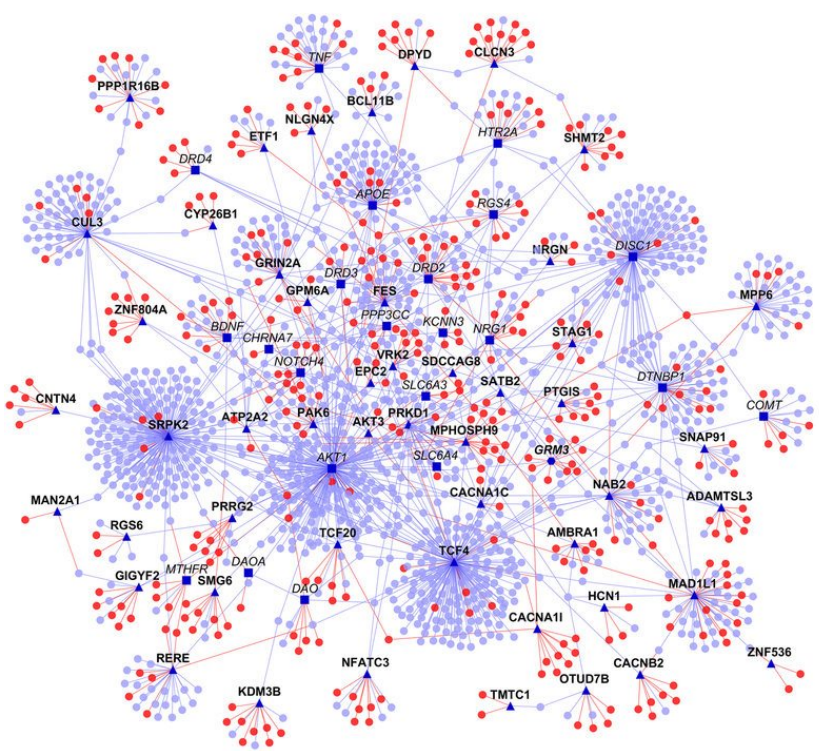
Explainability in Graph Machine Learning

Megha Khosla (TU Delft)

https://khosla.github.io
m.khosla@tudelft.nl

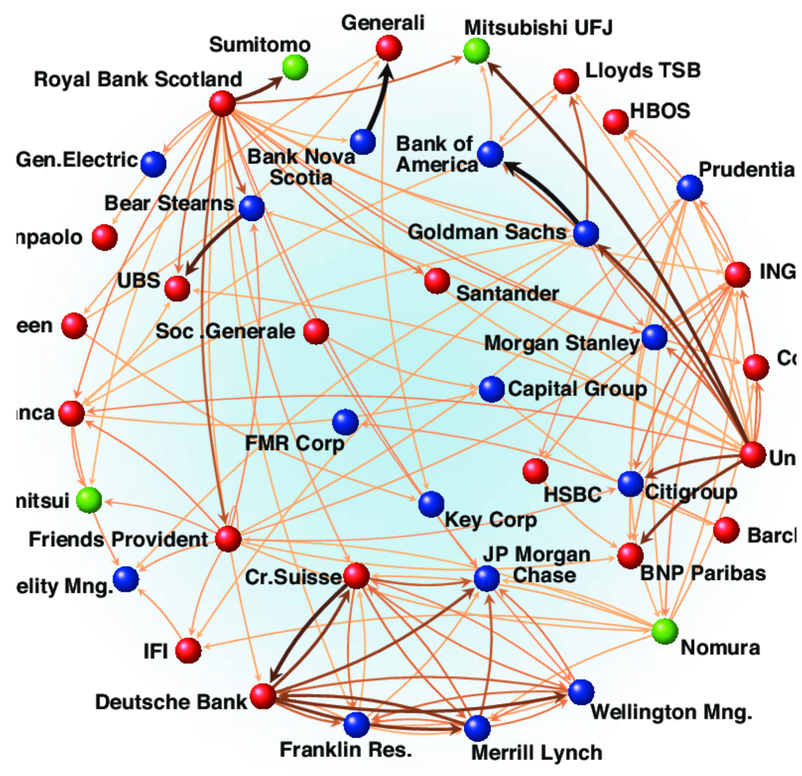
TU Delft

Graphs are everywhere



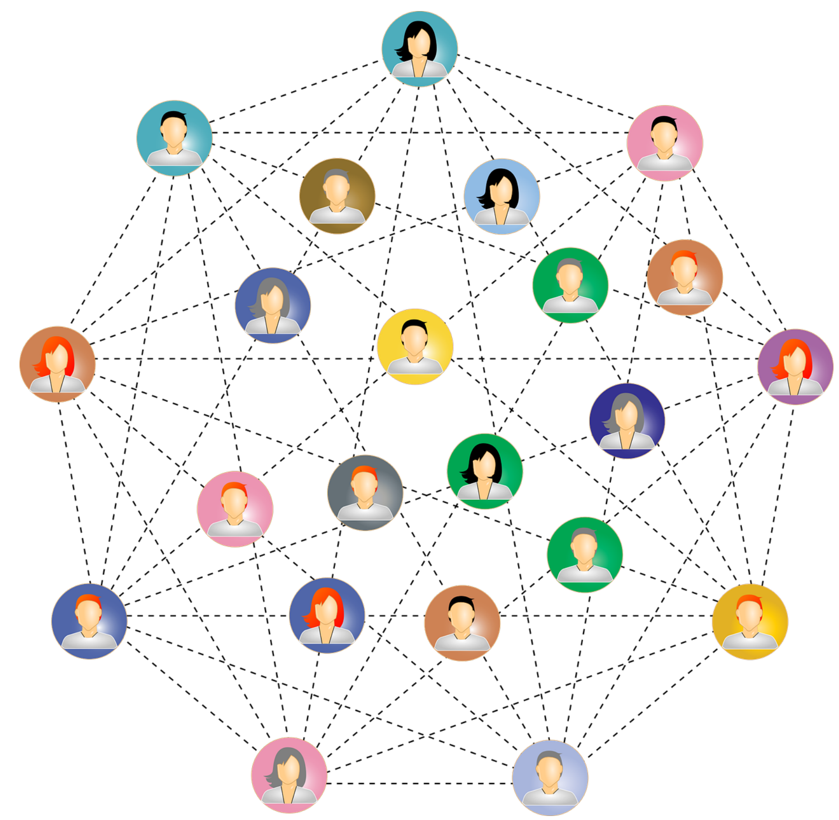
Protein interaction network

Image Source : wikipedia



Financial network

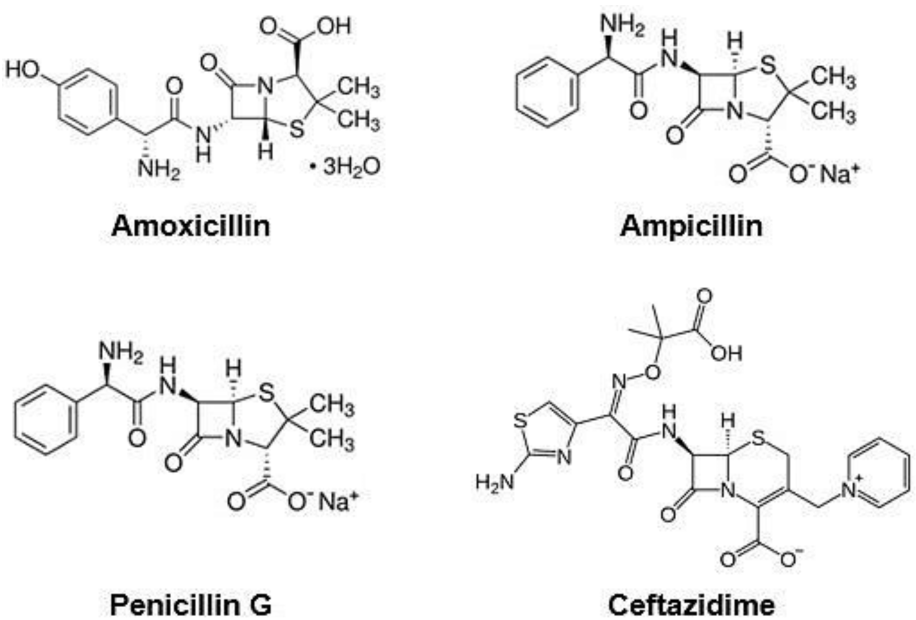
Image Source : Schweitzer et al. 2009



Social network

Image Source : Medium

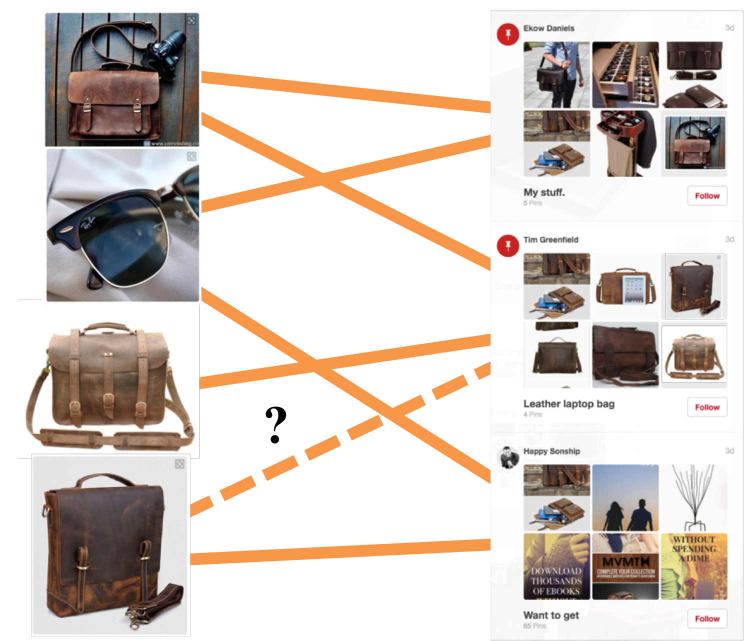
Success of Graph Machine Learning

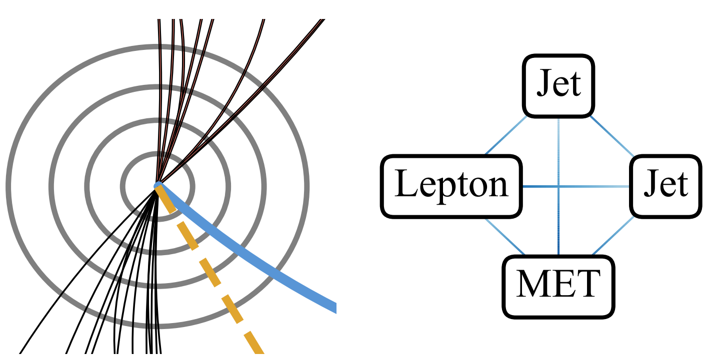


discover **novel antibiotics** (Stokes *et al.*, Cell'20)

Image Source : Coman et al. 2017

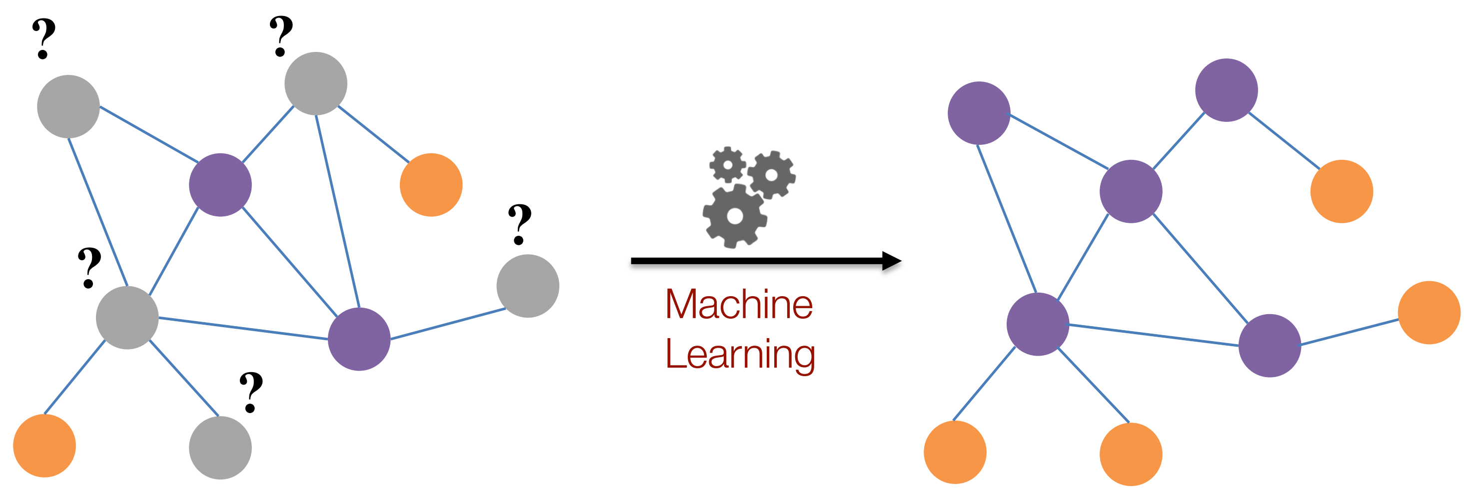
power **web-scale recommender systems** (Ying *et al.*, KDD'18; Pal *et al.*, KDD'20)





assist **particle physicists** (Shlomi *et al.*, Mach. Learn.: Sci. Technol'21)

