
Improving Graph Out-of-distribution Generalization Beyond Causality

Can Xu

East China Normal University
Shanghai, China

Yao Cheng

East China Normal University
Shanghai, China

Jiangxiang Yu

East China Normal University
Shanghai, China

Haosen Wang

Southeast University
Nanjing, China

Jingsong Lv

Zhejiang Lab
Hangzhou, China

Yao Liu

East China Normal University
Shanghai, China

Xiang Li*

East China Normal University
Shanghai, China

Abstract

Existing causal methods for graph out-of-distribution (OOD) generalization primarily assume the causal relationships between invariant sub-graphs and labels, which neglect the unignorable role of environment. However, in complex real-world applications, the assumption is easily violated. To address the problem, this paper presents the theorems of environment-label dependency and mutable rationale invariance, where the former characterizes the usefulness of environments in determining graph labels while the latter refers to the mutable importance of graph rationales. Based on analytic investigations, a novel variational inference based method DEROG (Probability Dependency on Environments and Rationales for OOD Graphs) is introduced. To alleviate the adverse effect of unknown prior knowledge on environments and rationales, DEROG utilizes generalized Bayesian inference. Further, DEROG employs an EM-based algorithm for optimization. Finally, extensive experiments on benchmark datasets under different distribution shifts are conducted to show the superiority of DEROG. Our code is publicly available at <https://github.com/LEOXC1571/DEROG>.

1 Introduction

Generalizing graph neural networks (GNNs) to out-of-distribution (OOD) graphs has recently received increasing attention. Most existing studies [1, 2] tackle the problem from the perspective of causal inference and invariant learning, which aim at dissecting graphs into two non-overlapping parts, i.e., *causal sub-graphs* and *spurious sub-graphs*. Despite the difference in generating spurious sub-graphs, previous methods [3–5] generally assume that causal sub-graphs are invariant under distribution shifts and are used to determine properties / labels of graphs, while spurious sub-graphs characterize different distribution shifts and exert marginal impact on graph label prediction. However, in practice, the relationships among causal sub-graphs (graph rationales²), environments, and labels are intricate. Therefore, existing methods could encounter major challenges when applied in complex scenarios.

*Corresponding author

²In causal methods, they use causal sub-graphs to denote the graph rationales. Since our proposed method is not causal-based, we directly use graph rationales for distinction.

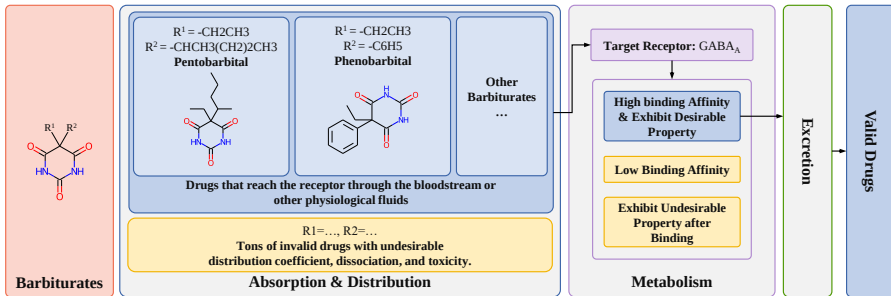


Figure 1: The process of drugs to take effect through binding with receptors. This figure demonstrates that there are a number of factors other than pharmacophore that determine validity and properties of drugs in real-world scenarios.

A crucial real-world application for graph OOD generalization is molecular learning. To illustrate, Fig. 1 shows an example of the process for structurally specific drugs to take effect. At the stage of absorption and distribution, for drugs with the same pharmacophore (i.e., Barbiturates) as the sub-structure with the maximum efficacy information, both R^1 and R^2 are motifs that are considered superfluous according to previous studies [3]. In fact, R^1 (R^2) determines whether drugs are able to reach target receptors and serves as the antecedent of valid drugs. This indicates that instead of being independent, properties of graphs are dependent on environments at certain level in real-world scenarios. At the metabolism stage, once drugs with the same pharmacophore successfully reach target receptors, factors like experimental assays, electron density distributions, geometry conformations, and metabolites also influence the validity of drugs. This demonstrates that causal sub-graphs could fail to exhibit rigorous invariance across different environments in predicting properties of graphs.

In this paper, we delve into graph OOD generalization beyond causality and propose a probabilistic graphical model DEROG, which **D**epends on both **E**nvironments and **R**ationales to predict labels for **O**OD **G**raphs. Specifically, we model environments and graph rationales as hidden variables to capture the shift and consistency of data distributions, respectively. Different from previous causal methods that assume the independence of labels on environments and the rigorous invariance of rationales, DEROG characterizes the complex relationships among environments, rationales, and graph labels from three aspects. First, the distribution of environments is characterized as the non-linear transformation of the joint distribution of graphs and labels. Second, due to the loose invariance of causal sub-graphs across various environments, we construct rationales based on both graphs and environments. Third, since environments can also determine labels, we jointly utilize environments, rationales, and graphs to predict overall properties.

Technically, DEROG adopts an Expectation-Maximization (EM) based optimization framework. In the E-step, the parameters w.r.t. environments and rationales are optimized, while in the M-step, environments and rationales are fixed and used to improve the final classifier. Due to the limited knowledge regarding the prior distributions of environments and rationales, we leverage the generalized Bayesian inference technique, which substitutes the KL-divergence terms in the objective function with negative entropy. Further, we upgrade DEROG by adding an environment alignment loss and a node-level contrastive loss, respectively. For the former, each graph could have a corresponding environment label that can be used to generate high-quality environment embeddings. For the latter, we sample positive and negative nodes based on the influence they exert in the final classifier, and the optimization of the contrastive loss leads to more expressive graph rationales. Finally, we summarize our main contributions as follows:

- We perform systematic analyses on the relationships among environments, graph rationales, and labels in the context of graph OOD generalization. We identify two pitfalls when performing environment inference and sub-graph extraction, and propose *environment-label dependency* and *mutable rationale invariance* that better suit the conditions on real-world data.
- We present a novel probabilistic graphical model based on both environment-label dependency and mutable rationale invariance. We leverage generalized Bayesian inference for hidden variables and put forward an EM-based framework for optimization.
- We conduct extensive experiments to evaluate the efficacy of DEROG on various real-world benchmark datasets under different distribution shifts. Our results show the significant advantages of

DEROG over other competitors. Also, DEROG can perform well on synthetic benchmarks, although these datasets are specially tailored for different SOTA models.

2 Related Work

The quest for a more robust and effective deep neural network leads to wide research interest on OOD generalization. Common methods for OOD generalization include robust optimization [6, 7], domain adaption [8–10] and invariant learning [11–13].

For the graph OOD problem, a number of causal inference methods [4, 14–16] aim at extracting invariant parts of graphs and then maximizing the correlations between these sub-structures and labels. For example, LECI [17] makes thorough exploration on the independence relation among invariant sub-graphs, environments and labels. Instead of focusing on extracting causal sub-graphs, some studies explore the roles of environments and spurious sub-graphs in invariant learning. GIL [18] addresses the importance of inferring environment labels, which provide prior assumptions on distributions. GALA [19] addresses two pitfalls in causal learning and combines CIGA [3] with pre-trained ERM to generate faithful environments. On node-level tasks, CaNet [20] points out that relations between ego-graphs and labels exhibit instability caused by confounding bias of latent environments. Therefore, a pseudo environment estimator is introduced to mitigate the gap. MoleOOD [21] is a recent variational inference based approach. However, it does not take the inter-dependency between variables into account when inferring environments. Also, it assumes consistent invariance of causal substructure across environments.

Further, there are also methods that tackle OOD generalization by data augmentation. For example, AIA [22] improves performance under covariate shift by augmenting input graphs with various spurious structures. IGM [23] combines invariant sub-graph with variant patterns to reduce the reliance on accurate environment labels. From a data-centric perspective, G-Splice [16] proposes an environment-aware OOD graph generation method with non-Euclidean-space linear extrapolation designed in both structural and feature space.

3 Preliminary

3.1 Problem Formulation

Let $\mathcal{Y}$, $\mathcal{G}$, and $\mathcal{E}$ denote graph label set, graph set, and environment set, respectively. The problem of graph OOD generalization aims to predict labels $Y \in \mathcal{Y}$ of corresponding graphs $G \in \mathcal{G}$ collected from various environments $E \in \mathcal{E}$. Distribution shifts on $P(G, Y) = P(G)P(Y|G)$ can be categorized into *covariate shift* and *concept shift*, depending on whether the shift occurs on $P(G)$ or $P(Y|G)$. Since shifts on the joint distribution $P(G, Y)$ exist between training and test set, an optimal GNN classifier $\Phi(G)$ is supposed to predict labels with the minimum amount of risks, which is represented as $\Phi(\cdot) = \arg \min_{\Phi} \mathbb{E}_{G, Y} [l(\Phi(G), Y)]$.

