Neurosymbolic Methods for Explainable Graph Neural Networks: A Survey

Kislay Raj[1,2,*], Alessandra Mileo[2]

[1]Research Ireland Centre's for Research Training in Artificial Intelligence, Dublin City University, Dublin, Ireland
[2]Insight SFI Research Centre for Data Analytics, School of Computing, Dublin City University, Dublin, Ireland

Abstract

This survey examines the role of neurosymbolic AI (NeSy) in enhancing the explainability of graph neural networks (GNNs). By combining neural and symbolic approaches, NeSy methods aim to mitigate the black-box nature of GNNs and provide transparent and interpretable decision making. We categorise explainability techniques, including rule learning, subgraph based methods, and knowledge graph integration, and evaluate their applications in domains such as biomedicine and fraud detection. The survey also compares instance level and model level explanation methods, highlighting their respective strengths and limitations. Finally, we discuss open challenges and future directions for advancing NeSy in GNN explainability.

Keywords

Neurosymbolic AI, Graph Neural Network, RuleLearning, Explainable AI

1. Introduction

NeSy is a promising approach that combines neural networks, which excel in learning complex patterns from data, with symbolic reasoning, which provides interpretability and representation of structured knowledge [1, 2]. This combination addresses significant issues that purely neural systems face, such as the lack of interpretability and the scalability challenges encountered by traditional Rule-based AI [3]. Two important research areas within NeSy are rule learning, which involves extracting logical rules from data or trained neural networks, and explainability in GNNs, which aims to understand the predictions made by graph based deep learning systems [4, 5].

The early pre-2020 GNN used gradient-based methods [1, 3], which were limited by relational networks and symbolic methods, and were computationally intensive. 2020-2025, neurosymbolic approaches emerged, integrating neural and symbolic paradigms. GNNExplainer [6], a key method in the realm of explainable AI (XAI), introduced subgraph-based explanations for GNNs. In addition to this, INSIDE-GNN [7] focused on mining activation rules, while Logic-Guided GNNs [8] integrated knowledge graph (KG) rules. These approaches mark a significant evolution in Rule-based explainability techniques for GNNs. They combine subgraph extraction with symbolic rule induction to improve model transparency. Rule learning techniques, such as differentiable inductive logic programming (ILP) [9] and neural-symbolic knowledge distillation [1], allow AI systems to generate logical rules while leveraging gradient based learning methods. These approaches are critical in domains where transparency is essential, such as healthcare and legal decision-making. GNNs have become very effective for working with relational and graph-based data, but their complexity makes it challenging to explain their decisions clearly. Explainability methods, such as GNNExplainer [6] and PGExplainer [10], identify significant subgraphs and node features to help users understand GNN decisions. However, most recent approaches rely heavily on soft masks, making explanations potentially less stable and less reliable. In addition, their explanations typically focus on individual predictions rather than on proving the overall predictions of the model. NeSy has emerged as a promising approach, integrating neural networks with symbolic reasoning to combine the power of data-driven models with human understandable logic.

Recent work has sought to overcome these limitations. Methods that extract symbolic rules from hidden layers of GNNs offer more precise and more reliable explanations [11, 12, 7]. Approaches such as semantic graphs for layer-wise analysis and logic-guided enhance GNN interpretability by integrating symbolic rules into neural predictions. These hybrid methods are central to our survey and highlight the role of Rule-based reasoning in GNN explainability. However, key challenges remain, namely improving scalability, ensuring fidelity of explanations, and developing standardised evaluation metrics.

This survey focusses on recent advances in GNN explainability, human readable rule learning, and exploring their integration within NeSy. Instead of broadly categorising hybrid architectures, we focus on methods that effectively integrate neural and symbolic techniques within GNNs to enhance their transparency, trustworthiness, and interpretability. This paper aims to bridge the gap between symbolic reasoning and GNN explainability, with a focus on human-interpretable rule extraction and integration of symbolic knowledge.

Although existing surveys have made valuable contributions to the explainability of GNN [5, 13] and NeSy [14, 15] in isolation, they exhibit three critical gaps, which this work uniquely bridges in both domains, systematically analysing how rule learning enhances GNN transparency while maintaining predictive performance. Previous surveys treat GNN explainability and symbolic rule learning as separate domains [5] focus purely on GNN methods, [14] on symbolic reasoning. To our knowledge, this is the first survey to systematically examine the integration of symbolic rule learning with GNN explainability within a unified NeSy framework. Unlike previous work, we incorporate recent advancements such as differentiable rule mining [9] and temporal graph explainers [16], which are key to improving the explainability and scalability of GNN models. While [17] surveys graph explainability broadly and [18] examines healthcare applications, we uniquely bridge technical mechanisms, integration frameworks, and domain applica-

ANSyA 2025: 1st International Workshop on Advanced Neuro-Symbolic Applications, co-located with ECAI 2025.
*Corresponding author.
kislay.raj2@mail.dcu.ie (K. Raj);
alessandra.mileo@insight-centre.org (A. Mileo)
0000-0003-0089-6866 (K. Raj); 0000-0002-6614-6462 (A. Mileo)

Table 1
Comparison of our survey with existing survey papers

Survey Paper	GNN Explainability	Rule-based- NeSy	Survey Focus	Key Techniques
A Comprehensive Survey on GNNs [5]	✓	✗	Taxonomy of GNN Explainability, Methods	GNNExplainer, PGExplainer, GraphMask
A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability [13]	✓	✗	Trustworthiness in GNNs: Privacy, Robustness, Fairness, and Explainability	Trustworthy GNNs, Privacy-Preserving, Robustness, Fairness
NeSy for Reasoning Over Knowledge Graphs: A Survey [14]	✗	✓	NeSy and Reasoning over Knowledge Graphs	Rule Learning, Embedding Approaches, Logical Constraints
Graph-Based Explainable AI: A Comprehensive Survey [17]	✓	✗	Graph Explainability Methods Beyond GNNs	Graph-Based Learning Models, Knowledge Graphs, GNN Models
Neurosymbolic AI: Explainability, Challenges, and Future Trends [15]	✗	✓	Classification of Explainability in NeSy	Implicit/Explicit Representations, Unified Representations
A Survey of Neurosymbolic Visual Reasoning with Scene Graphs and Common Sense Knowledge [19]	✗	✓	Knowledge-Based Neurosymbolic Approaches for Scene Representation	Scene Graph Generation, Visual Reasoning, Common Sense Knowledge Integration
A Study on Neurosymbolic Artificial Intelligence: Healthcare Perspectives [18]	✗	✓	Neurosymbolic AI in Healthcare Applications	Rule-based Explainability, Knowledge Representation, Machine Learning
Neurosymbolic Methods for Explainable Graph Neural Networks: A Survey	✓	✓	**Rule Learning and GNN Explainability in NeSy**	**Symbolic Reasoning, GNN Explainability**

