STABLEGX: STABLE GLOBAL EXPLANATIONS FOR GNNS ON LARGE HETEROGENEOUS AND HETEROPHILIC GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Neural Networks (GNNs) are increasingly deployed on large real-world graphs, where global explanations should remain consistent across the different graph samples required for scalable analysis. Existing global GNN explainers can produce different explanations from different samples of data for the same black-box GNN model and face additional challenges on heterogeneous and heterophilic graphs with large, irregular neighborhoods and rich node features. We introduce StableGX, a global explainer for node-classification GNNs that targets both fidelity and stability across explanation samples. StableGX maps each node and its multi-hop neighborhood to a fixed-dimensional, interpretable representation of node features and neighborhood statistics, and identifies representative exemplar populations in the GNN embedding space. It then learns logical signatures for these populations using boundary-aware sampling, which identifies candidate points near the population boundary and applies max–min selection to improve boundary coverage. We theoretically relate boundary approximation error to fidelity and stability, and establish a bound on the approximation error induced by the proposed sampling procedure. Experiments across homogeneous and heterogeneous, homophilic and heterophilic graphs show that StableGX produces faithful and stable global explanations. In particular, it improves over GNNXemplar on large heterogeneous and heterophilic graphs, while GNNXemplar remains stronger on simpler homogeneous graph benchmarks.

1 INTRODUCTION

Graph Neural Networks (GNNs) are increasingly used to model large real-world systems in domains such as finance, social networks, and cybersecurity (Liu et al., 2024; Lv et al., 2021). These systems are often naturally heterogeneous, containing multiple types of entities and relations whose interactions are captured by modern GNN architectures. Understanding the underlying behavior of these GNN models, which are now routinely applied to increasingly large and complex graphs, becomes important for establishing trust, diagnosing failures, and enabling meaningful human oversight. This has motivated a growing body of work on GNN explainability Azzolin et al. (2023); Lucic et al. (2022); Luo et al. (2020); Zhang et al. (2020). However, existing explainers have largely been developed and evaluated on smaller and structurally simpler graphs, leaving their scalability, reliability, and practical applicability to large heterogeneous real-world graphs insufficiently understood.

GNN explainers are commonly studied at two complementary levels: local methods explain individual predictions Luo et al. (2020); Yuan et al. (2021); Wang et al. (2023), whereas global methods summarize recurring model behavior across multiple instances (Armgaan et al., 2025; 2024; Azzolin et al., 2023; Yuan et al., 2020). This work focuses on global explanations, with the objective of identifying class-level patterns that characterize how a trained GNN makes predictions across groups of nodes. Such characterization becomes particularly challenging on heterogeneous graphs, where nodes of the same predicted class can occur in different structural and attribute contexts. For instance, in a fraud-detection use-case on financial graph, fraudulent transactions may involve different combinations of users, devices, merchants, and interaction patterns such that a global explanation must capture multiple recurring behaviors rather than a single representative structure.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing global GNN explainers approach this problem through generated graphs, recurring substructures, or exemplar-based descriptions (Yuan et al., 2020; Azzolin et al., 2023; Armgaan et al., 2025); however, applying these approaches to large heterogeneous graphs introduces challenges related to scalability and constructing explanations that are interpretable across structurally diverse neighborhoods.

Large real-world graphs introduce two additional challenges for global explanation that must be addressed to be acceptable in practice. First, explanation learning must operate on sampled subsets of nodes and neighborhoods as processing the full graph is computationally impractical. An explainer while processing a sampled batch encounters different nodes and neighborhood structures, leading to unstable extraction of patterns across multiple batches. Prior work has documented such sensitivity of GNN explanations to variations in the input, model, and explanation procedure (Kosan et al., 2024; Ribeiro et al., 2024). Here, we focus on the stability across samples used for explanation learning to ensure independently learned explanations remain consistent in the behavior they characterize even when their exact logical forms may differ. Second, the explanation modality itself must remain interpretable at scale. Multi-hop neighborhoods can be large, variable-sized, and heterogeneous, making direct subgraph-based characterizations difficult to summarize and compare. This motivates a compact, fixed-dimensional representation that preserves interpretable node- and neighborhood-level information.

To address the above challenges, we propose StableGX (Stable Global eXplainer), a global explainer for node-classification GNNs on large real-world graphs. StableGX first maps each node and its multi-hop neighborhood to a fixed-dimensional, interpretable representation of node attributes and neighborhood statistics. It then identifies representative exemplar populations in the GNN embedding space and learns a logical signature for each population. To improve the fidelity and stability of these signatures, we introduce boundary-aware sampling that identifies candidate points near the exemplar-population boundary and applies max–min selection to obtain broad coverage of that boundary. The selected points are used to train depth-constrained decision trees, whose positive paths define exemplar signatures; these signatures are then combined through Boolean symbolic regression to obtain class-level explanations. We further show theoretically that the boundary approximation error controls signature fidelity and stability, and that the proposed boundary-aware selection bounds this error in terms of boundary-candidate accuracy and finite-budget coverage.

In summary, our main contributions are:

- We study stability under sampling for global GNN explanations, focusing on whether explanations learned from different samples of the same graph retain consistent prediction behavior.
- We introduce StableGX, a scalable global explainer for node-classification GNNs on large real-world graphs. By representing multi-hop neighborhoods using fixed-dimensional interpretable statistics and learning exemplar signatures with depth-constrained decision trees, StableGX avoids operating directly on large variable-size neighborhoods during signature extraction and produces compact logical explanations.
- We propose a boundary-aware sampling strategy for exemplar-signature learning that identifies candidate points near the exemplar-population boundary and uses max–min selection to improve boundary coverage.
- We provide a theoretical analysis showing that boundary approximation error controls signature fidelity and stability, and derive a bound on this error for our boundary-aware sampling procedure, under some assumption.
- We evaluate StableGX across homogeneous and heterogeneous, homophilic and heterophilic node-classification benchmarks, including large real-world graphs. StableGX remains applicable at scales where existing global explainers face practical limitations and improves fidelity and stability on large heterogeneous and heterophilic graphs.

2 Related Work

Global GNN explanations. Most GNN explanation methods are local, identifying important features, edges, or subgraphs for individual predictions (Ying et al., 2019; Luo et al., 2020; Yuan et al.,

2021). Global explainers instead summarize recurring model behavior across multiple instances. XGNN and GNNInterpreter generate class-representative graphs (Yuan et al., 2020; Wang & Shen, 2024), while GLGExplainer derives class-level logical descriptions by aggregating local subgraph explanations (Azzolin et al., 2023). GNNXemplar Armgaan et al. (2025), the closest work to ours, identifies representative exemplar populations in the GNN embedding space and describes them using logical signatures. StableGX adopts the exemplar-based view but differs in two key aspects: it represents node neighborhoods through fixed-dimensional interpretable statistics and explicitly targets stable signature learning across different explanation samples using boundary-aware sampling, avoiding the expensive LLM-based signature extraction used by GNNXemplar.

Stability and heterogeneous GNN explanations. Existing work on heterogeneous GNN explainability is largely local or structure-based, including path-, subgraph-, and perturbation-based explanations Li et al. (2023); Deng & Shen (2024); Lu et al. (2026); Das & Heindorf (2026). Prior studies have shown that GNN explanations can be sensitive to perturbations in the input, model, or explanation procedure Kosan et al. (2024); Ribeiro et al. (2024). Our setting differs in that we study the consistency of *global* explanations learned from different sampled observations of the same fixed GNN. We focus particularly on large heterogeneous and heterophilic graphs, where both scalable explanation learning and stable characterization of recurring model behavior become challenging.

3 Definitions and Problem Setup

We now define some important definitions, formalize the problem statement and its key challenges.

Definition 1 (Graph) *A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi, X)$ over a node set $\mathcal{V}$, directed edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, where $\phi : \mathcal{V} \to T$ is the node type mapping function and $\psi : \mathcal{E} \to S$ is the edge type mapping function. T and S denote the predefined node types and edge types, where $|S|, |T| \geq 1$, and $X = \{x_v | v \in \mathcal{V}\}$ are node attributes where $x_v \in \mathbb{R}^d$ is the set of features characterizing each node.*

For simplicity, we present the methodology for a homogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$, with $|S| = |T| = 1$, and describe the necessary extensions for heterogeneous graphs where relevant.

