From Abductive Explanations to Global Logical Rules for Node Classification in SGCs

Bryan Lima Cavalcante, Thiago Alves Rocha

Instituto Federal do Ceará (IFCE)
Brazil

bryan.lima.cavalcante07@aluno.ifce.edu.br

thiago.alves@ifce.edu.br ✉

Abstract. *Graph Neural Networks (GNNs) have achieved remarkable performance in node classification tasks, motivating growing interest in methods capable of explaining their predictions. Recent logic-based approaches, such as LogicXGNN, derive global logical rules for Graph Neural Networks (GNNs) from collections of explanatory subgraphs. While informative, these subgraphs may contain redundant structural information that is specific to individual nodes, potentially limiting the generality of the extracted rules. In this work, we propose a logic-based framework for node classification in Simple Graph Convolution (SGC) networks that uses minimal abductive explanations as an intermediate representation for rule extraction. For each node, we compute a minimal set of node-feature pairs sufficient to preserve the predicted class. These explanations are then used to train decision trees from which global logical rules are extracted. Experiments on benchmark datasets show that the proposed framework produces compact global rules while maintaining high fidelity to the original SGC model.*

1. Introduction

Graph Neural Networks (GNNs) have become the standard tool for node classification on attributed graphs, with applications ranging from document categorization in citation networks to entity classification in knowledge graphs [Ju et al. 2024, Hu et al. 2020]. As these models inform decisions in scientific and operational pipelines, both explaining individual node predictions and characterising overall model behaviour have become central concerns [Yuan et al. 2023]. Two families of methods coexist in the literature. *Instance-level* methods explain a single prediction by highlighting a supporting subgraph or feature subset, but do not generalise across the graph. *Model-level* methods extract global rules that describe the classifier as a whole, but currently rely on heuristic aggregation and carry no formal sufficiency guarantee at the node level. This work targets the intersection of the two: global logical rules for node classification whose per-node components carry an explicit sufficiency certificate.

The closest existing work in this direction is LogicXGNN [Geng et al. 2026], which extracts per-class DNF rules from a GNN by characterising each node through two complementary predicates. The first is a structural component, a Weisfeiler-Lehman hash of the node's multi-hop receptive field. The second is an embedding component, obtained by binarising a small set of informative dimensions of the learned node representation

against thresholds selected by an auxiliary decision tree on graph-level mean-pooled embeddings. Correctly classified instances are then encoded as binary vectors over these predicates, and a decision tree fit against the model's prediction distils a per-class set of rules; a follow-up grounding step links each predicate back to input features. However, its main exposition and experiments target graph classification, with baselines such as GLGExplainer and GraphTrail [Azzolin et al. 2023, Armgaan et al. 2024] evaluated in the same setting; the extension to node classification is only sketched. Moreover, the predicates are derived from statistical criteria and carry no formal guarantee that a fired rule preserves the model's prediction under perturbations of the remaining inputs.

A parallel line of work uses abductive explanations (AXps) [Ignatiev et al. 2019, Marques-Silva and Ignatiev 2022], which provide exactly this kind of guarantee in other domains: an AXp is a minimal subset of input assignments whose values, alone, entail the predicted class regardless of how the remaining inputs are set. AXps are computable in polynomial time when the classifier is linear in the inputs being explained [Marques-Silva et al. 2020], but have not been directly applied to GNNs, in part because most GNN scores are not linear in the input features. Simple Graph Convolution (SGC) [Wu et al. 2019] is a natural target for closing this gap: it matches the accuracy of GCN and other state-of-the-art GNNs on standard node classification benchmarks while its score reduces to a linear function of the input features after a fixed feature propagation. We exploit this linearity to bring formal per-node guarantees to rule-based global explanation for GNNs.

To address the limitations of these previous lines of work, we propose AXSGC (Abduction-based Explanations for SGC), a method that operates in three stages. First, for each node v, we compute a node-feature abductive explanation (NF-AXp): a minimal set of node-feature pairs (u, j) sufficient to preserve the SGC prediction at v under any perturbation of the remaining pairs. Second, we encode each NF-AXp as a Boolean vector over distance-indexed predicates of the form "feature j appears in the explanation of v through a node at hop distance k", which abstracts away the concrete identity of the contributing neighbour while preserving its structural distance from v. Third, we fit a single shallow decision tree over these vectors with the SGC prediction as target; its root-to-leaf paths yield per-class DNF rules over the predicate vocabulary.

The empirical evaluation is organised around three research questions. **(Q1)** Does the global rule set extracted reproduce the predictions of the underlying SGC, and at what cost in number of rules? **(Q2)** How large are the NF-AXps that AXSGC manipulates internally and the resulting global rules? **(Q3)** Are individual rules informative enough to admit a structural reading, and how much of the data does a handful of them cover? We evaluate on four standard benchmarks (BAShapes, Cora, Citeseer, PubMed) against LogicXGNN [Geng et al. 2026] under matched decision-tree capacity.