3.2 Preliminary Study

In Section 1, we claimed that causal sub-graphs cannot independently determine properties of graphs under distribution shifts by a toy example in computational chemistry. We next further substantiate the following two claims with preliminary experimental results. First, causal sub-graph extraction is not a sufficient condition for desired OOD generalization performance. Second, the capability of SOTA causal learning models in extracting causal sub-graphs does not meet the expectations of researchers when applied to complex real-world datasets.

As the current best performer in graph OOD domain, LECI [17] extracts causal sub-graphs through the following process. First, it uses a GNN to obtain node representations. Then, it combines node representations from both ends of an edge and aggregates them through a Multi-Layer Perceptron (MLP) to derive a one-dimensional edge score. Subsequently, such edge score is passed through a Gumbel-Sigmoid activation function to obtain a mask weight that is close to 0 or 1, which can be used to well separate causal sub-graphs from spurious ones. Ideally, for edges that connect nodes in causal sub-graphs, their weights should be close to 1, while that of other edges in the graph should be close to 0. Finally, these edge mask weights are used to control the strength of message propagation in subsequent GNN classifiers.

We next inspect edge mask weights derived by LECI under different distribution shifts on three real-world graph OOD benchmark datasets. Details on these distribution shifts and datasets will be given in Sec. 5. Table 1 summarizes the *min*, *max* and *mean* values of mask weights w.r.t. all the edges for graphs in the test set. From the table, we see that LECI assigns relatively large weights to all edges and the variance is small in all cases. Notably, in the HIV-Scaffold-Covariate experiment, LECI assigns identical weights to 166,694 edges across 4,108 graphs from the test set. These results show that all the edges are considered important and the whole graphs are considered as causal sub-graphs. However, **this phenomenon contradicts with the presence of distribution shifts**. Further, LECI can still achieve the best performance among all causal methods (see results in Table 2 later). All these observations show that superior OOD generalization performance does not necessarily require causal sub-graph extraction. Additionally, existing causal learning methods are less capable in effectively extracting causal sub-graphs that determine graph properties. More preliminary study results are given in Appendix B. *These observations motivates us to study non-causal methods that do not distinguish causal and spurious sub-graphs.*

3.3 Environments and Rationales

Previous invariant learning methods encounter two major pitfalls: one is the idealized assumption of independence between environments and labels, while the other is the excessive emphasis on the invariance of causal sub-graphs. These assumptions can be easily satisfied on most synthetic datasets, where certain sub-graphs determine the overall properties and remaining parts of graphs are completely superfluous.

Table 1: Statistics of Edge Mask Weights in LECI

Dataset	HIV		LBAP-core-IC50	Twitter
Domain	Scaffold	Size	Assay	Length
Shift	Covariate	Concept	Covariate	Concept
Min.	0.7068	0.8392	0.9042	0.9031
Max.	0.7068	0.8874	0.9768	0.9957
Mean	0.7068	0.8443	0.9434	0.9822

However, real-world cases are more complex where these assumptions are more likely to fail (See Fig. 1). Therefore, this paper places great importance on the influence of environments on invariant graph pattern learning and we consider E as a variable highly correlated with the joint distribution (G, Y) . Formally, we have:

Proposition 3.1. *According to the definition of covariate shift and concept shift [17], as elaborated in Proof C.1, shifts on distributions of environment reflect shifts on distributions of samples. We have $P(E) := P(G)$ under covariate shift and $P(E) := P(Y|G)$ under concept shift. Therefore, $P(E) := P(G, Y)$ holds for distribution shifts on graphs.*

The widely used real-world datasets for graph OOD generalization are molecule and NLP-related datasets [24, 25]. For the former, we have illustrated the correlations between environments and labels, and the inconsistent invariance of causal sub-graphs in Fig. 1 with an illustration of the process of drug efficacy. Compared to other structural components within a molecule, pharmacophores indeed play a decisive role in binding with receptors and triggering desirable properties. But the efficacy of molecules with the same pharmacophoric group is dependent on a number of influential factors. For the latter, text sequences are converted into graphs where nodes represent words and edges stand for correlations between words. Due to the existence of irony, paradox, and contextual information, phenomena that labels are not entirely determined by invariant sub-graphs are also widely observed. To sum up, graph rationales are not the sole decisive factor for labels and environments are not completely irrelevant to labels either. Therefore, we next draw Theorem 3.2 and 3.3 that describe the non-negligible dependency of labels on environments and the mutable invariance of graph rationales across varying environments, respectively.

Theorem 3.2. (Environment-label dependency) *Let (G_i, Y_i) and (G_j, Y_j) denote the graph-label pairs sampled from the training and test sets, respectively. Further, $Q_{\Phi_{\mathcal{E}}}$ represents the invariant rationale extractor optimized on the training set. Given the rationales $\hat{G}_i = Q_{\Phi_{\mathcal{E}}}(G_i)$ and $\hat{G}_j = Q_{\Phi_{\mathcal{E}}}(G_j)$, there exists an optimal $Q_{\Phi_{\mathcal{E}}^*}$ such that $\Phi_{\mathcal{E}}^* = \arg \max I(\hat{G}_i; Y_i)$ and $I(G_j \setminus \hat{G}_j; Y_j) > 0$.*

Generally speaking, even if optimal graph rationales are obtained by a sub-graph extractor trained on data across multiple environments, the remaining sub-graphs still influence the graph labels at a small but non-negligible level. Put it briefly, the stability of invariance exhibited by graph rationales varies across different conditions.

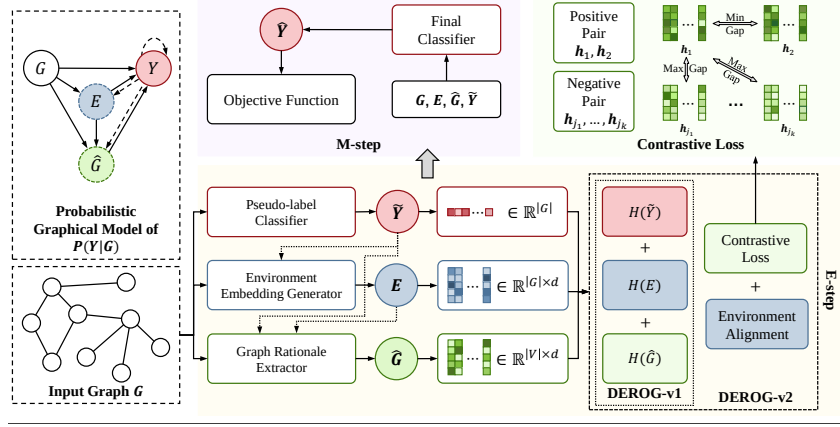


Figure 2: The overall framework of the proposed method DEROG.

Theorem 3.3. (Mutable rationale invariance) Given a graph G with rationale $\hat{G}$, $\exists \mathcal{E}_{sub} \subseteq \mathcal{E}$ and $E_k \in \mathcal{E} \setminus \mathcal{E}_{sub}$, $\forall E_i, E_j \in \mathcal{E}_{sub}$, we have $P(Y|\hat{G}, E_i) = P(Y|\hat{G}, E_j) \neq P(Y|\hat{G}, E_k)$.

According to Theorem 3.3, graph rationales possess certain levels of invariance if and only if graphs are drawn from similar environments. Distribution shifts on inputs (G, Y) result in shifts on $P(E)$, which lead to varying degrees of impact of graph rationales on labels across environments. Given the complex nature of distribution shifts in real-world scenarios, our method imposes no independence assumptions and makes a relaxed inference on environments and graph rationales.

