Local Geometry Improves Explanation Robustness for Graph Neural Networks

Mengting Diao[1] , **Li Sun**[1] , **Sen Su**[1*]
[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
{diaomengting, lsun, susen}@bupt.edu.cn

Abstract

Graph Neural Networks (GNNs) are widely used in high-stakes decision-making, where explanation robustness is crucial. Yet, even minor structural perturbations can drastically change explanations without changing predictions. Existing methods primarily characterize structural perturbations by their edit distance or budget, and enforce robustness through uniform constraints or averaging mechanisms over this perturbation set, which fail to capture the inherently direction-dependent effects of structural changes. In this paper, we propose GEOXGNN (Local Geometry Improves Explanation Robustness for Graph Neural Networks), a robust explanation framework that explicitly models the directional sensitivity of structural perturbations from a geometric perspective. GEOXGNN introduces local, direction-aware metrics in the node representation space to characterize how perturbations along different directions affect node embeddings and explanation outcomes. Based on the learned local metrics, GEOXGNN further enforces directional constraints, encouraging explanations to remain robust along high-sensitivity directions while preserving explanatory fidelity. Extensive experiments on multiple node-level and graph-level benchmark datasets demonstrate that GEOXGNN significantly improves explanation robustness over existing methods.

1 Introduction

Graph Neural Networks (GNNs) have been widely applied to structured data scenarios such as molecular property prediction, financial risk control, and social network analysis [Wale *et al.*, 2008; Yang *et al.*, 2019; Kipf and Welling, 2017; Yanardag and Vishwanathan, 2015a]. As GNNs are increasingly deployed in high-stakes decision-making tasks, a wide range of explanation methods have been developed and adopted to support understanding and trust in their predictions [Ying *et al.*, 2019; Luo *et al.*, 2020; Wang *et al.*, 2023a; Wang *et al.*, 2023b]. In this context, a central challenge lies in ensuring the robustness of explanations under structural perturbations.

Recent studies have shown that even when the model prediction remains unchanged, small perturbations to the input graph structure (e.g., adding or deleting a few edges) may lead to drastic changes in the produced explanations [Li *et al.*, 2024; Dombrowski *et al.*, 2019]. Such non-robustness not only undermines the credibility of explanations but also introduces potential risks in practical applications[Lakkaraju and Bastani, 2020]. For instance, in molecular graphs, minor modifications to non-critical chemical bonds may preserve the prediction label while significantly altering the key substructures highlighted by the model; in financial relational networks, slight rewiring of edges may shift risk-related explanations from truly relevant entities to irrelevant ones. Prior work has explored robustness in GNN explanations from multiple perspectives. One line of work adopts smoothing-based mechanisms, either through stochastic averaging (e.g., randomized smoothing) [Dombrowski *et al.*, 2019; Tan and Tian, 2023] or geometry-informed regularization [Liu *et al.*, 2025], which improve stability by uniformly suppressing explanation sensitivity under admissible perturbations. Another line of work follows an ensemble-based strategy, such as partitioning the graph into multiple substructures and aggregating explanations through voting, thereby achieving robustness via uniform aggregation over perturbation outcomes [Li *et al.*, 2025]. Despite their algorithmic differences, these approaches largely share a common characteristic: robustness is enforced by uniformly suppressing or averaging explanation sensitivity over admissible structural perturbations, without explicitly accounting for the inherently anisotropic effects induced by different structural perturbation directions. While reducing explanation variance, such uniform treatment can inadvertently dampen discriminative signals, leading to overly conservative explanations and a degradation in explanation fidelity.

However, this assumption overlooks a pervasive yet long-underestimated fact in real-world graph structures: the impact of structural perturbations on explanations is intrinsically direction-sensitive. Consider mutagenicity prediction for molecular graphs as an example. In such tasks, model decisions are often dominated by a small number of key functional groups, such as nitro groups ($-NO_2$). Under structural modifications that preserve the overall prediction out-

*Corresponding author.

come (e.g., remaining classified as toxic), explanations may nevertheless exhibit different responses. For instance, without altering the prediction result, introducing a nitroso group ($-NO$) may lead to a pronounced shift in the explanation. Although this group is not part of the original toxic core, its close relation to the nitro group in terms of electronic structure and reactivity can cause the explanation to incorrectly transfer importance to the newly introduced substructure. By contrast, introducing a cyano group ($-CN$), despite involving a structural modification of comparable topological scale, typically does not induce a similar degree of explanation drift, with the explanation remaining more strongly focused on the original nitro group. Although these two modifications are comparable in scale and preserve the model prediction, they induce substantially different explanation responses. This phenomenon indicates that explanation robustness is not determined solely by the size or location of a structural perturbation, but rather depends on whether the perturbation engages directions to which the model's discriminative evidence is highly sensitive. As a result, structural perturbations of equal size can elicit highly non-uniform effects on representations and explanations, which cannot be captured by isotropic perturbation modeling and motivates explicit direction-sensitive mechanisms. This motivates the need for an explicit mechanism to model direction-sensitive responses, enabling characterization of perturbation effects along different directions.

In this paper, we propose GeoXGNN (Local Geometry Improves Explanation Robustness for Graph Neural Networks), a robust explanation framework grounded in local, direction-dependent metrics. The key idea is to introduce a local metric structure in the node representation space that explicitly captures the heterogeneous sensitivity of explanations to structural perturbations across different directions, and to directly leverage this information in explanation construction and regularization.

