Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs

Yongqiang Chen[1]*, **Yonggang Zhang**[2], **Yatao Bian**[3], **Han Yang**[1], **Kaili Ma**[1], **Binghui Xie**[1]
[1]The Chinese University of Hong Kong [2]Hong Kong Baptist University
{yqchen,hyang,klma,bhxie21,jcheng}@cse.cuhk.edu.hk yatao.bian@gmail.com
Tongliang Liu[4], **Bo Han**[2], **James Cheng**[1]
[3]Tencent AI Lab [4]TML Lab, The University of Sydney
tongliang.liu@sydney.edu.au {csygzhang,bhanml}@comp.hkbu.edu.hk

Abstract

Despite recent success in using the invariance principle for out-of-distribution (OOD) generalization on Euclidean data (e.g., images), studies on graph data are still limited. Different from images, the complex nature of graphs poses unique challenges to adopting the invariance principle. In particular, distribution shifts on graphs can appear in a variety of forms such as attributes and structures, making it difficult to identify the invariance. Moreover, domain or environment partitions, which are often required by OOD methods on Euclidean data, could be highly expensive to obtain for graphs. To bridge this gap, we propose a new framework, called **C**ausality **I**nspired **I**nvariant **G**raph Le**A**rning (CIGA), to capture the invariance of graphs for guaranteed OOD generalization under various distribution shifts. Specifically, we characterize potential distribution shifts on graphs with causal models, concluding that OOD generalization on graphs is achievable when models focus *only* on subgraphs containing the most information about the causes of labels. Accordingly, we propose an information-theoretic objective to extract the desired subgraphs that maximally preserve the invariant intra-class information. Learning with these subgraphs is immune to distribution shifts. Extensive experiments on 16 synthetic or real-world datasets, including a challenging setting – DrugOOD, from AI-aided drug discovery, validate the superior OOD performance of CIGA[1].

1 Introduction

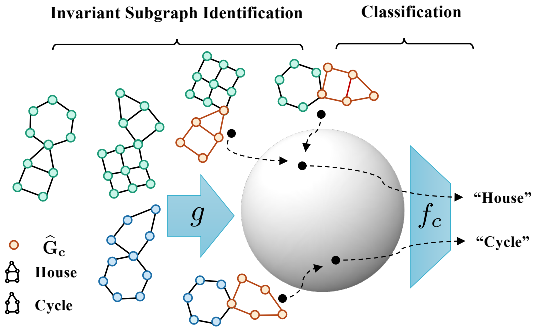
Graph representation learning with graph neural networks (GNNs) has gained great success in tasks involving relational information [45, 35, 99, 106, 107]. However, it assumes that the training and test graphs are drawn from the same distribution, which is often violated in reality [37, 47, 38, 40]. The mismatch between training and test distributions, i.e., *distribution shifts*, introduced by some underlying environmental factors related to data collection or processing, could seriously degrade the performance of deployed models [7, 24]. Such *out-of-distribution* (OOD) generalization failures become the major roadblock for practical applications of graph representation learning [40].

Meanwhile, enabling OOD generalization on regular Euclidean data has received surging attention and several solutions were proposed [4, 81, 10, 49, 23, 48, 2]. In particular, the invariance principle from causality is at the heart of those works [76, 74, 79]. The principle leverages the Independent Causal Mechanism (ICM) assumption [74, 77] and implies that, model predictions that only focus on the causes of the label can stay invariant to a large class of distribution shifts [76, 4, 2].

[*] Work done during an internship at Tencent AI Lab.

[1] Code is available at https://github.com/LFhase/CIGA.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a)

Algorithm	OOD Guarantee	Regime	E Known	SCM Support
IRM [4]	Yes	$\mathbb{R}$	Yes	PIIF
IB-IRM [2]	Yes	$\mathbb{R}$	Yes	PIIF&FIIF
EIIL [23]	Yes	$\mathbb{R}$	No	PIIF
DANN [31]	N/A	$\mathbb{R}$	Yes	N/A
MatchDG [61]	N/A	$\mathbb{R}$	Yes	FIIF
GroupDro [81]	N/A	$\mathbb{R}$	Yes	N/A
CNC [124]	N/A	$\mathbb{R}$	No	N/A
GIB [120]	Yes	$\mathcal{G}$	No	FIIF
DIR [104]	No	$\mathcal{G}$	No	FIIF
CIGA (Ours)	**Yes**	$\mathcal{G}$	**No**	**PIIF&FIIF**

(b)

Figure 1: (a) Illustration of **C**ausality Inspired **I**nvariant **G**raph Le**A**rning (CIGA): GNNs need to classify graphs based on the specific motif ("House" or "Cycle"). The featurizer g will extract an (orange colored) subgraph $\widehat{G}_c$ from each input for the classifier f_c to predict the label. The training objective of g is implemented in a contrastive strategy where the distribution of $\widehat{G}_c$ at the latent sphere will be optimized to maximize the intra-class mutual information, hence predictions will be invariant to distribution shifts; (b) An overview of potential algorithms for OOD generalization on graphs.

Despite the success of the invariance principle on Euclidean data, the complex nature of graphs raises several new challenges that prohibit direct adoptions of the principle. First, distribution shifts on graphs are more complicated. They can happen at both attribute-level and structure-level, and be observed in multiple forms such as graph sizes, subgraph densities and homophily [113, 11, 102]. On the other hand, each of the shifts can spuriously correlate with labels in different modes [4, 71, 2]. Consequently, the entangled complex distribution shifts make it more difficult to identify and capture the invariance on graphs. Second, OOD algorithms developed and analyzed on Euclidean data often require additional environment (or domain) labels for distinguishing the sources of distribution shifts [4]. However, the environment labels could be highly expensive to obtain and thus often unavailable for graphs, as collecting the labels usually requires expert knowledge due to the abstraction of graphs [37]. These challenges render the problem studied in this paper even more challenging:

How could one generalize the invariance principle to enable OOD generalization on graphs?

To solve the above problem, we propose **C**ausality Inspired **I**nvariant **G**raph Le**A**rning (CIGA), a new framework for capturing the invariance of graphs to enable guaranteed OOD generalization under different distribution shifts. Specifically, we build three Structural Causal Models (SCMs) [74] to characterize the distribution shifts that could happen on graphs: one is to model the graph generation process, and the other two are to model two possible interactions between invariant and spurious features during the graph generation, i.e., Fully Informative Invariant Feature (FIIF) and Partially Informative Invariant Feature (PIIF) (Sec. 2.2). Then, we generalize the invariance principle to graphs for OOD generalization: GNN models are invariant to distribution shifts if they focus only on an invariant and critical subgraph G_c that contains the most of the information in G about the underlying causes of the label. Thus, the problem of achieving OOD generalization on graphs can be rephrased into two processes: invariant subgraph identification and label prediction. Accordingly, shown as Fig. 1(a), we introduce a prototypical invariant graph learning algorithm that decomposes a GNN into: a) a featurizer g for identifying the underlying invariant subgraph G_c from G; b) a classifier f_c for making predictions based on G_c. To extract the desired subgraph G_c, we derive an information-theoretic objective for the featurizer to identify subgraphs that maximally preserves the invariant intra-class information across a set of different (unknown) environments. We theoretically show that this approach can provably identify the underlying G_c under mild assumptions (Sec. 3).

Experiments on 16 synthetic and real-world datasets with various distribution shifts, including a challenging setting from AI-aided drug discovery [40], show that CIGA can significantly outperform all of existing methods up to 10%, demonstrating its promising OOD generalization ability (Sec. 4).

Related Work. We review existing methods that might improve the OOD generalization on graphs, summarize the main differences between our solution and them in Table 1(b), and leave thorough discussions to Appendix B.2. On Euclidean data, Invariant Learning [4, 23, 2], Group Distributionally Robust Optimization [49, 81, 124], Domain Adaption and Domain Generalization [31, 93, 52, 27, 61, 100] are three widely adopted approaches to enable OOD generalization. However, they all have their own limitations when being applied to graphs. First, previous invariant learning methods are mostly developed and analyzed for Euclidean data [4, 2, 23], or under specific SCM assumptions [4], making the theoretical results hardly able to generalize to the complicated graph data [80] that can have multiple types of distribution shifts [71]. Group Distributionally Robust Optimization that minimizes the gap between worst group risk and average risk [49, 81, 124], and Domain Adaption/Generalization methods that aim to learn class-conditional domain invariant representations [31, 93, 52, 27, 100], cannot guarantee a min-max optimal predictor without additional assumptions [126, 4, 2]. Moreover, most existing methods require environment labels that are however expensive to obtain in graphs, which limits their applications to graphs [4, 49, 2, 81, 31, 93, 27, 61]. In contrast, we aim to develop OOD algorithms for graphs that are provably generalizable under different types of distribution shifts.

Another line of relevant works is about GNN explainability that aims to find a subgraph of the input as the explanation for a GNN prediction [116, 122]. Although some may leverage causality to justify the generated explanation [53], they mostly focus on understanding the predictions of GNNs instead of for OOD generalization. The closest works to ours are two interpretable GNNs that aim to explicitly extract a subgraph for both predictions and explanations guided by information theory [120] and causality [104], respectively. However, they focus on graphs and shifts generated under a specific SCM. Although one of them can provide theoretical guarantee for OOD generalization [120] by using the information bottleneck criteria [2], they would inevitably fail to generalize to graphs generated under different SCMs. More discussions about the failure are deferred to Appendix D.4. Besides, Bevilacqua et al. [11] also discuss OOD generalization on graphs but limited to a specific graph family and graph size shifts. Wu et al. [103] propose OOD generalization algorithms on graphs for the task of node classification, also limited to graphs and shifts under a specific SCM.

To the best of our knowledge, there is no existing work that could handle more comprehensive graph distribution shifts than CIGA, while also achieving provable OOD generalization performance.

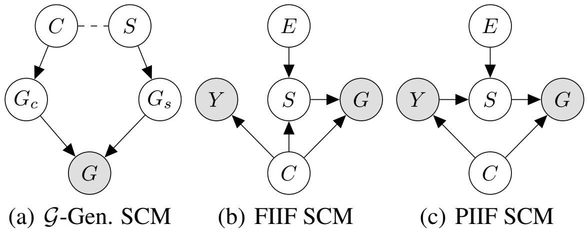
2 OOD Generalization on Graphs through the Lens of Causality

2.1 Problem Setup

In this work, we focus on OOD generalization in graph classification. Specifically, we are given a set of graph datasets $\mathcal{D} = \{\mathcal{D}^e\}_e$ collected from multiple environments $\mathcal{E}_{\text{all}}$. Samples $(G_i^e, Y_i^e) \in \mathcal{D}^e$ from the same environment are considered as drawn independently from an identical distribution $\mathbb{P}^e$. A GNN $\rho \circ h$ generically has an encoder $h : \mathcal{G} \to \mathbb{R}^h$ that learns a meaningful representation h_G for each graph G to help predict the label $\hat{Y}_G = \rho(h_G)$ with a downstream classifier $\rho : \mathbb{R}^h \to \mathcal{Y}$. The goal of OOD generalization on graphs is to train a GNN $\rho \circ h$ with data from training environments $\mathcal{D}_{\text{tr}} = \{\mathcal{D}^e\}_{e \in \mathcal{E}_{\text{tr}} \subseteq \mathcal{E}_{\text{all}}}$ that generalizes well to all (unseen) environments, i.e., to minimize $\max_{e \in \mathcal{E}_{\text{all}}} R^e$, where R^e is the empirical risk of $\rho \circ h$ under environment e [97, 4]. We leave more details about the background of GNN for graph classification and invariant learning in Appendix B.1.

It is known that OOD generalization is impossible without assumptions on the environments $\mathcal{E}_{\text{all}}$ [74, 2]. Thus, we will first formulate the data generation process with structural causal model and latent-variable model [74, 77, 50], to characterize the distribution shifts that could happen on graphs. Then, we investigate whether the existing methods are generalizable under these distribution shifts.

2.2 Graph Generation Process



(a) $\mathcal{G}$-Gen. SCM (b) FIIF SCM (c) PIIF SCM

Figure 2: SCMs on graph distribution shifts.

We take a latent-variable model perspective on the graph generation process and assume that the graph is generated through a mapping $f_{\text{gen}} : \mathcal{Z} \to \mathcal{G}$, where $\mathcal{Z} \subseteq \mathbb{R}^n$ is the latent space and $\mathcal{G} = \cup_{N=1}^{\infty} \{0,1\}^N \times \mathbb{R}^{N \times d}$ is the graph space. Let E denote environments. Following previous works [50, 2], we partition the latent variable from $\mathcal{Z}$ into an invariant part $C \in \mathcal{C} = \mathbb{R}^{n_c}$ and a varying part

$S \in \mathcal{S} = \mathbb{R}^{n_s}$, s.t., $n = n_c + n_s$, according to whether they are affected by E or not. Similarly in images, C and S can represent content and style while E can refer to the locations where the images are taken [7, 125, 50]. Furthermore, C and S control the generation of the observed graphs (Assumption 2.1) and can have multiple types of interactions at the latent space (Assumptions 2.2, 2.3).

Graph generation model. We elaborate the SCM for the graph generation process in Assumption 2.1 and Fig. 2(a), where noises in the structural equations are omitted for simplicity [77].

Assumption 2.1 (Graph Generation Structural Causal Model)**.**

$$G_c := f_{\text{gen}}^{G_c}(C), \qquad G_s := f_{\text{gen}}^{G_s}(S), \qquad G := f_{\text{gen}}^{G}(G_c, G_s).$$

In Assumption 2.1, f_{gen} is decomposed into $f_{\text{gen}}^{G_c}$, $f_{\text{gen}}^{G_s}$ and f_{gen}^{G} to control the generation of G_c, G_s, and G, respectively. Among them, G_c inherits the invariant information of C that would not be affected by the interventions (or changes) of E [74, 77]. For example, certain properties of a molecule can usually be described by a sub-molecule, or a functional group, which is invariant across different species or assays [12, 92, 40]. On the contrary, the generation of G_s and G will be affected by the environment E through S. Thus, graphs collected from different environments (or domains) can have different distributions of structure-level properties (e.g., graph sizes [11, 102]) as well as feature-level properties (e.g., homophily [62, 17]). Therefore, the subgraph G_s inherits the spurious feature about Y [125]. In fact, Assumption 2.1 is compatible with many graph generation models by specifying the function classes of $f_{\text{gen}}^{G_c}$, $f_{\text{gen}}^{G_s}$ and f_{gen}^{G} [89, 57, 117, 59]. Since our goal is to characterize the potential distribution shifts in Assumption 2.1, we focus on building a general SCM that is compatible to many graph families and leave graph family specifications and their implications to OOD generalization in future works. More discussions are provided in Appendix C.

Interactions at latent space. Following previous works [4, 2], we categorize the latent interactions between C and S into Fully Informative Invariant Features (FIIF, Fig. 2(b)) and Partially Informative Invariant Features (PIIF, Fig. 2(c))[2], depending on whether the latent invariant part C is fully informative about label Y, i.e., $(S, E) \perp\!\!\!\perp Y|C$. Formal definitions of the corresponding SCMs are given as follows, where noises are omitted for simplicity [74, 77].

Assumption 2.2 (FIIF Structural Causal Model)**.** $Y := f_{\text{inv}}(C),\ S := f_{\text{spu}}(C, E),\ G := f_{\text{gen}}(C, S).$

Assumption 2.3 (PIIF Structural Causal Model)**.** $Y := f_{\text{inv}}(C),\ S := f_{\text{spu}}(Y, E),\ G := f_{\text{gen}}(C, S).$

In the two SCMs above, f_{gen} corresponds to the graph generation process in Assumption 2.1, and f_{spu} is the mechanism describing how S is affected by C and E at the latent space. By definition, S is directly controlled by C in FIIF and indirectly controlled by C through Y in PIIF, which can exhibit different behaviors in the observed distribution shifts. In practice, performances of OOD algorithms can degrade dramatically if one of FIIF or PIIF is excluded [5, 71]. This issue can be more serious in graphs, since different distribution shifts can have different interaction modes at the latent space. Moreover, $f_{\text{inv}} : \mathcal{C} \rightarrow \mathcal{Y}$ indicates the labelling process, which assigns labels Y for the corresponding G merely based on C. Consequently, $\mathcal{C}$ is better clustered than $\mathcal{S}$ when given Y [13, 15, 86, 87], which also serves as the necessary separation assumption for a classification task [69, 16, 65].

Assumption 2.4 (Better Clustered Invariant Features)**.** $H(C|Y) \leq H(S|Y)$.

2.3 Challenges of OOD Generalization on Graphs

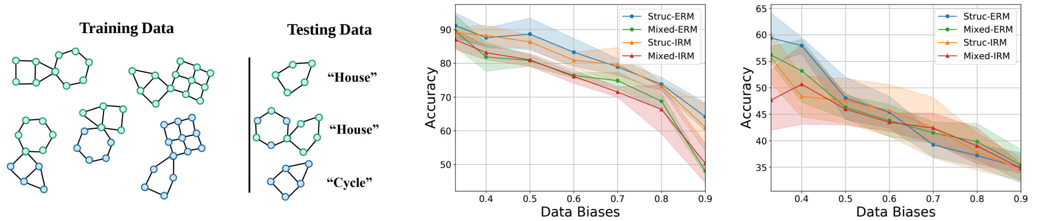
Built upon the graph generation process, we can formally derive the desired GNN that is able to generalize to OOD graphs under different distribution shifts, which implies the invariant GNN below[3].

Definition 2.5 (Invariant GNN)**.** Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{\text{all}}$ that follow the same graph generation process in Sec. 2.2, considering a GNN $\rho \circ h$ that has a permutation invariant graph encoder $h : \mathcal{G} \rightarrow \mathbb{R}^h$ and a downstream classifier $\rho : \mathbb{R}^h \rightarrow \mathcal{Y}$, $\rho \circ h$ is an invariant GNN if it minimizes the worst case risk among all environments, i.e., $\min \max_{e \in \mathcal{E}_{\text{all}}} R^e$.

Can existing methods produce a desired invariant GNN model? We find the answers to be negative unfortunately. Based on the synthetic BAMotif graph classification task [58, 104] shown in Fig. 3,

[2] Note that FIIF and PIIF can be mixed as Mixed Informative Invariant Features (Appendix 6(d)) in several ways, while our analysis will focus on the axiom ones for the purpose of generality.

[3] A discussion on Def. 2.5 and its relation to the SCMs is provided in Appendix E.1.



(a) Failure cases for existing methods. (b) Structure and attribute shifts. (c) Mixed with graph size shifts.

Figure 3: Failures of OOD generalization on graphs: (a) GNNs are required to classify whether the graph contains a "house" or "cycle" motif, where the colors represent node features. However, distribution shifts in the training data exist at both structure-level (from left to right: "house" mostly co-occur with a hexagon), attribute-level (from upper to lower: nodes are mostly colored green if the graph contains a "house", or colored blue if the graph contains a "cycle"), and graph sizes, making GNNs hard to capture the invariance. Consequently, *ERM can fail* for leveraging the shortcuts and predicting graphs that have a hexagon or have nodes mostly colored green as "house". *IRM can fail* as the test data are not sufficiently supported by the training data. (b) GCNs optimized with neither ERM nor IRM can generalize to OOD graphs under structure-level shifts (Struc-) or mixed with feature shifts (Mixed-). (c) When more complex shifts presented, GNNs can fail more seriously.

we theoretically and empirically analyze whether existing methods could produce an invariant GNN, through the investigation of the following aspects. More details and results are given in Appendix D.

Can GNNs trained with ERM generalize to OOD graphs? As shown in Fig. 3, we find that GNNs trained with the standard empirical risk minimization (ERM) algorithm [97] are not able to generalize to OOD graphs. As the data biases grows stronger, the performances of GNNs drop dramatically. Furthermore, when graph size shifts are mixed in the data, GNNs can have larger variance at low data biases, indicating the instability of learning the desired relationships for the task. The reason is that ERM tends to overfit to the shortcuts or spurious correlations presented in specific substructures or attributes in the graphs [33]. This phenomenon has also been shown to exist in GNNs equipped with more sophisticated architectures such as attention mechanisms [99], under graph size shifts [46].

Can OOD objectives improve OOD generalization of GNNs? Meanwhile, as shown in Fig. 3, OOD objectives primarily developed on Euclidean data such as invariant risk minimization (IRM) [4] also cannot alleviate the problem. On the contrary, IRM can fail catastrophically at non-linear regime if without sufficient support overlap for the test environments, i.e., $\cup_{e\in\mathcal{E}_{\text{te}}}\text{supp}(\mathbb{P}^e) \not\subseteq \cup_{e\in\mathcal{E}_{\text{tr}}}\text{supp}(\mathbb{P}^e)$ [80]. In addition to IRM, the failure would also happen for alternative objectives [49, 9, 2] as proved by Rosenfeld et al. [80]. Besides, different distribution shifts on graphs can be nested with each other where each one can have distinct spurious correlation type, e.g., FIIF or PIIF. OOD objectives will also fail seriously if either of the correlation types is not supported [5, 71]. Moreover, non-trivial environment partitions or labels are required for performance guarantee of these OOD objectives [4, 49, 81, 2]. However, collecting meaningful environment partitions of graphs requires expert knowledge about graph data. Thus, the environment labels can be expensive to obtain and are usually not available [67, 28, 37]. Alternative options such as random partitions tend not to alleviate the issue [23, 55], as it can be trivially deemed as mini-batching.

Challenges of OOD generalization on graphs. The aforementioned failure analysis reveals that existing methods or objectives fail to elicit an invariant GNN primarily due to the following two challenges: a) Distribution shifts on graphs are more complicated where different types of spurious correlations can be entangled via different graph properties; b) Environment labels are usually not available due to the abstraction of graphs. Despite these challenges, we are still highly motivated to address the following research question: *Would it be possible to learn an invariant GNN that is generalizable under various distribution shifts by lifting the invariance principle to the graph data?*

3 Invariance Principle for OOD Generalization on Graphs

We provide affirmative answers to the previous question by proposing a new framework, CIGA: **C**ausality Inspired **I**nvariant **G**raph Le**A**rning. Specifically, built upon the SCMs in Sec. 2.2, we generalize the invariance principle to graphs and instantiate the principle with theoretical guarantees.

3.1 Invariance for OOD Generalization on Graphs

Towards extending the invariance principle to graphs under SCMs in Sec. 2.2, we need to identify a set of variables that have stable causal relationship with Y under both FIIF and PIIF (Assumption 2.2, 2.3). According to the ICM assumption [77], the labeling process $C \to Y$ is not informed nor influenced by other processes, implying that the conditional distribution $P(Y|C)$ remains invariant to the interventions on the environment latent variable E [74]. Consequently, for a GNN with a permutation invariant encoder $h : \mathcal{G} \to \mathbb{R}^h$ and a downstream classifier $\rho : \mathbb{R}^h \to \mathcal{Y}$, if h can recover the information of C from G in the learned graph representations, then the learning of ρ resembles traditional ERM [97] and can achieve the desired min-max optimality required by an invariant GNN (Def. 2.5). However, recovering C from G is particularly difficult, since the generation of G from C involves two causal mechanisms $f_{\text{gen}}^{G_c}$ and f_{gen}^{G} in Assumption 2.1. The unavailability of E further adds up the difficulty of enforcing the independence between the learned representations and E.

