate the most explanatory subgraph across the whole dataset.
*It learns the distribution of the underlying explanatory graphs using **a parametrized subgraph generator** trained on all the data* → holistic approach!

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log P_{Y^*}(G_e|\theta, G),$$

probability that the generated graph is a valid explanation for the target label Y*

<u>Optimization objective:</u>

$$\min_{\theta} -\log P_{Y^*}(G_e|\theta, G) + \mathcal{L}_{\text{INFO}}(G_e, G) := \mathcal{L}_{\text{ATTR}}(G_e, Y^*) + \mathcal{L}_{\text{INFO}}(G_e, G).$$

*measures how much the explanatory subgraph capture the important substructures for the target label Y**

assures sparsity and conciseness of the explanation

Generative Explainability Methods

Method	Generator	Information Constraint	Level	Scenario	Output
PGExplainer [28]	Mask Generation	size	instance	factual	E
GIB [50]	Mask Generation	mutual information	instance	factual	N
GSAT [30]	Mask Generation	variational	instance	factual	E
GNNInterpreter [44]	Mask Generation	size	model	factual	N / E / NF
GEM [26]	VGAE	size	instance	factual	E
CLEAR [29]	VGAE	size	instance	counterfactual	E / NF
OrphicX [27]	VGAE	variational & size	instance	factual	E
D4Explainer	Diffusion	size	instance & model	counterfactual	E
GANExplainer [25]	GAN	-	instance	factual	E
RCExplainer [43]	RL-MDP	size	instance	factual	SUBGRAPH
XGNN [51]	RL-MDP	size	model	factual	SUBGRAPH
GFlowExplainer [23]	RL-DAG	size	instance	factual	SUBGRAPH

PART 3: EVALUATION OF EXPLAINABILITY

Problem

- GraphLIME
- GradCAM
- EB
- GBP
- SA
- PGExplainer
- Gem
- SubgraphX
- XGNN
- PGM-Explainer
- GNNExplainer
- GNN-LRP
- DeepLift

Methods have…
Different settings → Methods are...
Not comparable

How do these GNN explanation methods compare with each other?

How should we evaluate these GNN explanation methods?

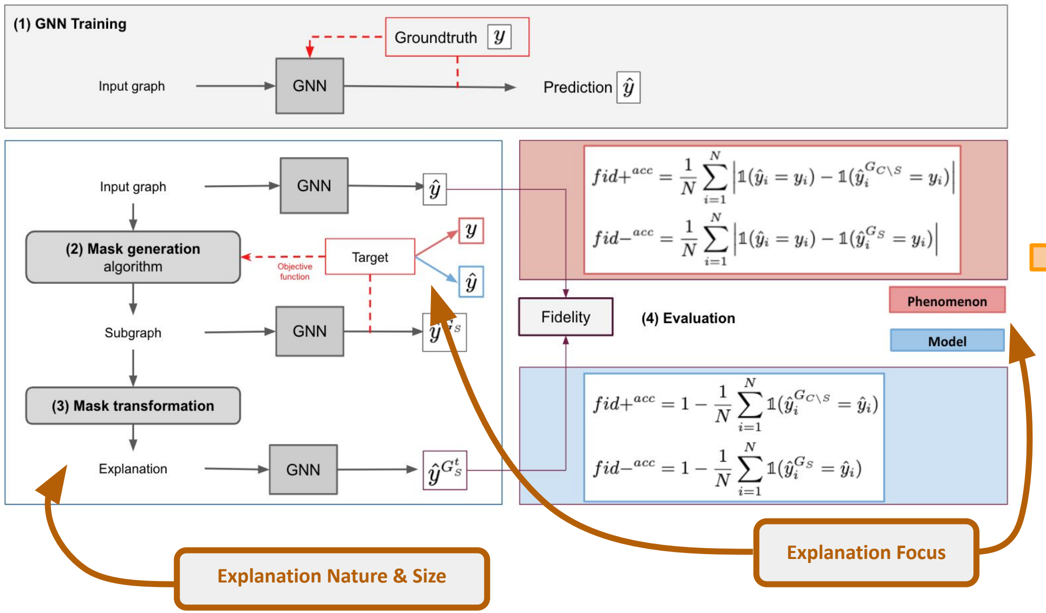
What is the optimal method to my problem?

