TopInG: Topologically Interpretable Graph Learning via Persistent Rationale Filtration

[1]Cheng Xin*, [2]Fan Xu*, [2]Xin Ding, [1]Jie Gao and [2]Jiaxin Ding

[1]Rutgers University, [2]Shanghai Jiao Tong University

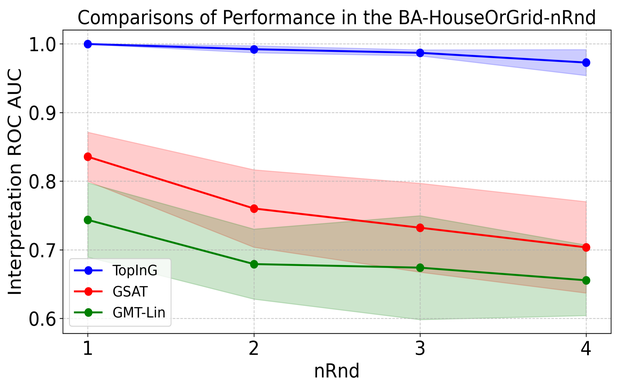
Problem and Contributions

Problem: Learn a model simultaneously predict the label and identify a **rational subgraph.**

- **TopInG** integrates TDA into GNNs for interpreting graphs by learning rationale filtration.
- Propose a new loss, **topological discrepancy**, measuring statistical difference on topological invariants --- **persistent homology**.
- Provide theoretical guarantees and a tractable approximation of our model.
- Experiments show **TopInG** improves performance on multiple benchmark datasets up to **20%+.**

Challenges

- Trade-off between Prediction & Interpretation
- Spurious correlations beyond rationale subgraph
- [New] Variform Rationale Subgraphs



Method

Main Idea: modeling an underlying graph generating process determined by an ordering $f : G \to [0, 1]$ consistent with the partition $G = G_X^* \bigsqcup G_\epsilon^*$ i.e.: $\exists t \approx 0.5, G_{<t} \approx G_X^*, G_{\geq t} \approx G_\epsilon^*$

Persistent Homology: on a graph filtration $\mathcal{F}(G) := \{G_{\leq t} \mid t \in f(E)\}$ a persistent homology is a chain of induced homology vector spaces connected by linear maps

$$H_p(\mathcal{F}(G)) : 0 \to \cdots \to H_p(G_{\leq t_1}) \to H_p(G_{\leq t_2}) \to H_p(G_{\leq t_3}) \to \cdots \to H_P(G)$$

Intuitively, using **topological discrepancy** to enlarge *persistent topological gap*

$$d_{\text{topo}}(\mathcal{P}(G_X), \mathcal{P}(G_\epsilon)) \triangleq \inf_{\pi \in \Pi(\mathcal{P}(G_X), \mathcal{P}(G_\epsilon))} \mathbb{E}_{(P,Q)\sim\pi}[d_{\text{B}}(P, Q)]$$

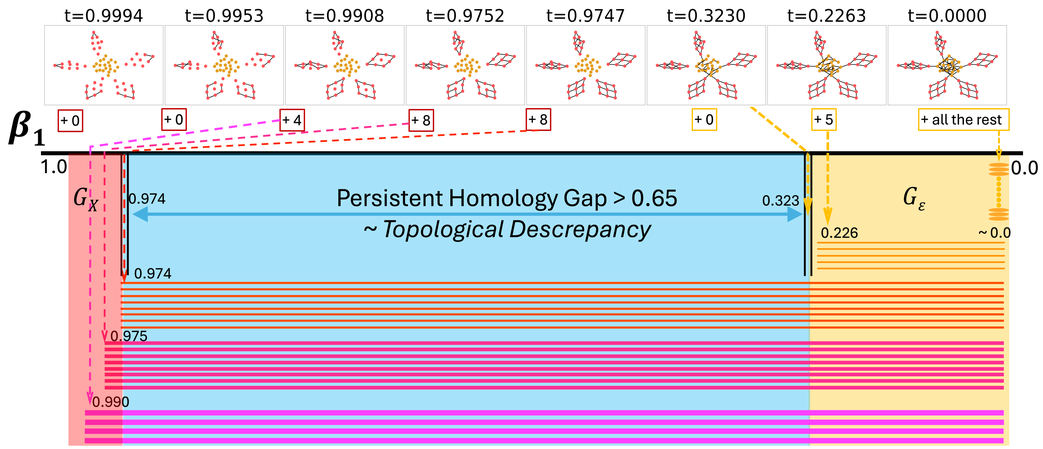
Theory

A tractable lower bound; upper bound by a functional Gromov-Hausdorff distance

$$\max_{\psi \in \Psi} |\mathbb{E}_{P\sim\mathcal{P}(G_X)}[\psi(P)] - \mathbb{E}_{Q\sim\mathcal{P}(G_\epsilon)}[\psi(Q)]| \leq d_{topo}(\mathcal{P}(G_X), \mathcal{P}(G_\epsilon)) \leq C \cdot d_{FGH}(G_X, G_\epsilon)$$

Theoretical Guarantee: optimized by ground truth rationale subgraphs

Theorem 3.4. *Assume $\forall G$, $|E_X| < |E_\epsilon|$, and G_X^* is minimal with respect to y_G in the sense that any subgraph $G_X \subset G_X^*$ losses some information of label, then $\mathcal{L}(\phi)$ is uniquely optimized by $f_\phi^*(e) = 1\{e \in G_X^*\}$.*