Typical ML Tasks on Graphs



Node classification

Link prediction

Graph classification

Community detection

Graph Machine Learning (GraphML)

Shallow Node Embedding Methods

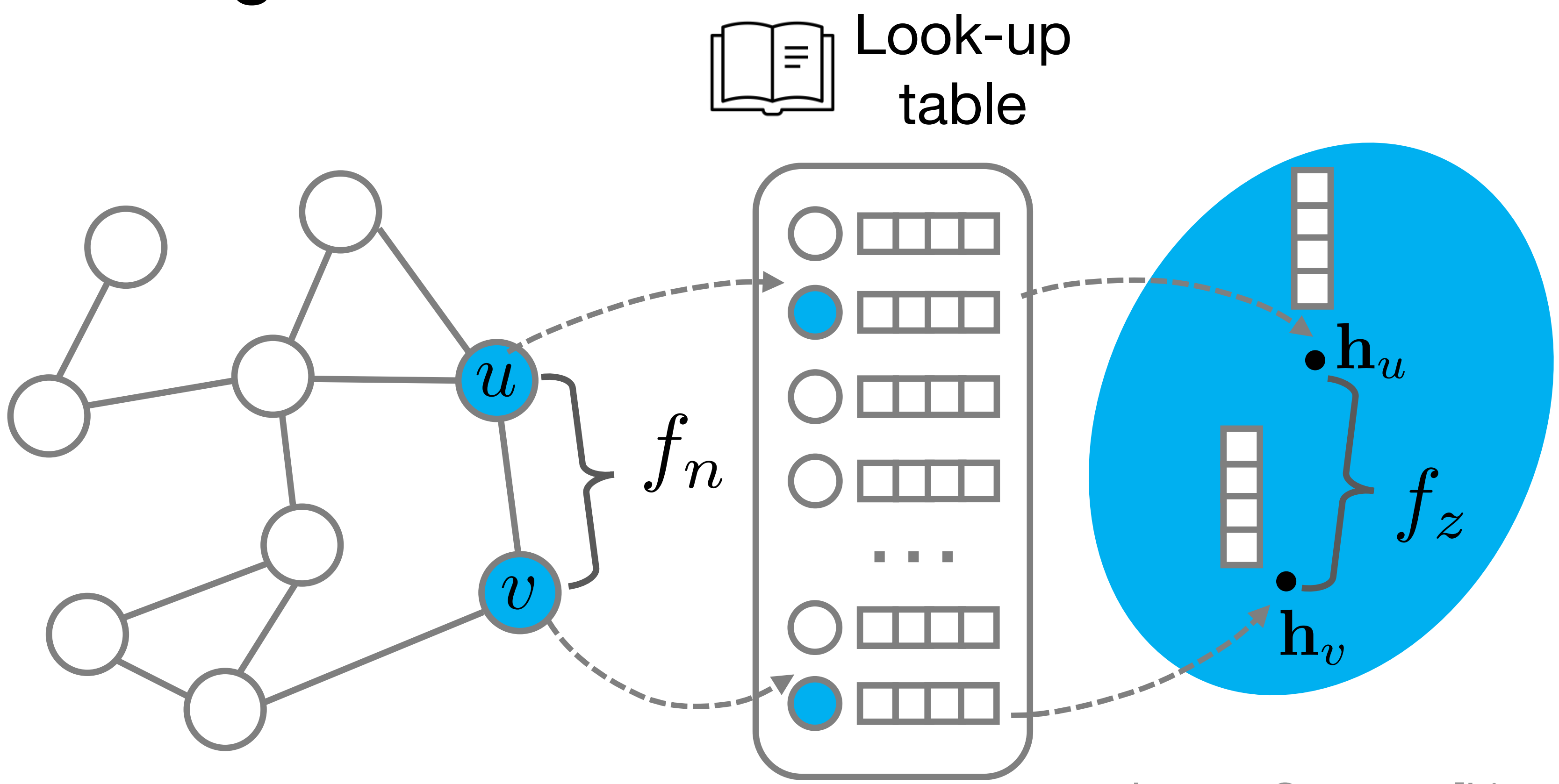


Image Source: [Li et al., 2022]

- Generate a look up table for node representations
- Similar nodes get embedded closer

Examples :

DeepWalk, Node2Vec, NERD, HOPE

Message Passing Graph Neural Networks (GNNs)

$$z_i^{(\ell)} = \text{AGGREGATE}\left(\left\{x_i^{(\ell-1)}, \left\{x_j^{(\ell-1)} \mid j \in \mathcal{N}_i\right\}\right\}\right)$$

$$x_i^{(\ell)} = \text{TRANSFORM}\left(z_i^{(\ell)}\right)$$

Examples :

GCN, GAT, GIN

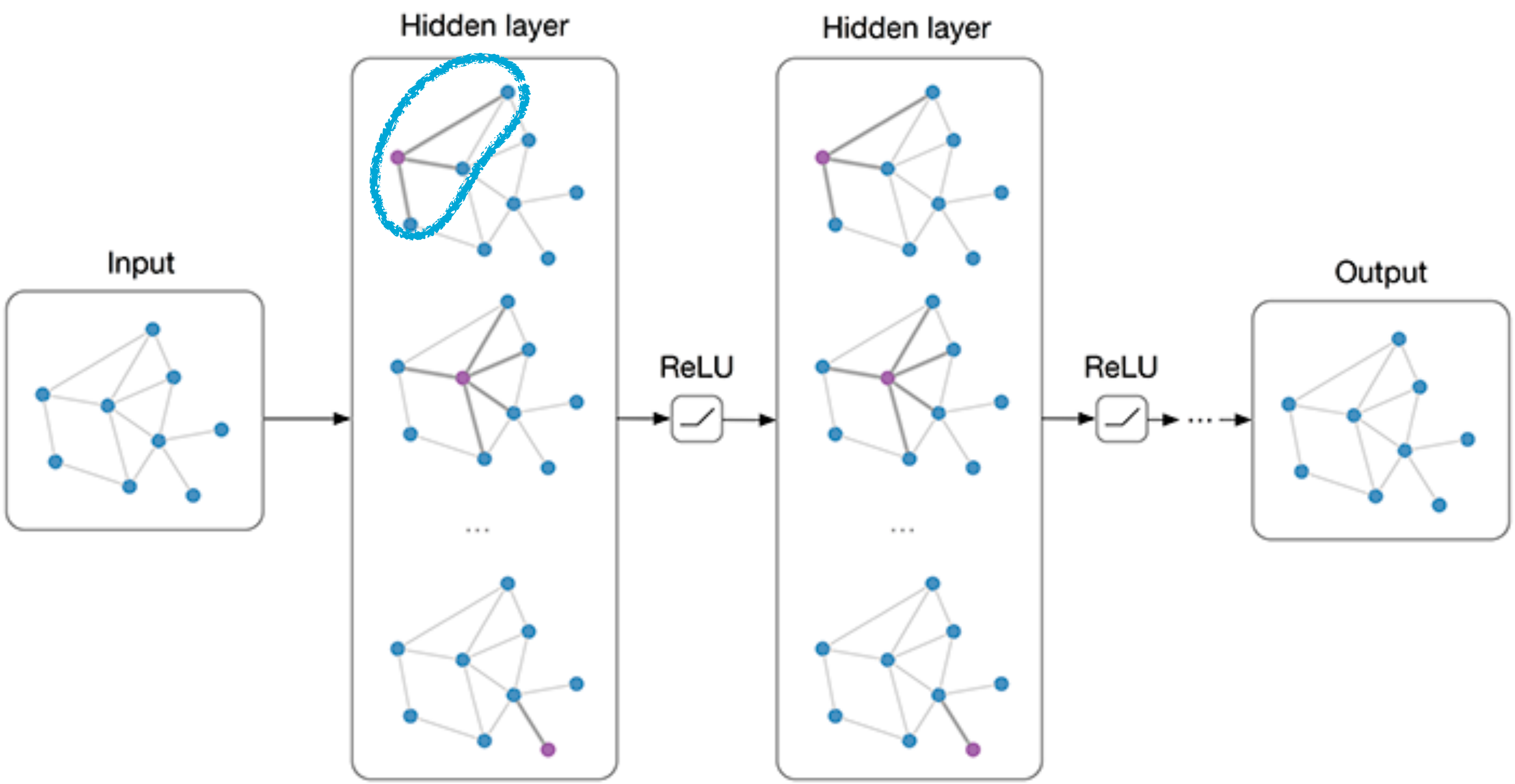


Image Source : https://tkipf.github.io/graph-convolutional-networks/

Computational Graph for GNNs

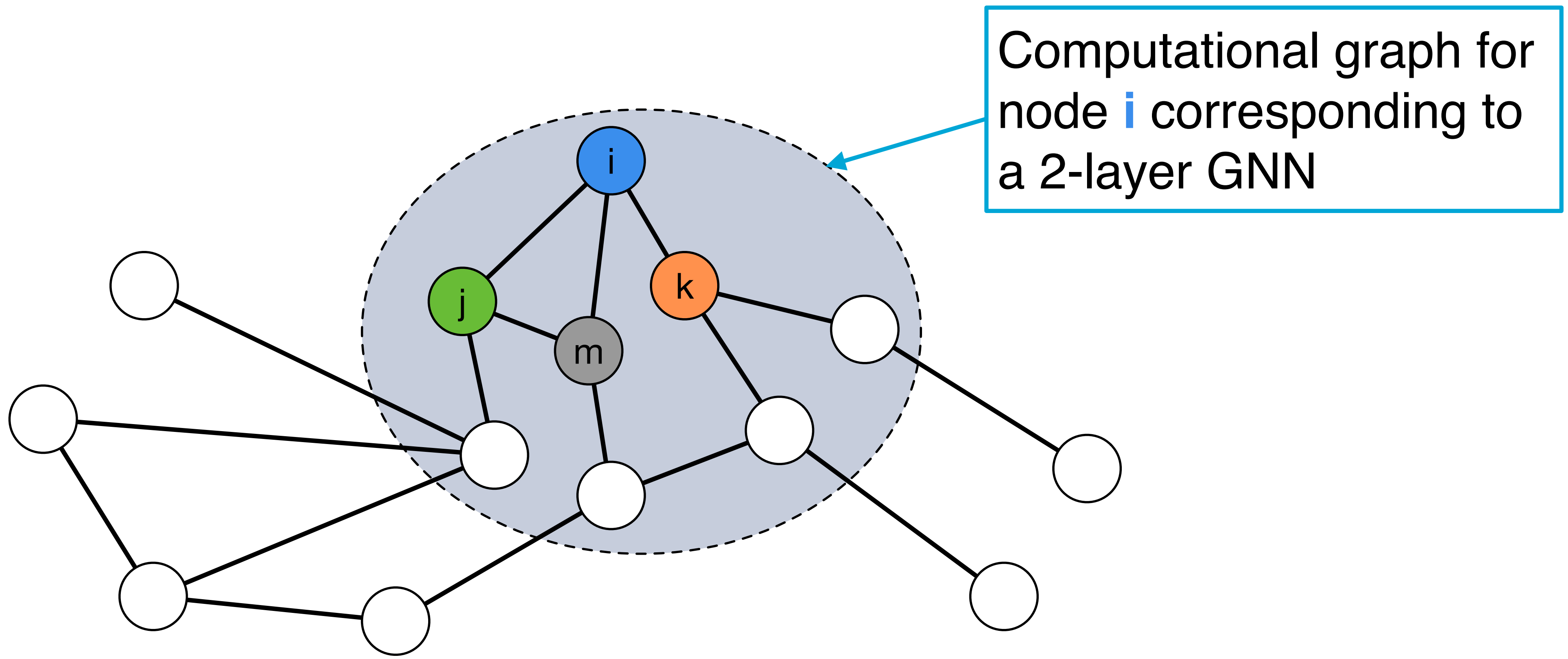


Image Source: [Lin et al., 2021]

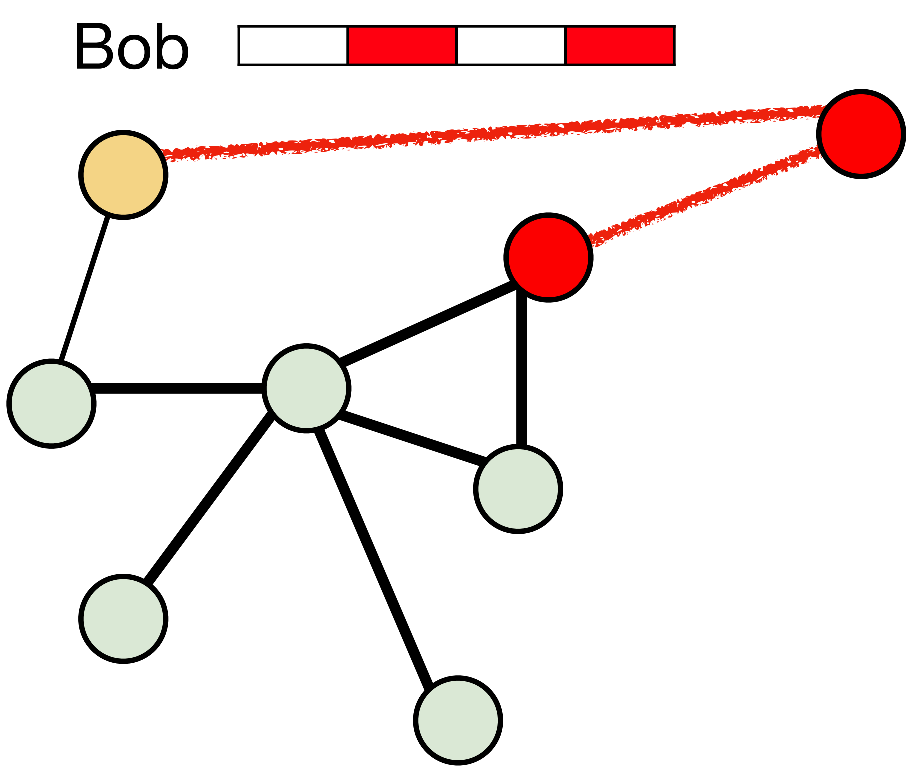
At inference time decision of a GNN on a particular node can be attributed to important nodes/edges and their features in its computational graph.