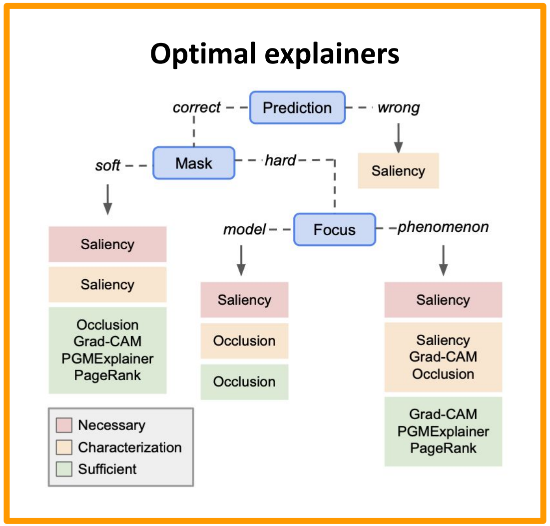
GraphFramEx

GraphFramEx

A systematic evaluation framework for explainability methods of graph neural networks.

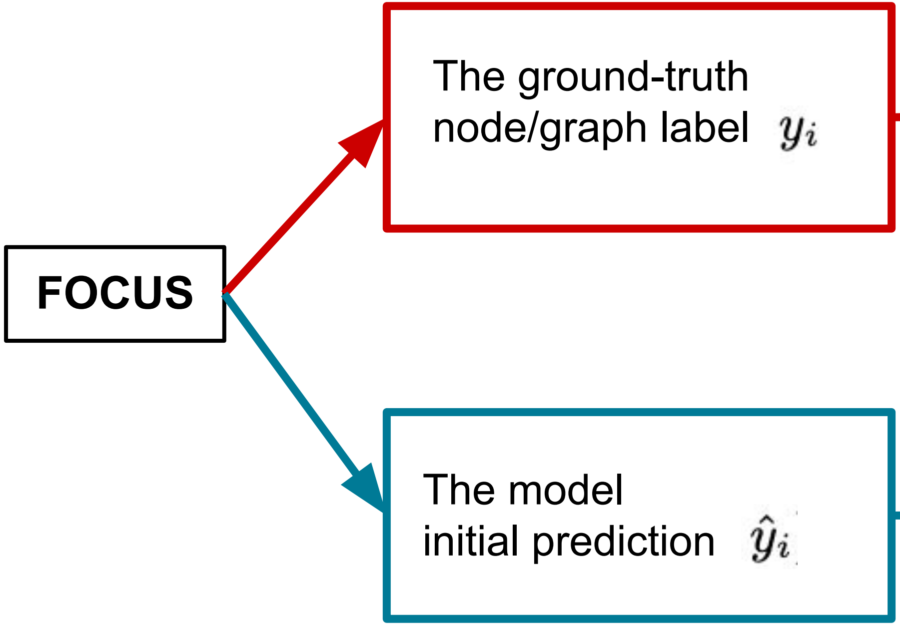
- Consider ***users need*** in the evaluation protocol on ***3 aspects***: explanation focus, mask nature, and mask transformation
- Distinguish ***2 types*** of explanations: ***necessary*** or ***sufficient***
- Investigate the ***influence of GNN accuracy*** on explainability