tions with a consistent emphasis on human-interpretable rule extraction. As in Table 1, this chapter highlights the gap in the discussion of the interplay between the learning of neurosymbolic rules and the explainability of GNN.

Table 3 compares the scope of our survey with previous works in key dimensions. Whereas existing surveys tend to focus on isolated aspects of either GNN explainability or symbolic AI, our work provides a unified perspective that bridges these domains and addresses emerging needs in explainability. Comparison of explainability methods by scope (Instance-Level, Model-Level, Rule-based, Concept-Based) narrows down the focus to surveys that discuss key explainability methods in GNNs and NeSy, with particular attention to evaluation protocols and method categorisation. This survey employs a systematic literature review (SLR) methodology with the objective of comprehensively identifying, selecting, and synthesising all relevant research related to the application of NeSy techniques for explaining GNNs. The process encompassing identification, screening, eligibility, and inclusion is detailed in the following and summarised in Figure 1. This analysis highlights three critical gaps that our survey addresses. As discussed in Section 2, current methods in instance-level and model-level GNN explainability are reviewed. Specifically: (1) the integration of low-level GNN explanations with high-level rule learning covered in Sections 3–4, the exploration of temporal and dynamic graph scenarios, which are essential for capturing evolving relationships in graphs and are discussed in Section 3 and Section 5 the provision of practical guidance, validated through domain case studies in Section 6. Temporal graph scenarios benefit significantly from symbolic rule learning, a feature often overlooked by traditional GNN explainability methods. This survey identifies key gaps across these three dimensions and proposes directions for future research to address them.

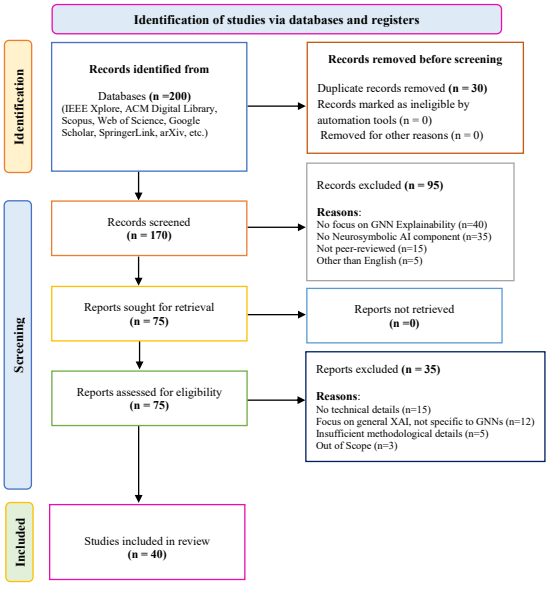

Figure 1: PRISMA Flow Diagram illustrating the study selection process.

2. GNN Explainability Methods

Understanding and interpreting GNN decisions has become a critical challenge, as these models achieve state-of-the-art results on tasks based on non-Euclidean data, such as node classification and graph classification. Unlike images

or text where gradient based visualisation heuristics are widely used, the discrete, relational nature of graphs means that applying such approaches can disrupt key structural properties and produce misleading explanations [5]. To address these shortcomings, researchers have developed four complementary families of explainability methods; instance-level, model-level, Rule-based (intrinsic or post-hoc), and concept-based. These methods collectively aim to identify influential components (nodes, edges, or features), extract significant substructures, and present them in a human readable form [20]. NeSy combines GNNs with symbolic reasoning to enhance model explainability; By integrating GNN explainability methods with symbolic rule learning, NeSy provides logical, human readable explanations while aiming to preserve predictive performance. Graph relations and subgraph motifs align naturally with symbolic rule preconditions, motivating a NeSy focus for GNN explainability.
Figure 2 illustrates the taxonomy of GNN explainability, which categorises methods into two main types: factual and counterfactual explanations. Factual explanations identify key features that significantly influence model predictions, using techniques such as gradient based methods and subgraph extraction. In contrast, counterfactual explanations focus on determining the minimal changes to the input graph that would alter the prediction of the model, helping to pinpoint characteristics whose modification can lead to a different outcome. This taxonomy addresses the challenges of improving GNN transparency through Rule-based reasoning in NeSy. Unlike broader explainability taxonomies, it specifically explores the intersection of NeSy and GNN XAI, offering a framework that combines symbolic reasoning to enhance interpretability and identify research gaps. While methods like GNNExplainer focus on instance-level explanations, Logic-Guided GNNs take a distinct approach by incorporating external knowledge, such as knowledge graphs. In addition, we explore hybrid methods that combine elements from different approaches to further enhance explainability.

2.1. Instance-Level Explanation Methods

Table 2 categorises the representative GNN explainability methods according to their scope: instance (local), hybrid (both), and model (global). It also highlights how each method supports key aspects such as classification saliency, knowledge extraction, and graph generation, providing a clearer understanding of their respective contributions to GNN transparency. Post hoc instance-level methods explain individual predictions without altering the trained GNN's parameters. Gradient-based techniques, such as Guided Backpropagation and CAM/ Grad-CAM, assign saliency scores to nodes, edges, or features, but often produce noisy and unstable explanations [21, 22]. Perturbation-based methods, notably GNNExplainer [6] and GraphMask [23], optimise discrete masks on graph elements to maximise fidelity and sparsity. The surrogate model approaches of PGExplainer [10] and GraphLime [24] fit interpretable models to local graph neighbourhoods. These instance-level methods provide detailed insights crucial for personalised applications such as medical diagnosis and fraud detection, although they can be computationally demanding and may lack generalisation across diverse inputs.

