



Empower Post-hoc Graph Explanations with Information Bottleneck: A Pre-training and Fine-tuning Perspective

Jihong Wang
wang1946456505@stu.xjtu.edu.cn
Xi'an Jiaotong University

Minnan Luo*
minnluo@xjtu.edu.cn
Xi'an Jiaotong University

Jundong Li
jundong@virginia.edu
University of Virginia

Yun Lin
lin_yun@sjtu.edu.cn
Shanghai Jiao Tong University

Yushun Dong
yd6eb@virginia.edu
University of Virginia

Jin Song Dong
dcsdjs@nus.edu.sg
National University of Singapore

Qinghua Zheng
qhzheng@mail.xjtu.edu.cn
Xi'an Jiaotong University

ABSTRACT

Researchers recently investigated to explain Graph Neural Networks (GNNs) on the access to a task-specific GNN, which may hinder their wide applications in practice. Specifically, task-specific explanation methods are incapable of explaining pretrained GNNs whose downstream tasks are usually inaccessible, not to mention giving explanations for the transferable knowledge in pretrained GNNs. Additionally, task-specific methods only consider target models' output in the label space, which are coarse-grained and insufficient to reflect the model's internal logic. To address these limitations, we consider a two-stage explanation strategy, *i.e.*, explainers are first pretrained in a task-agnostic fashion in the representation space and then further fine-tuned in the task-specific label space and representation space jointly if downstream tasks are accessible. The two-stage explanation strategy endows post-hoc graph explanations with the applicability to pretrained GNNs where downstream tasks are inaccessible and the capacity to explain the transferable knowledge in the pretrained GNNs. Moreover, as the two-stage explanation strategy explains the GNNs in the representation space, the fine-grained information in the representation space also empowers the explanations. Furthermore, to achieve a trade-off between the fidelity and intelligibility of explanations, we propose an explanation framework based on the Information Bottleneck principle, named *Explainable Graph Information Bottleneck* (EGIB). EGIB subsumes the task-specific explanation and task-agnostic explanation into a unified framework. To optimize EGIB objective, we derive a tractable bound and adopt a simple yet effective explanation generation architecture. Based on the unified

framework, we further theoretically prove that task-agnostic explanation is a relaxed sufficient condition of task-specific explanation, which indicates the transferability of task-agnostic explanations. Extensive experimental results demonstrate the effectiveness of our proposed explanation method.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Graph Neural Networks; Explanation; Information Bottleneck

ACM Reference Format:

Jihong Wang, Minnan Luo, Jundong Li, Yun Lin, Yushun Dong, Jin Song Dong, and Qinghua Zheng. 2023. Empower Post-hoc Graph Explanations with Information Bottleneck: A Pre-training and Fine-tuning Perspective. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, August 6–10, 2023, Long Beach, CA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3580305.3599330>

1 INTRODUCTION

Graph Neural Networks (GNNs) have emerged as a promising learning paradigm and demonstrated superior learning performance on different graph learning tasks, such as node classification [9, 15, 17, 30], graph classification [41, 43], and link prediction [5, 47]. Despite their strengths, GNNs are usually treated as black box models and thus cannot provide human-intelligible explanations [18, 42]. Such opaqueness impedes their broad adoption in many decision-critical applications pertaining to fairness, privacy, and safety [7]. To better understand the working mechanisms of GNNs, researchers started to investigate the GNN explanation problem recently: *what knowledge does the GNN model extract to make a specific decision?*

To answer the above question, many post-hoc explanation methods have been proposed. According to the taxonomy provided in a recent survey [45], these methods can be subsumed into four technical route lines: the gradient-based [2, 22], perturbation-based [18, 24, 27, 33, 42], decomposition-based [2, 23], and surrogate model-based [11, 32] methods. However, almost all the existing explanation methods are task-specific, *i.e.*, tailored for explaining a specific learning task. Expressly, the explanation methods usually assume

*Corresponding author: Minnan Luo, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

KDD '23, August 6–10, 2023, Long Beach, CA, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0103-0/23/08...\$15.00

<https://doi.org/10.1145/3580305.3599330>

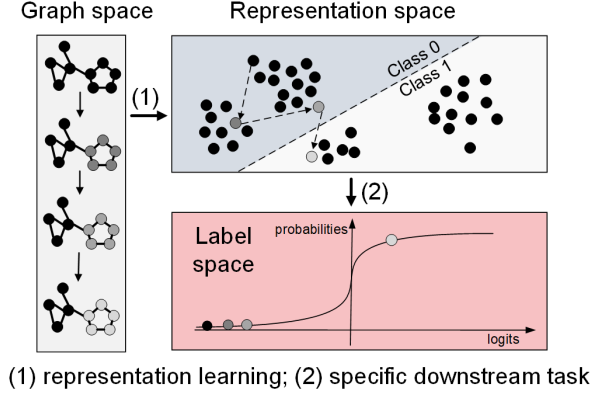


Figure 1: An intuitive view of the insufficiency of task-specific explanations. Graphs are first embedded into the representation space with different embedding logic, and then decisions are made in the label space. One may fail to capture the embedding logic in the label space since the output probabilities stay similar for different graphs.

we have access to a target GNN model which is well-tuned on a specific learning task. They attribute the target model’s decision on the specific task to the input graph space and find the most crucial components (e.g., nodes, edges, or subgraphs) that decide the GNN model’s output. However, the dependency on task-specific target models may hinder the explanation methods’ application in practice because of the following reasons. First, task-specific explanation methods are incapable of explaining GNNs that are pretrained in an unsupervised way. Recently, unsupervised pretrained GNNs have attracted tremendous research interest because of their freedom of manual labels and transferability to multiple downstream tasks [37]. Nonetheless, task-specific explanation methods cannot explain the transferable knowledge in these pretrained GNNs since downstream tasks are usually inaccessible. Moreover, even if some downstream tasks are given, task-specific explanation methods can only explain task-specific knowledge, while knowledge transferable to other tasks is not explained. Second, task-specific explanation methods only consider target models’ output in the label space, which are coarse-grained and insufficient to reflect the model’s internal logic. GNNs can be regarded as a non-injective projection from the graph space to the label space. Thus disparate graphs may lead to similar prediction results in the label space. For example, in Fig. 1, we perturb the graph gradually with decreasing edge weights (deep color indicates larger edge weight). Although the model embeds the graphs into distinctive representations, the behavior of graphs in the label space stays similar except for the last one. That is, in the label space, we cannot capture the logic of how graphs are embedded. Consequently, the coarse-grained information in the label space is insufficient to supervise the explanations.

To overcome the limitations of task-specific explanation methods, we propose to investigate the task-agnostic explanations for GNNs with the transferable and fine-grained information in the representation space. Specifically, we consider a two-stage explanation strategy which is applicable whether downstream tasks are

accessible. In the first stage where downstream tasks are inaccessible, we pretrain a task-agnostic explainer in the representation space yielded by the target pretrained GNNs. In the second stage where downstream tasks are accessible, we fine-tune the pretrained explainer to be task-specific jointly in the label space and representation space. As mentioned above, the representation space can provide fine-grained supervision compared with the coarse-grained label space, thus benefiting the fidelity of explanations. Note that the second stage is not compulsory. Experimental results demonstrate that we can achieve superior performance even without any fine-tuning.

There are mainly two challenges for explanation in the two-stage strategy. First, there is no theoretical support for the transferability of task-agnostic explanations. The transferability of explanations indicates the capacity they can explain the transferable knowledge in the pretrained GNNs. Transferable explanations should incorporate graph information that is crucial for multiple downstream tasks. As far as we know, there is no work theoretically analyzing the transferability of task-agnostic explanations and thus it’s still unexplored whether we can explain the transferable knowledge of pretrained GNNs in the representation space. Second, it is unclear how to supervise the explanations in the representation space. Typical task-specific explanation methods [18, 33, 42] usually supervise the explanations in the label space by employing Mutual Information (MI) as a relevance metric to measure the performance of explanations. In this paper, we extend this idea to identify the crucial subgraphs in the representation space. Nonetheless, different from the label space, the representation space is usually high-dimensional and semantically entangled [29], which exacerbates the optimization difficulties of MI since MI requires the integral on the corresponding space.

