# A New Concept for Explaining Graph Neural Networks

Anna Himmelhuber[1,2], Stephan Grimm[1], Sonja Zillner[1], Martin Ringsquandl[1], Mitchell Joblin[1] and Thomas Runkler[1,2]

[1]Siemens AG, Munich, Germany
[2]Technical University of Munich, Munich, Germany

**Abstract**

Graph neural networks (GNNs), similarly to other connectionist models, lack transparency in their decision-making. A number of sub-symbolic approaches, such as generating importance masks, have been developed to provide insights into the decision making process of such GNNs. These are first important steps on the way to model explainability, but leaving the interpretation of these sub-symbolic explanations to human analysts can be problematic since humans naturally rely on their background knowledge and therefore also their biases about the data and its domain. To overcome this problem we introduce a conceptual approach by suggesting model-level explanation rule extraction through a standard white-box learning method from the generated importance masks.

**Keywords**

Graph Neural Networks, Explainable AI, Decision Trees

## 1. Introduction

Many important real-world data sets come in the form of graphs or networks. These include social networks, knowledge graphs, protein-interaction networks, the World Wide Web and many more. Graph neural networks designed to leverage relational inductive biases induced by a graph via a neural message passing strategy. Unlike standard neural networks, graph neural networks encode relational information in addition to node and edge feature information [1]. Similarly to other connectionist models, GNNs lack transparency in their decision-making. Since the unprecedented levels of performance of such AI methods lead to their increasingly dominant role, there is an emerging need to understand the decision-making process of such systems [2]. The growing concern regarding potential bias in these models creates demand for model transparency and interpretability. In other words, model explainability is a prerequisite for building trust and the adoption of an AI system in high stakes domains requiring reliability and safety such as healthcare and mission critical industrial applications with significant economic implications, e.g. predictive maintenance. This includes being able to explain the decision making processes of GNNs in learning from complex graph structure. While the increasing demand for explainable AI is a positive development, explanatory models are often built for AI researchers, making them hard to understand for non-experts. Our approach targets experts of the application domain. Explainable AI and with it the widespread

NeSy'21

application of AI models are more likely to succeed if the evaluation of these models is focused more on the user's needs [3].

While a variety of explainable models for graph neural networks are being developed [4] [5] [6], a common weakness among these approaches is the user's responsibility for compiling and comprehending the symbols in the final step, relying on their own implicit form of knowledge and reasoning about them [7]. How the input, the output and the emitted symbols relate to each other, is often an implicit cognitive intuition on the user side, with differing grades of tolerance in their interpretation. Humans utilize their background knowledge and therefore also their biases about the data when drawing conclusions. This can lead to the deduction of inconsistent or meaningless explanations.

In this position paper, we address this weakness of existing approaches by proposing a post-processing rule-based companion to such a sub-symbolic explainer method, with the conceptual schema shown in Figure 1. Thereby we want to complement the sub-symbolic instance-level explanations with model-level rules. By extracting and aggregating global rule-based explanations through a standard white-box machine learning method from the generated explainer subgraph, we reduce the amount of additional interpretation needed by the user and provide a model-level explanation, that captures explanations about the global behavior of a model by investigating what input patterns can lead to a specific prediction. As an example of such an approach, we developed a novel method known as SUBGREX. We take the output of the state-of-the-art explainer method as input as well as graph-specific attributes such as node distances and network motifs and use decision trees to generate rule-based explanations.

