Extended Abstract: Towards Robust Fidelity for Evaluating Explainability of Graph Neural Networks

Xu Zheng[1*], Farhad Shirani[1*], Tianchun Wang[2], Wei Cheng[3], Zhuomin Chen[1], Haifeng Chen[3], Hua Wei[4], Dongsheng Luo[1]

[1]Florida International University, [2]Pennsylvania State University, [3]NEC Laboratories America,[4]Arizona State University

{xzhen019,fshirani,zchen051,dluo}@fiu.edu,tkw5356@psu.edu,{weicheng,haifeng}@nec-labs.com,hua.wei@asu.edu

ABSTRACT

Graph Neural Networks (GNNs) have emerged as pivotal architectures in analyzing graph-structured data, and their expansive application in sensitive domains requires a comprehensive understanding of their decision-making processes — necessitating a framework for GNN explainability. An explanation function for GNNs takes a pre-trained GNN along with a graph as input, to produce a 'sufficient statistic' subgraph with respect to the graph label. A main challenge in studying GNN explainability is to provide measures that evaluate the performance of these explanation functions. A popular family of measurements is *Fidelity*, including Fid_+, Fid_-, and Fid_Δ. Fidelity measures the faithfulness of an explanation by calculating the difference in the model output when keeping (Fid_-) or removing (Fid_+) the explanation. In this paper, we argue that Fidelity measurements are unreliable due to potential distribution shifts between original graphs and the subgraph generated by keeping or removing explanation subgraphs . Subsequently, a robust class of fidelity measures is introduced. Extensive empirical analysis on both synthetic and real datasets are provided to illustrate that the proposed metrics are better aligned with gold standard metrics. The project website is available at: https://trustai4s-lab.github.io/fidelity.html.

KEYWORDS

Graph Neural Networks, Trustworthy Learning, Explainable AI

1 INTRODUCTION

With the proliferation of Graph Neural Networks (GNNs) in sensitive sectors like healthcare and fraud detection, the demand for understanding their decision-making processes has grown significantly [14]. Recently, explanation techniques have been proposed for GNNs, most commonly focusing on identifying a subgraph that dominates the model's prediction in a post-hoc sense [18].

In the design and study of explainable GNNs, both model design and choice of evaluation metrics are important. While most efforts have primarily been made to develop new network architectures and optimization objectives to achieve more accurate explanations [11, 16, 18], in this paper, we underscore the critical importance of choosing the right evaluation metrics for the achieved explanations. In an ideal scenario, quantitative evaluation of an explanation subgraph can be achieved by comparing it with a gold standard or ground truth explanation [16]. However, in real-world applications, such ground truth explanation subgraphs are a rarity, often making direct comparisons impracticable. In lieu of this, *surrogate* fidelity metrics, namely Fid_+, Fid_-, and Fid_Δ, have been included to measure the faithfulness of explanation subgraphs. At its core, the intuition driving such metrics is straightforward: if a subgraph is discriminative to the model, the prediction should change significantly when it is removed from the input. Otherwise, the prediction should be maintained. Hence, Fid_+ is defined as the difference in accuracy (or predicted probability) between the original prediction and the new predictions of non-explanation subgraph which is obtained by masking out the explanation subgraph [13], and Fid_- measures the difference between predictions of the original graph and explanation subgraph [17]. As prevailing standards, these Fidelity metrics and their variants have been widely used in existing popular platforms, such as GraphFramEx [2], GraphXAI [1], GNNX-BENCH [9], and DIG [10].

Although intuitively correct, we argue that the aforementioned Fidelity metrics come with significant drawbacks due to the impractical assumption that the to-be-explained model can make accurate predictions of the explanation subgraph (in Fid_-) or non-explanation subgraph (in Fid_+). This does not hold in a wide range of real-world scenarios, because when edges are removed, the resultant subgraphs might be Out Of Distribution (OOD) [4]. For example, in MUTAG dataset [3], each graph is a molecule with nodes representing atoms and edges describing the chemical bonds. The functional group NO_2 is considered the dominating subgraph that causes a molecule to be positively mutagenic. The explanation subgraph only consists of 2 edges, which is much smaller than whole molecular graphs. Such disparities in properties introduce distribution shifts, putting the Fidelity metrics on shaky grounds, because of the violation of a key assumption in machine learning: the training and test data come from the same distribution [7].