To analyze the transferability of task-agnostic explanation, we provide a unified view of task-agnostic explanation and task-specific explanation based on the Information Bottleneck (IB) principle [1, 28]. Specifically, graph explanations are usually defined as the compact subgraphs that are most crucial for the GNNs’ output, which shares a similar ideology with IB on the tradeoff between informativeness and compactness. In this paper, we extend the IB to graph explanations and propose a novel explanation framework named *Explainable Graph Information Bottleneck* (EGIB). Our EGIB subsumes the task-agnostic and task-specific settings into a unified framework. Based on the unified framework, we theoretically prove that our task-agnostic explanation can transfer to downstream tasks and that task-agnostic explanation is a relaxed sufficient condition of task-specific explanation, which exhibits the validity and transferability of task-agnostic pretraining in the first stage. Furthermore, we derive a tractable bound and adopt a simple yet effective explanation generation architecture to optimize the EGIB objective which is capable of supervising the explanations in the representation space. Experimental results show that our proposed EGIB achieves superior performance even without any task-specific fine-tuning. Contributions of our paper can be summarized as follows:

- **A Two-stage Explanation Strategy.** We consider a two-stage explanation strategy following a pretraining and fine-tuning pipeline, which is applicable whether downstream tasks are accessible.

- **An Unified Framework.** We propose a unified framework based on the IB principle that subsumes the task-agnostic and task-specific settings with tractable bound for optimization. Based on the unified framework, we theoretically analyze the relationship between task-agnostic explanation and task-specific explanation, which demonstrates the validity and transferability of our method.
- **Experimental Evaluations.** We conduct extensive experiments on real-world datasets with multiple tasks and observe that our proposed EGIB outperforms existing methods on effectiveness and can be transferable to downstream tasks even without fine-tuning.

2 NOTATIONS AND PRELIMINARIES

In this section, we first elaborate on the notions used in this paper. Then we formally introduce our explanation pipeline.

2.1 Notations

In most cases, we denote random variables with upper-case letters (e.g., G and Z), and represent its support set by calligraphic letters (e.g., \mathcal{G} and \mathcal{Z}). Lower-case letters with subscript (e.g., g_i and z_i) refer to the instances of random variables (e.g., G and Z) correspondingly. The MI between random variables X and Z is formulated as $I(X; Z) = \int_{\mathcal{X}} \int_{\mathcal{Z}} p(x, z) \log \frac{p(x, z)}{p(x)p(z)} dx dz$. In this paper, we suppose that GNNs follow an embedding-prediction architecture. Graphs are first projected to node-level representations or graph-level representations and then followed by a downstream model, e.g., MLPs. Since node-level representation learning can be seen as a particular case of graph representation learning by considering the corresponding k -hop neighbor subgraphs of each node as a single graph instance, we unify the embedding procedure as $Z = f_e(G)$ where $G \in \mathcal{G}$ indicates graphs in the graph-level embedding procedure or k -hop neighbor subgraphs in the node-level embedding procedure. And $Z \in \mathcal{Z}$ denotes graph/node representations, correspondingly. Moreover, we denote downstream models as $\hat{Y} = f_d(Z)$ where $\hat{Y} \in \mathcal{Y}$ denotes the prediction result on specific downstream tasks.

2.2 Task-specific Explanation and Task-agnostic Explanation

Typical task-specific explanation methods consider the decision procedure of GNNs as end-to-end, i.e., they consider a composition of f_d and f_e : $f_t = f_d \circ f_e$. Then they optimize the explainer to identify the subgraphs $S \in \mathcal{S}$ that are the most crucial in the prediction of $\hat{Y} = f_t(G)$ in specific tasks where \mathcal{S} is the set that contains all possible subgraphs of G . However, the downstream models f_d may be inaccessible for pretrained GNNs. Moreover, they can only explain how the embedding model f_e behave on the specific task corresponding to f_d while the knowledge in f_e transferable to other tasks is not explained. And since they only consider the output \hat{Y} of the end-to-end model f_t in the label space, the fine-grained information in the representation space yielded by f_e may be filtered out by f_d and thus cannot be captured by the explanations.

Different from task-specific explanation, task-agnostic explanation aims to explain the embedding procedure $Z = f_e(G)$ without

any knowledge of downstream tasks. To overcome the above problems of task-specific explanations, we consider a two-stage explanation strategy where a task-agnostic explainer is pretrained in the representation space first and then fine-tuned in the label space and the representation space jointly. The two-stage explanation strategy can empower explanations with better transferability and fidelity while liberating them from the assumption on access to specific downstream tasks.

3 METHODOLOGY

In this section, we introduce our proposed Explainable Graph Information Bottleneck (EGIB) which unifies the task-agnostic and task-specific settings. Then we derive tractable bounds for the optimization of EGIB and introduce our explanation generation architecture. Finally, we analyze the transferability of task-agnostic explanations theoretically.

3.1 Explainable Graph Information Bottleneck

There are mainly two aspects for the assessment of graph explanations, i.e., fidelity and intelligibility [13]. Fidelity measures the importance of explanations in the decision procedure of the target model. Intelligibility requires the explanations to be as compact as possible. To generate explanatory subgraphs with both satisfactory fidelity and intelligibility, we propose a method based on the IB principle from a unified view of task-agnostic explanations and task-specific explanations. Specifically, our EGIB is mainly built on two crucial notions: sufficient subgraphs and ϵ -explanatory subgraphs. The sufficient subgraphs guarantee the fidelity of explanations while ϵ -explanatory subgraphs further guarantee the intelligibility of sufficient subgraphs by discarding superfluous information. Formally,

DEFINITION 1. *Sufficient subgraphs: Given graph G , let \mathcal{S} be the set of its subgraphs, and T be its output of GNNs. A subgraph $S \in \mathcal{S}$ is called the **sufficient subgraphs** of T if and only if S is sufficient for T i.e., $I(T; G|S) = 0$.*

We call S as *task-agnostic sufficient subgraphs* when T refers to graph/node representations Z , and *task-specific sufficient subgraphs* when T refers to downstream prediction results \hat{Y} . Sufficient subgraphs ensure the fidelity of explanations as S extract all information related with T , i.e.,

$$I(T; G|S) = I(T; G) - I(T; S) = 0. \quad (1)$$

Please refer to the details of the above equation in the Appendix. As we focus on the post-hoc explanations, the mutual information $I(T; G)$ can be regarded as a constant, which is decided by the target GNN model. According to Eq. (1), we can achieve sufficient subgraphs with desirable fidelity by the following objective:

$$\arg \max_{S \in \mathcal{S}} I(T; S). \quad (2)$$

However, there are trivial solutions for the sufficient subgraphs (e.g., $S = G$) as it overlooks the intelligibility and thus may involve superfluous information. Inspired by the IB principle, we define the ϵ -explanatory subgraphs by discarding the superfluous information of sufficient subgraphs:

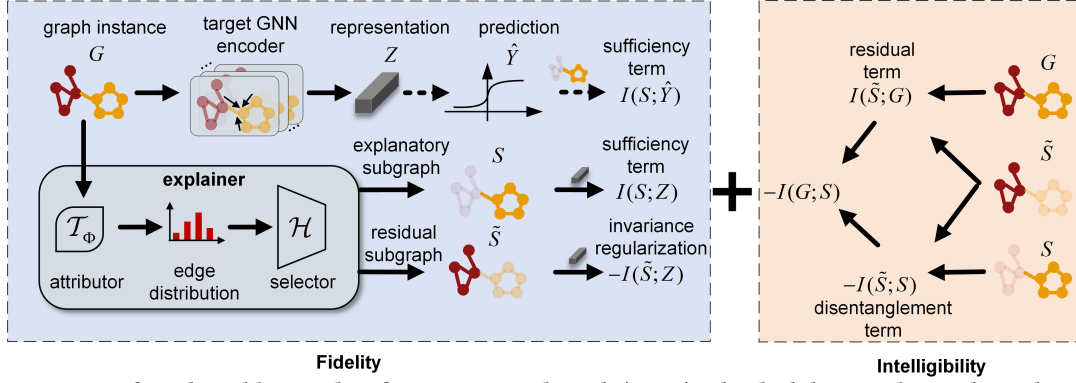


Figure 2: An overview of Explainable Graph Information Bottleneck (EGIB). The dash lines indicate the task-specific setting which is ignored in the task-agnostic pretraining. The objective can be roughly separated into two parts: the fidelity term and the intelligibility term. In the fidelity term, we employ an invariance regularization besides the sufficiency term. And the intelligibility term can be further decomposed to a residual term and a disentanglement term.