Overall, the contributions of this work are the following:

- A specialisation of the AXp to node-feature pairs (*NF-AXp*) in SGC classifiers, computable in polynomial time via the linearity of the SGC score.
- A *global rule extraction procedure* that aggregates per-node NF-AXps into a Boolean matrix and reads per-class DNF rules from a single shallow decision tree.
- An *empirical evaluation* on four benchmarks against LogicXGNN: fidelity up to 30.2 pp higher with up to 83.8% fewer rules.

2. Preliminaries

This section fixes the notation used in the remainder of the paper. We first recall the SGC classifier and isolate the property that makes it amenable to formal explanation. We then recall the abstract notion of abductive explanation from formal explainability, which serves as the conceptual foundation for the method developed in the next section.

2.1. Simple Graph Convolution

Let $G = (V, E)$ be an undirected graph with $n = |V|$ nodes, adjacency matrix $A \in \{0,1\}^{n\times n}$, and node-feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, where d is the number of node features and $X_{u,j}$ is the value of feature j at node u. In most datasets considered in this work, node attributes are represented as binary bag-of-words vectors. Hence, $X_{u,j} \in \{0,1\}$, where $X_{u,j} = 1$ indicates that term j occurs in the document associated with node u, and $X_{u,j} = 0$ indicates its absence. Then, throughout the paper, we assume Boolean node features. Let $I \in \mathbb{R}^{n\times n}$ be the identity matrix. The augmented adjacency $A + I$ adds a self-loop to every node, so each node aggregates its own features alongside those of its neighbours. Let $\tilde{D}$ be the diagonal degree matrix of $A + I$, with $\tilde{D}_{ii} = \sum_{k=1}^{n}(A + I)_{ik}$, where $\tilde{D}_{ii}$ is the degree of node i in the augmented graph. The symmetric normalisation by $\tilde{D}^{-1/2}$ on both sides bounds the spectrum of the propagation operator, defined as

$$\hat{A} \;=\; \tilde{D}^{-1/2}\,(A + I)\,\tilde{D}^{-1/2}. \tag{1}$$

The entry $\hat{A}_{v,u}$ encodes the one-step weighted influence of node u on node v after self-loop augmentation and symmetric normalisation.

Let $\Theta \in \mathbb{R}^{d\times C}$ be the learned weight matrix, where C is the number of classes and each column $\Theta_{:,c} \in \mathbb{R}^{d}$ is the linear classifier for class c, with $\Theta_{j,c}$ the weight assigned to feature j when scoring class c. Let $K \in \mathbb{N}$ be the number of times feature propagation is applied; K plays the same role as the number of message-passing layers in a GCN, controlling how far information can travel before classification. The score matrix is given by

$$\hat{Y} \;=\; \mathrm{softmax}\big(\hat{A}^{K}\,\mathbf{X}\,\Theta\big), \tag{2}$$

with row-wise softmax, where row v of $\hat{Y}$ is the predicted class distribution at node v. The predicted class of node v is

$$c^{*}(v) \;=\; \arg\max_{c\in\{1,\ldots,C\}} \big(\hat{A}^{K}\,\mathbf{X}\,\Theta\big)_{v,c}, \tag{3}$$

where $c^{*}(v)$ is the class SGC assigns to node v.

2.2. Abductive Explanations

An abductive explanation (AXp) [Ignatiev et al. 2019, Marques-Silva and Ignatiev 2022] is a minimal set of input assignments whose values determine the model's prediction.

Definition 1 (Abductive explanation)**.** *Let $\phi : \mathcal{X} \to \mathcal{Y}$ be a classifier whose feature space factors as $\mathcal{X} = \prod_{i\in F} \mathcal{X}_i$, where F is a finite index set of features and $\mathcal{X}_i$ is the domain of feature i, and let $x \in \mathcal{X}$ be an input with prediction $y = \phi(x)$. A subset $H \subseteq F$ is an* abductive explanation *(AXp) for $\phi(x) = y$ if (a) $\phi(x') = y$ for every $x' \in \mathcal{X}$ with $x'_i = x_i$ for all $i \in H$, and (b) no proper subset of H satisfies (a).*

In Definition 1, ϕ is the classifier whose prediction is being explained, F is the set of explainable inputs, $\mathcal{X}_i$ is the domain of input i, x is the observed input, $y = \phi(x)$ is the prediction at x, H is the subset of inputs the explanation holds fixed, and x' ranges over all alternative inputs that agree with x on H. Clause (a) says that as long as the values on H are pinned, no perturbation of the remaining inputs $F \setminus H$ can change the prediction. Clause (b) identifies the inputs the model actually relies on rather than a superset that also happens to be sufficient: without (b), $H = F$ would trivially satisfy (a).