To build an evaluation foundation for eXplainable AI (XAI) in the graph domain, In this paper, we investigate robust fidelity measurements for evaluating the correctness of explanations. There are several non-trivial challenges associated with this problem. First, the to-be-explained GNN model is usually evaluated as a black-box model, which cannot be re-trained to ensure the generalization capacity [7]. Second, the evaluation method is required to be stable and ideally deterministic. As a result, complex parametric methods, such as adversarial perturbations [6, 8], are not suitable as the results are affected by randomly initiated parameters. We propose a generalized class of surrogate fidelity measures that are robust to distribution shift issues in a wide range of scenarios based on information theory. Our contributions are summarized as follows.

- We identify the OOD issue in Fidelity measurement in the explainable graph learning domain.
- We introduce novel evaluation metrics that are resilient to distribution shifts, enhancing their applicability in real-world contexts.

*Equal Contribution.

- Through rigorous empirical analyses on a diverse mix of synthetic and real datasets, we validate that our approach resonates well with gold standard benchmarks.

2 PRELIMINARIES

2.1 Notations

We parameterize a labeled graph G by a tuple $(\mathcal{V}, \mathcal{E}; Y, \boldsymbol{X}, \boldsymbol{A})$, where i) $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ is the node set, ii) $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, iii) Y is the graph class label taking values from finite set of classes $\mathcal{Y}$, iv) $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the feature matrix, where the i-th row of $\boldsymbol{X}$, denoted by $\boldsymbol{X}_i \in \mathbb{R}^{1 \times d}$, is the d-dimensional feature vector associated with node $v_i, i \in [n]$, and v) $\boldsymbol{A} \in \{0,1\}^{n \times n}$ is the adjacency matrix. The graph parameters $(Y, \mathbf{A}, \boldsymbol{X})$ are generated according to the joint probability measure $P_{Y,\mathbf{A},\boldsymbol{X}}$. Note that the adjacency matrix determines the edge set $\mathcal{E}$, where $A_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$, and $A_{ij} = 0$, otherwise. We write $|G|$ and $|\mathcal{E}|$ interchangeably to denote the number of edges of G. Throughout this paper, we use lower-case letters, such as $g, y, \mathbf{x}$, and $\mathbf{a}$, to represent realizations of the random objects $G, Y, \mathbf{X}$ and $\mathbf{A}$, respectively. Given a labeled graph $G = (\mathcal{V}, \mathcal{E}; Y, \boldsymbol{X}, \boldsymbol{A})$, we denote the corresponding graph without label as $\overline{G}$, and parameterize it by $(\mathcal{V}, \mathcal{E}; \boldsymbol{X}, \boldsymbol{A})$.

Graph Classification Task: In the graph classification task under consideration:

- A set of labeled training data $\mathcal{T} = \{(\overline{G}_i, Y_i) | Y_i \in \mathcal{Y}, i \in [|\mathcal{T}|]\}$, where $(\overline{G}_i, Y_i)$ corresponds to the i-th graph and its associated class label. The pairs $(\overline{G}_i, Y_i), i \in [|\mathcal{T}|]$ are generated independently according to an identical joint distribution induced by $P_{Y,\boldsymbol{X},\mathbf{A}}$.
- A classification function (GNN model) $f(\cdot)$ trained to classify an unlabeled input graph $\overline{G}$ into its class Y. It takes $\overline{G}$ as input and outputs a probability distribution P_Y on alphabet $\mathcal{Y}$. The reconstructed label $\widehat{Y}$ is produced randomly based on P_Y.

In graph classification tasks, G_i is a random graph whose distribution is determined by the (general) joint distribution $P_{Y,\boldsymbol{A},\boldsymbol{Z}}$, with the GNN model $f(\cdot)$ trained to predict the label for graph G based on the learned representation of $\overline{G}$. Formally we define a classifier as follows.

*Definition 2.1 (**Classifier**).* For a classification task with underlying distribution $P_{Y,\boldsymbol{X},\mathbf{A}}$, a classier is a function $f : \overline{\mathcal{G}} \to \Delta_{\mathcal{Y}}$. For a given $\epsilon > 0$, the classifier is called ϵ-accurate if $P(\widehat{Y} \neq Y) \leq \epsilon$, where $\widehat{Y}$ is produced according to probability distribution $f(\overline{G})$.

Intuitively, an explanation in the context of graph learning is a **subgraph** which is an almost *sufficient statistic* of the input graph with respect to the output label. A high-level description of the explainability problem is provided in the following.