3.2 Invariant Graph Learning Framework

Causal algorithmic alignment. To enable a GNN to learn to extract the information about C from G, we propose the CIGA framework that *explicitly aligns with* the two causal mechanisms $f_{\text{gen}}^{G_c}$ and f_{gen}^{G} in Assumption 2.1. The idea of alignment in CIGA is motivated by the algorithmic reasoning results that a neural network can learn a reasoning process better if its computation structure aligns with the process better [108, 110]. Specifically, we realize the alignment by decomposing a GNN into two sub-components[4]: a) a featurizer GNN $g : \mathcal{G} \to \mathcal{G}_c$ aiming to identify the desired G_c; b) a classifier GNN $f_c : \mathcal{G}_c \to \mathcal{Y}$ that predicts the label Y based on the estimated G_c, where $\mathcal{G}_c$ refers to the space of subgraphs of G. Formally, the learning objectives of f_c and g can be formulated as:

$$\max\nolimits_{f_c,\, g} I(\widehat{G}_c; Y), \text{ s.t. } \widehat{G}_c \perp\!\!\!\perp E,\ \widehat{G}_c = g(G), \tag{1}$$

where maximizing $I(\widehat{G}_c; Y)$ is equivalent to minimizing a variational upper bound of $R(f_c(\widehat{G}_c))$ [3, 120] that takes $\widehat{G}_c$ as inputs to predict label Y for G through f_c and g, and $\widehat{G}_c$ is the estimated subgraph containing the information about C and hence needs to be independent of E. Moreover, the extracted G_c can either shares the same graph space with input G or has its own space with latent node and edge features, depending on the specific graph generation process. In practice, architectures from the literature of interpretable GNNs are compatible with CIGA [122], hence can serve as practical choices for the implementation of CIGA. More details are given in Appendix F.

Although we can technically align with the two causal mechanisms with g and f_c, trivially optimizing this architecture cannot satisfy $\widehat{G}_c \perp\!\!\!\perp E$. Formally, merely maximizing $I(\widehat{G}_c; Y)$ may include a subgraph from G_s in $\widehat{G}_c$ since G_s also shares certain mutual information with Y. Moreover, the unavailability of E prevents the direct usage of E in enforcing the independence that is often adopted by previous methods [4, 49, 81, 31, 93], making the identification of G_c more challenging.

Optimization objective. To mitigate this issue, we need to find and translate other properties of G_c into some differentiable and equivalent objectives to satisfy the independence constraint $\widehat{G}_c \perp\!\!\!\perp E$. *The goal of the desired objective.* We begin by considering a simplistic setting where all the invariant subgraphs G_c have the same size s_c, i.e., $|G_c| = s_c$[5]. When maximizing $I(\widehat{G}_c; Y)$ in Eq. 1, both FIIF and PIIF can introduce part of G_s into $\widehat{G}_c$. In FIIF (Fig. 2(b)), as G_c already contains the maximal possible information in G about Y, G_c is a solution to $\max I(\widehat{G}_c; Y)$. However, some subgraph of G_c can be replaced by some subgraph of G_s that is equally informative about Y. In PIIF (Fig. 2(c)), there also exists some subgraph of G_s that contains additional information about Y than G_c, hence $\widehat{G}_c$ is more likely to involve some subgraph of G_s. Thus, the new objective needs to eliminate the auxiliary subgraphs of $\widehat{G}_c$ from G_s such that the estimated $\widehat{G}_c$ can only contain G_c.

An important property of G_c. Under both FIIF and PIIF SCMs (Fig. 2), for $G_c^{e_1}, G_c^{e_2}$ that relate to the same causal factor c under two environments e_1 and e_2, the desired $\widehat{G}_c^{e_1}, \widehat{G}_c^{e_2}$ in e_1 and e_2 tend to have high mutual information, i.e., $(G_c^{e_1}, G_c^{e_2}) \in \arg\max I(\widehat{G}_c^{e_1}; \widehat{G}_c^{e_2})$. While for $G_c^{e_1}$

[4] The encoder of the GNN in CIGA can be regarded as the composition of g and the graph encoder in f_c.

[5] Throughout the paper, we use generalized set operators for the ease of understanding. They can have multiple implementations in terms of nodes, edges or attributes.

and another $G_{c'}^{e_1}$ corresponding to a different $c' \neq c$, under the same environment e_1, including any subgraph from $G_s^{e_1}$ in $\widehat{G}_c^{e_1}, \widehat{G}_{c'}^{e_1}$ will enlarge their mutual information, or in other words, $(G_c^{e_1}, G_{c'}^{e_1}) \in \arg\min I(\widehat{G}_c^{e_1}; \widehat{G}_{c'}^{e_1})$. Thus, we can derive an important property of G_c, that is, $\forall e_1, e_2 \in \mathcal{E}_{\text{all}}$,

$$G_c^{e_1} \in \arg\max_{\widehat{G}_c^{e_1}} I(\widehat{G}_c^{e_1}; \widehat{G}_c^{e_2} | C = c) - I(\widehat{G}_c^{e_1}; \widehat{G}_{c'}^{e_2} | C = c', c' \neq c), \tag{2}$$

where $\widehat{G}_c^{e_1}$ and $\widehat{G}_c^{e_2}$ are the estimated invariant subgraphs corresponding to the same causal factor c under environment e_1 and e_2, respectively, while $\widehat{G}_{c'}^{e_2}$ corresponds to a different causal factor c'.

Deriving CIGA*v1 based on the identified property of* G_c. In practice, C is not given. Nevertheless, since C and Y shares a stable causal relationship in both FIIF and PIIF SCMs, Y can serve as a proxy of C in Eq. 2. Moreover, as Eq. 2 holds for any $\forall e_1, e_2 \in \mathcal{E}_{\text{all}}$, the environment superscripts can be eliminated without affecting Eq. 2. Furthermore, when both $I(\widehat{G}_c^{e_1}; \widehat{G}_c^{e_2} | C = c)$ and $I(\widehat{G}_c; Y)$ are maximized, $I(\widehat{G}_c^{e_1}; \widehat{G}_{c'}^{e_1} | C = c', c' \neq c)$ is automatically minimized, otherwise all classes will collapse to trivial solutions which is contradictory given $I(\widehat{G}_c; Y)$ being maximized. Therefore, we can derive an alternative objective to Eq. 1 by leveraging Eq. 2 to replace the independence condition:

$$\text{(CIGAv1)} \qquad \max_{f_c, g} I(\widehat{G}_c; Y), \ \text{s.t.}\ \widehat{G}_c \in \operatorname*{arg\,max}_{\widehat{G}_c = g(G), |\widehat{G}_c| \le s_c} I(\widehat{G}_c; \widetilde{G}_c | Y), \tag{3}$$

where $\widetilde{G}_c = g(\widetilde{G})$ and $\widetilde{G} \sim \mathbb{P}(G|Y)$, i.e., $\widetilde{G}$ is sampled from training graphs that share the same label Y as G. In Theorem 3.1, we show how Eq. 3 is equivalent to Eq. 1. Nevertheless, Eq. 3 requires a strong assumption on the size of G_c. However, the size of G_c is usually unknown or changes for different Cs. In this circumstance, maximizing Eq. 2 without additional constraints will lead to the presence of part of G_s in $\widehat{G}_c$. For instance, $\widehat{G}_c = G$ is a trivial solution to Eq. 3 when $s_c = \infty$.

Deriving CIGA*v2 by resolving size constraint on* G_c *in* CIGA*v1.* To this end, we further resort to the properties of G_s. In both FIIF and PIIF SCMs (Fig. 2), G_s and G_c can share certain overlapped information about Y. When maximizing $I(\widehat{G}_c; \widetilde{G}_c | Y)$ and $I(\widehat{G}_c; Y)$, the appearance of partial G_s in $\widehat{G}_c$ will not affect the optimality. However, it can reduce the mutual information between the left part $\widehat{G}_s = G - \widehat{G}_c$ and Y, i.e., $I(\widehat{G}_s; Y)$. Therefore, by maximizing $I(\widehat{G}_s; Y)$, we can reduce including part of G_s into $\widehat{G}_c$. Meanwhile, to avoid trivial solution that $G_c \subseteq \widehat{G}_s$ during maximizing $I(\widehat{G}_s; Y)$, we can leverage the better clustering property of G_c implied by Assumption 2.4 to derive the constraint $I(\widehat{G}_s; Y) \le I(\widehat{G}_c; Y)$. Thus, we can obtain a new objective CIGAv2 as follows:

$$\begin{aligned} &\max_{f_c, g} I(\widehat{G}_c; Y) + I(\widehat{G}_s; Y), \ \text{s.t.}\ \widehat{G}_c \in \arg\max_{\widehat{G}_c = g(G)} I(\widehat{G}_c; \widetilde{G}_c | Y), \\ &\text{(CIGAv2)} \qquad\qquad I(\widehat{G}_s; Y) \le I(\widehat{G}_c; Y), \ \widehat{G}_s = G - g(G), \end{aligned} \tag{4}$$

where $\widehat{G}_c = g(G), \widetilde{G}_c = g(\widetilde{G})$ and $\widetilde{G} \sim \mathbb{P}(G|Y)$, i.e., $\widetilde{G}$ is sampled from training graphs that share the same label Y as G. We also prove the equivalence between Eq. 4 and Eq. 1 in Theorem 3.1.

3.3 Theoretical Analysis and Practical Discussions

Theorem 3.1 (CIGA Induces Invariant GNNs)**.** *Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{all}$ that follow the same graph generation process in Sec. 2.2, assuming that* (a) f_{gen}^G *and* $f_{gen}^{G_c}$ *in Assumption 2.1 are invertible,* (b) *samples from each training environment are equally distributed, i.e.,$|\mathcal{D}_{\hat{e}}| = |\mathcal{D}_{\tilde{e}}|$, $\forall \hat{e}, \tilde{e} \in \mathcal{E}_{tr}$, then:*

(i). If $\forall G_c, |G_c| = s_c$, then each solution to Eq. 3, elicits an invariant GNN (Def. 2.5).

(ii). Each solution to Eq. 4, elicits an invariant GNN (Def. 2.5).

We prove Theorem 3.1 (i) and (ii) in Appendix E.2, E.3, respectively.

Practical implementations of CIGA objectives. After showing the power of CIGA, we introduce the practical implementations of CIGAv1 and CIGAv2 objectives. Specifically, an exact estimate of the second term $I(\widehat{G}_c; \widetilde{G}_c | Y)$ could be highly expensive [96, 8]. However, contrastive learning with supervised sampling provides a practical solution for the approximation [42, 20, 82, 96, 8]:

$$I(\widehat{G}_c; \widetilde{G}_c | Y) \approx \mathbb{E}_{\substack{\{\widehat{G}_c, \widetilde{G}_c\} \sim \mathbb{P}_g(G | \mathcal{Y} = Y) \\ \{G_c^i\}_{i=1}^M \sim \mathbb{P}_g(G | \mathcal{Y} \neq Y)}} \log \frac{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})}}{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})} + \sum_{i=1}^M e^{\phi(h_{\widehat{G}_c}, h_{G_c^i})}}, \tag{5}$$

Table 1: OOD generalization performance on structure and mixed shifts for synthetic graphs.

	SPMotif-Struc†			SPMotif-Mixed†			
	bias=0.33	bias=0.60	bias=0.90	bias=0.33	bias=0.60	bias=0.90	Avg
ERM	59.49 (3.50)	55.48 (4.84)	49.64 (4.63)	58.18 (4.30)	49.29 (8.17)	41.36 (3.29)	52.24
ASAP	64.87 (13.8)	64.85 (10.6)	**57.29 (14.5)**	66.88 (15.0)	59.78 (6.78)	**50.45 (4.90)**	60.69
DIR	58.73 (11.9)	48.72 (14.8)	41.90 (9.39)	67.28 (4.06)	51.66 (14.1)	38.58 (5.88)	51.14
IRM	57.15 (3.98)	61.74 (1.32)	45.68 (4.88)	58.20 (1.97)	49.29 (3.67)	40.73 (1.93)	52.13
V-Rex	54.64 (3.05)	53.60 (3.74)	48.86 (9.69)	57.82 (5.93)	48.25 (2.79)	43.27 (1.32)	51.07
EIIL	56.48 (2.56)	60.07 (4.47)	55.79 (6.54)	53.91 (3.15)	48.41 (5.53)	41.75 (4.97)	52.73
IB-IRM	58.30 (6.37)	54.37 (7.35)	45.14 (4.07)	57.70 (2.11)	50.83 (1.51)	40.27 (3.68)	51.10
CNC	70.44 (2.55)	**66.79 (9.42)**	50.25 (10.7)	65.75 (4.35)	59.27 (5.29)	41.58 (1.90)	59.01
CIGAv1	**71.07 (3.60)**	63.23 (9.61)	51.78 (7.29)	**74.35 (1.85)**	**64.54 (8.19)**	49.01 (9.92)	**62.33**
CIGAv2	**77.33 (9.13)**	**69.29 (3.06)**	**63.41 (7.38)**	**72.42 (4.80)**	**70.83 (7.54)**	**54.25 (5.38)**	**67.92**
Oracle (IID)		88.70 (0.17)			88.73 (0.25)		

[†]Higher accuracy and lower variance indicate better OOD generalization ability.

where positive samples $(\widehat{G}_c, \widetilde{G}_c)$ are the extracted subgraphs of graphs that share the same label as G, negative samples are those having different labels, $\mathbb{P}_g(G|\mathcal{Y}=Y)$ is the push-forward distribution of $\mathbb{P}(G|\mathcal{Y}=Y)$ by featurizer g, $\mathbb{P}(G|\mathcal{Y}=Y)$ refers to the distribution of G given the label Y, $\mathbb{P}(G|\mathcal{Y}\neq Y)$ refers to the distribution of G given the label that is different from Y, $h_{\widehat{G}_c}, h_{\widetilde{G}_c}, h_{G_c^i}$ are the graph presentations of the estimated subgraphs, and ϕ is the similarity metric for graph representations. As $M \to \infty$, Eq. 5 approximates $I(\widehat{G}_c; \widetilde{G}_c|Y)$, which can be regarded as a non-parameteric resubstitution entropy estimator via the von Mises-Fisher kernel density [1, 41, 101]. Thus, plugging it into Eq. 3 and Eq. 4 can relieve the issue of approximating $I(\widehat{G}_c; \widetilde{G}_c|Y)$ in practice.

To implement $I(\widehat{G}_s; Y)$ given the constraint $I(\widehat{G}_s; Y) \leq I(\widehat{G}_c; Y)$ in CIGAv2, a practical choice is to adopt hinge loss that implement the constrained $I(\widehat{G}_s; Y)$ as $\frac{1}{N} R_{\widehat{G}_s} \cdot \mathbb{I}(R_{\widehat{G}_c} \leq R_{\widehat{G}_s})$, where N is the number of samples, $\mathbb{I}$ is an indicator function that outputs 1 when the inner condition is satisfied otherwise 0, and $R_{\widehat{G}_s}$ and $R_{\widehat{G}_c}$ are the empirical risk vector of the predictions for each sample based on the corresponding $\widehat{G}_s$ and $\widehat{G}_c$. More implementation details can be found in Appendix F.

Discussions and implications of CIGA. Although using contrastive learning to improve OOD generalization is not new in the literature [27, 61, 124], previous methods cannot yield OOD guarantees in graph circumstances due to the highly non-linearity and the unavailability of domain labels E. In particular, CIGA can *be reduced to directly applying contrastive learning* when without the decomposition for causal algorithmic alignment. However, in the experiments we found that merely using the contrastive objective, i.e., CNC [124], yields unsatisfactory OOD generalization performance, which further implies the necessity of the decomposition in CIGA.

Moreover, the architecture of CIGA can have multiple other implementations for both the featurizer and classifier, such as identifying G_c at the latent space [86, 87]. Since we cannot enumerate every possible implementation, in this work we choose interpretable GNN architectures as a prototype validation for CIGA and leave more sophisticated architectures as future works. In particular, when optimized with ERM objective, CIGA can *be reduced to interpretable GNNs*. However, merely using interpretable GNNs such as ASAP [78], GIB [120] or DIR [104] cannot yield satisfactory OOD performance. As shown in Table 1(b) and discussed in Appendix. D.4, GIB can only work for FIIF, while DIR *cannot* yield OOD guarantees for neither FIIF and PIIF SCMs. These results are also empirically validated in the experiments. We provide more detailed discussions in Appendix B.

4 Empirical Studies

We conduct extensive experiments with 16 datasets to verify the effectiveness of CIGA.

Datasets. We use the SPMotif datasets from DIR [104] where artificial structural shifts and graph size shifts are nested (SPMotif-Struc). Besides, we construct a harder version mixed with attribute shifts (SPMotif-Mixed). To examine CIGA in real-world scenarios with more complicated relationships and distribution shifts, we also use DrugOOD [40] from AI-aided Drug Discovery with Assay, Scaffold, and Size splits, convert the ColoredMNIST from IRM [4] using the algorithm from Knyazev et al. [46] to inject attribute shifts, and split Graph-SST [122] to inject degree biases. To compare with previous specialized OOD methods for graph size shifts [113, 11], we use the datasets in Bevilacqua et al. [11] that are converted from TU benchmarks [67]. More details can be found in Appendix G.1.

Baselines and our methods. Besides the ERM, we also compare with SOTA interpretable GNNs, GIB [120], ASAP Pooling [78], and DIR [104], to validate the effectiveness of the optimization

Table 2: OOD generalization performance on complex distribution shifts for real-world graphs.

DATASETS	DRUG-ASSAY	DRUG-SCA	DRUG-SIZE	CMNIST-SP	GRAPH-SST5	TWITTER	AVG (RANK)†
ERM	71.79 (0.27)	68.85 (0.62)	66.70 (1.08)	13.96 (5.48)	43.89 (1.73)	60.81 (2.05)	54.33 (6.00)
ASAP	70.51 (1.93)	66.19 (0.94)	64.12 (0.67)	10.23 (0.51)	44.16 (1.36)	60.68 (2.10)	52.65 (8.33)
GIB	63.01 (1.16)	62.01 (1.41)	55.50 (1.42)	15.40 (3.91)	38.64 (4.52)	48.08 (2.27)	47.11 (10.0)
DIR	68.25 (1.40)	63.91 (1.36)	60.40 (1.42)	15.50 (8.65)	41.12 (1.96)	59.85 (2.98)	51.51 (9.33)
IRM	72.12 (0.49)	68.69 (0.65)	66.54 (0.42)	31.58 (9.52)	43.69 (1.26)	63.50 (1.23)	57.69 (4.50)
V-REX	72.05 (1.25)	68.92 (0.98)	66.33 (0.74)	10.29 (0.46)	43.28 (0.52)	63.21 (1.57)	54.01 (6.17)
EIIL	72.60 (0.47)	68.45 (0.53)	66.38 (0.66)	30.04 (10.9)	42.98 (1.03)	62.76 (1.72)	57.20 (5.33)
IB-IRM	72.50 (0.49)	68.50 (0.40)	66.64 (0.28)	**39.86 (10.5)**	40.85 (2.08)	61.26 (1.20)	58.27 (5.33)
CNC	72.40 (0.46)	67.24 (0.90)	65.79 (0.80)	12.21 (3.85)	42.78 (1.53)	61.03 (2.49)	53.56 (7.50)
CIGAv1	**72.71 (0.52)**	**69.04 (0.86)**	**67.24 (0.88)**	19.77 (17.1)	**44.71 (1.14)**	**63.66 (0.84)**	**56.19 (2.50)**
CIGAv2	**73.17 (0.39)**	**69.70 (0.27)**	**67.78 (0.76)**	**44.91 (4.31)**	**45.25 (1.27)**	**64.45 (1.99)**	**60.88 (1.00)**
ORACLE (IID)	85.56 (1.44)	84.71 (1.60)	85.83 (1.31)	62.13 (0.43)	48.18 (1.00)	64.21 (1.77)	

†Averaged rank is also reported in the blankets because of dataset heterogeneity. Lower rank is better.

objective in CIGA. We use the same selection ratio (i.e., s_c) for all models. Moreover, to validate the effectiveness of the decomposition in CIGA, we compare CIGA with SOTA OOD objectives including IRM [4], v-Rex [49] and IB-IRM [2], for which we apply random environment partitions following [23]. We also compare CIGA with EIIL [23] and CNC [124] that do not require environment labels, where CNC [124] has a more sophisticated contrastive sampling strategy for combating subpopulation shifts. More implementation and comparison details are deferred to Appendix G.2.

Evaluation. We report the classification accuracy for all datasets, except for DrugOOD datasets where we use ROC-AUC following [40], and for TU datasets where we use Matthews correlation coefficient following [11]. We repeat the evaluation multiple times, select models based on the validation performances, and report the mean and standard deviation of the corresponding metric. For each dataset, we also report the "Oracle" performances that run ERM on the randomly shuffled data.

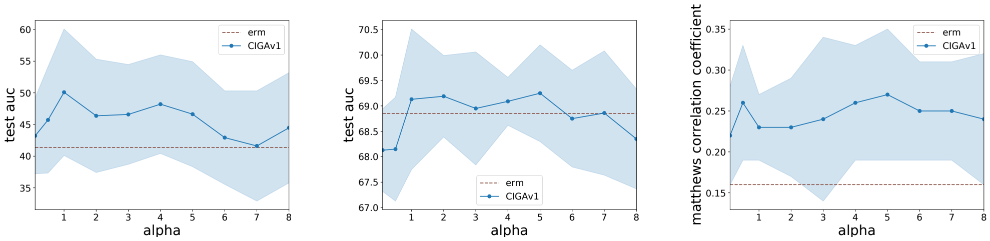
OOD generalization performance on structure and mixed shifts. In Table 1, we report the test accuracy of each method, where we omit GIB due to its poor convergence. Different biases indicate different strengths of the distribution shifts. Although the training accuracy of most methods converges to more than 99%, the test accuracy decreases dramatically as the bias increases and as more distribution shifts are mixed, which concurs with our discussions in Sec. 2.3 and Appendix D. Due to the simplicity of the task as well as the relatively high support overlap between training and test distributions, interpretable GNNs and OOD objectives can improve certain OOD performance, while they can have *high variance* since they donot have OOD generalization guarantees. In contrast, CIGAv1 and CIGAv2 outperform all of the baselines by a significant margin up to 10% with *lower variance*, which demonstrates the effectiveness and excellent OOD generalization ability of CIGA.

OOD generalization performance on realistic shifts. In Table 2 and Table 3, we examine the effectiveness of CIGA in real-world data and more complicated distribution shifts. Both averaged accuracy and ranks are reported because of the dataset heterogeneity. Since the tasks are harder than synthetic ones, interpretable GNNs and OOD objectives perform similar to or even under-perform the ERM baselines, which is also consistent to the observations in non-linear benchmarks [34, 40]. However, both CIGAv1 and CIGAv2 consistently and significantly outperform previous methods, including previous specialized methods Γ GNNs [11] for combating graph size shifts, demonstrating the generality and superiority of CIGA.