3. AXps for Selecting Node-Feature Pairs

The construction developed in this section turns the sufficiency and minimality conditions of Definition 1 into a concrete procedure for explaining SGC node predictions. We proceed in two steps. Subsection 3.1 reformulates the SGC classification rule as a system of linear inequalities over node-feature pairs, exposing the structure that the AXp computation exploits. Subsection 3.2 adapts Definition 1 to a node v of an SGC classifier and gives a greedy deletion procedure that returns an AXp at node-feature using a number of model queries linear in the size of the candidate set.

3.1. Linear inequalities for SGC predictions

Fix $\hat{A}$ and Θ. The entry $(\hat{A}^K)_{v,u}$ is the total weight with which the feature vector of u enters the score at v after K propagation steps; it is nonzero exactly when u is reachable from v in at most K hops on the augmented graph. This entry depends on u only through structural quantities of the augmented graph (degrees and the number of walks of length up to K from v to u), and in particular it is zero whenever the shortest-path distance from v to u exceeds K. We write $\mathcal{N}_K(v) := \{u \in V : (\hat{A}^K)_{v,u} \neq 0\}$ for the K-hop receptive field of v and note that $v \in \mathcal{N}_K(v)$. Only feature values at nodes in $\mathcal{N}_K(v)$ can change the score at v, so they are the only candidates for an explanation of $c^*(v)$.

Expanding $\hat{A}^K \mathbf{X}$ entrywise as $(\hat{A}^K \mathbf{X})_{v,j} = \sum_{u \in V} (\hat{A}^K)_{v,u} X_{u,j}$ and contracting with Θ, the unnormalised score at v for class c decomposes as a linear combination over node-feature pairs:

$$(\hat{A}^K \mathbf{X} \Theta)_{v,c} \;=\; \sum_{(u,j) \in \mathcal{N}_K(v) \times \{1,\dots,d\}} \alpha_{u,j}^{v,c} \, X_{u,j}, \qquad \alpha_{u,j}^{v,c} := (\hat{A}^K)_{v,u} \, \Theta_{j,c}. \tag{4}$$

The coefficient $\alpha_{u,j}^{v,c}$ is the fixed weight with which the Boolean input $X_{u,j}$ enters the unnormalised score of class c at v. It factors into a structural term $(\hat{A}^K)_{v,u}$, recording how strongly u reaches v, and a semantic term $\Theta_{j,c}$, recording how strongly feature j votes for class c. The coefficients depend only on $\hat{A}$ and Θ, not on the input features, so they can be precomputed once per node and reused for every explanation query each v. Softmax is strictly monotone, so the predicted class is determined by the argmax of the unnormalised scores. The prediction $c^*(v) = c$ therefore holds iff, for every competing class $c' \neq c$, the linear score induced by column $\Theta_{:,c}$ exceeds the linear score induced by column $\Theta_{:,c'}$ evaluated on the same feature values $X_{u,j}$ over the receptive field. The classification rule at v is thus a system of $C - 1$ linear inequalities in the Boolean inputs $\{X_{u,j}\}_{(u,j) \in \mathcal{N}_K(v) \times \{1,\dots,d\}}$:

$$\sum_{(u,j) \in \mathcal{N}_K(v) \times \{1,\dots,d\}} (\alpha_{u,j}^{v,c} - \alpha_{u,j}^{v,c'}) \, X_{u,j} \;>\; 0, \qquad \forall \, c' \neq c, \tag{5}$$

with margin coefficients $\alpha^{v,c}_{u,j} - \alpha^{v,c'}_{u,j}$ fixed by the model. We take $c^*(v)$ throughout the paper as the prediction to be explained.

3.2. NF-AXps for SGC

Adapting Definition 1 for SGC, the components correspond as follows: ϕ is the SGC argmax classifier of Eq. (3), $F = \mathcal{N}_K(v) \times \{1, \ldots, d\}$ is the set of node-feature pairs inside the receptive field of v, $\mathcal{X} = \{0,1\}^F$ is the Boolean cube indexed by these node-feature pairs, and $\mathcal{Y} = \{1, \ldots, C\}$ is the set of class labels. The observed input x is the current Boolean configuration on F, and $y = c^*(v)$ is the predicted class. A subset $H \subseteq F$ is an AXp at node-feature level for $c^*(v) = c$ iff fixing $X_{u,j} = x_{u,j}$ on H preserves, for every $c' \neq c$, the column-c-vs-column-c' inequality derived from (4), for all $\{0,1\}$-assignments of the remaining entries in $F \setminus H$. AXp sufficiency is thus a single linear inequality universally quantified over Boolean completions of $F \setminus H$, with coefficients $\alpha^{v,c}_{u,j}$ fixed by the model.