*Definition 2.2 (**Explanation** [11, 16]).* Given a graph classifier $f(\cdot)$ and an input graph $\overline{G}$, the explanation is a sub-graph $\overline{G}^{(exp)} = (\mathcal{V}^{(exp)}, \mathcal{E}^{(exp)})$, such that $\overline{G}^{(exp)}$ is minimal and sufficient.

2.2 Fidelity

Fidelity evaluates the faithfulness of an explanation by measuring the differences in the classifier's output when (only) keeping or removing the explanation subgraph [17]. Fid_+ is defined as the change of predicted probability between the original graph and the new predictions after removing explanations [13]. Fid_- quantifies the prediction difference by only keeping the explanation subgraph and masking out the remaining part. Fid_Δ is the difference between Fid_+ and Fid_-. Given an input graph $\overline{G}$ with label y, a classifier $f(\cdot)$ and an explanation $\overline{G}^{(exp)}$, these fidelity measurements are formally defined as follows.

$$\begin{aligned} Fid_+ &\triangleq f(\overline{G})_y - f(\overline{G} - \overline{G}^{(exp)})_y, \\ Fid_- &\triangleq f(\overline{G})_y - f(\overline{G}^{(exp)})_y, \\ Fid_\Delta &\triangleq Fid_+ - Fid_-. \end{aligned} \tag{1}$$

However, we argue that it is not well-behaved in a wide range of scenarios of interest due to the OOD issue mentioned in the introduction. To elaborate, for a good classifier, which has a low probability of error, the prediction of f on $\overline{G}$ is reliable. However, this is not necessarily true for the $f(\overline{G} - \overline{G}_s)$ and $f(\overline{G}_s)$ terms. The reason is that subgraphs $\overline{G} - \overline{G}^{(exp)}$ and $\overline{G}^{(exp)}$ are not typical realizations. For instance, in many applications, it is very unlikely or impossible to observe the explanation graph in isolation. As a result, the predictions made by f are unreliable.

3 ROBUST FIDELITY

Generally, in scenarios where $\overline{G}^{(exp)}$ and $\overline{G} - \overline{G}^{(exp)}$ are not typical with respect to the distribution of $\overline{G}$, the existing fidelity measure may not be well-behaved. We address this by introducing a class of modified fidelity measures by modifying the definitions of Fid_+ and Fid_- in equation 1. To this end, we define the stochastic graph sampling function $E_\alpha : \overline{G} \mapsto \overline{G}_\alpha$ with edge erasure probability $\alpha \in [0,1]$. That is, $E_\alpha(\cdot)$ takes a graph $\overline{G}$ as input, and outputs a sampled graph $\overline{G}_\alpha$ whose node set is the same as that of $\overline{G}$, and its edges are sampled from $\overline{G}$ such that each edge is included with probability α and erased with probability $1 - \alpha$, independently of all other edges. We introduce the following generalized class of surrogate fidelity measures, and show that they are robust to OOD issues in a wide range of scenarios:

$$Fid_{\alpha_1,+} \triangleq f(\overline{G})_y - \mathbb{E} f(\overline{G} - E_{\alpha_1}(\overline{G}^{(exp)}))_y, \tag{2}$$

$$Fid_{\alpha_2,-} \triangleq f(\overline{G})_y - \mathbb{E} f(\overline{G}^{(exp)} + E_{\alpha_2}(\overline{G} - \overline{G}^{(exp)}))_y, \tag{3}$$

$$Fid_{\alpha_1,\alpha_2,\Delta} \triangleq Fid_{\alpha_1,+} - Fid_{\alpha_2,-}, \tag{4}$$

where $\alpha_1, \alpha_2 \in [0,1]$. Note that if $\alpha_1 = 1$ and $\alpha_2 = 0$, we recover the original fidelity measures, i.e., $Fid_{1,+} = Fid_+$, $Fid_{0,-} = Fid_-$, and $Fid_{1,0\Delta} = Fid_\Delta$. On the other hand, if $\alpha_1 = 0$ and $\alpha_2 = 1$, we have $\overline{G} - E_{\alpha_1}(\overline{G}^{(exp)})) = E_{\alpha_2}(\overline{G} - \overline{G}^{(exp)}) + \overline{G}^{(exp)} = \overline{G}$. Consequently, in this case there would be no OOD issue, however, the resulting fidelity measure is not informative since $Fid_{0,1,\Delta} = 0$, for all classifiers f and explanations. In practice, $\alpha_1 < 1$ and $\alpha_2 > 0$ would yield a suitable fidelity measure as this choice alleviates the OOD problem.