Table 3: OOD generalization performance on graph size shifts for real-world graphs in terms of Matthews correlation coefficient.

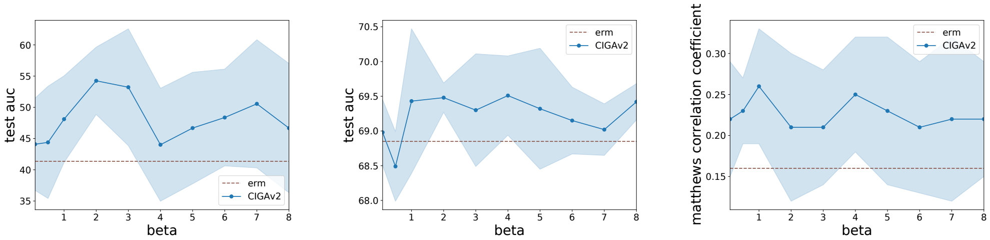
DATASETS	NCI1	NCI109	PROTEINS	DD	AVG
ERM	0.15 (0.05)	0.16 (0.02)	0.22 (0.09)	0.27 (0.09)	0.20
ASAP	0.16 (0.10)	0.15 (0.07)	0.22 (0.16)	0.21 (0.08)	0.19
GIB	0.13 (0.10)	0.16 (0.02)	0.19 (0.08)	0.01 (0.18)	0.12
DIR	0.21 (0.06)	0.13 (0.05)	0.25 (0.14)	0.20 (0.10)	0.20
IRM	0.17 (0.02)	0.14 (0.01)	0.21 (0.09)	0.22 (0.08)	0.19
V-REX	0.15 (0.04)	0.15 (0.04)	0.22 (0.06)	0.21 (0.07)	0.18
EIIL	0.14 (0.03)	0.16 (0.02)	0.20 (0.05)	0.23 (0.10)	0.19
IB-IRM	0.12 (0.04)	0.15 (0.06)	0.21 (0.06)	0.15 (0.13)	0.16
CNC	0.16 (0.04)	0.16 (0.04)	0.19 (0.08)	0.27 (0.13)	0.20
WL KERNEL	**0.39 (0.00)**	0.21 (0.00)	0.00 (0.00)	0.00 (0.00)	0.15
GC KERNEL	0.02 (0.00)	0.00 (0.00)	0.29 (0.00)	0.00 (0.00)	0.08
$\Gamma_{\text{1-HOT}}$	0.17 (0.08)	**0.25 (0.06)**	0.12 (0.09)	0.23 (0.08)	0.19
Γ_{GIN}	0.24 (0.04)	0.18 (0.04)	0.29 (0.11)	**0.28 (0.06)**	0.25
Γ_{RPGIN}	0.26 (0.05)	0.20 (0.04)	0.25 (0.12)	0.20 (0.05)	0.23
CIGAv1	0.22 (0.07)	**0.23 (0.09)**	**0.40 (0.06)**	**0.29 (0.08)**	**0.29**
CIGAv2	**0.27 (0.07)**	0.22 (0.05)	**0.31 (0.12)**	0.26 (0.08)	**0.27**
ORACLE (IID)	0.32 (0.05)	0.37 (0.06)	0.39 (0.09)	0.33 (0.05)	

Comparisons with advanced ablation variants. As discussed in Sec. 3.3, CIGA can be reduced to interpretable GNNs and contrastive learning approaches. However, across all experiments, we can observe that neither the advanced interpretable GNNs (DIR) nor sophisticated contrastive objectives with specialized sampling strategy (CNC) can yield satisfactory OOD performance, which serves as *strong evidence* for the necessities of the decomposition as well as the objective in CIGA. Furthermore, although CIGAv1 can outperform CIGAv2 when we may have a relatively accurate



(a) SPMotif-Mixed (bias=0.9) (b) DrugOOD-Scaffold (c) NCI109

Figure 4: Hyperparameter sensitivity analysis on the coefficient of contrastive loss (α).



(a) SPMotif-Mixed (bias=0.9, α=4) (b) DrugOOD-Scaffold (α=1) (c) NCI109 (α=1)

Figure 5: Hyperparameter sensitivity analysis on the coefficient of hinge loss (β).

s_c, the improvements in CIGAv1 are not as stable as CIGAv2 or even unsatisfactory when the assumption is violated. This phenomenon also reveals the superiority of CIGAv2 in practice.

Hyperparameter sensitivity analysis. To examine how sensitive CIGA is to the hyperparamters α and β for contrastive loss and hinge loss, respectively. We conduct experiments based on the hardest datasets from each table (i.e., SPMotif-Mixed with the bias of 0.9, DrugOOD-Scaffold and the NCI109 datasets from Table 1, Table 2, and Table 3, respectively.) with different α and β. When changing the value of β, we fix the α to a specific value under which the model has a relatively good performance (but not the best, to fully examine the robustness of CIGA in practice).

The results are shown in Fig. 4 and Fig. 5. It can be found that both CIGAv1 and CIGAv2 are robust to different values of α and β, respectively, across different datasets and distribution shifts. Besides, the results also reflect the effects of the additional penalty terms in CIGA. For example, in Fig. 16, when α is too small, the invariance of the identified invariant subgraphs $\widehat{G}_c$ may not be guaranteed, resulting worse performances. Similarly, as shown in Fig. 17, when β becomes too small, some part of the spurious subgraph may still appear in the estimated invariant subgraphs, which yields worse performances. Besides, when α and β become too large, the optimization of CIGA can be affected due to their intrinsic conflicts with ERM, hence a better optimization scheme for CIGA can be a promising future direction [18]. We provide more details and additional analysis on the efficiency of CIGA and single environment OOD generalization performance of CIGA in Appendix G.4, as well as the visualization examples of the identified invariant subgraph in Appendix G.5.

5 Conclusions

We studied the OOD generalization on graphs via graph classification, and propose a new solution CIGA through the lens of causality. By modeling potential distribution shifts on graphs with SCMs, we generalized and instantiated the invariance principle to graphs, which was shown to have promising theoretical and empirical OOD generalization ability under a variety of distribution shifts.

Acknowledgments and Disclosure of Funding

We thank the reviewers for their valuable comments. This work was supported by GRF 14208318 from the RGC of HKSAR and CUHK direct grant 4055146. TL was partially supported by Australian Research Council Projects DP180103424, DE-190101473, IC-190100031, DP-220102121, and FT-220100318. YZ and BH were supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202, Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652, and Tencent AI Lab Rhino-Bird Gift Fund.

References

[1] I. Ahmad and P.-E. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

[2] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2021.

[3] A. A. Alemi, I. Fischer, and J. V. D. and. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint*, arXiv:1907.02893, 2019.

[5] B. Aubin, A. Słowik, M. Arjovsky, L. Bottou, and D. Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

[6] P. W. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016.

[7] S. Beery, G. V. Horn, and P. Perona. Recognition in terra incognita. In *Computer Vision European Conference, Part XVI*, volume 11220, pages 472–489, 2018.

[8] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, volume 80, pages 531–540, 10–15 Jul 2018.

[9] A. Bellot and M. van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint*, arXiv:2007.10653, 2020.

[10] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

[11] B. Bevilacqua, Y. Zhou, and B. Ribeiro. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*, volume 139, pages 837–851, 18–24 Jul 2021.

[12] R. S. Bohacek, C. McMartin, and W. C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, 1996.

[13] D. Burshtein, V. D. Pietra, D. Kanevsky, and A. Nadas. Minimum impurity partitions. *The Annals of Statistics*, 20(3):1637–1646, 1992.

[14] S. Chang, Y. Zhang, M. Yu, and T. S. Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, volume 119, pages 1448–1458, 2020.

[15] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

[16] P.-H. Chen, C.-J. Lin, and B. Schölkopf. A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, 21(2):111–136, 2005.

[17] Y. Chen, H. Yang, Y. Zhang, K. Ma, T. Liu, B. Han, and J. Cheng. Understanding and improving graph injection attack by promoting unnoticeability. In *International Conference on Learning Representations*, 2022.

[18] Y. Chen, K. Zhou, Y. Bian, B. Xie, K. Ma, Y. Zhang, H. Yang, B. Han, and J. Cheng. Pareto invariant risk minimization. *arXiv preprint*, arXiv:2206.07766, 2022.

[19] Z. Chen, L. Chen, S. Villar, and J. Bruna. Can graph neural networks count substructures? In *Advances in Neural Information Processing Systems*, 2020.

[20] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546, 2005.

[21] C. Chuang, A. Torralba, and S. Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *International Conference on Machine Learning*, volume 119, pages 1984–1994. PMLR, 2020.

[22] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

[23] E. Creager, J. Jacobsen, and R. S. Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, volume 139, pages 2189–2200, 2021.

[24] A. J. DeGrave, J. D. Janizek, and S. Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[26] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Annual Meeting of the Association for Computational Linguistics*, pages 49–54, 2014.

[27] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019.

[28] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson. Benchmarking graph neural networks. *arXiv preprint*, arXiv:2003.00982, 2020.

[29] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[30] O.-E. Ganea, X. Huang, C. Bunne, Y. Bian, R. Barzilay, T. S. Jaakkola, and A. Krause. Independent SE(3)-equivariant models for end-to-end rigid protein docking. In *International Conference on Learning Representations*, 2022.

[31] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Mache Learning Research*, 17:59:1–59:35, 2016.

[32] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. E. Peters, M. Schmitz, and L. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *arXiv preprint*, arXiv:1803.07640, 2018.

[33] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[34] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[35] W. L. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[36] K. Han, B. Lakshminarayanan, and J. Z. Liu. Reliable graph neural networks for drug discovery under distributional shift. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[37] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, 2020.

[38] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. H. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[39] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37, pages 448–456, 2015.

[40] Y. Ji, L. Zhang, J. Wu, B. Wu, L.-K. Huang, T. Xu, Y. Rong, L. Li, J. Ren, D. Xue, H. Lai, S. Xu, J. Feng, W. Liu, P. Luo, S. Zhou, J. Huang, P. Zhao, and Y. Bian. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery – A Focus on Affinity Prediction Problems with Noise Annotations. *arXiv preprint*, arXiv:2201.09637, 2022.

[41] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and j. m. robins. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[42] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.

[43] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[44] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint*, arXiv:1611.07308, 2016.

[45] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[46] B. Knyazev, G. W. Taylor, and M. R. Amer. Understanding attention and generalization in graph neural networks. In *Advances in Neural Information Processing Systems*, pages 4204–4214, 2019.

[47] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning,*, pages 5637–5664, 2021.

[48] M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint*, arXiv:2008.01883, 2020.

[49] D. Krueger, E. Caballero, J. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826, 2021.

[50] J. V. Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, pages 16451–16467, 2021.

[51] Q. Li, Z. Han, and X. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, pages 3538–3545, 2018.

[52] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, pages 647–663, 2018.

[53] W. Lin, H. Lan, and B. Li. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning,*, pages 6666–6679, 2021.

[54] Y. Lin, H. Dong, H. Wang, and T. Zhang. Bayesian invariant risk minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[55] Y. Lin, S. Zhu, and P. Cui. ZIN: when and how to learn invariance by environment inference? *arXiv preprint arXiv:2203.05818*, 2022.

[56] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

[57] L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[58] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, pages 19620–19631, 2020.

[59] Y. Luo, K. Yan, and S. Ji. Graphdf: A discrete flow model for molecular graph generation. In *International Conference on Machine Learning*, pages 7192–7203, 2021.

[60] K. Ma, H. Yang, H. Yang, T. Jin, P. Chen, Y. Chen, B. F. Kamhoua, and J. Cheng. Improving graph representation learning by contrastive regularization. *arXiv preprint*, arXiv:2101.11525, 2021.

[61] D. Mahajan, S. Tople, and A. Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324, 2021.

[62] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[63] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. D. Veij, E. Felix, M. P. Magariños, J. F. Mosquera, P. Mutowo-Meullenet, M. Nowotka, M. Gordillo-Marañón, F. M. I. Hunter, L. Junco, G. Mugumbate, M. Rodríguez-López, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(Database-Issue):D930–D940, 2019.

[64] S. Miao, M. Liu, and P. Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pages 15524–15543, 2022.

[65] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48, 1999.

[66] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, pages 4602–4609, 2019.

[67] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint*, arXiv:2007.08663, 2020.

[68] C. Morris, Y. Lipman, H. Maron, B. Rieck, N. M. Kriege, M. Grohe, M. Fey, and K. M. Borgwardt. Weisfeiler and leman go machine learning: The story so far. *arXiv preprint*, arXiv:2112.09992, 2021.

[69] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181–201, 2001.

[70] R. L. Murphy, B. Srinivasan, V. A. Rao, and B. Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, pages 4663–4673, 2019.

[71] V. Nagarajan, A. Andreassen, and B. Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[72] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

[73] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[74] J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[75] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, feb 2019. ISSN 0001-0782.

[76] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[77] J. Peters, D. Janzing, and B. Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

[78] E. Ranjan, S. Sanyal, and P. P. Talukdar. ASAP: adaptive structure aware pooling for learning hierarchical graph representations. In *AAAI Conference on Artificial Intelligence*, pages 5470–5477, 2020.

[79] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[80] E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[81] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[82] R. Salakhutdinov and G. E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, pages 412–419, 2007.

[83] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. A. Riedmiller, R. Hadsell, and P. W. Battaglia. Graph networks as learnable physics engines for inference and control. In *International Conference on Machine Learning*, pages 4467–4476, 2018.

[84] A. Santoro, F. Hill, D. G. T. Barrett, A. S. Morcos, and T. P. Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, volume 80, pages 4477–4486, 2018.

[85] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019.

[86] B. Schölkopf. Causality for machine learning. *arXiv preprint*, arXiv:1911.10500, 2019.

[87] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[88] O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 525–536, 2018.

[89] T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. In *Journal of Classification*, volume 14, pages 75–100, 1997.

[90] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[91] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[92] T. Sterling and J. J. Irwin. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

[93] B. Sun and K. Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450, 2016.

[94] H. Tang, Z. Huang, J. Gu, B. Lu, and H. Su. Towards scale-invariant graph-related problem solving by iterative homogeneous gnns. In *Advances in Neural Information Processing Systems*, 2020.

[95] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[96] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint*, arXiv:1807.03748, 2018.

[97] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1991.

[98] P. Velickovic, R. Ying, M. Padovano, R. Hadsell, and C. Blundell. Neural execution of graph algorithms. In *International Conference on Learning Representations*, 2020.

[99] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[100] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: A survey on domain generalization. In *International Joint Conference on Artificial Intelligence*, pages 4627–4635, 2021.

[101] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939, 2020.

[102] B. Wu, J. Li, C. Hou, G. Fu, Y. Bian, L. Chen, and J. Huang. Recent advances in reliable deep graph learning: Adversarial attack, inherent noise, and distribution shift. *arXiv preprint arXiv:2202.07114*, 2022.

[103] Q. Wu, H. Zhang, J. Yan, and D. Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022.

[104] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022.

[105] L.-P. A. C. Xhonneux, A. Deac, P. Veličković, and J. Tang. How to transfer algorithmic reasoning knowledge to learn new algorithms? In *Advances in Neural Information Processing Systems*, pages 19500–19512, 2021.

[106] K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, and S. Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pages 5449–5458, 2018.

[107] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[108] K. Xu, J. Li, M. Zhang, S. S. Du, K. Kawarabayashi, and S. Jegelka. What can neural networks reason about? In *International Conference on Learning Representations*, 2020.

[109] K. Xu, M. Zhang, S. Jegelka, and K. Kawaguchi. Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. In *International Conference on Machine Learning*, pages 11592–11602, 2021.

[110] K. Xu, M. Zhang, J. Li, S. S. Du, K. Kawarabayashi, and S. Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021.

[111] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

[112] H. Yang, K. Ma, and J. Cheng. Rethinking graph regularization for graph neural networks. In *AAAI Conference on Artificial Intelligence*, pages 4573–4581, 2021.

[113] G. Yehudai, E. Fetaya, E. Meirom, G. Chechik, and H. Maron. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pages 11975–11986, 2021.

[114] R. Yeung. *Information Theory and Network Coding*. 01 2008. ISBN 978-0-387-79233-0.

[115] Z. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, pages 4805–4815, 2018.

[116] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, pages 9240–9251, 2019.

[117] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International Conference on Machine Learning*, pages 5694–5703, 2018.

[118] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, pages 5812–5823, 2020.

[119] Y. You, T. Chen, Y. Shen, and Z. Wang. Graph contrastive learning automated. In *International Conference on Machine Learning*, pages 12121–12132, 2021.

[120] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021.

[121] J. Yu, J. Liang, and R. He. Finding diverse and predictable subgraphs for graph domain generalization. *arXiv preprint*, arXiv:2206.09345, 2022.

[122] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint*, arXiv:2012.15445, 2020.

[123] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[124] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint*, arXiv:2203.01517, 2022.

[125] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, and K. Zhang. Adversarial robustness through the lens of causality. In *International Conference on Learning Representations*, 2022.

[126] H. Zhao, R. T. des Combes, K. Zhang, and G. J. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

[127] X. Zhou, Y. Lin, W. Zhang, and T. Zhang. Sparse invariant risk minimization. In *39th International Conference on Machine Learning*, pages 27222–27244, 2022.

[128] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. B.4 in the appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. A in the appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sec. C and Sec. E in the appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sec. E in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and the other required materials are provided in `https://github.com/LFhase/CIGA`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. G.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sec. G.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sec. G.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data used are all publicly available datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We didn't conduct research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We didn't conduct research with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We didn't conduct research with human subjects.

Appendix of CIGA

Contents

A Broader Impacts

Considering the wide applications and high sensitivity of GNNs to distribution shifts and spurious correlations, it is important to develop GNNs that are able to generalize to OOD data, especially for realistic scenarios such as AI-aided Drug Discovery where OOD data are ubiquitous. By formulating OOD generalization problem on graphs using causality, our work can serve as an initiate step towards tackling OOD generalization problem on graphs, with the hope to empower GNNs for broader applications and social benefits. Besides, this paper does not raise any ethical concerns. This study does not involve any human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues.

B More Discussions on Related Work and Future Directions

B.1 More backgrounds

We give more background introduction about GNNs and Invariant Learning in this section.

Graph Neural Networks. Let $G = (A, X)$ denote a graph with n nodes and m edges, where $A \in \{0,1\}^{n\times n}$ is the adjacency matrix, and $X \in \mathbb{R}^{n\times d}$ is the node feature matrix with a node feature dimension of d. In graph classification, we are given a set of N graphs $\{G_i\}_{i=1}^N \subseteq \mathcal{G}$ and their labels $\{Y_i\}_{i=1}^N \subseteq \mathcal{Y} = \mathbb{R}^c$ from c classes. Then, we train a GNN $\rho \circ h$ with an encoder $h : \mathcal{G} \to \mathbb{R}^h$ that learns a meaningful representation h_G for each graph G to help predict their labels $y_G = \rho(h_G)$ with a downstream classifier $\rho : \mathbb{R}^h \to \mathcal{Y}$. The representation h_G is typically obtained by performing pooling with a READOUT function on the learned node representations:

$$h_G = \text{READOUT}(\{h_u^{(K)} | u \in V\}), \tag{6}$$

where the READOUT is a permutation invariant function (e.g., SUM, MEAN) [107, 115, 70, 107, 19, 68], and $h_u^{(K)}$ stands for the node representation of $u \in V$ at K-th layer that is obtained by neighbor aggregation:

$$h_u^{(K)} = \sigma(W_K \cdot a(\{h_v^{(K-1)}\} | v \in \mathcal{N}(u) \cup \{u\})), \tag{7}$$

where $\mathcal{N}(u)$ is the set of neighbors of node u, $\sigma(\cdot)$ is an activation function, e.g., ReLU, and $a(\cdot)$ is an aggregation function over neighbors, e.g., MEAN.

Invariant Learning. Invariant learning typically considers a supervised learning setting based on the data $\mathcal{D} = \{\mathcal{D}^e\}_e$ collected from multiple environments $\mathcal{E}_{\text{all}}$, where $\mathcal{D}^e = \{G_i^e, y_i^e\}$ is the dataset from environment $e \in \mathcal{E}_{\text{all}}$. (G_i^e, y_i^e) from a single environment e are considered as drawn independently from an identical distribution $\mathbb{P}^e$. The goal of OOD generalization is to train a GNN $\rho \circ h : \mathcal{G} \to \mathcal{Y}$ with data from training environments $\mathcal{D}_{\text{tr}} = \{\mathcal{D}^e\}_{e \in \mathcal{E}_{\text{tr}} \subseteq \mathcal{E}_{\text{all}}}$, and generalize well to all (unseen) environments, i.e., to minimize:

$$\min_{\rho, h} \max_{e \in \mathcal{E}_{\text{all}}} R^e(\rho \circ h), \tag{8}$$

where R^e is the empirical risk under environment e [97, 76, 4]. More details can be referred in [2].

B.2 Detailed related work

GNN Explainability. Works in GNN explainability aim to find a subgraph of the input graph as the explanation for the prediction of a GNN model [116, 122]. Although some may leverage causality in explanation generation [53], they mostly focus on understanding the predictions of GNNs in a post-hoc manner instead of OOD generalization. Recently there are two works aiming to provide robust explanations under distribution shifts, i.e., GIB [120] and DIR [104], and both of them focus on tackling FIIF spurious correlations (Assumption C.2). The theoretical guarantees of GIB follows the theory of information bottleneck [95], while GIB can not solve PIIF spurious correlations (Assumption C.3). As both FIIF and PIIF widely exist in realistic scenarios, failing to solve either of them could result in severe performance degradation in practice [4, 2, 5, 71]. While for DIR, though as a generalization of Chang et al. [14] to graphs, can not provide any theoretical guarantees under FIIF spurious correlations as shown in Appendix D.4, nor under PIIF spurious correlations.

GNN Extrapolation. Recently there is a surge of attention in improving the extrapolation ability of GNNs and apply them to various applications, such as mathematical reasoning [84, 85], physics [6, 83], and graph algorithms [94, 98, 108, 105]. Xu et al. [110] study the neural network extrapolation ability from a geometrical perspective. Han et al. [36] improve OOD drug discovery by mitigating the overconfident misprediction issue. Knyazev et al. [46], Yehudai et al. [113] focus on the extrapolation of GNNs in terms of graph sizes, while making additional assumptions on the knowledge about ground truth attentions and access to test inputs. Bevilacqua et al. [11] study the graph size extrapolation problem of GNNs through a causal lens, while the induced invariance principle is built upon assumptions on the specific family of graphs. Different from these works, we consider the GNN extrapolation as a causal problem, establish generic SCMs that are compatible with several graph generation models, as well as, more importantly, different types of distribution shifts. Hence, the induced the invariance principle and provable algorithms built upon the SCMs in our work can generalize to multiple graph families and distribution shifts.