The natural unit of explanation is therefore a node-feature pair, rather than a whole node or a whole feature. As mentioned in Section 2.1, most datasets considered in our experiments represent node attributes as binary bag-of-words vectors, where $X_{u,j} = 1$ indicates the presence of term j at node u. We therefore adopt a presence-based explanation semantics. Following robust explanation approaches for text classification, explanations are defined in terms of subsets of the words present in the input [La Malfa et al. 2021]. Accordingly, we consider as explanatory candidates only node-feature pairs (u, j) for which the corresponding term is present, i.e., $X_{u,j} = 1$. The features that can possibly matter for the prediction at v are therefore exactly those whose pair (u, j) is both reachable and with $X_{u,j} = 1$. We collect these into the candidate set

$$\mathcal{C}_v \;=\; \big\{(u,j) \,:\, u \in \mathcal{N}_K(v),\; X_{u,j} = 1\big\}, \tag{6}$$

which serves as the search space for the NF-AXp of v. We assume: only pairs that are currently switched on $X_{u,j} = 1$ can be "removed" by setting the corresponding entry to zero, and pairs that are switched off $X_{u,j} = 0$ already contribute zero to the score under that baseline.

Let $c^* = c^*(v)$ denote the SGC prediction at v. We want a subset $\mathcal{S}_v \subseteq \mathcal{C}_v$ such that no Boolean completion of the remaining entries of $\mathbf{X}$ on the receptive field, restricted to $\{0,1\}$ on the indices outside $\mathcal{S}_v$ and inside $\mathcal{C}_v$, can flip the predicted class. Because the SGC score is linear in $\mathbf{X}$, the worst-case completion is computable explicitly. For every rival class $c' \neq c^*$, define the contribution of a single pair to the margin of c^* over c',

$$\delta_{u,j,c'} \;=\; (\hat{A}^K)_{v,u}\,\big(\Theta_{j,c^*} - \Theta_{j,c'}\big),$$

and the worst-case margin attainable under a fixed $\mathcal{S} \subseteq \mathcal{C}_v$ as

$$\mathrm{margin}(c', \mathcal{S}) \;=\; \sum_{(u,j)\in\mathcal{S}} \delta_{u,j,c'} \;+\; \sum_{(u,j)\in\mathcal{C}_v\setminus\mathcal{S}} \min\{0,\, \delta_{u,j,c'}\}. \tag{7}$$

The two sums admit a direct worst-case reading. Pairs in $\mathcal{S}$ are held at their observed value $X_{u,j} = 1$ by the explanation and therefore contribute exactly $\delta_{u,j,c'}$. Pairs outside $\mathcal{S}$ can take either Boolean value; the completion that minimises the margin sets $X_{u,j} = 0$ when

$\delta_{u,j,c'} > 0$, removing a contribution that would otherwise favour c^*, and keeps $X_{u,j} = 1$ when $\delta_{u,j,c'} < 0$, adding a contribution that reduces c^*. The minimum operator records exactly this worst-case assignment. Specialising Definition 1 to the SGC then yields an explicit sufficiency condition.

We call the AXp for SGC a node-feature AXp (NF-AXp), since it ranges over node-feature pairs $(u, j) \in \mathcal{C}_v$ rather than single feature indices.

Definition 2 (NF-AXp for SGC at v)**.** *A set $\mathcal{S}_v \subseteq \mathcal{C}_v$ is a node-feature abductive explanation for $c^*(v)$ if $\mathrm{margin}(c', \mathcal{S}_v) \geq 0$ for every $c' \neq c^*$. Moreover, no proper subset of $\mathcal{S}_v$ satisfies the same condition.*

The worst-case margin in Definition 2 accounts only for Boolean completions of $\mathcal{C}_v \setminus \mathcal{S}_v$: pairs (u, j) with $X_{u,j} = 0$ are held at zero throughout, in line with the presence-based semantics adopted for bag-of-words features [La Malfa et al. 2021] and with the family of *restricted* AXps [Izza et al. 2024], which restrict the quantification of Definition 1 to a neighbourhood of the observed input. Allowing $X_{u,j}$ to change from 0 to 1 would correspond to inserting a term that is absent from node u, whereas the adopted perturbation model only allows terms present in the input to be removed. This is analogous to robust explanations for text classifiers [La Malfa et al. 2021], which explain a prediction in terms of subsets of the words present in the input. Definition 2 thus adapts Definition 1 to the SGC setting under a restricted perturbation model, rather than departing from it.

A NF-AXp $\mathcal{S}_v$ is the analogue, for SGC node classification, of the AXp of Definition 1: an irredundant collection of present node-feature pairs whose value alone, regardless of how every other entry of $\mathbf{X}$ inside $\mathcal{C}_v$ is set within $\{0, 1\}$, already pins the prediction to c^*. Equivalently, $\mathcal{S}_v$ is a robustness certificate against Boolean erasures on its complement. The full candidate set $\mathcal{C}_v$ is trivially sufficient: substituting $\mathcal{S}_v = \mathcal{C}_v$ into Eq. (7) reduces the second sum to zero and recovers the actual margin computed by the model on the observed input, which is non-negative by the definition of c^*. So $\mathcal{C}_v$ is sufficient but not subset-minimal, and the question is how to compress it.