2.2. Model-Level Explanation Methods

Model-level explanations provide information on the overall decision logic of GNNs across an entire dataset, offering a high-level view of the model's behaviour. Techniques like XGNN [25] synthesise prototype graphs that help reveal class-specific structural motifs and trigger high-confidence predictions, contributing to a clearer understanding of the output of GNN. Another significant method, activation rule mining, extracts symbolic rules from hidden layer activations, generating global logic style summaries of the model decision-making process [26]. These symbolic rules play a crucial role in making the model's behaviour more interpretable, providing an additional layer of explanation compared to traditional GNN explainers. However, while these global insights support the verification against domain knowledge and enhance trust, they often rely on a small, representative set of graphs, which may miss rare but crucial decision pathways [27]. By focussing on symbolic rule extraction, our survey highlights how incorporating Rule-based reasoning into model-level explainability can capture more nuanced decision patterns and improve the comprehensiveness of GNN explanations. Figure 2 introduces a distinction between factual and counterfactual explanations, which are critical for understanding how different methods explain the predictions of GNN. While the categories in the figure focus on the nature of explanations, factual vs. counterfactual, the rest of the paper categorises methods by their scope, such as instance-level, model-level, and Rule-based methods. Hybrid methods combine intrinsic explainability with symbolic reasoning, such as rule extraction, to provide both transparent decision-making and logical reasoning.

Table 2 provides an overview of various GNN explainability methods, highlighting their approach, focus, and whether they incorporate Rule-based reasoning. These methods are evaluated on the basis of whether they incorporate rule-based reasoning, which is a crucial aspect of enhancing the transparency and interpretability of models. Rule-based reasoning helps in extracting human readable explanations, enabling users to better understand the model's decision-making process and ensuring that predictions align with domain specific knowledge. Some methods, particularly those related to concept-based explanations, also include elements of graph generation to provide global-level insights into the model's behaviour. Graph generation techniques are often employed in concept-based methods to provide a global-level understanding of GNN behaviour. Knowledge extraction refers to methods that aim to extract explicit knowledge from trained GNN models. These methods are closely related to rule-based and concept-based techniques, as they extract human readable explanations that can be validated by domain experts.

2.3. Neurosymbolic Explanation Methods for GNN XAI

Neurosymbolic approaches to GNN explainability introduce symbolic structure into the explanation pipeline and, where possible, extract human readable artefacts that experts can inspect. Together, these methods form a single area within NeSy for GNN XAI. Rule-based methods analyse trained models to derive symbolic if–then rules that summarise decision logic. Examples include RelEx [33], which extracts relational clauses from model behaviour, GLGExplainer [35], which induces global logical formulas over learnt concepts,

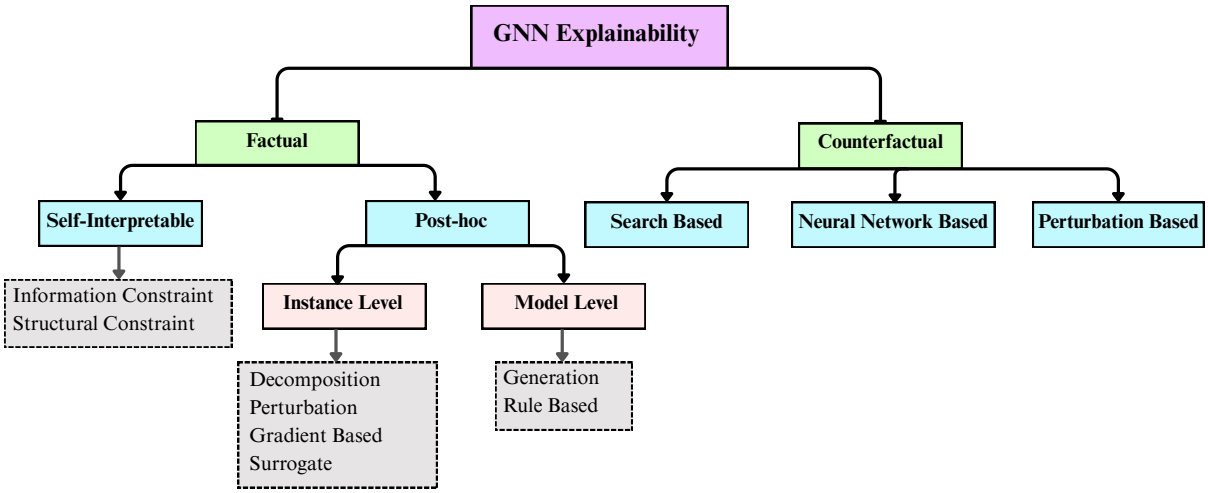

Figure 2: Overview of the taxonomy of GNN explainability methods [10]

Table 2
Comparison of GNN explainability methods by scope and supported functionalities. *Type (scope): Instance*-level (local, per-node/sample), *Model*-level (global), or *Hybrid* (supports both). *Classification:* the method natively explains predictive tasks (e.g., node/graph classification) by attributing importance to features/subgraphs (✓= supported; ✗= not applicable/not primary). *Knowledge extraction:* explicit symbolic artefacts (e.g., rules, Boolean formulas, constraints). *Graph generation:* synthesises exemplar/prototype graphs or subgraphs (beyond masking or extraction). Abbreviations (e.g., SA, CAM, LRP) follow the original papers; ticks indicate capability, not comparative quality.