Specifically, GeoXGNN departs from the common assumption that structural perturbations act as isotropic or uniformly scaled noise. First, it performs local perturbation response analysis to estimate a direction-aware sensitivity metric for each node, characterizing how small perturbations along different perturbation-induced directions affect node representations and explanation outcomes. Second, building on the learned local metric, GeoXGNN incorporates discriminative-aware robustness modeling, imposing direction-sensitive stabilization constraints that selectively suppress explanation instability along directions that are both highly sensitive and relevant to the model's decision-making process. Finally, these components are integrated into a single optimization objective for robust explanation learning. Importantly, GeoXGNN is designed to be both a *self-contained robust explanation framework* and a *plug-in robustness module* that can be seamlessly integrated with existing graph explanation methods to enhance their robustness. Extensive experiments demonstrate that GeoXGNN substantially improves explanation robustness across multiple node- and graph-level benchmark datasets, while consistently preserving explanation fidelity.

In summary, our main contributions are as follows:

- **Direction-sensitive perspective on explanation robustness.** We are the first to study the robustness of GNN explanations from a direction-sensitive perspective, revealing that explanation non-robustness under structural perturbations exhibits pronounced anisotropic behavior that cannot be captured by uniform perturbation modeling.
- **Metric-aware explanation framework.** We propose GeoXGNN, a metric-aware explanation framework that leverages local, direction-dependent sensitivity metrics to guide explanation construction, explicitly accounting for heterogeneous responses to structural perturbations.
- **Evaluation and analysis.** We conduct extensive experiments on multiple node- and graph-level benchmark datasets, demonstrating that GeoXGNN consistently improves explanation robustness while preserving explanation fidelity, and provide supporting theoretical analysis.

2 Problem Definition

2.1 Notations

Let $G = (V, E, X)$ denote an attributed graph, where $V = \{v_1, \ldots, v_n\}$ is the node set, $E \subseteq V \times V$ is the edge set, and $X \in \mathbb{R}^{n \times d}$ denotes node features. We consider both node-level and graph-level prediction tasks.

Let f be a trained Graph Neural Network. Given an input graph G, the model produces a latent representation

$$z = f(G) \in \mathbb{R}^D,$$

which is used to derive the final prediction. Throughout this work, we consider a pre-trained target GNN whose parameters are fixed during explanation.

2.2 Explanation Robustness Under Structural Perturbations

An explanation method assigns importance scores to structural elements of a graph. In this work, we formalize an explanation as a scoring function:

$$s_G : E \to \mathbb{R}.$$

Based on these scores, an explanation subgraph is constructed by selecting the top-K edges with the highest scores, denoted by $\mathcal{E}_K(G)$. We denote the corresponding explanation graph as $G_s = (V_s, E_s)$, where $E_s = \mathcal{E}_K(G)$ and $V_s \subseteq V$ is the set of nodes incident to E_s.

We study explanation robustness under *structural perturbations* of the input graph. Let $G' = (V', E')$ denote a perturbed graph generated by adding and/or removing edges from G, where $V' \subseteq V$ and $E' \subseteq V' \times V'$. Given a perturbation budget k, we define the admissible perturbation set as

$$\mathcal{B}_k(G) = \{G' \mid d_E(G, G') \leq k\},$$

where $d_E(\cdot, \cdot)$ denotes the edge edit distance, defined as the minimum number of edge additions or deletions required to transform G into G'.

We focus on prediction-preserving structural perturbations. Specifically, given a perturbed graph $G' \in \mathcal{B}_k(G)$ such that

$f(G') = f(G)$, we are interested in how the explanation subgraph $\mathcal{E}_K(G')$ compares to $\mathcal{E}_K(G)$.

Intuitively, an explanation method is considered robust if the top-K explanation subgraph remains stable under small, prediction-preserving structural perturbations.

3 Methodology

Existing robust graph explanation methods typically enforce robustness by treating structural perturbations of comparable graph-edit distance as uniformly scaled noise. As discussed in the introduction, such isotropic treatment fails to capture the highly uneven, direction-dependent effects induced by structural perturbations in graph neural networks.

To explicitly model and control this phenomenon, we propose GeoXGNN, a metric-aware robust explanations for graph neural networks that characterizes explanation robustness from a direction-sensitive perspective. We first model local geometric sensitivity (Sec 3.1), then introduce a discriminative-aware robustness modeling framework (Sec 3.2), and finally formulate a robust explanation objective(Sec 3.3). Figure 1 shows an overview of our GeoXGNN.

3.1 Local Geometric Sensitivity Modeling

To explicitly characterize local geometric sensitivity, we analyze how small structural perturbations deform learned representations and model the resulting anisotropic responses through a perturbation-induced local geometric metric. This metric provides a direction-dependent characterization of representation deformation under unit-scale graph edits.

Structural Perturbations and Representation Geometry. Let f be a pre-trained and fixed GNN. For node-level tasks, we denote the node representation as $z_v(G) \in \mathbb{R}^D$; for graph-level tasks, $z(G) \in \mathbb{R}^D$ denotes the graph representation. Given an original graph G and a structurally perturbed graph G', we define the representation shift

$$\Delta z = z(G') - z(G). \tag{1}$$

Although perturbations may share the same graph-edit distance, they can induce representation shifts along different directions with highly uneven magnitudes. As a result, measuring perturbation effects using a fixed Euclidean norm $\|\Delta z\|_2$ fails to reflect direction-dependent local sensitivity.

To model direction-dependent sensitivity in the representation space, we introduce a *local geometric metric* that characterizes how small structural perturbations deform representations along different directions. Although graph structures are discrete, the learned representation $z(G) \in \mathbb{R}^D$ lies in a continuous space, and admissible structural perturbations induce a family of local representation shifts around $z(G)$. From a Riemannian geometry perspective, when the perturbation budget is small, the induced representation shifts are dominated by first-order effects and can be locally approximated by a linear subspace, which serves as a tangent space at $z(G)$.