Explainable GraphML

Supervised GNNS

Why was a node/edge/graph assigned a particular label?

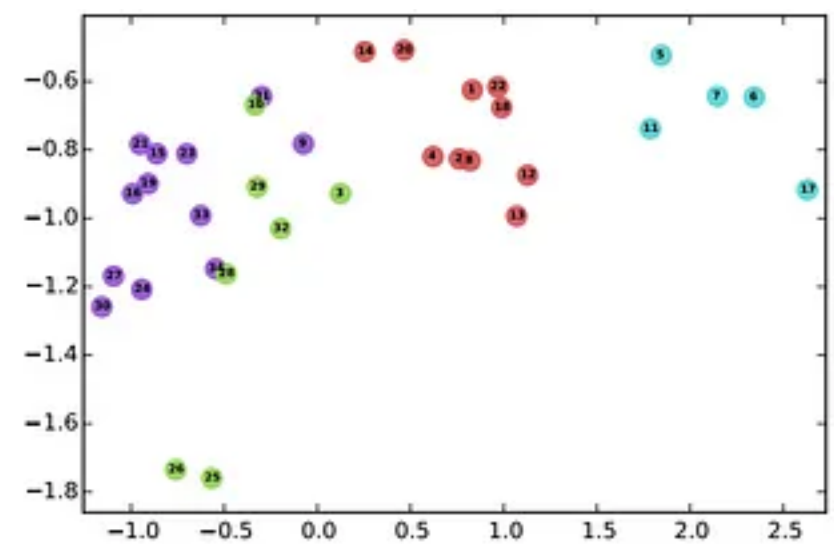
Decision has to be explained not only in terms of features but also graph structure. General explainability methods cannot be trivially applied for graphs.



When features are themselves uninterpretable?

<u>Unsupervised node embeddings</u>

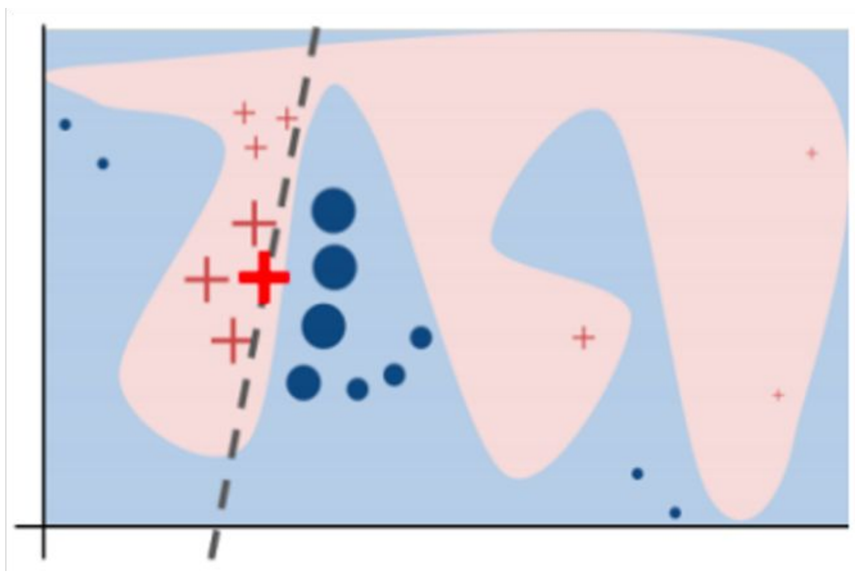
What do node embeddings encode?



No task information. Need to decode/explain embeddings in terms of input graph structure. **What should an explanation look like?**

Post hoc explanations Vs. Self explaining models

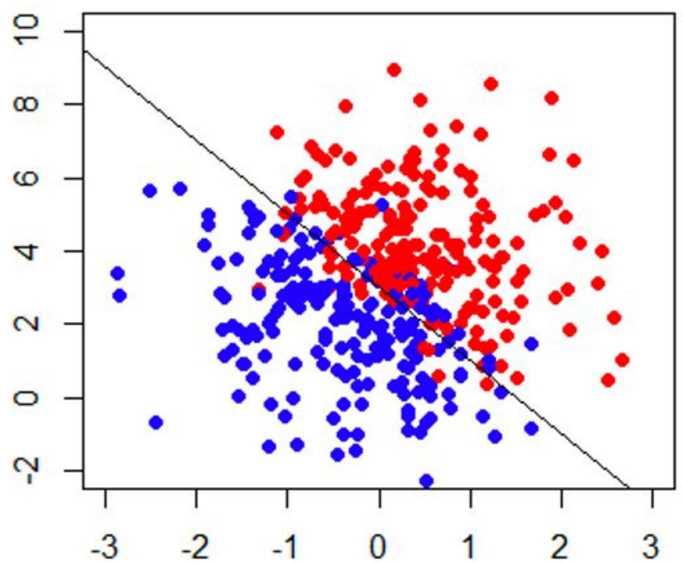
Post-hoc explanations



Explaining an already trained complex model
does not affect its performance

Explanations might not be faithful to the model

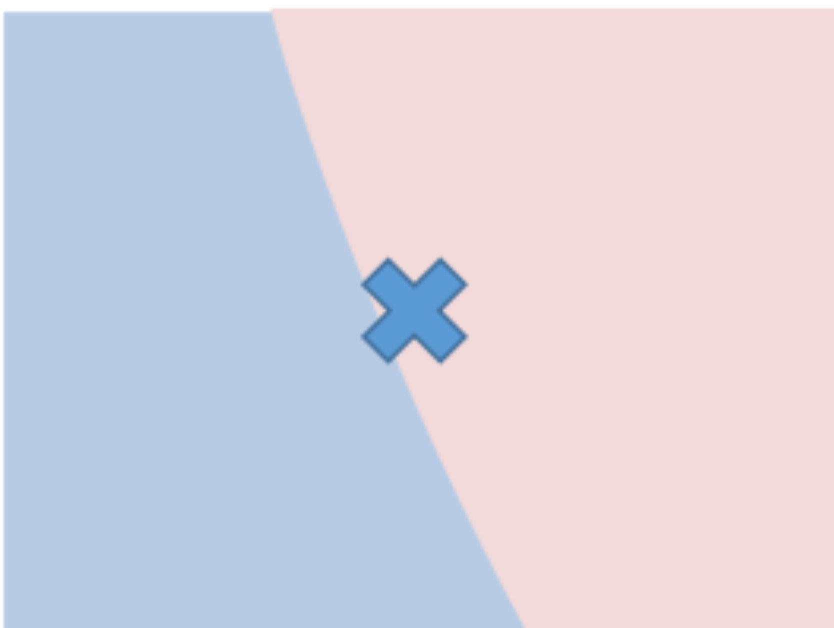
Self explaining models



There is usually a tradeoff between interpretability and performance

Explanations are by design faithful to the model

Local Vs Global Explanation



Vs.

Local or instance level explanations for explaining Individual predictions

Global explanations should ideally should explain complete model behaviour

Key Challenges

How to **define** explanations?

Uncover effect of various input elements in decision making

User of the explanation should be able to understand the explanation

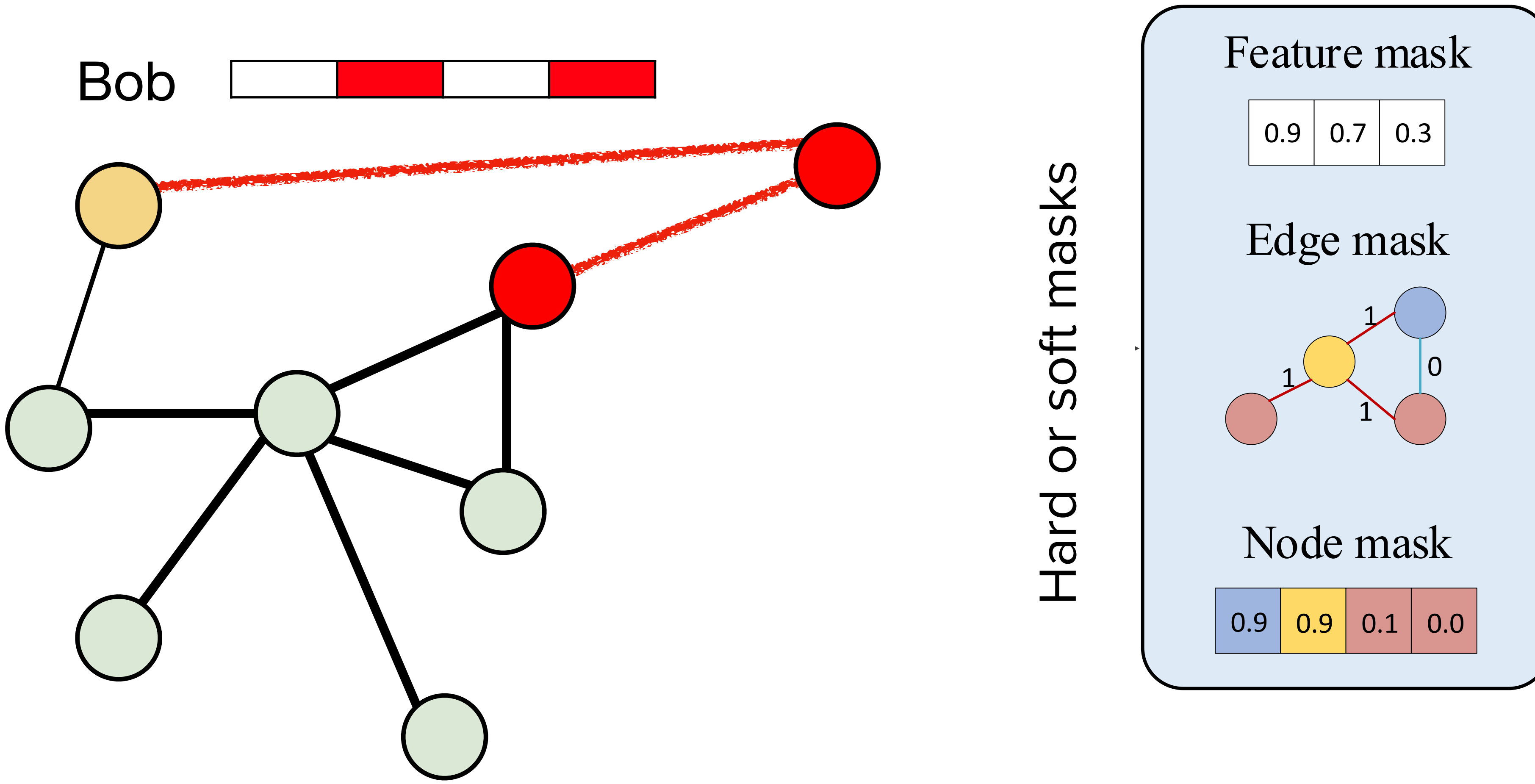
How to **evaluate** the **explainer** and the **explanations**?

