An Explainable AI Library for Benchmarking Graph Explainers

Chirag Agarwal*
Media and Data Science Research Lab, Adobe

Owen Queen*
University of Tennessee, Knoxville

Himabindu Lakkaraju
Harvard University

Marinka Zitnik
Harvard University

ABSTRACT

With Graph Neural Network (GNN) explainability methods increasingly used to understand GNN predictions in critical real-world applications, it is essential to reliably evaluate the correctness of generated explanations. However, assessing the quality of GNN explanations is challenging as existing evaluation strategies depend on specific datasets with no or unreliable ground-truth explanations and GNN models. Here, we introduce G-XAI Bench, an open-source graph explainability library providing a systematic framework in *PyTorch* and *PyTorch Geometric* to compare and evaluate the reliability of GNN explanations. G-XAI Bench provides comprehensive programmatic functionality in the form of data processing functions, GNN model implementations, collections of synthetic and real-world graph datasets, GNN explainers, and performance metrics to benchmark any GNN explainability method. We introduced G-XAI Bench to support the development of novel methods with a strong bent towards developing the foundations of which GNN explainers are most suitable for specific applications and why.

KEYWORDS

Graph Neural Networks, Explainability, Benchmarks

ACM Reference Format:
Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2018. An Explainable AI Library for Benchmarking Graph Explainers. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Workshop on Graph Learning Benchmarks (GLB), WWW).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As Graph Neural Networks (GNNs) are being increasingly used for learning representations of graph-structured data in high-stakes applications, such as criminal justice [1], molecular chemistry [19] and biological networks [10, 28], it becomes critical to ensure that the relevant stakeholders can understand and trust their functionality. To this end, previous work developed several methods to explain predictions made by GNNs [4, 8, 13, 16–18, 20, 23, 25].

With the increase in newly proposed GNN explanation methods, it is critical to ensure their reliability. However, explainability in graph machine learning is still a nascent area and lacks both standardized evaluation strategies as well as reliable benchmarks to evaluate, test, and compare GNN explanations [2]. As a result, current approaches tend to base their analysis on specific real-world [19] and synthetic [7] datasets with limited ground-truth explanations. Further, GNN explainability research suffers from the following: i) evaluation strategies are methodologically weak as they can be solved using trivial baselines (e.g., random nodes or edges as explanations) [2]; ii) evaluation strategies do not provide a standard toolkit for benchmarking different kinds of explanation methods. While previous studies developed specific benchmark datasets [7, 19], relying on those benchmarks and associated ground-truth explanations is insufficient as they are not indicative of diverse real-world applications [2]. This gets further complicated by mismatches between GNN explanations as they are used in a real-world application versus in a benchmark.

To address the above challenges, we introduce G-XAI Bench, an explainability toolkit that provides the research community with a comprehensive and diverse resource to systematically access, evaluate, and compare GNN explanations across the entire range of GNN explainers and underlying GNN predictors. G-XAI Bench provides a set of versatile data loaders, data processing functions, visualizers, real-world graph datasets with ground-truth explanations, and evaluation metrics to reliably benchmark GNN explanations.

Relationship to existing graph benchmarks. Prior benchmarks in graph machine learning literature, such as Open Graph Benchmark (OGB) [11], Graph Robustness Benchmark (GRB) [27], GNNMark [5], GraphGT [6], MalNet [9], Therapeutics Data Commons [12], and EFO-1-QA [24], focus on providing resources to compare and evaluate GNN predictors, quantify stability/robustness of GNN predictors, scalability to very large graphs, etc. As these benchmarks already provide great support for the development and benchmarking of GNN predictors, G-XAI Bench aims to support the study of GNN explainers instead of the underlying GNN predictors. To this end, prior research on evaluating GNN explanations mainly leveraged ground-truth explanations associated with specific datasets [19]. As such, it cannot be used for benchmarking GNN explainers because of numerous reasons, including the pitfalls outlined by Faber et al. [7]. In contrast, G-XAI Bench provides a broader ecosystem for benchmarking state-of-the-art GNN explainers on diverse datasets and performance metrics.