Experiments

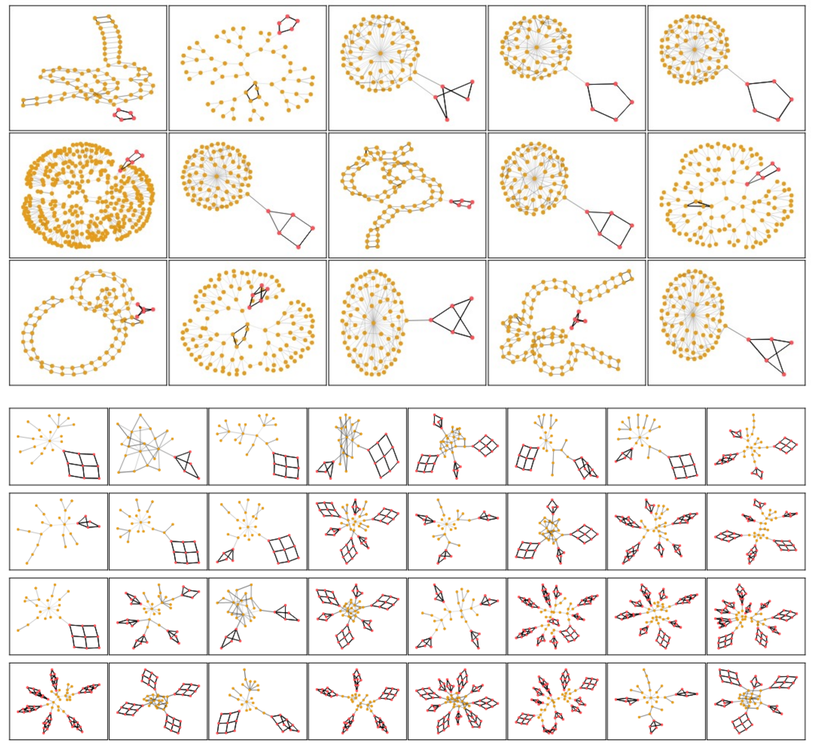
- Table 1: Interpretation Performance (AUC) on benchmark datasets.

	SingleMotif				MultipleMotif		RealDataset	
Method	**BA-2Motifs**	**BA-HouseGrid**	**SPmotif0.5**	**SPMotif0.9**	**BA-HouseAndGrid**	**BA-HouseOrGrid**	**Mutag**	**Benzene**
GNNExplainer	67.35 ± 3.29	50.73 ± 0.34	62.62 ± 1.35	58.85 ± 1.93	53.04 ± 0.38	53.21 ± 0.36	61.98 ± 5.45	48.72 ± 0.14
PGExplainer	84.59 ± 9.09	50.92 ± 1.51	69.54 ± 5.64	72.34 ± 2.91	10.36 ± 4.37	3.14 ± 0.01	60.91 ± 17.10	4.26 ± 0.36
MatchExplainer	86.06 ± 28.37	64.32 ± 2.32	57.29 ± 14.35	47.29 ± 13.39	81.67 ± 0.48	79.87 ± 1.61	91.04 ± 6.59	55.65 ± 1.16
MAGE	79.81 ± 2.27	82.69 ± 4.78	76.63 ± 0.95	74.38 ± 0.64	99.95 ± 0.06	99.93 ± 0.07	99.57 ± 0.47	96.03 ± 0.63
DIR	82.78 ± 10.97	65.50 ± 15.31	78.15 ± 1.32	49.08 ± 3.66	64.96 ± 14.31	59.71 ± 21.56	64.44 ± 28.81	54.08 ± 13.75
GSAT	98.85 ± 0.47	98.58 ± 0.59	74.49 ± 4.46	65.25 ± 4.42	92.92 ± 2.03	77.52 ± 3.71	99.38 ± 0.25	91.57 ± 1.48
GMT-Lin	97.72 ± 0.59	85.68 ± 2.79	76.26 ± 5.07	69.08 ± 10.14	76.12 ± 7.47	74.36 ± 5.41	$\mathbf{99.87 \pm 0.09}$	83.90 ± 6.07
TopING	$\mathbf{100.00 \pm 0.00}$	$\mathbf{99.87 \pm 0.13}$	$\mathbf{95.08 \pm 0.82}$	$\mathbf{90.82 \pm 4.95}$	$\mathbf{100.00 \pm 0.00}$	$\mathbf{100.00 \pm 0.00}$	96.38 ± 2.56	$\mathbf{100.00 \pm 0.00}$

- Table 2: Prediction Accuracy (Acc) on benchmark datasets.

	RealDataset		SpuriousMotif		
	Mutag	**Benzene**	**b=0.5**	**b=0.7**	**b=0.9**
DIR	68.72 ± 2.51	50.67 ± 0.93	45.49 ± 3.81	41.13 ± 2.62	37.61 ± 2.02
GSAT	$\mathbf{98.28 \pm 0.78}$	$\mathbf{100.00 \pm 0.00}$	47.45 ± 5.87	43.57 ± 2.43	45.39 ± 5.02
GMT-Lin	91.20 ± 2.75	$\mathbf{100.00 \pm 0.00}$	51.16 ± 3.51	53.11 ± 4.12	47.60 ± 2.06
TopING	92.92 ± 7.02	$\mathbf{100.00 \pm 0.00}$	$\mathbf{79.30 \pm 3.92}$	$\mathbf{75.46 \pm 7.62}$	$\mathbf{65.64 \pm 4.98}$

Visualization



Project Page

SCAN ME

https://jackal092927.github.io/publication/TopInG_ICML2025

C. Xin and J. Gao acknowledge funding from IIS-22298766, DMS-2220271, DMS-2311064, CRCNS2207440, CCF-2208663 and CCF-2118953. F. Xu and J. Ding were supported by NSF China under Grant No. T2421002, 62202299, 62020106005, 62061146002.