Next, we provide the pseudo-code for computing the proposed $Fid_{\alpha_1,+}$ and $Fid_{\alpha_2,-}$ in Alg. 1 and Alg. 2, respectively. Suppose that we have a set of input graphs, $\{\overline{G}_i\}_{i=1}^{\mathcal{T}}$. For each graph $\overline{G}_i$, the explanation subgraph to be evaluated is denoted by $\overline{G}_i^{(exp)}$. The

model to be explained is denoted by $f(\cdot)$. We have two hyper-parameters, M and α_1 in computing $Fid_{\alpha_1,+}$. M is the number of samples and α_1, introduced in equation 2, is the ratio of edges sampled from explanation subgraph. For $Fid_{\alpha_2,-}$, we have another hyper-parameter α_2 instead, which indicates the ratio of edges sampled from non-explanation subgraph.

Algorithm 1 Computating $Fid_{\alpha_1,+}$

Input: A set of input graphs and their subgraphs $\{(\overline{G}_i, \overline{G}_i^{(exp)}\}_{i=1}^{\mathcal{T}}$, a GNN model $f(\cdot)$, hyperparameters M and α_1.
Output: $Fid_{\alpha_1,+}$ of $\{\overline{G}_i^{(exp)}\}_{i=1}^{\mathcal{T}}$.
for each pair $(\overline{G}_i, \overline{G}_i^{(exp)})$ **do**
  for m from 1 to M **do**
    $E_{\alpha_1}(\Psi(\overline{G}_i))) \leftarrow$ sample α_1 edges from $\overline{G}_i^{(exp)}$
    $\overline{G}_{i,m} \leftarrow \overline{G}_i - E_{\alpha_1}(\overline{G}_i^{(exp)})$
    $Fid_{\alpha_1+}[i,m] \leftarrow f(\overline{G}_i)_{y_i} - f(\overline{G}_{i,m})_{y_i}$
  end for
  $Fid_{\alpha_1+}[i] \leftarrow \frac{1}{M}\sum_{m=1}^{M} Fid_{\alpha_1,+}[i,m]$
end for
$Fid_{\alpha_1,+} \leftarrow \frac{1}{\mathcal{T}}\sum_{i=1}^{\mathcal{T}} Fid_{\alpha_1,+}[i]$
Return $Fid_{\alpha_1,+}$.

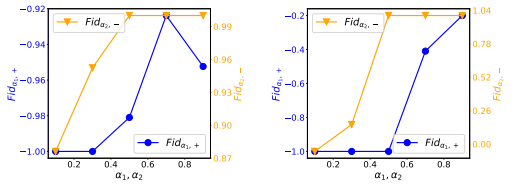
Algorithm 2 Computating $Fid_{\alpha_2,-}$

Input:A set of input graphs and their subgraphs $\{(\overline{G}_i, \overline{G}_i^{(exp)}\}_{i=1}^{\mathcal{T}}$, a GNN model $f(\cdot)$, hyperparameters M and α_2.
Output: $Fid_{\alpha_2,-}$ of $\{\overline{G}_i^{(exp)}\}_{i=1}^{\mathcal{T}}$.
for each pair $(\overline{G}_i, \overline{G}_i^{(exp)})$ **do**
  for m from 1 to M **do**
    $\overline{G}_i^c \leftarrow \overline{G}_i - \overline{G}_i^{(exp)}$
    $E_{\alpha_2}(\overline{G}_i^c) \leftarrow$ sample α_2 edges from $\overline{G}_i^c$
    $Fid_{\alpha_2,-}[i,m] \leftarrow f(\overline{G}_i)_{y_i} - f(E_{\alpha_2}(\overline{G}_i^c) + \overline{G}_i^{(exp)})_{y_i}$
  end for
  $Fid_{\alpha_2,-}[i] \leftarrow \frac{1}{M}\sum_{m=1}^{M} Fid_{\alpha_2,-}[i,m]$
end for
$Fid_{\alpha_2,-} \leftarrow \frac{1}{\mathcal{T}}\sum_{i=1}^{\mathcal{T}} Fid_{\alpha_2,-}[i]$
Return $Fid_{\alpha_2,-}$.

It can be noted that robust fidelity cannot avoid the OOD problem during the evaluation processing. Compared to original Fidelity, our method can alleviate the effect of OOD and estimate the true fidelity score by using sampling methods.