DEFINITION 2. ϵ -*explanatory subgraphs*: Given a sufficient subgraph S of GNNs' output T , S is called the ϵ -**explanatory subgraphs** of T if and only if $I(S; G|T) \leq \epsilon$ where ϵ constraints the compactness of S .

Analogously, we define *task-agnostic ϵ -explanatory subgraphs* and *task-specific ϵ -explanatory subgraphs* when T refers to representations and prediction results, separately. ϵ -explanatory subgraphs restrict the sufficient subgraphs to be compact by discarding information irrelevant with T . Similar to Eq. (1), it can be easily proved that:

$$I(S; G|T) = I(S; G) - I(S; T) \leq \epsilon. \quad (3)$$

In tandem with Eq. (2), we can optimize explanatory subgraphs by

$$\arg \max_{S \in \mathcal{S}} I(Z; S) \text{ s.t. } I(S; G) - I(S; T) \leq \epsilon. \quad (4)$$

To address the above constrained optimization problem, we adopt the Lagrangian relaxation algorithm and rewrite Eq. (4) as:

$$\arg \max_{S \in \mathcal{S}} I(S; T) - \alpha(I(S; G) - I(S; T)), \quad (5)$$

Letting $\beta = \frac{\alpha}{1+\alpha}$, we arrive at the IB-based optimization objective:

$$\arg \max_{S \in \mathcal{S}} I(T; S) - \beta I(S; G), \quad (6)$$

where $I(T; S)$ is named the sufficiency term since it requires the subgraphs to be sufficient to the given T . And $-I(S; G)$ is named the intelligibility term as it guarantees the intelligibility of subgraphs. The hyper-parameter β makes a trade-off between fidelity and intelligibility, just like the typical IB [1, 28] which balances the informativeness and compression. Typical explanation methods usually compress the subgraphs to be compact by l_1 norm regularization which restricts the size of subgraphs. However, compared with our intelligibility term, the l_1 norm regularization may lead to a biased assumption since explanations may vary in size [20]. Specifically, some subgraphs may miss crucial edges, while others may involve superfluous edges with the same size restriction. Instead, our intelligibility term is capable of discarding the superfluous information without any assumption on the size of explanations.

In Eq. (6), we achieve a unified framework that subsumes the task-specific explanations and task-agnostic explanations. When T

refers to the representations Z , we can pretrain an explainer in a task-agnostic fashion. In contrast, when T refers to the task-specific predictions \hat{Y} , we can train a task-specific explainer. In this paper, we adopt a two-stage explanation strategy based on our framework. In the first stage, we employ a learnable explainer to generate explanatory subgraphs, i.e., $S \sim q_\Phi(S|G)$ in a task-agnostic fashion:

$$\arg \max_{S \sim q_\Phi(S|G)} I(Z; S) - \beta I(S; G). \quad (7)$$

And in the second stage, the explainer is fine-tuned to generate task-specific explanatory subgraphs

$$\arg \max_{S \sim q_\Phi(S|G)} I(Z; S) + \gamma I(\hat{Y}; S) - \beta I(S; G). \quad (8)$$

Note that in the second stage, instead of only explaining downstream predictions \hat{Y} , representations are also involved as they can provide fine-grained information.

Invariance regularization. Previous self-explainable work [35] demonstrates that it is crucial to ensure the invariance of GNNs' predictions across different superfluous information. In this paper, we extend this idea to post-hoc explanation by restricting $I(\tilde{S}; Z|S) = 0$ where \tilde{S} is the complementary subgraphs of S and $G = (S, \tilde{S})$. We call \tilde{S} as the *residual subgraph* in this paper. $I(\tilde{S}; Z|S) = 0$ indicates that the representations are independent of other information except for the explanatory subgraph S . Since $I(\tilde{S}; Z|S) = I(\tilde{S}; Z) - I(\tilde{S}; Z; S) \leq I(\tilde{S}; Z)$, we adopt $I(\tilde{S}; Z)$ as an *invariance regularization* term to guarantee the invariance of Z across different superfluous information. The objective in two stages can be further rewritten as:

$$\arg \max_{S \sim q_\Phi(S|G)} I(Z; S) - \beta I(S; G) - \lambda I(\tilde{S}; Z), \quad (9)$$

$$\arg \max_{S \sim q_\Phi(S|G)} I(Z; S) + \gamma I(\hat{Y}; S) - \beta I(S; G) - \lambda I(\tilde{S}; Z). \quad (10)$$

3.2 Optimization

Optimization of mutual information is notoriously intractable as it involves the integration of high-dimensional data. There are copious works devoted to mutual information estimation and optimization [3, 21, 34]. In this paper, we adopt the well-known InfoNCE [8]

loss, which is widely used in self-supervised learning [26, 31]. Formally, $I(Z; S)$ in Eq. (9) and Eq. (10) can be estimated by InfoNCE as:

$$\mathcal{L}_{NCE}(Z; S; \Phi) = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(\langle z_i, f_e(s_i) \rangle)}{\sum_{j \neq i} \exp(\langle z_i, f_e(s_j) \rangle)} \right], \quad (11)$$

where N is the number of samples in a mini-batch, z_i and s_i denotes the samples of Z and S , respectively. Note that s_i is sampled from the learnable distribution $q_\Phi(S|G)$. $\langle z_i, f_e(s_i) \rangle$ denotes the inner product of two vectors. $I(\tilde{S}; Z)$ can be computed similarly to $I(Z; S)$.

The remaining problem is how to optimize the task-specific term $I(\hat{Y}; S)$ and the intelligibility term $I(S; G)$ in Eq. (9) and Eq. (10). The former term can be addressed following previous works, e.g., PGExplainer [18]:

$$\mathcal{L}_{ts}(\hat{Y}; S; \Phi) = -\sum_{i=1}^N \sum_{c=1}^C P(f_t(g_i) = c) \log P(f_t(s_i) = c), \quad (12)$$

where $f_t = f_d \circ f_e$ denotes the composition of GNN-based encoder f_e and downstream model f_d . C is the total number of labels, g_i is the i -th sample of graph G .

Directly minimizing the intelligibility term $I(S; G)$ usually leads to a min-max optimization objective and thus suffers concerns about the instability and convergence during training [44]. Instead, we bypass this problem by decomposing the intelligibility term $I(S; G)$ to a residual term and a disentanglement term:

$$I(S; G) = -I(\tilde{S}; G) + I(\tilde{S}; S) + H(G). \quad (13)$$

From the above decomposition, we find that we can guarantee the explanatory subgraphs' intelligibility by maximizing the residual term $I(\tilde{S}; G)$ and minimizing the disentanglement term $I(\tilde{S}; S)$ simultaneously. $I(\tilde{S}; G)$ is called the residual term because it measures the mutual information between G and the residual subgraph \tilde{S} , i.e., the complement of the explanatory subgraph S . We maximize $I(\tilde{S}; G)$ by employing an additional learnable GNN encoder f_θ to learn representations of G^1 and then minimizing the InfoNCE loss:

$$\mathcal{L}_{NCE}(\tilde{S}; G; \theta; \Phi) = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(\langle f_\theta(g_i), f_\theta(\tilde{s}_i) \rangle)}{\sum_{j \neq i} \exp(\langle f_\theta(g_j), f_\theta(\tilde{s}_i) \rangle)} \right].$$

$I(\tilde{S}; S)$ in Eq. (22) is named disentanglement term since it measures the relevance between two complementary subgraphs, which indicates the explanatory subgraphs should be disentangled with the residual subgraphs. We minimize $I(\tilde{S}; S)$ by maximizing the InfoNCE loss:

$$\mathcal{L}_{NCE}(\tilde{S}; S; \Phi) = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(\langle f_e(s_i), f_e(\tilde{s}_i) \rangle)}{\sum_{j \neq i} \exp(\langle f_e(s_j), f_e(\tilde{s}_i) \rangle)} \right].$$

The overall loss function for the first-stage can be summarized as:

$$\begin{aligned} \mathcal{L}_1(\Phi; \theta) = & \mathcal{L}_{NCE}(Z; S; \Phi) + \beta * (\mathcal{L}_{NCE}(\tilde{S}; G; \theta; \Phi) \\ & - \mathcal{L}_{NCE}(\tilde{S}; S; \Phi)) - \lambda * \mathcal{L}_{NCE}(Z; \tilde{S}; \Phi). \end{aligned} \quad (14)$$

In the second-stage, we fine-tune the explainer by simply employing an additional term $\mathcal{L}_{ts}(\hat{Y}; S; \Phi)$ based on the first-stage:

$$\mathcal{L}_2(\Phi; \theta) = \mathcal{L}_1(\Phi; \theta) + \gamma \mathcal{L}_{ts}(\hat{Y}; S; \Phi). \quad (15)$$

¹Note that different from optimization of $I(Z; S)$ in Eq. (11), we do not directly employ the target GNN model f_e but adopt a learnable GNN encoder f_θ . This is because that $I(S; G)$ will degrade to $I(S; Z)$ if we embed graphs with f_e .