We obtain a NF-AXp $\mathcal{S}_v$ by greedy deletion [Ignatiev et al. 2019]. The procedure starts from $\mathcal{S}_v \leftarrow \mathcal{C}_v$ and processes the elements of $\mathcal{C}_v$ in a fixed deterministic order. For each pair (u, j) in turn, it tentatively removes the pair from $\mathcal{S}_v$ and recomputes the worst-case margin in Eq. (7); if $\mathrm{margin}(c', \mathcal{S}_v \setminus \{(u, j)\}) \geq 0$ holds for every rival class, the pair is dropped permanently, otherwise it is restored. Because the dependence of margin on $\mathcal{S}$ is monotone, in the sense that removing a pair can only weaken the margin, a pair that fails the test once cannot become removable later, so a single linear pass suffices. The result is a NF-AXp at a cost of $|\mathcal{C}_v|$ margin checks per node, each of which is itself a sum over $|\mathcal{C}_v|$ contributions and so runs in time linear in the candidate set. The total cost of an explanation at v is therefore $\mathcal{O}\big((C-1)\,|\mathcal{C}_v|^2\big)$ arithmetic operations.

By construction, the returned $\mathcal{S}_v$ inherits the two formal properties of Definition 2. Sufficiency holds because the last accepted state of $\mathcal{S}_v$ passes the margin test against every rival; subset-minimality holds because every pair in the final $\mathcal{S}_v$ was tested and could not be removed. Both properties are exact, in the strong sense that they hold for the original SGC decision function. These per-node sets are the NF-AXps of AXSGC, and they are the raw material for the global step.

4. Extracting global rules from AXps

We now lift the per-node NF-AXps produced in Section 3 into a small set of global logical rules. We proceed in two steps. Subsection 4.1 aggregates the per-node sets into a single propositional representation by replacing concrete node identities with distance-indexed predicates over features, producing a Boolean matrix that summarises the entire dataset. Subsection 4.2 reads global per-class rules out of this matrix by fitting one shallow decision tree per class against the SGC prediction and converting its root-to-leaf paths into DNF rules whose agreement with the model is measured by fidelity.

4.1. Distance-indexed predicates and the explanation matrix

The NF-AXps $\{\mathcal{S}_v\}_{v\in V}$ are detailed certificates but they are written in terms of concrete neighbour identities. Two nodes with structurally similar receptive fields can produce explanations that share no pair (u, j) at all, simply because their neighbours u are different nodes. To generalise across the graph we abstract over node identity while keeping the hop distance at which a contribution arrives. Concretely, for every feature index $j \in \{1, \ldots, d\}$ and every hop distance $k \in \{0, 1, \ldots, K\}$ we define a Boolean predicate over nodes:

$$\mathrm{P}_{k,j}(v) \;=\; 1 \iff \exists\, u \in V : \mathrm{dist}_G(v, u) = k \text{ and } (u, j) \in \mathcal{S}_v \tag{8}$$

where dist_G is the unweighted shortest-path distance in G. In words, $\mathrm{P}_{k,j}(v)$ asserts that "feature j appears in the NF-AXp of v through at least one node located exactly k hops away from v". The hop truncation $k \leq K$ is consistent with the SGC receptive field: pairs at greater distance have $(\hat{A}^K)_{v,u} = 0$ and therefore cannot appear in any $\mathcal{S}_v$.

Two design choices in Eq. (8) deserve explicit comment. First, the predicate aggregates over neighbours by existential quantification, which is what allows nodes with isomorphic-but-not-identical neighbourhoods to share predicates. Second, the predicate is gated by membership in $\mathcal{S}_v$, it fires only when the active occurrence was actually used by the abductive explanation. This second condition is what differentiates the construction from a generic "bag of features per hop" encoding, since the predicate selects exactly the contributions that survived the procedure in Subsection 3.2.

Collecting the predicates yields a global representation of the dataset. We index columns by pairs (k, j) in lexicographic order and define the explanation matrix

$$M \in \{0, 1\}^{n\times(K+1)\,d}, \qquad M_{v,\,(k,j)} \;=\; \mathrm{P}_{k,j}(v). \tag{9}$$

Each row of M summarises the NF-AXp of a single node in the predicate vocabulary; columns correspond to elementary structural-semantic events of the form "feature j in a node at hop k". The width $(K + 1)\, d$ is bounded a priori by the model and is independent of how the receptive fields of different nodes overlap. The matrix is sparse in practice, since most $\mathcal{S}_v$ are small relative to d, and it is the only object the global stage of AXSGC needs to consult.