Additionally, Wu et al. [103] propose causal models as well as specialized objectives to extrapolate nodes with different neighbors. However, their formulation is limited to node classification task and specific spurious correlation type. In contrast, the induced invariance principle in Wu et al. [103], can be seen as a extension of CIGA for node classification, where we cab identify an invariant subgraph from the K-hop neighbor graph of each node, and making predictions based on it, i.e., $Y \perp\!\!\!\perp E | G_c^{\text{ego}} \subseteq G_u^{\text{ego}}$ for node u. We leave specific formulation and implementation to future works.

Causality and OOD Generalization. Causality comes to the stage for demystifying and improving the huge success of machine learning algorithms to further advances [75, 86, 87]. One of the most widely applied concept from causality is the Independent Causal Mechanism (ICM) that assumes conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions [74, 77]. The invariance principle is also induced from the ICM assumption. Once proper assumptions about the underlying data generation process via Structural Causal Models (SCM) are established, it is promising to apply the invariance principle to machine learning models for finding an invariant representation about the causal relationship between the underlying causes and the label [76, 4]. Consequently, models built upon the invariant representation can generalize to unseen environments or domains with guaranteed performance [76, 79, 4, 81, 10, 48, 34, 49, 23, 2]. The arguably first formulation of invariance principle was introduced by Peters et al. [76]. Arjovsky et al. [4] propose a novel formulation of learning causal invariance in representation learning, i.e., IRM, show how it connects with existing areas such as distributional robust optimization [72] and generalization [123], and prove its effectiveness in addressing PIIF spurious correlations (Assumption C.3). However, in practice, both PIIF and FIIF (Assumption C.2) can appear in data, while IRM can fail in these cases [5, 71]. Ahuja et al. [2] then propose to add information bottleneck criteria into the IRM formulation to address the issue. However, their results are restricted to linear regime and also require environment partitions to distinguish the sources of distribution shifts. Recently, Creager et al. [23] and Lin et al. [55] propose new OOD objectives to relieve the needs for environment partitions, but limited to PIIF spurious types and linear regime. Besides, Lin et al. [54] identify the overfitting problem as a key challenge when applying IRM on large neural networks. Zhou et al. [127] propose to alleviate this problem by imposing sparsity constrain.

In parallel invariant learning approaches, Sagawa* et al. [81] propose to regularize the worst group in group distributionally robust optimization (GroupDro). Zhang et al. [124] propose a contrastive approach to tackle GroupDro when the group partitions are not available. However, minimizing the gap between worst group risk and averaged risk can not yield a OOD generalizable predictors in our circumstances. Besides, traditional approaches to tackle OOD generalization also include Domain Adaption, Transfer Learning and Domain Generalization[79, 21, 31, 93, 52, 27, 61, 100], which aim to learn the class conditional invariant representation shared across source domain and target domain. However, they all require a stronger assumption on the availability of target domain data or the ground truth predictors [34, 2], hence are not able to yield predictors with OOD generalization guarantees. We refer interested readers to Pearl [75], Schölkopf [86], Schölkopf et al. [87] for an in-depth understanding, and Gulrajani and Lopez-Paz [34], Ahuja et al. [2] for a thorough overview.

B.3 More discussions on connections of CIGA with existing work

Although primarily serving for graph OOD generalization problem, our theory complements the identifiability study on graphs through contrastive learning, and aligns with the discoveries in the image domain that contrastive learning learns to isolate the content (C) and style (S) [128, 50]. Moreover, our results also partially explain the success of graph contrastive learning [118, 60, 119], where GNNs may implicitly learn to identify the underlying invariant subgraphs for prediction.

On expressivity of graph encoder in CIGA. The expressivity of CIGA is essentially constrained by the encoders embedded for learning graph representations. During isolating G_c from G, if the encoder can not differentiate two isomorphic graphs G_c and $G_c \cup G_s^p$ where $G_s^p \subseteq G_s$, then the featurizer will fail to identify the underlying invariant subgraph. Moreover, the classifier will also fail if the encoder can not differentiate two non-isomorphic G_cs from different classes. Thus, adopting more powerful graph representation encoders into CIGA can improve the OOD generalization.

On CIGA and graph information bottleneck. Under the FIIF assumption on latent interaction, the independence condition derived from causal model can also be rewritten as $Y \perp\!\!\!\perp S|C$ (similar to that in DIR [104] as they also focus on FIIF), which further implies $Y \perp\!\!\!\perp S|\widehat{G}_c$. Hence it is natural to use Information Bottleneck (IB) objective [95] to solve for G_c:

$$\begin{aligned} &\min_{f_c, g} R_{G_c}(f_c(\widehat{G}_c)), \\ &\text{s.t. } G_c = \underset{\widehat{G}_c = g(G) \subseteq G}{\arg\max}\; I(\widehat{G}_c, Y) - I(\widehat{G}_c, \mathcal{G}), \end{aligned} \tag{9}$$

which explains the success of many existing works in finding predictive subgraph through IB [120]. However, the estimation of $I(\widehat{G}_c, G)$ is notoriously difficult due to the complexity of graph, which can lead to unstable convergence as observed in our experiments. In contrast, optimization with contrastive objective in CIGA as Eq. 5 induces more stable convergence.

On CIGA for node classifications. As the task of node classification can be viewed as graph classification based on the ego-graphs of a node, our analysis and discoveries can generalize to node classification. More specifically, the invariance principle for node classification can be implemented by identifying an invariant subgraph from the K-hop neighbor graph of each node, and making predictions based on it, i.e., $Y \perp\!\!\!\perp E|G_c^{\text{ego}} \subseteq G_u^{\text{ego}}$ for node u [103].

B.4 Discussions on limitations of CIGA and future directions

Better graph generation modeling. Compared to Bevilacqua et al. [11], we do not specify a specific graph family in the SCM for graph generation process. Since our focus is to describe the potential distribution shifts with SCMs, in Assumption 2.1, we aim to build a SCM that is compatible to many graph generation processes [89, 57, 117, 59]. However, it is often the case that practitioners have certain inductive knowledge about the graph generation process, which may imply useful leads and invariance in modeling the generation process [111, 30, 56]. In Appendix C.1, we provide an example about incorporating the graphon [57] knowledge into the SCMs, which derives similar solutions as in the literature [113, 11]. Therefore, we believe it is promising to leverage more additional knowledge for more precise graph generation modeling and better OOD generalization on graphs.

Better contrastive sampling. Typical contrastive or graph contrastive learning approaches leverage augmentation techniques as well as sophisticated sampling strategies during the positive or negative pairs selection [20, 82, 96, 118, 119]. A better augmentation or sampling strategy can benefit the OOD generalization in general as shown by Kügelgen et al. [50] and Zhang et al. [124]. Since our implementation of CIGA in this work aims to verify the theoretical findings, we do not apply sophisticated augmentation or sampling during the sampling while simply using the supervised contrastive approach [42]. Nevertheless, it is promising to leverage better augmentation and contrastive strategy to improve the generalization ability in CIGA [121].

More sophisticated architectures/parameter tunning. The CIGA framework introduced in Sec. 3 can have multiple implementations. We choose interpretable architectures in our experiments for the purpose of concept verification. Essentially, different architectures can have different advantages and limitations. For the interpretable GNNs used in our experiments, it can provide interpretability for the results (as shown in Appendix G.5), but still requires more training time (as shown in Appendix G.4).

Therefore, it may not be applicable to some resource-limited scenarios such as Edge-AI. Besides, the approximation may also be limited to the chosen architectures. More sophisticated architectures can be incorporated, such as identifying and disentangling G_c at the latent space [86, 87]. Moreover, as shown in Appendix G.4, CIGA still requires certain additional tunning efforts for the objectives. Hence we believe it is also a promising future direction to reduce the parameter tunning by leveraging better optimization techiniques [88, 18]

C Full Structural Causal Models on Graph Generation

Due to the space constraints in the main paper, we make some simplifications when giving the SCMs on the graph generation process. Hence in this section, supplementary to the graph generation process in Sec. 2.2, we provide full SCMs on the graph generation process in this section as shown in Fig. 6. Formal descriptions are given as Assumptions C.1, C.2, C.3, C.4.

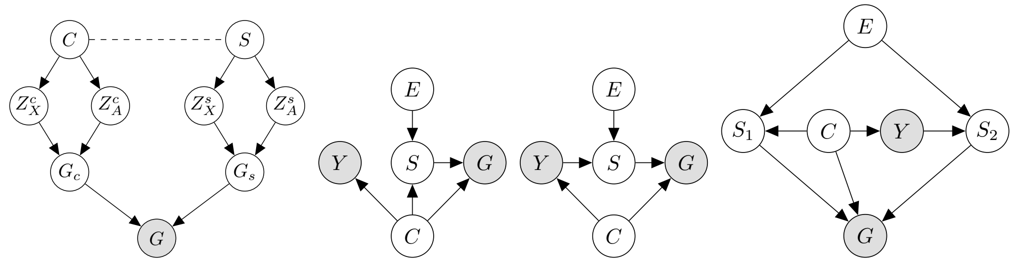
To begin with, we take a latent-variable model perspective on the graph generation process and assume that the graph is generated through a mapping $f_{\text{gen}} : \mathcal{Z} \rightarrow \mathcal{G}$, where $\mathcal{Z} \subseteq \mathbb{R}^n$ is the latent space and $\mathcal{G} = \cup_{N=1}^{\infty}\{0,1\}^N \times \mathbb{R}^{N\times d}$ is the graph space. Let E denote environments. Following previous works [50, 2], we partition the latent variable from $\mathcal{Z}$ into an invariant part $C \in \mathcal{C} = \mathbb{R}^{n_c}$ and a varying part $S \in \mathcal{S} = \mathbb{R}^{n_s}$, s.t., $n = n_c + n_s$, according to whether they are affected by E. Similarly in images, C and S can represent content and style while E can refer to the locations where the images are taken [7, 125, 50]. While in graphs, C can be the latent variable that controls the generation of functional groups in a molecule, which can not be affected by the changes of environments, such as species (or scaffolds), experimental environment for examining the chemical property (or assays) [40]. On the contrary, the other latent variable S inherits environment-specific information thus can further affect the finally generated graphs. Besides, C and S can have multiple types of interactions at the latent space with environments E and labels Y, which will generate different types of spurious correlations [2].

Assumption C.1 (Graph generation SCM)**.**

$$\begin{aligned}
&(Z_A^c, Z_X^c) := f_{\text{gen}}^{(A,X)^c}(C),\ G_c := f_{\text{gen}}^{G_c}(Z_A^c, Z_X^c),\\
&(Z_A^s, Z_X^s) := f_{\text{gen}}^{(A,X)^s}(S),\ G_s := f_{\text{gen}}^{G_s}(Z_A^s, Z_X^s),\\
&G := f_{\text{gen}}^{G}(G_c, G_s).
\end{aligned}$$

Specifically, the graph generation process is shown as Fig. 6(a). The generation mapping f_{gen} is decomposed into $f_{\text{gen}}^{(A,X)^c}$,$f_{\text{gen}}^{G_c}$, $f_{\text{gen}}^{(A,X)^s}$,$f_{\text{gen}}^{G_s}$ and f_{gen}^{G} to control the generation of (Z_A^c, Z_X^c), G_c, (Z_A^s, Z_X^s), G_s, and G, respectively. Given the variable partitions C and S at the latent space $\mathcal{Z}$, they control the generation of the adjacency matrix and features for the invariant subgraph G_c and spurious subgraph G_s through two pairs of latent variables (Z_A^c, Z_X^c) and (Z_A^s, Z_X^s), respectively. Z_A^c and Z_A^s will control the structure-level properties in the generated graphs, such as degrees, sizes, and subgraph densities. While Z_X^c and Z_X^s mainly control the attribute-level properties in the generated graphs, such as homophily. Then, G_c and G_s are entangled into the observed graph G through f_{gen}^{G}. It can be a simply JOIN of a G_c with one or multiple G_s, or more complex generation processes controlled by the latent variables [89, 57, 117, 59, 11]. Note that since our focus is to describe the potential distribution shifts with SCMs, in Assumption 2.1, we aim to build a SCM that is compatible to many graph generation processes [89, 57, 117, 59]. In fact, in Appendix C.1, we showcase how our SCMs can generalize to specific graph families studied in the literature [11, 104, 103], when given more additional knowledge about the graph generation process. Nevertheless, we believe integrating specific graph generation processes and their implications to improving OOD generalization on graphs would be a promising future direction, as discussed in Appendix B.4.

Due to the correlation between E and G, graphs collected from different environments can have different structure-level properties such as degrees, graph sizes, and subgraph densities, as well as feature-level properties such as homophily [46, 113, 11, 17]. Meanwhile, all of them can spuriously correlated with the labels depending on how the underlying latent variables are interacted with each others. The interaction types can be further divided into two axiom types FIIF and PIIF, as well as the mixed one MIIF. Previous OOD methods such as GIB [120] and DIR [104] mainly focus on FIIF case, while others such as IRM [4] mainly focuses on the PIIF case. Evidences show that



(a) Graph generation SCM (b) FIIF SCM (c) PIIF SCM (d) MIIF SCM

Figure 6: Full SCMs on Graph Distribution Shifts.

failing to model either of them when developing the OOD objectives can have serious performance degenerations in practice [5, 71]. That is why we aim to model both of them in our solution.

Assumption C.2 (FIIF SCM)**.**

$$Y := f_{\text{inv}}(C),\ S := f_{\text{spu}}(C, E),\ G := f_{\text{gen}}(C, S).$$

Assumption C.3 (PIIF SCM)**.**

$$Y := f_{\text{inv}}(C),\ S := f_{\text{spu}}(Y, E),\ G := f_{\text{gen}}(C, S).$$

Assumption C.4 (MIIF SCM)**.**

$$Y := f_{\text{inv}}(C),\ S_1 := f_{\text{spu}}(C, E),\ S_2 := f_{\text{spu}}(Y, E),\ G := f_{\text{gen}}(C, S_1, S_2).$$

As for the interactions between C and S at the latent space, we categorize the interaction modes into Fully Informative Invariant Features (FIIF, Fig. 6(b)), and Partially Informative Invariant Features (PIIF, Fig. 6(c)), depending on whether the latent invariant part C is fully informative about label Y, i.e., $(S, E) \perp\!\!\!\perp Y|C$. It is also possible that FIIF and PIIF are entangled into a Mixed Informative Invariant Features (MIIF,Fig. 6(d)). We follow Arjovsky et al. [4], Ahuja et al. [2] to formulate the SCMs for FIIF and PIIF, where we omit noises for simplicity [74, 77]. Since MIIF is built upon FIIF and PIIF, we will focus on the axiom interaction modes (FIIF and PIIF) in this paper, while most of our discussions can be extended to MIIF or more complex interactions built upon FIIF and PIIF.

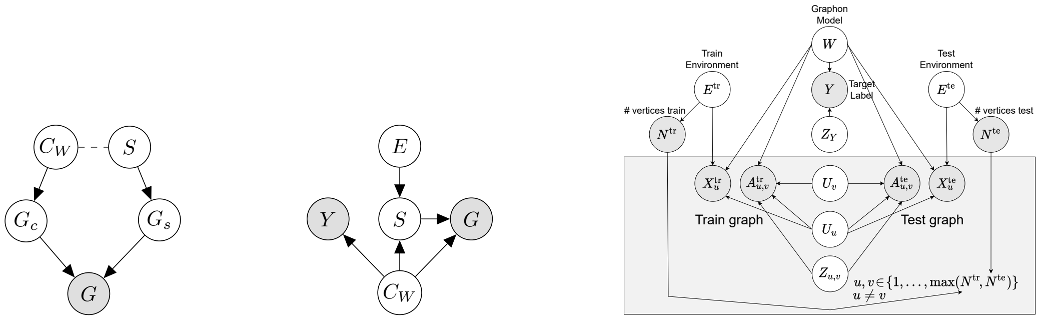
Among all of the interaction modes, f_{gen} corresponds to the graph generation process in Assumption C.1. f_{spu} is the mechanism describing how S is affected by C and E at the latent space. In FIIF, S is directly controlled by C while in PIIF, indirectly controlled by C through Y, which can exhibit different behaviors in practice [2, 71]. Additionally, in MIIF, S is further partitioned into S_1 and S_2 depending on whether it is directly or indirectly controlled by C, respectively. Moreover, $f_{\text{inv}} : \mathcal{C} \rightarrow \mathcal{Y}$ indicates the labeling process, which assigns labels Y for the corresponding G merely based on C. Consequently, $\mathcal{C}$ is better clustered than $\mathcal{S}$ when given Y [13, 15, 86, 87], which also serves as the necessary separation assumption for a classification task [69, 16, 65].

Assumption C.5 (Latent Separability)**.** $H(C|Y) \leq H(S|Y)$.

C.1 Discussions on specific cases of the SCMs

Although our primary focus in this work is to characterize general graph distribution shifts that could happen in practice without any additional knowledge about the underlying graph family, and derive the corresponding solutions, our SCMs (Fig. 6) can generalize to specific cases studied in previous works, when incorporating more inductive biases about the underlying graph family [11, 104, 103]. Specifically, we illustrate the specialized SCMs in Fig. 7 for the SCM studied in [11] which assumes the graphs are generated following a graphon model [57].

When with the additional knowledge about the underlying graph generative model, the graph generation SCM (Fig. 6(a)) and the FIIF SCM (Fig. 6(b)) together generalizes to the graphon SCM studied in [11]. We now give a brief description in the below.



(a) $\mathcal{G}$-Gen. SCM

(b) FIIF SCM

(c) Graphon SCM from [11].

Figure 7: Specialized graph generation SCMs when incorporating additional knowledge.

Specifically, shown as in Fig. 7(a), C now is instantiated as a graphon model $C_W \sim \mathbb{P}(C_W)$, where $C_W : [0,1]^2 \rightarrow [0,1]$ is a random symmetric measurable function sampled from the set of all symmetric measurable functions [57]. Besides, the label Y is determined according to C_W. Then, C_W will further control the generation of the adjcency matrix $G_c = A^c$ through graphon generative process:

$$A^c_{u,v} := \mathbb{I}(Z_{u,v} > C_W(U_u, U_v)),\ \forall u, v \in V,$$

where $Z_{u,v}$ is an independent uniform noises on $[0,1]$ for each possible edge (u,v) in the graph. Bascially, Z and U are inherited from the graphon SCM as Fig. 7(c).

On the other hand, as S does not imply any information about Y in this case, it resembles the FIIF SCM (Fig. 6(b)). In other words, $(S, E) \perp\!\!\!\perp Y | C$ still holds. Moreover, the node attributes $G_s = X^s$ are generated jointly influenced by the environment E and the graphon C_W through S:

$$X_v := f^s_{\text{gen}}(S),\ S := f_{\text{spu}}(E, C_W),$$

which resembles the attribute generation in Fig. 7(c).

Then, both G_c and G_s are concatenated together. In a simplistic case intuitively, we can regard G_c only contains the edges in G and G_s only contains the node attributes. Since the graphon model mainly controls the edge connection, the edge connection patterns, e.g., motif appearance frequency or subgraph densities, acts as a informative indicator for the label Y. In contrast, the node attributes and its numbers would be affected by the environments. A GNN model is prone to the changes of the environments if it overfits to some spurious patterns about the graph sizes or the attributes. While if the GNN model can leverage the connection patterns to make predictions, it remain invariant to the changes of environments, or the spurious patterns such as graph sizes and node attributes, which resembles the solutions derived in [113, 11]. Besides, it also partially explains why CIGA can generalize to OOD graphs studied in these works [113, 11].

In addition to the graphon SCM, essentially, the SCM studied in [104] resembles the FIIF SCM, and that of [103] resembles PIIF SCM, which also serves as partial evidence for the superiority OOD generalization performances of CIGA.

D More Details about Failure Case Studies in Sec. 2.3

In this section, we provide details on failure case studies in Sec. 2.3. We first elaborate the empirical evaluation setting where we construct a synthetic graph datasets to probe the behaviors of existing methods in OOD generalization on graphs.

D.1 More empirical details about failure case study in Sec. 2.3

To begin with, we construct 3-class synthetic datasets based on BAMotif [58] and follow Wu et al. [104] to inject spurious correlations between motif graph and base graph during the generation. In this graph classification task, the model needs to tell which motif the graph contains, e.g., “House” or “Cycle” motif, as shown in Fig. 8. We inject the distribution shifts in the training data while keeping the test data and validation data without the biases. For structure-level shifts, we introduce the artificial bias based on FIIF, where the motif and the base graph are spuriously correlated with a probability of various bias. For mixed shifts, we additionally introduced attribute-level shifts based on FIIF, where all of the node features are spuriously correlated with a probability of various bias. The number of training graphs is 600 for each class and the number of graphs in validation and test set is 200 for each class. More construction details are given in Appendix G.

For the GNN encoders, by default, we use 3-layer GCN [45] with mean readout, a hidden dimension of 64, and JK jump connections [106] at the last layer. During training, we use a batch size of 32, learning rate of $1e-3$ with Adam optimizer [43], and batch normalization between hidden layers [39]. Meanwhile, to stabilize the training, we also use dropout [91] of 0.1 and early stop the training when the validation accuracy does not increase till 5 epoch after first 20 epochs. All of the experiments are repeated 5 times, and the mean accuracy as well as variance are reported and plotted. When using IRM objective [4], as the environment partitions are not available, we generate 2 environments with random partitions.

D.2 More discussions about failure case study in Sec. 2.3

In Fig. 9, 10, 11, 12, we investigate whether existing training objectives (ERM and IRM), adding more message passing, as well as using expressive GNNs, can improve the OOD generalization ability on graphs. Here we also provide a additional discussion in complementary to the discussions on OOD generalization performance of ERM and IRM objectives in Sec. 2.3.

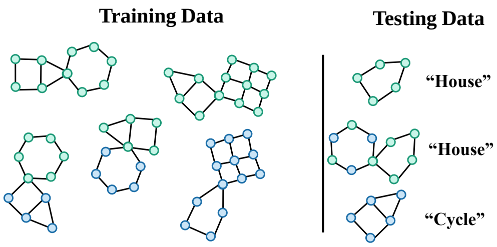


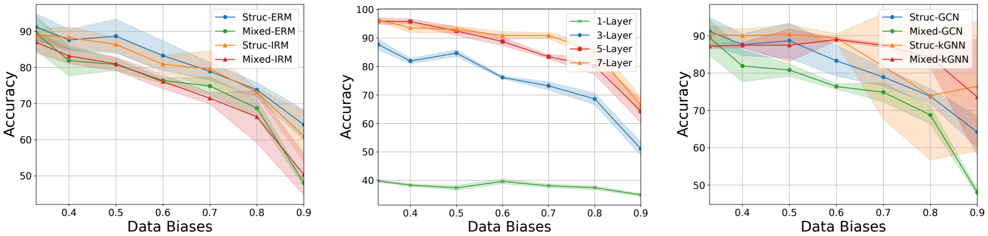
Figure 8: Failure cases of existing methods. GNNs are required to classify whether the graph contains a "house" or "cycle", where the colors represent node features. However, distribution shifts in the training exists at both structure level (From left to right: "house" mostly co-occur with a hexagon), attribute level (From upper to lower: graphs nodes are mostly green colored if they contain "house", or blued colored if they contain "cycle"), and graph sizes, making GNNs hard to capture the invariance. *ERM can fail* for leveraging the shortcuts and predict graphs that have a hexagon or have mostly green nodes as "house". *IRM can fail* when test data is not sufficiently supported by the training data.