3.3 Explanation Generation Architecture

In this subsection, we elaborate on how we design the explainer $q_\Phi(S|G)$ and how to optimize the explainer with the loss functions in Eq. (14) and Eq. (15). In practice, we decompose the explainer $q_\Phi(S|G)$ into a learnable attributor \mathcal{T}_Φ and a selector \mathcal{H} . The attributor with parameters Φ estimates the distribution of every edge, i.e., $q_\Phi(e_{ij}|G)$ where e_{ij} is a binary random variable indicating the mask for edge (i, j) . Specifically, if $e_{ij} = 1$, we retain the edge (i, j) to construct explanatory subgraphs, and otherwise, we mask out the edge to be superfluous. Then the selector samples edges from $q_\Phi(e_{ij}|G)$ to instantiate an explanatory subgraph s_k . Formally, we have

$$s_k = \mathcal{H} \left(\bigcup_{(i,j) \in \mathcal{E}} \{e_{ij}\} \right) \text{ where } e_{ij} \sim q_\Phi(e_{ij}|G) = \mathcal{T}_\Phi(G), \quad (16)$$

where \mathcal{E} denotes the edges in G .

It is intractable to optimize the explainer with the loss functions in Eq. (14) and Eq. (15) because the e_{ij} are discrete and the sampling procedure in \mathcal{H} is indifferentiable. To tackle this problem, existing graph explanation methods [18, 33, 39] usually assume e_{ij} follows the Bernoulli distribution. And then, they employ the Gumbel-Softmax reparameterization trick [19] to sample edges independently. The reparameterization trick relaxes the edges to be continuous and makes the sampling procedure derivable. However, the assumption on Bernoulli distribution is unreasonable since it assumes that every edge contributes to the GNNs' output independently and discards their correlations.

Instead, we assume that the edges follow a Categorical distribution, i.e., we assume $q_\Phi(e_{ij}|G) = \frac{\exp(w_{ij}/\tau)}{\sum_{(i,j) \in \mathcal{E}} \exp(w_{ij}/\tau)}$ where logits w_{ij} is calculated by the attributor \mathcal{T}_Φ which is implemented by a multilayer perceptron (MLP)², and $\tau > 0$ is the temperature value. Different from the Bernoulli distribution, the Categorical distribution assumption measures the contribution of edge (i, j) among all edges in the whole graph G without overlooking their relevance. With the Categorical assumption, we generate the explanatory subgraphs by sampling k edges from the distribution and set the value of $e_{i,j}$ of corresponding selected edges to be 1 while masking out other edges. To this end, we adopt the Gumbel top- k trick which is provably equivalent to sampling without replacement from $q_\Phi(e_{ij}|G)$ [16, 38]. Formally, edges are sampled by $\text{argtopk}_{(i,j) \in \mathcal{E}}(w_{ij} - \log(-\log U_e))$ where $U_e \sim \text{Uniform}(0, 1)$.

Nonetheless, the operator $\text{argtopk}(\cdot)$ is indifferentiable. To tackle this problem, we choose the straight-through (ST) estimator [4, 12], which can approximately estimate the gradient for discrete variables. Specifically, instead of directly setting the values of mask $e_{i,j}$ corresponding to the selected edges to be 1, we use:

$$e_{i,j} := 1 - \text{no_grad}(q_\Phi(e_{ij}|G)) + q_\Phi(e_{ij}|G), \quad (17)$$

where $\text{no_grad}(\cdot)$ means that the back-propagation signals will not go through this term. Experimental results demonstrate the superiority of the Categorical distribution assumption combined with the ST estimator.

²Details of the implementation can be found in the appendix.

3.4 Theoretical Analysis

In this section, we theoretically analyze the connection between task-agnostic explanation and task-specific explanation, which demonstrates the validity and transferability of our proposed two-stage explanation.

THEOREM 1. *Let $S \in \mathcal{S}$ be a subgraph of G , Z be the representations of G . Given any potential task-specific predictions \hat{Y} , we have:*

- (1) *If S is a task-agnostic sufficient subgraph corresponding to Z , then S must be a task-specific sufficient subgraph corresponding to \hat{Y} .*
- (2) *If S is a task-agnostic ϵ -explanatory subgraph corresponding to Z , then S must be a task-specific ϵ' -explanatory subgraph corresponding to \hat{Y} where $\epsilon' = \epsilon + I(G; Z|\hat{Y})$.*

See the proof details in the Appendix. The above theorem reveals that the task-agnostic sufficient subgraph is a sufficient condition for the task-specific sufficient subgraph. And the task-agnostic explanatory subgraph is a relaxed sufficient condition for the task-specific explanatory subgraph where the compactness constraint is relaxed from ϵ to ϵ' . That is, task-agnostic ϵ -explanatory subgraphs can be used to explain predictions on any potential downstream tasks \hat{Y} since they are sufficient for \hat{Y} . And their compactness on specific tasks is bounded by ϵ and the transferability knowledge in the representation space (knowledge except for \hat{Y} , i.e., $I(G; Z|\hat{Y})$). This theorem demonstrates the transferability of our task-agnostic explanation in the first stage. Moreover, in the second stage, the compactness of task-agnostic explanations on specific tasks can be further tightened by fine-tuning.

4 EXPERIMENTS

In this section, we conduct extensive experimental studies to evaluate our proposed EGIB method. We first introduce the details of our adopted datasets, baseline methods, and experimental settings. And then, we discuss the results of the experiments. Specifically, we aim to answer the following questions: **RQ1:** Can the fine-grained information in the representation space enhance the explanations? **RQ2:** Can our pretrained task-agnostic explanations be transferable to downstream tasks? **RQ3:** Can our proposed invariance regularization term, the Categorical distribution assumption, and the IB-based intelligibility term benefit the effectiveness of EGIB?

4.1 Datasets

To evaluate the effectiveness and transferability of EGIB on both node-level and graph-level tasks, we adopt two groups of real-world datasets that contain multiple tasks. The statistics of the used datasets can be found in the appendix.

MoleculeNet. The MoleculeNet [36] is a large-scale benchmark for molecular machine learning. In a molecular graph, each node indicates an atom, and each edge denotes a bond connecting atoms. The prediction of molecular graphs' properties can be treated as graph-level tasks. We adopt four graph classification tasks in MoleculeNet to evaluate graph-level explanations: BACE, BBBP, SIDER, and HIV.

PPI. The PPI [49] dataset collects the physical interactions between proteins in 24 different human tissues. In a PPI graph, each node indicates a protein, while edges denote interactions among

proteins. Each protein in the datasets has 121 binary labels associated with its functions. The prediction of each protein function can be regarded as a node-level task. We adopt four tasks in PPI to evaluate the node-level explanation.

4.2 Baselines

We adopt three task-specific explanation methods and a task-agnostic explanation method as baselines. Specifically, 1) GNNExplainer [42] is a transductive explanation method that learns edge masks for every single instance. 2) PGExplainer [18] is an inductive explanation method. Compared with GNNExplainer, PGExplainer provides a global understanding of predictions made by GNNs. 3) Refine [33] is an inductive explanation method that learns the multi-grained explanations with class-wise attributors and contrastive learning. It employs a two-stage training strategy to generate explanations with both local and global understanding of the predictions. 4) TAGE [39] is a task-agnostic explanation method that learns the explainer based on conditioned contrastive learning. Moreover, we refer to EGIB-TA as a variant of EGIB without the second fine-tuning stage.