On this basis, local sensitivity around a point $z(G)$ can be characterized by two components: (i) a local tangent space that captures admissible infinitesimal directions of variation induced by structural perturbations, and (ii) an inner product (metric tensor) defined on this space that quantifies the relative scale of deformation along different directions [Tron *et al.*, 2024]. Such a metric assigns different weights to different tangent directions, thereby enabling explicit characterization of anisotropic deformation behavior.

Formally, we define a local geometric metric as a symmetric positive semidefinite matrix

$$\mathbf{M}_G \in \mathbb{R}^{D\times D}, \tag{2}$$

which induces a local norm

$$\|\Delta z\|_{\mathbf{M}_G} = \sqrt{\Delta z^\top \mathbf{M}_G \Delta z}. \tag{3}$$

to measure the effective scale of representation variation.

In the following, we describe how such tangent directions and the associated local metric can be induced from perturbation-based response analysis.

Perturbation-Induced Tangent Directions. To capture direction-dependent sensitivity, we adopt the *Adaptive Orthogonal Frame (AOF)* [Sun *et al.*, 2026] mechanism to construct local tangent directions in representation space. Specifically, AOF generates perturbation-induced tangent vectors by introducing structured sparse structural perturbations and observing their induced representation changes.

Following the AOF construction, we introduce M virtual perturbation nodes and connect each to k selected real nodes, producing an augmented graph $\widehat{G}$. For each virtual perturbation, we obtain a tangent candidate

$$t_m = z(\widehat{G}^{(m)}) - z(G), \qquad m = 1, \dots, M. \tag{4}$$

Applying an orthogonal decomposition with length recovery yields an orthogonal frame

$$W = [w_1, \dots, w_M] \in \mathbb{R}^{D\times M}, \tag{5}$$

where each w_m represents a local tangent direction induced by structural perturbations.

Based on the AOF frame, we define the local geometric metric as

$$\mathbf{M}_G = W^\top W = \mathrm{diag}\big(\|w_1\|^2, \dots, \|w_M\|^2\big). \tag{6}$$

which is symmetric positive definite on the spanned tangent subspace. The induced norm Eq. (3) measures the effective scale of representation variation under local structural perturbations. Intuitively, directions corresponding to larger $\|w_m\|$ encode higher structural sensitivity revealed by perturbation responses, while insensitive directions are naturally down-weighted.

Why a Local Metric. Unlike raw perturbation sampling or isotropic norms, the induced metric $\mathbf{M}_G$ provides a coordinate-free and direction-aware characterization of local geometry. It enables consistent comparison of representation shifts across different perturbation-induced directions and forms the foundation for robustness analysis used to control explanation non-robustness in subsequent sections.

Lemma 1 (Local Sensitivity Characterization). *Let $W = [w_1, \dots, w_M]$ be the AOF basis constructed from perturbation-induced tangent vectors. Then the metric $\mathbf{M}_G = W^\top W$ characterizes the anisotropic sensitivity of representation shifts in the neighborhood of $z(G)$, with each direction weighted according to its local structural sensitivity.*

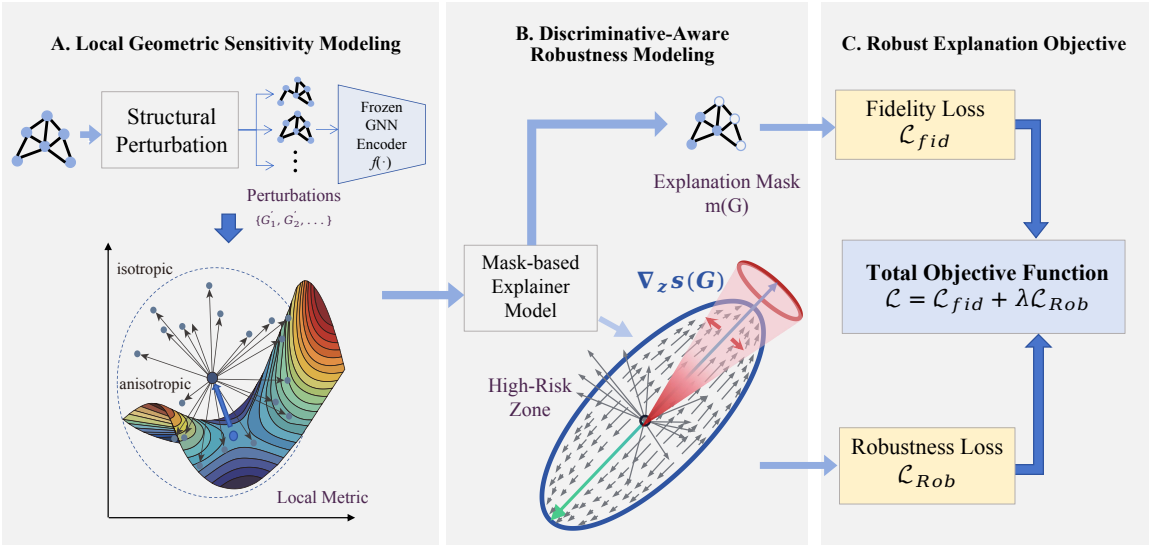

Figure 1: The overall framework of GeoXGNN.

3.2 Discriminative-Aware Robustness Modeling

Local geometric sensitivity characterizes how small structural perturbations are amplified in the representation space. However, not all locally geometrically sensitive directions should be suppressed for robustness: only those directions whose variations are relevant to the model's discriminative decision constitute robustness-critical targets. This motivates a *discriminative-aware* robustness modeling principle, which selectively constrains explanation sensitivity along discriminative-relevant directions rather than indiscriminately penalizing all local variations.