Can better architectures improve OOD generalization of GNNs?

Adding more message passing turns. It is a common practice in GNNs to denoise the signals by aggregating more neighbors with higher layers, or enhance the expressive power with more powerful readout functions [106, 107, 112]. Aggregating neighbor information with more layers to denoise the input signal, or enhancing the expressivity with more powerful readout functions, are two common choices in GNNs to improve the generalization ability [106, 51, 107, 112]. However, in the experiments next, we empirically found that GCNs with more layers and more powerful readout operations are still sensitive to distribution shifts. In particular, stacking more layers helps denoising certain shifts, while the OOD performance would drop more sharply when the bias increases. Intuitively, if the spurious features from nodes cannot be eliminated by the denoising property of a deeper GNN, they would spread among the whole graph more widely, which in turn leads to stronger spurious correlations. Besides, the spurious correlations would be more difficult to be disentangled if there are distribution shifts at both structure-level and attribute-level. Since the node representations from hidden layers can also encode graph topology features [107], distribution shifts introduced through Z_A^s and Z_X^s will doubly mix at the learned features. In the worst case, the information about Z_A^c and Z_X^c could be partially covered by or even replaced by Z_A^s and Z_X^s. This will make OOD generalization of message passing GNNs trained through ERM much more difficult or even impossible. Besides, as the node representations of $1 \leq i \leq k$-th layer can also encode graph topology features [107], which, if spuriously correlated with labels through Z_A^s and entangled with part of invariant node features, i.e., Z_X^c, in the worst case, can greatly improve the difficulty or even make the OOD generalization impossible for neighbor aggregation GNNs trained with ERM.

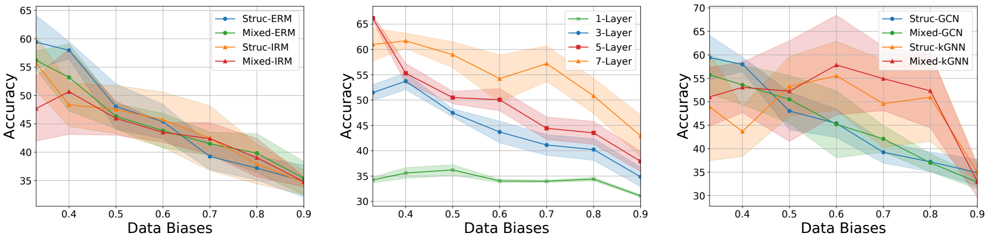
Using more expressive GNNs. Previous results on the expressivity of GNNs show that GNNs are limited to distinguish isomorphic graphs at most as 1-WL/2-WL test can distinguish [107]. After that, many follow-up variants are proposed to improve the expressivity of GNNs [68]. However, if the labels are spuriously correlated with certain subgraphs, even the GNN has high expressivity can still be prone to distribution shifts. In a idealistic case, when classifying a graph with a highly expressive GNN, it reduces to the linear or discrete feature case on the Euclidean regime. In this case, there exists many evidences showing that neural networks can fail to generalize to OOD data without a proper objective [7, 24, 4, 81, 10, 49, 23, 48, 2]. Empirically, we use k-GNNs [66] to verify the intuition and observe similar failures for this provably more expressive GNN as basic GNN variants.

D.3 More empirical results about failure case study in Sec. 2.3



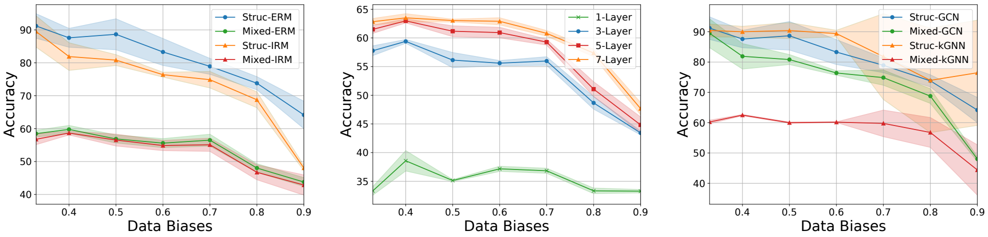
(a) Failures of training objectives. (b) Failures of deeper GNNs. (c) Failures of expressive GNNs.

Figure 9: Failure of existing methods on SPMotif with FIIF attribute shifts.



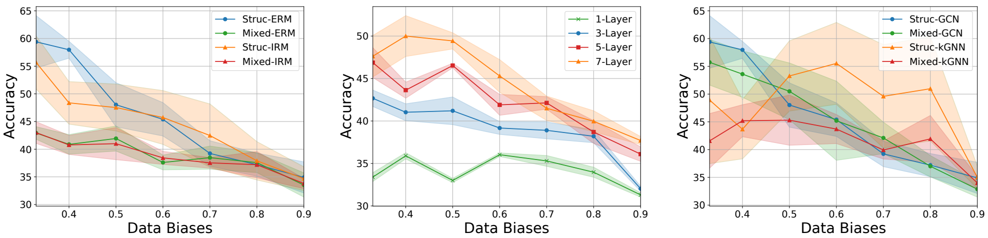
(a) Failures of training objectives. (b) Failures of deeper GNNs. (c) Failures of expressive GNNs.

Figure 10: Failure of existing methods on SPMotif with FIIF attribute shifts and graph size shifts.



(a) Failures of training objectives. (b) Failures of deeper GNNs. (c) Failures of expressive GNNs.

Figure 11: Failure of existing methods on SPMotif with PIIF attribute shifts.



(a) Failures of training objectives. (b) Failures of deeper GNNs. (c) Failures of expressive GNNs.

Figure 12: Failure of existing methods on SPMotif PIIF attribute shifts with graph size shifts.

To explore the behaviors of aforementioned methods against complicated distribution shifts on graphs, we first modify construction method in Wu et al. [104] to construct dataset for Fig. 9, where only FIIF structure-level spurious correlations are injected. Then we also inject FIIF attribute-level shifts, by setting the node attributes to constant vectors which is spuriously correlated with the labels. Furthermore, in Fig. 10, graph size shifts are added, which is exactly the SPMotif datasets used in DIR [104]. Besides, in Fig. 11, we can also change the FIIF attribute-level shifts to PIIF attribute-level shifts, where we flip the labels by a probability of 5% and let the flipped label to be spuriously correlated with the node features, following the PIIF SCM in Fig. 6. Graph size shifts can also be injected in this case, shown as Fig. 12. Next, we summarize our findings from the experiments.

Observation I: All existing methods are sensitive to distribution shifts. From the Fig. 9, 10, 11, 12, we can observe that *all* GNNs are sensitive to distribution shifts. As the intensity of spurious correlation grows, GNNs are more likely to overfit to shortcuts presented either in the structure-level or attribute-level, which is similar to general deep learning models [33].

Observation II: Higher variance also indicates unstable OOD performance. Although GNNs show certain robustness against single distribution shifts, e.g., performances do not decrease sharply at the beginning in Fig. 9, when the spurious correlation grows stronger, the OOD performance become more *unstable*, e.g., higher variance. The reason is that, GNNs sometimes can directly learn about the desired information at some random initializations, since the task is relatively simple compared to reality. Hence the performance will be highly sensitive to the quality of initialized points at the beginning. Consequently, the performances from multiple runs would exhibit high variance. However, when the task becomes more difficult, GNNs will consistently be prone to distribution shifts, and the variance will be smaller, as shown in experiments (Sec. 4).

Observation III: Entangling more distribution shifts can degenerate more GNN performance. As implied by the graph generation SCMs in Fig. 6, distribution shifts can happen at both structure-level and attribute-level, and each of them can have different type of spurious correlation with the label. In Fig. 9, we can find that, when the attribute-level distribution shifts are mixed, the performance will be worse and more unstable. When the graph size shifts are mixed, this phenomenon will be more obvious, as shown in Fig. 10. This phenomenon also verifies the observations in Knyazev et al. [46] that attention mechanism in GNN is also sensitive to graph size shifts and can hardly learn the desired attention distributions without further guidance. Moreover, when the structure-level and attribute-level shifts have different spurious correlation types, i.e., when FIIF structure-level shifts and PIIF attribute-level shifts are both presented, the performance drop will be more serious, by comparing Fig. 9 to Fig. 11, as well as Fig. 10 to Fig. 12.

Observation IV: Using more powerful architectures can not improve the OOD performance. From the sub-figures (b) and (c) in Fig. 9, 10, 11, 12, we can also observe that neither adding more message passing turns nor using more expressive GNN architectures can be immune to distribution shifts. On the contrary, they also exhibit similar behaviors like basic GNN architectures. Specifically, adding more message passing runs show certain robustness against distribution shifts since they are more likely to learn the desired information during the optimization [109]. However, when the intensity of spurious correlation grows stronger, deeper GNNs are more likely to overfit to shortcuts hence their performances will drop more sharply. On the other hand, using provably more expressive GNN architectures can not improve the OOD performance, either. In Fig. 9, 10, 11, 12 we use 1-2-3-GNN following the algorithm of k-GNNs which is provably more expressive than 2-WL test [66]. When there are no graph size shifts, k-GNNs will have higher performance at the beginning. When there are graph size shifts, k-GNNs will have a lower initial performance at the beginning. Then, as the spurious strength grows, k-GNNs can suddenly become seriously unstable, though k-GNNs can have higher averaged performance, which reflects unsatisfactory OOD performance as Observation II implies. When the intensity of spurious correlations grows even stronger, similar to deeper GNNs, OOD performances of k-GNNs will be more unstable and go down to similar level as that of normal GNN architectures. Hence, it calls for better optimization objectives as well as a suitable architectures to help improve the OOD generalization performance.

Beyond the empirical studies in previous section, we aim to accompany more formal discussions for explaining the failures of existing optimization objectives and architectures in the next sections.

D.4 Theoretical discussions for failure case study in Sec. 2.3

A motivating example. To begin with, we follow Ahuja et al. [2] to introduce a formal example on the failures of GNNs optimized with ERM or IRM [97, 4] via a linear binary classification problem:

Definition D.1 (Linear classification structural equation model (FIIF))**.**

$$\begin{aligned} &Y := (w^*_{\text{inv}} \cdot C) \oplus N,\ N \sim \text{Ber}(q),\ N \perp\!\!\!\perp (C, S),\\ &X \leftarrow S(C, S), \end{aligned}$$

where $w^*_{\text{inv}} \in \mathbb{R}^{n_c}$ with $\|w^*_{\text{inv}}\| = 1$ is the labeling hyperplane, $C \in \mathbb{R}^{n_c}$, $S \in \mathbb{R}^{n_s}$ are the corresponding invariant and varying latent variables, N is Bernoulli binary noise with a parameter of q and identical across all environments, $\oplus$ is the XOR operator, S is invertible.

Given data generation process as Assumption C.1, and latent space interaction as Assumption C.2 or C.3, and strictly separable invariant features 2.4, consider a k-layer linearized GNN $\rho \circ h$ using mean as READOUT for binary graph classification, if $\cup_{e\in\mathcal{E}_{\text{te}}}\text{supp}(\mathbb{P}^e) \not\subseteq \cup_{e\in\mathcal{E}_{\text{tr}}}\text{supp}(\mathbb{P}^e)$:

(i) For graphs features generated as Definition D.1, $\rho \circ h$ optimized with ERM or IRM will fail to generalize OOD (Eq. 8) almost surely;

(ii) For graphs with more than two nodes, globally same node features generated as Definition D.1, and graph labels that are the same as global node labels, $\rho \circ h$ optimized with ERM or IRM will fail to generalize OOD (Eq. 8) almost surely;

For graph classification, if the number of nodes is fixed to one, it covers the linear classification as above. When $\cup_{e\in\mathcal{E}_{\text{te}}}\text{supp}(\mathbb{P}^e) \not\subseteq \cup_{e\in\mathcal{E}_{\text{tr}}}\text{supp}(\mathbb{P}^e)$, it implies the S from training environments $\mathcal{E}_{\text{tr}}$ does not cover S from testing environments, while C can be covered. Moreover, the condition of strictly separable training data now can be formulated as $\min_{C\in\cup_{e\in\mathcal{E}_{\text{tr}}}(C\subseteq G^e)} \text{sgn}(w^*_{\text{inv}} \cdot C)(w^*_{\text{inv}} \cdot C) > 0$. Recall that ERM trains the model by minimizing the empirical risk (e.g., 0-1 loss) over all training data, and IRM formulates OOD generalization as:

$$\begin{aligned} &\min_{\theta, f_c} \frac{1}{|\mathcal{E}_{\text{tr}}|}\sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\rho \circ h)\\ &\text{s.t. } \rho \in \arg\min_{\hat{\rho}} R^e(\hat{\rho} \circ h),\ \forall e \in \mathcal{E}_{\text{tr}}. \end{aligned} \tag{10}$$

However, both ERM and IRM can not enable OOD generalization, i.e., finding the ground truth w^*_{inv}, following the Theorem 3 from Ahuja et al. [2]:

Theorem D.2 (Insufficiency of ERM and IRM)**.** *Suppose each $e \in \mathcal{E}_{all}$ follows Definition. D.1, C are strictly separable, bounded and satisfy the support overlap between $\mathcal{E}_{tr}$ and $\mathcal{E}_{te}$, and S are bounded, if S does not support the overlap, then both ERM and IRM fail at solving the OOD generalization problem.*

The reason is that, when C from all environments are strictly separable, there can be infinite many Bayes optimal solutions given training data $\{G^e, y^e\}_{e\in\mathcal{E}_{\text{tr}}}$, while there is only one optimal solution that does not rely on S. Hence, the probability of generalization to OOD (finding the optimal solution) tends to be 0 in probability.

As for case (ii), when the GNN uses mean readout to classify more than one node graphs, assuming the graph label is determined by the node label and all of the nodes have the same label that are determined as Definition D.1, then GNN optimized with ERM and IRM will also fail because of the same reasons as case (i).

Discussions on the failures of previous OOD related solutions. First of all, for IRM or similar objectives [81, 49, 2, 9] that require environment information or non-trivial data partitions, they can hardly be applied to graphs due to the lack of such information. The reason is that obtaining such information can be expensive due to the abstraction of graphs. Moreover, as proved in Theorem 5.1 of Rosenfeld et al. [80], when there is not sufficient support overlap between training environments and testing environments, the IRM or similar objectives can fail catastrophically when being applied to non-linear regime. The only OOD objective EIIL [23] that does not require environment labels, also rely on similar assumptions on the support overlap. We also empirically verify their failing behaviors in our experiments.

Moreover, since part of explainability works also try to find a subset of the inputs for interpretable prediction robustly against distribution shifts. Here we also provide a discussion for these works. The first work following this line is InvRAT [14], which develops an information-theoretic objective (we re-formulate it to suit with OOD generalization problem on graphs):

$$\min_{g,f_c}\max_{f_s} R(f_c\circ g, Y) + \lambda h(R(f_c\circ g, Y) - R_e(f_s\circ g, Y, E)). \tag{11}$$

However, it also requires extra environment labels for optimization that are often unavailable in graphs. Besides, the corresponding assumption on the data generation for guaranteed performance is essentially PIIF if applied to our case, while it can not provide any theoretical guarantee on FIIF.

We also notice a recent work, DIR [104], as a generalization of InvRAT to graphs while studying FIIF spurious correlations, that proposes an alternative objective which does not require environment label:

$$\min \mathbb{E}_s[R(h, Y|\text{do}(S=s))] + \lambda\text{Var}_s(\{R(h, Y|\text{do}(S=s))\}). \tag{12}$$

However, the theoretical justification established for DIR (Theorem 1 to Corollary 1 in Wu et al. [104]) essentially depends on the quality of the generator g which can be prone to spurious correlations. Thus, DIR can hardly provide any theoretical guarantees when applied to our case, neither for FIIF nor PIIF. In experiments, we empirically find the unstable and relatively high sensitivity of DIR to spurious correlations, which verifies our finding. More details about empirical behaviors of DIR can be found in Appendix G.

In contrast to DIR, GIB [120] that focuses on discovering a informative subgraph for explanation, essentially can provide theoretical guarantees for FIIF spurious correlations. Theoretically, (we copy the discussion in Appendix F here to provide an overview of relationships between GIB and DIR.) Under the FIIF assumption on latent interaction, the independence condition derived from causal model can also be rewritten as $Y \perp\!\!\!\perp S|C$ (similar to that in DIR [104] as they also focus on FIIF), which further implies $Y \perp\!\!\!\perp S|\widehat{G}_c$. Hence it is natural to use Information Bottleneck (IB) objective [95] to solve for G_c:

$$\begin{aligned}&\min_{f_c,g} R_{G_c}(f_c(\widehat{G}_c)),\\ &\text{s.t. } G_c = \mathop{\arg\max}_{\widehat{G}_c=g(G)\subseteq G} I(\widehat{G}_c, Y) - I(\widehat{G}_c, \mathcal{G}),\end{aligned} \tag{13}$$

which explains the success of many existing works in finding predictive subgraph through IB [120]. However, the estimation of $I(\widehat{G}_c, G)$ is notoriously difficult due to the complexity of graph, which can lead to unstable convergence as observed in our experiments. In contrast, optimization with contrastive objective in CIGA as Eq. 5 induces more stable convergence.

D.5 Challenges of OOD generalization on graphs.

From the aforementioned analysis, we can summarize some key challenges revealed by the failures of both existing optimization objectives and GNN architectures. In particular, we are facing two main challenges a) Distribution shifts on graphs are more complicated where different types of spurious correlations can be entangled via different graph properties; b) Environment labels are usually not available due to the abstract graph data structure.

E Theory and Discussions

In this section, we provide proofs for propositions and theorems mentioned in the main paper.

E.1 More discussions on Definition 2.5 for Invariant GNNs

Definition 2.5 is motivated by applying the invariance principle to the established SCMs in Sec. 2.2, following the literature of invariant learning [76]. In this section, we will present Proposition E.2 and Proposition E.3 to illustrate how satisfying the minmax objective in Definition E.1 is equivalent to identifying the underlying invariant subgraph G_c that contains all of the information about causal factor C in G, under both FIIF and PIIF SCMs (Fig. 2(b) and Fig. 2(c)).

Definition E.1 (Invariant GNN)**.** Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{\text{all}}$ that follow the same graph generation process in Sec. 2.2, considering a GNN $\rho \circ h$ that has a permutation invariant graph encoder $h : \mathcal{G} \rightarrow \mathbb{R}^h$ and a downstream classifier $\rho : \mathbb{R}^h \rightarrow \mathcal{Y}$, $\rho \circ h$ is an invariant GNN if it minimizes the worst case risk among all environments, i.e., $\min \max_{e \in \mathcal{E}_{\text{all}}} R^e$.

First, we show that using the invariant subgraphs G_c to predict Y can satisfy the minmax objective $\min \max_{e \in \mathcal{E}_{\text{all}}} R^e$ in Proposition E.2.

Proposition E.2. *Let $\mathcal{G}_c$ denote the subgraph space for G_c, given a set of graphs with their labels $\mathcal{D} = \{G^{(i)}, y^{(i)}\}_{i=1}^N$ and $\mathcal{E}_{all}$ that follow the graph generation process in Sec. 2.2 (or Sec. C), a GNN $\rho \circ h : \mathcal{G}_c \rightarrow \mathcal{Y}$ that takes G_c of G as the input to predict Y, and solves the following objective can generalize to OOD graphs, i.e., solving the minmax objective in Def. E.1:*

$$\min_\theta R_{\mathcal{G}_c}(\rho \circ h),$$

where $R_{\mathcal{G}_c}$ is the empirical risk over $\{G_c^{(i)}, y^{(i)}\}_{i=1}^N$ and $G_c^{(i)}$ is the underlying invariant subgraph G_c for $G^{(i)}$.

Proof. We establish the proof with independent causal mechanism (ICM) assumption in SCM [74, 77]. In particular, given the data generation assumption, i.e., for both FIIF (Assumption 2.2) and PIIF (Assumption 2.3), we have: $\forall e$,

$$\begin{aligned}
P(Y|C) &= P(Y|C, E = e) \\
P(Y|G_c) \sum_{G_c} P(G_c|C) &= P(Y|G_c) \sum_{G_c} P(G_c|C, E = e) \\
P(Y|G_c) \sum_{G_c} P(G_c|C) &= P(Y|G_c, E = e) \sum_{G_c} P(G_c|C) \\
P(Y|G_c) &= P(Y|G_c, E = e),
\end{aligned} \tag{14}$$

where we use ICM for the first three equalities. From Eq. 14, it suffices to know $P(Y|G_c)$ is invariant across different environments. Hence, a GNN predictor $\rho \circ h : \mathcal{G}_c \rightarrow \mathcal{Y}$ optimized with empirical risk given G_c, essentially minimizes the empirical risk across all environments, i.e., $\min R_{\mathcal{G}_c} = \min \max R^e$. Thus, if $\rho \circ h$ solves $\min R_{\mathcal{G}_c}$, it also solves $\min \max R^e$, hence it elicits a invariant GNN predictor according to Definition. E.1. □

Besides, we show in Proposition E.3 that only using the underlying invariant subgraphs G_c to make predictions can satisfy the minmax objectives. Or equivalently, a GNN predictor solving the minmax objective can only rely on the underlying invariant subgraph G_c to predict Y.

Proposition E.3. *Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{all}$ that follow the same graph generation process in Sec. 2.2, considering a GNN $\rho \circ h$ that has a permutation invariant graph encoder $h : \mathcal{G} \rightarrow \mathbb{R}^h$ and a downstream classifier $\rho : \mathbb{R}^h \rightarrow \mathcal{Y}$, $\rho \circ h$ that minimizes the worst case risk among all environments, i.e., $\min \max_{e \in \mathcal{E}_{all}} R^e$, can not rely on any part of G_s, i.e., $\rho \circ h(G) \perp\!\!\!\perp G_s$.*

Proof. The proof for Proposition E.3 is straightforward. Assuming that $\rho \circ h(G) \not\perp\!\!\!\perp G_s$, as E is influenced by the changes of E through S in both FIIF and PIIF SCMs (Fig. 2(b) and Fig. 2(c)), then $\rho \circ h(G) \not\perp\!\!\!\perp E$ as well. Consequently, there exists some graph G corresponding to G_c, G_s^e and $\rho \circ h(G) = Y$ under an environment e, such that we can always find a proper e' to make $\rho \circ h(G) \neq Y$. In contrast, the prediction of a GNN that satisfies $\rho \circ h(G) \perp\!\!\!\perp G_s$ remains invariant against arbitrary changes of environments. Thus, it leads to a contradiction to the condition that $\min \max_{e' \in \mathcal{E}_{\text{all}}} R^{e'}$. Therefore, a GNN that solves $\min \max_{e \in \mathcal{E}_{\text{all}}} R^e$ must satisfy $\rho \circ h(G) \perp\!\!\!\perp G_s$. □

Combining Proposition E.2 and Proposition E.3, we are highly motivated to find the underlying invariant subgraphs to make predictions about the original graphs, which converges to Eq. 1. Tackling Eq. 1 under the unavailability of E brings us two variants of CIGA solutions, as illustrated in Section 3.