4.3 Experimental Settings and Metrics

Target model settings. In this paper, we consider target models trained in a two-stage fashion. Specifically, we first pretrain a GNN-based encoder to learn graph/node representations in a self-supervised way. Then we freeze the encoder and train MLP-based downstream models with the learned representations in a supervised way. For MoleculeNet, we adopt the pretraining strategies proposed in previous work [10] and pretrain a 5-layer GIN [40] encoder on ZINC-2M [25]. ZINC-2M is a large unlabeled dataset in MoleculeNet, which contains 2 million molecules. Then, we train a 2-layer MLP for every downstream task (i.e., BACE, BBBP, SIDER, and HIV) in a supervised way. For PPI, we employ GRACE [48] to train a 2-layer GCN [14] as the encoder on all graphs from PPI. Downstream models are also adopted as a 2-layer MLP.

Explanation settings. The task-agnostic explainers, i.e., EGIB and TAGE are trained consistently with target models. For MoleculeNet, the explainer is first pretrained on ZINC-2M and then fine-tuned on downstream tasks, i.e., BACE, BBBP, SIDER, and HIV. For PPI, the explainer is first pretrained in a task-agnostic way on the whole graphs and then fine-tuned with specific tasks. Previous work [39] shows that task-specific explainers like PGExplainer cannot transfer to other tasks. Thus, we only evaluate task-specific explainers' performance on the task they are trained with.

Metric. To evaluate the performance of explanation quantitatively, we adopt the metrics of fidelity score and the sparsity score following previous works [39, 45, 46]. The fidelity score measures the change of prediction probability when edges are removed. The larger the fidelity score is, the more critical the removed edges are. And the sparsity score measures the fraction of edges selected as important by explanation methods. The larger the sparsity score, the fewer edges identified as explanations. Details of the two metrics can be found in the appendix.

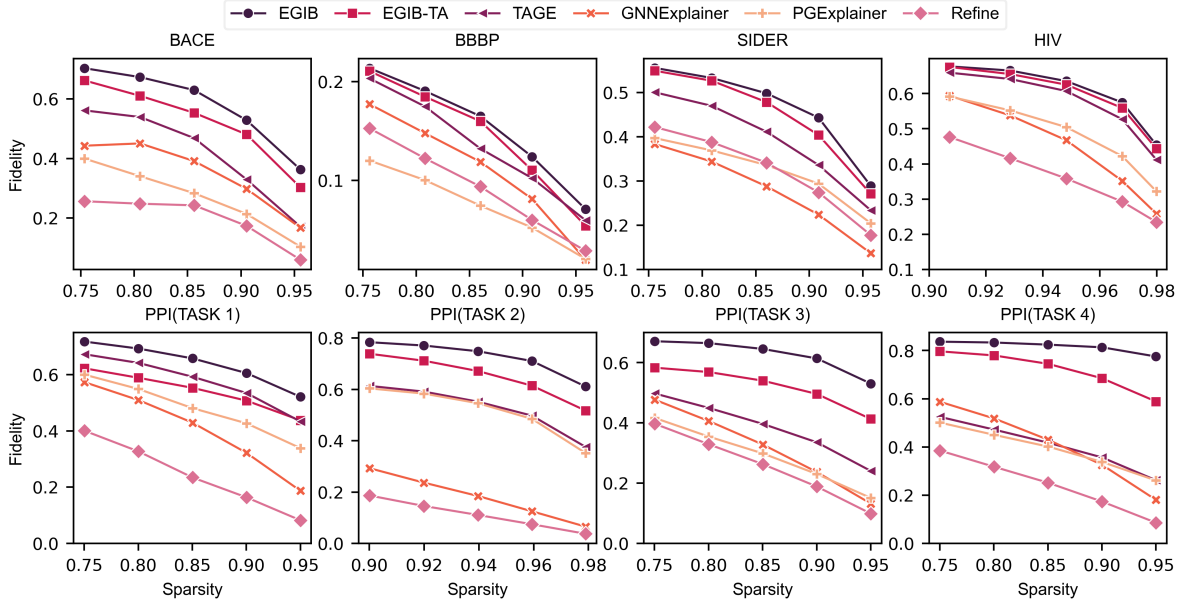


Figure 3: Quantitative results of EGIB against other baselines.

4.4 Experimental Results

4.4.1 Effectiveness and transferability of EGIB. We quantitatively analyze the effectiveness of EGIB on multiple downstream tasks to answer **RQ1** and **RQ2**. In Fig. 3, we demonstrate how the fidelity scores change when the sparsity scores of explanations vary. We observe that 1) Our proposed methods supervised by the representation space (*i.e.*, EGIB and EGIB-TA) outperform task-specific explanation methods (*i.e.*, GNNExplainer, PGExplainer, and Refine). We attribute the observation to the fine-grained information in node/graph representations. The decision procedure of GNNs can be regarded as a projection from the high-dimensional graph space to the low-dimensional label space, where information in the graphs is filtered layer by layer. Compared with task-specific predictions, intermediate representations provide more fine-grained information to supervise task-agnostic explainers. 2) Our proposed pretrained task-agnostic explanations, *i.e.*, EGIB-TA, can transfer to multiple downstream tasks without access to downstream models during pretraining, which indicates that we can explain the transferable knowledge in the pretrained GNNs with EGIB. Specifically, we can observe that EGIB-TA outperforms other baseline methods except for a slight gap with TAGE on task 1 of PPI dataset. A possible interpretation for the slight gap is that the downstream models involved in TAGE provide more information. 3) The task-specific fine-tuning can enhance the performance of EGIB. From the figure, we can observe that EGIB with fine-tuning outperforms EGIB-TA thanks to the task-specific information in the label space.

4.4.2 Ablation studies. To answer **RQ3**, we conduct ablation studies to investigate whether the invariance regularization term, the Categorical distribution assumption, and the IB-based intelligibility term benefit the performance of EGIB. Specifically, we consider three variants of our task-agnostic explanation EGIB-TA: 1) We

eliminate the invariance regularization term and refer to the invariant as EGIB /wo IR. 2) We adopt the typical Bernoulli distribution assumption [18] with GumbelSoftmax reparameterization trick instead of our Categorical assumption and refer to this variant as EGIB /wo Cat. 3) Based on EGIB /wo Cat, we further replace the IB-based intelligibility term with the typical l_1 norm regularization and refer it to EGIB /wo IB. Note that the l_1 norm regularization cannot apply to the Categorical assumption because the l_1 norm of edge logits in the Categorical assumption is fixed as 1. The detailed comparison among different settings can be found in Table 1. We control the sparsity of explanations on the same level and report the fidelity score in Table 2. From the table, we can observe that: 1) EGIB-TA outperforms EGIB wo/ IR and EGIB wo/ Cat, which indicates that both the Categorical assumption and the invariance regularization benefit the performance of our EGIB-TA. 2) EGIB wo/ Cat with the IB-based intelligibility constraint outperforms EGIB wo/ IB with l_1 norm. We attribute this observation to the fact that l_1 norm may lead to a biased assumption on explanations' size [20]. Specifically, explanations can vary in size with the same size restriction. Some explanations may miss crucial edges, while others may involve additional irrelevant edges. Instead, we bypass the biased assumption with IB-based intelligibility constraint by minimizing the superfluous information in Eq. (4) while keeping the relevant information.

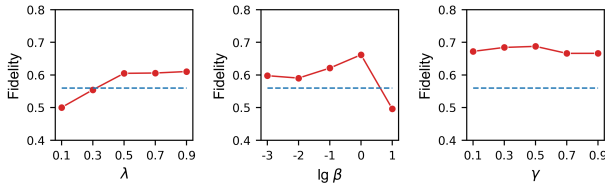
4.4.3 Parameter studies. In Fig. 4, we investigate the influence of different choices hyper-parameters in Eq. (14) and Eq. (15), *i.e.*, λ , β , and γ on BACE task. For λ and β , we report the performance without fine-tuning, *i.e.*, EGIB-TA. Specifically, for λ , we fix $\beta = 1.0$ and vary λ from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. For β , we fix $\lambda = 1.0$ and vary β from $\{0.001, 0.01, 0.1, 1.0, 10.0\}$. For γ , we report the performance after fine-tuning, *i.e.*, EGIB. We fix $\alpha = 1.0$, $\beta = 1.0$ and vary γ from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The blue line indicates TAGE,

Table 1: Settings of ablation studies.

	Invariance regularization	Categorical assumption	Intelligibility constraint
EGIB wo/ IR	×	✓	IB
EGIB wo/ IB	✓	×	l_1
EGIB wo/ Cat	✓	×	IB
EGIB-TA	✓	✓	IB

Table 2: Fidelity score of of ablation studies. Bold indicates the best results on the same level of sparsity.