Explainer Sensitivity in Representation Space. Let $s(G)$ denote the output of a graph explanation method applied to G, and let $z(G)$ be the corresponding representation produced by the frozen GNN. We write $s(G) = s(z(G))$ to emphasize this dependency. For sufficiently small perturbations, explanation variation admits a first-order approximation:

$$s(G') - s(G) \approx \nabla_z s(G)^\top \Delta z, \Delta z = z(G') - z(G). \quad (7)$$

This expression highlights that explanation instability depends not only on the Δz, but also on its alignment with the explainer gradient $\nabla_z s(G)$.

Discriminative Subspace Projection of Local Geometry. Let $f(z)$ denote the prediction score of the frozen GNN for the predicted class. The gradient

$$g = \nabla_z f(z) \quad (8)$$

defines the most discriminative direction in the representation space, as it characterizes the direction of maximal first-order change in the decision score. We define the discriminative subspace as

$$\mathcal{S}_{\text{disc}} = \text{span}\{g\}. \quad (9)$$

Let $\{w_m\}_{m=1}^{M}$ denote locally geometrically sensitive directions identified by the local metric construction. We retain only their discriminative-relevant components by projecting them onto $\mathcal{S}_{\text{disc}}$:

$$\tilde{w}_m = P_{\text{disc}} w_m, \qquad P_{\text{disc}} = \frac{gg^\top}{\|g\|_2^2}. \quad (10)$$

This projection filters out geometrically sensitive but discriminative-irrelevant variations, ensuring that robustness regularization targets only directions that directly interact with the model decision.

Lemma 2 (Discriminative-Aware Local Stability Bound). *Assume that the explanation function $s(z)$ is locally Lipschitz in a neighborhood of $z(G)$. For sufficiently small perturbations with representation shift Δz, the explanation variation satisfies*

$$|s(G') - s(G)| \leq \|\nabla_z s(G)\|_{\mathbf{M}_G^{-1}} \cdot \|\Delta z\|_{\mathbf{M}_G}. \quad (11)$$

where the local metric $\mathbf{M}_G$ is constructed from discriminative-projected sensitivity directions $\{\tilde{w}_m\}$. Consequently, explanation instability is controlled only along discriminative-relevant, geometrically amplified directions, rather than across the entire representation space.

3.3 Robust Explanation Objective

We incorporate the geometry- and discriminative-aware direction information into a unified explanation objective, so that the explainer is trained to reduce explanation non-robustness primarily along robustness-critical directions, while preserving discriminative evidence and explanation fidelity.

Direction-Wise Robustness Risk. Let $\{w_m\}_{m=1}^{M}$ denote the principal sensitivity directions induced by the local metric $\mathbf{M}_G$, and define $u_m = w_m/\|w_m\|$. We quantify the robustness risk along direction m as

$$r_m(G) = \|w_m\|^2 \cdot \big(\langle \nabla_z s(G), u_m \rangle\big)^2, \tag{12}$$

which is large only when a direction is both geometrically sensitive and strongly coupled with the explainer gradient.

Robust Explanation Objective. Aggregating direction-wise risks yields the robustness objective

$$\mathcal{L}_{\text{rob}}(G) = \sum_{m=1}^{M} r_m(G) = \big\|W^\top \nabla_z s(G)\big\|_2^2, \tag{13}$$

where $W = [w_1, \ldots, w_M]$. This objective penalizes explanation sensitivity only along geometrically amplified directions that interfere with discriminative stability.

Overall Objective. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{fid}} + \lambda\, \mathcal{L}_{\text{rob}}(G), \tag{14}$$

where $\mathcal{L}_{\text{fid}}$ enforces explanation fidelity and λ controls the robustness–fidelity trade-off.

Theorem 1 (Direction-Selective Robustness Guarantee). *Under the local Lipschitz assumption of $s(z)$, the proposed objective selectively bounds explanation variation along discriminative-relevant, geometrically sensitive directions. Compared to isotropic regularization, it yields strictly tighter bounds when local sensitivity is anisotropic.*

Computational Complexity. Let d denote the representation dimension and M the number of local sensitivity directions. Constructing the local metric incurs a one-time cost $\mathcal{O}(Md^2)$ per graph in the offline stage. During explainer training, computing the robustness objective requires $\mathcal{O}(Md)$ additional operations per iteration, introducing negligible overhead compared to standard gradient-based explanation training.

4 Experiments

4.1 Experimental Setups

Tasks and Datasets. We evaluate explanation robustness on both *node classification* and *graph classification* tasks. For node classification, we use three citation networks: Cora [McCallum *et al.*, 2000], Citeseer [Giles *et al.*, 1998], and PubMed [Sen *et al.*, 2008] . For graph classification, we consider three real-world datasets, namely MUTAG [Morris *et al.*, 2020], PROTEINS [Borgwardt *et al.*, 2005], and IMDB-BINARY [Yanardag and Vishwanathan, 2015b], as well as three synthetic datasets, including SG+House, SG+Diamond, and SG+Wheel [Agarwal *et al.*, 2023]. We use a 70/10/20 split throughout each dataset. We select GCN [Kipf and Welling, 2017] with 3-layer as the target GNN and kept fixed during explanation and robustness evaluation. For explainer, we use a mask-based explainer, which outputs edge-level importance scores $m_e(G) \in [0, 1]$ for each edge $e \in E$. The explanation subgraph is constructed by selecting the top-K edges with the highest scores. We construct the local geometric metric using an AOF-style virtual-node augmentation.