We consider node classification on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ with label space $\mathcal{Y}$. A trained GNN $f_\theta(\cdot)$ produces, for each node $v \in \mathcal{V}$, a latent representation $\boldsymbol{h}_v \in \mathbb{R}^{d_h}$ and a predicted label $\hat{y}_v \in \mathcal{Y}$, i.e.,

$$f_\theta(v; \mathcal{G}) \mapsto (\boldsymbol{h}_v, \hat{y}_v).$$

For our explanation task, we consider a sampled explanation subgraph $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, X_r)$ for run r, where $\mathcal{V}_r \subseteq \mathcal{V}$, $\mathcal{E}_r \subseteq \mathcal{E}$, and X_r contains the attributes of nodes in $\mathcal{V}_r$. The prediction $\hat{y}_v$ and latent representation $\boldsymbol{h}_v$ are available for every $v \in \mathcal{V}_r$.

Definition 2 (Exemplars Armgaan et al. (2025)) *Exemplars refer to representative nodes within the graph that embody prototypical structural and attribute characteristics shared by many other nodes in the same GNN predicted class. In the context of node classification, these exemplars serve as anchors for human-understandable explanations. Each exemplar, e, is associated with a population $\mathcal{P}_e \subset \mathcal{V}$, which represents the subset of nodes that e exemplifies. A node $e \in \mathcal{V}$ is called an exemplar for the class $c \in Y$ if $|\mathcal{P}_e| \geq \tau$, where*

$$\mathcal{P}_e = \left\{ v \in \mathcal{V} \;\middle|\; \begin{array}{l} \hat{y}_v = \hat{y}_e = c, \\ d_h(\boldsymbol{h}_v, \boldsymbol{h}_e) < \epsilon \end{array} \right\} \tag{1}$$

where $d_h(\cdot, \cdot)$ is a distance function on embeddings of nodes from GNN, $\epsilon \in \mathbb{R}$ is a distance threshold, and $\tau \in \mathbb{N}$ is a minimum population size hyperparameter.

We follow the exemplar identification method of Armgaan et al. (2025) which automatically discovers the thresholds from data by exploring $k-$nearest neighbors along with a greedy selection algorithm.

For each node v, we associate an interpretable serialized representation $\boldsymbol{s}_v \in \mathbb{R}^{d_s}$ that summarizes the attributes of v and its k-hop neighborhood. We defer its construction to Section 5.1.

Definition 3 (Exemplar Signature) *For an exemplar e with associated population $\mathcal{P}_e$, an* exemplar signature σ_e *is an interpretable function over the serialized feature space, $\sigma_e = \mathbb{R}^{d_s} \to \{0, 1\}$, that characterized the membership in $\mathcal{P}_e$, i.e.,*

$$\sigma_e(\boldsymbol{s}_v) \approx \mathbb{I}[v \in \mathcal{P}_e] = m_e(\boldsymbol{s}_v)$$

where $m_e : \mathbb{R}^{d_s} \to \{0, 1\}$ is the true membership function.

Definition 4 (Fidelity and Stability) *Let $X \sim \mu$ denote an input drawn from the data distribution. For an exemplar e,*

$$\mathrm{Fid}_e = \Pr\left[\sigma_e(X) = m_e(X)\right], \tag{2}$$

$$\mathrm{Stab}_e = \Pr\left[\sigma_e^{(1)}(X) = \sigma_e^{(2)}(X)\right], \tag{3}$$

where $\sigma_e^{(1)}$ and $\sigma_e^{(2)}$ are signatures learned for the same exemplar in two different explanation runs. At the class level, let $m_c(X) = \mathbb{I}[f_\theta(X) = c]$. Then

$$\mathrm{GFid}_c = \Pr\left[g_c(X) = m_c(X)\right], \tag{4}$$

$$\mathrm{GStab}_c = \Pr\left[g_c^{(1)}(X) = g_c^{(2)}(X)\right], \tag{5}$$

where $g_c^{(1)}$ and $g_c^{(2)}$ are class-level explanations for class c learned in two different explanation runs.

Problem Setup. Given a trained GNN $f_\theta(\cdot)$ and a sampled explanation subgraph $\mathcal{G}_r$, we consider a set of exemplars $\mathcal{Z}$, where each exemplar e represents a population $\mathcal{P}_e$ of nodes with similar behavior in the GNN representation space. Our goal is to learn an interpretable signature σ_e that characterizes membership in each exemplar population and combine these signatures into a class-level explanation g_c that approximates the GNN's prediction behavior for class c. We require both *fidelity*, i.e., agreement with the corresponding GNN behavior, and *stability*, i.e., consistency across explanations learned from different samples $\mathcal{G}_r$.

Achieving these properties on large real-world graphs introduces two key challenges:

1 **Interpretable explanation modality at scale.** A target node's k-hop neighborhood can contain hundreds or thousands of nodes, vary substantially in size, and include multiple node and edge types. We therefore require a fixed-dimensional, interpretable representation of the node and its neighborhood that retains information relevant to the GNN's prediction.

2 **Stable exemplar-signature learning.** The signature σ_e should reliably characterize membership in $\mathcal{P}_e$ across different sampled graphs $\mathcal{G}_r$. This is challenging because samples can over-represent redundant dense regions while missing sparse regions important for defining the population boundary, causing the learned signature to depend on the particular sample.

4 Theoretical Motivation

We show how accurate representation of an exemplar population boundary controls the fidelity and stability of its learned signatures, independent of a signature extraction learner (like decision tree, symbolic regression, etc).

For exemplar e, let $(\mathcal{X}, d)$ be a metric space with membership function $m_e : \mathcal{X} \to \{0, 1\}$. Define its member region and boundary as

$$\mathcal{P}_e = \{x \in \mathcal{X} : m_e(x) = 1\}, \qquad \Gamma_e = \partial\mathcal{P}_e. \tag{6}$$

Let $\sigma_e : \mathcal{X} \to \{0, 1\}$ be the learned exemplar signature, and let $X \sim \mu$ denote a sample from the underlying data distribution. We measure the membership error of the signature as

$$R_e = \Pr\left[\sigma_e(X) \neq m_e(X)\right]. \tag{7}$$

For $r > 0$, let $T_r(\Gamma_e) = \{x \in \mathcal{X} : d(x, \Gamma_e) \leq r\}$ denote the set of points whose distance from the true boundary is at most r.

To quantify how accurately the boundary is represented during signature learning, let $\mathcal{S}_e^+$ and $\mathcal{S}_e^-$ denote the member and non-member points used to represent the population boundary. We define their boundary approximation error as

$$\eta_e = \max_{s\in\{+,-\}} d_H(\mathcal{S}_e^s, \Gamma_e), \tag{8}$$

where d_H denotes the Hausdorff distance. The following analysis treats η_e abstractly; Section 5 describes the construction of these boundary points and how the selection procedure controls η_e.

We make three assumptions, stated formally in Appendix G.

- A1 **Controlled interior error:** away from a fixed-width neighborhood of the true boundary, the signature's membership error is bounded by ξ.
- A2 **Local boundary recovery:** for sufficiently small boundary approximation error η_e, errors made by the learned signature near the boundary are confined to an $L\eta_e$ neighborhood of the true boundary, where $L > 0$ is a learner-dependent constant. This is a local response property of the learner: improved boundary approximation leads to a correspondingly smaller region in which boundary-related errors can occur.
- A3 **Bounded boundary mass:** the probability mass within distance r of the true boundary grows at most as Cr^α.

The corresponding technical conditions and complete formal statements are provided in Appendix G.

Theorem 1 (Boundary approximation controls signature fidelity and stability) *Under the above conditions, if $\eta_e \le \eta_0$, then*

$$R_e \le \xi + C(L\eta_e)^\alpha. \tag{9}$$

Consequently, the exemplar-membership fidelity satisfies

$$\mathrm{Fid}_e \ge 1 - \xi - C(L\eta_e)^\alpha. \tag{10}$$

Moreover, let $\sigma_e^{(1)}$ and $\sigma_e^{(2)}$ be signatures learned in two different runs for the same membership function m_e, with corresponding membership errors $R_e^{(1)}$ and $R_e^{(2)}$. Then their stability satisfies

$$\mathrm{Stab}_e \ge 1 - R_e^{(1)} - R_e^{(2)}. \tag{11}$$

The result follows by decomposing the membership error into an interior component, controlled by ξ, and a boundary component, controlled by the probability mass within an $L\eta_e$-neighborhood of the true boundary. Stability follows because two signatures can disagree only if at least one disagrees with the common membership function. Complete proofs are provided in Appendix G.2.

Thus, smaller boundary approximation error η_e directly improves the fidelity bound and, consequently, the stability bound. In the following section, we introduce a boundary-aware selection procedure and show how its boundary-identification and coverage properties control η_e.