2 LIBRARY OVERVIEW

G-XAI Bench is a general-purpose library that provides a comprehensive list of functions to systematically evaluate the quality of GNN explanations. For a given GNN model trained on a graph dataset, the library allows the use of any state-of-the-art GNN explanation method to generate explanations on the model's prediction.

*Both authors contributed equally to this research.

Workshop on Graph Learning Benchmarks (GLB), WWW, April 25–29, 2022, Virtual

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

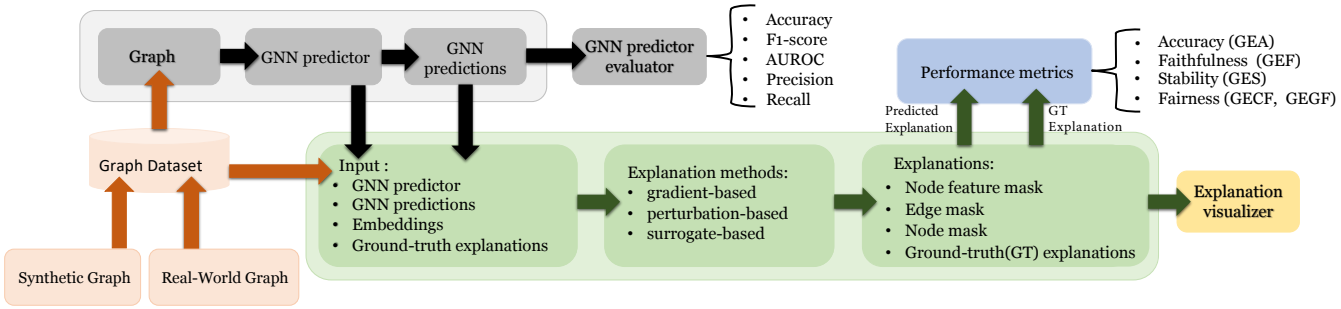

Figure 1: Overview of G-XAI Bench: G-XAI Bench provides i) dataloader classes for GraphXAI-ready synthetic and real-world datasets with ground-truth explanations for evaluating GNN Explainers; ii) implementation of explanation methods compatible with deep learning frameworks, such as *PyTorch* and *PyTorch Geometric* libraries; iii) visualization functions for GNN explainers; iv) utility functions to support new GNN explainers; and v) a diverse set of performance metrics to evaluate the reliability of explanations generated by GNN explainers.

To this end, G-XAI Bench provides the complete evaluation framework with: i) a dataset generator that can handle both real-world and synthetic datasets with/without ground-truth explanations, ii) GNN explanation method(s) that takes a prediction from the underlying GNN model and generates an explanation for it, and iii) quantifying the reliability of the output explanations using performance metrics.

G-XAI Bench library is thoroughly documented and includes test scripts for distinct use-cases and graph machine learning tasks, including comparing the reliability of state-of-the-art GNN explainers using performance metrics like faithfulness, stability, and fairness, and visualizing output explanations on a node-, edge-, or graph-level. Further, G-XAI Bench provides graph and explanation functions compatible with deep learning frameworks, such as *PyTorch* and *PyTorch Geometric* libraries. Our explainability library provides a pipeline where new datasets (both real-world and synthetic), explanation methods, and performance metrics can be easily integrated. Finally, the first version of our graph explainability library focuses on the node- and graph-classification downstream tasks, but in the planned future releases, we are working towards extending the applicability of the library to other graph downstream tasks like link-prediction.

3 LIBRARY DESIGN

We design G-XAI Bench as an open-source library that substitutes existing scattered and complex evaluation strategies with an ecosystem of Graph XAI-ready datasets, models, evaluation metrics, and visualization scripts with minimal dependency on external packages and easy-to-use classes with minimal implementation efforts (Figure 1). Below we describe diverse functionality useful for systematic access and evaluation of GNN explainers.