Method	Type	Classification	Knowledge Extraction	Graph Generation
SA [22]	Instance	✓	✗	✗
Guided BP [21]	Instance	✓	✗	✗
CAM [22]	Instance	✓	✗	✗
Grad-CAM [22]	Instance	✓	✗	✗
GNNExplainer [6]	Instance	✓	✗	✗
PGExplainer [10]	Instance	✓	✗	✗
GraphMask [23]	Instance	✓	✗	✗
ZORRO [28]	Instance	✓	✗	✗
Causal Scr. [29]	Instance	✓	✗	✗
SubgraphX [30]	Instance	✓	✗	✗
LRP [31]	Instance	✓	✗	✗
Excitation BP [22]	Instance	✓	✗	✗
GNN-LRP [32]	Instance	✓	✗	✗
GraphLime [24]	Instance	✓	✗	✗
RelEx [33]	Instance	✓	✓	✗
PGM-Explainer [34]	Instance	✓	✗	✗
INSIDE-GNN [7]	Hybrid	✓	✗	✗
FSAM [26]	Model	✓	✗	✓
Logic-Guided [8]	Hybrid	✓	✓	✗
GLGExplainer [35]	Model	✓	✓	✗
GraphTrail [36]	Model	✓	✓	✗
XGNN [25]	Model	✓	✗	✓

and GraphTrail [36], which derives model level rule traces for global understanding. Rule quality is commonly assessed by fidelity, that is, agreement with the underlying GNN, and by interpretability, captured through expert judgement or rule complexity [37]. Although such methods often provide clearer insight than purely neural attributions, they face scalability challenges and frequently require pruning or visualisation to manage complexity [38]. Concept-based methods link predictions to human defined or automatically discovered concepts rather than raw features. Graph concept bottleneck models [20] predict intermediate concepts, such as functional groups in molecules, before the final classification, allowing concept-level debugging. Graph CAV [27] adapts concept activation vectors to graphs, identifying subgraph patterns whose presence or absence most influences decisions. Grounding explanations in domain concepts facilitates expert validation and can support downstream rule learning. Other NeSy variants include knowledge graph integration and logical regularizers that inject constraints during training; activation-level mining and distillation, where frequent activation patterns are translated into compact clauses; and diagnostic mapping of internal representations, such as FSAM [26], which maps semantic structure across layers and can inform subsequent symbolic extraction. Intrinsic masking frameworks, such as INSIDE-GNN [7], also improve traceability without necessarily pro-

Table 3
Comparative analysis of survey coverage on key neurosymbolic explainability aspects. ✓= **Full coverage**, **Partial = Limited coverage**, ✗= **Not addressed**

Survey Focus	**[5]**	**[14]**	**[13]**	**[15]**	**Our Work**
GNN Explainability Methods	✓	✗	✗	✗	✓
Symbolic Rule Learning	✗	✓	✗	✓	✓
Neurosymbolic Integration	✗	Partial	✗	✓	✓
human readable Rules	✗	✗	✗	Partial	✓
Temporal Graphs	✗	✗	✗	✗	✓
Application Case Studies	✗	Partial	✗	Partial	✓

ducing explicit rules. These variants improve transparency without relying solely on attribution of local characteristics.

3. Evaluation Metrics for NeSy Methods in GNN Explainability

To compare and benchmark the diverse families of GNN explainability and rule learning methods, a unified set of evaluation metrics is essential. These metrics are specifically tailored to evaluate the effectiveness and interpretability of NeSy methods in the context of GNN explainability. We summarise below the core evaluation criteria used across the literature, noting that how each metric specifically suits the evaluation of explainability in GNN models.

Fidelity: Measures the agreement between an explanation or extracted rule set and the original GNN output. In instance-level methods, Fidelity is often used as a measure of classification accuracy of a surrogate model or masked graph relative to the base GNN [6, 10]. In rule extraction, Fidelity quantifies the percentage of GNN predictions that are accurately reproduced by the distilled rules [12].

Sparsity: Quantifies the compactness of an explanation, such as the number of nodes, edges, or features included in an instance-level mask or the total count of generated rules. Sparse explanations are preferred for human comprehension, but must balance the loss of fidelity [6, 23].

Rule Complexity: Evaluates the interpretability of extracted logic rules. It includes metrics such as the number of predicates of average rule length, tree depth, or total rule count. Lower complexity generally implies easier human validation [37].

Concept Completeness and Purity: For concept-based methods, completeness measures how well the discovered concepts cover the model decision space, while purity assesses the semantic coherence of each concept cluster [27]. High Completeness and Purity indicate that concepts accurately and precisely represent the underlying decision factors.

Prototype Faithfulness: In prototype graph generation, this metric measures how representative the graphs generated are of the target class. It is assessed by the confidence drop when real inputs are replaced with prototypes in the GNN inference pipeline [25].

Stability: Reflects the robustness of explanations under small perturbations of the input graph. Stable methods produce consistent explanations for similar inputs, an important property of trustworthy AI [21].

Although these structured metrics provide a foundation for evaluating NeSy methods in GNN explainability, there is a significant gap left. Currently, no single benchmark captures all dimensions of neurosymbolic explainability, such as local and global fidelity, rule complexity, concept coherence, and prototype-graph quality, within a unified framework. The development of such comprehensive evaluation standards is essential for advancing the effective integration of NeSy AI into GNN explainability.

4. Taxonomy of Neurosymbolic Methods

The taxonomy of neurosymbolic methods for GNN explainability classifies approaches by the integration mechanism, the explanatory objective, and the applicability to graph types. Integration mechanisms include rule activation mining, rule extraction, knowledge graph integration, and hybrid reasoning methods such as Logic Tensor Networks [7]. Emerging techniques include privacy preservation and temporal rule induction. Explanatory objectives focus on fidelity, robustness, sparsity, and user trust. Applicability covers homogeneous, heterogeneous, temporal, and attributed graphs [19]. Temporal graphs model relationships that evolve over time and pose distinct challenges for explainability. Examples include social networks and disease progression, where explanations must reflect dynamic change. Traditional GNN explainers often assume static structure and, therefore, overlook temporal evolution. In practice, intrinsic masking and diagnostic tracing, as in INSIDE GNN, improve transparency in domains such as biomedicine and fraud detection. Subgraph saliency methods (for example, GNNExplainer) can serve as a precursor to post hoc rule distillation, while logic-guided GNNs inject knowledge graph constraints during training for applications in recommender systems and biomedicine. Logic Tensor Networks support relational reasoning with differentiable logical constraints, and temporal rule mining has been explored for dynamic graphs in traffic and finance.