5 Methodology

5.1 Interpretable Representation and Exemplars

Given a node $v \in \mathcal{V}_r$, we map its K-hop neighborhood to a fixed-dimensional interpretable representation $\boldsymbol{s}_v$. Let $\mathcal{N}_k(v)$ denote the set of nodes at exactly k hops from v in $\mathcal{G}_r$. For each hop $k \in \{1, \dots, K\}$, we compute the neighborhood feature summary

$$\boldsymbol{a}_k(v) = \mathrm{Agg}\left(\{\boldsymbol{x}_u : u \in \mathcal{N}_k(v)\}\right), \tag{12}$$

where Agg computes the mean for continuous features and category-wise counts for categorical features. Any structural attributes included in the node features, such as node degree, are aggregated in the same manner. We additionally retain the neighborhood size

$$b_k(v) = |\mathcal{N}_k(v)|. \tag{13}$$

The serialized representation is then

$$\boldsymbol{s}_v = \boldsymbol{x}_v \parallel \boldsymbol{a}_1(v) \parallel \cdots \parallel \boldsymbol{a}_K(v) \parallel b_1(v) \parallel \cdots \parallel b_K(v). \tag{14}$$

Thus, $\boldsymbol{s}_v$ provides a fixed-dimensional summary of the target node and its neighborhood in which every dimension corresponds to an explicit node- or neighborhood-level statistic, making it directly suitable for learning logical rules. For heterogeneous graphs, we compute these statistics separately by node type; details are provided in Appendix F.

Similar neighborhood-level summary statistics have been shown to provide effective representations for graph learning (Tang et al., 2023). They are also compatible with the stochastic neighborhood sampling commonly used by scalable GNNs (Hamilton et al., 2018; Hu et al., 2020); here, we use them as an interpretable modality for characterizing the behavior of the trained GNN.

Exemplar identification. We identify representative exemplars in the GNN embedding space following GNNXemplar (Armgaan et al., 2025). For each predicted class $c \in \mathcal{Y}$, candidate nodes with $\hat{y}_v = c$ are ranked by reverse-KNN representativeness, which measures how frequently a node appears among the nearest neighbors of other same-class nodes. A greedy coverage procedure then selects a compact set of exemplars $\mathcal{Z}$. Each exemplar $e \in \mathcal{Z}$ is associated with its population $\mathcal{P}_e$, as defined in Section 3. The complete selection procedure is provided in Appendix E.

5.2 Boundary-Aware Sampling

For each exemplar e, we seek a compact set of members and non-members that characterizes the boundary of its population $\mathcal{P}_e$. Random sampling although can reduce the internal error R_{int} since interior points are generally well represented in a sufficiently large sample, however it's less effective in reducing the boundary errors R_{bdry}. They does not explicitly prioritize regions where population membership changes, and may therefore provide limited coverage of the population boundary. We therefore first identify boundary candidates using local population membership in the serialized feature space and then apply max–min selection to improve their coverage.

For a node $v \in \mathcal{V}_r$, let

$$m_e(v) = \mathbb{I}[v \in \mathcal{P}_e],$$

and let $\mathcal{K}_q(v)$ denote its q nearest neighbors in the serialized feature space, excluding v itself, under $d_{\text{ser}}(u, v) = \|\boldsymbol{s}_u - \boldsymbol{s}_v\|_2$. We estimate the local membership probability as

$$\hat{p}_e(v) = \frac{1}{|\mathcal{K}_q(v)|} \sum_{u \in \mathcal{K}_q(v)} m_e(u). \tag{15}$$

Values near 0.5 indicate locally mixed neighborhoods containing both members and non-members. We therefore construct the two boundary-candidate sets

$$\mathcal{B}_e^+ = \{v \in \mathcal{P}_e : |\hat{p}_e(v) - 0.5| < \epsilon_b\}, \tag{16}$$

$$\mathcal{B}_e^- = \{v \in \mathcal{V}_r \setminus \mathcal{P}_e : |\hat{p}_e(v) - 0.5| < \epsilon_b\}, \tag{17}$$

where $\epsilon_b \in (0, 0.5)$ is a hyperparameter and controls the width of the locally mixed region.

Since we have limited sampling budget and the candidate sets may still contain many nearby points, we apply greedy max–min selection independently to each side. For either $\mathcal{B} \in \{\mathcal{B}_e^+, \mathcal{B}_e^-\}$, we initialize with the candidate closest to equal local membership,

$$v_1 = \arg\min_{v \in \mathcal{B}} |\hat{p}_e(v) - 0.5|,$$

and subsequently select

$$v_j = \arg \max_{v \in \mathcal{B} \setminus \{v_1, \ldots, v_{j-1}\}} \min_{u \in \{v_1, \ldots, v_{j-1}\}} d_{\text{ser}}(u, v). \tag{18}$$

Let $\text{MM}_M(\mathcal{B})$ denote the first $\min\{M, |\mathcal{B}|\}$ points selected by this procedure. The resulting member and non-member boundary sets are

$$\mathcal{D}_e^+ = \text{MM}_M(\mathcal{B}_e^+), \qquad \mathcal{D}_e^- = \text{MM}_M(\mathcal{B}_e^-), \tag{19}$$

where $M \in \mathbb{N}$ is the sampling budget. Max–min selection therefore retains boundary candidates that are distributed across different regions of the locally estimated population boundary.

Proposition 1 (Boundary approximation under max–min selection) *Suppose the boundary-candidate sets $\mathcal{B}_e^+$ and $\mathcal{B}_e^-$ constructed using Equations 16 and 17 satisfy*

$$d_H(\mathcal{B}_e^s, \Gamma_e) \le \delta_B, \qquad s \in \{+, -\}. \tag{20}$$

For each $s \in \{+, -\}$, define the optimal M-point covering radius as

$$\rho_M^{s,\star} = \min_{\substack{\mathcal{C} \subseteq \mathcal{B}_e^s \\ |\mathcal{C}| \le M}} \sup_{b \in \mathcal{B}_e^s} \min_{c \in \mathcal{C}} d(b, c),$$

and let

$$\rho_M^\star = \max_{s \in \{+,-\}} \rho_M^{s,\star}.$$

Then, using the max–min selected sets $\mathcal{D}_e^+$ and $\mathcal{D}_e^-$ to instantiate the boundary sets in the definition of η_e, we have

$$\eta_e \le \delta_B + 2\rho_M^\star. \tag{21}$$

Combining Proposition 2 with Theorem 1, whenever $\delta_B + 2\rho_M^\star \le \eta_0$, the signature fidelity satisfies

$$\mathrm{Fid}_e \ge 1 - \xi - C\left[L\left(\delta_B + 2\rho_M^\star\right)\right]^\alpha. \tag{22}$$

The proposition separates the boundary approximation error into two components: the boundary-candidate error δ_B and the finite-budget covering error $\rho_M^\star$. Constructing a sufficiently rich boundary-candidate set can improve the representation of the true boundary, while greedy max–min selection achieves a covering radius within a factor of two of the optimal M-point covering radius. A detailed statement and proof are provided in Appendices G.1 and G.2, respectively.

5.3 Signature Extraction and Class-Level Explanation

Signature extraction. Given a total sampling budget $\bar{M}$, we allocate a fraction ρ to boundary-selected points and the remaining fraction $1 - \rho$ to randomly sampled points. Let $\mathcal{R}_e^+$ and $\mathcal{R}_e^-$ denote the random samples from members and non-members, respectively, and let $\mathcal{D}_e^+$ and $\mathcal{D}_e^-$ denote the corresponding boundary-selected sets, with the total number of sampled points equal to $\bar{M}$. We train a depth-constrained decision tree $\mathcal{T}_e$ on $\mathcal{R}_e^+ \cup \mathcal{R}_e^- \cup \mathcal{D}_e^+ \cup \mathcal{D}_e^-$ in the serialized feature space to predict exemplar-population membership,

$$\mathcal{T}_e(\boldsymbol{s}_v) \approx m_e(v).$$

Each internal node of $\mathcal{T}_e$ defines a predicate over one dimension of $\boldsymbol{s}_v$, while each root-to-leaf path forms a conjunction of such predicates. Let Π_e denote the set of paths terminating at leaves that predict membership in $\mathcal{P}_e$. The exemplar signature is

$$\sigma_e(\boldsymbol{s}_v) = \bigvee\nolimits_{\gamma \in \Pi_e} \bigwedge\nolimits_{\pi \in \gamma} \pi(\boldsymbol{s}_v). \tag{23}$$

Thus, σ_e provides a compact Boolean characterization of $\mathcal{P}_e$ in terms of interpretable node and neighborhood statistics, with maximum tree depth d_t controlling its complexity.