Agreement with the decision logic of the model

Should be human understandable

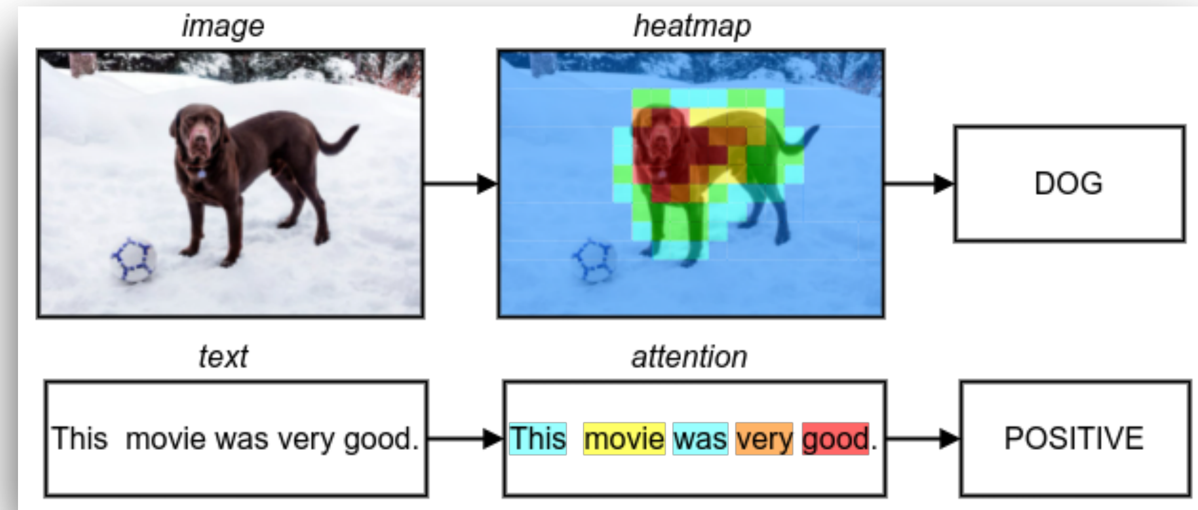
Post-hoc explanations for supervised GNNs

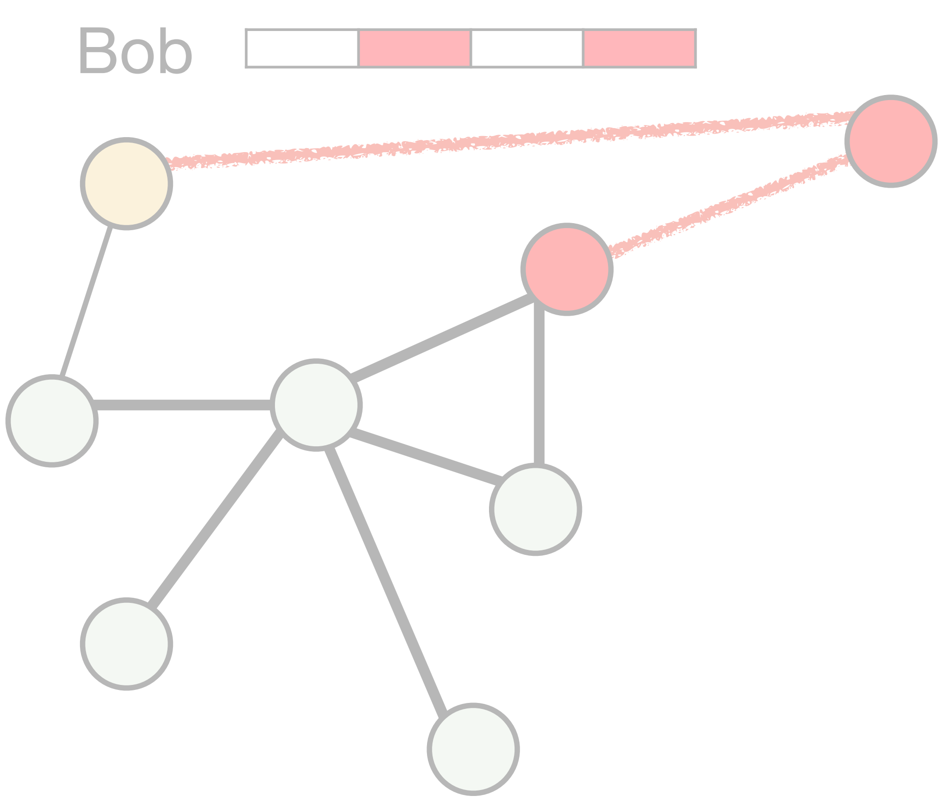
Substructure and feature importances

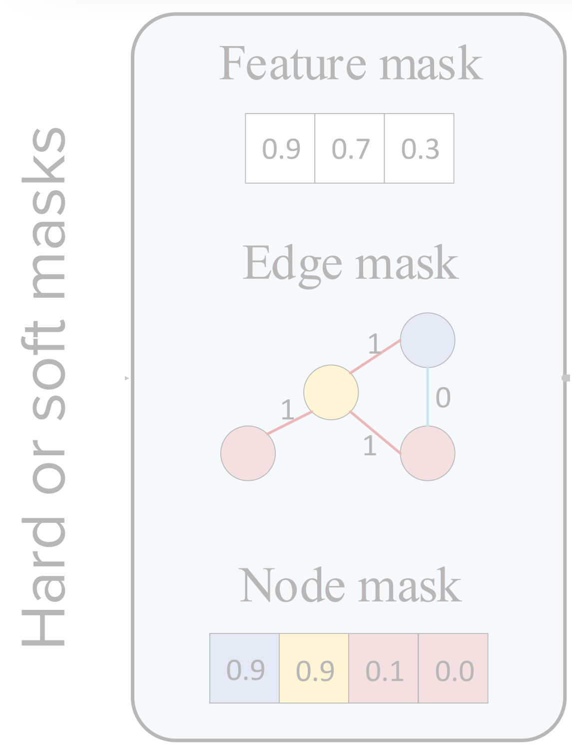


Substructures and subset of input features

Substructure and feature importances



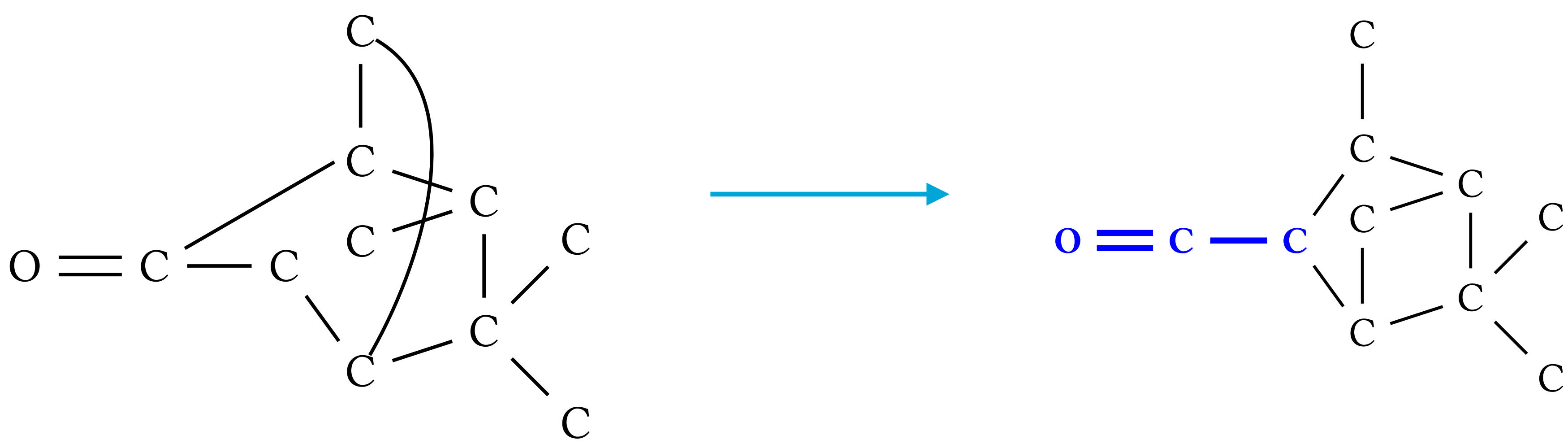




Substructures and subset of input features

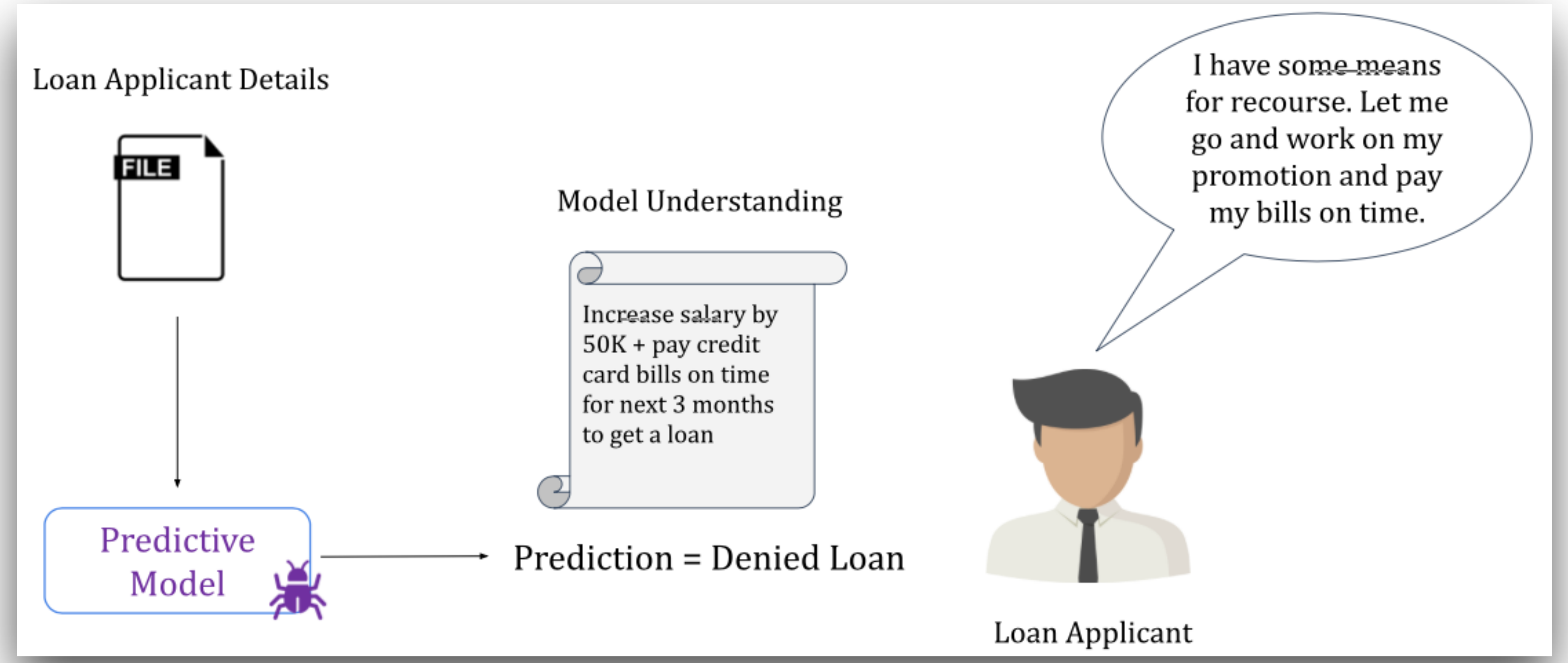
Counterfactuals

Smallest amount of perturbation on the input graph which result in change in GNN’s prediction

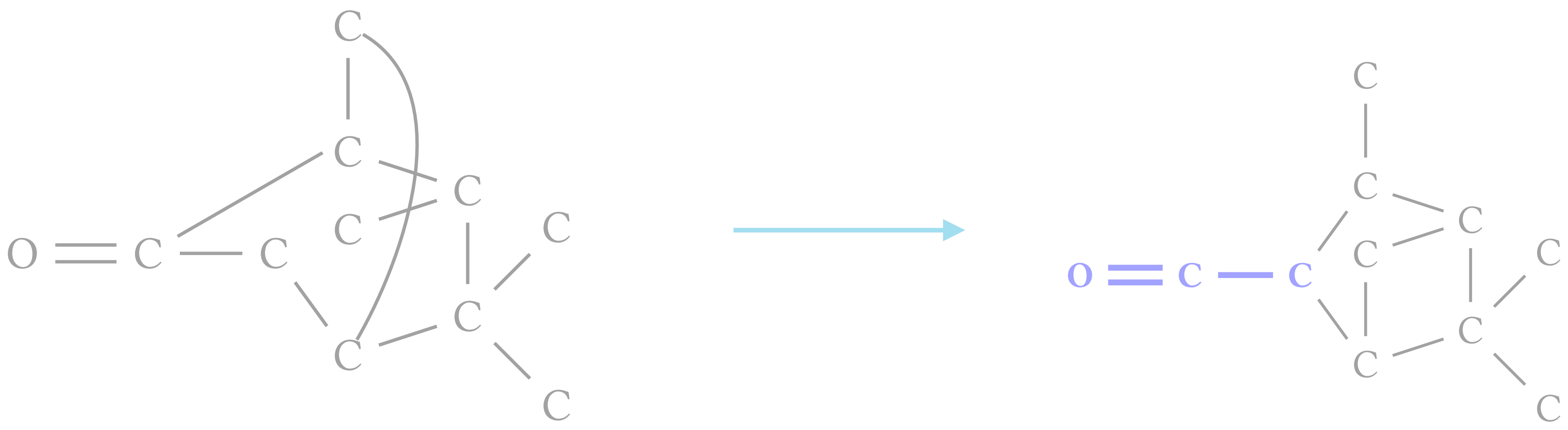


Recourse rules for improving molecules to combat HIV (Image adapted from Huang et al., 2023)

Counterfactuals



Smallest amount of perturbation on the input graph which result in change in GNN's prediction



Recourse rules for improving molecules to combat HIV (Image adapted from Huang et al., 2023)

Concepts

Concepts are small higher level units of information that can be interpreted by humans

Examples : motifs in graphs or specific properties like "node degree > 6" or "node next to carbon atom"

