Trustworthy Machine Learning for Graph Data

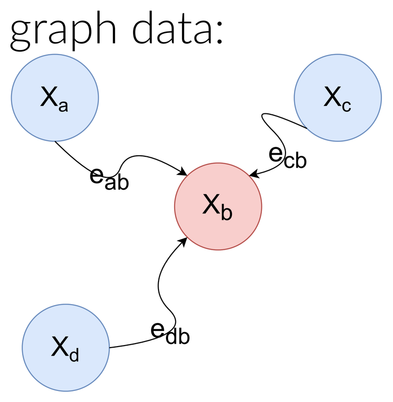
Steve Azzolin [1] Antonio Longa [1] Sagar Malhotra [2]
Stefano Teso [1] Andrea Passerini [1]

[1]University of Trento [2]TU Wien

Background

Graph Neural Networks (GNNs) are Neural Networks for graph data:

$$h_b^0 = x_b \in \mathbb{R}^d$$
$$h_b^l = Upd^l(h_b^{l-1}, Aggr^l(\{\{h_u^{l-1} : u \in N(b)\}\}))$$



Where *Upd* is a Neural Network and *Aggr* is any permutation invariant function.

Motivation

GNNs lack interpretability, thus hindering understanding, debugging, and human trust:

Figure 1. Aristolochic acids are a family of carcinogenic, mutagenic, and nephrotoxic phytochemicals commonly found in the flowering plant family Aristolochiaceae.

As popular post-hoc methods have been found to fall short in reliably explaining trained GNNs [5, 4], new **explainable by-design** architectures have been proposed:

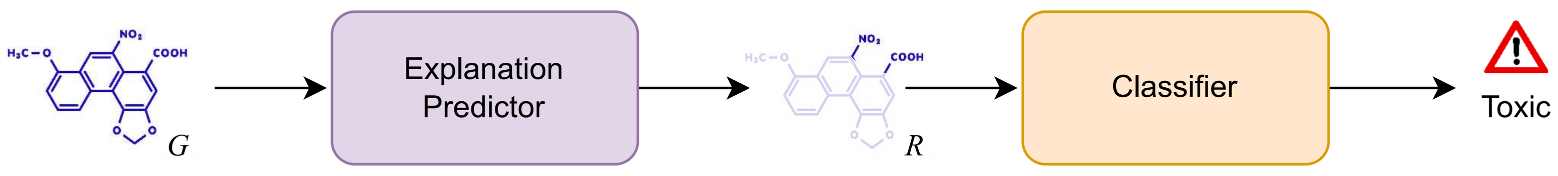


Figure 2. Pipeline of **Self-Explainable GNNs** (SEGNNs).

⚠️ Nonetheless, some SEGNNs are found to be more **faithful** to random explanations than to their true explanations [3]. ⚠️

Our contribution

We aim to study the root of this issue while providing insights into how to build more reliable SEGNNs:

- **RQ1**: What characterizes a good explanation?
- **RQ2**: How good are SEGNNs?
- **RQ3**: Can we go beyond subgraph-based explanations?

RQ1: What characterizes a good explanation? [1]

Current literature measures *how much the model adheres to its explanation* by measuring the **faithfulness** of explanations:

- *sufficient*, i.e., keeping it fixed shields the model's output from changes to its complement $C = G \setminus R$
$$SUF_{d,p_R}(R) = \mathbb{E}_{G' \sim p_R}[\Delta_d(G, G')],$$
- *necessary*, i.e., altering it affects the model's output even with C fixed
$$NEC_{d,p_C}(R) = \mathbb{E}_{G' \sim p_C}[\Delta_d(G, G')]$$

ℹ️ We provide a taxonomy of the current faithfulness metric:

Table 1. *SUF* and *NEC* recover existing faithfulness metrics for appropriate choices of divergence d and interventional distributions p_R and p_C.

Metric	Estimates	Divergence d	Allowed changes
Unf	Suf	$\mathrm{KL}(p_\theta(\cdot \mid G), p_\theta(\cdot \mid G'))$	zero out all irrelevant features
Fid-		$\lvert p_\theta(\hat{y} \mid G) - p_\theta(\hat{y} \mid G')\rvert$	zero out all irrelevant features, delete all irrelevant edges
RFid-		"	delete a random subset of irrelevant edges
PS		$\mathbb{1}\{p_\theta(\hat{y} \mid G) = p_\theta(\hat{y} \mid G')\}$	multiply all irrelevant elements by relevance scores
Fid+	Nec	$\lvert p_\theta(\hat{y} \mid G) - p_\theta(\hat{y} \mid G')\rvert$	zero out all relevant features, delete all relevant edges
RFid+		"	delete a random subset of relevant edges
PN		$\mathbb{1}\{p_\theta(\hat{y} \mid G) \neq p_\theta(\hat{y} \mid G')\}$	multiply all relevant elements by relevance scores

ℹ️ Metrics are not interchangeable in the sense that metric values across different metric parameters are not consistent.