Phenomenon VS Model Focus

Given an explanation, the GNN model has to reproduce …



Phenomenon

$$fid_+ = \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = y_i)\right|$$

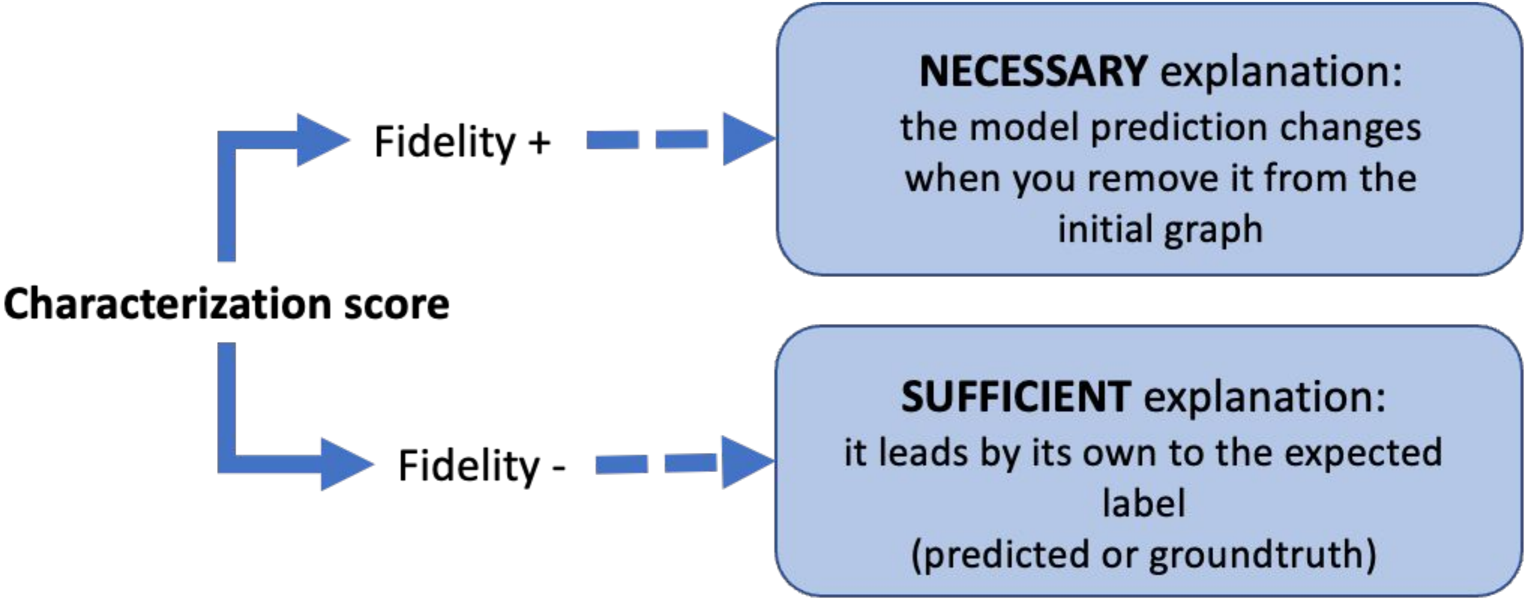
$$fid_- = \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_S} = y_i)\right|$$