Dataset	GNNExpl.	PGExpl.	GNN-LRP	Curvature	**GeoXGNN**
Cora	0.406	0.522	**0.288**	0.312	0.454
Citeseer	0.602	0.688	0.512	0.509	**0.483**
PubMed	0.699	0.721	0.810	0.821	**0.691**
MUTAG	0.226	0.245	0.267	0.232	**0.215**
PROTEINS	0.256	**0.112**	0.123	0.156	0.180
IMDB-BINARY	0.302	0.255	0.233	0.201	**0.187**
SG+House	0.812	0.786	**0.712**	0.723	0.737
SG+Diamond	0.610	0.724	0.502	0.566	**0.424**
SG+Wheel	0.563	0.573	0.624	0.610	**0.533**

Table 1: Fidelity-KL on node and graph classification.

For each instance, we introduce $M_{\text{AOF}} = 16$ virtual nodes, and connect each virtual node to $k_{\text{AOF}} = 16$ real nodes sampled uniformly to generate local tangent directions.

Baselines. We compare against four representative explainers: GNNExplainer [Ying *et al.*, 2019], PGExplainer [Luo *et al.*, 2020], GNN-LRP [Schnake *et al.*, 2022], and Curvature [Liu *et al.*, 2025]. To assess the *plug-in* property, we further integrate our GeoXGNN into each baseline, forming enhanced variants (e.g., GNNExplainer+GeoXGNN), and evaluate them under the same settings.

Structural Perturbations and Evaluation. We evaluate explanation robustness under *edge-edit perturbations*, including both edge deletions and additions. For graph classification, we consider perturbed graphs G' satisfying $d_E(G, G') \le k$, where k is set to $\lfloor 0.1|E| \rfloor$. For node classification, we fix $k = 5$ and perform perturbations within the 3-hop neighborhood of the target node. Our focus is explanation robustness *when the model prediction does not change*, when predictions differ, corresponding explanations are expected to change accordingly and should not be constrained to be consistent.

4.2 Metrics

Fidelity-KL. We follow prior work [Yuan *et al.*, 2023; Liu *et al.*, 2025] and measure fidelity by the KL divergence between the predictive distributions of the original graph and the masked explanation graph: $\text{Fidelity-KL}(G) = \text{KL}\big(p(G) \,\|\, p(G \odot m)\big)$, where $p(\cdot)$ denotes the prediction of f and $G \odot m$ is the explainer-produced masked graph. Lower values indicate that the explanation preserves the model behavior more faithfully.

Robustness Evaluation. We evaluate explanation robustness under prediction-preserving structural perturbations using two metrics: AUC and Top-K Edge Jaccard. AUC reflects whether the perturbed explanation preserves the *relative ranking* of important edges, Jaccard quantifies the *hard set overlap* of the selected top-K subgraphs. Following [Bajaj *et al.*, 2021], we quantifying how much an explanation changes after adding noise to the input graph. We treat the top-K edges of the explanation on G as positives and use the edge scores produced on a prediction-preserving perturbed graph G' as ranking scores. We then compute the ROC-AUC over candidate edges, and report the mean AUC.

We measure the stability of the extracted top-K explanation subgraphs by the Jaccard overlap between the top-K

Datasets	GNNExpl.		PGExpl.		GNN-LRP		Curvature		**GeoXGNN**	
	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑
Cora	0.607 ± 0.012	0.592 ± 0.215	0.623 ± 0.232	0.453 ± 0.089	0.719 ± 0.038	0.621 ± 0.132	0.810 ± 0.121	0.835 ± 0.112	$\mathbf{0.828 \pm 0.168}$	$\mathbf{0.854 \pm 0.056}$
Citeseer	0.512 ± 0.021	0.644 ± 0.182	0.581 ± 0.112	0.458 ± 0.109	0.433 ± 0.021	0.527 ± 0.214	0.678 ± 0.021	0.572 ± 0.017	$\mathbf{0.814 \pm 0.102}$	$\mathbf{0.719 \pm 0.096}$
PubMed	0.523 ± 0.113	0.559 ± 0.198	0.458 ± 0.122	0.642 ± 0.128	0.665 ± 0.121	0.589 ± 0.156	0.828 ± 0.021	0.834 ± 0.168	$\mathbf{0.866 \pm 0.035}$	$\mathbf{0.837 \pm 0.112}$

Table 2: Node classification robustness comparison. We report AUC (mean±std) and Jaccard (mean±std).

Datasets	GNNExpl.		PGExpl.		GNN-LRP		Curvature		**GeoXGNN**	
	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑	AUC↑	Jaccard↑
MUTAG	0.612 ± 0.123	0.452 ± 0.211	0.633 ± 0.065	0.621 ± 0.116	0.723 ± 0.078	0.698 ± 0.102	0.821 ± 0.154	0.762 ± 0.135	$\mathbf{0.832 \pm 0.104}$	$\mathbf{0.867 \pm 0.083}$
PROTEINS	0.322 ± 0.134	0.511 ± 0.189	0.623 ± 0.115	0.562 ± 0.178	0.712 ± 0.157	0.621 ± 0.121	0.700 ± 0.112	$\mathbf{0.723 \pm 0.104}$	$\mathbf{0.722 \pm 0.085}$	0.684 ± 0.131
IMDB-BINARY	0.389 ± 0.162	0.358 ± 0.163	0.431 ± 0.158	0.325 ± 0.176	0.345 ± 0.021	0.462 ± 0.089	$\mathbf{0.612 \pm 0.147}$	0.487 ± 0.021	0.532 ± 0.126	$\mathbf{0.494 \pm 0.080}$
SG+House	0.418 ± 0.211	0.511 ± 0.136	0.512 ± 0.173	0.622 ± 0.043	0.601 ± 0.123	0.577 ± 0.168	0.819 ± 0.083	0.832 ± 0.074	$\mathbf{0.837 \pm 0.082}$	$\mathbf{0.868 \pm 0.023}$
SG+Diamond	0.598 ± 0.026	0.612 ± 0.024	0.542 ± 0.145	0.573 ± 0.042	0.623 ± 0.072	0.523 ± 0.033	0.755 ± 0.035	0.623 ± 0.046	$\mathbf{0.812 \pm 0.087}$	$\mathbf{0.707 \pm 0.088}$
SG+Wheel	0.553 ± 0.036	0.587 ± 0.167	0.612 ± 0.029	0.568 ± 0.013	0.528 ± 0.065	0.486 ± 0.052	0.798 ± 0.121	0.652 ± 0.047	$\mathbf{0.825 \pm 0.055}$	$\mathbf{0.742 \pm 0.035}$