Class-level explanation. The exemplar signatures capture distinct populations represented in the GNN embedding space. For each class c, we construct its explanation using signatures from *all* exemplars in $\mathcal{Z}$, irrespective of their classes:

$$g_c(\boldsymbol{s}_v) = \Psi_c\left(\{\sigma_e(\boldsymbol{s}_v) \mid e \in \mathcal{Z}\}\right), \tag{24}$$

where Ψ_c is learned via symbolic regression over Boolean operators AND, OR, and NOT. Thus, g_c combines the presence and absence of exemplar signatures to approximate $f_\theta(\cdot)$ for class c.

6 Experiments

We evaluate the fidelity and stability of StableGX and the effect of boundary-aware sampling on exemplar-signature learning near the population boundary and overall.

6.1 Experimental Setup

Datasets. We evaluate on twelve node-classification datasets spanning homogeneous and heterogeneous, homophilic and heterophilic settings. IEEE-CIS, Pokec, PDNS, and RCDD are taken from H2GB Lin et al. (2025); the remaining datasets follow the evaluation suite of GNNXemplar Armgaan et al. (2025): TAGCora, Citeseer, WikiCS, ogbn-arxiv, Amazon-Ratings, Questions, Minesweeper, and BA-Shapes. Dataset statistics are provided in Appendix A.3.

GNN models. We first train a node-classification GNN and explain its predicted labels. We use HGT Hu et al. (2020) for the main heterogeneous-graph experiments and GAT/GCN for homogeneous graphs, following GNNXemplar configurations where applicable. The predictive performance of the explained models are reported in Appendix B

Baselines and protocol. We compare against GNNXemplar Armgaan et al. (2025) and GLGExplainer Azzolin et al. (2023), both of which produce global explanations for node-classification GNNs. Explanations are learned from independently sampled explanation subgraphs and evaluated on a common held-out set of target nodes not used during explanation construction. For sampling ablations, we compare boundary-aware sampling against random sampling under the same sampling budget.

We evaluate signature's and global fidelity and stability using the definitions in Section 3. The corresponding probabilities are estimated empirically on the held-out evaluation set and averaged over five independent explanation runs. For Additional implementation details see Appendix A.

6.2 Global Fidelity and Stability

Table 1 compares StableGX with GNNXemplar and GLGExplainer in terms of global fidelity and stability. Global fidelity is measured by agreement between the global explanation and the GNN predictions on held-out nodes, while global stability measures agreement between explanations learned in different runs (5) and is averaged over all pairs of runs. OOM, OOC, and NF denote out-of-memory (for GLGExplainer), out-of-context (context limit for LLM reached in GNNXemplar), and no-valid-explanation outcomes (for GLGExplainer), respectively.

The results show different behavior across graph settings. On several homogeneous and homophilic datasets, GNNXemplar attains higher fidelity, while StableGX remains competitive on homogeneous heterophilic graphs and achieves stronger fidelity and stability on the large heterogeneous and heterophilic datasets. GLGExplainer completes on only 2 of the evaluated datasets and attains lower fidelity where valid explanations are obtained. GNNXemplar's stronger performance on TAGCora and ogbn-arxiv may stem from its LLM-based refinement, which can effectively leverage textual attributes and construct compact set-membership predicates over categorical features. We also observe distinct scalability limitations in the baselines. For GNNXemplar, serialized heterogeneous neighborhoods can exceed the available LLM context, preventing signature extraction on some large graphs. GLGExplainer frequently encounters memory or runtime failures in its subgraph-isomorphism stage.

6.3 Effect of Boundary-Aware Sampling

We next isolate the effect of boundary-aware sampling on exemplar-signature learning. For each fixed exemplar population P_e, we learn signatures using either random sampling or our boundary-aware sampling strategy. Across the two settings, we keep the exemplar populations, total sampling budget $\bar{M}$, decision-tree depth, and all other hyperparameters unchanged; thus, the only difference is how the training points used for signature extraction are selected.

We evaluate the resulting signatures in two complementary regions. First, to measure performance where boundary-aware sampling is explicitly intended to help, we consider a held-out set of locally mixed nodes,

$$V_{e,\mathrm{bdry}} = \{v \in V_{\mathrm{eval}} : |\hat{p}_e(v) - 0.5| < \epsilon_b\},$$

which is not used during signature learning. We measure signature fidelity on these boundary-region nodes and stability across independent explanation runs (see Figure 1 a). Boundary-aware sampling

Table 1: Global-fidelity and -stability across graph settings. Fidelity measures agreement with GNN predictions, while stability measures agreement across independent explanation runs. Stability is compared only when fidelity is at least 55%. Best fidelity and eligible stability values are bolded.

	Fidelity			Global Stability	
Dataset	**StableGX**	**GNNXemplar**	**GLGExplainer**	**StableGX**	**GNNXemplar**
Heterogeneous and heterophilic graphs					
IEEE-CIS	$\mathbf{0.76 \pm 0.03}$	0.41 ± 0.01	OOM	**0.83**	0.99
Pokec	$\mathbf{0.59 \pm 0.00}$	OOC	OOM	**0.833**	OOC
PDNS	$\mathbf{0.82 \pm 0.02}$	0.60 ± 0.03	NF	**0.856**	0.80
RCDD	$\mathbf{0.91 \pm 0.01}$	0.55 ± 0.02	OOM	**0.90**	0.78
Homogeneous and heterophilic graphs					
Amazon-Ratings	$\mathbf{0.83 \pm 0.01}$	0.82 ± 0.01	NF	**0.87**	0.86
Questions	$\mathbf{0.94 \pm 0.05}$	0.92 ± 0.01	OOM	**0.99**	0.97
Minesweeper	$\mathbf{0.95 \pm 0.02}$	0.86 ± 0.02	0.22 ± 0.07	0.87	**0.97**
BA-Shapes	0.92 ± 0.01	$\mathbf{0.93 \pm 0.00}$	0.30 ± 0.09	0.89	**0.96**
Homogeneous and homophilic graphs					
TAGCora	0.50 ± 0.01	$\mathbf{0.83 \pm 0.01}$	NF	0.35	**0.74**
Citeseer	0.82 ± 0.02	$\mathbf{0.92 \pm 0.03}$	NF	0.78	**0.81**
WikiCS	0.70 ± 0.04	$\mathbf{0.78 \pm 0.01}$	OOM	**0.74**	0.69
ogbn-arxiv	0.51 ± 0.04	$\mathbf{0.84} \pm 0.01$	OOM	0.35	**0.77**

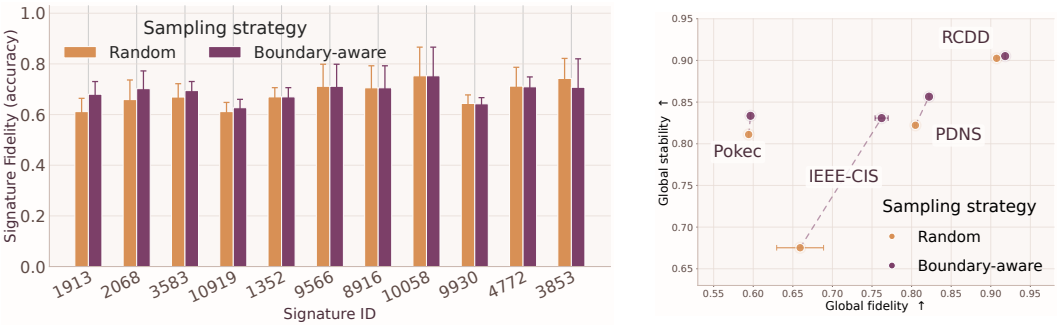


Figure 1: Effect of boundary-aware sampling in StableGX: (a) signature fidelity near the exemplar-population boundary of IEEE-CIS dataset, and (b) global-fidelity and -stability for random vs boundary-aware sampling.

yields higher fidelity in these locally mixed regions (bars are sorted by population sizes of exemplar), indicating that explicitly representing the population boundary improves the learned signatures.

Second, we evaluate these signatures on the held-out set V_{eval} to test whether gains extend beyond the boundary region. Figure 1 (b) shows that, under the same budget, boundary-aware sampling improves fidelity and stability, indicating that better boundary coverage improves not only locally ambiguous regions but also overall explanation quality and consistency.