4 Methodology

4.1 The Variational Framework

Based on Theorem 3.2, and 3.3, to capture environment-label dependence and mutable rationale invariance, we propose a novel probabilistic graphical model, which is shown in the upper-left corner of Fig. 2. From Fig. 2, we rewrite $P(Y|G)$ as:

$$P(Y|G) = \iiint P(Y|\hat{G}, E, \tilde{Y}, G) P(\hat{G}|E, \tilde{Y}, G) P(E|\tilde{Y}, G) P(\tilde{Y}|G) d\hat{G} dE d\tilde{Y}, \quad (1)$$

where $\hat{G}$, E , and $\tilde{Y}$ are hidden variables denoting graph rationales, environments, and pseudo-labels, respectively. From Proposition 3.1, we know that $P(E) := P(G)$ captures covariate shift while $P(E) := P(Y|G)$ describes concept shift. To characterize both shifts, in this paper, we use $P(E) := P(G, Y)$. Since Y is to be predicted, we introduce pseudo-labels $\tilde{Y}$ and use G and $\tilde{Y}$ to infer E in Eq. 1. Due to the intractability of latent variables regarding $P(Y|G)$, we cannot directly maximize its likelihood. Hence, we resort to the variational inference (VI) technique and maximize the evidence lower bound (ELBO) of $P(Y|G)$.

Proposition 4.1. The ELBO of $P(Y|G)$ takes the form of:

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_{Q_{\Phi_{\hat{G}}}, Q_{\Phi_E}, Q_{\Phi_{\tilde{Y}}}} \left[\log P_{\theta_Y}(Y|\hat{G}, E, \tilde{Y}, G) \right] - \text{KL} \left[Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) || P(\hat{G}|E, \tilde{Y}, G) \right] - \\ & \text{KL} \left[Q_{\Phi_E}(E|\tilde{Y}, G) || P(E|\tilde{Y}, G) \right] - \text{KL} \left[Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) || P(\tilde{Y}|G) \right], \end{aligned} \quad (2)$$

where $Q_{\Phi_{\hat{G}}}$, Q_{Φ_E} , and $Q_{\Phi_{\tilde{Y}}}$ are variational distributions parameterized by $\Phi_{\hat{G}}$, Φ_E , and $\Phi_{\tilde{Y}}$, respectively. θ_Y also denotes learnable parameters of the final classifier.

Pseudo-label classifier. Since graph labels are not accessible during training, we introduce a pseudo-label classifier $Q_{\Phi_{\tilde{Y}}}$. Specifically, given a graph $G = (A, X)$ with the adjacency matrix A and the node feature matrix X , we first use a GNN to generate node embeddings and then employ a Readout function to derive the graph embedding. After that, with c classes in total, the embedding is further mapped to the probability distribution of pseudo-labels with a MLP. Formally, we have:

$$\tilde{Y} = \text{MLP}(\text{Readout}(\text{GNN}(X, A))) \in \mathbb{R}^c. \quad (3)$$

Environment embedding generator. In our probabilistic graphical model, environments are derived based on both graphs and labels. Therefore, given a graph G with n nodes, to generate the environment embedding, we first enrich node features by concatenating X with the pseudo-label of graph, and then fed into a GNN encoder. Note that since $\tilde{Y} \in \mathbb{R}^c$ is a graph-level feature, before concatenation, we need to expand the dimension of $\tilde{Y}$ and replicate it n times by row. After L layers, we further add an additional gradient reverse layer (GRL) [26] to promote the adversarial training of environments. The overall procedure is given as follows:

$$\mathbf{H}_E^L = \text{GRL}(\text{GNN}([X, \text{Linear}(\tilde{Y})], A)) \in \mathbb{R}^{n \times d}, E = \text{Readout}(\mathbf{H}_E^L) \in \mathbb{R}^d, \quad (4)$$

where d is the dimensionality of hidden embeddings and $\text{Linear}(\cdot)$ represents the linear transformation.

Graph rationale extractor. Once the environment of G is obtained, the graph rationale can be encoded similarly:

$$\hat{G} = \sigma(\text{GNN}([X, E, \text{Linear}(\tilde{Y})], A)) \in \mathbb{R}^{n \times d}, \quad (5)$$

where σ is the sigmoid function. Similarly, since E and $\tilde{Y}$ are graph-level embeddings, we need to first expand and replicate them before concatenation with X . Different from the environment, we represent the graph rationale in the form of node-level representation. This is because nodes contribute differently to the rationale and node-level representation can retain the importance of each node to the rationale.

Final classifier. To implement the final classifier P_{θ_Y} , we jointly leverage environments, graph rationales, and pseudo-labels. We first use a L -layer GNN encoder to generate node embeddings $\mathbf{H}_Y^L$. Then, $\mathbf{H}_Y^L$ is further combined with $\hat{G}$ to memorize rationale information. Finally, we take $\mathbf{H}_Y^L$ and $\hat{G}$ as input, and use an MLP to predict graph labels. Specifically, we have:

$$\mathbf{H}_Y^L = \text{GNN}([X, E, \text{Linear}(\tilde{Y})], A) \in \mathbb{R}^{n \times d}, \hat{Y} = \text{MLP}(\text{Readout}(\mathbf{H}_Y^L \odot \hat{G})) \in \mathbb{R}^c. \quad (6)$$

4.2 E-step: Inference on Environments, Graph Rationales, and Pseudo-labels

The VI-based graph OOD generalization problem has been formulated as maximizing the ELBO of $P(Y|G)$, which is given in Eq. 2. In the E-step, we aim to optimize the parameters regarding environments, rationales, and pseudo-labels, which reduces the objective function in Eq. 2 to:

$$\max \left[-\text{KL} \left(Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) || P(\hat{G}|E, \tilde{Y}, G) \right) - \text{KL} \left(Q_{\Phi_E}(E|\tilde{Y}, G) || P(E|\tilde{Y}, G) \right) - \text{KL} \left(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) || P(\tilde{Y}|G) \right) \right]. \quad (7)$$

However, the prior distributions of $\tilde{Y}$, E , and $\hat{G}$ are unknown. It is impracticable for us to induce a faithful prior distribution $P(\cdot)$ in a non-parametric manner like some node-level tasks [27]. To address the problem, we adopt the core principle of generalized Bayesian inference [28] and substitute the KL divergence with a convex function, which takes the form of negative entropy in this study. Then we have:

Proposition 4.2. *The optimization objective in the E-step can be transformed into:*

$$\mathcal{L}_{\mathbf{E}_{v1}} = -\lambda_{\Phi_{\hat{G}}} H(Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G)) - \lambda_{\Phi_E} H(Q_{\Phi_E}(E|\tilde{Y}, G)) - \lambda_{\Phi_{\tilde{Y}}} H(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G)), \quad (8)$$

where $H(\cdot)$ is the entropy function.

Inspired by β -VAE [29], we use λ to balance the scale of each term. In our implementations, the ratio between $\lambda_{\Phi_{\hat{G}}}$, λ_{Φ_E} , and $\lambda_{\Phi_{\tilde{Y}}}$ is fixed at 1:1:10. By directly optimizing the objective $\mathcal{L}_{\mathbf{E}_{v1}}$, we propose the first version of DEROG, denoted as DEROG-v1.

To further enrich the connotation of latent environments and enhance the salience of important nodes, we further introduce an environment alignment loss and a contrastive loss. For the former, each environment can be generally associated with an id, which serves as the label of distribution shift. Therefore, we adopt the cross-entropy loss to drive the learned environment embeddings to be easily separable. In line with Proposition 3.1, this term enhances the model’s capability in distinguishing various distribution shifts. Given a graph G and its environment id Y_E , we derive:

$$\mathcal{L}_{env} = \text{CrossEntropy}(Y_E, \text{Linear}(E)). \quad (9)$$

On the other hand, it is expected that a well-learned rationale extractor should learn more (less) information from key (irrelevant) nodes. To achieve this goal, considering that the learned $\hat{G} \in \mathbb{R}^{n \times d}$ contains the importance of each node to the rationale, we add a contrastive loss, which pushes embeddings of key nodes away from others. Given a graph G , we consider that nodes that are more likely to contribute to the rationale are positive samples, while others are negative ones. Since the sigmoid function used to generate $\hat{G}$ ensures that each entry in $\hat{G}$ ranges from 0 to 1, we simply measure the importance score of each node v_i by computing the L_1 norm of its embedding $\hat{g}_i$. Then nodes are sorted in descending order by their scores and divided into two halves. We next show how to construct positive and negative pairs as follows. We first randomly sample a node v_1 as anchor. Then we select a random node v_2 as positive sample from the same half where v_1 is located. After that, we randomly select k nodes from the other half as negative samples. Finally, the contrastive loss is given by:

$$\mathcal{L}_{cl} = -\log \frac{\exp(h_1^T h_2 / \tau)}{\sum_{v_j \in \mathcal{V}_-} \exp(h_1^T h_j / \tau)}, \quad (10)$$

where $\mathcal{V}_-$ represents the negative node set with k samples. We add both $\mathcal{L}_{env}$ and $\mathcal{L}_{cl}$ to $\mathcal{L}_{E_{v1}}$, and formulate a new loss function:

$$\mathcal{L}_{E_{v2}} = \mathcal{L}_{E_{v1}} + \lambda_{env} \mathcal{L}_{env} + \lambda_{cl} \mathcal{L}_{cl}. \quad (11)$$

Optimizing $\mathcal{L}_{E_{v2}}$ leads to the second version of DEROG, denoted as DEROG-v2.