Model

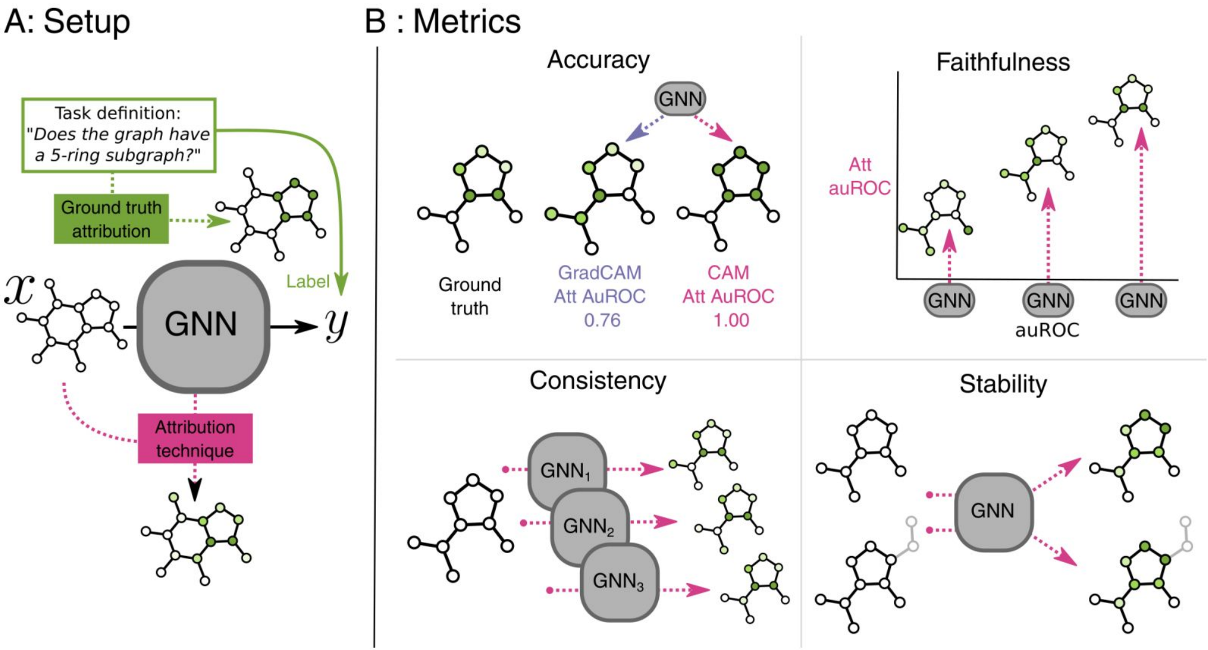
$$fid_+ = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = \hat{y}_i)$$

$$fid_- = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i^{G_S} = \hat{y}_i)$$

Necessary VS Sufficient Explanations



Popular Evaluation Metrics in GNN Explainability



Accuracy *measures how well an attribution matches ground-truth.*

Faithfulness *measures how well the performance of an attribution method matches model performance.*

Consistency *measures how accuracy varies across different hyperparameters of a model.*

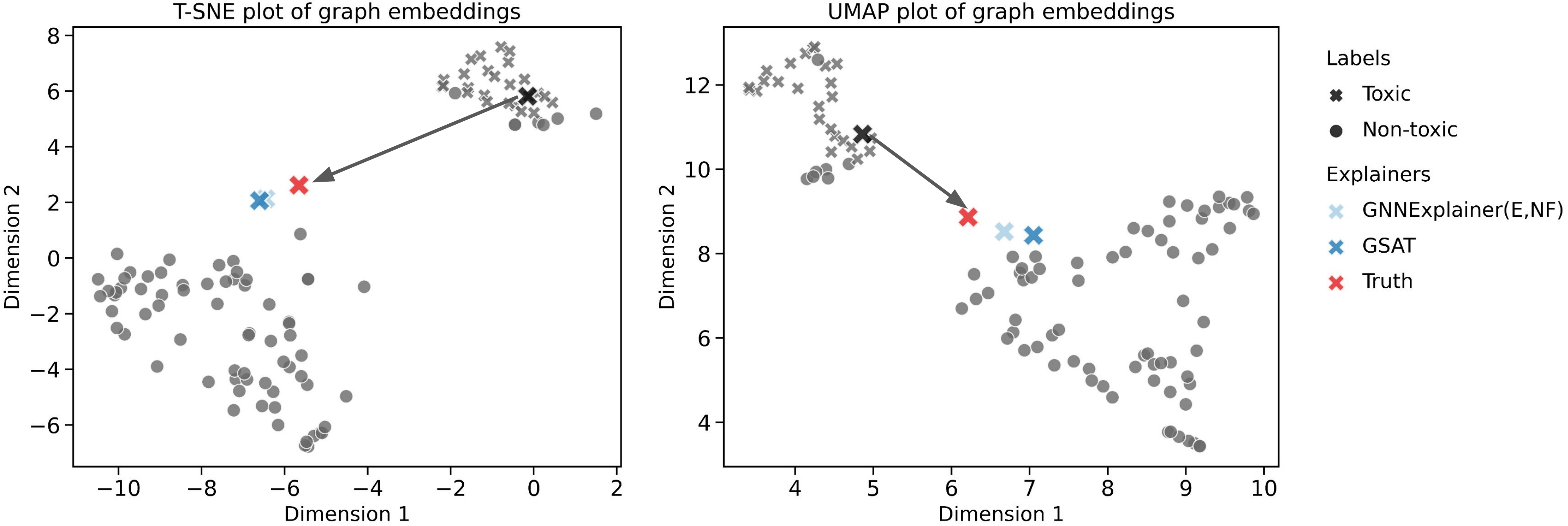
Stability *measures how attributions change when the input is perturbed*