7 Conclusion

We introduce StableGX, a global GNN explainer for stable and interpretable class-level explanations on large real-world graphs. StableGX summarizes sampled nodes using interpretable hop-wise feature and node-type statistics, identifies representative exemplar populations, and learns signatures with boundary-aware sampling and depth-constrained decision trees. Across homogeneous, heterogeneous, homophilic, and heterophilic node-classification datasets, StableGX is particularly effective on large heterogeneous and heterophilic graphs, where existing global explainers lose fidelity or fail to scale. We further show theoretically and empirically that boundary-aware sampling improves explanation fidelity and stability.

AI USE STATEMENT

In this work, we used generative AI tools for coding and rephrasing sentences. We do not use generative AI tools for autonomously running codes or generating entire draft.

Generative AI tools were additionally used for doing some preliminary background study. All AI-assisted outputs were reviewed, verified, and edited by the authors. In particular, all AI assisted text and codes are verified manually by the authors and literature review is thoroughly checked.

The authors take full responsibility for the final content of the paper, including all text, claims, code, analyses, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our theoretical and empirical results throughout the paper and appendix. The complete assumptions, formal statements, and proofs of our theoretical results are provided in Appendix G. The Appendix A further documents the datasets, GNN architectures, preprocessing and neighborhood-serialization procedure, exemplar identification, boundary-aware sampling, decision-tree and symbolic-regression settings, and all relevant hyperparameters used in our experiments. We report results across multiple independent explanation runs and specify the evaluation protocol used for fidelity and stability. Additional ablations and implementation details are also included in the Appendix A.2 to facilitate reproduction of the reported results. An anonymized implementation of StableGX, together with scripts for reproducing the experiments, is available at `https://anonymous.4open.science/r/ICLR_StableGXL-5B06/README.md`.

REFERENCES

Burouj Armgaan, Manthan Dalmia, Sourav Medya, and Sayan Ranu. Graphtrail: Translating GNN predictions into human-interpretable logical rules. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=fzlMza6dRZ`.

Burouj Armgaan, Eshan Jain, Harsh Pandey, Mahesh Chandran, and Sayan Ranu. Gnnxemplar: Exemplars to explanations – natural language rules for global gnn interpretability, 2025. URL `https://arxiv.org/abs/2509.18376`.

Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of gnns via logic combination of learned concepts, 2023. URL `https://arxiv.org/abs/2210.07147`.

Pallabee Das and Stefan Heindorf. Discrete diffusion-based model-level explanation of heterogeneous gnns with node features. In *Proceedings of the ACM Web Conference 2026*, pp. 1274–1285. ACM, April 2026. doi: 10.1145/3774904.3792518. URL `http://dx.doi.org/10.1145/3774904.3792518`.

Jiale Deng and Yanyan Shen. Self-interpretable graph learning with sufficient and necessary explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11749–11756, 2024. doi: 10.1609/aaai.v38i10.29059.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(85)90224-5. URL `https://www.sciencedirect.com/science/article/pii/0304397585902245`.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018. URL `https://arxiv.org/abs/1706.02216`.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer, 2020. URL `https://arxiv.org/abs/2003.01332`.

Mert Kosan, Samidha Verma, Burouj Armgaan, Khushbu Pahwa, Ambuj Singh, Sourav Medya, and Sayan Ranu. Gnnx-bench: Unravelling the utility of perturbation-based gnn explainers through in-depth benchmarking, 2024. URL https://arxiv.org/abs/2310.01794.

Tong Li, Jiale Deng, Yanyan Shen, Luyu Qiu, Yongxiang Huang, and Caleb Chen Cao. Towards fine-grained explainability for heterogeneous graph neural network, 2023. URL https://arxiv.org/abs/2312.15237.

Junhong Lin, Xiaojie Guo, Shuaicheng Zhang, Yada Zhu, and Julian Shun. When heterophily meets heterogeneity: Challenges and a new large-scale graph benchmark. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5607–5618, 2025. doi: 10.1145/3711896.3737421. URL https://doi.org/10.1145/3711896.3737421.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks, 2024. URL https://arxiv.org/abs/2310.00149.

Yifan Lu, Pengfei Jiao, Xuan Guo, Ziyun Zou, Yiwei Wang, Mengzhou Gao, Huaming Wu, and Imran Razzak. Towards robust heterogeneous graph explanations under structural perturbations. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz (eds.), *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pp. 1262–1273. ACM, 2026. doi: 10.1145/3774904.3792516. URL https://doi.org/10.1145/3774904.3792516.

Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks, 2022. URL https://arxiv.org/abs/2102.03322.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network, 2020. URL https://arxiv.org/abs/2011.04573.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking, and refining heterogeneous graph neural networks, 2021. URL https://arxiv.org/abs/2112.14936.

José Ribeiro, Lucas Cardoso, Vitor Santos, Eduardo Carvalho, Níkolas Carneiro, and Ronnie Alves. How reliable and stable are explanations of xai methods?, 2024. URL https://arxiv.org/abs/2407.03108.

Jianheng Tang, Fengrui Hua, Ziqi Gao, Peilin Zhao, and Jia Li. Gadbench: Revisiting and benchmarking supervised graph anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29628–29653, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5eaafd67434a4cfb1cf829722c65f184-Paper-Datasets_and_Benchmarks.pdf.

Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2297–2309, February 2023. ISSN 1939-3539. doi: 10.1109/tpami.2022.3170302. URL http://dx.doi.org/10.1109/TPAMI.2022.3170302.

Xiaoqi Wang and Han-Wei Shen. Gnninterpreter: A probabilistic generative model-level explanation for graph neural networks, 2024. URL https://arxiv.org/abs/2209.07924.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019. URL https://arxiv.org/abs/1903.03894.

Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining*, KDD '20, pp. 430–438. ACM, August 2020. doi: 10.1145/3394486.3403085. URL http://dx.doi.org/10.1145/3394486.3403085.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations, 2021. URL https://arxiv.org/abs/2102.05152.

Yue Zhang, David Defazio, and Arti Ramesh. Relex: A model-agnostic relational model explainer, 2020. URL https://arxiv.org/abs/2006.00305.

A Experimental Setup Details

This section covers details for the experiments and implementation details of the method along with hyperparameters used and other details like similarity metric used, etc.

A.1 Initial Setup

Hardware and library version used. We run all the experiments of this paper on a system with linux OS (Ubuntu 22.04.6 LTS), 4 Nvidia A30 GPU with (24GB VRAM), 503 GB RAM, Intel(R) Xeon(R) Silver 4314 CPU. We have cuda version 13.0 for Nvidia GPUs.

Hypr-parameters. To select exemplar per class we directly use the exemplar selection algorithm from Armgaan et al. (2025) also defined in App E. This algorithm uses k_{knn} and picks at max top k_{top} exemplar per class using cosine similarity (i.e., $d(\cdot,\cdot)$) until 95% coverage is reached for that class. For all H2GB datasets we pick $k_{\text{knn}} = 10$ and $k_{\text{top}} = 10$, for GNNXemplar dataset we reuse the exact values from their experiment ($k_{\text{knn}} = 5$ and $k_{\text{top}} = 5$). For a class c this gives us b_c number of points, where $0 < b_c \leq k_{top}$ thereby giving us total exemplars $b = \sum_c b_c$. Next we select $d_{\text{ser}}(\cdot,\cdot) = \text{cosine_similarity}$, we did do an ablation on this in App C. In serialization for $k - \text{hop}$ we use exactly the same number of layer for which the GNN is trained, ensuring our explanations only has information which the GNN uses for prediction. For B_j we pick it equal to size of positive population used for training Decision Tree $|\mathcal{D}_j^+|$. We use decision tree of depth $3 - 4$ across all datasets. We consistently use $\rho = 0.1$ across all datasets.

A.2 Implementation Details

Use of Random Forest to filter our irrelevant features. After serialization for heterogenous and heterophilous graph we often get a very high dimensional vector representation of a node and it's $k - hop$ neighborhood. To filter out few of the irrelevant dimensions we use random forest classifier. We train a random forest classifier with input as the serialized points and output and the GNN's prediction, we do this with 5-fold cross validation and finally discard the features which lowest features importance (calculated using sklearn's built-in method which uses mean decrease in impurity to calculate the importance of features).

Symbolic Regression. For symbolic regression we use 'AND', 'OR' and 'NOT' gates to combine multiple signatures together. This allows to use slightly complex however still highly interpretable signature combinations to make predictions aligned with GNN. Note that this symbolic regression layer tries to mimic the final classification head of the GNNs and hence just using simple 'OR' across all signature is often less powerful than what a classification head (generally few layer MLP) can do.

Sampling of population. Instead of working on the complete graph, we sample a subset of graph and work with that. So the extracted exemplars and signatures everything is restricted to that sample of the graph. This is done for heterogeneous-heterophilous graphs which are very huge in size.