E.2 Proof for theorem 3.1 (i)

Theorem E.4 (CIGAv1 Induces Invariant GNNs). *Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{all}$ that follow the same graph generation process in Sec. 2.2, assuming that* (a) f_{gen}^G *and* $f_{gen}^{G_c}$ *in Assumption 2.1 are invertible,* (b) *samples from each training environment are equally distributed, i.e.,$|\mathcal{D}_{\hat{e}}| = |\mathcal{D}_{\tilde{e}}|$, $\forall \hat{e}, \tilde{e} \in \mathcal{E}_{tr}$, if $\forall G_c, |G_c| = s_c$, then a GNN $f_c \circ g$ solves Eq. 4, is an invariant GNN (Def. 2.5).*

Proof. We re-write the objective as follows:

$$\max_{f_c, g} I(\widehat{G}_c; Y), \text{ s.t. } \widehat{G}_c \in \underset{\widehat{G}_c = g(G), |\widehat{G}_c| \le s_c}{\arg\max} I(\widehat{G}_c; \widetilde{G}_c | Y), \tag{15}$$

where $\widehat{G}_c = g(G), \widetilde{G}_c = g(\widetilde{G})$ and $\widetilde{G} \sim \mathbb{P}(G|Y)$, i.e., $\widetilde{G}$ and G have the same label.

The proof of Theorem E.4 is essentially to show the estimated $\widehat{G}_c$ through Eq. 15 is the underlying G_c, then the maximizer of $I(\widehat{G}_c; Y)$ in Eq. 15 can produce most informative and stable predictions about Y based on G, hence is an invariant GNN (Definition. E.1).

In the next, we are going to take an information-theoretic view of the first term $I(\widehat{G}_c; Y)$ and the second term $I(\widehat{G}_c; \widetilde{G}_c | Y)$ to conclude the proof. We begin by introducing the following lemma:

Lemma E.5. *Given the same conditions as Thm. E.4, $I(\widehat{G}_c; Y)$ is maximized if and only if $I(\widehat{G}_c; Y | E = e)$ is maximized, $\forall e \in \mathcal{E}_{tr}$.*

The proof for Lemma E.5 is straightforward, given the condition that samples from each training environment are equally distributed, i.e.,$|\mathcal{D}_{\hat{e}}| = |\mathcal{D}_{\tilde{e}}|$, $\forall \hat{e}, \tilde{e} \in \mathcal{E}_{\text{tr}}$. Obviously, $\widehat{G}_c = G_c$ is a maximizer of $I(\widehat{G}_c; Y) = I(C; Y) = H(Y)$, since $f_{\text{gen}}^c : \mathcal{C} \to \mathcal{G}_c$ is invertible and C causes Y. However, there might be some subset $G_s^p \subseteq G_s$ from the underlying G_s that entail the same information about label, i.e., $I(G_c^p \cup G_s^p; Y) = I(G_c; Y)$ where $\widehat{G}_c = G_c^p \cup G_s^p$ and $G_c^p = G_c \cap \widehat{G}_c$. For FIIF (Assumption 6(b)), it can not happen, otherwise, let $G_c^l = G_c - G_c^p$, then we have:

$$\begin{aligned}
I(\widehat{G}_c; Y) = I(G_c^p \cup G_s^p; Y) &= I(G_c^p \cup G_c^l; Y) = I(G_c; Y) \\
I(G_c^p; Y) + I(G_s^p; Y | G_c^p) &= I(G_c^p; Y) + I(G_c^l; Y | G_c^p) \\
I(G_s^p; Y | G_c^p) &= I(G_c^l; Y | G_c^p) \\
H(Y | G_c^p) - H(Y | G_c^p, G_s^p) &= H(Y | G_c^p) - H(Y | G_c^p, G_c^l) \\
H(Y | G_c^p) - H(Y | G_c^p, G_s^p) &= H(Y | G_c^p), \\
H(Y | G_c^l, G_s^p) &= 0,
\end{aligned} \tag{16}$$

where the second last equality is due to $C \to Y$ and the invertibility of $f_{\text{gen}}^c : \mathcal{C} \to \mathcal{G}_c$ in FIIF, i.e., $H(Y|C) = H(Y|G_c) = H(Y|G_c^p, G_c^l) = 0$. However, in PIIF, it can hold since conditioning on G_c^p, G_s^p can not determine Y, as $S \not\perp\!\!\!\perp Y | C$. In other words, $G_s \not\perp\!\!\!\perp Y | G_c$, which means G_s can imply some information about Y that is equivalent to $I(G_c^l; Y | G_c^p)$.

To avoid the presence of spuriously correlated G_s in $\widehat{G}_c$, we will use the second term to eliminate it:

$$\begin{aligned}&\max_{f_c,g} I(\widehat{G}_c;\widetilde{G}_c|Y),\\ &\qquad = H(\widehat{G}_c|Y) - H(\widehat{G}_c|\widetilde{G}_c,Y),\end{aligned} \tag{17}$$

where $\widehat{G}_c = g(G), \widetilde{G}_c = g(\widetilde{G})$ are two positive samples drawn from the same class (i.e., condition on the same Y). Since the all of the training environments are equally distributed, maximizing $I(\widehat{G}_c;\widetilde{G}_c|Y)$ is essentially maximizing $I(\widehat{G}_c, E=\hat{e};\widetilde{G}_c, E=\tilde{e}|Y), \forall \hat{e},\tilde{e} \in \mathcal{E}_{\text{tr}}$. Hence, we have:

$$\begin{aligned}&\max_{f_c,g} I(\widehat{G}_c;\widetilde{G}_c|Y),\\ &\qquad = I(\widehat{G}_c, E=\hat{e};\widetilde{G}_c, E=\tilde{e}|Y)\\ &\qquad = H(\widehat{G}_c, E=\hat{e}|Y) - H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y).\end{aligned} \tag{18}$$

We claim Eq. 18 can eliminate any potential subsets from G_s in the estimated $\widehat{G}_c$.

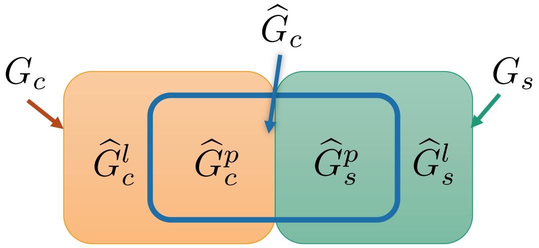


Figure 13: Illustration of the notation. G_c and G_s are two disjoint sets. $\widehat{G}_c$ may contain certain subsets from G_c and G_s. The subsets from G_c and G_s contained in $\widehat{G}_c$ are denoted as $\widehat{G}_c^p$ and $\widehat{G}_s^p$, respectively. While the left subsets in G_c and G_s are denoted as $\widehat{G}_c^l$ and $\widehat{G}_s^l$, respectively.

Otherwise, suppose there are some subsets $\widehat{G}_s^p \subseteq \widehat{G}_s$ and $\widetilde{G}_s^p \subseteq \widetilde{G}_s$ contained in the estimated $\widehat{G}_c, \widetilde{G}_c$, where $\widehat{G}_s, \widetilde{G}_s$ be the corresponding underlying G_ss for $\widehat{G}_c, \widetilde{G}_c$. Let $\widehat{G}_c^*$ and $\widetilde{G}_c^*$ be the ground truth invariant subgraph G_cs of $\widehat{G}$ and $\widetilde{G}$, $\widehat{G}_c^l = \widehat{G}_c^* - \widehat{G}_c$ and $\widetilde{G}_c^l = \widetilde{G}_c^* - \widetilde{G}_c$ be the **l**eft (un-estimated) subsets from corresponding ground truth G_cs, and $\widehat{G}_c^p = \widehat{G}_c^* - \widehat{G}_c^l$ and $\widetilde{G}_c^p = \widetilde{G}_c^* - \widetilde{G}_c^l$ be the complement, or equivalently, the **p**artial $\widehat{G}_c^*, \widetilde{G}_c^*$ that are estimated in $\widehat{G}_c, \widetilde{G}_c$, respectively. We can also define similar counterparts for G_s: $\widehat{G}_s^p, \widetilde{G}_s^p$ are the partial $\widehat{G}_s, \widetilde{G}_s$s contained in the estimated $\widehat{G}_c, \widetilde{G}_c$ while $\widehat{G}_s^l, \widetilde{G}_s^l$ are the left subsets $\widehat{G}_s, \widetilde{G}_s$, respectively.

Recall the constraint that $|G_c| = s_c$, hence if $\widehat{G}_c^p \subseteq \widehat{G}_c$, then a corresponding $\widehat{G}_c^l = \widehat{G}_c^* - \widehat{G}_c^p$ will be replaced by $\widehat{G}_s^p$ in $\widehat{G}_c$. In this case, we have:

$$\begin{aligned}H(\widehat{G}_c, E=\hat{e}|Y) &= H(E=\hat{e}|\widehat{G}_c, Y) + H(\widehat{G}_c|E=\hat{e}, Y)\\ &= H(\widehat{G}_c^p \cup \widehat{G}_s^p|E=\hat{e}, Y)\\ &= H(\widehat{G}_c^p|E=\hat{e}, Y) + H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y)\end{aligned} \tag{19}$$

where the second equality is due to $E=\hat{e}$ is determined so that $H(E=\hat{e}|\widehat{G}_c, Y) = 0$. Compared Eq. 19 to that when $\widehat{G}_c = \widehat{G}_c^*$, we have the entropy change as:

$$\begin{aligned}\Delta H(\widehat{G}_c, E=\hat{e}|Y) &= H(\widehat{G}_c, E=\hat{e}|Y) - H(\widehat{G}_c^*, E=\hat{e}|Y),\\ &= H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y) - H(\widehat{G}_c^l|\widehat{G}_c^p, E=\hat{e}, Y).\end{aligned} \tag{20}$$

Let $\epsilon = H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y)$. In a idealistic setting, when the noise of the generation process $S := f_{\text{spu}}(Y, E)$ in PIIF tends to be 0, i.e., $\epsilon \to 0$, S is determined conditioned on E, Y, hence G_s and any subsets of G_s are all determined. Then, it suffices to know that in Eq. 20, $H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y) = 0$ while $H(\widehat{G}_c^l|\widehat{G}_c^p, E=\hat{e}, Y) > 0$ since $\widehat{G}_c^l$ can not be determined when given $\widehat{G}_c^p, E=\hat{e}, Y$. Thus, when some subset from G_s is included in $\widehat{G}_c$, it will minimize $H(\widehat{G}_c, E=\hat{e}|Y)$.

However in practice, it is usual that $\epsilon > 0$. Therefore, in the next, we will show how $\epsilon = H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y)$ can be cancelled thus leading to a smaller $H(\widehat{G}_c, E=\hat{e}|Y)$, by considering the second term $H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y)$.

As for $H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y)$, without loss of generality, we can divide all of the possible cases into two:

(i) One of $\widehat{G}_c$ and $\widetilde{G}_c$ contains some subset of G_s, i.e., $\widehat{G}_c$ contains some $\widehat{G}_s^p \subseteq \widehat{G}_s$;

(ii) Both $\widehat{G}_c$ and $\widetilde{G}_c$ contain some $\widehat{G}_s^p \subseteq \widehat{G}_s$ and $\widetilde{G}_s^p \subseteq \widetilde{G}_s$, respectively.

For (i), we have:

$$\begin{aligned} H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) &= H(\widehat{G}_c^p, \widehat{G}_s^p, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) \\ &= H(\widehat{G}_s^p|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}) + H(\widehat{G}_c^p, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y), \end{aligned} \tag{21}$$

Thus, we can write the change of $H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y)$ between $\widehat{G}_c = \widehat{G}_c^p \cup \widehat{G}_s^p$ and $\widehat{G}_c = \widehat{G}_c^*$ as:

$$\begin{aligned} \Delta H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) &= H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) - H(\widehat{G}_c^*, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y), \\ &= H(\widehat{G}_s^p|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}) \\ &\quad - H(\widehat{G}_c^l|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}). \end{aligned} \tag{22}$$

Combing $\Delta H(\widehat{G}_c, E=\hat{e}|Y)$, we have:

$$\begin{aligned} \Delta I(\widehat{G}_c, E=\hat{e}; \widetilde{G}_c, E=\tilde{e}|Y) &= \Delta H(\widehat{G}_c, E=\hat{e}|Y) - \Delta H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) \\ &= \left\{ H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y) - H(\widehat{G}_s^p|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}) \right\} \\ &\quad + \left\{ -H(\widehat{G}_c^l|\widehat{G}_c^p, E=\hat{e}, Y) + H(\widehat{G}_c^l|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}) \right\}, \\ &= -H(\widehat{G}_c^l|\widehat{G}_c^p, E=\hat{e}, Y) + H(\widehat{G}_c^l|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}), \end{aligned} \tag{23}$$

where the last equality is because of the independence of $\widehat{G}_s^p$ between $\widetilde{G}_c, E=\tilde{e}$ conditioned on $Y, E=\hat{e}$. Since conditioning will lower the entropy for both discrete and continuous variables [22, 114], we have:

$$\Delta I(\widehat{G}_c, E=\hat{e}; \widetilde{G}_c, E=\tilde{e}|Y) < 0, \tag{24}$$

which implies the existence of $\widehat{G}_s^p$ in $\widehat{G}_c$ will lower down the second term in Eq. 15 for the case (i).

For (ii), we have:

$$\begin{aligned} H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) &= H(\widehat{G}_c^p, \widehat{G}_s^p, E=\hat{e}|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, Y) \\ &= H(\widehat{G}_s^p|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e}) \\ &\quad + H(\widehat{G}_c^p, E=\hat{e}|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, Y), \end{aligned} \tag{25}$$

Similar to (i), $H(\widehat{G}_s^p|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, Y, \widehat{G}_c^p, E=\hat{e})$ can be cancelled out with $H(\widehat{G}_s^p|\widehat{G}_c^p, E=\hat{e}, Y)$. Then, we have:

$$\begin{aligned} \Delta I(\widehat{G}_c, E=\hat{e}; \widetilde{G}_c, E=\tilde{e}|Y) &= \Delta H(\widehat{G}_c, E=\hat{e}|Y) - \Delta H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) \\ &= -H(\widehat{G}_c^l|\widehat{G}_c^p, E=\hat{e}, Y) + H(\widehat{G}_c^l|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, \widehat{G}_c^p, Y, E=\hat{e}). \end{aligned} \tag{26}$$

Since additionally conditioning on $\widehat{G}_s^p$ in $H(\widehat{G}_c^l, E=\hat{e}|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, Y)$ can not lead to new information about $\widehat{G}_c^l$, we have:

$$\begin{aligned} H(\widehat{G}_c^l|\widetilde{G}_c^p, \widetilde{G}_s^p, E=\tilde{e}, \widehat{G}_c^p, Y, E=\hat{e}) &= H(\widehat{G}_c^l|\widetilde{G}_c^p, E=\tilde{e}, \widehat{G}_c^p, Y, E=\hat{e}) \\ &< H(\widehat{G}_c^l|\widehat{G}_c^p, Y, E=\hat{e}), \end{aligned} \tag{27}$$

which follows that $\Delta I(\widehat{G}_c, E=\hat{e}; \widetilde{G}_c, E=\tilde{e}|Y) < 0$.

To summarize, the ground truth G_c is the only maximizer of the objective (Eq. 15), hence solving for the objective (Eq. 15) can elicit an invariant GNN.

E.3 Proof for theorem 3.1 (ii)

Theorem E.6 (CIGAv2 Induces Invariant GNNs)**.** *Given a set of graph datasets $\{\mathcal{D}^e\}_e$ and environments $\mathcal{E}_{all}$ that follow the same graph generation process in Sec. 2.2, assuming that* (a) f_{gen}^{G} *and* $f_{gen}^{G_c}$ *in Assumption 2.1 are invertible,* (b) *samples from each training environment are equally distributed, i.e.,* $|\mathcal{D}_{\hat{e}}| = |\mathcal{D}_{\tilde{e}}|,\ \forall \hat{e}, \tilde{e} \in \mathcal{E}_{tr}$*, a GNN* $f_c \circ g$ *solves Eq. 4, is an invariant GNN (Def. 2.5).*

Proof. We re-write the objective as follows:

$$\begin{aligned} \max_{f_c, g}\ I(\widehat{G}_c;Y) + I(\widehat{G}_s;Y),\ \text{s.t. } \widehat{G}_c \in \mathop{\arg\max}_{\widehat{G}_c=g(G),\widetilde{G}_c=g(\widetilde{G})} I(\widehat{G}_c;\widetilde{G}_c|Y), \\ I(\widehat{G}_s;Y) \leq I(\widehat{G}_c;Y),\ \widehat{G}_s = G - g(G). \end{aligned} \tag{28}$$

where $\widehat{G}_c = g(G), \widetilde{G}_c = g(\widetilde{G})$ and $\widetilde{G} \sim \mathbb{P}(G|Y)$, i.e., $\widetilde{G}$ and G have the same label.

Similar to the proof for Theorem E.4, to prove Theorem E.6 is essentially to show the estimated $\widehat{G}_c$ through Eq. 28 is the underlying G_c, hence the minimizer of Eq. 28 elicits an invariant GNN predictor (Definition. E.1).

In the next, we also begin with a lemma:

Lemma E.7. *Given data generation process as Theorem E.6, for both FIIF and PIIF, we have:*

$$I(C;Y) \geq I(S;Y),$$

hence $I(G_c;Y) \geq I(G_s;Y)$.

Proof for Lemma E.7. For both FIIF and PIIF, Assumption 2.4 implies that $H(C|Y) \leq H(S|Y)$. It follows that $I(C;Y) = H(Y) - H(C|Y) \geq H(Y) - H(S|Y) = I(S;Y)$. Then, since $f_{\text{gen}}^{G_c} : \mathcal{C} \to \mathcal{G}_c$ is invertible, we have $I(G_c;Y) = I(C;Y) \geq I(S;Y) \geq I(G_s;Y)$. □

Given Lemma E.7, we know $\widehat{G}_c$ at least contains some subset of the underlying G_c, otherwise the constraint $I(\widehat{G}_s;Y) \leq I(\widehat{G}_c;Y)$ will be violated since $G_c \subseteq \widehat{G}_s$ in this case.

Assuming there are some subset of G_s contained in $\widehat{G}_c$, without loss of generality, we can divide all of the possible cases about $\widehat{G}_c$ into two:

(i) $\widehat{G}_c$ only contains a subset of the underlying G_c;

(ii) $\widehat{G}_c$ contains a subset of the underlying G_c as well as part of the underlying G_s;

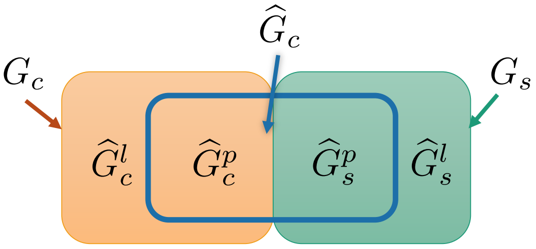


Figure 14: Illustration of the notation for estimated $\widehat{G}_c$ from G. G_c and G_s are two disjoint sets. $\widehat{G}_c$ may contain certain subsets from G_c and G_s. The subsets from G_c and G_s contained in $\widehat{G}_c$ are denoted as $\widehat{G}_c^p$ and $\widehat{G}_s^p$, respectively. While the left subsets in G_c and G_s are denoted as $\widehat{G}_c^l$ and $\widehat{G}_s^l$, respectively. Similar notations are also applicable for the estimated $\widetilde{G}_c$ from $\widetilde{G}$.

Before the discussion, let us inherit the notations of subsets of G_c, G_s from the proof for Theorem E.4: Let $\widehat{G}_c^*$ and $\widetilde{G}_c^*$ be the ground truth invariant subgraph G_cs of $\widehat{G}$ and $\widetilde{G}$, $\widehat{G}_c^l = \widehat{G}_c^* - \widehat{G}_c$ and $\widetilde{G}_c^l = \widetilde{G}_c^* - \widetilde{G}_c$ be the **l**eft (un-estimated) subsets from corresponding ground truth G_cs, and $\widehat{G}_c^p = \widehat{G}_c^* - \widehat{G}_c^l$ and $\widetilde{G}_c^p = \widetilde{G}_c^* - \widetilde{G}_c^l$ be the complement, or equivalently, the **p**artial $\widehat{G}_c^*, \widetilde{G}_c^*$ that are estimated in $\widehat{G}_c, \widetilde{G}_c$, respectively. Similarly, $\widehat{G}_s^p, \widetilde{G}_s^p$ are the partial $\widehat{G}_s, \widetilde{G}_s$s contained in the estimated $\widehat{G}_c, \widetilde{G}_c$ while $\widehat{G}_s^l, \widetilde{G}_s^l$ are the left subsets $\widehat{G}_s, \widetilde{G}_s$, respectively.

First of all, case (i) cannot hold because, when maximizing $I(\widehat{G}_c;\widetilde{G}_c|Y)$, if $\exists \widehat{G}_c^l = \widehat{G}_c^* - \widehat{G}_c$, as shown in the proof for Theorem E.4, including $\widehat{G}_c^l$ into $\widehat{G}_c$ can always enlarge $I(\widehat{G}_c;\widetilde{G}_c|Y)$, while not affecting the optimality of $I(\widehat{G}_s;Y)+$

$I(\widehat{G}_c;Y)$ by re-distributing $\widehat{G}_c^l$ from $\widehat{G}_s$ to $\widehat{G}_c$. Consequently, $\widehat{G}_c^*$ must be included in $\widehat{G}_c$, i.e., $\widehat{G}_c^* \subseteq \widehat{G}_c$.

As for case (ii), recall that, by the condition of equally distributed training samples from each training environment, maximizing $I(\widehat{G}_c;\widetilde{G}_c|Y)$ is essentially maximizing $I(\widehat{G}_c, E=\hat{e};\widetilde{G}_c, E=\tilde{e}|Y)$, $\forall \hat{e}, \tilde{e} \in \mathcal{E}_{\text{tr}}$, hence, we have:

$$\begin{aligned}
&\max_{g,f_c} I(\widehat{G}_c;\widetilde{G}_c|Y),\\
&\qquad = I(\widehat{G}_c, E=\hat{e};\widetilde{G}_c, E=\tilde{e}|Y)\\
&\qquad = H(\widehat{G}_c, E=\hat{e}|Y) - H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y).
\end{aligned} \tag{29}$$

We claim Eq. 29 can eliminate any potential subsets in the estimated $\widehat{G}_c$. Similarly, we have:

$$\begin{aligned}
H(\widehat{G}_c, E=\hat{e}|Y) &= H(E=\hat{e}|\widehat{G}_c, Y) + H(\widehat{G}_c|E=\hat{e}, Y)\\
&= H(\widehat{G}_c^* \cup \widehat{G}_s^p|E=\hat{e}, Y)\\
&= H(\widehat{G}_c^*|E=\hat{e}, Y) + H(\widehat{G}_s^p|\widehat{G}_c^*, E=\hat{e}, Y)\\
&= H(\widehat{G}_c^*|Y) + H(\widehat{G}_s^p|\widehat{G}_c^*, E=\hat{e}, Y)
\end{aligned} \tag{30}$$

where the second equality is due to $E=\hat{e}$ is determined. Compared to the case that $\widehat{G}_c = \widehat{G}_c^*$, we have:

$$\begin{aligned}
\Delta H(\widehat{G}_c, E=\hat{e}|Y) &= H(\widehat{G}_c, E=\hat{e}|Y) - H(\widehat{G}_c^*, E=\hat{e}|Y),\\
&= H(\widehat{G}_s^p|\widehat{G}_c^*, E=\hat{e}, Y).
\end{aligned} \tag{31}$$

Then, as for $H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y)$, without loss of generality, we can divide all of the possible cases into two:

(a) $\widehat{G}_c$ contains some $\widehat{G}_s^p \subseteq \widehat{G}_s$;

(b) Both $\widehat{G}_c$ and $\widetilde{G}_c$ contain some $\widehat{G}_s^p \subseteq \widehat{G}_s$ and $\widetilde{G}_s^p \subseteq \widetilde{G}_s$, respectively.