Table 3: Graph classification robustness comparison. We report AUC (mean±std) and Jaccard (mean±std).

edge sets $\mathcal{E}_k(G)$ and $\mathcal{E}_k(G')$:

$$\mathrm{Jaccard}(G, G') = \frac{|\mathcal{E}_k(G) \cap \mathcal{E}_k(G')|}{|\mathcal{E}_k(G) \cup \mathcal{E}_k(G')|}.$$

Higher AUC and Jaccard indicate more robust explanations.

4.3 Main Results

We report **fidelity on original graphs** (Fidelity-KL in Table 1) and **robustness under prediction-preserving perturbations** (AUC/Jaccard in Tables 2 and 3). GeoXGNN maintains competitive fidelity across datasets, achieving the best or near-best KL on several benchmarks (e.g., MUTAG, IMDB-BINARY, SG+Wheel, and Citeseer) and remaining close to strong baselines on Cora and PubMed, indicating robustness gains are not obtained by sacrificing faithfulness. Meanwhile, GeoXGNN consistently improves robustness: it attains the highest AUC and Jaccard on all three node datasets, and achieves the best results on MUTAG and all synthetic graph datasets while staying competitive on PROTEINS and IMDB-BINARY. Overall, GeoXGNN yields more stable explanations under prediction-preserving edge edits without degrading fidelity on clean graphs.

4.4 Direction-Dependent Analysis

Under the same edit budget and prediction-preserving constraint, different perturbations can induce representation shifts along different local directions.

To verify this direction-dependent effect, we group perturbations by their alignment with the leading eigen-direction of M_G and compare robustness across groups. For each test graph G, we first compute its local metric M_G and perform eigendecomposition $M_G = U\Lambda U^\top$.

We use the leading eigen-direction u_1 as the *most sensitive direction*. We then sample N prediction-preserving perturbed graphs $\{G'_j\}_{j=1}^N$ with the same edit budget. For each perturbation, we compute $\Delta z_j = z(G'_j) - z(G)$ and its normalized direction $\widehat{\Delta z}_j$, and define the alignment score: $a_j = \left|\left\langle \widehat{\Delta z}_j, u_1 \right\rangle\right|$. We bucket perturbations into *High-align* (top 30% by a_j) and *Low-align* (bottom 30%). We report group-wise robustness by averaging Jaccard@k within each bucket. Table 4 shows that *High-align* perturbations cause larger explanation drift for isotropic baselines, while GeoXGNN substantially reduces the gap (*Low-align* − *High-align*), indicating direction-dependent stabilization along the most sensitive geometry direction.

Method	High-align ↑	Low-align ↑	Gap ↓
Euclidean / no-metric	0.523	0.632	0.109
GeoXGNN	0.839	0.866	0.027

Table 4: Direction-dependent robustness (Avg Jaccard@k).

4.5 Evaluation As Plug-In

Table 5 evaluates GeoXGNN as a plug-in module by integrating it into GNNExplainer, PGExplainer, and GNN-LRP. Overall, GeoXGNN improves robustness for most (dataset, explainer) pairs across both node and graph classification. Gains are more consistent for optimization-based methods (e.g., GNNExplainer) than for amortized PGExplainer. Curvature is omitted since it is itself a plug-in method and is compared as a baseline in the main tables.

4.6 Ablation and Sensitivity Analysis

We study three key components: (i) the local geometric metric, (ii) the discriminative-aware risk weighting, and (iii) the robustness weight λ.

Effect of Local Metric. As shown in Table 6, replacing the local metric with an isotropic Euclidean treatment causes a clear robustness drop ($0.867 \rightarrow 0.655$ in Avg Jaccard@k) and a notable fidelity degradation (KL $0.235 \rightarrow 0.379$). This is because isotropic regularization tends to suppress changes indiscriminately, thus harming fidelity.

Effect of Discriminative-Aware Risk. Removing discriminative-aware weighting reduces robustness from 0.867 to 0.709 and increases KL from 0.235 to 0.291. While