A.3 Data Statistic

We report the complete statistics of the datasets used in Section 6 in the Tables 2 and 3.

B Training details of GNN

To train GNNs used for H2GB dataset, we follow the recommended pipeline by H2GB with their default configurations. They clearly specify the training test splits, learning algorithms and dimensions of exact layers. For IEEE-CIS its a 2 layer HGT and for the rest of H2GB datasets we use 3 layer HGT's default configuration. For the GNNXemplar dataset we use the exact same configuration used by them in their paper, they use layer 2 GAT for TAGCora dataset layer 2 GCN for WikiCS and a layer 3 GCN for the rest, further details are mentioned in Appendix of Armgaan et al.

Table 2: Dataset Statistics for Homogeneous Datasets

Homogenous-Heterophilous				
Dataset	**#nodes**	**#edges**	**#features**	**#cls**
Amazon-ratings	24,492	93,050	300	5
Minesweeper	10,000	39,402	7	2
Questions	48,921	153,540	301	2
BA-Shapes	300	4,110	0	4
Homogenous-Homophilous				
Dataset	**#nodes**	**#edges**	**#features**	**#cls**
TAGCora	2,708	10,556	1433	7
Citeseer	3,327	9,104	3703	6
WikiCS	11,701	216,123	300	10
obgn-arxiv	169,343	1,166,243	128	40

Table 3: Dataset Statistic for Heterogeneous Datasets

Heterogeneous-Heterophilous							
Dataset	**#nodes**	(types)	**#edges**	(types)	**#features**	**#classes**	(ratio)
IEEE-CIS	153,880	(12)	2,873,472	(22)	4823	2	(12:1)
Pokec	1,731,977	(16)	51,774,836	(31)	66	2	(1:1)
PDNS	1,173,558	(2)	76,797,104	(4)	10	2	(1:2)
RCDD	13,806,619	(7)	157,814,864	(14)	256	2	(11:1)

(2025). To train the GNNs for GNNXemplar dataset we use the exact same setting which is used by the authors and same train-test splits. The performance of the trained models are reported in Table 4.

C Ablation Study

In this sections we do some important ablation study for StableGX. This measures the overall impact of essential parameters choices for StableGX.

C.1 Ablation on size of boundary-aware samples

For a total budget of $\bar{M}$, we choose ρ fraction of points for boundary-aware sampling and rest $(1-\rho)$ fraction of points for random sapling. We do an ablation on ρ to understand it's effect on overall performance of StableGX. The results are presented in Table 5.

C.2 Ablation on Tree Depth

The depth of the decision tree controls the complexity of an exemplar signature. Recall that σ_e is obtained as the disjunction of all positive root-to-leaf paths of T_e, with each path represented as a conjunction of split predicates. Hence, constraining the tree depth bounds the number of predicates along each path and limits the complexity of the resulting signature. Shallower trees yield simpler explanations, whereas larger depths allow more complex exemplar-population boundaries to be represented.

We analyze this trade-off on IEEE-CIS in Figure 2. To isolate the effect of tree depth, we keep the serialized representations, exemplar populations, boundary-aware samples $\mathcal{D}_e^+$ and $\mathcal{D}_e^-$, and randomly sampled points $\mathcal{R}_e^+$ and $\mathcal{R}_e^-$ fixed across all configurations. For each exemplar $e \in \mathcal{Z}$, we then train decision trees with depths $d \in \{d_1, \ldots, d_D\}$ on the same training set

$$\mathcal{D}_e^+ \cup \mathcal{D}_e^- \cup \mathcal{R}_e^+ \cup \mathcal{R}_e^-,$$

Table 4: Performance of Black Box GNNs used for explaining on Datasets

Dataset	Arch.	Performance (Acc./F1)
IEEE-CIS	R-GAT	24.15 (F1)
Pokec	R-GAT	68.18 (Acc)
PDNS	R-GAT	96.42 (F1)
RCDD	R-GAT	77.24 (F1)
TAGCora	GAT	73.84 (Acc)
Citeseer	GCN	66.20 (Acc.)
WikiCS	GCN	78.40 (Acc.)
obgn-arxiv	GCN	62.58 (Acc.)
Amazon-ratings	GCN	46.43 (Acc.)
Minesweeper	GCN	80.36 (Acc.)
Questions	GCN	97.07 (Acc.)
BA-Sahpes	GCN	95.71 (Acc.)

Table 5: Effect of the boundary-sampling ratio ρ on global fidelity and stability.

ρ	Global Stability	Global Fidelity
0.05	0.830 ± 0.021	0.806 ± 0.026
0.10	$\mathbf{0.857 \pm 0.024}$	$\mathbf{0.822 \pm 0.031}$
0.20	0.834 ± 0.015	0.810 ± 0.017
0.40	0.841 ± 0.013	0.812 ± 0.026
0.70	0.836 ± 0.016	0.813 ± 0.026

and extract the corresponding signatures. We compare the resulting signatures using the signature-fidelity metric defined in Section 3.

D Qualitative Evaluations

Figures 3 and 4 show examples of signature-level rules extracted for two different exemplars from IEEE-CIS dataset. We also show the symbolic rules over the identified exemplar signatures to obtain class-level explanations. Although the exemplar-identification procedure often identifies more than five exemplars, symbolic regression typically selects only few of their signatures. For example, the symbolic rule learned for the IEEE-CIS dataset is

$$(\sigma_3 \wedge \sigma_5) \vee \sigma_8 \vee \sigma_{10},$$

where σ_j denotes the signature of the j-th exemplar. For this dataset, ten exemplars were identified for class 0. Rather than forming a disjunction over all ten signatures, symbolic regression selects

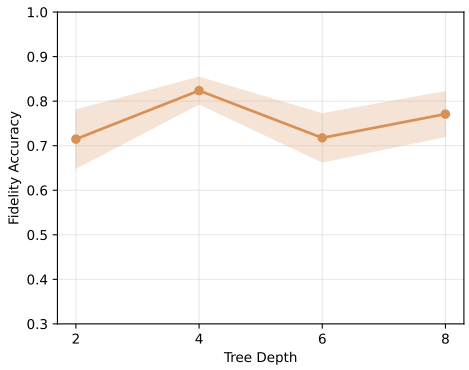


Figure 2: Effect of decision-tree depth on exemplar-signature fidelity on IEEE-CIS.

$\mathcal{R}_5$

$$V246 \leq 0, \quad V207 \leq 0,$$
$$C11 \leq 0, \quad V190 > 0$$

$\mathcal{R}_7$

$$V246 \leq 0, \quad V207 \leq 0,$$
$$C11 > 0, \quad V310 \leq 0$$

$\mathcal{R}_{12}$

$$V246 \leq 0, \quad V207 > 0,$$
$$C11 \leq 0, \quad id_{19} = 384$$

$\mathcal{R}_{14}$

$$V246 \leq 0, \quad V207 > 0,$$
$$C11 > 0, \quad C14 \leq 0$$

$\mathcal{R}_{19}$

$$V246 > 0, \quad V287 \leq 0,$$
$$V280 \leq 0, \quad V279 \leq 0$$

$\mathcal{R}_{22}$

$$V246 > 0, \quad V287 \leq 0,$$
$$V280 > 0, \quad V207 \leq 0$$

$\mathcal{R}_{23}$

$$V246 > 0, \quad V287 \leq 0,$$
$$V280 > 0, \quad V207 > 0$$

Figure 3: Rules extracted for fraudulent transactions in the IEEE-CIS dataset, for a fraudulent exemplar (node id = 5215).

only σ_3, σ_5, σ_8, and σ_{10}. Moreover, in the branch involving σ_3 and σ_5, the class is predicted only when both signatures are present, thereby capturing interactions between exemplar-level patterns.

Another example is the symbolic rule learned for class 1 of the PDNS dataset:

$$\neg\sigma_1 \vee (\sigma_4 \wedge \neg\sigma_0)\,.$$

For class 0, the exemplar-selection procedure identifies three exemplars, indexed by 0, 1, and 2. The class-1 explanation uses σ_4 as positive evidence and the negations of the opposing-class signatures σ_0 and σ_1 as negative evidence. This example illustrates how symbolic regression can construct more expressive class-level explanations while excluding signatures that are redundant or irrelevant to the prediction.

E Exemplar Selection Algorithm

Exemplars are selected independently for each GNN-predicted class. Let

$$\mathcal{V}_c = \{v \in \mathcal{V}^r_{\text{tar}} \mid \hat{y}_v = c\}$$

denote the target nodes predicted as class c.