## 2. Rule-Based Explanations of Subgraphs

The goal is to generate systems that can provide model-level rule-based explanations, which don't rely on the user's understanding of the domain. We can achieve this by combining the results of a sub-symbolic explainer method with a white-box rule generator which satisfies the representation needs for human comprehensibility and reasoning. The rule-based explanation generation isn't a stand-alone approach, but an add-on post-processing method in order to enhance the explanations and make them more user-centric. After training a GNN, the GNNs decision making process is interpreted by identifying a sparse receptive field containing influential elements. Our post-processing approach consists of taking these initial symbolic explanations and lifting them to the level of rules. The proposed process can be seen in Algorithm 1 for a node classification task, where an edge mask $M_E$ and node feature mask $M_X$ are generated by a sub-symbolic explainer model $F_{ex}$, and subsequently rules for edge and node features are created by the white-box models $D_E$ and $D_X$. The rules are created through a classification process, where the individual edges and features are assigned binary labels "influential" or "not-influential" based on their masking value. The white-box model is trained using the binary masks as the target variable. The generated rules then function as a model-level explanation for

the respective classes of the original classification problem and furthermore, introduce a verbalization element that can make explanations more comprehensible for the user.

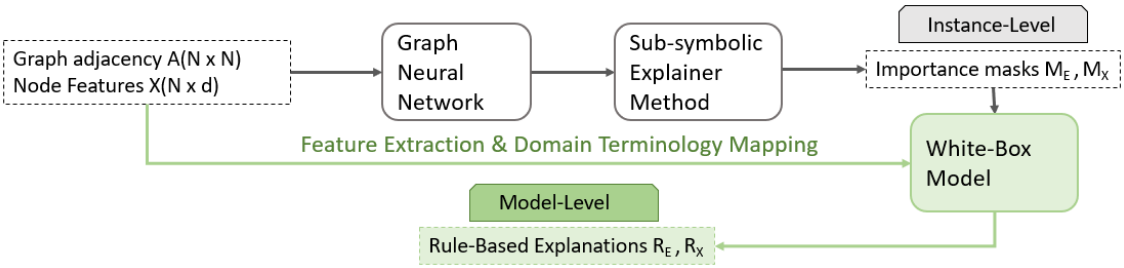


**Figure 1:** Conceptual schema of generating rule-based explanations

An important part of generating an user-centric explanation is to make explanation customizable and include provenance, e.g. by including information about the domain knowledge utilized by the system to increase user-understandability and acceptability. Such domain knowledge can be included by extracting attributes from the subgraphs depending on the domain and classification task. An example of such inclusion of domain knowledge are graph-specific attributes, such as motifs mapped to the domain language (see Algorithm 1). In a more formalized setting, relational rules could also be included e.g. to represent domain-specific constraints.

## 2.1. SUBGREX Model

To test this conceptual approach, we propose our SUBGREX model, which enables the generation of model-level explanations. Since the state-of-the-art method outperforms alternative baseline approaches by 43.0% in explanation accuracy [4], we chose the GNNExplainer as the sub-symbolic explainer method $F_{ex}$. As a method for extracting the rules, the standard machine learning mechanism decision tree is employed with $D_E = ID3_E, D_X = ID3_X$. The decision trees can be linearized into decision rules $R_E$ and $R_X$. The classification by the decision tree is binary and based on whether the node $v_i$ is considered influential by the chosen subgraph generation method stemming from the edge mask $M_E$, which is a 0/1-valued vector and therefore our target attribute for the white-box model $D_E$. The attributes $a_1, ...a_L$ for the $D_E$ input are extracted from node, edge, and graph meta information from the explanation subgraphs. Since graph architecture offers more information than tabular data, it is vital to take graph-specific data such as network motifs into account. In order to enhance the user-friendliness of our approach, the attributes are personalized to the domain by mapping them to the respective domain terminology as shown in Algorithm 1. The feature selection rules are generated analogously with the corresponding node feature mask $M_X$.