4.3 M-step: Classifier Optimization

In the M-step, we fix the learned parameters of $Q_{\Phi_{\hat{Y}}}$, $Q_{\Phi_{\hat{G}}}$, and Q_{Φ_E} , and optimize the final classifier P_{θ_Y} . According to Proposition 4.1, the objective in the M-step is to minimize the negative log-likelihood of $P(Y|\hat{G}, E, G)$. The loss function can be formally represented by:

$$\mathcal{L}_M = \text{CrossEntropy}(Y_T, \hat{Y}), \quad (12)$$

where Y_T represents ground-truth graph labels. We iteratively perform E-step and M-step until model convergence.

[Convergence discussion.] We next discuss the convergence of EM algorithm in DEROG. Our objective is to optimize Eq. 2. In both E- and M-steps, the corresponding objectives are optimized towards a direction of decreasing values, which also decreases the value of Eq. 2. Further, since KL divergence of latent variables and negate of reconstruction probability both have a clear lower bound of 0, Eq. 2 will also be lower bounded by 0. Therefore, with a clear lower bound and each step optimized towards smaller values, the EM algorithm guarantees convergence in our method.

[Time complexity analysis.] While DEROG adopts an EM-based framework, its major time complexity comes from GNNs and MLPs implemented, which are both linear to the number of nodes in the graph. Therefore, the overall time complexity of our method is still linear to the number of nodes in the graph. Due to the space limitation, detailed proofs for Proposition 4.1 and 4.2 can be found in Appendix C and the pseudocode of DEROG is given in Algorithm 1 of Appendix D.

5 Experiments

5.1 Setup

To make fair comparison between DEROG and SOTA methods, experiments are conducted on molecular learning and NLP tasks in GOOD [24] and DrugOOD [25]. To further assess the model performance under various distributional shifts, this study presented results from both covariate and concept shift scenarios in the GOOD benchmark. However, since DrugOOD benchmark does not include concept shifts, only the results from covariate shift are reported.

For fairness, we select representative general OOD baselines and graph-specific baselines. The former includes ERM, IRM [12], VREx [6], and DANN [8], while the latter consists of DIR [4], GSAT [15], CIGA [3], GALA [19], AIA [22], iMoLD [30], EQuAD [31], IGM [23], and LECI [17]. Details of these datasets and baselines are elaborated in Appendix E. Due to fundamental distinctions between methods that focus on graph-level tasks and node-level ones, baselines for node-level predictions like

Table 2: Performance comparisons in real-world scenarios. The evaluation metric employed for the two molecule datasets is ROC-AUC, whereas classification accuracy is adopted for NLP datasets. The best result is emphasized in bold while the second-best one is underscored.

Task		Molecule Learning							Sentiment Analysis			
Dataset	Domain	GOODHIV				DrugOOD-LBAP-core-IC50			GOODTwitter		GOODSST2	
		Scaffold		Size		Scaffold		Assay	Length		Length	
		Covariate	Concept	Covariate	Concept	Covariate	Covariate	Covariate	Covariate	Concept	Covariate	Concept
ERM		72.01(1.02)	67.23(2.41)	63.05(2.80)	61.88(4.27)	70.03(0.59)	67.02(0.21)	69.93(0.40)	57.77(0.75)	45.44(1.85)	78.94(0.55)	71.35(0.56)
IRM		71.05(1.25)	64.05(2.30)	67.94(1.07)	70.92(7.60)	67.78(0.51)	65.37(0.51)	71.33(0.73)	55.89(1.41)	49.68(3.13)	80.21(1.13)	72.17(1.47)
VREx		67.78(1.80)	65.68(3.42)	64.29(1.98)	57.85(1.45)	68.52(0.46)	65.09(1.50)	71.06(0.93)	57.01(0.08)	49.93(3.13)	78.62(1.83)	71.59(0.20)
DANN		72.45(0.90)	67.54(2.26)	64.61(4.41)	64.23(3.80)	67.85(1.20)	65.93(0.57)	70.86(0.38)	56.51(1.33)	48.31(0.65)	79.71(0.25)	73.87(1.74)
DIR		70.72(0.73)	68.59(1.78)	60.28(2.81)	76.73(2.83)	68.23(0.58)	63.53(1.41)	71.28(1.46)	55.94(1.00)	49.39(1.39)	80.02(1.16)	62.46(2.90)
GSAT		70.96(1.36)	68.23(1.41)	63.78(1.36)	63.94(7.00)	69.14(0.44)	66.73(0.36)	71.94(0.76)	57.14(0.65)	48.07(1.78)	80.78(1.98)	71.76(0.25)
CIGA		69.19(1.01)	69.89(0.57)	61.78(0.45)	75.90(1.28)	67.47(1.12)	64.62(0.90)	71.35(0.50)	57.49(1.04)	49.79(3.85)	79.67(2.16)	68.70(1.88)
GALA		68.00(4.88)	70.93(4.17)	60.82(5.48)	75.18(2.48)	66.97(0.21)	65.36(0.42)	70.46(0.76)	57.35(0.70)	49.31(2.41)	81.36(0.63)	67.70(0.72)
AIA		67.56(1.96)	65.18(1.10)	61.64(3.37)	69.90(1.37)	65.50(0.76)	63.53(0.12)	70.89(0.81)	56.20(0.86)	46.66(1.92)	78.80(2.13)	69.43(2.32)
iMoLD		69.90(1.55)	70.65(0.22)	63.63(2.07)	73.76(0.45)	68.68(0.90)	66.09(0.27)	71.00(0.29)	56.78(1.41)	48.33(2.99)	56.78(1.41)	73.14(0.48)
EQuAD		67.38(1.47)	63.85(0.84)	60.86(1.70)	48.52(2.01)	64.48(0.75)	62.93(0.64)	69.75(0.92)	57.79(1.25)	43.81(1.63)	75.40(1.92)	67.06(1.38)
IGM		66.20(1.01)	63.90(2.47)	61.30(1.82)	66.00(0.79)	67.80(1.60)	70.80(2.04)	81.60(2.53)	57.90(1.44)	52.70(0.73)	80.26(2.57)	72.06(1.32)
LECI		72.50(0.85)	69.48(1.39)	64.86(2.40)	55.91(2.76)	69.09(0.58)	66.50(0.40)	72.62(0.26)	59.64(0.60)	53.62(2.66)	83.46(0.33)	72.97(1.62)
DEROG-v1		75.51(0.66)	70.45(1.53)	68.88(1.28)	76.33(1.99)	70.56(0.18)	67.88(0.51)	72.78(0.22)	60.15(0.58)	54.90(0.79)	81.57(0.17)	73.58(0.56)
DEROG-v2		76.87(0.80)	71.40(1.50)	70.48(0.30)	80.65(8.35)	70.87(0.21)	68.07(0.75)	<u>73.03(0.21)</u>	60.54(0.43)	<u>54.35(0.33)</u>	82.00(0.42)	75.48(1.17)

EAGLE [32], CaNet [20], and MARIO [33] are not included. Summary of hyperparameter settings of DEROG are shown in Appendix E.4. Note that differing from previous studies that primarily conduct experiments under covariate shifts, this paper extensively investigates generalizing graphs under more challenging concept shifts on real-world data, thereby providing a thorough model evaluation. More details on these two shifts are provided in Proof C.1. Further, hyperparameter sensitivity analysis is included in Appendix G.

5.2 Main Results

We conduct experiments on real-world benchmarks, which are molecule learning and text sentiment analysis. Results are given in Table 2. From the table, we observe that:

Molecule learning. DEROG-v2 surpasses other competitors on multiple benchmarks, while its simplified version, DEROG-v1, also achieves commendable performance, ranking second in most benchmarks. All these results verify the effectiveness of environment-label dependency and mutable rationale invariance. Among baselines, LECI achieves the best performance in covariate shift scenarios. It obtains environment through encoding the input graph G into graph-level embedding and imposes a stringent assumption on the independence among causal graphs, environments, and labels. Since covariate shift only happens on $P(G)$ and affects less on graph labels Y , the independence assumption between environments and labels improves the performance of LECI. However, concept shift happens on $P(Y|G)$, which means the spuriousness of environment is weakened. As a result, LECI performs poorly under concept shift conditions due to the strict independence assumption on Y and E . While IGM achieves the best results on LBAP-IC50-Assay and -Size benchmarks, it does not perform well on GOOD benchmarks.