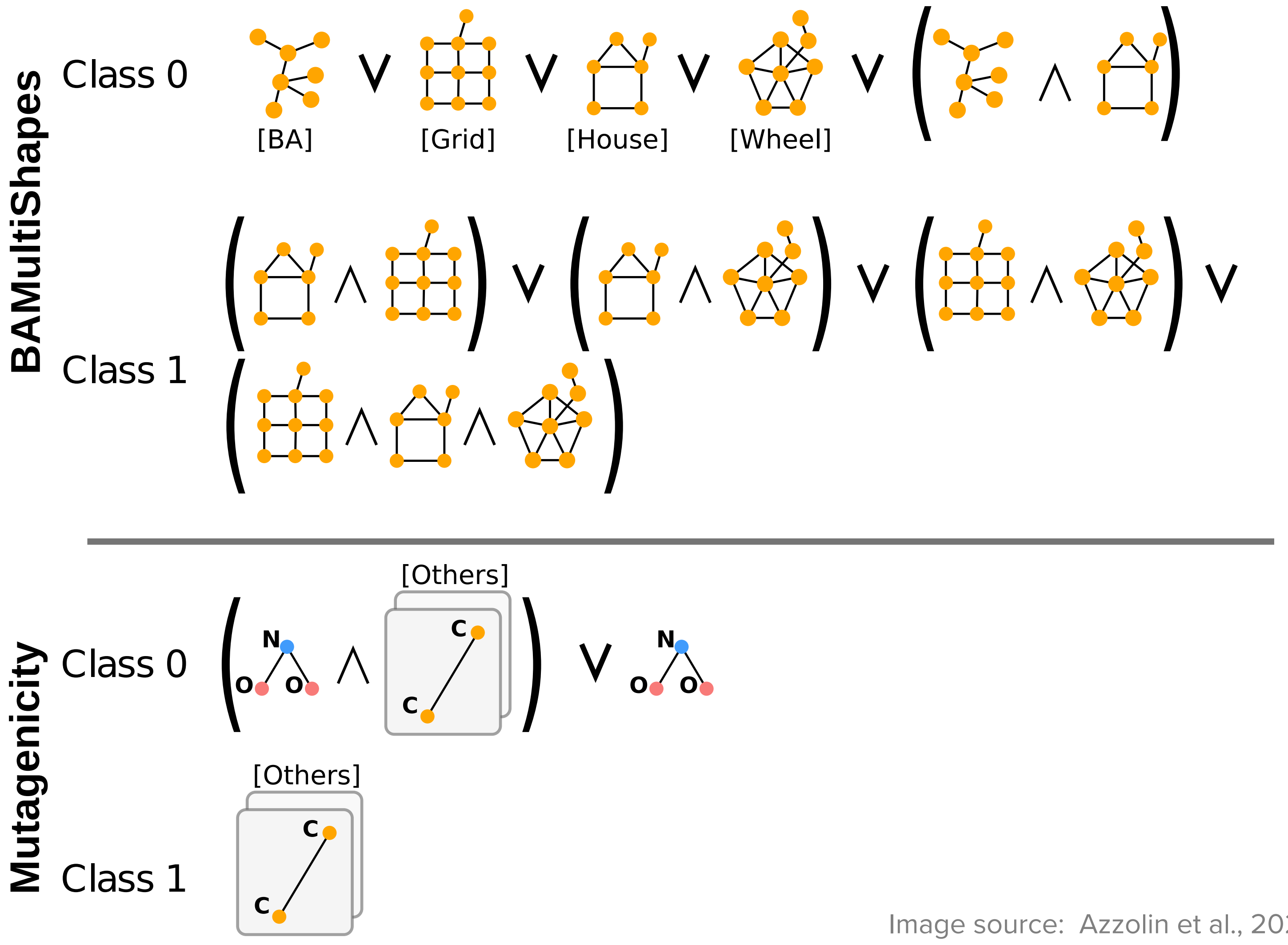
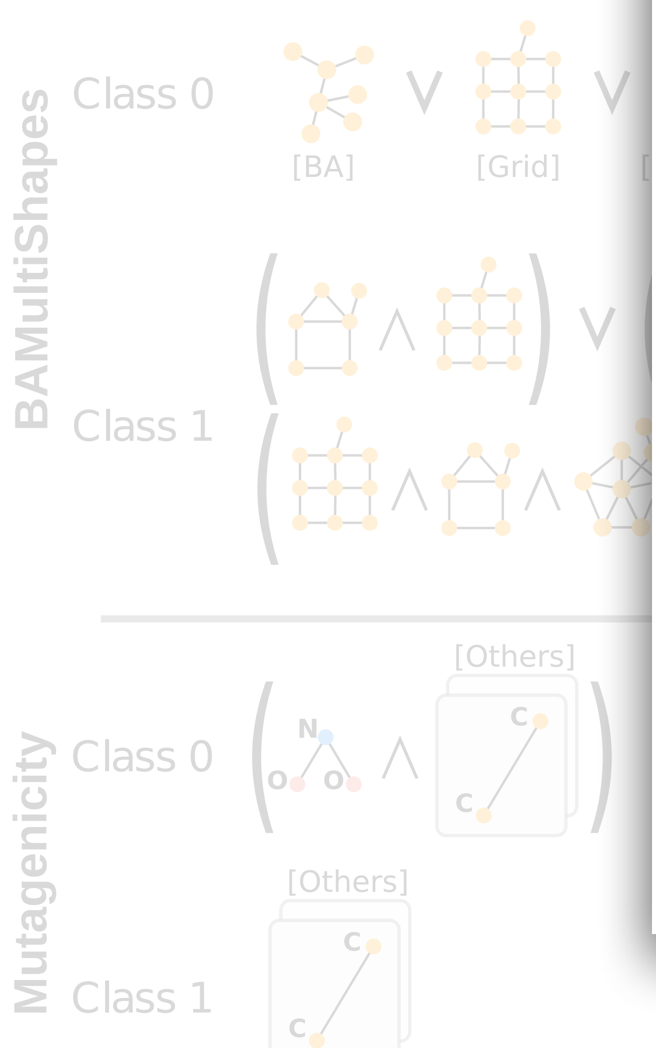


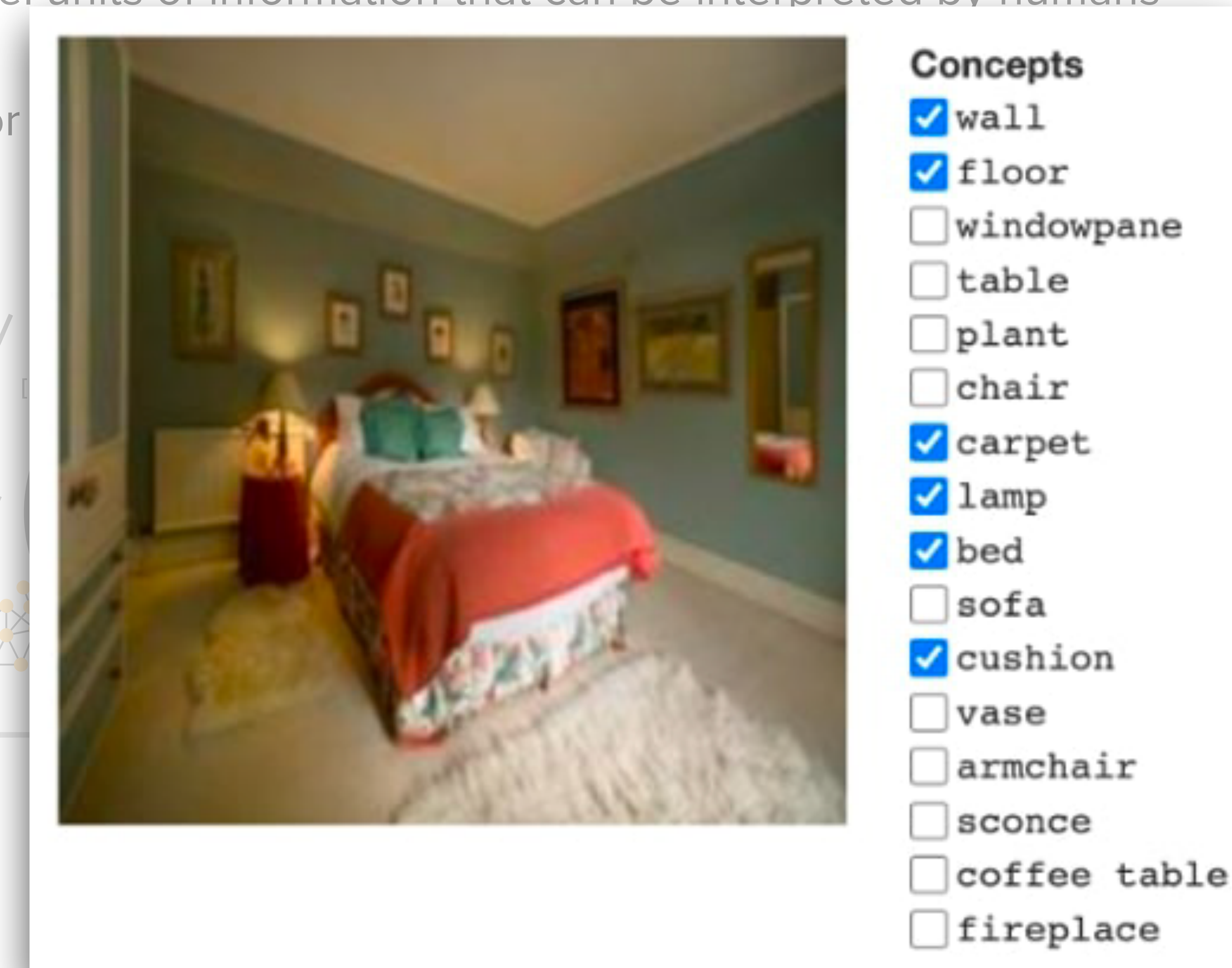
Image source: Azzolin et al., 2023

Concepts

Concepts are small higher level units of information that can be interpreted by humans

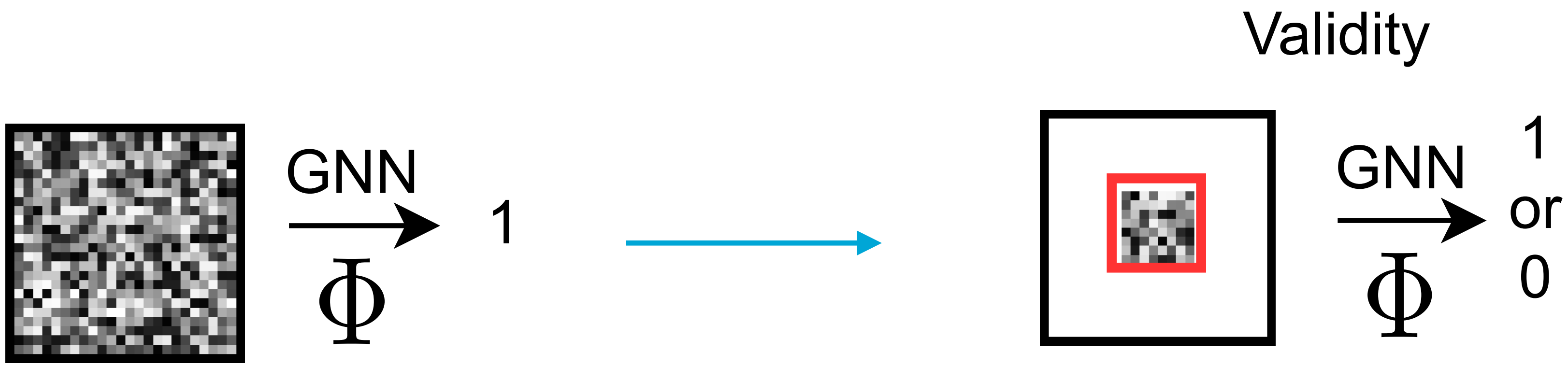
Examples : motifs in graphs or
carbon atom"





Substructure and feature explanations

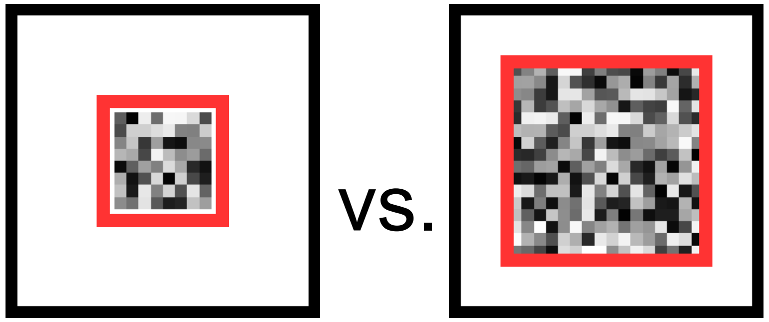
Valid Explanation



A **subset** of the input such that the prediction while just using the input stays the same as the original prediction is a **valid** explanation

Sparsity

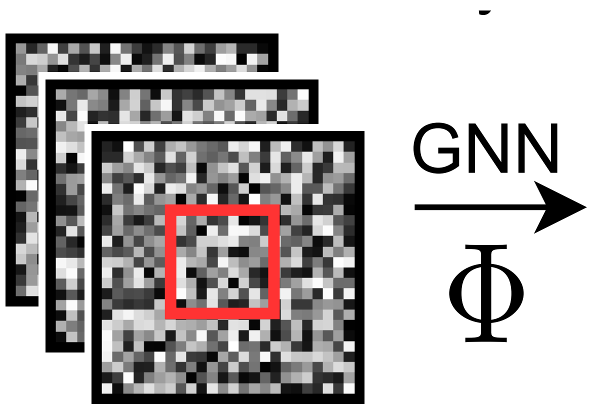
But a complete input is also a valid explanation



The chosen subset (explanation) should be sparse

Stability

What happens to the not selected part of the input?