Faithfulness Limitations

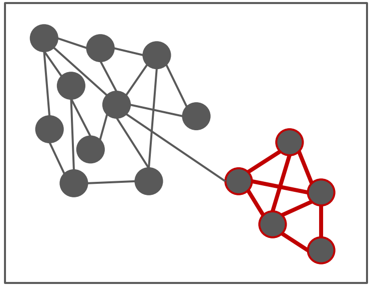
Fidelity metrics …

1. … evaluate **out-of-distribution** explanations
2. … are inconsistent with the **accuracy** metric
3. … lead to divergent conclusions across **datasets**
4. … depend on the **edge removal** strategy

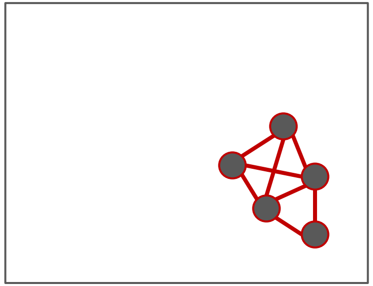
The Out-Of-Distribution Problem



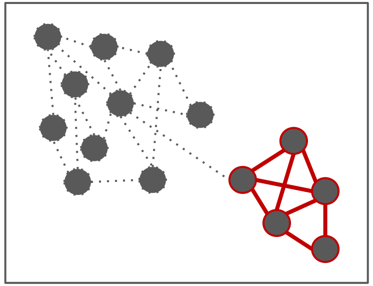
Edge Removal Strategy



Hard Explanation → *Explanatory subgraph containing only the important edges. Only the nodes connected are kept.*

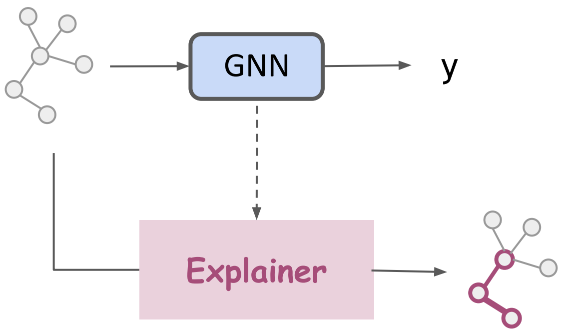


Soft Explanation → *Weighted graph where important edges have a weight of 1; the others a weight of 0. It preserves the whole graph structure with all nodes and edge indices*