Text sentiment analysis. Our methods DEROG-v2 and DEROG-v1 lead other baselines on all the tasks except GOODSST2 with covariate shift. In this task, as pointed out in [17], due to the presence of significant amount of noise in graph labels and environment labels, hyperparameters in LECI that enforce the independence between environment and graph labels are set to very small values or even zero. This again provides evidence for Theorem 3.2, which illustrates that environment information could affect graph labels.

To fairly evaluate the effectiveness of DEROG, we also conduct experiments on synthetic benchmark datasets [4, 24], where the results are given in Appendix F.1 due to the space limitation. Although these datasets are human-crafted towards different methods, our approach DEROG-v2 can still achieve the overall best performance, which further shows its superiority against other competitors.

5.3 Ablation Study

We next comprehensively validate the effectiveness of main components in DEROG, including training techniques, regularization terms and inferred latent variables. Variants of DEROG are categorized into three classes. The first class includes w/ OBI, w/o EM, and w/o GRL. These variants are implemented to validate different techniques used in the model framework. Specifically, the w/ OBI means DEROG with the original Bayesian inference, where latent environments and graph rationales are optimized based on KL divergence. The w/o EM indicates model parameters are

updated simultaneously instead of the EM algorithm. The w/o GRL is DEROG with no gradient reverse layer attached after the GNN encoder in the environment embedding generator. Further, the second class of variants consist of w/o $\mathcal{L}_{env}$, w/o $\mathcal{L}_{cl}$, and w/o $H(\tilde{Y})$, which are adopted to measure the importance of each term in the objective function of Eq. 11 on the overall model performance. To verify the role of the latent environment and graph rationale in model inference, we implement the third class of variants of DEROG to evaluate the expressiveness of inferred variables. Variants include w/o E , w/o $\hat{G}$, and w/o $E \& \hat{G}$. They replace latent environments, graph rationales, or both with standard Gaussian noises, respectively. Note that all the experiments are based on DEROG-v2.

Table 3 summarizes the experimental results. From a global perspective, we observe that DEROG outperforms all its variants. We next provide detailed analysis on these variants.

For the first class of variants with different training techniques, w/ OBI achieves the worst performance. This verifies the rationality of adopting generalized Bayesian inference in the absence of reliable prior knowledge. The performance advantage of DEROG over w/o EM shows the effectiveness of EM algorithm in optimizing variational inference problems. The w/o GRL variant also performs worse than DEROG, which indicates the importance of gradient reverse layer in latent variable inference.

For the second class, we see that the environment alignment (w/o $\mathcal{L}_{env}$) has the minimal impact on overall performances, which leads to the smallest performance drop. This further demonstrates the small reliance of model performance on the environment labels. Hence, when environment labels are absent, we can remove the term $\mathcal{L}_{env}$ from our objective. Also, the performance drops of both w/o $\mathcal{L}_{cl}$ and w/o $H(\tilde{Y})$ indicate that the contrastive loss and the negative entropy of pseudo-labels are essential.

For the third class, the replacement of inferred variables with random noises results in significant performance penalties. This shows the necessity of both environments and graph rationales. Among them, w/o $\hat{G}$ leads to a larger performance drop than w/o E , which indicates that graph rationales are of more important. However, environment is also unignorable.

5.4 Efficiency Study

To investigate the model efficiency, we show the training and validation loss of DEROG over time on two benchmark datasets. We further take LECI and CIGA as reference because the former is the runner-up method while the latter is simple and thus efficient. The results are drawn in Fig. 4. We observe that CIGA is most efficient, while DEROG converges faster than LECI, in particular, on the HIV-Size-Concept case. However, both CIGA and LECI converge to a lower loss on **in-distribution training set**, but a higher and divergent loss on **OOD validation set**, while DEROG stands in the opposite. The stable convergence curve of DEROG on the validation set further explains its effectiveness in graph OOD generalization; also, it is efficient.

Figure 3: Ablation study on DEROG

Dataset	HIV		Twitter	SST2
Domain	Scaffold	Size	Length	Length
Shift	Covariate	Concept	Covariate	Concept
ERM	72.01(1.02)	61.88(4.27)	57.77(0.75)	71.35(0.56)
LECI	72.50(0.85)	55.91(2.76)	59.64(0.60)	72.97(1.62)
DEROG-v1	75.51(0.66)	76.33(1.99)	60.15(0.58)	73.58(0.56)
DEROG-v2	76.87(0.80)	80.65(8.35)	60.54(0.43)	75.48(1.17)
w/ OBI	73.02(1.89)	62.67(7.21)	56.37(1.41)	71.26(1.04)
w/o EM	74.39(1.76)	61.63(2.97)	56.84(0.68)	72.70(1.57)
w/o GRL	75.69(0.77)	64.77(4.63)	57.52(0.21)	73.68(0.83)
w/o $\mathcal{L}_{env}$	76.12(0.73)	77.90(1.23)	60.31(0.50)	74.42(1.10)
w/o $\mathcal{L}_{cl}$	75.99(0.96)	77.78(2.40)	60.25(0.52)	73.80(1.36)
w/o $H(\tilde{Y})$	74.10(1.48)	63.86(3.19)	59.83(0.08)	72.63(0.96)
w/o E	69.36(0.36)	58.68(5.26)	58.00(0.79)	69.24(1.63)
w/o $\hat{G}$	68.94(3.42)	56.33(7.69)	57.81(1.55)	67.68(0.82)
w/o $E \& \hat{G}$	61.08(2.34)	54.48(1.53)	56.42(1.07)	64.41(2.19)

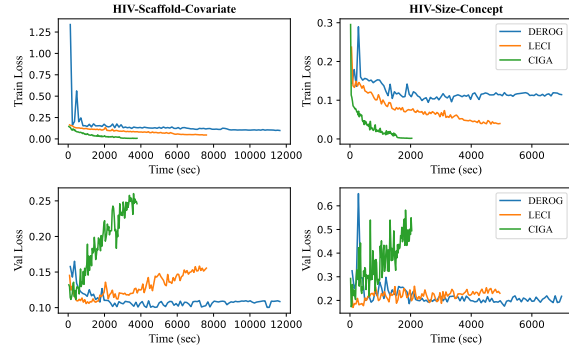


Figure 4: Efficiency study

6 Conclusion

This paper addresses the graph OOD generalization problem. We first provided systematic analyses on the restricted applicability of existing invariant learning methods in complex real-world scenarios, specifically regarding the underlying assumptions made for environments, invariant sub-graphs, and labels. After that, we re-evaluated the roles of environment and graph rationales and constructed a probabilistic graphical model DEROG concerning the conditional distribution of labels. Experimental results under complicated real-world distribution shifts show that DEROG can consistently provide superior performance.

References

- [1] Ling Yang, Jiayi Zheng, Heyuan Wang, Zhongyi Liu, Zhilin Huang, Shenda Hong, Wentao Zhang, and Bin Cui. Individual and structural graph information bottlenecks for out-of-distribution generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12, 2023.
- [2] Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang. Generalizing graph neural networks on out-of-distribution graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):322–337, 2024.
- [3] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 22131–22148. Curran Associates, Inc., 2022.
- [4] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022.
- [5] Xin Sun, Liang Wang, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. Dive: Subgraph disagreement for graph out-of-distribution generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’24, page 2794–2805, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671878. URL <https://doi.org/10.1145/3637528.3671878>.
- [6] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021.
- [7] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR, 17–23 Jul 2022.
- [8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, Jan 2016.
- [9] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450. Springer International Publishing, 2016.
- [10] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023.
- [11] Jonas Peters, Peter Bühlmann, and Nicolai 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, Oct 2016.