Table 4 highlights the mechanisms and neurosymbolic methods of explainability, providing an overview of the landscape and its gaps. This taxonomy focusses exclusively on approaches that use symbolic or semantic information *produce, consume, or constrain* for explanation. Purely neural attributions (for example, gradient or perturbation scores without a semantic interface) are treated as baselines and are not part of the taxonomy. We organise methods by *where and how* symbolic knowledge enters the pipeline, and we keep *scope* (instance level, model level, or hybrid) and *integration stage* (intrinsic or post hoc) as orthogonal tags used elsewhere in the paper. The aim is to guide NeSy researchers towards classes that deliver human readable artefacts (rules, concepts, constraints, or semantically annotated prototypes) or inject structured knowledge during learning. *Rule activation mining* groups methods that anal-

Table 4
Characterisation of neurosymbolic methods for GNN explainability

Method Category	Example	Mechanism	Metrics	Use Cases	Graph Types
Rule Activation Mining	INSIDE-GNN [7]	FORSIED activation mining	Fidelity: High Sparsity: Moderate	Biomedicine, Fraud Detection	Homogeneous, Attributed
Rule Extraction	GNNExplainer[6] + Distillation [39]	Subgraph saliency + clauses	Fidelity: High User Trust: High	Recommender Systems, Social Networks	Homogeneous, Attributed
KG Integration	Logic-Guided GNN [8]	Datalog regularizers	Fidelity: High Robustness: Moderate	Biomedicine, Knowledge Graphs	Heterogeneous
Hybrid Reasoning	Logic Tensor Networks [7]	Differentiable constraints	Fidelity: High User Trust: High	Relational Reasoning	Heterogeneous
Attention-Based Rule Selection	RuleFormer-GNN [40]	Attention-based rule selection	User Trust: High Robustness: Moderate	Social Networks	Homogeneous, Temporal
Privacy-Preserving Extraction	Differential Privacy Rules [41]	Privacy-guaranteed rule mining	Fidelity: Moderate User Trust: High	Healthcare	Attributed
Temporal Rule Induction	Temporal Rule Mining [16]	Rules over evolving graphs	Robustness: Low Sparsity: Moderate	Traffic Networks, Finance	Temporal

yse internal activations to frequent surface patterns that can be aligned with semantic conditions or forwarded to a rule learner. These methods improve traceability and can precede explicit rule induction (for example, INSIDE GNN). *Rule extraction* contains post hoc pipelines that distil trained GNNs into symbolic clauses or rule lists, for example, subgraph importance followed by clause induction, or global rule tracing. *Knowledge graph (KG) integration* covers intrinsic approaches that inject constraints or relations from a KG into training (for example, logic-guided objectives), shaping representations in a knowledge-aware way. *Hybrid reasoning* includes differentiable logical frameworks that couple neural encoders with soft constraints for relational reasoning (for example, Logic Tensor Networks), yielding predictions that are amenable to symbolic inspection. *Attention based rule selection* captures models that learn to select or weight candidate rules to form concise, context aware explanations. *Privacy preserving extraction* comprises methods that mine rules under formal privacy guarantees so that explanations can be shared in sensitive domains. *Temporal rule induction* collects methods that learn or apply rules to evolving graphs, ensuring that explanations remain coherent over time.
Each branch corresponds to a distinct *integration point* for semantics: mining hidden states; extracting rules from behaviour; injecting KG constraints during learning; reasoning with differentiable logic; selecting rules with attention; enforcing privacy during extraction; and handling temporal dynamics with time-aware rules.

5. Neurosymbolic Integration for GNN Explainability

We consider neurosymbolic integration for GNN explainability along the following features: (i) *symbolic rule learning* (intrinsic constraints during training or post hoc extraction); (ii) *scope of explanation* (instance level, model level, or hybrid); (iii) *explanatory artefacts* (masks or subgraphs, human readable rules, prototype graphs or generated graphs, and concepts); (iv) *integration stage* (intrinsic or post hoc); and (v) *temporal and counterfactual support* (the ability to reason over evolving graphs and hypothetical edits). Table 2 situates existing explanations against these features.

Combining symbolic rule learning with GNN explainability fosters neurosymbolic frameworks that retain the representational power of neural models while offering symbolic transparency [11, 20, 26]. Table 2 shows a clear gap: no single method currently provides both instance-level and model- level explanations together with human readable rule extraction and prototype graph generation. Although current hybrid approaches perform well on saliency for classification and on rule derivation, they typically omit graph synthesis. Incorporating graph generation into Rule-based explainers therefore, a next step for NeSy. This would not only produce symbolic rules, but also produce prototype graphs that present learnt knowledge in a visual and structured form. Such a combination would support detailed case-level justifications, by explaining individual predictions, and global symbolic insights, by summarising class-level patterns. In practice, the choice of features should align with the target workflow. For safety critical domains, hybrid explainers that provide transparent case-level reasoning together with validated rules for audit are preferable. For exploratory analysis of class structure, model-level generators such as XGNN can reveal global patterns, which may then be formalised as symbolic constraints and fed back into the training loop. This interplay of saliency, rule induction, and graph synthesis underpins robust and interpretable neurosymbolic systems.

5.1. Neurosymbolic Methods for GNN Explainability

The neurosymbolic methods in GNN explainability focus on how neural and symbolic components can be integrated to enhance transparency and adaptability. It presents key integration mechanisms such as rule activation mining, rule

extraction, and knowledge graph integration. Rule activation mining extracts symbolic rules from GNN activations, which can be done through single-layer or multilayer mining [37]. Rule extraction involves deriving human readable rules from trained GNNs, using subgraph-based or embedding-based methods. Knowledge graph integration enriches the GNN explainability by embedding structured knowledge from KGs, either through KG-guided training or KG-augmented explanations. Hybrid reasoning combines neural and symbolic modules to enable bidirectional interaction, exemplified by methods like neural-symbolic distillation and attention-based rule selection. Emerging methods, such as extraction of privacy-preserving rules and induction of temporal rules, address new challenges in the field [26]. Defines explainability goals such as fidelity, robustness, sparsity, and user trust, which are critical to ensuring that GNNs provide meaningful and reliable explanations. The authors discuss the adaptability of these methods for different graph types, including homogeneous, heterogeneous, temporal, and attributed graphs. INSIDE-GNN [7] for rule activation mining and GNNExplainer [6] for rule extraction, alongside emerging techniques such as differential privacy in rule extraction, it emphasises the importance of scalable KG integration, standardised benchmarks for temporal graphs, and ethical considerations such as bias mitigation.