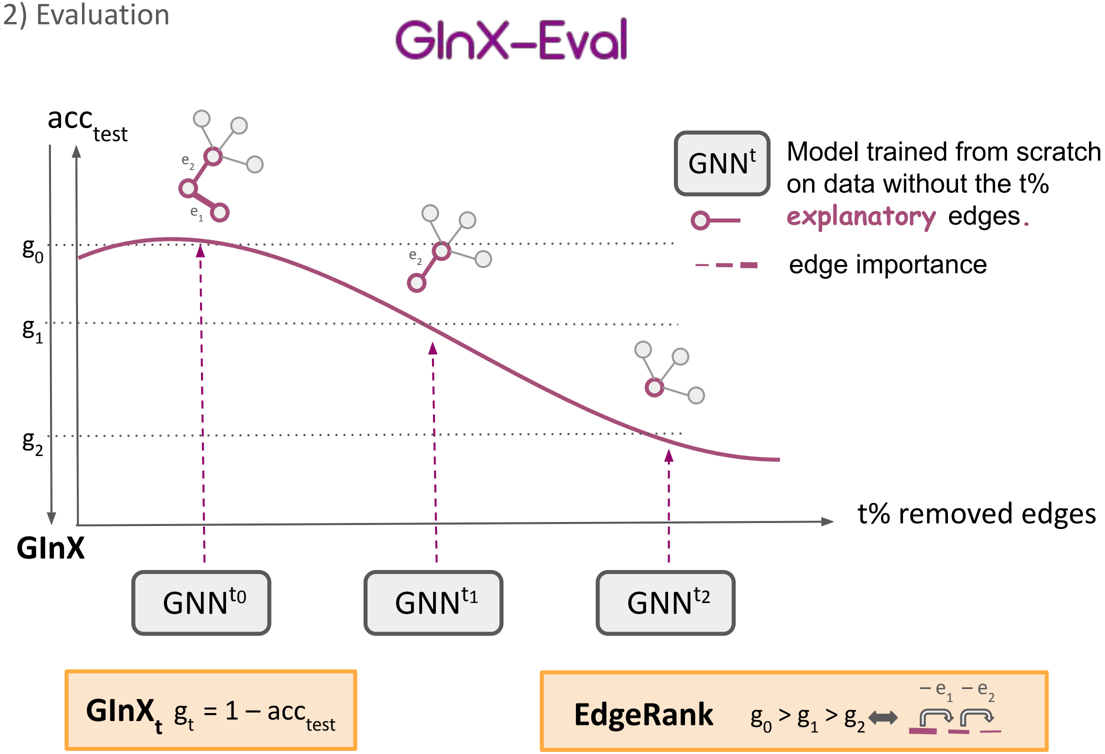


GInX-Eval Evaluation Procedure for GNN explainability method

(1) Explainability



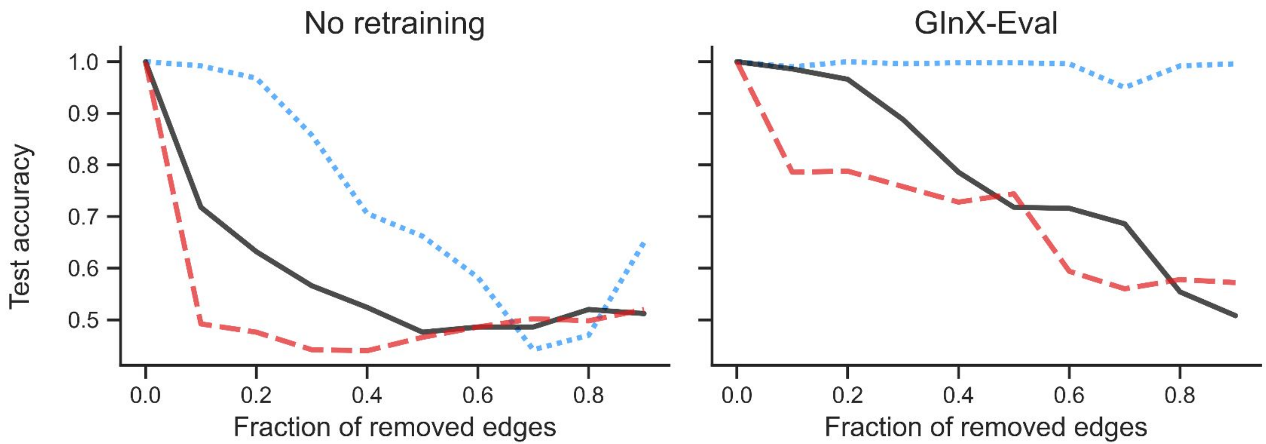
(2) Evaluation



GInX *score measures the drop of performance when removing edges*
→ edge informativeness to the model

EdgeRank *score measures if edges are removed according to their edge importance*
→ correct edge ordering.

GInX–Eval Overcomes the Out-Of-Distribution Problem



Baselines

Inverse = worst case scenario where edges are assigned the inverted ground-truth weights.

Random = random edge importance ~ uniform distribution

Truth = pre-defined ground-truth edge importance

GInX-Eval What it is and what it is not…

GInX-Eval is:

- a **validation tool** of ground-truth explanations → model-based xAI aligns with human-based xAI
- a **meta-evaluation** of new metrics → check agreement with GInX-Eval
- an insightful evaluation of **edge ranking power** of xAI methods.