Reverse-neighbor popularity. For every $u \in \mathcal{V}_c$, we compute its q nearest neighbors in the GNN embedding space, denoted by $\mathrm{KNN}_q(u)$. The reverse nearest-neighbor set of a candidate node v is

$$\mathcal{R}_q(v) = \{u \in \mathcal{V}_c \mid v \in \mathrm{KNN}_q(u)\}.$$

Its popularity is then

$$p(v) = |\mathcal{R}_q(v)|.$$

A high popularity indicates that v lies near many similar nodes (w.r.t the GNN because of the use of GNN embedding) and can therefore represent a large region of the embedding space.

$\mathcal{R}_5$

$$V232 \leq 0, \quad V186 \leq 0,$$
$$V186 > 0, \quad V285 \leq 0$$

$\mathcal{R}_9$

$$V232 \leq 0, \quad V186 > 0,$$
$$V318 \leq 0, \quad V274 \leq 0$$

$\mathcal{R}_{13}$

$$V232 \leq 0, \quad V186 > 0,$$
$$V318 > 0, \quad V290 > 0$$

$\mathcal{R}_{15}$

$$V232 > 0, \quad V257 \leq 0$$

$\mathcal{R}_{17}$

$$V232 > 0, \quad V257 > 0,$$
$$V192 \leq 0$$

Figure 4: Rules extracted for fraudulent transactions in the IEEE-CIS dataset, for a different exemplar (node id = 6046)

Greedy diverse selection. Candidates are considered in decreasing order of popularity. Let $\mathcal{U}$ denote the set of nodes already covered, initially $\mathcal{U} = \varnothing$. When a candidate v is selected as an exemplar, it covers itself and its reverse-neighbor set:

$$\mathcal{U} \leftarrow \mathcal{U} \cup \{v\} \cup \mathcal{R}_q(v).$$

Candidates already contained in $\mathcal{U}$ are skipped, preventing the selection of redundant exemplars from an already represented region. The procedure continues until

$$\frac{|\mathcal{U}|}{|\mathcal{V}_c|} \geq 0.95.$$

The selected nodes form the exemplar set for class c, and $\{v\} \cup \mathcal{R}_q(v)$ defines the population represented by each exemplar.

F Heterogeneous Specific Changes

In Section 5.1 the serialization is only shown for homogeneous graphs, for heterogeneous graphs we take the below mentioned approach.

Notice for a heterogeneous graph we also have a type t associated with every node ($\phi(\cdot)$), so we get

$$\boldsymbol{a}_{k,t}(v) = \frac{1}{|\mathcal{N}_{k,t}(v)} \sum_{u \in \mathcal{N}_{k,t}(v)} \boldsymbol{x}_u,$$
$$b_{k,t}(v) = |\mathcal{N}_{k,t}(v)|.$$

where $\mathcal{N}_{k,t}(v) = \{u \in \mathcal{N}_k(v) | \phi(u) = t\}$, concatenating them gives us

$$\boldsymbol{a}(v) = \Big\|_{t \in \{1,\ldots,T\}} \left(\|_{k \in \{1,\ldots,K\}} \boldsymbol{a}_{k,t}(v) \right),$$
$$\boldsymbol{b}(v) = \Big\|_{t \in \{1,\ldots,T\}} \left(\|_{k \in \{1,\ldots,K\}} b_{k,t}(v) \right).$$

After this, we can use equation 14 to get the serialized representation $\boldsymbol{s}_v$.

G Theoretical Proofs

G.1 Boundary Error, Fidelity, and Stability of Exemplars

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a metric space with metric d, and let

$$m_e : \mathcal{X} \to \{0, 1\}$$

denote the true membership function associated with exemplar e. Define the member region and its boundary as

$$A_e = \{x \in \mathcal{X} \mid m_e(x) = 1\}, \qquad \Gamma_e = \partial A_e. \tag{25}$$

For $r > 0$, let

$$T_r(\Gamma_e) = \{x \in \mathcal{X} \mid d(x, \Gamma_e) \leq r\} \tag{26}$$

denote the radius-r tube around the true boundary, and let $X \sim \mu$ denote an input drawn from the underlying data distribution.

Let $\mathcal{D}_e^+$ and $\mathcal{D}_e^-$ denote the positive- and negative-side boundary points used to learn σ_e. We define their boundary approximation error as

$$\eta_e := \max_{s \in \{+,-\}} d_H(\mathcal{D}_e^s, \Gamma_e), \tag{27}$$

where d_H denotes the Hausdorff distance. We later show how the proposed boundary-selection procedure controls η_e.

Assumption A1 (Controlled interior error). For a fixed boundary-region width $\tau > 0$, the learned signature σ_e has bounded error away from the boundary,

$$R_{\text{int}} := \Pr[\sigma_e(X) \neq m_e(X),\ X \notin T_\tau(\Gamma_e)] \leq \xi. \tag{28}$$

This is an assumption on the learned signature away from the boundary and bounds the component of membership error not attributed to boundary recovery.

Remark (Interpretation of Assumption A1). Assumption A1 isolates the component of the prediction error arising away from the decision boundary. This assumption is intended to capture the fact that the interior classification problem can be comparatively well supported by passive sampling. In particular, suppose that the training examples $X_1, \ldots, X_n$ are drawn i.i.d. from $\mathcal{P}$, and let

$$p_{\text{int}} := \Pr[X \notin T_\tau(\Gamma_e)]\,.$$

Then the number of sampled interior points,

$$N_{\text{int}} = \sum_{i=1}^{n} \mathbf{1}\{X_i \notin T_\tau(\Gamma_e)\}\,,$$

satisfies

$$N_{\text{int}} \sim \text{Binomial}(n, p_{\text{int}}).$$

Hence, whenever $p_{\text{int}} > 0$, the number of available interior observations grows linearly with the sample size and concentrates around np_{int}. For a learning procedure whose error decreases with the number of representative interior observations, passive sampling can therefore yield a small interior error. Assumption A1 abstracts this property of the learner rather than deriving it from Assumptions A2–A3 and some additional learnability / generalization assumptions on learned signature.

Assumption A2 (Local boundary recovery). There exist constants $L > 0$ and $\eta_0 > 0$ such that, whenever the boundary approximation error satisfies $\eta_e \leq \eta_0$, the mistakes made by the learned signature within the boundary region are confined to an $L\eta_e$-neighborhood of the true boundary,

$$\{x \in \mathcal{X} \mid \sigma_e(x) \neq m_e(x)\} \cap T_\tau(\Gamma_e) \subseteq T_{L\eta_e}(\Gamma_e). \tag{29}$$

This is an assumption on the *local* behavior of the learned signature and links improved boundary approximation to a smaller region in which boundary-related mistakes can occur. Importantly, it does not assume low risk directly; rather, it characterizes how the learning algorithm responds to increasingly accurate boundary information. If a learner cannot exploit improved boundary information in a given setting, no corresponding reduction in local error can be expected; our analysis therefore concerns regimes in which the learner is responsive to improvements.

Assumption A3 (Boundary mass). There exist constants $C > 0$, $\alpha > 0$, and $r_0 > 0$ such that, for all $0 < r \le r_0$,

$$\Pr\left[X \in T_r(\Gamma_e)\right] \le C r^{\alpha}. \tag{30}$$

This is an assumption on the data distribution around the true boundary

We choose the small-error regime η_0 such that

$$L\eta_0 \le \min\{\tau, r_0\}. \tag{31}$$

Theorem 2 (Theorem 1, Boundary error controls fidelity and stability) *Suppose Assumptions A1–A3 hold and $\eta_e \le \eta_0$. Then the exemplar-membership error*

$$R_e := \Pr[\sigma_e(X) \neq m_e(X)]$$

satisfies

$$R_e \le \xi + C(L\eta_e)^{\alpha}.$$

Consequently, the membership fidelity satisfies

$$\mathrm{Fid}_e := \Pr[\sigma_e(X) = m_e(X)] \ge 1 - \xi - C(L\eta_e)^{\alpha}. \tag{32}$$

Furthermore, consider two independently learned signatures $\sigma_e^{(1)}$ and $\sigma_e^{(2)}$ for the same membership function m_e, with corresponding interior-error bounds ξ_1, ξ_2 from Assumption A1 and boundary approximation errors $\eta_e^{(1)}, \eta_e^{(2)} \le \eta_0$. Their functional disagreement satisfies

$$\Pr\left[\sigma_e^{(1)}(X) \neq \sigma_e^{(2)}(X)\right] \le \xi_1 + \xi_2 + CL^{\alpha}\left[(\eta_e^{(1)})^{\alpha} + (\eta_e^{(2)})^{\alpha}\right]. \tag{33}$$

Thus, decreasing the boundary approximation error improves both membership fidelity and functional agreement across independently learned signatures, provided the interior error remains controlled.