Dataset	Metric	GNNExplainer		PGExplainer		GNN-LRP	
		Base	+GeoXGNN	Base	+GeoXGNN	Base	+GeoXGNN
Cora	Avg Jaccard	0.592 ± 0.215	$\mathbf{0.659 \pm 0.083}$	0.453 ± 0.089	$\mathbf{0.467 \pm 0.099}$	0.621 ± 0.132	$\mathbf{0.636 \pm 0.078}$
Citeseer	Avg Jaccard	$\mathbf{0.644 \pm 0.182}$	0.642 ± 0.058	0.458 ± 0.109	$\mathbf{0.543 \pm 0.067}$	0.527 ± 0.214	$\mathbf{0.533 \pm 0.084}$
PubMed	Avg Jaccard	0.559 ± 0.198	$\mathbf{0.589 \pm 0.048}$	$\mathbf{0.642 \pm 0.128}$	0.640 ± 0.037	0.589 ± 0.156	$\mathbf{0.611 \pm 0.084}$
MUTAG	Avg Jaccard	0.452 ± 0.211	$\mathbf{0.512 \pm 0.071}$	0.621 ± 0.116	$\mathbf{0.626 \pm 0.076}$	0.698 ± 0.102	$\mathbf{0.712 \pm 0.067}$
PROTEINS	Avg Jaccard	0.511 ± 0.189	$\mathbf{0.523 \pm 0.099}$	0.562 ± 0.178	$\mathbf{0.572 \pm 0.085}$	0.621 ± 0.121	$\mathbf{0.625 \pm 0.067}$
IMDB-BINARY	Avg Jaccard	0.358 ± 0.163	$\mathbf{0.375 \pm 0.066}$	0.325 ± 0.176	$\mathbf{0.345 \pm 0.056}$	0.462 ± 0.089	$\mathbf{0.482 \pm 0.077}$
SG+House	Avg Jaccard	0.511 ± 0.136	$\mathbf{0.523 \pm 0.056}$	0.622 ± 0.043	$\mathbf{0.632 \pm 0.041}$	0.577 ± 0.168	$\mathbf{0.612 \pm 0.054}$
SG+Diamond	Avg Jaccard	0.612 ± 0.024	$\mathbf{0.634 \pm 0.081}$	0.573 ± 0.042	$\mathbf{0.589 \pm 0.026}$	0.523 ± 0.033	$\mathbf{0.567 \pm 0.073}$
SG+Wheel	Avg Jaccard	0.587 ± 0.167	$\mathbf{0.612 \pm 0.022}$	$\mathbf{0.568 \pm 0.013}$	0.564 ± 0.023	0.486 ± 0.052	$\mathbf{0.495 \pm 0.032}$

Table 5: Plug-and-play evaluation. We report Avg Jaccard@k (higher is better).

Variant	Local Metric	Disc. Risk	λ	Robustness↑	Fidelity↓
w/o Robustness	✗	✗	0	0.620 ± 0.075	0.219 ± 0.019
Euclidean (no metric)	✗	✓	1.0	0.655 ± 0.054	0.379 ± 0.135
w/o Disc. Risk ($\alpha = 1$)	✓	✗	1.0	0.709 ± 0.035	0.291 ± 0.109
Full GeoXGNN	✓	✓	1.0	0.867 ± 0.083	0.235 ± 0.032

Table 6: Ablation study. We report robustness (Avg Jaccard@k, higher is better) and fidelity (e.g., KL divergence, lower is better) on MUTAG.

λ	Robustness↑	Fidelity↓
0	0.620 ± 0.075	0.219 ± 0.019
0.1	0.620 ± 0.074	0.215 ± 0.018
0.5	0.701 ± 0.064	0.234 ± 0.021
1.0	0.867 ± 0.083	0.235 ± 0.032
2.0	0.859 ± 0.139	0.243 ± 0.065

Table 7: Sensitivity to λ. We report robustness (Avg Jaccard@k) and fidelity (e.g., KL divergence) on MUTAG.

the local metric alone already improves robustness over the unregularized baseline, discriminative-aware weighting further strengthens robustness by prioritizing regulation on perturbations that are more decision-relevant.

Effect of Robustness Weight λ. Table 7 shows that robustness improves as λ increases up to $\lambda = 1.0$, while fidelity remains relatively stable in the moderate range ($\lambda = 0.5$–1.0, KL $\approx$ 0.234–0.235). When λ becomes too large (2.0), robustness saturates with larger variance and fidelity slightly degrades, indicating mild over-regularization.

5 Related Work

Graph Explainability: To enhance the transparency of graph neural networks, many explanation techniques have been proposed in recent years. These methods can be categorized into instance-level methods and model-level methods [Yuan *et al.*, 2023; Zhang and Chen, 2020]. For instance-level methods [Ying *et al.*, 2019; Luo *et al.*, 2020; Schnake *et al.*, 2022; Wang *et al.*, 2021; Zhang *et al.*, 2024; Liu and Xie, 2025], Ying et al. [Ying *et al.*, 2019] proposed GNNExplainer, which treats explanation generation as a mask optimization problem based on perturbation. PGExplainer [Luo *et al.*, 2020] adopted a reparameterization technique to learn an approximate discrete mask that maximizes the mutual information between key structures and the prediction. GNN-LRP [Schnake *et al.*, 2022] attributes predictions to relevant graph walks via layer-wise relevance propagation, enabling fine-grained and robust higher-order explanations. For model-level methods [Shin *et al.*, 2024; Magister *et al.*, 2021; Yuan *et al.*, 2020; Wang and Shen, 2023], they give high-level explanations which reflect patterns of behavior over the whole model. However, most existing explanation methods have been shown to be vulnerable to small perturbations, leading to non-robust explanations.

Explanation Robustness: Recent works have raised growing concerns about the robustness of existing explainability methods. Some obustness-oriented explanation improve explanation robustness by enforcing consistency across perturbed inputs through *smoothing-based* mechanisms, such as randomized smoothing [Dombrowski *et al.*, 2019; Tan and Tian, 2023; Zhao *et al.*, 2023; Chen *et al.*, 2019] or geometry-informed regularization [Liu *et al.*, 2025]. Other approaches enhance robustness from an *ensemble-based* perspective, for example by partitioning the graph into multiple substructures and aggregating explanations via voting over perturbation outcomes [Li *et al.*, 2025]. However, these methods typically assume an *isotropic* perturbation neighborhood and enforce robustness by smoothing, consistency regularization, or aggregation over perturbed inputs, without modeling direction-dependent representation sensitivity. This may attenuate discriminative signals, yielding overly conservative explanations with reduced fidelity.