5.2. Rule-Guided GNNs

Rule-guided GNNs represent a key advance in the integration of symbolic reasoning into GNNs by embedding logical constraints into the learning process, thus improving interpretability, consistency, and accuracy. Logic-guided GNN [8] employs Datalog-style clauses as soft regularizers, aligning predictions with domain rules. Applied to data sets such as WordNet, it achieved a 5% F1 score improvement, particularly in noisy settings, demonstrating the ability of symbolic reasoning to mitigate data inconsistencies. RuleFormer-GNN [5] further advances this approach by using attention mechanisms to dynamically apply context-relevant rules. This led to a 40% reduction in rule violations in benchmark tasks, highlighting its effectiveness in managing complex graph structures. Ongoing developments combine symbolic rule learning with machine learning strategies, such as reinforcement learning, to enhance adaptability while preserving interpretability, paving the way for more robust and transparent rule-guided GNNs. Symbolic rule learning bridges the gap between neural networks and human-understandable logic by extracting interpretable rules from the learnt representations of the network. In GNNs, which handle complex graph-structured data, this approach is particularly valuable, as it provides an explainable layer over the model's decision-making. Recent advances in neural symbolic AI have integrated symbolic rule learning with GNNs, enabling the generation of human readable rules while preserving high predictive accuracy.

5.3. Post-hoc Rule Extraction and Iterative Refinement

Post hoc rule extraction converts GNN trained opaque decisions into explicit, human readable 'if-then' rules, while iterative refinement closes the loop by using those rules to guide further model training. Early pipelines combine the saliency of the subgraph of GNNExplainer [6] with neural-symbolic distillation [1] to produce concise rule sets that justify node or graph predictions in the recommendation and knowledge graph systems. Functional Semantic Activation Mapping (FSAM) [26] advances this by tracking neuron activations across layers and distilling them into natural language narratives or logic clauses, revealing both which embedding dimensions drive each class decision and architectural issues such as oversmoothing. Crucially, this becomes a closed-loop extracted rule serving as soft constraints or diagnostic feedback during subsequent training, improving both model accuracy and interpretability. Counterfactual validation tests rule the necessity, concept bottleneck layers ground extraction in human-meaningful abstractions, and advanced pruning heuristics distil large rule spaces into concise, high-value rule sets. Scaling this iterative process to large, dynamic graphs will require streaming rule mining algorithms and sparse GNN architectures, but it promises fully neurosymbolic systems that learn continuously and explain transparently.

5.4. Hybrid Architectures

Hybrid neurosymbolic architectures combine neural network pattern recognition with symbolic reasoning to enhance the explainability of GNN. These architectures enable bidirectional information flow, making GNN output more interpretable [4]. For example, neural modules learn graph embeddings, which symbolic modules use to generate human readable rules. This synergy addresses the "black box" nature of GNN, providing accurate and logical explanations. Key methods include logic tensor networks [7] and neural theorem provers (NTP) [42]. LTNs integrate logical constraints into neural networks, improving interpretability by enforcing domain knowledge during training, while NTPs use symbolic deduction to reason over graph structures. Both methods excel in relational reasoning tasks, achieving accuracy up to 98%. Other methods like INDIE GNN and XNNN also improve the quality of explanations by optimising discrete scores or generating prototype graphs [17]. Hybrid approaches improve GNN explainability by combining neural and symbolic reasoning. They improve factual and counterfactual explanations, addressing limitations in purely neural or symbolic methods. Applications in biomedicine and fraud detection show their practical benefits, although challenges such as computational complexity and scalability remain [20].

6. Applications and Future Directions

In previous sections, we examined the challenges and limitations of GNN explainability and NeSy AI, including issues related to scalability, rule complexity, and the integration of symbolic reasoning with neural networks. The neurosymbolic integration of rule learning and GNN explainability has already shown impact in multiple domains. In biomedicine, hybrid pipelines that combine GNN-derived embeddings with differentiable rule miners have surfaced interpretable associations between molecules, genes, and diseases, supporting drug repurposing and personalised therapy design [14, 11]. For example, in a noisy SARS-CoV-2 interaction network, a neurosymbolic method rediscovered the relationship 'hydroxychloroquine inhibits SARS-CoV-2' with

greater precision 89%, while also providing an explicit rule that clinicians could inspect and validate [37]. In social networks and recommender systems, explainable GNNs extract human readable if–then recommendations—for example, "if the user has liked items in category A and belongs to community C, then recommend product X" improving both accuracy and user trust [2, 7]. In autonomous driving, dynamic rule learners encode traffic regulations and safety constraints as symbolic rules that adapt in real time to new sensor inputs, enabling safe planning alongside high fidelity perception [15]. Neural XAI methods for GNNs, such as gradient based techniques, provide model explanations but struggle with instability and domain level interpretability. Symbolic XAI offers transparency, but faces challenges with scalability and deployment in real time. NeSy combines Rule-based reasoning with neural methods, offering a more interpretable and potentially scalable solution. Nevertheless, three challenges remain central to real world use: first, existing rule mining methods struggle to scale to graphs with millions of nodes and edges; second, many frameworks depend heavily on domain specific priors or ontologies, limiting applicability to new contexts; and third, there is no unified standard for evaluating the joint quality of neural predictions, symbolic rule fidelity, and explanation completeness. These gaps can be addressed by optimising algorithms to reduce computational complexity, exploring hierarchical or distributed frameworks for scalability, and using large language models (LLMs) for improved natural language translation of rules, thereby enhancing accessibility for domain experts.