- [12] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.
- [13] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021.
- [14] Junchi Yu, Jian Liang, and Ran He. Mind the label shift of augmentation-based graph ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11620–11630, Jun 2023.
- [15] Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15524–15543. PMLR, 17–23 Jul 2022.
- [16] Xiner Li, Shurui Gui, Youzhi Luo, and Shuiwang Ji. Graph structure extrapolation for out-of-distribution generalization. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=Xgrey8uQhr>.
- [17] Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [18] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 35, pages 11828–11841. Curran Associates, Inc., 2022.
- [19] Yongqiang Chen, Yatao Bian, Kaiwen Zhou, Binghui Xie, Bo Han, and James Cheng. Does invariant graph learning via environment augmentation learn invariance? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [20] Qitian Wu, Fan Nie, Chenxiao Yang, Tianyi Bao, and Junchi Yan. Graph out-of-distribution generalization via causal intervention. In *Proceedings of the ACM on Web Conference 2024*, WWW ’24, page 850–860. Association for Computing Machinery, 2024.
- [21] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. In *Advances in Neural Information Processing Systems*, volume 35, pages 12964–12978. Curran Associates, Inc., 2022.
- [22] Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, JUN ZHOU, Xiang Wang, and Xiangnan He. Unleashing the power of graph data augmentation on covariate distribution shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [23] Tianrui Jia, Haoyang Li, Cheng Yang, Tao Tao, and Chuan Shi. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8562–8570, Mar 2024.
- [24] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, pages 2059–2073. Curran Associates, Inc., 2022.
- [25] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, Houtim Lai, Wei Liu, Junzhou Huang, Shuigeng Zhou, Ping Luo, Peilin Zhao, and Yatao Bian. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery – a focus on affinity prediction problems with noise annotations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8023–8031, Jun. 2023.
- [26] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

- [27] Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. In *The Eleventh International Conference on Learning Representations*, 2023.
- [28] A. P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.
- [29] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- [30] Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning invariant molecular representation in latent discrete space. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [31] Tianjun Yao, Yongqiang Chen, Zhenhao Chen, Kai Hu, Zhiqiang Shen, and Kun Zhang. Empowering graph invariance learning with deep spurious infomax. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56860–56884. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/yao24a.html>.
- [32] Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. Environment-aware dynamic graph learning for out-of-distribution generalization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49715–49747. Curran Associates, Inc., 2023.
- [33] Yun Zhu, Haizhou Shi, Zhenshuo Zhang, and Siliang Tang. Mario: Model agnostic recipe for improving ood generalization of graph contrastive learning. In *Proceedings of the ACM on Web Conference 2024*, WWW ’24, page 300–311. Association for Computing Machinery, 2024.
- [34] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors, 2019.
- [35] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, M West, et al. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7(453-464):210, 2003.
- [36] Dimitris G. Tzikas, Aristidis C. Likas, and Nikolaos P. Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.
- [37] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.
- [38] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, Jan 2018.
- [39] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5782–5799, 2023.
- [40] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54. Association for Computational Linguistics, Jun 2014.
- [41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Jun 2019.

A Notations

In this paper, a number of notations are employed to elucidate our theories and the proposed method. Table 3 presents notations used in this paper along with their corresponding descriptions. The notations introduced in Section 3 and Section 4 are separated by a horizontal line.

Table 3: Notations used in this paper

Symbols	Descriptions
$\mathcal{G}$	The graph space
$\mathcal{Y}$	The label space
$\mathcal{E}$	The environment space
G	A single piece of graph
E	Environment
$\hat{G}_i = \Phi_{\hat{G}}(G)$	A sub-graph of graph G_i
$G_i \setminus \hat{G}_i$	Supplementary part of G_i
$Q_{\Phi\tilde{Y}}$	Pseudo-label classifier
$Q_{\Phi E}$	Latent environment generator
$Q_{\Phi\hat{G}}$	Graph rationale extractor
P_{θ_Y}	Final classifier
c	Number of classes of labels in a dataset
n	Number of nodes in a graph
d	Embedding size
$\tilde{Y}$	Pseudo label
$\hat{G}$	Graph rationale
$\mathbf{H}^l$	Node feature at layer l
$\hat{Y}$	Predicted label of final classifier
v_1	Positive anchor node when calculating contrastive loss
v_2	Positive node excluding v_1 when calculating contrastive loss
k	Number of negative sample when calculating contrastive loss
$\mathcal{V}_-$	Negative node set when calculating contrastive loss
$h_{[\cdot]}$	Features of positive or negative node samples when calculating contrastive loss
λ	Loss weight
Y_T	Ground-truth label
Y_E	Ground-truth environment id

It’s worth noting that $\hat{G}_i$ represents a sub-graph of G_i , which is used to explain pitfalls in theories proposed by previous invariant learning methods. But $\hat{G}$ proposed in Section 4 represents graph rationales in latent space instead of certain parts of graphs. The specific entities represented by $\hat{G}$ or $\hat{G}_i$ in different methods may differ, but their core essence is to represent the important parts in graphs.

B Additional Preliminary Study

To further illustrate the environment-label dependency in OOD scenarios quantitatively, we conduct additional experiments using causal-based baselines. Since LECI does not perform sub-graph partitioning, we use CIGA, which divides graphs into causal and spurious parts, and are then respectively used for label prediction and yield c_{out} and s_{out} . We compute the loss between c_{out} and y , and that between s_{out} and y on the test set, whose results are shown in Table 4. We observe that the spurious graph, which has been previously considered to represent the environment distribution,

achieves similar levels of accuracy in label prediction as the causal sub-graph. This further suggests that the causal and spurious sub-graphs identified by existing methods on real datasets, in fact, exhibit comparable levels of mutual information with the labels. This also explains why we study non-causal method in this paper.

Table 4: Preliminary studies on c_{out} and s_{out} from CIGA

Dataset	HIV		Twitter
Domain	Scaffold	Size	Length
Shift	Covariate	Concept	Covariate
$loss(c_{out}, y)$	0.1035	0.2218	2.5274
$loss(s_{out}, y)$	0.1035	0.2220	2.4446

C Proofs of Theorems and Propositions

C.1 Proof of Proposition 3.1

Proposition C.1. (Restatement of Proposition 3.1) According to the definition of covariate shift and concept shift [17], as elaborated in Proof C.1, shifts on distributions of environment reflect shifts on distributions of samples. We have $P(E) := P(G)$ under covariate shift and $P(E) := P(Y|G)$ under concept shift. Therefore, $P(E) := P(G, Y)$ holds for distribution shifts on graphs.

Definition C.2. (Covariate shift) For $\mathcal{D} = \{(G, Y)\}$, $\mathcal{D}_{train} \subset \mathcal{D}$, $\mathcal{D}_{test} \subset \mathcal{D}$, and $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$, we say the distribution shift on (G, Y) is covariate shift if $P_{train}(G) \neq P_{test}(G)$ and $P_{train}(Y|G) = P_{test}(Y|G)$.

Definition C.3. (Concept shift) For $\mathcal{D} = \{(G, Y)\}$, $\mathcal{D}_{train} \subset \mathcal{D}$, $\mathcal{D}_{test} \subset \mathcal{D}$, and $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$, we say the distribution shift on (G, Y) is concept shift if $P_{train}(G) = P_{test}(G)$ and $P_{train}(Y|G) \neq P_{test}(Y|G)$.

Proof. Since the environment is considered as a variable describing distribution shift, it satisfies $P(E) := P(G)$ under covariate shift conditions and $P(E) := P(Y|G)$ under covariate shift conditions. With $P(G, Y) = P(G)P(Y|G)$, we have $P(E) := P(G, Y)$. $\square$

C.2 Proof of Theorem 3.2

Theorem C.4. (Restatement of Theorem 3.2) (**Environment-label dependency**) Let (G_i, Y_i) and (G_j, Y_j) denote the graph-label pairs sampled from the training and test sets, respectively. Further, $Q_{\Phi_{\hat{G}}}$ represents the invariant rationale extractor optimized on the training set. Given the rationales $\hat{G}_i = Q_{\Phi_{\hat{G}}}(G_i)$ and $\hat{G}_j = Q_{\Phi_{\hat{G}}}(G_j)$, there exists an optimal $Q_{\Phi_{\hat{G}}}$ such that $\Phi_{\hat{G}}^* = \arg \max I(\hat{G}_i; Y_i)$ and $I(G_j \setminus \hat{G}_j; Y_j) > 0$.