4.2. Global rule extraction

The matrix M from Subsection 4.1 reduces global explanation to a tabular learning problem on Boolean inputs, where each input is a node and the target is the SGC prediction at that node. We train a single shallow decision tree T on M with the multi-class target

$c^*(v) \in \{1, \ldots, C\}$. Crucially, the supervision signal is the SGC prediction and not the dataset label: the tree models the behaviour of the SGC classifier on the predicate vocabulary, so any disagreement between T and c^* is a deficiency of the rules and not of the model, and the corresponding gap is what we will later quantify as fidelity. Splits are scored by Gini impurity and the depth of T is fixed a priori.

The tree T partitions the node set according to short conjunctions of predicate tests. Every internal node of T branches on a single column of M, so each root-to-leaf path corresponds to a conjunction of literals of the form $\mathrm{P}_{k,j}(v) = 0$ or $\mathrm{P}_{k,j}(v) = 1$, and the leaf at the end of the path carries a single class label $c \in \{1, \ldots, C\}$. Reading the literals along a path that terminates at a leaf labelled with class c yields a rule of the form

$$\bigwedge_{(k,j)\in S^+} \mathrm{P}_{k,j}(v) \;\wedge \bigwedge_{(k,j)\in S^-} \neg\, \mathrm{P}_{k,j}(v) \implies c^*(v) = c, \tag{10}$$

where S^+ and S^- collect the (hop, feature) tests asserted respectively true and false along the path. Grouping the leaves of T by their class label and taking the disjunction within each group yields one DNF formula per class; the collection of these formulae forms the global rule set returned by AXSGC. The depth of T is bounded by a small constant so that every rule remains short enough to inspect; this also acts as an implicit regulariser against memorising rare predicate combinations. Each literal carries a structural reading: a positive literal $\mathrm{P}_{k,j}(v) = 1$ asserts that some node at distance k from v contributed feature j to the explanation, while a negative literal asserts the absence of any such contribution at distance exactly k from v.

The agreement between the resulting rule set and the SGC model is measured directly. Writing $T(M_{v,\cdot})$ for the class label predicted by T when applied to the row of M indexed by v, we define the fidelity of the rule set on a test split $V_{\mathrm{test}} \subseteq V$ as

$$\text{Fidelity} \;=\; \Pr_{v\in V_{\mathrm{test}}}\big[\, T(M_{v,\cdot}) = c^*(v) \,\big], \tag{11}$$

i.e. the fraction of test nodes on which the rule set reproduces the SGC prediction. Fidelity is the central quantity of the global stage because it isolates the contribution of the rule extractor from the contribution of the underlying classifier: the SGC accuracy with respect to the dataset labels is unaffected by the extraction step, while any disagreement between the rules and c^* is attributable to the abstraction introduced by the predicates and to the capacity limits imposed on the trees.

5. Experiments

This section answers the three research questions Q1–Q3 stated in the Introduction. Section 5.1 fixes the protocol. Sections 5.2, 5.3 and 5.4 address the three questions in turn.

5.1. Setup

Four standard node-classification benchmarks are used, spanning the range of regimes AXSGC is intended to cover. BAShapes is a synthetic graph of 700 nodes built from a Barabási–Albert base augmented with house motifs, with 32 one-hot degree features and 4 classes; it serves as a controlled sanity check where ground-truth motifs are known. Cora (2708 nodes, 1433 binary bag-of-words features, 7 classes) and Citeseer (3327 nodes,

Tabela 1. Test accuracy of the SGC classifier and fidelity / rule count of the two global rule-based explainers under identical tree depth 12.

		AXSGC		LogicXGNN	
Dataset	SGC acc. (%)	Fidelity (%)	rules	Fidelity (%)	rules
BAShapes	100.00	100.00	6	100.00	4
Cora	74.40	90.70	77	78.60	318
Citeseer	61.70	83.50	72	76.60	445
PubMed	99.90	99.90	31	69.70	153

3703 binary bag-of-words features, 6 classes) are citation networks with natively binary inputs. PubMed (19717 nodes, 500 TF-IDF features, 3 classes) is an order of magnitude larger and is binarised at zero so that the feature matrix $\mathbf{X} \in \{0,1\}^{n\times d}$ matches the Boolean signature on which the predicates $\mathrm{P}_{k,j}$ operate.

The base classifier is the SGC of [Wu et al. 2019] with propagation depth $K = 2$, trained with learning rate 10^{-2}, weight decay $5 \cdot 10^{-4}$, 200 epochs and seed 42. Once $\hat{A}$ and the learned weights are fixed, NF-AXps $\mathcal{S}_v$ are obtained from the propagated representation $\hat{A}^K\mathbf{X}$ by the procedure of Section 3.2. The predicates $\mathrm{P}_{k,j}$ of Section 4.1 then populate the explanation matrix M, and the global stage of Section 4.2 fits a single multi-class decision tree T on M with target $c^*(v)$, using the Gini criterion and maximum depth 12. Each root-to-leaf path of T yields one rule of the form prescribed by Eq. (10).