Now for Proposition 1 we break it here into Assumption A4 and Proposition 2. This clearly separates the assumption required and with the rest of proposition.

Assumption A4 (Boundary-candidate accuracy). For $s \in \{+, -\}$, let $\mathcal{B}_e^s$ denote the boundary-candidate set identified on side s of the membership boundary. We assume that each candidate set approximates the true boundary with Hausdorff error at most δ_B,

$$d_H(\mathcal{B}_e^s, \Gamma_e) \le \delta_B, \qquad s \in \{+, -\}. \tag{34}$$

This is an assumption on the boundary-identification procedure. It requires both that the identified candidates lie close to the true boundary and that every region of the true boundary is represented by nearby candidates from each side.

For a selection budget M, we obtain the positive- and negative-side boundary sets as

$$\mathcal{D}_e^s = \mathrm{MM}_M(\mathcal{B}_e^s), \qquad s \in \{+, -\}, \tag{35}$$

where $\mathrm{MM}_M(\cdot)$ denotes greedy max–min selection.

Define the optimal M-point covering radius for side s as

$$\rho_M^{s,\star} = \min_{\substack{\mathcal{C} \subseteq \mathcal{B}_e^s \\ |\mathcal{C}| \le M}} \sup_{b \in \mathcal{B}_e^s} d(b, \mathcal{C}),$$

and let

$$\rho_M^{\star} = \max_{s \in \{+,-\}} \rho_M^{s,\star}.$$

Proposition 2 (Proposition 1 Boundary approximation under max–min selection) *Suppose Assumption A4 holds and d is a metric. Then greedy max–min selection with budget M satisfies*

$$\eta_e \le \delta_B + 2\rho_M^{\star}. \tag{36}$$

Thus, the boundary approximation error η_e is controlled by the boundary-candidate error δ_B and the optimal finite-budget covering radius $\rho_M^{\star}$.

Combining Theorems 2 and Proposition 2, whenever

$$\delta_B + 2\rho_M^\star \le \eta_0, \tag{37}$$

the exemplar-membership fidelity satisfies

$$\mathrm{Fid}_e \ge 1 - \xi - C\left[L\left(\delta_B + 2\rho_M^\star\right)\right]^\alpha. \tag{38}$$

Thus, fidelity is governed by two complementary components:

- the interior error ξ, which captures errors away from the membership boundary; and
- the boundary-error term, which decreases as the boundary-candidate error δ_B and the finite-budget covering radius $\rho_M^\star$ decrease.

G.2 Proofs for theorems

Proof 1 (Proof of Theorem 2) *Define the boundary-region error as*

$$R_{\mathrm{bdry}} := \Pr\left[\sigma_e(X) \ne m_e(X),\ X \in T_\tau(\Gamma_e)\right].$$

Since $T_\tau(\Gamma_e)$ and its complement partition the input space, the total membership error decomposes as

$$R_e = R_{\mathrm{int}} + R_{\mathrm{bdry}}.$$

By Assumption A1,

$$R_{\mathrm{int}} \le \xi.$$

By Assumption A2, whenever $\eta_e \le \eta_0$, all mistakes within the boundary region are confined to $T_{L\eta_e}(\Gamma_e)$. Therefore,

$$R_{\mathrm{bdry}} \le \Pr\left[X \in T_{L\eta_e}(\Gamma_e)\right].$$

Since $\eta_e \le \eta_0$ and $L\eta_0 \le r_0$, we have $L\eta_e \le r_0$. Hence, by Assumption A3,

$$R_{\mathrm{bdry}} \le C(L\eta_e)^\alpha.$$

For $\eta_e = 0$, the same result follows by taking $r \downarrow 0$ in Assumption A3.

Combining the interior and boundary terms gives

$$R_e \le \xi + C(L\eta_e)^\alpha.$$

Since

$$\mathrm{Fid}_e = 1 - R_e,$$

we obtain

$$\mathrm{Fid}_e \ge 1 - \xi - C(L\eta_e)^\alpha.$$

For two independently learned signatures $\sigma_e^{(1)}$ and $\sigma_e^{(2)}$ corresponding to the same membership function m_e, observe that

$$\left\{\sigma_e^{(1)}(X) \ne \sigma_e^{(2)}(X)\right\} \subseteq \left\{\sigma_e^{(1)}(X) \ne m_e(X)\right\} \cup \left\{\sigma_e^{(2)}(X) \ne m_e(X)\right\}.$$

Indeed, if both signatures agree with $m_e(X)$, then they must agree with each other. Applying the union bound therefore gives

$$\Pr\left[\sigma_e^{(1)}(X) \ne \sigma_e^{(2)}(X)\right] \le R_e^{(1)} + R_e^{(2)}.$$

Applying the membership-error bound to each run,

$$R_e^{(i)} \le \xi_i + CL^\alpha\left(\eta_e^{(i)}\right)^\alpha, \qquad i \in \{1,2\},$$

and substituting these bounds yields

$$\Pr\left[\sigma_e^{(1)}(X) \ne \sigma_e^{(2)}(X)\right] \le \xi_1 + \xi_2 + CL^\alpha\left[\left(\eta_e^{(1)}\right)^\alpha + \left(\eta_e^{(2)}\right)^\alpha\right].$$

This proves the result.

Proof 2 (Proof of Proposition 2) *Fix $s \in \{+, -\}$ and define the covering radius achieved by the max–min selected set as*

$$\rho_M^s := \sup_{b \in \mathcal{B}_e^s} d\left(b, \mathcal{D}_e^s\right).$$

We now show an already established results from Gonzalez (1985), which proves that the greedy max-min selection satisfies,

$$\rho_M^s \leq 2\rho_M^{s,\star}.$$

If $|\mathcal{B}_e^s| \leq M$, all candidates are selected, so $\rho_M^s = 0$ and the claim holds trivially. Otherwise, let $v_1, \ldots, v_M$ denote the points selected by greedy max–min, and let v_{M+1} denote the point that would be selected next. By the greedy selection rule,

$$\rho_M^s = \min_{1 \leq i \leq M} d(v_{M+1}, v_i).$$

Moreover, every pair among $v_1, \ldots, v_{M+1}$ is separated by at least ρ_M^s.

Now consider an optimal set of at most M centers achieving covering radius $\rho_M^{s,\star}$. These centers induce at most M balls of radius $\rho_M^{s,\star}$ that cover $\mathcal{B}_e^s$. Since there are $M + 1$ greedy points, the pigeonhole principle implies that two of them, say v_i and v_j, lie in the same optimal ball. By the triangle inequality,

$$d(v_i, v_j) \leq 2\rho_M^{s,\star}.$$

Since every pair of greedy points is separated by at least ρ_M^s,

$$\rho_M^s \leq d(v_i, v_j) \leq 2\rho_M^{s,\star}.$$

Thus,

$$\rho_M^s \leq 2\rho_M^{s,\star}.$$

We now relate this covering radius to the true boundary. Since

$$\mathcal{D}_e^s \subseteq \mathcal{B}_e^s,$$

Assumption A4 implies

$$\sup_{z \in \mathcal{D}_e^s} d(z, \Gamma_e) \leq \delta_B.$$

Conversely, consider any $x \in \Gamma_e$. Assumption A4 guarantees that there exists $b \in \mathcal{B}_e^s$ such that

$$d(x, b) \leq \delta_B.$$

By the definition of ρ_M^s, there exists $z \in \mathcal{D}_e^s$ satisfying

$$d(b, z) \leq \rho_M^s.$$

Therefore, by the triangle inequality,

$$d(x, z) \leq d(x, b) + d(b, z) \leq \delta_B + \rho_M^s.$$

Hence,

$$\sup_{x \in \Gamma_e} d(x, \mathcal{D}_e^s) \leq \delta_B + \rho_M^s.$$

Combining the two directed distances gives

$$d_H(\mathcal{D}_e^s, \Gamma_e) \leq \delta_B + \rho_M^s \leq \delta_B + 2\rho_M^{s,\star}.$$

Finally, taking the maximum over $s \in \{+, -\}$ and using the definitions of η_e and $\rho_M^\star$ gives

$$\begin{aligned} \eta_e &= \max_{s \in \{+,-\}} d_H(\mathcal{D}_e^s, \Gamma_e) \\ &\leq \delta_B + 2 \max_{s \in \{+,-\}} \rho_M^{s,\star} \\ &= \delta_B + 2\rho_M^\star. \end{aligned}$$

This proves the result.

You may include other additional sections here.