Table 2. Model ranking and *SUF* results according to different p_R.

Split	Model	Motif2 $p_R^{id_1}$	Motif2 $p_R^{id_2}$
ID	LECI	1 (81 $\pm$ 03)	2 (82 $\pm$ 03)
	GSAT	2 (78 $\pm$ 01)	1 (84 $\pm$ 02)
	CIGA	3 (65 $\pm$ 07)	3 (73 $\pm$ 06)

RQ1: What characterizes a good explanation? (cont.) [1]

ℹ️ Previous *Necessity* metrics do not penalize useless explanations

ℹ️ We propose a new necessity metric that penalizes overly large explanations

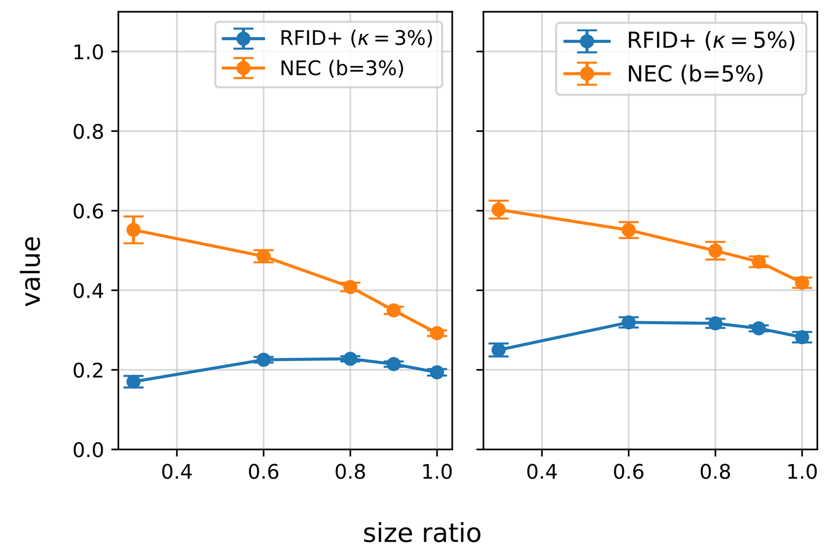


Figure 3. Our proposed **Nec** is sensitive to the number of irrelevant items in the explanation, whereas **RFid+** is not.

RQ2: How good are SE-GNNs? [1]

We identified some architectural design choices favoring **un-faithfulness** and fixed them:

- **Hard Scores** (HS): give exact zero importance to information outside of R;
- **Explanation Readout** (ER): aggregate only over R for the final prediction.

Table 3. Test set accuracy and faithfulness of some augmented SE-GNNs.

Dataset	BaMS Acc	BaMS Faith	Motif2 Acc	Motif2 Faith	Motif-Size Acc	Motif-Size Faith	BBBP Acc	BBBP Faith
GSAT	100 $\pm$ 00	35 $\pm$ 03	92 $\pm$ 01	61 $\pm$ 01	90 $\pm$ 01	60 $\pm$ 02	79 $\pm$ 04	27 $\pm$ 08
GSAT + ER	100 $\pm$ 00	35 $\pm$ 03	92 $\pm$ 01	63 $\pm$ 01	90 $\pm$ 01	65 $\pm$ 01	80 $\pm$ 02	33 $\pm$ 04
GSAT + HS	98 $\pm$ 01	21 $\pm$ 06	53 $\pm$ 02	24 $\pm$ 05	54 $\pm$ 03	22 $\pm$ 05	71 $\pm$ 01	31 $\pm$ 09
GSAT + ER + HS	99 $\pm$ 01	24 $\pm$ 04	57 $\pm$ 04	37 $\pm$ 03	56 $\pm$ 07	29 $\pm$ 09	73 $\pm$ 02	33 $\pm$ 02
GISST	100 $\pm$ 00	25 $\pm$ 03	92 $\pm$ 01	53 $\pm$ 02	92 $\pm$ 00	50 $\pm$ 02	84 $\pm$ 03	23 $\pm$ 11
GISST + ER	–	–	–	–	–	–	85 $\pm$ 06	27 $\pm$ 06
GISST + HS	–	–	–	–	–	–	83 $\pm$ 05	19 $\pm$ 07
GISST + ER + HS	–	–	–	–	–	–	81 $\pm$ 07	15 $\pm$ 09
RAGE	96 $\pm$ 01	33 $\pm$ 05	83 $\pm$ 02	64 $\pm$ 04	74 $\pm$ 09	63 $\pm$ 07	82 $\pm$ 01	33 $\pm$ 04
RAGE + ER	96 $\pm$ 02	33 $\pm$ 02	85 $\pm$ 06	66 $\pm$ 03	71 $\pm$ 09	55 $\pm$ 07	84 $\pm$ 01	33 $\pm$ 05
RAGE + HS	97 $\pm$ 01	46 $\pm$ 03	85 $\pm$ 01	65 $\pm$ 02	78 $\pm$ 07	65 $\pm$ 09	84 $\pm$ 02	46 $\pm$ 02
RAGE + ER + HS	96 $\pm$ 01	46 $\pm$ 04	83 $\pm$ 04	64 $\pm$ 04	75 $\pm$ 08	62 $\pm$ 12	82 $\pm$ 01	43 $\pm$ 03