The reference baseline is LogicXGNN [Geng et al. 2026], the closest existing rule-based global method for graph neural networks. It is run on the same SGC predictor and the same propagated features, and configured with maximum tree depth 12 so that the two decision trees have identical capacity. The primary metric is the fidelity of the rule set with respect to the SGC, as defined in Eq. (11), namely the fraction of test nodes on which the class-conditioned DNF rules reproduce $c^*(v)$. The size of the global rule set is measured by the total number of extracted rules. Our implementation is publicly available.[1]

5.2. Fidelity and Rule Count (Q1)

In addition to fidelity, we report the number of extracted rules; together they characterise how faithful and how compact the global rule set is. Unlike the per-node guarantees of Subsection 3.2, the global rules carry no explicit sufficiency certificate. A high fidelity is empirical evidence that the abductive vocabulary captures the SGC decision boundary.

A rule-based explainer is only useful insofar as it faithfully tracks the model it claims to describe, and that faithfulness should be obtained with a rule set small enough to inspect. Table 1 reports, for each dataset, the SGC test accuracy together with the fidelity and rule count of the two global explainers. The SGC accuracy is included as a reference point so that fidelity is read against the predictive quality of the model being explained, not against ground-truth labels.

On BAShapes both methods reach perfect fidelity, with rule counts of the same order as the number of latent motif classes: both methods simply rediscover those motifs, and AXSGC does so with six rules against four. On Cora, AXSGC reaches 90.7% fidelity with 77 rules, against 78.6% with 318 rules for LogicXGNN. Citeseer shows the same

[1] `https://github.com/bryanlimac/AXSGC`

Tabela 2. Distribution of the NF-AXp size $|\mathcal{S}_v|$ across all nodes of each dataset.

Dataset	n	mean $\lvert\mathcal{S}_v\rvert$	median $\lvert\mathcal{S}_v\rvert$	max $\lvert\mathcal{S}_v\rvert$
BAShapes	700	4.0	2	31
Cora	2708	67.4	32	2426
Citeseer	3327	61.7	35	1238
PubMed	19717	629.2	292	45288

Tabela 3. Mean rule length (number of conditions per rule) of the two global rule-based explainers under identical tree depth 12.

Dataset	AXSGC	LogicXGNN
BAShapes	2.83 ± 0.90	1.75 ± 0.43
Cora	8.88 ± 2.57	9.41 ± 2.20
Citeseer	8.82 ± 2.74	9.64 ± 2.18
PubMed	8.06 ± 3.29	7.62 ± 0.71

qualitative pattern at lower absolute fidelity for both methods, with AXSGC reaching 83.5% fidelity against 76.6% for LogicXGNN, a gain of 6.9 percentage points while using 83.8% fewer rules (72 vs 445). On PubMed the gap is the largest of the four benchmarks: 99.9% fidelity with 31 rules against 69.7% with 153. A plausible reading is that the NF-AXp step produces candidate node-feature pairs tailored to each node before aggregation. A single rule can then subsume many NF-AXps once they share a common minimal core. Decision-tree path enumeration on the raw propagated features, in contrast, does not exploit this redundancy.

5.3. NF-AXp and Rule Sizes (Q2)

The global rule set is the externally visible artefact, but it is built by aggregating the NF-AXps $\mathcal{S}_v$ produced by greedy deletion. Their size controls both the cost of AXSGC and the readability of the intermediate evidence that an inspector might want to query for individual nodes. Table 2 reports the mean, median and maximum of $|\mathcal{S}_v|$ over the whole node set of each dataset.

The mean NF-AXp size scales jointly with the receptive-field density induced by $\hat{A}^K$ and with the feature dimension: BAShapes' sparse one-hot encoding yields median size 2; Cora and Citeseer lie in the same bag-of-words regime; PubMed's denser TF-IDF signature produces the largest NF-AXps. In all three non-synthetic datasets the median of $|\mathcal{S}_v|$ is markedly smaller than the mean, and both are orders of magnitude smaller than the maximum. This long-tailed shape is precisely the regime in which the aggregation through M is most beneficial: many short, overlapping NF-AXps collapse into a few global rules, while the rare long ones are absorbed by the same aggregation step.

Table 3 reports the mean rule length of both methods. LogicXGNN produces slightly shorter rules on BAShapes, while on Cora, Citeseer and PubMed the two methods yield rules of comparable length, with all differences within one standard deviation of each other.

5.4. A Qualitative Example and Coverage (Q3)

Fidelity numbers and average sizes do not by themselves indicate whether individual rules carry semantic content. As a concrete illustration, consider the highest-coverage rule

Tabela 4. Top-1 and top-10 coverage of AXSGC and LogicXGNN rules per dataset.