6 Conclusion

In this paper, we studied the robustness of GNN explanations under structural perturbations and highlighted the importance of direction-sensitive effects that are missed by isotropic perturbation modeling. We proposed GeoXGNN, a metric-aware framework that leverages local, direction-dependent sensitivity metrics to stabilize explanations along sensitive directions. Experiments show that GeoXGNN consistently improves explanation robustness while preserving fidelity.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62072052), and the National Key Research and Development Program of China (2024YFF0907401).

References

[Agarwal *et al.*, 2023] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023.

[Bajaj *et al.*, 2021] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. *Advances in neural information processing systems*, 34:5644–5655, 2021.

[Borgwardt *et al.*, 2005] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47–i56, 2005.

[Chen *et al.*, 2019] Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14300–14310, 2019.

[Dombrowski *et al.*, 2019] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13567–13578, 2019.

[Giles *et al.*, 1998] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Lakkaraju and Bastani, 2020] Himabindu Lakkaraju and Osbert Bastani. "how do I fool you?": Manipulating user trust via misleading black box explanations. In Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington, editors, *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 79–85. ACM, 2020.

[Li *et al.*, 2024] Jiate Li, Meng Pang, Yun Dong, Jinyuan Jia, and Binghui Wang. Graph neural network explanations are fragile. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[Li *et al.*, 2025] Jiate Li, Meng Pang, Yun Dong, Jinyuan Jia, and Binghui Wang. Provably robust explainable graph neural networks against graph perturbation attacks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[Liu and Xie, 2025] Yazheng Liu and Sihong Xie. Explanations of GNN on evolving graphs via axiomatic layer edges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[Liu *et al.*, 2025] Yazheng Liu, Xi Zhang, Sihong Xie, and Hui Xiong. Robust explanations of graph neural networks via graph curvatures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Luo *et al.*, 2020] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Magister *et al.*, 2021] Lucie Charlotte Magister, Dmitry Kazhdan, Vikash Singh, and Pietro Liò. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *CoRR*, abs/2107.11889, 2021.

[McCallum *et al.*, 2000] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[Morris *et al.*, 2020] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[Schnake *et al.*, 2022] Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schutt, Klaus-Robert Muller, and Gregoire Montavon. Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596, November 2022.

[Sen *et al.*, 2008] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[Shin *et al.*, 2024] Yong-Min Shin, Sun-Woo Kim, and Won-Yong Shin. PAGE: prototype-based model-level explanations for graph neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(10):6559–6576, 2024.

[Sun *et al.*, 2026] Li Sun, Zhenhao Huang, Silei Chen, Lanxu Yang, Junda Ye, Sen Su, and Philip S. Yu. Multi-domain riemannian graph gluing for building graph foundation models, 2026.

[Tan and Tian, 2023] Zeren Tan and Yang Tian. Robust explanation for free or at the cost of faithfulness. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 33534–33562. PMLR, 2023.

[Tron *et al.*, 2024] Eliot Tron, Rita Fioresi, Nicolas Couellan, and Stéphane Puechmorel. Cartan moving frames and the data manifolds. *CoRR*, abs/2409.12057, 2024.

[Wale *et al.*, 2008] Nikil Wale, Ian A. Watson, and George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowl. Inf. Syst.*, 14(3):347–375, 2008.

[Wang and Shen, 2023] Xiaoqi Wang and Han-Wei Shen. Gnninterpreter: A probabilistic generative model-level explanation for graph neural networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Wang *et al.*, 2021] Xiang Wang, Ying-Xin Wu, An Zhang, Xiangnan He, and Tat-Seng Chua. Towards multi-grained explainability for graph neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18446–18458, 2021.

[Wang *et al.*, 2023a] Senzhang Wang, Jun Yin, Chaozhuo Li, Xing Xie, and Jianxin Wang. V-infor: A robust graph neural networks explainer for structurally corrupted graphs. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Wang *et al.*, 2023b] Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):2297–2309, 2023.

[Yanardag and Vishwanathan, 2015a] Pinar Yanardag and S. V. N. Vishwanathan. Deep graph kernels. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1365–1374. ACM, 2015.

[Yanardag and Vishwanathan, 2015b] Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1365–1374, 2015.

[Yang *et al.*, 2019] Yiying Yang, Zhongyu Wei, Qin Chen, and Libo Wu. Using external knowledge for financial event prediction based on graph neural networks. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2161–2164. ACM, 2019.

[Ying *et al.*, 2019] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9240–9251, 2019.

[Yuan *et al.*, 2020] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. XGNN: towards model-level explanations of graph neural networks. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 430–438. ACM, 2020.

[Yuan *et al.*, 2023] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5782–5799, 2023.

[Zhang and Chen, 2020] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101, 2020.

[Zhang *et al.*, 2024] Jiaxing Zhang, Zhuomin Chen, Hao Mei, Longchao Da, Dongsheng Luo, and Hua Wei. Reg-explainer: Generating explanations for graph neural networks in regression tasks. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[Zhao *et al.*, 2023] Tianxiang Zhao, Dongsheng Luo, Xiang Zhang, and Suhang Wang. Towards faithful and consistent explanations for graph neural networks. In Tat-Seng Chua, Hady W. Lauw, Luo Si, Evimaria Terzi, and Panayiotis Tsaparas, editors, *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023*, pages 634–642. ACM, 2023.