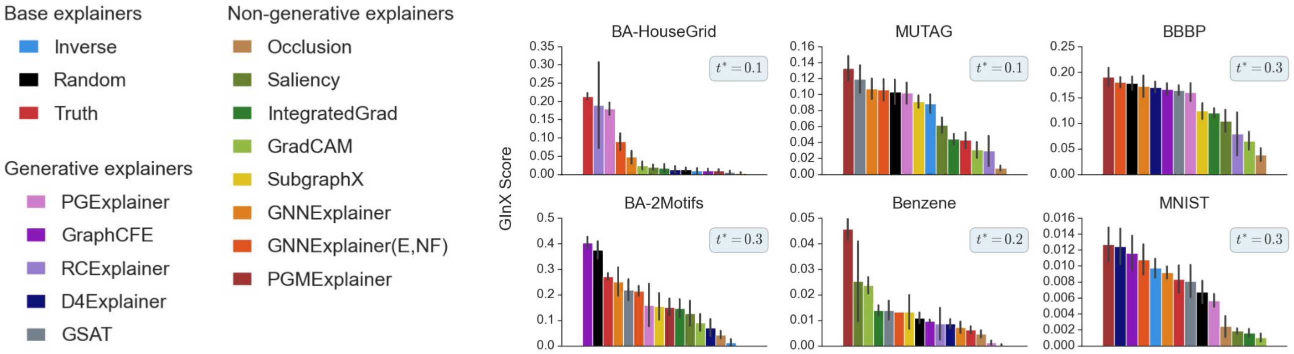
GInX-Eval is NOT:

- a **systematic** evaluation metric
- a **computationally scalable** metric

GInX-Eval

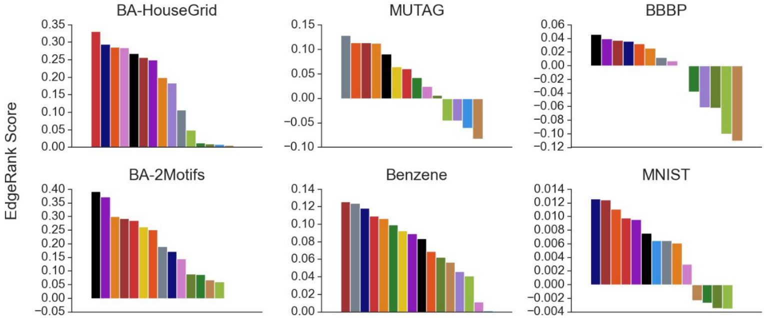
Results with GInX_{t^*} score and EdgeRank score

GInX score



Observation 1: gradient-based methods and Occlusion are the worse methods at capturing informative edges

EdgeRank score



Observation 2: gradient-based methods and Occlusion, RCExplainer, PGExplainer are the worse methods at correctly ordering edges by their importance.

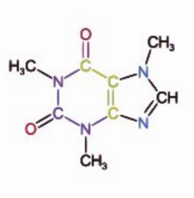
PART 4: CHALLENGES

Challenge 1: Explainability & Dataset

How do different types of data affect the explanations?