FSAM[26], which maps the semantic structure between GNN layers, motivates a complementary direction: the automated evaluation of GNN explanations using neurosymbolic reasoning. Existing protocols largely rely on fidelity, sparsity and stability, but rarely assess whether explanations are logically consistent, free of redundancy, and semantically meaningful. A neurosymbolic evaluator could formalise rules and subgraphs into a symbolic representation and then check consistency, redundancy, coverage and robustness, thereby complementing fidelity based measures with logical and semantic assessment.

6.1. Human Readable Knowledge Extraction

Current neurosymbolic methods can produce rules that mimic model behaviour, but these rules often remain complex or overly numerous. Future work should focus on learning compact and semantically meaningful rule sets that align with domain knowledge. For example, a healthcare rule such as "if a patient is older than 60 with heart disease, the probability of a cardiovascular event is higher" is clinically interpretable. Reducing the complexity of the rules improves usability by focussing on valid and relevant rules. Techniques such as hierarchical rule induction, concept bottleneck modules, and interactive pruning interfaces will be essential for distilling extensive rule collections into a smaller, high-value subset that captures the most critical decision drivers [43, 44, 45].

6.2. Prototype Graph Generation for Global Explanations

While instance-level explanations highlight local decision factors and rule-based methods capture symbolic structure, few approaches generate prototypical graphs that illustrate class-level behaviour in a visually intuitive form [44]. Extending model-level explainers, such as XGNN with semantic annotations that bind generated subgraphs to extracted rules, would enable practitioners to see both the structural motifs and the logical conditions underlying each class. Such prototype graphs, annotated with human readable captions, could serve as useful tools for model validation, teaching and regulatory compliance. By pursuing these directions, streamlined human readable rule extraction, annotated prototype graph generation, and automated neurosymbolic evaluation systems can deliver not only predictive performance but also transparent, actionable insights for deployment in critical domains.

7. Conclusion

This survey outlines the integration of the explainability of NeSy and GNN, addressing both theoretical and practical challenges. Our analysis of Table 2 reveals a critical gap: no existing approach simultaneously delivers local saliency, global summaries, human readable rule sets, and prototype graph generation. We propose three directions for advancing NeSy-based GNN explainability: (1) distilled rule extraction to produce compact, domain-specific logic; (2) dual mode explainers that provide both local and global insight by integrating annotated prototype graphs with symbolic conditions; and (3) unified evaluation benchmarks that assess fidelity, interpretability, and scalability. Pursuing these directions will enable neurosymbolic GNNs to match state-of-the-art performance on graph structured tasks while also providing transparent, actionable insight. A unified evaluation framework is needed for GNN explainability that combines fidelity analysis with assessment of human readable rule extraction and usefulness. We anticipate that models capable of generating prototype graph outputs linked to concise rules will accelerate adoption in regulated domains such as healthcare, finance, and autonomous systems by offering case level transparency and global rule audits. Establishing comprehensive benchmarks and human in the loop validation protocols will be essential to ensure these systems are robust and trustworthy in real world deployments.

Acknowledgements

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

[1] A. D. Garcez, L. C. Lamb, Neurosymbolic ai: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. doi:10.1007/s10462-023-10448-w.

[2] B. C. Colelough, W. Regli, Neuro-symbolic ai in 2024: A systematic review, arXiv (2025). doi:10.48550/arxiv.2501.05435.

[3] D. Yu, B. Yang, D. Liu, H. Wang, S. Pan, A survey on neural-symbolic learning systems, Neural Networks 166 (2023) 105–126. doi:10.1016/j.neunet.2023.06.028.

[4] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, T. P. Singh, Neuro-symbolic artificial intelligence: a sur-

vey, Neural Computing and Applications 36 (2024) 12809–12844. doi:10.1007/s00521-024-09960-z.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (2020) 4–24. doi:10.1109/tnnls.2020.2978386.

[6] R. Ying, D. Bourgeois, J. You, M. Zitnik, Gnnexplainer: Generating explanations for graph neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019, pp. 1–4. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf.

[7] L. Veyrin-Forrer, A. Kamal, S. Duffner, M. Plantevit, C. Robardet, On gnn explainability with activation rules, Data Mining and Knowledge Discovery 38 (2022) 3227–3261. doi:10.1007/s10618-022-00870-z.

[8] M. Grohe, The logic of graph neural networks, in: Proceedings of the 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS '21), IEEE, 2021, pp. 1–17. doi:10.1109/LICS52264.2021.9470677.

[9] R. Evans, E. Grefenstette, Learning explanatory rules from noisy data, Journal of Artificial Intelligence Research 61 (2018) 1–64. doi:10.1613/jair.5714.

[10] J. Kakkad, J. Jannu, K. Sharma, C. Aggarwal, S. Medya, A survey on explainability of graph neural networks, arXiv preprint arXiv:2306.01958 (2023) 3.

[11] V. T. Ho, D. Stepanova, M. H. Gad-Elrab, E. Kharlamov, G. Weikum, Rule learning from knowledge graphs guided by embedding models, in: Lecture notes in computer science, 2018, pp. 72–90. doi:10.1007/978-3-030-00671-6_5.

[12] C. Geng, Z. Zhao, Z. Wang, H. Ye, X. Si, Extracting interpretable logic rules from graph neural networks, 2025. URL: https://arxiv.org/abs/2503.19476, arXiv preprint arXiv:2503.19476.

[13] H. Zhang, B. Wu, X. Yuan, S. Pan, H. Tong, J. Pei, Trustworthy graph neural networks: Aspects, methods, and trends, Proceedings of the IEEE 112 (2024) 97–139. doi:10.1109/jproc.2024.3369017.

[14] L. N. DeLong, R. F. Mir, J. D. Fleuriot, Neurosymbolic ai for reasoning over knowledge graphs: A survey, IEEE Transactions on Neural Networks and Learning Systems (2024) 1–21. doi:10.1109/tnnls.2024.3420218.

[15] X. Zhang, V. S. Sheng, Neuro-symbolic ai: Explainability, challenges, and future trends, arXiv preprint arXiv:2411.04383 (2024). doi:10.48550/arXiv.2411.04383.