Dataset class. The Dataset class in G-XAI Bench comprises of the: i) NodeDataset class, and ii) GraphDataset classes. These classes provide several utility function for easier and faster incorporation of new datasets in the evaluation pipeline of GNN explainers. The dataset classes in G-XAI Bench have been developed to output Graph XAI-ready datasets, i.e., we provide ground-truth explanations along with every node or graph output by the dataloader. Every synthetic and real-world datasets inherit this Dataset classes, e.g., for the MUTAG dataset, we have class MUTAG(GraphDataset).

Explanation class. The library is centered on the Explanation class capable of storing multiple types of explanations produced by GNN explainers. G-XAI Bench provides a _BaseExplainer class that is a parent class to all explanation methods in our current release. In particular, the _BaseExplainer class contains several functions, such as _get_embedding(), _set_masks() for setting the output explanation, _predict() for getting model predictions, and get_explanation_node() for storing the output explanation using the underlying model, helpful for evaluating an output explanation.

Visualization. G-XAI Bench provides diverse functions that supports the visualization of explanation from all state-of-the-art GNN explainers. In particular, for a given explanation, a user can leverage the visualization function to qualitatively compare both node- and graph-level explanations. In addition, all function implementations are parameterized and user-friendly, e.g., researchers and practitioners can change the color and weight interpretation of an output explanation. In Figure 2, we show the output explanation from four different GNN explainers as produced by our visualization function. For simplicity, we include the visualization function inside the Explanation class described above.

Evaluation class. In contrast to existing benchmarks, the proposed G-XAI Bench provides an extensive list of performance metrics pertaining to key desiderata of GNN explainers as described in Agarwal et al. [2], i.e., accuracy, faithfulness, stability, and fairness. In particular, all evaluation metrics leverage predicted explanations, ground-truth explanations, and other user-controlled parameters like top-k features. G-XAI Bench package all these metrics and utility functions inside the Metrics class.

Figure 3 shows a code snippet for evaluating the correctness of output explanations for a given GNN prediction in G-XAI Bench.

4 DATASETS

G-XAI Bench incorporates synthetic and real-world graph datasets with ground-truth explanations to benchmark the quality of any GNN explainer. In addition, our well-documented dataset class (detailed in Section 3) allows researchers and practitioners to integrate