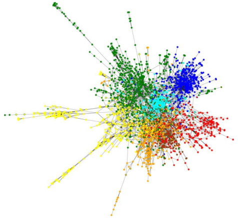
small-scale

ex: molecules

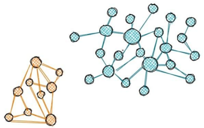


large-scale

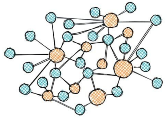
ex: citation networks



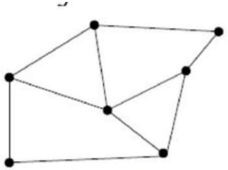
Homophily



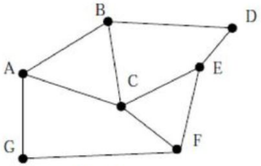
Heterophily



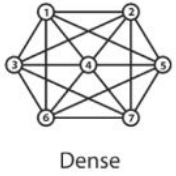
Unlabeled



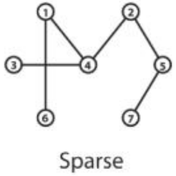
Labeled



Dense



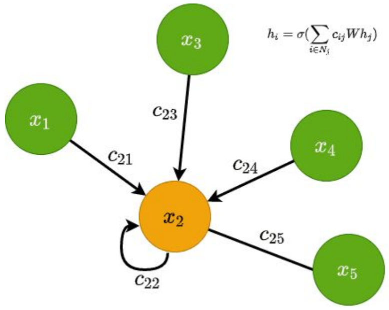
Sparse



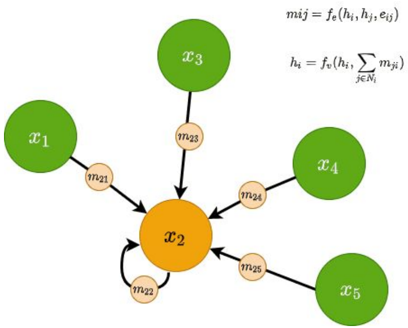
Challenge 2: Explainability & GNN model

What is the easiest architecture to explain?

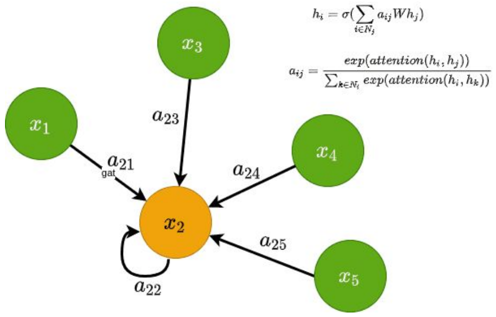
GCN



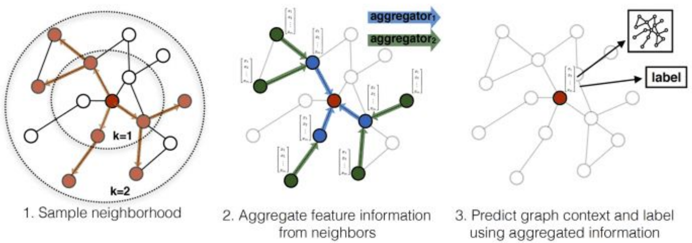
MPNN



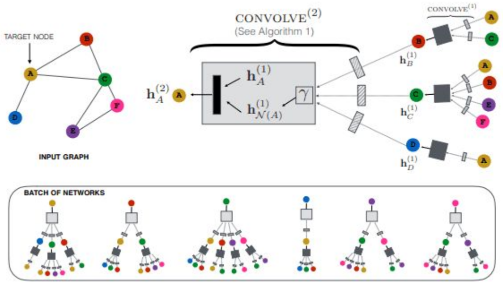
GAT



GraphSAGE



PinSAGE



Challenge 3: Explainability VS Interpretability

Explainability = explain a problem related to the model output, the data → *phenomenon-focus*

Challenge: Explanation without Ground Truths
Ground truth explanations for GNN are usually inaccessible => 1 main challenge:
→ select a suitable explanation among various methods in the absence of ground truth guidance.
… Potential solutions include involving human evaluations and creating synthetic datasets

$$\neq$$

Interpretability = explain the model inner workings, parameters and key steps → *model-focus*

Challenge: Dependent on the model architecture, not generalizable.

Challenge 4: Align model and human rationales

Questions to design explanations:

- Should the explanation be human-intelligible ?
- Should the explanation contain informative entities to the model?
- Is there an explanation for everything? that score high on both human-based and model-based evaluation?

Human-Model evaluation: Faithfulness? Accuracy? GInX-Eval!

Thank you for your attention

Questions?

Kenza AMARA
ETH AI Center & DS3Lab
kenza.amara@ai.ethz.ch
linkedin.com/in/kenza-amara
@KenzaAMARA4
github.com/k-amara

ETH AI CENTER

ETH zürich IVIA lab