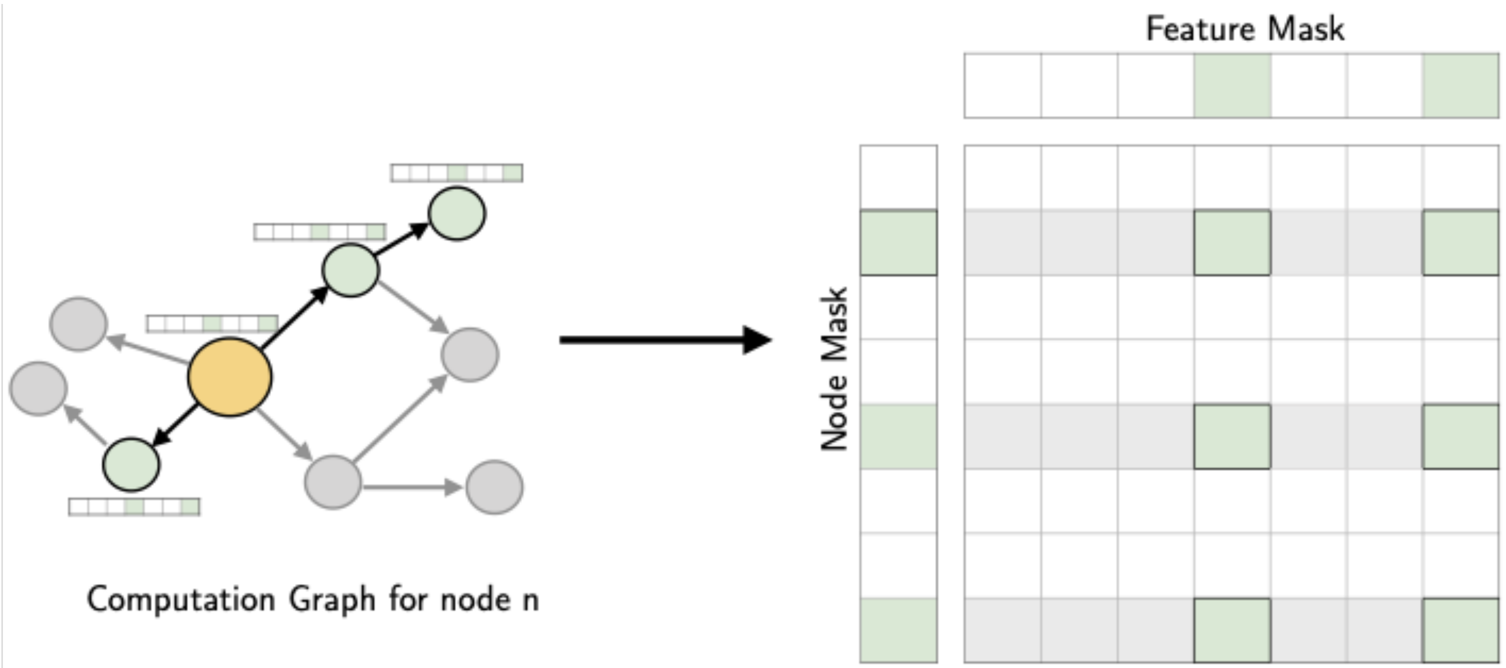


$\mathbb{E}$ 1

- Set the not selected part by some noisy values.
- Check the expected prediction over multiple such perturbations.

A stable explanation is one which achieves in expectation a close prediction to that of the original prediction

Constructing a perturbed input



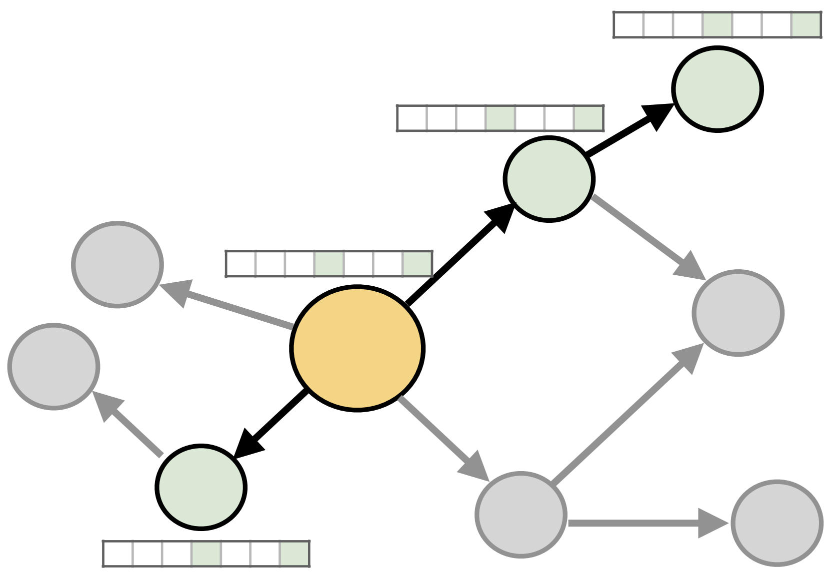
Selected nodes and features are marked green

Construct a perturbed input by setting selected features of selected nodes (the **green** cells) to their true values and others to random noisy values

If **M**(S) corresponds to product of feature and node masks, we obtain the perturbed input as

$$\mathbf{X}_S = \mathbf{X} \odot \mathbf{M}(S) + \mathbf{Z} \odot (\mathbf{1} - \mathbf{M}(S)), \mathbf{Z}_{ij} \sim \mathcal{N}$$

RDT-Fidelity of an explanation



Computation Graph for node n

$$F(S) = \mathbb{E}_{\mathbf{X}_S|Z\sim\mathcal{N}}\left[\mathbb{1}_{\Phi(\mathbf{X})=\Phi(\mathbf{X}_S)}\right]$$

with

$$\mathbf{X}_S = \mathbf{X} \odot \mathbf{M}(S) + \mathbf{Z} \odot (\mathbf{1} - \mathbf{M}(S)), \mathbf{Z}_{ij} \sim \mathcal{N}$$

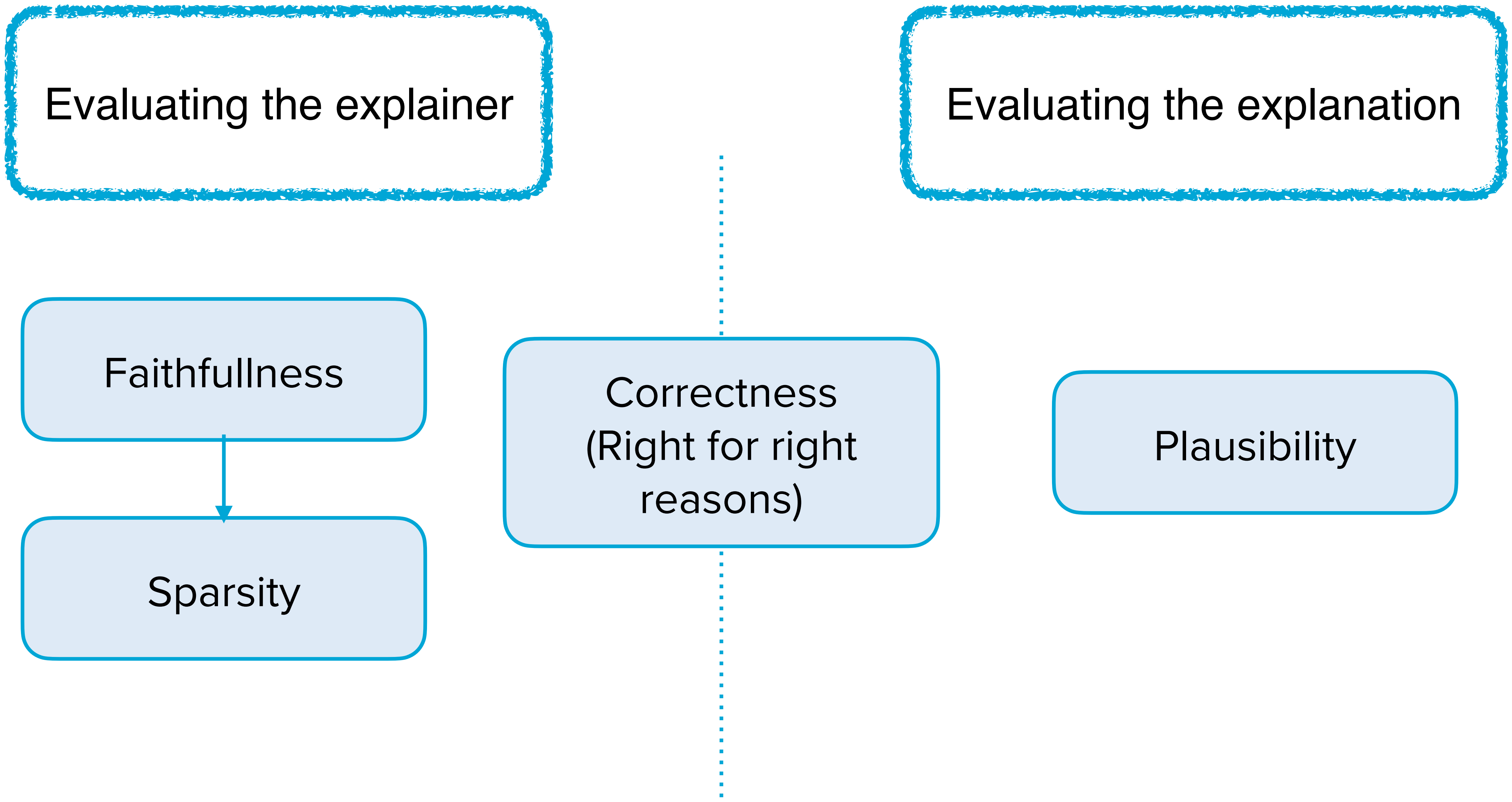
Zorro

Find the sparsest explanation such that its RDT-fidelity is maximised.

Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks**.** Funke, Khosla et al. TKDE 2022

Evaluating Post-Hoc Explanations

Evaluating Post-Hoc Explanations



[BAGEL Benchmark, Rathee et al. 2022]

https://github.com/Mandeep-Rathee/Bagel-benchmark

Faithfullness

Take 1: Check *sufficiency* and *comprehensiveness* of the explanation

Sufficiency

Keep the most important features/nodes/edges and check if they alone can predict the original decision.

Comprehensiveness

Remove the features/nodes/edges not in the explanation and check if the original prediction changes.

Faithfullness

How to compute sufficiency and comprehensiveness for soft masks?

What happens when you cannot remove features?

Take 2: Use RDT-Fidelity to check if the explanation is predictive and stable

$$F(S) = \mathbb{E}_{\mathbf{X}_S|Z\sim\mathcal{N}}\left[\mathbf{1}_{\Phi(\mathbf{X})=\Phi(\mathbf{X}_S)}\right]$$

Where

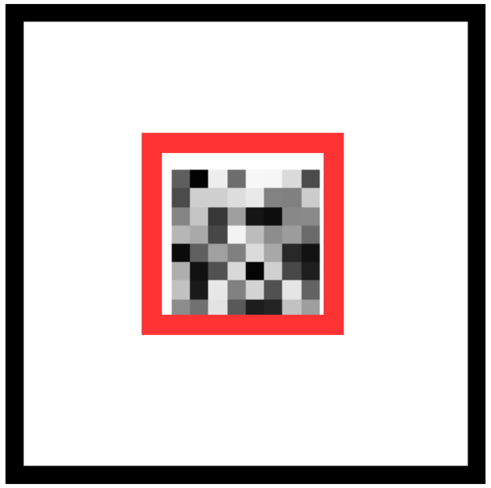
$$\mathbf{X}_S = \mathbf{X} \odot \mathbf{M}(S) + \mathbf{Z} \odot (\mathbf{1} - \mathbf{M}(S)), \mathbf{Z}_{ij} \sim \mathcal{N}$$

Sparsity

But the full input is also a faithful explanation

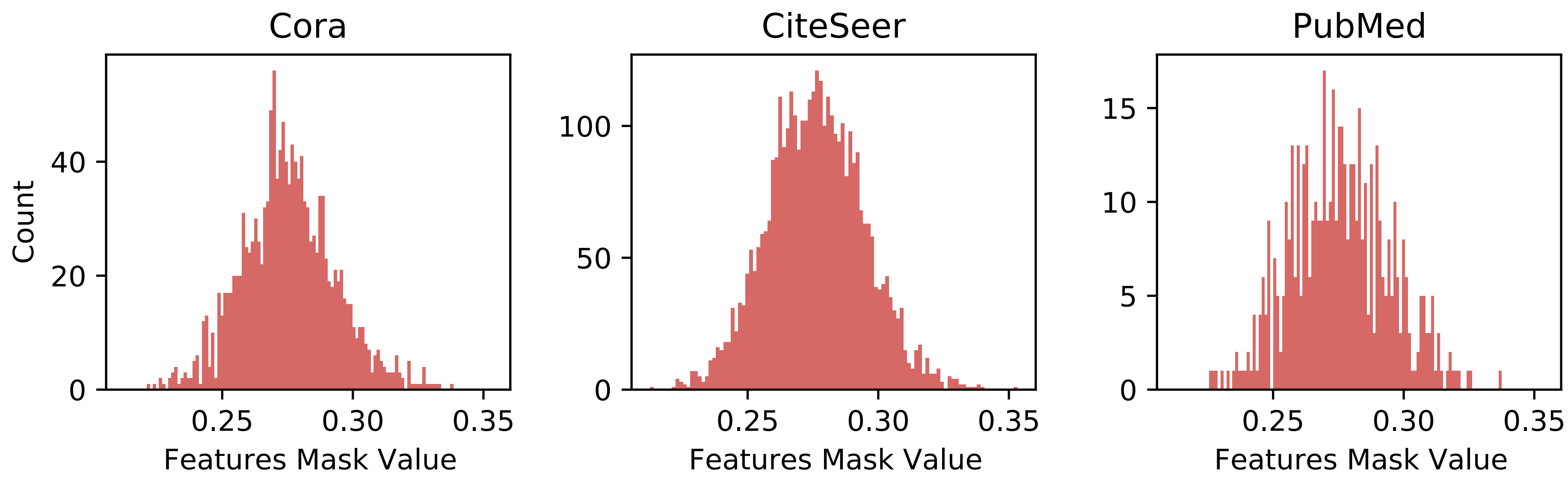
Are the explanations non-trivial?