Proof. Consider $(G_i, Y_i) \in \mathcal{D}_{train}$ and $(G_i, Y_i) \in \mathcal{D}_{test}$. Given the invariant sub-graph extractor $\Phi_{\hat{G}}$ is the optimized on the training set, its parameters are optimized to the optimal state if and only if

$$\text{for } \forall (G_i, Y_i) \in \mathcal{D}_{train}, \text{ s.t. } \Phi_{\hat{G}}^* = \arg \max I(\hat{G}_i; Y_i). \quad (13)$$

We have $I(\hat{G}_i; Y_i) \leq 1$. Let E_i and E_j denote any environment from $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$, respectively. According to [19, 21], $I(\hat{G}_i; Y_i) = I(\hat{G}_j; Y_j)$ if and only if ‘‘Variation Consistency’’ and ‘‘Variation Sufficiency’’ are satisfied. However, according to the real-world examples of molecules to be effective and tons of failure cases, such consistency and sufficiency of the distribution shift of E doesn’t exist in real-world scenarios. Therefore, $I(\hat{G}_i; Y_i) > I(\hat{G}_j; Y_j)$ holds, and we have $I(\hat{G}_j \setminus G_j; Y_j) > 0$. $\square$

C.3 Proof of Theorem 3.3

Theorem C.5. (Restatement of Theorem 3.3) (**Mutable rationale invariance**) Given a graph G with rationale $\hat{G}$, $\exists \mathcal{E}_{sub} \subseteq \mathcal{E}$ and $E_k \in \mathcal{E} \setminus \mathcal{E}_{sub}$, $\forall E_i, E_j \in \mathcal{E}_{sub}$, we have $P(Y|\hat{G}, E_i) = P(Y|\hat{G}, E_j) \neq P(Y|\hat{G}, E_k)$.

Proof. For an $\mathcal{E}_{sub} \subseteq \mathcal{E}$ and $E_i, E_j \in \mathcal{E}_{sub}$, if E_i and E_j satisfied ‘‘Variation Consistency’’ and ‘‘Variation Sufficiency’’, $P(Y|\hat{G}, E_i) = P(Y|\hat{G}, E_j)$. According to Theorem 3.2, conditional probability of Y across environments is inconsistent, and we have $P(Y|\hat{G}, E_i) \neq P(Y|\hat{G}, E_k)$. $\square$

C.4 Proof of Proposition 4.1

Proposition C.6. (Restatement of Proposition 4.1) *The ELBO of $P(Y|G)$ takes the form of:*

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_{Q_{\Phi_{\hat{G}}}, Q_{\Phi_E}, Q_{\Phi_{\tilde{Y}}}} \left[\log P_{\theta_Y}(Y|\hat{G}, E, \tilde{Y}, G) \right] - \\ & \text{KL} \left[Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) || P(\hat{G}|E, \tilde{Y}, G) \right] - \\ & \text{KL} \left[Q_{\Phi_E}(E|\tilde{Y}, G) || P(E|\tilde{Y}, G) \right] - \\ & \text{KL} \left[Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) || P(\tilde{Y}|G) \right], \end{aligned} \quad (14)$$

where $Q_{\Phi_{\hat{G}}}$, Q_{Φ_E} , and $Q_{\Phi_{\tilde{Y}}}$ are variational distributions parameterized by $\Phi_{\hat{G}}$, Φ_E , and $\Phi_{\tilde{Y}}$, respectively. θ_Y also denotes learnable parameters of the final classifier.

Proof. Consider the conditional probability distribution of Y given input graph G $P(Y|G)$. According to Jensen’s Inequality, we have:

$$P(Y|G) = \log \iiint Q(\hat{G}, E, \tilde{Y}|G) \frac{P(Y, \hat{G}, E, \tilde{Y}|G)}{Q(\hat{G}, E, \tilde{Y}|G)} d\hat{G}dEd\tilde{Y} \quad (15)$$

$$\geq \iiint Q(\hat{G}, E, \tilde{Y}|G) \log \left[\frac{P(Y, \hat{G}, E, \tilde{Y}|G)}{Q(\hat{G}, E, \tilde{Y}|G)} \right] d\hat{G}dEd\tilde{Y} \quad (16)$$

$$= \text{ELBO}(P(Y|G)). \quad (17)$$

Since $Q(\hat{G}, E, \tilde{Y}|G) = Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G)$ and $P(Y, \hat{G}, E, \tilde{Y}|G) = P(\hat{G}|E, \tilde{Y}, G)P(E|\tilde{Y}, G)P(\tilde{Y}|G)P(Y|\hat{G}, E, \tilde{Y}, G)$, the evidence lower bound (ELBO) can be further transformed into:

$$\begin{aligned} & \iiint Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G) \log \left[P(Y|\hat{G}, E, \tilde{Y}, G) \right] d\hat{G}dEd\tilde{Y} + \\ & \iiint Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G) \log \left[\frac{P(E|\tilde{Y}, G)}{Q(E|\tilde{Y}, G)} \right] d\hat{G}dEd\tilde{Y} + \\ & \iiint Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G) \log \left[\frac{P(\tilde{Y}|G)}{Q(\tilde{Y}|G)} \right] d\hat{G}dEd\tilde{Y} + \\ & \iiint Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G) \log \left[\frac{P(\hat{G}|E, \tilde{Y}, G)}{Q(\hat{G}|E, \tilde{Y}, G)} \right] d\hat{G}dEd\tilde{Y} \end{aligned} \quad (18)$$

$$\begin{aligned} & = \iiint Q(\hat{G}|E, \tilde{Y}, G)Q(E|\tilde{Y}, G)Q(\tilde{Y}|G) \log \left[P(Y|\hat{G}, E, \tilde{Y}, G) \right] d\hat{G}dEd\tilde{Y} + \\ & \int Q(\hat{G}|E, \tilde{Y}, G) \log \left[\frac{P(\hat{G}|E, \tilde{Y}, G)}{Q(\hat{G}|E, \tilde{Y}, G)} \right] d\hat{G} + \\ & \int Q(E|\tilde{Y}, G) \log \left[\frac{P(E|\tilde{Y}, G)}{Q(E|\tilde{Y}, G)} \right] dE + \\ & \int Q(\tilde{Y}|G) \log \left[\frac{P(\tilde{Y}|G)}{Q(\tilde{Y}|G)} \right] d\tilde{Y}. \end{aligned} \quad (19)$$

Our proposed method employ parameterized networks P_θ , Q_{Φ_E} , $Q_{\Phi_{\tilde{Y}}}$, and $Q_{\Phi_{\hat{G}}}$ to predict the conditional probability of Y and infer latent variables E , $\hat{G}$, and $\tilde{Y}$. Therefore, ELBO($P(Y|G)$) takes the form of:

$$\begin{aligned}
& \underbrace{\mathbb{E}_{Q_{\Phi_{\hat{G}}}, Q_{\Phi_E}, Q_{\Phi_{\tilde{Y}}}} \left[\log P_{\theta_Y}(Y|\hat{G}, E, \tilde{Y}, G) \right]}_{\text{Average log-loss in reconstruction of } Y \text{ from latent variables}} - \\
& \underbrace{\text{KL} \left[Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) || P(\hat{G}|E, \tilde{Y}, G) \right] - \text{KL} \left[Q_{\Phi_E}(E|\tilde{Y}, G) || P(E|\tilde{Y}, G) \right]}_{\text{Divergence between variational posterior and prior}} - \\
& \underbrace{\text{KL} \left[Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) || P(\tilde{Y}|G) \right]}_{\text{Divergence between variational posterior and prior}}. \tag{20}
\end{aligned}$$

□

C.5 Proof of Proposition 4.2

Proposition C.7. (Restatement of Proposition 4.2) *The optimization objective in the E-step can be transformed into:*

$$\begin{aligned}
\mathcal{L}_{\mathbf{E}_{v1}} = & -\lambda_{\Phi_{\hat{G}}} H(Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G)) - \lambda_{\Phi_E} H(Q_{\Phi_E}(E|\tilde{Y}, G)) - \\
& \lambda_{\Phi_{\tilde{Y}}} H(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G)), \tag{21}
\end{aligned}$$

where $H(\cdot)$ is the entropy function.

Adhering the original definition of KL divergence, MoleOOD [21] assumes uniform or Gaussian distributions as the prior distribution of latent variables, and optimizes the posterior distribution Q . However, the assumed prior distribution does not align with the true prior distribution of latent variables. Such an approach not only requires additional computations but also fails to infer plausible posterior distribution of latent variables.

Due to the absence of reliable methods for inferring prior distribution on graph-level tasks, this paper introduces the theory of generalized Bayesian inference to optimize the latent variables.