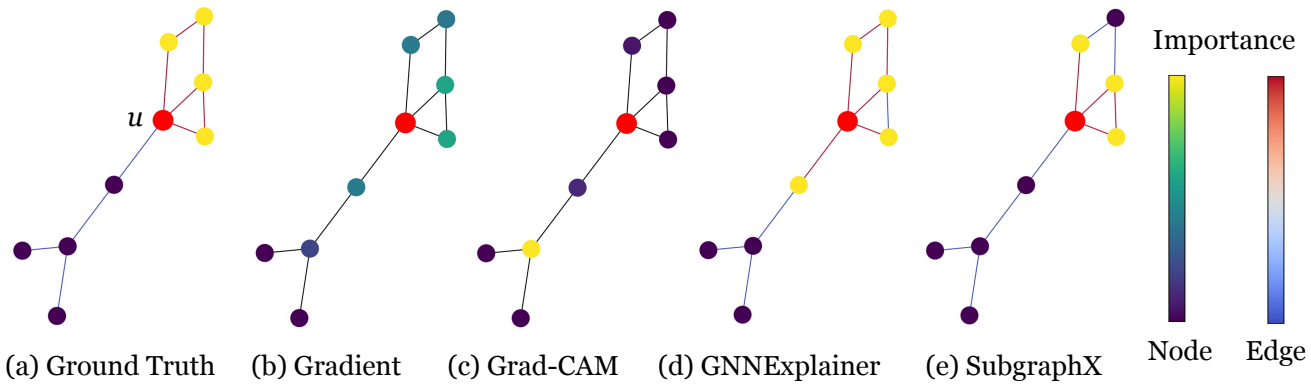

Figure 2: Visualization of four different explainers from the G-XAI Bench library on the BA-Shapes dataset. The visualization is for explaining the prediction of node u. We show the $L+1$-hop around node u, where L is the number of layers of the GNN model predicting on the dataset. Two colorbars indicate the intensity of attribution scores for the node and edge explanations. Note that edge importance is not defined for every method, so edges are simply set to black to indicate that the method does not provide edge scores. The visualization tools in G-XAI Bench allow users to compare the explanations of different GNN explainers, such as gradient-based methods (Gradient and Grad-CAM) and perturbation-based methods (GNNExplainer and SubgraphX).

```
import graphxai
from graphxai.explainers import GradCAM
from graphxai.metrics import graph_exp_faith

dataset = graphxai.get_dataset('BAHouses')
data = dataset.get_graph()

# Train a model ...

gcam = GradCAM(model, loss_function)
exp = gcam.get_explanation_node(data, node_idx)

feat_faith, node_faith, edge_faith = \
    graph_exp_faith(exp, data, model)

exp.visualize_node(show = True)
```

Figure 3: An example of explaining a prediction in the G-XAI Bench pipeline. With just a few lines of code, one can calculate an explanation for a node or graph, calculate metrics based on that explanation, and visualize the explanation.

new datasets into G-XAI Bench. Below we describe few existing datasets from our library.

4.1 Synthetic Graphs

In the initial release of G-XAI Bench, we follow Ying et al. [25] and incorporate BA-Shapes node classification dataset. We start with a base Barabasi-Albert (BA) [3] graph using N nodes (e.g., $N = 300$) and a set of K (e.g., $K = 80$) five-node "house"-structured motifs randomly attached to nodes of the base graph. The final graph is perturbed by adding random edges. The nodes in the output graph are categorized into four classes corresponding to nodes at the top, middle, bottom of houses, and nodes that do not belong to a house.

4.2 Real-world Graphs

In addition to synthetic datasets, G-XAI Bench library includes real-world graph datasets with ground-truth explanations. Here, we incorporate popular benchmark datasets from molecular chemistry and biology employed in previous works [15, 19]. We integrate these datasets as they contain a specific pattern (e.g., a certain chemical group in a molecule) which represents ground-truth explanations. Below, we discuss the details of each of the real-world datasets that we employ and their ground-truth explanations:

MUTAG [15] dataset contains 188 molecular graphs labeled into two classes according to their mutagenic properties, i.e., effect on the Gram-negative bacterium *S. typhimuriuma*. During training, a GNN can chose either NH2 or NO2 chemical groups to learn to predict *mutagenicity*. Therefore, any combination of these molecules can be used as a ground-truth explanation for evaluating the quality of an output explanation.

Alkane-Carbonyl [19] dataset contains 1125 molecular graphs labeled into two classes where a positive sample indicates a molecule that contains an unbranched alkane and a carbonyl (C=O) functional group. The ground-truth explanations consist of any combinations of both alkane and carbonyl functional groups within a given molecule.

Recidivism [14] dataset has 18,876 nodes representing defendants who got released on bail at the U.S. state courts during 1990-2009. Defendants are connected based on the similarity of past criminal records and demographics. The goal is to classify defendants into bail vs. no bail considering race information as the protected attribute.

5 EXPERIMENTS AND RESULTS

To demonstrate the capabilities and utility of G-XAI Bench, we systematically evaluate and compare the quality of 9 state-of-the-art

Dataset	Method	GEA (↑)	GEF (↓)
BA-Shapes	Random	0.170±0.028	0.185±0.015
	Grad	0.303±0.044	0.163±0.016
	GradCAM	0.603±0.043	0.170±0.015
	GuidedBP	**0.825**±0.034	0.168±0.017
	Integrated Grad (IG)	0.527±0.050	0.182±0.013
	GNNExplainer	0.538±0.025	0.177±0.013
	PGMExplainer	0.480±0.032	0.172±0.012
	PGExplainer	0.319±0.032	**0.156**±0.014
	SubgraphX	0.223±0.020	0.165±0.010
MUTAG	Random	0.081±0.043	0.383±0.077
	Grad	0.001±0.001	0.732±0.070
	GradCAM	0.418±0.077	0.385±0.078
	GuidedBP	0.003±0.003	0.740±0.069
	Integrated Grad (IG)	**0.423**±0.052	**0.118**±0.051
	GNNExplainer	0.149±0.017	0.403±0.079
	PGMExplainer	0.010±0.009	0.403±0.079
	PGExplainer	0.222±0.016	0.403±0.079
	SubgraphX	0.027±0.015	0.515±0.077
Alkane-Carbonyl	Random	0.034±0.006	0.295±0.030
	Grad	0.011±0.003	0.327±0.031
	GradCAM	0.005±0.003	0.536±0.033
	GuidedBP	0.028±0.003	0.572±0.033
	Integrated Grad (IG)	0.027±0.004	**0.001**±0.001
	GNNExplainer	0.048±0.006	0.326±0.031
	PGMExplainer	0.016±0.005	0.234±0.028
	PGExplainer	**0.067**±0.007	0.183±0.026
	SubgraphX	0.020±0.005	0.419±0.032