Take 1: Sparsity for hard masks = Selection size / total



Sparsity

What about soft masks ?



A uniform distribution of normalised mask distribution implies complete input

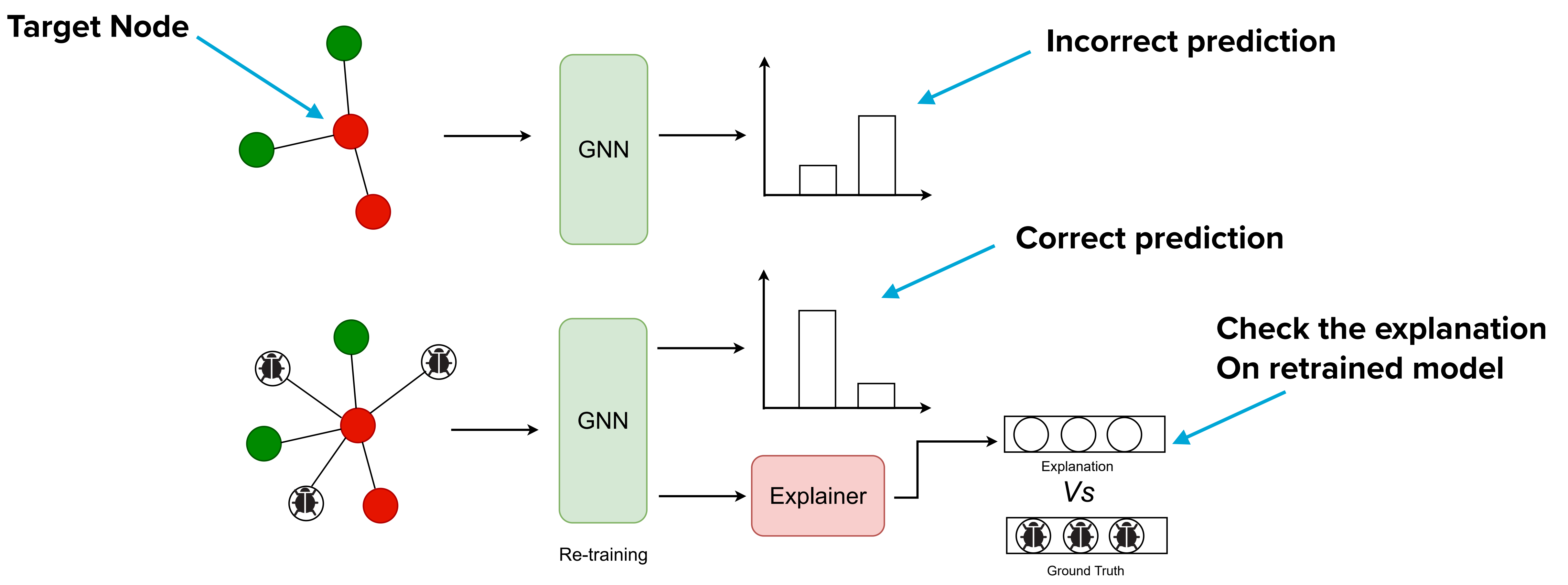
Take 2: Check Entropy of normalised distribution of masks

Lower the entropy sparser the explanation

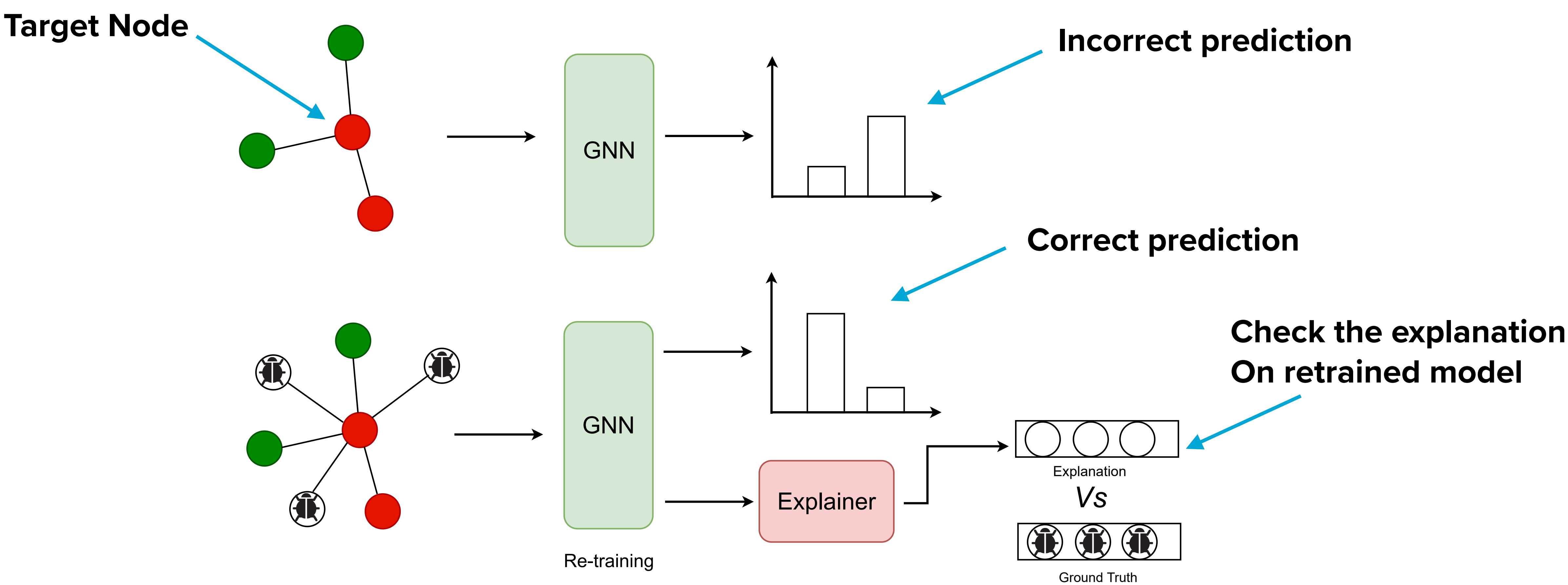
Correctness

Can the explainer detect any injected correlations responsible for altering model's behavior ?

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



Correctness



Drawbacks :

(i) Choosing correlations is tricky in the first place
(ii) Requires model retraining

Plausibility

How close are the explanations to human rationales ?

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...

Compute agreement of explanation with human rationales

Metrics : F1 score for hard masks, AUPRC score for soft masks

Plausibility

Should be used in conjunction with a suitable faithfulness metric

First ensure that the explanation is in fact approximating model's decision

GCN	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...
GAT	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...
APPNP	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...

Given the explainer is faithful to the model one can use plausibility to compare GNN models for the agreement of their decision making process with human rationales.

Other Evaluation schemes

Measuring agreement (explanation accuracy) with planted subgraph in a synthetic graph

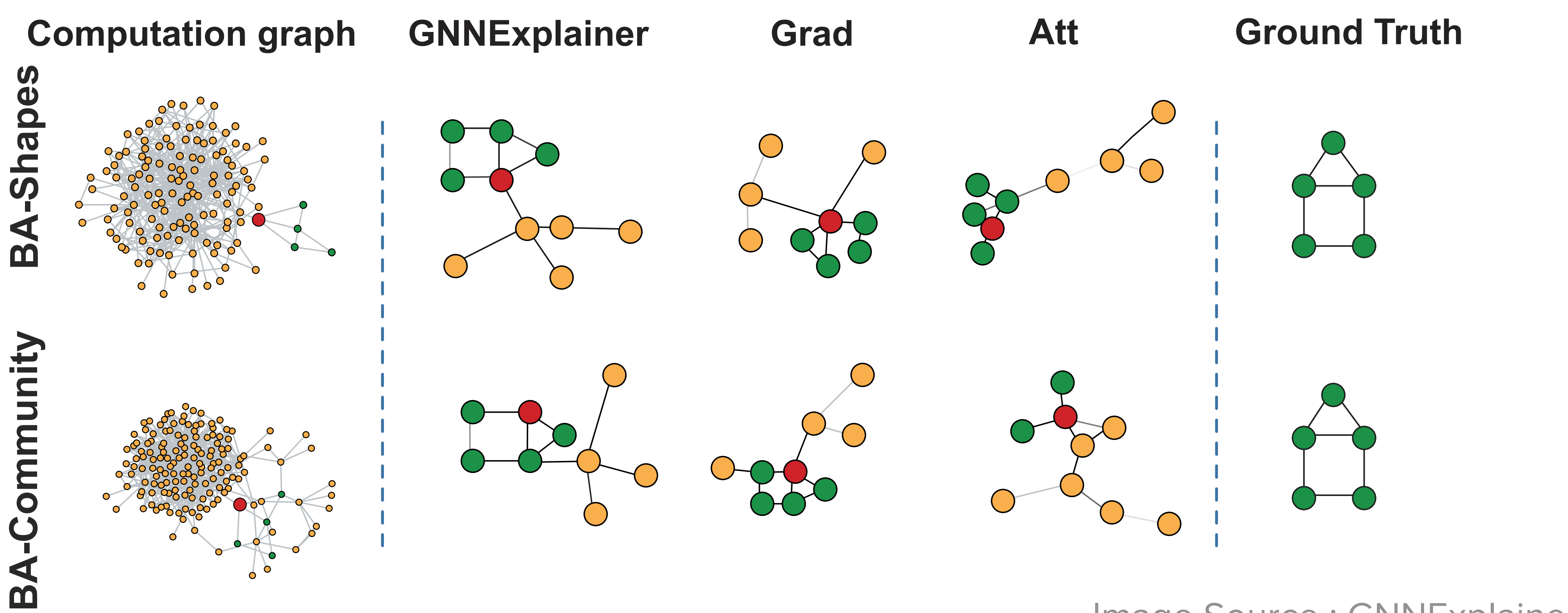


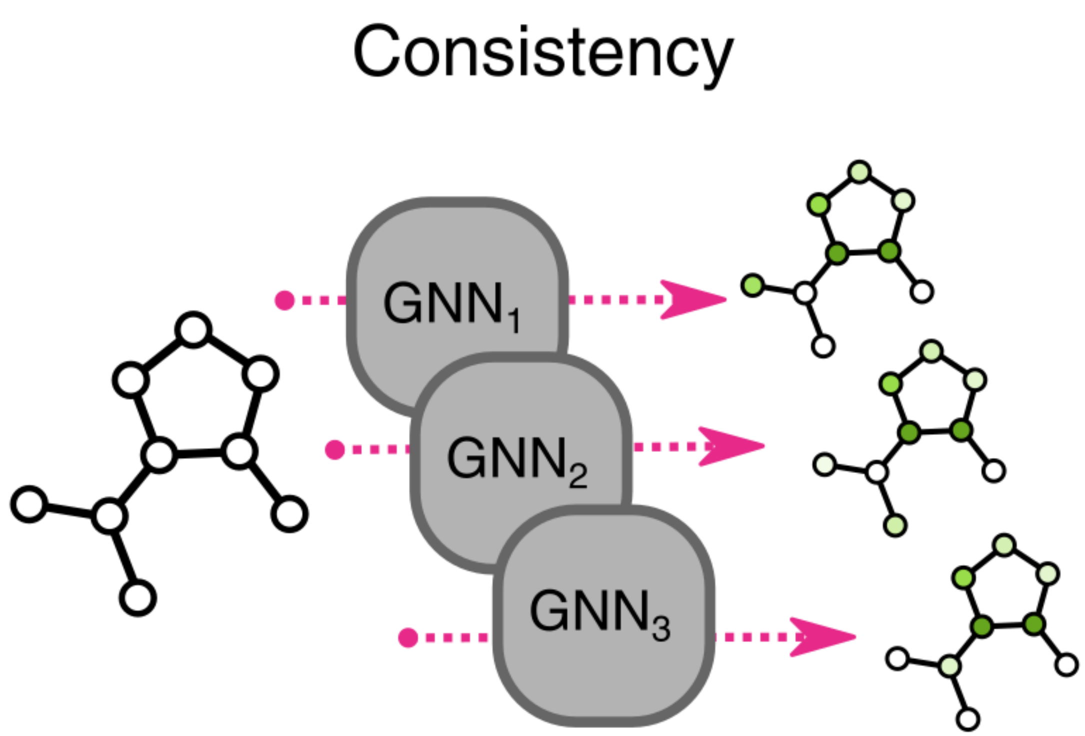
Image Source : GNNExplainer

Drawback/Issue : How to be sure if the model picked the planted subgraph?