Graph-level	BACE	BBBP	SIDER	HIV
EGIB wo/ IR	0.424 ± 0.271	0.089 ± 0.130	0.249 ± 0.289	0.485 ± 0.326
EGIB wo/ IB	0.532 ± 0.266	0.191 ± 0.157	0.472 ± 0.300	0.574 ± 0.319
EGIB wo/ Cat	0.567 ± 0.260	0.202 ± 0.159	0.506 ± 0.307	0.603 ± 0.317
EGIB-TA	0.662 ± 0.205	0.211 ± 0.151	0.549 ± 0.297	0.676 ± 0.296
Node-level	TASK 1	TASK 2	TASK 3	TASK 4
EGIB wo/ IR	0.413 ± 0.452	0.413 ± 0.481	0.371 ± 0.439	0.488 ± 0.449
EGIB wo/ IB	0.212 ± 0.381	0.206 ± 0.395	0.266 ± 0.401	0.364 ± 0.421
EGIB wo/ Cat	0.332 ± 0.435	0.351 ± 0.467	0.345 ± 0.429	0.461 ± 0.442
EGIB-TA	0.437 ± 0.452	0.517 ± 0.488	0.413 ± 0.449	0.589 ± 0.423

**Figure 4: Results of parameter studies.**

which performs the second best among all methods. We can observe that 1) the invariance regularization benefits the performance of the explanation, and our method keeps stable when $\lambda \geq 0.5$. 2) The Lagrangian multiplier β makes a trade-off between fidelity and compression. Either too-large or too-small β will degrade the performance. 3) The fine-tuning keeps stable across different values of γ .

4.4.4 Visualization. As there are no ground-truth explanations for our above-used multi-task datasets, we provide more visualization results on two additional single-task datasets with explanation ground-truths, *i.e.*, MUTAG [6] and BA-2MOTIFS [18] as well as quantitative results. Among them, MUTAG is a molecular mutagenic property prediction dataset. Here, carbon rings with chemical groups NO_2 groups and NH_2 are widely known to be mutagenic [18, 20]. BA-2MOTIFS is a synthetic dataset where house motifs and cycle motifs give class labels and thus are regarded as ground-truth explanations. Note that the single-task datasets can only reflect the effectiveness of EGIB while the transferability of EGIB to multiple tasks is not presented. We remain it as future work to construct multi-task datasets with ground-truth explanations to facilitate the research of task-agnostic transferable explanations. Although there is only a binary classification task in both datasets, experimental results still demonstrate the effectiveness of our proposed method. Specifically, we follow the experimental setting in previous work [20] and report ROC-AUC averaged over 10 times tests with different random seeds. From Table 3, we can observe

that our proposed EGIB outperforms other methods quantitatively and qualitatively³. Specifically, our EGIB can correctly identify the crucial edges, while other methods fail and miss some important edges. For example, in the first row of explanations, all baselines miss the NH_2 group while our method correctly captures the mutagenic group. And in the fourth row, we can find some edges in the house motif is missed by baseline methods while our EGIB can identify the complete motif.

5 RELATED WORKS

In this section, we briefly introduce existing post-hoc explanation methods on graphs, which can be categorized into task-specific and task-agnostic methods.

Task-specific Explanation on Graphs. According to a recent survey [45], most task-specific explanation methods on graphs can be pigeonholed into four technique routes: (1) Gradient-based methods [2, 22] usually involve gradient-like scores as heuristics to quantify the importance of edges, nodes, or node features. For example, SA [2] directly employs the squared values of gradients as the importance scores of different graph nodes, edges, or node features. Gradient-based methods usually suffer from saturation problems, *i.e.*, when model output changes minimally w.r.t. any input change, the gradients can hardly reflect the contributions of inputs. (2) Surrogate methods [11, 32] employ a simple and interpretable surrogate model to approximate the predictions of the GNNs for the neighboring areas of the input example. For instance, GraphLime [11] considers the N -hop neighboring nodes and their predictions as a local dataset. Then a kernel-based feature selection algorithm HISC Lasso is employed to fit the local dataset. The weights of different features in HISC Lasso are used to select important features. However, GraphLime can only explain node features and is incapable of making an explanation with graph structure. (3) Decomposition methods [2, 23] build score decomposition rules to distribute the prediction scores to the input space. GNN-LRP [23] makes a high-order Taylor decomposition of GNNs to develop the score decomposition rule. It proves that each term in the Taylor decomposition corresponds to a graph walk, and such terms can be regarded as importance scores. However, the method is computationally expensive as each walk is considered separately. (4) Perturbation-based methods [18, 24, 27, 33, 42] generate masks with a parametrized explainer model. Then the explanatory subgraphs are identified by the masks combined with the input graphs. To name a few, GNNExplainer [42] learns soft masks with a local view and applies the masks to the adjacency matrix.

Task-Agnostic Explanations on Graphs. As far as we know, only one existing work [39] named TAGE attempted to explain GNNs in a task-agnostic way. Specifically, they decompose the explainer as a representation explainer and downstream explainer. The representation explainer is first trained using a self-supervised training framework without knowledge of downstream tasks. Then the representation explainer generates the final explanation in cooperation with the downstream explainer. There are mainly three

³Note that on BA-2MOTIFS and MUTAG, GNNExplainer and PGExplainer work worse than results reported in previous work [18] as we do not cherry pick the target model. Instead, we evaluate the performance of explanation methods on target models with different random seeds. Similar settings can be found in work [20]

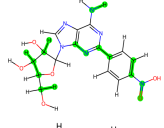
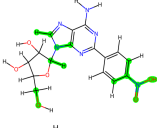
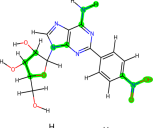
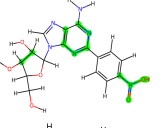
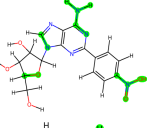
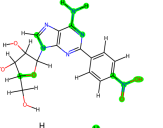
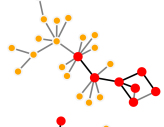
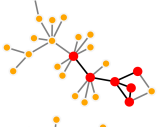
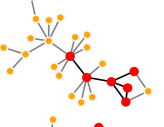
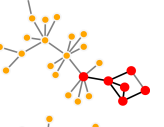
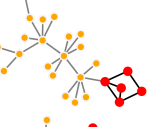
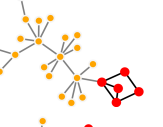
	GNNExplainer	PGExplainer	Refine	TAGE	EGIB-TA	EGIB
MUTAG						
BA-2MOTIFS						
	Explanation AUC					
MUATG	59.77 \pm 2.80	80.38 \pm 3.77	71.48 \pm 7.76	89.49 \pm 6.90	98.21 \pm 1.42	98.31 \pm 1.33
BA-2MOTIFS	66.22 \pm 10.42	75.05 \pm 11.36	73.27 \pm 9.36	75.81 \pm 16.70	82.81 \pm 11.12	88.09 \pm 10.83

Table 3: Qualitative and quantitative results of different explanation methods on datasets with explanation groundtruths. The green edges in MUTAG and black edges in BA-2MOTIFS indicate the explanations.

differences between our EGIB and TAGE. First, in the training time of the representation explainer in TAGE, they mimic the behavior of the downstream explainer with a multivariate Laplace distribution, which may lead to a distribution shift in the inference time. Instead, our EGIB does not rely on any assumptions about downstream tasks. Second, TAGE assumes that the edges in explanatory subgraphs follow a Bernoulli distribution which overlooks the relevance among edges. Moreover, they guarantee the intelligibility of explanations by simply restricting the size of explanatory subgraphs with an l_1 norm regularization, which usually leads to a biased assumption [20]. In contrast, our EGIB makes a Categorical distribution assumption that measures the contribution of an edge among all edges without overlooking their relevance. Additionally, our IB-based intelligibility term does not assume the size of explanations. Third, our EGIB theoretically subsumes the task-agnostic explanation and task-specific explanation into a unified framework. Based on the framework, we further analyze the transferability of task-agnostic explanations.