For (a), we have:

$$\begin{aligned}
H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) &= H(\widehat{G}_c^*, \widehat{G}_s^p, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y)\\
&= H(\widehat{G}_s^p|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^*, E=\hat{e}) + H(\widehat{G}_c^*, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y),
\end{aligned} \tag{32}$$

Similarly to the proof for Theorem E.4, when considering $\Delta I(\widehat{G}_c;\widetilde{G}_c|Y)$, the effects of $H(\widehat{G}_s^p|\widetilde{G}_c, E=\tilde{e}, Y, \widehat{G}_c^*, E=\hat{e})$ is cancelled out by $H(\widehat{G}_s^p|\widehat{G}_c^*, E=\hat{e}, Y)$. Hence, we have:

$$\Delta I(\widehat{G}_c;\widetilde{G}_c|Y) = 0.$$

For (b), we have:

$$\begin{aligned}
H(\widehat{G}_c, E=\hat{e}|\widetilde{G}_c, E=\tilde{e}, Y) &= H(\widetilde{G}_c^*, \widetilde{G}_s^p, E=\hat{e}|\widetilde{G}_c^*, \widetilde{G}_s^p, E=\tilde{e}, Y)\\
&= H(\widehat{G}_s^p|\widetilde{G}_c^*, \widetilde{G}_s^p, E=\tilde{e}, Y, \widehat{G}_c^*, E=\hat{e})\\
&\quad + H(\widehat{G}_c^*|\widetilde{G}_c^*, \widetilde{G}_s^p, E=\tilde{e}, Y, E=\hat{e}),
\end{aligned} \tag{33}$$

Similarly, $H(\widehat{G}_s^p|\widetilde{G}_c^*, \widetilde{G}_s^p, E=\tilde{e}, Y, \widehat{G}_c^*, E=\hat{e}) = 0$ can also be cancelled out by $H(\widehat{G}_s^p|\widehat{G}_c^*, E=\hat{e}, Y)$. Moreover, for $H(\widehat{G}_c^*|\widetilde{G}_c^*, \widetilde{G}_s^p, E=\tilde{e}, Y, E=\hat{e})$, $\widetilde{G}_s^p$ can not bring no additional information about $\widehat{G}_c^*$, when conditioning on $\widetilde{G}_c^*, Y, E=\tilde{e}$. Hence, we also have:

$$\Delta I(\widehat{G}_c;\widetilde{G}_c|Y) = 0.$$

To summarize, when maximizing $I(\widehat{G}_c;\widetilde{G}_c|Y)$, including any $\widehat{G}_s^p \subseteq \widehat{G}_s^*$ can not bring additional benefit while affecting the optimality of $I(\widehat{G}_s;Y) + I(\widehat{G}_c;Y)$. More specifically, when considering the changes to $I(\widehat{G}_s;Y) + I(\widehat{G}_c;Y)$, $\forall G_s^p \subseteq G_s$, we have

$$I(G - \widehat{G}_c^* - G_s^p;Y) \leq I(G - \widehat{G}_c^*;Y),\ \forall G_s^p \subseteq G_s,$$

while $I(Y;\widehat{G}_c^*,G_s^p)=I(Y;\widehat{G}_c^*)+I(Y;\widehat{G}_s^p|\widehat{G}_c^*),\ \forall e\in\mathcal{E}_{\text{tr}}$. Consequently,

$$\begin{aligned}\Delta I(\widehat{G}_s;Y)+I(\widehat{G}_c;Y)&=-I(\widehat{G}_s^p;Y|\widehat{G}_s^l)+I(\widehat{G}_s^p;Y|\widehat{G}_c^*)\\&=-I(\widehat{G}_s^p;Y)+I(\widehat{G}_s^p;Y|\widehat{G}_c^*)\leq 0.\end{aligned}\tag{34}$$

Hence, only the underlying G_c is the solution to Eq. 28, which implies that solving for the objective (Eq. 28) can elicit an invariant GNN.

F Details of Prototypical CIGA Implementation

In fact, the CIGA framework introduced in Sec. 3 can have multiple implementations. We choose interpretable architectures in our experiments for the purpose of concept verification. More sophisticated architectures can be incorporated. Experimental results in Sec. 4 also demonstrates that, even equipped with basic GNN architectures, CIGA already has the excellent OOD generalization ability, hence it is promising to incorporate more advanced architectures from the prosperous GNN literature.

We now introduce the details of the architectures used in our experiments. Recall that CIGA decomposes a GNN model for graph classification into two modules, i.e., a featurizer: $g:\mathcal{G}\to\mathcal{G}_c$ and a classifier $f_c:\mathcal{G}_c\to\mathcal{Y}$. Specifically, for the implementation of Featurizer, we choose one of the common practices GAE [44] for calculating the sampled weights for each edge. More formally, the soft mask is predicted through the following equation:

$$Z=\text{GNN}(G)\in\mathbb{R}^{n\times h},\ M=\text{a}(Z,A)\in\mathbb{R}^{n\times n},$$

where a calculates the sampling weights for each edge using a MLP: $M_{ij}=\text{MLP}([Z_i,Z_j])$.

If a sampling ratio s_c is predetermined, we sample s_c of total edges with the largest predicted weights as a soft estimation of $\widehat{G}_c$. Then, the estimated $\widehat{G}_c$ will be forwarded to the classifier f_c for predicting the labels of the original graph. Although Theorem E.4 assumes s_c is known, in real applications we do not know the specific s_c. Hence, in experiments, we select s_c according to the validation performance. To thoroughly study the effects of $I(\widehat{G}_s;Y)$ comparing to CIGAv1, we stick to using the same s_c and sampling process for CIGAv2, while CIGAv2 essentially requires less specific knowledge about ground truth r_c hence achieving better empirical performance. Moreover, once the sampled edges are determined, the classifier GNN can take either the original feature of the input graph or the learned feature from the featurizer as the new node attributes for $\widehat{G}_c$. We select the architecture according to the validation performance from some random runs.

For the implementation of the information theoretic objectives, we will use CIGAv2 for elaboration while the implementation of CIGAv1 can be obtained via removing the third term from CIGAv2. Recall that CIGAv2 has the following formulation:

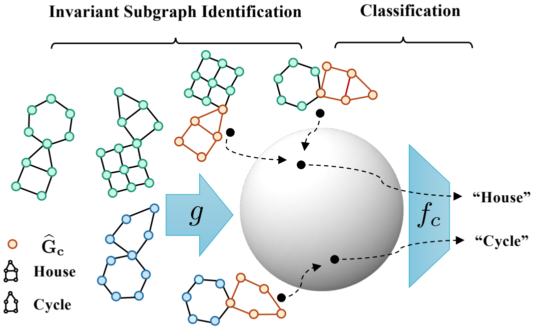


Figure 15: Illustration of **C**ausality **I**nspired **I**nvariant **G**raph Le**A**rning (CIGA): GNNs need to classify graphs based on the specific motif ("House" or "Cycle"). The featurizer g will extract an (orange colored) subgraph $\widehat{G}_c$ from each input for the classifier f_c to predict the label. The training objective of g is implemented in a contrastive strategy where the distribution of $\widehat{G}_c$ at the latent sphere will be optimized to maximize the intra-class mutual information. With the identified invariant subgraph G_c, the predictions made by classifier f_c based on G_c are invariant to distribution shifts;

$$\begin{aligned}\max_{f_c,g}\ I(\widehat{G}_c;Y)+I(\widehat{G}_s;Y),\ \text{s.t.}\ \widehat{G}_c\in&\mathop{\arg\max}_{\widehat{G}_c=g(G),\widetilde{G}_c=g(\widetilde{G})}I(\widehat{G}_c;\widetilde{G}_c|Y),\\&I(\widehat{G}_s;Y)\leq I(\widehat{G}_c;Y),\ \widehat{G}_s=G-g(G).\end{aligned}\tag{35}$$

where $\widehat{G}_c = g(G), \widetilde{G}_c = g(\widetilde{G})$ and $\widetilde{G} \sim P(G|Y)$, i.e., $\widetilde{G}$ and G have the same label. In Sec. 3.3, we introduce a contrastive approximation for $I(\widehat{G}_c; \widetilde{G}_c|Y)$:

$$I(\widehat{G}_c; \widetilde{G}_c|Y) \approx \mathbb{E}_{\substack{\{\widehat{G}_c, \widetilde{G}_c\} \sim \mathbb{P}_g(G|\mathcal{Y}=Y) \\ \{G_c^i\}_{i=1}^M \sim \mathbb{P}_g(G|\mathcal{Y}\neq Y)}} \log \frac{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})}}{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})} + \sum_i^M e^{\phi(h_{\widehat{G}_c} h_{G_c^i})}}, \tag{36}$$

where positive samples $(\widehat{G}_c, \widetilde{G}_c)$ are the extracted subgraphs of graphs that have the same label of G, negative samples are those with different labels, $\mathbb{P}_g(G|\mathcal{Y} = Y)$ is the pushforward distribution of $\mathbb{P}(G|\mathcal{Y} = Y)$ by featurizer g, $\mathbb{P}(G|\mathcal{Y} = Y)$ refers to the distribution of G given the label Y, $h_{\widehat{G}_c}, h_{\widetilde{G}_c}, h_{G_c^i}$ are the graph presentations of the estimated subgraphs, and ϕ is the similarity metric for the graph presentations. As $M \to \infty$, Eq. 36 approximates $I(\widehat{G}_c; \widetilde{G}_c|Y)$ which can be regarded as a non-parameteric resubstitution entropy estimator via the von Mises-Fisher kernel density [1, 41, 101].

While for the third term $I(\widehat{G}_s; Y)$ given the constraint $I(\widehat{G}_s; Y) \leq I(\widehat{G}_c; Y)$, a straightforward implementation is to imitate the hinge loss:

$$\frac{1}{N} R_{\widehat{G}_s} \cdot \mathbb{I}(R_{\widehat{G}_s} \leq R_{\widehat{G}_c}), \tag{37}$$

where N is the number of samples, $\mathbb{I}$ is a indicator function that outputs 1 when the interior condition is satisfied otherwise 0, and $R_{\widehat{G}_s}$ and $R_{\widehat{G}_c}$ are the empirical risk vector of the predictions for each sample based on $\widehat{G}_s$ and $\widehat{G}_c$ respectively. One can also formulate Eq. 35 from game-theoretic perspective [14].

Finally, we can derive the specific loss for the optimization of CIGAv2 combining Eq. 36 and Eq. 37:

$$\begin{aligned} &R_{\widehat{G}_c} + \alpha \mathbb{E}_{\substack{\{\widehat{G}_c, \widetilde{G}_c\} \sim \mathbb{P}_g(G|\mathcal{Y}=Y) \\ \{G_c^i\}_{i=1}^M \sim \mathbb{P}_g(G|\mathcal{Y}\neq Y)}} \log \frac{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})}}{e^{\phi(h_{\widehat{G}_c}, h_{\widetilde{G}_c})} + \sum_i^M e^{\phi(h_{\widehat{G}_c} h_{G_c^i})}} \\ &+ \beta \frac{1}{N} R_{\widehat{G}_s} \cdot \mathbb{I}(R_{\widehat{G}_c} \leq R_{\widehat{G}_s}), \end{aligned} \tag{38}$$

where $R_{\widehat{G}_c}, R_{\widehat{G}_s}$ are the empirical risk when using $\widehat{G}_c, \widehat{G}_s$ to predict Y through the classifier. Typically, we use a additional MLP downstream classifier ρ_s for $\widehat{G}_s$ in the classifier GNN. $h_{\widehat{G}_c}$ is the graph representation of $\widehat{G}_c$ which can be induced from the GNN encoder either in the featurizer or in the classifier. α, β are the weights for $I(\widehat{G}_c; \widetilde{G}_c|Y)$ and $I(\widehat{G}_s; Y)$, and ϕ is implemented as cosine similarity. The optimization loss for CIGAv1 merely contains the first two terms in Eq. 38.

The detailed algorithm for CIGA is given in the Algorithm 1, assuming the $h_{\widehat{G}_c}$ is obtained via the graph encoder in f_c. Fig. 15 also shows a illustration of the working procedure of CIGA.

G Detailed Experimental Settings

In this section, we provide more details about our experimental settings in Sec. 4, including the dataset preparation, dataset statistics, implementations of baselines, selection of models and hyperparameters as well as evaluation protocols.

G.1 Details about the datasets

We provide more details about the motivation and construction method of the datasets that are used in our experiments. Statistics of the datasets are presented in Table 4.

SPMotif datasets. We construct 3-class synthetic datasets based on BAMotif [116, 58] following [104], where the model needs to tell which one of three motifs (House, Cycle, Crane) that the graph contains. For each dataset, we generate 3000 graphs for each class at the training set, 1000 graphs for each class at the validation set and testing set, respectively. During the construction, we merely inject the distribution shifts in the training data while keep the testing data and validation data without the biases. For structure-level shifts (**SPMotif-Struc**), we introduce the bias based

Algorithm 1 Pseudo code for the CIGA framework.

```
**Input:** Training graphs and labels $\mathcal{D}_{\text{tr}} = \{G_i, Y_i\}_{i=1}^N$; learning rate $l$; loss weights $\alpha, \beta$ required by Eq. 38; number of training epochs $e$; batch size $b$;
Randomly initialize parameters of $g, f_c, \rho_s$;
**for** $i = 1$ **to** $e$ **do**
    Sample a batch of graphs $\{G^j, Y^j\}_{j=1}^b$;
    Estimate the invariant subgraph for the batch: $\{\widehat{G}_c^j\}_{j=1}^b = g(\{G^j, Y^j\}_{j=1}^b)$;
    Make predictions based the estimated invariant subgraph: $\{\widehat{Y}^j\}_{j=1}^b = f_c(\{\widehat{G}_c^j\}_{j=1}^b)$;
    Calculate the empirical loss $R_{\widehat{G}_c}$ with $\{\widehat{Y}^j\}_{j=1}^b$;
    Fetch the graph representations of invariant subgraphs from $f_c$ as $\{h_{\widehat{G}_c^j}\}_{j=1}^b$;
    Calculate the contrastive loss $R_c$ with Eq. 36, where positive samples and negative samples are constructed from the batch;
    Obtain $\widehat{G}_s$ for the batch: $\{\widehat{G}_s^j\}_{j=1}^b = \{G^j - \widehat{G}_c^j\}_{j=1}^b$;
    Make predictions based on the $\widehat{G}_s$: $\{\widehat{Y}_s^j\}_{j=1}^b = \rho_s(\{\widehat{G}_s^j\}_{j=1}^b)$;
    Calculate the empirical loss $R_{\widehat{G}_s}$ with $\{\widehat{Y}_s^j\}_{j=1}^b$, and weighted as Eq. 37;
    Update parameters of $g, f_c, \rho_s$ with respect to $R_{\widehat{G}_c} + \alpha R_c + \beta R_{\widehat{G}_s}$ as Eq. 38;
**end for**
```

Table 4: Information about the datasets used in experiments. The number of nodes and edges are taking average among all graphs. MCC indicates the Matthews correlation coefficient.

Datasets	# Training	# Validation	# Testing	# Classes	# Nodes	# Edges	Metrics
SPMotif	9,000	3,000	3,000	3	44.96	65.67	ACC
PROTEINS	511	56	112	2	39.06	145.63	MCC
DD	533	59	118	2	284.32	1,431.32	MCC
NCI1	1,942	215	412	2	29.87	64.6	MCC
NCI109	1,872	207	421	2	29.68	64.26	MCC
SST5	6,090	1,186	2,240	5	19.85	37.70	ACC
Twitter	3,238	694	1,509	3	21.10	40.20	ACC
CMNIST-sp	40,000	5,000	15,000	2	56.90	373.85	ACC
DrugOOD-Assay	34,179	19,028	19,032	2	32.27	70.25	ROC-AUC
DrugOOD-Scaffold	21,519	19,041	19,048	2	29.95	64.86	ROC-AUC
DrugOOD-Size	36,597	17,660	16,415	2	30.73	66.90	ROC-AUC

on FIIF, where the motif and one of the three base graphs (Tree, Ladder, Wheel) are artificially (spuriously) correlated with a probability of various biases, and equally correlated with the other two. Specifically, given a predefined bias b, the probability of a specific motif (e.g., House) and a specific base graph (Tree) will co-occur is b while for the others is $(1-b)/2$ (e.g., House-Ladder, House-Wheel). We use random node features for SPMotif-Struc, in order to study the influences of structure level shifts. Moreover, to simulate more realistic scenarios where both structure level and topology level have distribution shifts, we also construct **SPMotif-Mixed** for mixed distribution shifts. We additionally introduced FIIF attribute-level shifts based on SPMotif-Struc, where all of the node features are spuriously correlated with a probability of various biases by setting to the same number of corresponding labels. Specifically, given a predefined bias b, the probability that all of the node features of a graph has label y (e.g., $y = 0$) being set to y (e.g., $\boldsymbol{X} = \mathbf{0}$) is b while for the others is $(1-b)/2$ (e.g., $P(\boldsymbol{X} = \mathbf{1}) = P(\boldsymbol{X} = \mathbf{2}) = (1-b)/2$). More complex distribution shift mixes can be studied following our construction approach, which we will leave for future works.

TU datasets. To study the effects of graph sizes shifts, we follow Yehudai et al. [113], Bevilacqua et al. [11] to study the OOD generalization abilities of various methods on four of TU datasets [67], i.e., **PROTEINS, DD, NCI1, NCI109**. Specifically, we use the data splits generated by Yehudai et al. [113] and use the Matthews correlation coefficient as evaluation metric following [11] due to the class imbalance in the splits. The splits are generated as follows: Graphs with sizes smaller than the 50-th percentile are assigned to training, while graphs with sizes larger than the 90-th percentile are assigned to test. A validation set for hyperparameters tuning consists of 10% held out examples from training. We also provide a detailed statistics about these datasets in table 5.

Table 5: Detailed statistics of selected TU datasets. Table from Yehudai et al. [113], Bevilacqua et al. [11].

	NCI1			NCI109		
	all	Smallest 50%	Largest 10%	all	Smallest 50%	Largest 10%
Class A	49.95%	62.30%	19.17%	49.62%	62.04%	21.37%
Class B	50.04%	37.69%	80.82%	50.37%	37.95%	78.62%
Num of graphs	4110	2157	412	4127	2079	421
Avg graph size	29	20	61	29	20	61

	PROTEINS			DD		
	all	Smallest 50%	Largest 10%	all	Smallest 50%	Largest 10%
Class A	59.56%	41.97%	90.17%	58.65%	35.47%	79.66%
Class B	40.43%	58.02%	9.82%	41.34%	64.52%	20.33%
Num of graphs	1113	567	112	1178	592	118
Avg graph size	39	15	138	284	144	746

Graph-SST datasets. Inspired by the data splits generation for studying distribution shifts on graph sizes, we split the data curated from sentiment graph data [122], that converts sentiment sentence classification datasets **SST5** and **SST-Twitter** [90, 26] into graphs, where node features are generated using BERT [25] and the edges are parsed by a Biaffine parser [32]. Our splits are created according to the averaged degrees of each graph. Specifically, we assign the graphs as follows: Those that have smaller or equal than 50-th percentile averaged degree are assigned into training, those that have averaged degree large than 50-th percentile while smaller than 80-th percentile are assigned to validation set, and the left are assigned to test set. For SST5 we follow the above process while for Twitter we conduct the above split in an inversed order to study the OOD generalization ability of GNNs trained on large degree graphs to small degree graphs.

CMNIST-sp. To study the effects of PIIF shifts, we select the ColoredMnist dataset created in IRM [4]. We convert the ColoredMnist into graphs using super pixel algorithm introduced by Knyazev et al. [46]. Specifically, the original Mnist dataset are assigned to binary labels where images with digits $0-4$ are assigned to $y=0$ and those with digits $5-9$ are assigned to $y=1$. Then, y will be flipped with a probability of 0.25. Thirdly, green and red colors will be respectively assigned to images with labels 0 and 1 an averaged probability of 0.15 (since we do not have environment splits) for the training data. While for the validation and testing data the probability is flipped to 0.9.

DrugOOD datasets. To evaluate the OOD performance in realistic scenarios with realistic distribution shifts, we also include three datasets from DrugOOD benchmark. DrugOOD is a systematic OOD benchmark for AI-aided drug discovery, focusing on the task of drug target binding affinity prediction for both macromolecule (protein target) and small-molecule (drug compound). The molecule data and the notations are curated from realistic ChEMBL database [63]. Complicated distribution shifts can happen on different assays, scaffolds and molecule sizes. In particular, we select `DrugOOD-lbap-core-ic50-assay`, `DrugOOD-lbap-core-ic50-scaffold`, and `DrugOOD-lbap-core-ic50-size`, from the task of Ligand Based Affinity Prediction which uses `ic50` measurement type and contains `core` level annotation noises. For more details, we refer interested readers to Ji et al. [40].

G.2 Training and Optimization in Experiments

During the experiments, we do not tune the hyperparameters exhaustively while following the common recipes for optimizing GNNs. Details are as follows.

GNN encoder. For fair comparison, we use the same GNN architecture as graph encoders for all methods. By default, we use 3-layer GNN with Batch Normalization [39] between layers and JK residual connections at last layer [106]. For the architectures we use the GCN with mean readout [45] for all datasets except Proteins where we empirically observe better validation performance with a GIN and max readout [107], and for DrugOOD datasets where we follow the backbone used in the paper [40], i.e., 4-layer GIN with sum readout. The hidden dimensions are fixed as 32 for SPMotif, TU datasets, CMNIST-sp, and 128 for SST5, Twitter and DrugOOD datasets.