RQ3: Beyond subgraph-based explanations [2]

Theorem: *Given a classifier g expressible as a purely existentially quantified first-order logic formula and a positive instance G of any size, then a Trivial Explanation for $g(G)$ is also a Prime Implicant explanation for $g(G)$.*

ℹ️ Subgraph-based explanations are *optimal* for motif-based tasks;

😥 But we do not know when we are explaining motif-based tasks;

😌 **Enhance standard SEGNNs with an interpretable side channel** and let the optimization pick the best alternative (Occam's razor).

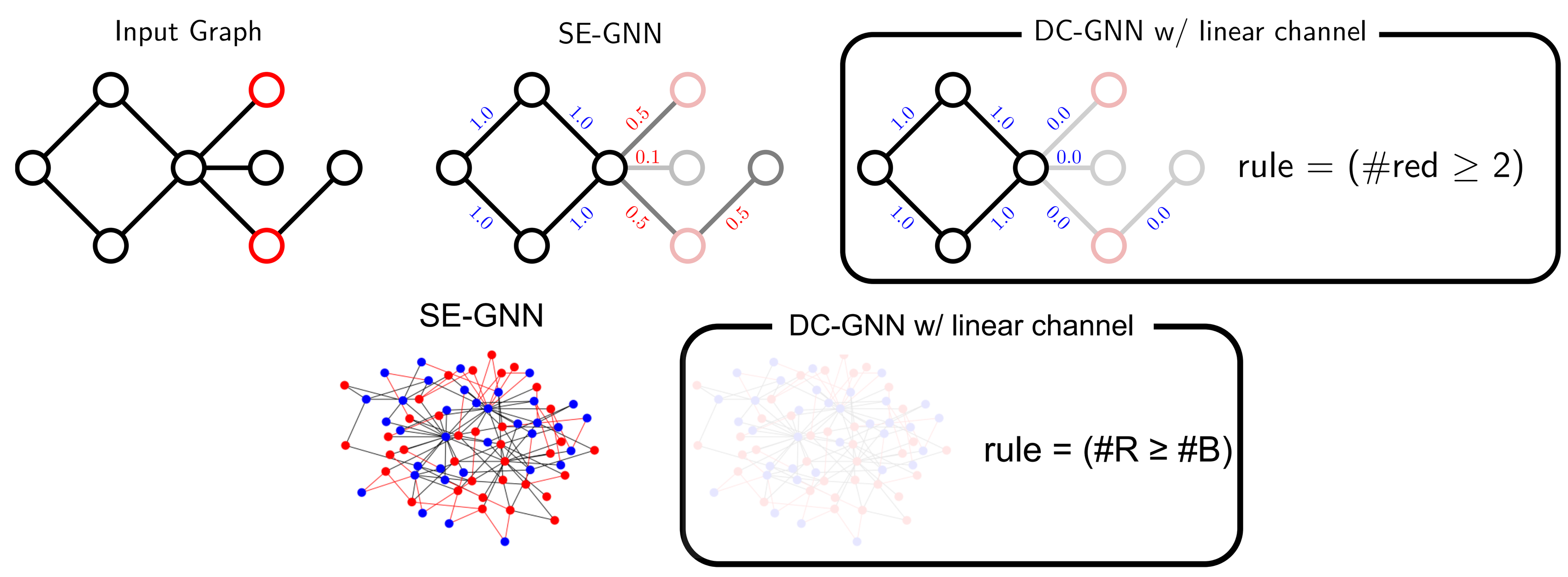


Figure 4. Examples of the proposed Dual-Channel SEGNN.

References

[1] Steve Azzolin, Antonio Longa, Stefano Teso, and Andrea Passerini. Reconsidering faithfulness in regular, self-explainable and domain invariant GNNs. 2025.

[2] Steve Azzolin, Sagar Malhotra, Andrea Passerini, and Stefano Teso. Beyond topological self-explainable gnns: A formal explainability perspective, 2025.

[3] Marc Christiansen, Lea Villadsen, Zhiqiang Zhong, Stefano Teso, and Davide Mottin. How faithful are self-explainable gnns? 2023.

[4] Zhong Li, Simon Geisler, Yuhang Wang, Stephan Günnemann, and Matthijs van Leeuwen. Explainable graph neural networks under fire, 2024.

[5] Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Lio, Bruno Lepri, and Andrea Passerini. Explaining the explainers in graph neural networks: a comparative study. 2024.

steve.azzolin@unitn.it