Table 1: Benchmarking state-of-the-art GNN explainers for synthetic and molecular datasets with ground-truth explanations. Arrows (↑/↓) indicate the direction of better performance. Note that stability and fairness performance metrics do not apply here because generating plausible perturbations for synthetic and molecular graphs is non-trivial and they have no protected features.

GNN explainers on both synthetic and real-world graphs. In particular, we benchmark GNN explainers using different performance metrics to quantify the quality of explanations.

GNN Explainers. Our current G-XAI Bench release incorporates 9 GNN explanation methods, including Grad [21], GradCAM [18], GuidedBP [4], Integrated Gradients [22], GNNExplainer [25], PGExplainer [17], SubgraphX [26]; PGMExplainer [23]. Finally, we follow Agarwal et al. [2] and consider random explanations as a controlled baseline in our experiments.

Implementation details. The G-XAI Bench provides flexibility to incorporate any state-of-the-art GNN predictors. For brevity, we use a 2-layer GIN model as the GNN predictor for our experiments and show how can we utilize our library to evaluate GNN explanations. The GNN model comprises of two GIN convolution layers with ReLU non-linear activation function and a fully-connected linear classification layer with Softmax activations. The hidden dimensionality of the layers is set to 16. We use an Adam optimizer with a learning rate of 1×10^{-2}, weight decay of 1×10^{-5}, and the number of epochs to 1000 for training our GIN models. Following prior works [2, 13], we select top-k (k = 25%) important nodes, node features, or edges for generating explanations for all graph explainability methods and all other hyperparameters were set following the authors' guidelines. All codes and datasets are available here.

Method	GEF (↓)	GES (↓)	GECF (↓)	GEGF (↓)
Random	0.322±0.003	0.794±0.004	0.751±0.002	0.156±0.004
Grad	**0.305**±0.004	0.643±0.003	0.050±0.006	0.583±0.003
GradCAM	0.538±0.004	**0.085**±0.001	**0.005**±0.000	**0.000**±0.003
GuidedBP	0.414±0.003	0.167±0.002	0.008±0.001	0.167±0.008
IG	0.636±0.004	0.161±0.002	0.032±0.004	0.000±0.004
GNNExplainer	0.404±0.004	0.716±0.002	0.604±0.003	0.200±0.001

Table 2: Evaluation of GNN explainers on Recidivism graph dataset based on node explanation masks. Arrows (↑/↓) indicate the direction of better performance. GradCAM method, on average, produces most reliable explanations when evaluated across all four performance metrics.

Performance metrics. G-XAI Bench follows standard practices for measuring accuracy and Agarwal et al. [2], and considers four broad category of performance metrics: i) *Graph Explanation Accuracy (GEA)* measures the correctness of an explanation using the ground-truth explanation of the input graph dataset, ii) *Graph Explanation Faithfulness (GEF)* quantifies the degree of faithfulness of an output explanation to an underlying GNN predictor, iii) *Graph Explanation Stability (GES)* measures whether the output explanations of a given graph and its perturbed counterpart (generated by making infinitesimally small perturbations to the node feature vector and associated edges) are similar, and iv) *Graph Explanation Fairness (GEC(G)F)* which reports counterfactual fairness and group fairness mismatch of a generated explanation.