Optimization and model selection. By default, we use Adam optimizer [43] with a learning rate of $1e-3$ and a batch size of 32 for all models at all datasets. Except for DrugOOD datasets, we use a batch size of 128 following the original paper [40]. To avoid underfitting, we pretrain models for 20 epochs for all datasets, except for CMNIST and Twitter where we pretrain 5 epochs and for SST5 we pretrain 10 epochs, because of the dataset size and the difficulty of the task. To avoid overfitting, we also employ an early stopping of 5 epochs according to the validation performance. Meanwhile, dropout [91] is also adopted for some datasets. Specifically, we use a dropout rate of 0.5 for CMNIST, SST5, Twitter, DrugOOD-Assay and DurgOOD-Scaffold, 0.1 for DrugOOD-Size according to the validation performance, and 0.3 for TU datasets following the practice of Bevilacqua et al. [11].

Implementations of baselines. For implementations of the interpretable GNNs, we use the author released codes [120, 78], where we use the codes provided by the authors[6] for DIR c[104] which is the same as the author released codes. During the implementation, we use the same s_c for all interpretable GNN baselines, chosen from $\{0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ according to the validation performances, and set to 0.25 for SPMotif following Wu et al. [104], 0.3 for Proteins and DD, 0.6 for NCI1, 0.7 for NCI109, 0.8 for CMNIST-sp, 0.5 for SST5 and Twitter, and 0.8 for DrugOOD datasets, respectively. Empirically, we observe that the optimization process in GIB can be unstable during its nested optimization for approximating the mutual information of the predicted subgraph and the input graph. We use a larger batch size of 128 or reduce the nested optimization steps to be lower than 20 for stabilizing the performance. If the optimization failed due to the instability during training, we will select the results with best validation accuracy as the final outcomes. Although SPMotif-Struc is also evaluated in DIR, we find the results are inconsistent to the results reported by the author, because DIR adopts `Last Epoch Model Selection` which is *different* from the claim that they select models according to `the validation performance`, i.e., `line` 264 to `line` 278 in `train/spmotif_dir.py` from the commit `4b975f9b3962e7820d8449eb4abbb4cc30c1025d` of `https://github.com/Wuyxin/DIR-GNN`. We select the hyperparamter for the proposed DIR regularization from $\{0.01, 0.1, 1, 10\}$ according to the validation performances at the datasets, while we stick to the authors claimed hyperparameters for the datasets they also experimented with.

For invariant learning, we refer to the implementations in DomainBed [34] for IRM [4], V-Rex [49] and IB-IRM [2]. Since the environment information is not available, we perform random partitions on the training data to obtain two equally large environments for these objectives. Moreover, we select the weights for the corresponding regularization from $\{0.01, 0.1, 1, 10, 100\}$ for these objectives according to the validation performances of IRM and stick to it for others, since we empirically observe that they perform similarly with respect to the regularization weight choice. For EIIL [23], we use the author released implementations about assigning different samples the weights for being put in each environment and calculating the IRM loss.

Besides, for CNC [124], we follow the algorithm description to modify the sampling strategy in supervised contrastive loss [42] based on a pretrained GNN optimized with ERM, and choose the weight for contrastive loss using the same grid search as for CIGA.

Implementations of CIGA. For fair comparison, CIGA uses the same GNN architecture for GNN encoders as the baseline methods. We did not do exhaustive hyperparameters tuning for the loss Eq. 38. By default, we fix the temperature to be 1 in the contrastive loss, and merely search α from $\{0.5, 1, 2, 4, 8, 16, 32\}$ and β from $\{0.5, 1, 2, 4\}$ according to the validation performances. For CMNIST-sp, we find larger β are required to get rid of intense spurious node features hence we expand the search range for β to $\{0.5, 1, 2, 4, 16, 32\}$, For Graph-SST datasets, we search α from $\{0.5, 1, 2, 4\}$ as we empirically find that increasing α does not help increase the performance with few random runs. Besides, we also have various implementation options for obtaining the features in $\widehat{G}_c$, for obtaining $h_{\widehat{G}_c}$, as well as for obtaining predictions based on $\widehat{G}_s$. By default, we feed the graph representations of featurizer GNN to the classifier GNN, as well as to the contrastive loss. For classifying G based on $\widehat{G}_s$, we use a separate MLP downstream classifier in the classifier GNN f_c. The only exception is for the CMNIST-sp dataset where the spurious correlation is stronger than the invariant signal. Directly feeding the graph representations from the featurizer GNN can easily overfit to the shortcuts hence we instead feed the original features to the downstream classifier GNN. There can be more other options, such as using separate graph convolutions on $\widehat{G}_s$ or $\widehat{G}_c$, which we leave for future work.

Evaluation protocol. We run each experiment 10 on TU datasets and 5 times for others where the random seeds start from 1 to the number of total repeated times. During each run, we select the

[6] `https://anonymous.4open.science/r/DIR/`

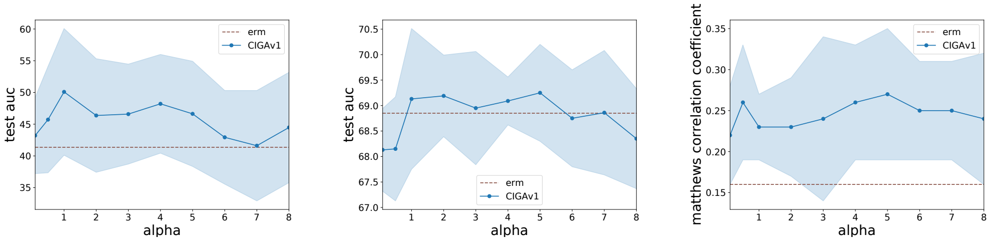
model according to the validation performance and report the mean and standard deviation of the corresponding metrics.

G.3 Software and Hardware

We implement our methods with PyTorch [73] and PyTorch Geometric [29]. We ran our experiments on Linux Servers with 40 cores Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz, 256 GB Memory, and Ubuntu 18.04 LTS installed. GPU environments are varied from 4 NVIDIA RTX 2080Ti graphics cards with CUDA 10.2, 2 NVIDIA RTX 2080Ti and 2 NVIDIA RTX 3090Ti graphics cards with CUDA 11.3, and NVIDIA TITAN series with CUDA 11.3.

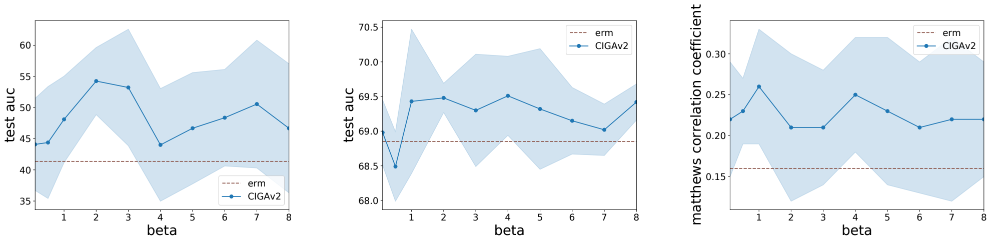
G.4 Additional Analysis

Hyperparameter sensitivity analysis. To examine how sensitive CIGA is to the hyperparamters α and β for contrastive loss and hinge loss, respectively, under different distribution shifts. We conduct experiments based on the hardest datasets from each table (i.e., SPMotif-Mixed with the bias of 0.9, DrugOOD-Scaffold and the NCI109 datasets from Table 1, Table 2, and Table 3, respectively.) To increase the difficulty, we search for more fine-grained spaces for both parameters, i.e., $\{0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8\}$. During changing the value of β, we will fix the α to a specific value under which the model has a relatively good performance (but not the best, to fully examine the robustness of CIGA in practice). During the sensitivity tests, we follow the evaluation protocol as that used for the main experiments. The results are shown in Fig. 16 and Fig. 17.



(a) SPMotif-Mixed under bias= 0.9 (b) DrugOOD-Scaffold (c) NCI109

Figure 16: Hyperparameter sensitivity analysis on the coefficient of contrastive loss (α).



(a) SPMotif-Mixed under bias= 0.9 with $\alpha = 4$ (b) DrugOOD-Scaffold with $\alpha = 1$ (c) NCI109 with $\alpha = 1$

Figure 17: Hyperparameter sensitivity analysis on the coefficient of hinge loss (β).

From the results above, we can see that both CIGAv1 and CIGAv2 are robust to different values of α and β, respectively, across different datasets and distribution shifts. Notably, in Fig. 16, when the coefficient α for the contrastive loss become too small, the invariance of the identified invariant subgraphs $\widehat{G}_c$ may not be guaranteed, resulting worse performances. Moreover, when α becomes too large, it may affect the optimization and yield worse performances. In SPMotif datasets, the worse performances can be observed via the large variances as well. Similarly for β, as shown in Fig. 17, when β becomes too small, some part from the spurious subgraph may still be contained in the estimated invariant subgraphs. While if β becomes too large, there might be part of $\widehat{G}_c$ being eliminated. Although both CIGAv1 and CIGAv2 are robust to the changes of α and β, the intrinsic difficult optimization in OOD generalization algorithms including the proposed CIGA in our work, still require a more proper and smooth optimization process [18].

Table 6: Averaged training time (sec.) per epoch of various methods on DrugOOD-Scaffold.

Methods	ERM	ASAP	GIB	DIR	IRM	EIIL	CNC	CIGAv1	CIGAv2
Running time	8.055	15.578	300.304	106.919	8.73	69.664	9.795	40.065	46.181
OOD Performance	68.85	66.19	62.01	63.91	68.69	68.45	67.24	69.04	69.7
Avg. Rank	2	5.5	9	8	3	6	4.5	3.5	3.5

Running time analysis. To examine how much computational overhead is induced by the architecture and the additional objectives in CIGA, we analyze and compare the averaged training time of different methods on DrugOOD-Scaffold. Factors that could affect the running time such as GNN backbone, batch size, and the running devices (NVIDIA RTX 2080Ti, Linux Servers with 40 cores Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz, 256 GB Memory, and Ubuntu 18.04 LTS), are fixed the same during the testing. The results are shown as in Table. 6. It can be found that CIGA is the only OOD method that outperforms ERM by a non-trivial margin with a relatively low additional computational overhead.

Table 7: Performances of different methods on Drug-Assay under single environment OOD generalization (i).

Methods	ERM	ASAP	GIB	DIR	CIGAv1	CIGAv2	Oracle (IID)
OOD Performance	63.29(2.67)	63.41(0.70)	62.72(0.59)	62.56(0.79)	**63.86 (0.57)**	**64.31 (0.92)**	84.71 (1.60)
Rank	5	4	8	9	2	1	

Table 8: Performances of different methods on Drug-Assay under single environment OOD generalization (ii).

Methods	ERM	IRM	V-Rex	EIIL	IB-IRM	CNC	CIGAv1	CIGAv2	Oracle (IID)
OOD Performance	63.29(2.67)	63.25(1.45)	62.18(1.71)	62.95(1.37)	61.95(1.72)	63.61(0.96)	**63.86 (0.57)**	**64.31 (0.92)**	84.71 (1.60)
Rank	5	6	10	7	11	3	2	1	

Single environment OOD generalization. The theory of invariant learning fundamentally assume the presence of multiple environments [76, 4]. However in practice, it does not always hold, which would inevitably fail all of the invariant learning solutions [4, 49, 23, 2], including CIGA.

Nevertheless, to examine how CIGA performs under various realistic scenarios, we conduct an additional experiment based on DrugOOD-Assay. We select samples that are from the largest assay group (i.e., the biochemical functionalities of these molecules are tested and reported under the same experimental setup in the lab) [40]. The results are separated and shown in Table 7 and Table 8. Besides the baselines, we also show the "Oracle" performances from the main table, to demonstrate the performance gaps.

From the Table 7 and Table 8, we can see that, both CIGAv1 and CIGAv2 maintain their state-of-the-art performances even in the single training environment setting. We hypothesize that enforcing the mutual information between the estimated $\widehat{G}_c$ also helps to retain the invariance even under the single training environment setting. That may partially explain why CNC can bring some improvements. We believe it is an interesting and promising future direction to develop in-depth understanding and better solutions under this circumstance.

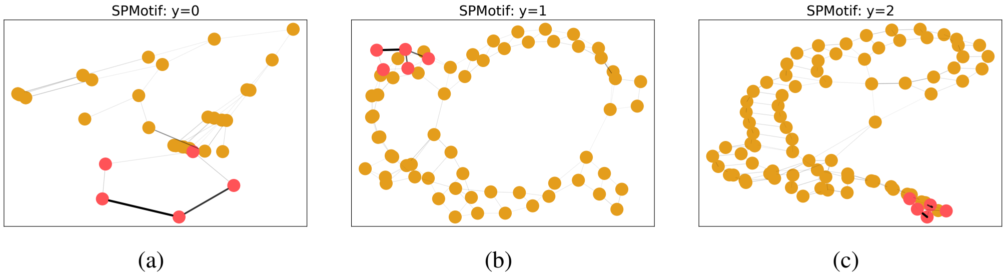
G.5 Interpretation Visualization

Since we use the interpretable GNN architecture to implement CIGA[7], it brings an additional benefit that provides certain interpretation for the predictions automatically, which may facilitate human understanding in practice.

[7] We use the code provided by [64].

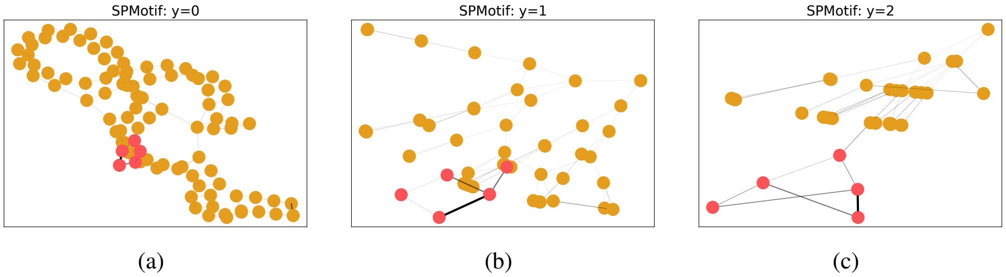
First, we provide some interpretation visualizations in SPMotif-Struc and SPMotif-Mixed datasets, under the biases of 0.6 and 0.9. Shown in Fig. 18 to Fig. 21, we use pink to color the ground truth nodes in G_c, and denote the relative attention strength with edge color intensities.

Besides, we also provide some interpretation visualization examples in DrugOOD datasets. Shown in Fig. 22 to Fig. 27, we use the edge color intensities to denote the attentions of models that pay to the corresponding edge. Some interesting patterns can be found in the molecules shared with the same label, which could provide insights to the domain experts when developing new drugs. We believe that, because of its superior OOD generalization performance on graphs, CIGA can have high potential to push forward the developments of AI-Assisted Drug Discovery, and enrich the AI tools for facilitating the fundamental practice of science in the future.



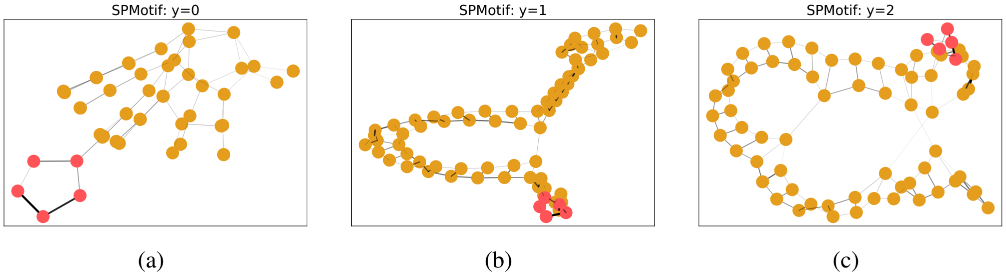
(a) (b) (c)

Figure 18: Interpretation visualization of examples from SPMotif-Struc under bias$= 0.6$.



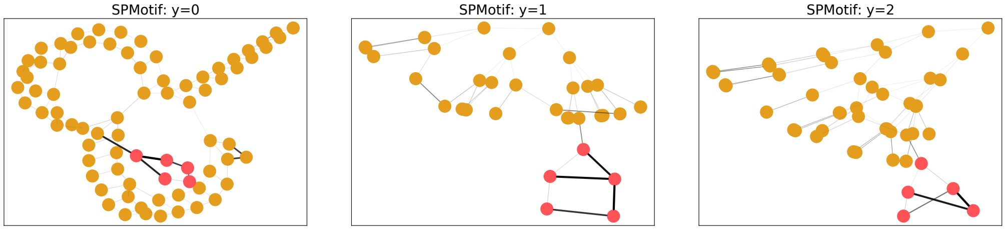
(a) (b) (c)

Figure 19: Interpretation visualization of examples from SPMotif-Struc under bias$= 0.9$.



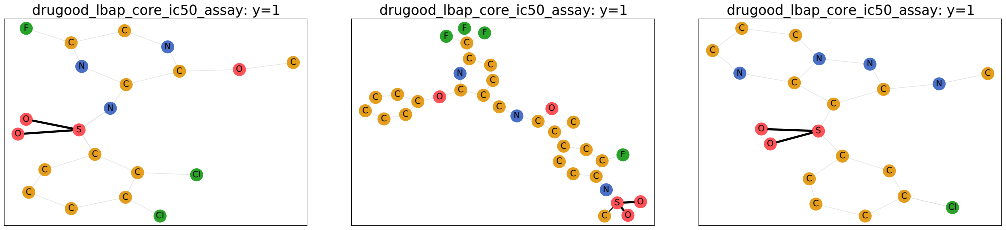
(a) (b) (c)

Figure 20: Interpretation visualization of examples from SPMotif-Mixed under bias$= 0.6$.



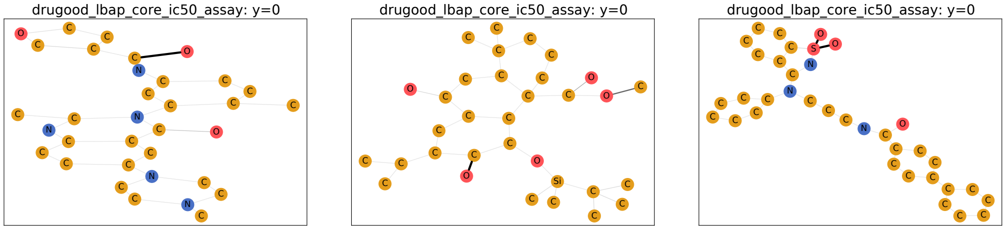
(a) (b) (c)

Figure 21: Interpretation visualization of examples from SPMotif-Mixed under bias$= 0.9$.


drugood_lbap_core_ic50_assay: y=1
drugood_lbap_core_ic50_assay: y=1
drugood_lbap_core_ic50_assay: y=1

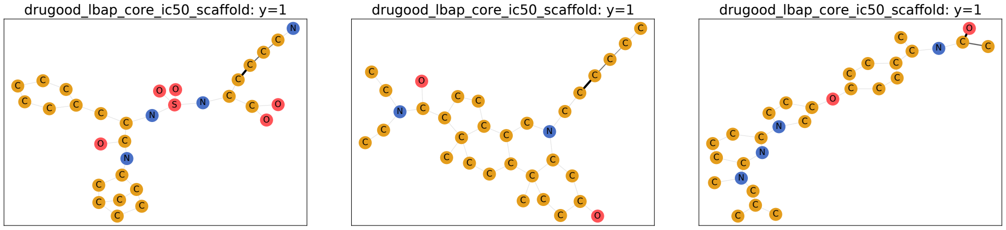
(a) (b) (c)

Figure 22: Interpretation visualization of activate examples ($y = 1$) from DrugOOD-Assay.


drugood_lbap_core_ic50_assay: y=0
drugood_lbap_core_ic50_assay: y=0
drugood_lbap_core_ic50_assay: y=0

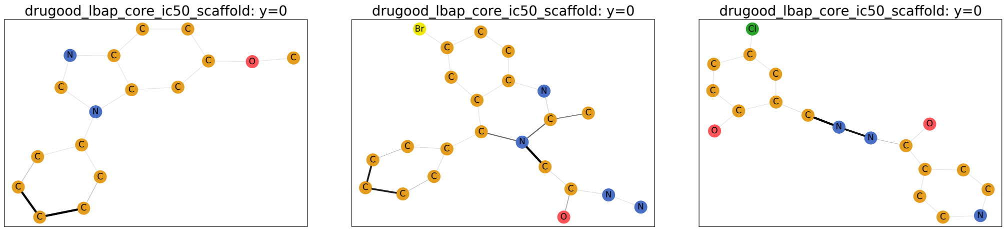
(a) (b) (c)

Figure 23: Interpretation visualization of inactivate examples ($y = 0$) from DrugOOD-Assay.


drugood_lbap_core_ic50_scaffold: y=1
drugood_lbap_core_ic50_scaffold: y=1
drugood_lbap_core_ic50_scaffold: y=1

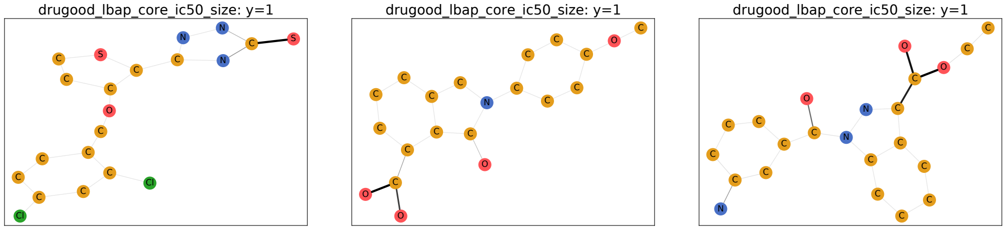
(a) (b) (c)

Figure 24: Interpretation visualization of activate examples ($y = 1$) from DrugOOD-Scaffold.


drugood_lbap_core_ic50_scaffold: y=0
drugood_lbap_core_ic50_scaffold: y=0
drugood_lbap_core_ic50_scaffold: y=0

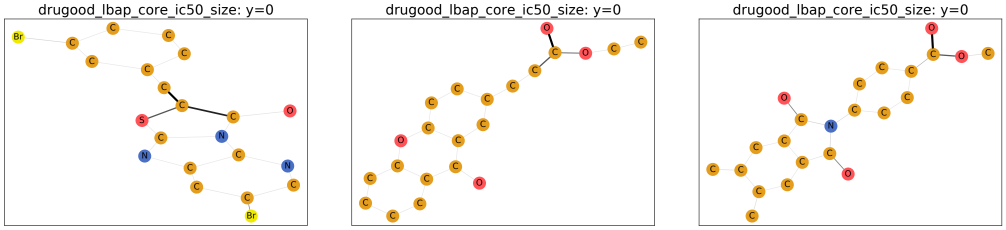
(a) (b) (c)

Figure 25: Interpretation visualization of inactivate examples ($y = 0$) from DrugOOD-Scaffold.


drugood_lbap_core_ic50_size: y=1
drugood_lbap_core_ic50_size: y=1
drugood_lbap_core_ic50_size: y=1

(a) (b) (c)

Figure 26: Interpretation visualization of activate examples ($y = 1$) from DrugOOD-Size.


drugood_lbap_core_ic50_size: y=0
drugood_lbap_core_ic50_size: y=0
drugood_lbap_core_ic50_size: y=0

(a) (b) (c)

Figure 27: Interpretation visualization of inactivate examples ($y = 0$) from DrugOOD-Size.