4 EXPERIMENTS

In this section, we empirically verify the effectiveness of the generalized class of surrogate fidelity measures. Two benchmark datasets with ground truth explanations are used for evaluation, BA-2motifs [11] and MUTAG [3]. We consider both GCN and GIN architectures [15, 16] as the models to be explained. We evaluate the accuracy performance of GNN models on training, validation, and test sets. Both GCN and GIN achieve good performances in these datasets, with most test accuracy scores above 0.9. Following routinely adopted settings [11, 16, 18], we can safely assume that both models can correctly use the informative components(motifs) in the input graphs to make predictions.



(a) BA-2motifs (b) MUTAG

Figure 1: Parameter studies on the effects of α_1 and α_2.

4.1 Quantitative Evaluation by Comparing to the Gold Standard

In adopted datasets, motifs are included which determine node labels or graph labels. Thus, the relationships between graph examples and data labels are well-defined by humans. The correctness of an explanation can be evaluated by comparing it to the ground truth motif. Previous studies [11, 16] usually model the evaluation as an edge classification problem. Specifically, edges in the ground-truth explanation are treated as labels, and importance weights given by the explainability method are viewed as prediction scores. Then, AUC scores are considered as the metric for correctness. In this section, we use a more tractable metric, edit distance [5] to compare achieved explanations with the ground-truth motifs as a Gold Standard metric.

Consider an input graph $\overline{G}_i$ and let $\overline{G}_i^{(gt)}$ be the ground-truth explanation subgraph, i.e., the motif. We construct a set of explanation functions, with varying qualities, to evaluate the well-behavedness of the proposed fidelity measures. To elaborate, for a given $\beta_1, \beta_2 \in [0,1]$, we construct an explanation function $\Psi_{\beta_1,\beta_2}(\cdot)$ by random IID sampling of the explanation subgraph edges, and the non-explanation subgraph edges, with sampling rates β_1 and β_2, respectively. That is, to construct $\Psi_{\beta_1,\beta_2}(\overline{G}_i)$, we remove β_1 ratio of edges from ground-truth explanation $\overline{G}_i^{(gt)}$ via random IID sampling, and randomly add $\beta_2 \in [0,1]$ ratio of edges from $\overline{G}_i - \overline{G}_i^{(gt)}$ to $\overline{G}_i^{(gt)}$ by random IID sampling from the non-explanation subgraph. Clearly, the explanation function should receive a better fidelity score for smaller (β_1, β_2). We sweep β_1 and β_2 in the range $[0, 0.1, 0.3, 0.5, 0.7, 0.9]$, and for each combination (β_1, β_2), we randomly sample 10 candidate explanations. We adopt the proposed $Fid_{\alpha_1,+}$, $Fid_{\alpha_2,-}$, $Fid_{\alpha_1,\alpha_2,\Delta}$, where we have taken $\alpha_1 = 1 - \alpha_2 = 0.1$, as well as their counterparts to evaluate their qualities. For each combination, we calculate the average metric scores.

As analyzed in previous works [18], fidelity measurements ignore the size of the explanation. Thus, redundant explanations are usually with high Fid_+ and low Fid_- scores. In the extreme case, with the whole input graph as the explanation, fidelity measures achieve the trivial optimal scores. This limitation is inherent and cannot not solved with the proposed metrics. To fairly compare the proposed metrics with the original ones, for each β_2, given a fidelity measurement, we use the Spearman correlation coefficient [12] between it and the gold standard edit distance to quantitatively evaluate the quality of the metric. Then, we report the average correlation

Table 1: Spearman correlation coefficient between metric and gold standard edit distance.

	Dataset	$Fid_+ \downarrow$	$Fid_{\alpha_1,+} \downarrow$	$Fid_- \uparrow$	$Fid_{\alpha_2,-} \uparrow$	$Fid_\Delta \downarrow$	$Fid_{\alpha_1,\alpha_2,\Delta} \downarrow$	AUC
GCN	BA-2motifs	-0.924	**-1.000**	0.819	**1.000**	-0.990	**-1.000**	-1.000
	MUTAG	-0.190	**-1.000**	-0.276	**1.000**	-0.105	**-1.000**	-1.000
GIN	BA-2motifs	-0.838	**-1.000**	0.905	**1.000**	**-1.000**	**-1.000**	-1.000
	MUTAG	**-1.000**	**-1.000**	0.886	**1.000**	-0.990	**-1.000**	-1.000

scores in Table 1. The number of sampling in our measurements are set to 50.