Results. We evaluate the performance of GNN explanation methods on synthetic and real-world datasets. Across all three datasets in Table 1, we find that gradient-based explanation methods like GuidedBP and Integrated Gradient generate accurate and faithful explanations. In particular, GuidedBP obtains a high Graph Explanation Accuracy score for BA-Shapes dataset and outperform all other GNN explainers by 52.08%. Whereas, on average across datasets, Integrated Gradient generates the least unfaithful explanations as compared to other GNN explainers. In addition, we show the behavior of GNN explanation methods on Recidivism which has *Race* as a protected attribute. In Table 2, we find that, on average across all performance metrics, GradCAM generates the most reliable explanation as compared to other explanation methods.

6 CONCLUSION

We introduce G-XAI Bench, an open-science resource to access and evaluate the quality of GNN explanations output by state-of-the-art GNN explainers. G-XAI Bench provides a comprehensive framework that comprises data loaders, data processing functions, visualizers, real-world graph datasets with ground-truth explanations, and evaluation metrics to reliably benchmark GNN explanations across the entire range of GNN explainers. G-XAI Bench provides a simple and transparent framework for evaluating explainability methods and increasing reproducibility. We believe that G-XAI Bench can help the Graph XAI community in both developing and evaluating new GNN explainers.

References.

[1] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *UAI*, 2021.

[2] Chirag Agarwal et al. Probing GNN explainers: A rigorous theoretical and empirical analysis of GNN explanation methods. In *AISTATS*, 2022.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 2002.

[4] Federico Baldassarre and Hossein Azizpour. Explainability techniques for graph convolutional networks. In *ICML Workshop on LRGR*, 2019.

[5] Trinayan Baruah et al. GNNMark: a benchmark suite to characterize graph neural network training on gpus. In *ISPASS*, 2021.

[6] Yuanqi Du et al. GraphGT: machine learning datasets for graph generation and transformation. In *NeurIPS Datasets and Benchmarks*, 2021.

[7] Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. When comparing to ground truth is wrong: On evaluating GNN explanation methods. In *KDD*, 2021.

[8] Lukas Faber et al. Contrastive graph neural network explanation. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

[9] Scott Freitas et al. A large-scale database for graph representation learning. In *NeurIPS Datasets and Benchmarks*, 2021.

[10] Deisy Morselli Gysi et al. Network medicine framework for identifying drug repurposing opportunities for COVID-19. *arXiv*, 2020.

[11] Weihua Hu et al. Open Graph Benchmark: datasets for machine learning on graphs. In *NeurIPS*, 2020.

[12] Kexin Huang et al. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks*, 2021.

[13] Qiang Huang et al. GraphLIME: local interpretable model explanations for graph neural networks. In *arXiv:2001.06216*, 2020.

[14] Kareem L. Jordan and Tina L. Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, 2015. doi: 10.1080/15377938.2014.984045. URL https://doi.org/10.1080/15377938.2014.984045.

[15] Jeroen Kazius et al. Derivation and validation of toxicophores for mutagenicity prediction. In *Journal of Medicinal Chemistry*, 2005.

[16] Ana Lucic et al. CF-GNNExplainer: counterfactual explanations for graph neural networks. *arXiv:2102.03322*, 2021.

[17] Dongsheng Luo et al. Parameterized explainer for graph neural network. In *NeurIPS*, 2020.

[18] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *CVPR*, 2019.

[19] Sanchez-Lengeling et al. Evaluating attribution for graph neural networks. *NeurIPS*, 2020.

[20] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for nlp with differentiable edge masking. In *ICLR*, 2021.

[21] Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[23] Minh N Vu and My T Thai. PGM-Eexplainer: probabilistic graphical model explanations for graph neural networks. In *NeurIPS*, 2020.

[24] Zihao Wang, Hang Yin, and Yangqiu Song. Benchmarking the combinatorial generalizability of complex query answering on knowledge graphs. 2021.

[25] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: generating explanations for graph neural networks. In *NeurIPS*, 2019.

[26] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *ICML*, 2021.

[27] Qinkai Zheng et al. Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning. In *NeurIPS Datasets and Benchmarks*, 2021.

[28] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. In *Bioinformatics*, 2018.