[16] A. Longa, V. Lachi, G. Santin, M. Bianchini, B. Lepri, P. Lio, F. Scarselli, A. Passerini, Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities, arXiv preprint arXiv:2302.01018 (2023). doi:10.48550/arXiv.2302.01018.

[17] M. Bugueño, R. Biswas, G. De Melo, Graph-based explainable ai: A comprehensive survey, https://hal.science/hal-04660442v1, 2024. HAL preprint.

[18] D. Hossain, J. Y. Chen, A study on neurosymbolic artificial intelligence: Healthcare perspectives, arXiv preprint arXiv:2503.18213 (2025). doi:10.48550/arXiv.2503.18213, 18 pages.

[19] M. J. Khan, F. Ilievski, J. G. Breslin, E. Curry, A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge, Deleted Journal (2024) 1–24. doi:10.3233/nai-240719.

[20] P. Barbiero, F. Giannini, G. Ciravegna, M. Diligenti, G. Marra, Relational concept bottleneck models, Advances in Neural Information Processing Systems 37 (2024) 77663–77685.

[21] F. Baldassarre, H. Azizpour, Explainability techniques for graph convolutional networks, arXiv preprint arXiv:1905.13686 (2019).

[22] P. E. Pope, et al., Explainability methods for graph convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10772–10781.

[23] M. S. Schlichtkrull, N. D. Cao, I. Titov, Interpreting graph neural networks for nlp with differentiable edge masking, arXiv preprint arXiv:2010.00577 (2020).

[24] Q. Huang, M. Yamada, Y. Tian, D. Singh, Y. Chang, Graphlime: Local interpretable model explanations for graph neural networks, IEEE Transactions on Knowledge and Data Engineering 35 (2022) 6968–6972. doi:10.1109/TKDE.2022.3187455.

[25] H. Yuan, J. Tang, X. Hu, S. Ji, Xgnn: Towards model-level explanations of graph neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20), ACM, 2020, pp. 430–438. doi:10.1145/3394486.3403085.

[26] K. Raj, A. Mileo, Towards understanding graph neural networks: Functional-semantic activation mapping, in: Neural-Symbolic Learning and Reasoning: 18th International Conference, NeSy 2024, Barcelona, Spain, September 9–12, 2024, Proceedings, Part II, Springer, 2024, pp. 98–106. doi:10.1007/978-3-031-71170-1_11.

[27] M. Nandan, S. Mitra, D. De, Graphxai: A survey of graph neural networks (gnns) for explainable ai (xai), Neural Computing and Applications 37 (2025) 10949–11000. doi:10.1007/s00521-025-11054-3.

[28] T. Funke, M. Khosla, A. Anand, Hard masking for explaining graph neural networks, in: Advances in Neural Information Processing Systems, 2020, pp. 10–15.

[29] A. Behnam, B. Wang, Graph neural network causal explanation via neural causal models, in: European Conference on Computer Vision, Springer, 2024, pp. 410–427.

[30] H. Yuan, et al., On explainability of graph neural networks via subgraph explorations, in: International Conference on Machine Learning, PMLR, 2021, pp. 12241–12252.

[31] R. Schwarzenberg, et al., Layerwise relevance visualization in convolutional text graph classifiers, arXiv preprint arXiv:1909.10911 (2019).

[32] T. Schnake, et al., Higher-order explanations of graph neural networks via relevant walks, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 7581–7596.

[33] Y. Zhang, D. Defazio, A. Ramesh, Relex: A model-agnostic relational model explainer, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 1042–1049.

[34] M. Vu, M. T. Thai, Pgm-explainer: Probabilistic graphical model explanations for graph neural networks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 12225–12235.

[35] S. Azzolin, A. Longa, P. Barbiero, P. Liò, A. Passerini, Global explainability of gnns via logic combination

of learned concepts, arXiv preprint arXiv:2210.07147 (2022).

[36] B. Armgaan, M. Dalmia, S. Medya, S. Ranu, Graphtrail: Translating gnn predictions into human-interpretable logical rules, Advances in Neural Information Processing Systems 37 (2024) 123443–123470.

[37] E. Tsamoura, T. Hospedales, L. Michael, Neural-symbolic integration: A compositional perspective, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 5051–5060. doi:10.1609/aaai.v35i6.16639.

[38] T. Dash, A. Srinivasan, L. Vig, Incorporating symbolic domain knowledge into graph neural networks, Machine Learning 110 (2021) 1609–1636. doi:10.1007/s10994-021-05966-z.

[39] A. Pluska, P. Welke, T. Gärtner, S. Malhotra, Logical distillation of graph neural networks, arXiv preprint arXiv:2406.07126 (2024). doi:10.48550/arXiv.2406.07126.

[40] Z. Xu, P. Ye, H. Chen, M. Zhao, H. Chen, W. Zhang, Ruleformer: Context-aware differentiable rule mining over knowledge graph, arXiv preprint arXiv:2209.05815 (2022). doi:10.48550/arXiv.2209.05815.

[41] X. Pei, X. Deng, S. Tian, K. Xue, Efficient privacy preserving graph neural network for node classification, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10096911.

[42] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, L. De Raedt, Deepproblog: Neural probabilistic logic programming, arXiv preprint arXiv:1805.10872 (2018). doi:10.48550/arXiv.1805.10872.

[43] C. Glanois, X. Feng, Z. Jiang, P. Weng, M. Zimmer, D. Li, W. Liu, Neuro-symbolic hierarchical rule induction, arXiv preprint arXiv:2112.13418 (2021). URL: https://arxiv.org/abs/2112.13418.

[44] B. Liang, Y. Wang, C. Tong, Ai reasoning in deep learning era: From symbolic ai to neural–symbolic ai, Mathematics 13 (2025) 1707. doi:10.3390/math13111707.

[45] R. Vavekanand, G. Kumar, S. Kurbanova, A lightweight physics-conditioned diffusion multi-model for medical image reconstruction, Biomedical Engineering Communications 5 (2026) 12. doi:10.53388/BMEC2026012.