Algorithm 1 Rule-Based Explanation Generation

Inputs: Explanation Subgraph Model: $F_{ex}$; Graph adjacency: $A(N \times N)$; Node features: $X(N \times d)$; Set of attributes: $L = \{a_1, ...a_L\}$; Attribute mapping dictionary: $T = \{a_1 : t_1, ...a_L : t_T\}$; White-box model: $D_E$, $D_X$
for $category = 1, 2, \ldots, K$ do
    for $node = 1, 2, \ldots, N$ do
        $M_X$, $M_E = F_{ex}(node, A, X)$
        for support_node in $M_E$ do
            If $M_E(support\_node) \rightarrow y_{support}$
            $x_{support}$ = Extract-Attributes($support\_node$, $M_X$, $M_E$, $L$)
            $x_{support_m}$ = Map-Attributes($x_{support}$, $T$)
            $D_E$.fit($x_{support_m}$, $y_{support}$)
        end for
        for support_node in $M_X$ do
            If $M_E(support\_node) \rightarrow y_{support}$
            $x_{support}$ = Extract-Attributes($support\_node$, $M_X$, $M_E$, $L$)
            $x_{support_m}$ = Map-Attributes($x_{support}$, $T$)
            $D_X$.fit($x_{support_m}$, $y_{support}$)
        end for
    end for
end for

## 2.2. Preliminary Results

The MUTAG dataset, which is comprised of molecule graphs labeled according to their mutagenic effec [8] is used. We carry out graph classification with a vanilla 2-layer Graph Convolutional Network. In the next step, we generate a corresponding edge mask $M_E$ and node feature mask $M_X$ with the GNNExplainer. For our experiment, we use a subset of 30 subgraphs that are symmetrically either classified as *mutagen* or *nonmutagen*, with respectively 14 feature attributes. Decision tree training and testing data follows a random 80%-20% split. In the following, linearized rule-based explanations from the respective decision trees are listed:

$R_X^{Mutagen}$ ={`If molecule contains atom C AND atom O; If molecule contains atom C AND atom S AND no atom O; If molecule contains atom H AND no atom C`} Sensitivity (Sens): 0.98; Accuracy (Acc): 0.83
$R_E^{Mutagen}$ ={ `If atom has more than 2 bonds AND if atom is part of an atom ring; If atom has only one bond AND is not part of an atom ring`} Sens: 0.72; Acc: 0.64
$R_X^{Nonmutagen}$ ={`If molecule contains no atom N AND no atom H` } Sens: 0.95: Acc: 0.94

The extracted node rules give the user an comprehensible idea of the network motifs that are influential in the GNN's decision making, either being a part of a cycle motif (atom ring) with a high degree (in this case, more than 2 bonds) or not being part of

the cycle motif and having only one bond. Also the node feature rules represent clear explanations which atoms, e.g. atom H, are influential in the respective classification. We can see, that SUBGREX correctly identifies, that atom N and atom H not being in the molecule is influential for classifying it as nonmutagen, with chemical group $NH_2$ being known to be mutagenic [8]. No node selection rules with a low Gini impurity could be extracted for the *nonmutagen* category. This indicates that no feature is influential enough to generate a model-level explanation of when a molecule is classified as nonmutagen. We report the sensitivity to show to which extent our explanations can function as a model-level explanation. Sensitivity indicates the ability of the decision tree to correctly classify a node or feature as influential. The sensitivity lies between 72% for mutagen node selection rules and 98% for mutagen feature selection rules and indicate a a high level of model explainability.

## 3. Conclusion

We have proposed a conceptual vision for how to approach generating enhanced, more user-centric rule-based explanations from sub-symbolic instance-level explanations, which improve model-level understanding. We also report on initial experiments that demonstrate the validity of our method. Even with the rather simple SUBGREX method we show some surprisingly effective results in terms of meaningfulness of explanations and high sensitivity. In further research we plan to evaluate and compare the effectiveness of different white-box models including semantic web technologies such as inductive logic programming.

This work was supported by the German Federal Ministry of Economics and Technology (BMWI) in the project RAKI (no. 01MD19012D).

## References

[1] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[3] T. Miller, P. Howe, L. Sonenberg, Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences, IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI) (2017).

[4] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, in: Advances in neural information processing systems, 2019, pp. 9244–9255.

[5] 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, 2020, pp. 430–438.

[6] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10772–10781.

[7] D. Doran, S. Schulz, T. R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, arXiv preprint arXiv:1710.00794 (2017).

[8] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, C. Hansch, Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity, Journal of medicinal chemistry 34 (1991) 786–797.