We have the following observations in Table 1. First, $Fid_{\alpha_1,+}$ consistently yields correlation scores near -1.0 with the edit distance. It signifies a robust inverse relationship between the two metrics. In contrast, the original Fid_+ metric exhibits mixed results with half-positive and half-negative correlations. This inconsistency in Fid_+ underscores the potential superiority and consistency of our proposed $Fid_{\alpha_1,+}$ in aligning closely with the edit distance across various datasets. Moreover, the proposed $Fid_{\alpha_2,-}$ is strongly positively related to gold-standard edit distance compared to the original Fid_-. we have similar observations in $Fid_{\alpha_1,\alpha_2,\Delta}$ and Fid_Δ. Third, we observe that the AUC score of edge classification, which is used in previous papers, is perfectly aligned with the gold standard edit distance, which verifies the correctness of using AUC as the metric.

4.2 Effects of α_1 and α_2.

As shown in our theoretical analysis, α_1 is the rate of removing edges from explanation subgraphs in $Fid_{\alpha_1,+}$ and α_2 is the rate of retaining edges from non-explanation subgraphs in $Fid_{\alpha_2,-}$. To empirically verify the effects of these two parameters, we use the GCN model and vary these two hyper-parameters in the range $[0.1, 0.3, 0.5, 0.7, 0.9]$. Results of Spearman correlation scores are shown in Figure 1. We observe that when $\alpha_1 = 0.1$ and $\alpha_2 = 0.9$, the proposed $Fid_{\alpha_1,+}$ and $Fid_{\alpha_2,-}$ are strongly aligned with the gold standard edit distance. As α_1 increases, the number of removing edges from explanation subgraphs increases, leading to more a severe distribution shifting problem. Thus, the Spearman correlation coefficient between $Fid_{\alpha_1,+}$ and edit distance increases. The similar phenomena can be observed in $Fid_{\alpha_2,-}$. As α_2 decreases, a smaller number of edges will be added from non-explanation subgraphs, which leads to the distribution shift problem.

5 CONCLUSION

In this paper, we have explored the limitations intrinsic in widely used evaluation measures for GNN explainers, including the Fidelity surrogate measures, and identified several pitfalls of conventional fidelity metrics, particularly their vulnerability to distribution shifts. We have introduced a set of evaluation metrics, which utilize a sampling technique to mitigate such distribution shift issues and avoid evaluating out-of-distribution inputs during evaluation. We have provided extensive empirical evaluation to validate the peformance of the proposed metrics across various datasets, demonstrating their alignment with gold standard benchmarks.

REFERENCES

[1] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating Explainability for Graph Neural Networks. *Scientific Data* 10, 144 (2023).

[2] Kenza Amara, Zhitao Ying, Zitao Zhang, Zhichao Han, Yang Zhao, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. 2022. GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks. In *LOG*.

[3] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry* 34, 2 (1991), 786–797.

[4] Junfeng Fang, Xiang Wang, An Zhang, Zemin Liu, Xiangnan He, and Tat-Seng Chua. 2023. Cooperative Explanations of Graph Neural Networks. In *WSDM*. 616–624.

[5] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. 2010. A survey of graph edit distance. *Pattern Analysis and applications* 13 (2010), 113–129.

[6] Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. *NeurIPS* 34 (2021), 3650–3666.

[7] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *NeurIPS* 32 (2019).

[8] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Kumar Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. 2021. Evaluations and Methods for Explanation through Robustness Analysis. In *ICLR*.

[9] Mert Kosan, Samidha Verma, Burouj Armgaan, Khushbu Pahwa, Ambuj Singh, Sourav Medya, and Sayan Ranu. 2023. GNNX-BENCH: Unravelling the Utility of Perturbation-based GNN Explainers through In-depth Benchmarking. *arXiv preprint arXiv:2310.01794* (2023).

[10] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, et al. 2021. DIG: A turnkey library for diving into graph deep learning research. *The Journal of Machine Learning Research* 22, 1 (2021), 10873–10881.

[11] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. *NeurIPS* 33 (2020), 19620–19631.

[12] Jerome L Myers, Arnold D Well, and Robert F Lorch Jr. 2013. *Research design and statistical analysis*. Routledge.

[13] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. 2019. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10772–10781.

[14] Bingzhe Wu, Jintang Li, Junchi Yu, Yatao Bian, Hengtong Zhang, CHaochao Chen, Chengbin Hou, Guoji Fu, Liang Chen, Tingyang Xu, et al. 2022. A survey of trustworthy graph learning: Reliability, explainability, and privacy protection. *arXiv preprint arXiv:2205.10014* (2022).

[15] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *ICLR*.

[16] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *NeurIPS* 32 (2019).

[17] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[18] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. 2021. On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*. PMLR, 12241–12252.