6 CONCLUSION

We investigate a two-stage explanation strategy since typical task-specific explanation methods are incapable of explaining pretrained GNNs where downstream tasks are inaccessible, not to mention explaining the transferable knowledge in pretrained GNNs. Moreover, the coarse-grained information in the label space may be insufficient to reflect the internal logic of GNNs. To overcome these limitations, we propose a unified framework named Explainable Graph Information Bottleneck (EGIB) based on IB which subsumes the task-specific explanations and task-agnostic explanations. Based

on the unified framework, we derive a tractable bound for optimization and adopt a simple yet effective graph generation architecture. Furthermore, we theoretically prove that the task-agnostic explanation is a relaxed sufficient condition of task-specific explanation, demonstrating the transferability of task-agnostic pretrained explanation. Our experiments demonstrate that our proposed EGIB outperforms other baseline methods on effectiveness. We discuss the limitations and potential social impact in the appendix.

ACKNOWLEDGMENTS

Jihong Wang, Minnan Luo, and Qinghua Zheng are supported by the National Key Research and Development Program of China (No. 2022YFB3102600), National Nature Science Foundation of China (No. 62192781, No. 62272374, No. 62202367, No. 62250009, No. 62137002, No. 61937001), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education, China (IRT_17R86), Project of China Knowledge Center for Engineering Science and Technology, and Project of Chinese academy of engineering “The Online and Offline Mixed Educational Service System for ‘The Belt and Road’ Training in MOOC China”. Yun Lin and Jin Song Dong are supported by the Minister of Education, Singapore (MOET32020-0004), the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity Research and Development Programme (Award No. NRF-NCR_TAU_2021-0002) and the Cyber Security Agency under its National Cybersecurity RD Programme (NCRP25-P04-TAICeN). Minnan Luo also would like to express their gratitude for the support of K. C. Wong Education Foundation.

REFERENCES

- [1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).
- [2] Federico Baldassarre and Hossein Azizpour. 2019. Explainability Techniques for Graph Convolutional Networks. In *International Conference on Machine Learning (ICML) Workshops, 2019 Workshop on Learning and Reasoning with Graph-Structured Representations*.
- [3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.
- [4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).
- [5] Lei Cai, Jundong Li, Jie Wang, and Shuiwang Ji. 2021. Line graph neural networks for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [6] 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.
- [7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [8] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.
- [9] Will Hamilton, Zhitaoying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).
- [10] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).
- [11] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. 2020. Graphlime: Local interpretable model explanations for graph neural networks. *arXiv preprint arXiv:2001.06216* (2020).
- [12] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
- [13] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 805–815.
- [14] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
- [16] Wouter Kool, Herke Van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*. PMLR, 3499–3508.
- [17] Meng Liu, Hongyang Gao, and Shuiwang Ji. 2020. Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 338–348.
- [18] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. *Advances in neural information processing systems* 33 (2020), 19620–19631.
- [19] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).
- [20] Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*. PMLR, 15524–15543.
- [21] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*. PMLR, 5171–5180.
- [22] 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.
- [23] Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T Schutt, Klaus-Robert Mueller, and Gregoire Montavon. 2021. Higher-Order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [24] Patrick Schwab and Walter Karlen. 2019. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems* 32 (2019).
- [25] Teague Sterling and John J Irwin. 2015. ZINC 15—ligand discovery for everyone. *Journal of chemical information and modeling* 55, 11 (2015), 2324–2337.
- [26] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. 2019. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *International Conference on Learning Representations*.
- [27] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*. 1018–1027.
- [28] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
- [29] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. 2019. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625* (2019).
- [30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations* (2018). <https://openreview.net/forum?id=rjXmpikCZ> accepted as poster.
- [31] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep Graph Infomax. *ICLR (Poster)* 2, 3 (2019), 4.
- [32] Minh Vu and My T Thai. 2020. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems* 33 (2020), 12225–12235.
- [33] Xiang Wang, Ying-Xin Wu, An Zhang, Xiangnan He, and Tat-Seng Chua. 2021. Towards Multi-Grained Explainability for Graph Neural Networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems*.
- [34] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. 2020. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149* (2020).
- [35] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872* (2022).
- [36] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.
- [37] Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z Li. 2022. A survey of pretraining on graphs: Taxonomy, methods, and applications. *arXiv preprint arXiv:2202.07893* (2022).
- [38] Sang Michael Xie and Stefano Ermon. 2019. Reparameterizable subset sampling via continuous relaxations. *arXiv preprint arXiv:1901.10517* (2019).
- [39] Yaochen Xie, Sumeet Katariya, Xianfeng Tang, Edward W Huang, Nikhil Rao, Karthik Subbian, and Shuiwang Ji. 2022. Task-Agnostic Graph Explanations. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=s_Q6pLNVHoh
- [40] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
- [41] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ryGs6iA5Km>
- [42] Zhitaoying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 32 (2019).
- [43] Zhitaoying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems* 31 (2018).
- [44] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. 2021. Recognizing predictive substructures with subgraph information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [45] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445* (2020).
- [46] 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.
- [47] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. *Advances in neural information processing systems* 31 (2018).
- [48] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131* (2020).
- [49] Marinka Zitnik and Jure Leskovec. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* 33, 14 (2017), i190–i198.

A DERIVATION DETAILS

A.1 Derivation for Eq. (1) and Eq. (3)

In this subsection, we show how Eq. (1) and Eq. (3) are derived. In Eq. (1), we have

$$I(T; G|S) = I(T; G) - I(T; S) = 0 \quad (18)$$

which can be derived as:

$$\begin{aligned} I(T; G|S) &= I(T; G) - I(T; G; S) \\ &= I(T; G) - (I(T; S) - I(T; S|G)) \\ &= I(T; G) - I(T; S) \end{aligned} \quad (19)$$

The third equation holds because S is a subgraph of G , i.e., S is conditionally independent of T given G . Moreover, in Eq. (3) we have:

$$I(S; G|T) = I(S; G) - I(S; T) \quad (20)$$

Similar to Eq. (1), the Eq. (3) can be proved as :

$$\begin{aligned} I(S; G|T) &= I(S; G) - I(T; G; S) \\ &= I(S; G) - (I(S; T) - I(T; S|G)) \\ &= I(S; G) - I(S; T) \end{aligned} \quad (21)$$

A.2 Derivation for Eq. (22)

In Eq. (22), we have

$$I(S; G) = -I(\tilde{S}; G) + I(\tilde{S}; S) + H(G) \quad (22)$$

which can be proved as:

$$\begin{aligned} I(S; G) &= -I(\tilde{S}; G|S) + I(S; \tilde{S}; G) \\ &= -I(\tilde{S}; G) + I(\tilde{S}; G; S) + H(G) \\ &= -I(\tilde{S}; G) + I(\tilde{S}; S) - I(\tilde{S}; S|G) + H(G) \\ &= -I(\tilde{S}; G) + I(\tilde{S}; S) + H(G) \end{aligned} \quad (23)$$

The second equation holds because $G = (\tilde{S}; S)$, and thus $I(S; \tilde{S}; G) = I(G; G) = H(G)$.

A.3 Proof for Theorem 1

THEOREM. Given a subgraph $S \in \mathcal{S}$ where \mathcal{S} is the set of subgraphs of G , Z denotes the representations of G , \hat{Y} denotes the downstream prediction based on Z . We have:

- (1) If S is a task-agnostic sufficient subgraph corresponding to Z , then S must be a task-specific sufficient subgraph corresponding to \hat{Y} .
- (2) If S is a task-agnostic ϵ -explanatory subgraph corresponding to Z , then S must be a task-specific ϵ' -explanatory subgraph corresponding to \hat{Y} where $\epsilon' = \epsilon + I(G; Z|\hat{Y})$.

PROOF. The conclusion (1) can be formulated as $I(Z; G|S) = 0 \Rightarrow I(\hat{Y}; G|S) = 0$. We prove the conclusion by:

$$\begin{aligned} &I(\hat{Y}; G|S) - I(Z; G|S) \\ &= I(G; \hat{Y}) - I(S; \hat{Y}) - (I(G; Z) - I(S; Z)) \quad (\text{Substitute in Eq. (1)}) \\ &= I(G; \hat{Y}; Z) + I(G; \hat{Y}|Z) - (I(S; \hat{Y}; Z) + I(S; \hat{Y}|Z)) - I(G; Z) + I(S; Z) \\ &= I(G; \hat{Y}; Z) - I(S; \hat{Y}; Z) - I(G; Z) + I(S; Z) \\ &= -I(G; Z|\hat{Y}) + I(S; Z|\hat{Y}) \\ &= -I(G; Z|\hat{Y}) + I(G; S; Z|\hat{Y}) + I(S; Z|\hat{Y}, G) \\ &= -I(G; Z|\hat{Y}, S) \leq 0 \end{aligned} \quad (24)$$

The third equation holds because \hat{Y} denotes the predictions based only on Z and thus \hat{Y} is conditionally independent of G and S given Z , i.e., $I(G; \hat{Y}|Z) = 0$ and $I(S; \hat{Y}|Z) = 0$. The final equation holds because S is the subgraph of G , and thus S is independent of Z given G and \hat{Y} , i.e., $I(S; Z|\hat{Y}, G) = 0$. The proof of conclusion (1) is completed.