		AXSGC		LogicXGNN	
Dataset	n	top-1	top-10	top-1	top-10
BAShapes	700	280	700 (100.0%)	300	700 (100.0%)
Cora	2708	374	2162 (79.8%)	232	936 (34.6%)
Citeseer	3327	740	2846 (85.5%)	125	679 (20.4%)
PubMed	19717	6783	19477 (98.8%)	2045	10528 (53.4%)

extracted on BAShapes, denoted R_0 and assigned to class 0 (the Barabási–Albert base nodes). It fires on 280 of the 700 nodes (40% of the dataset).

$$R_0: \quad P_{0,31}(v) = 1 \ \wedge \ P_{1,2}(v) = 0 \ \Rightarrow \ c^*(v) = 0.$$

On BAShapes feature 31 is the highest-degree bin of the one-hot degree encoding and feature 2 is the degree-two bin, so the rule asserts that a node belongs to the BA base whenever its own zero-hop representation activates the high-degree bin and no node at one hop activates the degree-two bin. The condition is structural rather than feature-driven and matches the way the synthetic graph is generated, which is the kind of correspondence one would hope to recover from a transparent model.

Coverage is also informative on the citation networks, where exhaustive inspection of the rule list is not realistic. Table 4 reports the top-1 and top-10 coverage for both methods. On BAShapes both reach 100% with their ten top rules. On the larger benchmarks, AXSGC rule sets are markedly more concentrated: the top 10 rules cover 79.8% of Cora, 85.5% of Citeseer and 98.8% of PubMed, against 34.6%, 20.4% and 53.4% for LogicXGNN — gaps of $2.3\times$, $4.2\times$ and $1.85\times$ respectively. AXSGC therefore concentrates coverage in a small subset of high-coverage rules, whereas LogicXGNN spreads it more thinly across its larger pool, consistent with its higher rule counts and lower fidelity (Table 1).

6. Conclusion

We presented AXSGC, a logic-based framework for explaining SGC node predictions. For each node, AXSGC derives an AXp with an explicit sufficiency margin from the closed form of Simple Graph Convolution. These per-node AXps are then aggregated into a global per-class ruleset via a decision tree over distance-indexed predicates. Across four benchmarks, our method improves fidelity by up to 30.2% over LogicXGNN while extracting up to 83.8% fewer rules. AXSGC delivers per-node certified sufficiency and empirically faithful, model-aligned global rules. We plan to extend AXSGC in two directions. First, we will apply it to other GNN architectures by composing it with a distillation step to SGC [Pereira et al. 2023]. Second, we will extend our method and the NF-AXps directly to other GNN architectures.

Referências

Armgaan, B., Dalmia, M., Medya, S., and Ranu, S. (2024). GraphTrail: Translating GNN predictions into human-interpretable logical rules. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*.

Azzolin, S., Longa, A., Barbiero, P., Lio, P., and Passerini, A. (2023). Global explainability of GNNs via logic combination of learned concepts. In *11th ICLR*.

Geng, C., Zhao, Z., Wang, Z., Ye, H., Jiang, Y., and Si, X. (2026). LogicXGNN: Grounded logical rules for explaining graph neural networks. In *14th ICLR*.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 22118–22133.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). Abduction-Based Explanations for Machine Learning Models. In *AAAI*, volume 33, pages 1511–1519.

Izza, Y., Huang, X., Morgado, A., Planes, J., Ignatiev, A., and Marques-Silva, J. (2024). Distance-restricted explanations: Theoretical underpinnings & efficient implementation. *arXiv preprint arXiv:2405.08297*.

Ju, W., Fang, Z., Gu, Y., Liu, Z., Long, Q., Qiao, Z., Qin, Y., Shen, J., Sun, F., Xiao, Z., Yang, J., Yuan, J., Zhao, Y., Wang, Y., Luo, X., and Zhang, M. (2024). A comprehensive survey on deep graph representation learning. *Neural Networks*, 173:106207.

La Malfa, E., Michelmore, R., Zbrzezny, A. M., Paoletti, N., and Kwiatkowska, M. (2021). On guaranteed optimal robust explanations for NLP models. In *IJCAI-21*.

Marques-Silva, J., Gerspacher, T., Cooper, M. C., Ignatiev, A., and Narodytska, N. (2020). Explaining naive Bayes and other linear classifiers with polynomial time and delay. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Marques-Silva, J. and Ignatiev, A. (2022). Delivering trustworthy AI through formal XAI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12342–12350. Senior Member Presentation Track.

Pereira, T., Nascimento, E., Resck, L. E., Mesquita, D., and Souza, A. (2023). Distill n' explain: Explaining graph neural networks using simple surrogates. In *26th AISTATS*.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Q. (2019). Simplifying graph convolutional networks. In *36th ICML*, pages 6861–6871.

Yuan, H., Yu, H., Gui, S., and Ji, S. (2023). Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5782–5799.