Other Evaluation schemes

Measuring attribution (explanation) consistency across high performing models

[Sanchez-Lengeling et al. 2020]

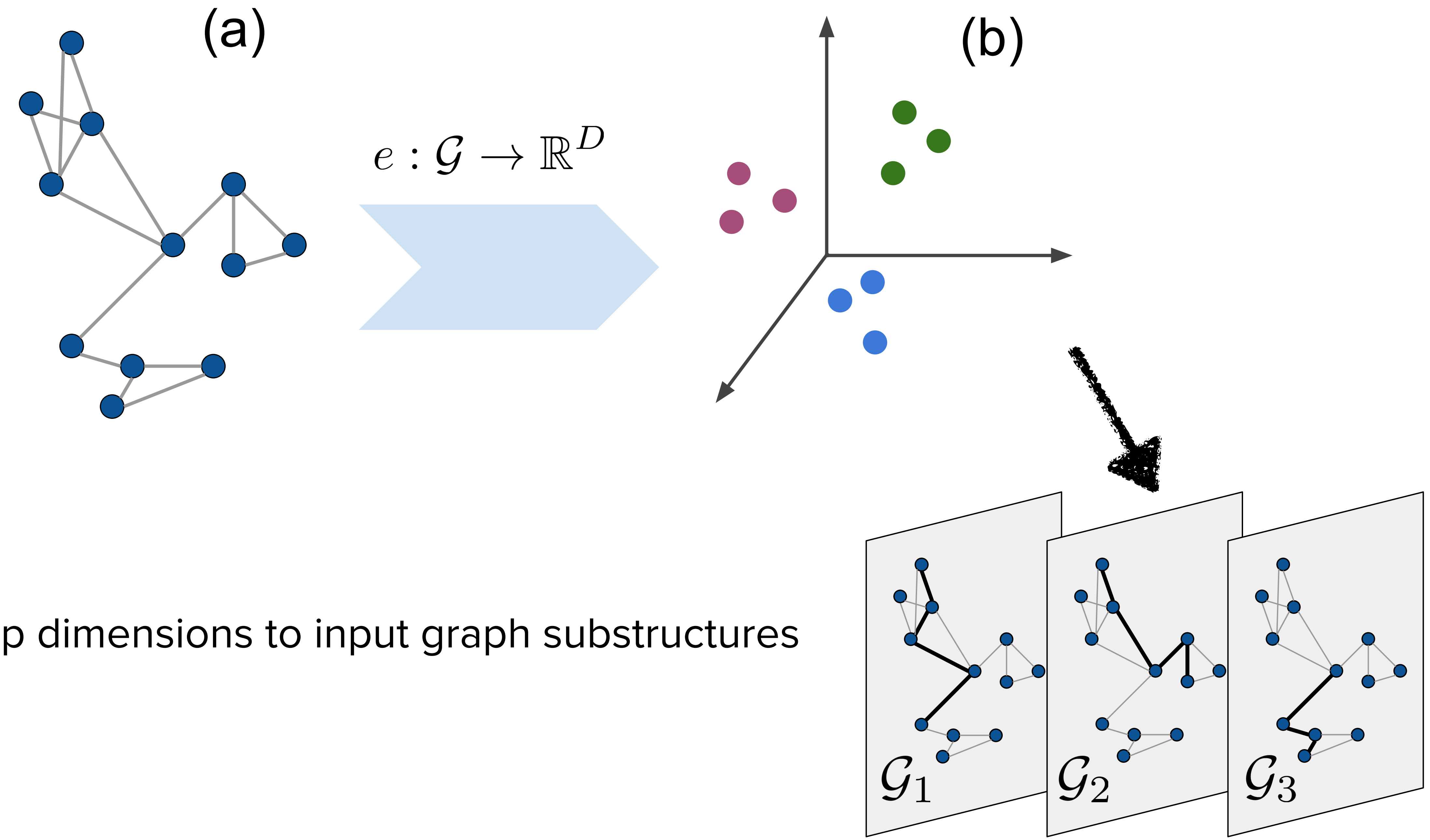


Quantifies the variability in explanation accuracy using the top 10% of models through a hyperparameter scan over model architectures

Drawback/Issue : How to be sure if the models used the intended explanation?

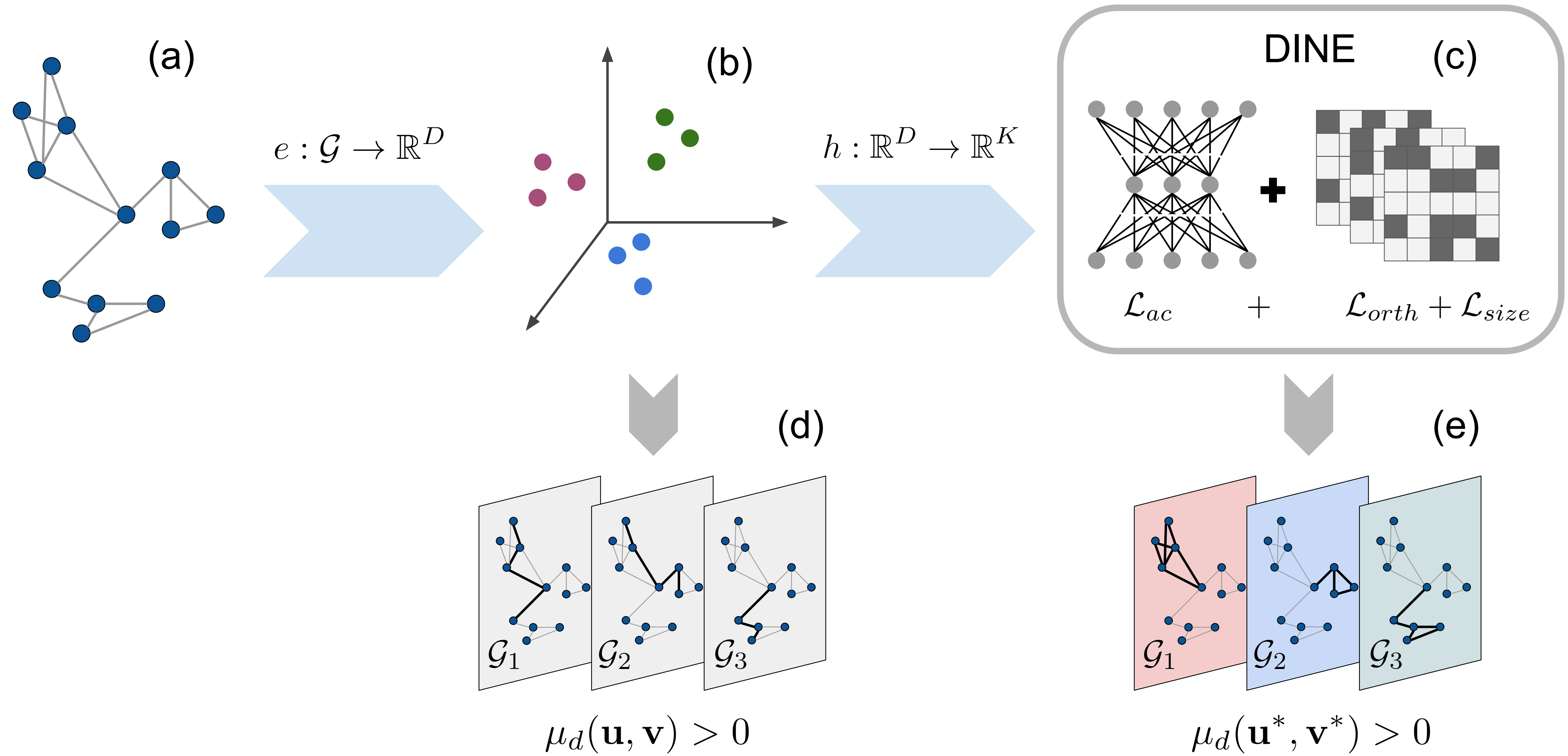
Explaining Node Embeddings

Global explanations for embedding dimensions



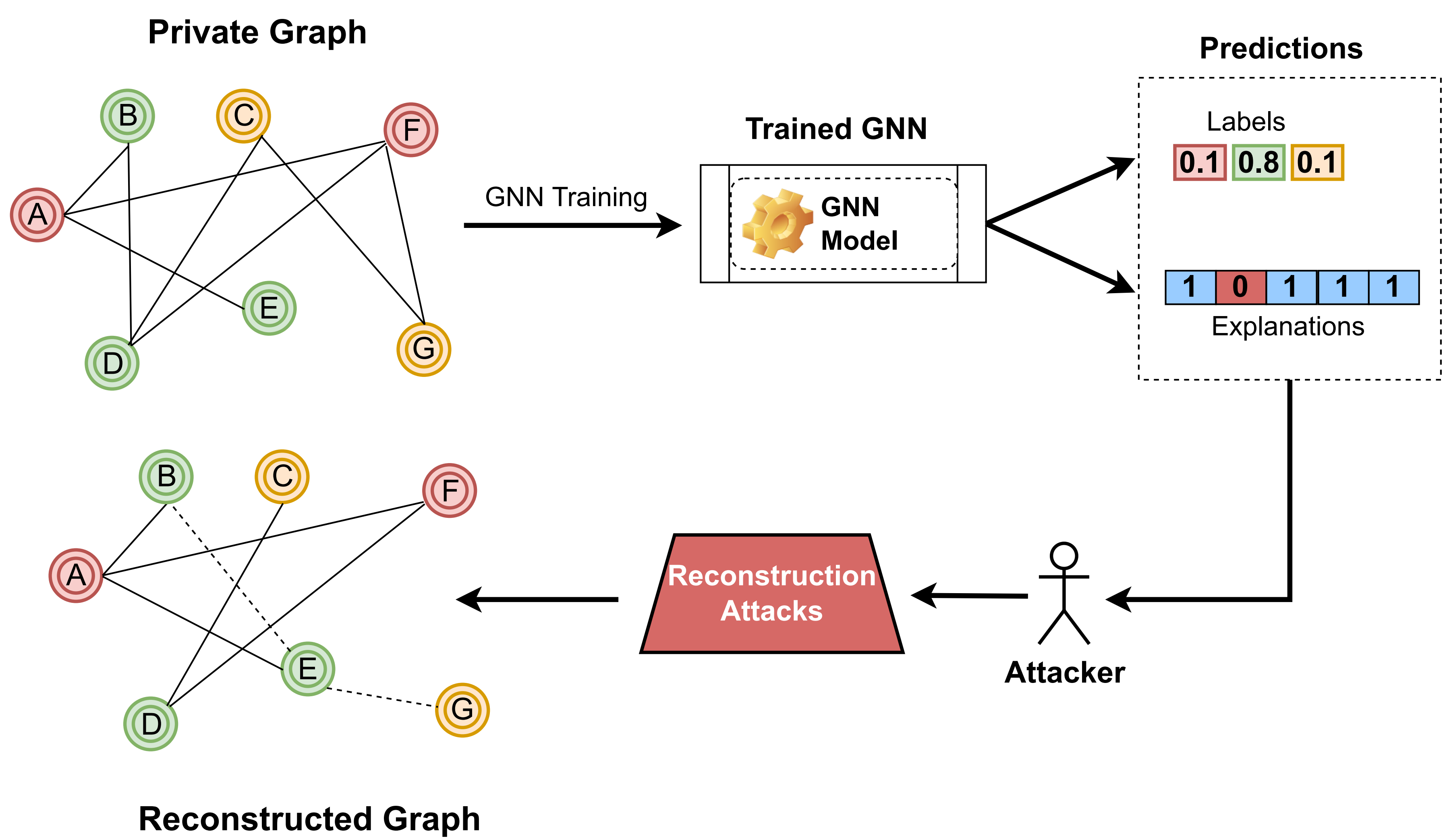
Map dimensions to input graph substructures

Global explanations for embedding dimensions



DINE: Dimensional Interpretability of Node Embeddings. *Piaggesi, Khosla et al. 2023.*

Explanations and privacy of training data



Private graph extraction via feature explanations. *Olatunji et al. PETS 2023*

Privacy and Transparency in Graph Machine Learning: A Unified Perspective. Khosla. AIMLAI 2022

Join us !

MLoG course together with Elvin Isufi

Participate in our workshop on Interplay of explainability and privacy in AI
on **8th and 9th February** 2024 in TU Delft