The conclusion (2) can be formulated as $I(S; G|Z) \leq \epsilon \Rightarrow I(S; G|\hat{Y}) \leq \epsilon + I(G; Z|\hat{Y})$. To prove the conclusion, we have:

$$\begin{aligned} &I(S; G|Z) - I(S; G|\hat{Y}) \\ &= I(S; \hat{Y}) - I(S; Z) \quad (\text{Substitute in Eq. (3)}) \\ &= I(G; \hat{Y}) - I(G; Z) \quad (\text{Substitute in Eq. (1)}) \\ &= I(G; \hat{Y}; Z) + I(G; \hat{Y}|Z) - I(G; Z) \\ &= I(G; \hat{Y}; Z) - I(G; Z) \\ &= -I(G; Z|\hat{Y}) \end{aligned} \quad (25)$$

The first equation holds because if S is a task-agnostic sufficient subgraph corresponding to Z , it must be a task-specific sufficient subgraph corresponding to \hat{Y} . Then substituting Eq. (3), we can prove the equation. According to the definition of ϵ -explanatory subgraph, we have $I(S; G|Z) \leq \epsilon$. Since $I(S; G|\hat{Y}) = I(S; G|Z) + I(G; Z|\hat{Y}) \leq \epsilon + I(G; Z|\hat{Y})$, we can say that S is a task-specific ϵ' -explanatory subgraph where $\epsilon' = \epsilon + I(G; Z|\hat{Y})$. The proof is completed. \square

B IMPLEMENTATION DETAILS

B.1 Implementations of Explainers

We adopt the multilayer perceptron (MLP) as the attributor \mathcal{T}_Φ to calculate the logits w_{ij} following PGExplainer [18]. Formally, for graph-level tasks, the logits of edge (i, j) , i.e., w_{ij} is calculated by the concatenation of its corresponding nodes' representations, z_i and z_j :

$$w_{ij} = \text{MLP}_g([z_i; z_j]; \Phi) \quad (26)$$

For node-level tasks, except for the corresponding nodes' representations of edge (i, j) , we also involve the target node's representation, z_t whose prediction is to be explained:

$$w_{ij} = \text{MLP}_n([z_i; z_j; z_t]; \Phi) \quad (27)$$

For both MLP_g and MLP_n , we adopt a 2-layer MLP with Relu activation to implement them.

B.2 Implementation of Target GNNs and Downstream Models

We adopt the same settings following previous work [39]. For MoleculeNet, we adopt a 5-layer GIN with Relu activation as the encoder. The hidden dimension is fixed as 600 for each layer. Batch Normalization is adopted to normalize the output of hidden layers. For PPI, we adopt a 2-layer GCN with Relu activation. For downstream models, 2-layer MLPs are adopted to predict specific tasks. For convenience of reproduction and comparison, we adopt the open-sourced parameters released by previous works⁴.

⁴<https://drive.google.com/drive/folders/1f41cVroWXtbACHfVgYkLo3Mj-gBdkexl>

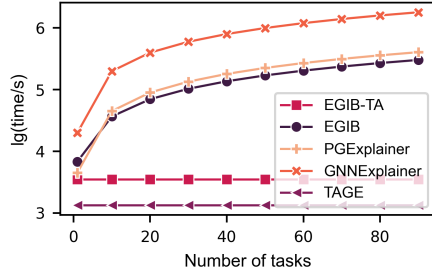


Figure 5: Training time of different explanation strategies on PPI. The y-axis indicates the time cost in seconds. Note that we take the logarithm for convenience of demonstration.

Table 4: Statistics of datasets

	MoleculeNet				PPI
	BACE	BBBP	SIDER	HIV	
# of Graphs	1513	2039	1427	41127	24
Avg. # of Nodes	34.12	24.05	33.64	25.53	56,944
Avg. # of Edges	36.89	25.94	35.36	27.84	818,716

B.3 Details of EGIB Training

For the training of EGIB, we adopt the Adam optimizer. For MoleculeNet, the learning rate is fixed as 0.001, the batch size is set as 256. The explainer is first pretrained on the unlabeled ZINC-2M dataset with 1 epoch and fine-tuned on specific tasks (BACE, BBBP, SIDER and HIV) with 10 epochs. For PPI dataset, the learning rate is fixed as $5e-6$ and the batch size is set as 4. The explainer is first pretrained on the whole dataset in a unsupervised fashion with 1 epoch and then further fine-tuned on specific tasks with 1 epoch. For the choice of k in the Categorical sampling, we simply set it as $0.1 * |\mathcal{E}|$ where $|\mathcal{E}|$ denotes the number of edges in the graphs or k -hop subgraphs.

C EXPERIMENTS

C.1 Datasets

The statistics of the used datasets are summarized in Table 4.

C.2 Metrics

Given a graph instance g_i , the fidelity score is calculated as:

$$\text{Fidelity} = \frac{1}{N} \sum_{i=1}^N \left(f_t(g_i)_{y_i} - f_t(\tilde{s}_i)_{y_i} \right) \quad (28)$$

where f_t is the target model, \tilde{s}_i denotes the residual subgraphs of which the explanation edges are removed. y_i denotes the label of graph g_i . The sparsity score is calculated as:

$$\text{Sparsity} = \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{|\mathcal{E}_i^s|}{|\mathcal{E}_i|} \right) \quad (29)$$

where $|\mathcal{E}_i^s|$ denotes the number of edges selected to be explanation and $|\mathcal{E}_i|$ is the number of edges in the graph samples.

C.3 Efficiency of EGIB

Besides the results on effectiveness, we also experimentally analyze the efficiency of our EGIB in the multi-task setting. With similar explainer architecture, we emphasize that the most crucial factor that decides the efficiency of explanation methods is the training strategy they adopt. In Fig. 5, we evaluate the efficiency of three different explanation strategies: the task-specific transductive explanation strategy (e.g., GNNExplainer), the task-specific inductive explanation strategy (e.g., PGExplainer) and our proposed two-stage task-agnostic explanation strategy in a multi-task setting. We record the training time (in seconds) on PPI dataset and take the logarithm as the y-axis. All experiments are conducted with the same machine with a single NVIDIA GeForce RTX 3090. We can observe that EGIB is the most efficient method among the three strategies in the multi-task setting because the explainer does not need to be trained from scratch and only requires fine-tuning with a few epochs. This observation demonstrates the superiority of our two-stage explanation strategy on efficiency in multi-task settings. Moreover, another task-agnostic method, TAGE, does not require any fine-tuning after pretraining and thus can be more efficient than our EGIB in a multi-task setting. However, our variant EGIB-TA can be comparable to TAGE without fine-tuning on specific tasks. The constant gap between EGIB-TA and TAGE on efficiency is independent of the task numbers and can be acceptable since EGIB-TA outperforms TAGE on effectiveness.

D DISCUSSION OF LIMITATIONS AND SOCIAL IMPACTS

Access to target models and datasets. Just like typical methods [18, 42], our method relies on access to the target model and the datasets. In other words, we consider a white-box setting. Nonetheless, this knowledge may be inaccessible in practice. When the target model is inaccessible, a straightforward strategy is to employ a surrogate model approximating the behavior of the target model. However, models may yield similar results with different logic. How to give post-hoc explanations with only some queries is an intractable problem. Moreover, since datasets can also be inaccessible, the training of inductive explanations may suffer the data scarcity problem.

Potential social impacts. Actually, most post-hoc explanation methods face reliable problems. Since we try to infer the possible logic of target models in a post-hoc fashion, the explanations may be unreliable. Wrong explanations may lead to potential negative and severe impacts, especially in fields like molecular property prediction or drug discovery. Thus, it is urgent and essential to study methods to verify the reliability of post-hoc explanations.