Proof. The optimization objective at E-step is:

$$\begin{aligned}
& \min \left[\text{KL} \left[Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) || P(\hat{G}|E, \tilde{Y}, G) \right] + \right. \\
& \quad \text{KL} \left[Q_{\Phi_E}(E|\tilde{Y}, G) || P(E|\tilde{Y}, G) \right] + \\
& \quad \left. \text{KL} \left[Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) || P(\tilde{Y}|G) \right] \right] \tag{22}
\end{aligned}$$

$$\begin{aligned}
& \Leftrightarrow \min \left[\int Q(\hat{G}|E, \tilde{Y}, G) \log Q(\hat{G}|E, \tilde{Y}, G) d\hat{G} + \right. \\
& \quad \int Q(E|\tilde{Y}, G) \log Q(E|\tilde{Y}, G) dE + \\
& \quad \int Q(\tilde{Y}|G) \log Q(\tilde{Y}|G) d\tilde{Y} - \\
& \quad \int Q(\hat{G}|E, \tilde{Y}, G) \log P(\hat{G}|E, \tilde{Y}, G) d\hat{G} - \\
& \quad \int Q(E|\tilde{Y}, G) \log P(E|\tilde{Y}, G) dE - \\
& \quad \left. \int Q(\tilde{Y}|G) \log P(\tilde{Y}|G) d\tilde{Y} \right]. \tag{23}
\end{aligned}$$

According to theorems of generalized Bayesian variation inference [28, 34–37], the KL divergence in the optimization objective could be generalized to convex functions. Hence, the optimization objective at E-step is transformed to

$$\Rightarrow \min \left[\lambda_{\Phi_{\hat{G}}} f \left(Q_{\Phi_{\hat{G}}}(\hat{G}|E, \tilde{Y}, G) \right) + \lambda_{\Phi_E} f \left(Q_{\Phi_E}(E|\tilde{Y}, G) \right) + \lambda_{\Phi_{\tilde{Y}}} f \left(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) \right) \right] \quad (24)$$

$$\Rightarrow \min \left[\lambda_{\Phi_{\hat{G}}} \int Q_{\hat{G}}(\hat{G}|E, \tilde{Y}, G) \log \left[Q_{\hat{G}}(\hat{G}|E, \tilde{Y}, G) \right] d\hat{G} + \lambda_{\Phi_E} \int Q_E(E|\tilde{Y}, G) \log \left[Q_E(E|\tilde{Y}, G) \right] dE + \lambda_{\Phi_{\tilde{Y}}} \int Q_{\tilde{Y}}(\tilde{Y}|G) \log \left[Q_{\tilde{Y}}(\tilde{Y}|G) \right] d\tilde{Y} \right] \quad (25)$$

$$\Leftrightarrow \min \left[-\lambda_{\Phi_{\hat{G}}} H(Q_{\Phi_{\hat{G}}}(\hat{G}|E, G, \tilde{Y})) - \lambda_{\Phi_E} H(Q_{\Phi_E}(E|G, \tilde{Y})) - \lambda_{\Phi_{\tilde{Y}}} H(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G)) \right]. \quad (26)$$

□

D Pseudocode of DEROG

The pseudocode of DEROG is shown in Algorithm 1.

E Implementation Details

E.1 Datasets

GOODHIV: A real-world molecule dataset derived from MoleculeNet [38], which are represented as molecule graphs with atom and bond features. Molecules are divided into various environments according to molecular scaffolds or sizes. The goal is to perform binary classification on molecule graphs based on whether they have the potential of suppressing HIV.

DrugOOD-LBAP-core-IC50: A molecule dataset adapted from DrugOOD [25]. The task is to perform binary classification for binding affinity with target proteins with the measurement of IC50. Apart from the two splitting strategies adopted in GOODHIV, molecules are split according to binding assays as well.

GOODTwitter & GOODSST2: Two NLP datasets [39, 40] for sentiment analysis, where sentences are broken down into tokens with features obtained from BERT [41]. Samples are split into different environments based on lengths of them. We investigate original texts of GOODTwitter and find out that the dataset itself is extremely noisy. There exist a significant amount of sarcastic and ambiguous expressions that are mistakenly labeled. Such phenomenon is addressed by other research [17] as well.

GOODMotif: A synthetic dataset derived from Spurious-Motif [4], where each sample is composed of a motif and a spurious graph. The label is solely dependent on the motif. Samples are categorized based on motifs or overall sizes.

GOODCMNIST: A semi-synthetic dataset derived from MNIST, a widely adopted dataset of hand-written digits. By assigning different colors to digits [24], digits are split into different categories to simulate covariate or concept shift.

Algorithm 1: Pseudo code of DEROG

```

1: Input: Training data  $\{G_i, Y_i\}_{i=1}^N$ , batch size  $|G|$ , number
   of nodes in a batch  $|V|$ , embedding size  $d$ , number of
   classes of labels  $c$ , number of epoch  $e$ .
2: Initialize  $\Phi_{\tilde{Y}}, \Phi_E, \Phi_{\hat{G}}, \theta_Y$ .
3: for  $i = 1$  to  $e$  do
4:   Sample a batch  $\{G_j, Y_j\}_{j=1}^b$ ;
5:   Generate pseudo-labels  $Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G) = \text{MLP}(\text{Readout}(\text{GNN}(G))) \in \mathbb{R}^{|G| \times c}$ ;
6:   Infer latent environments  $E = \text{Readout}(\text{GRL}(\text{GNN}[G, \text{Linear}(\tilde{Y})])) \in \mathbb{R}^{|G| \times d}$ ;
7:   Predict graph rationales  $\hat{G} = \text{Sigmoid}(\text{GNN}[G, E, \text{Linear}(\tilde{Y})]) \in \mathbb{R}^{|V| \times d}$ ;
8:   Calculate  $\mathcal{L}_E =$ 
       $\frac{1}{|G|} \sum_G [-\lambda_{\Phi_{\hat{G}}} H(Q_{\Phi_{\hat{G}}}(\hat{G}|E, G, \tilde{Y})) - \lambda_{\Phi_E} H(Q_{\Phi_E}(E|G, \tilde{Y})) - \lambda_{\Phi_{\tilde{Y}}} H(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G))];$ 
9:   if Model is DEROG-v2 then
10:     Calculate environment alignment  $\mathcal{L}_{env} = \frac{1}{|G|} \sum_G \text{CrossEntropy}(\text{Linear}(E), Y_E)$ ;
11:     Sample anchor node  $V_1$  and positive node  $v_2$ , and  $k$  negative nodes.
12:     Calculate contrastive loss  $\mathcal{L}_{cl} = -\frac{1}{|G|} \sum_G \log \frac{\exp(h_1^T h_2 / \tau)}{\sum_{v_j \in V_-} \exp(h_1^T h_j / \tau)}$ ;
13:     Calculate  $\mathcal{L}_E = \mathcal{L}_E + \lambda_{env} \mathcal{L}_{env} + \lambda_{cl} \mathcal{L}_{cl}$ ;
14:   end if
15:   Update  $\Phi_{\tilde{Y}}, \Phi_E, \Phi_{\hat{G}}$ ;
16:   Infer latent environments  $E$  and graph rationales  $\hat{G}$ 
      without gradient;
17:   Fuse  $E$  and  $\hat{G}$  into initialized features of nodes in  $\theta_Y$   $\mathbf{H}_Y^0 = (\mathbf{H}_Y^0 \odot \hat{G}, E, \tilde{Y}) \in \mathbb{R}^{|V| \times d}$ ;
18:   Predict graph labels  $\hat{Y} = \text{Readout}(\text{Pool}(\mathbf{H}_Y^L \odot \hat{G}))$ ,
      where  $\mathbf{H}_Y^L = \text{GNN}(G) \in \mathbb{R}^{|V| \times d}$ ;
19:   Calculate  $\mathcal{L}_M = \frac{1}{|G|} \sum_{(G, Y)} \text{CrossEntropy}(Y_T, \hat{Y})$ ;
20:   Update  $\theta_Y$ 
21: end for

```

Table 5: Number of graphs in the adopted datasets.

Shift Type		Covariate					Concept				
Dataset	Domain	Train	ID Val	ID Test	OOD Val	OOD Test	Train	ID Val	ID Test	OOD Val	OOD Test
GOODHIV	Scaffold	24692	4112	4112	4113	4108	15209	3258	3258	9365	10037
	Size	26169	4112	4112	2773	3961	14454	3096	3096	9956	10525
LBAPcore	Scaffold	21519	4920	30708	19041	19048	-	-	-	-	-
	Size	36597	12153	12411	17660	16415	-	-	-	-	-
	Assay	34179	11314	11683	19028	19032	-	-	-	-	-
GOODTwitter	Length	2590	554	554	1785	1457	2595	555	555	1530	1705
GOODSST2	Length	24744	5301	5301	17206	17490	27270	5843	5843	15142	15944
GOODMotif	Basis	18000	3000	3000	3000	3000	12600	2700	2700	6000	6000
	Size	18000	3000	3000	3000	3000	12600	2700	2700	6000	6000
GOODCMNIST	Background	42000	7000	7000	7000	7000	29400	6300	6300	14000	14000
	Color	42000	7000	7000	7000	7000	29400	6300	6300	14000	14000

GOOD datasets use the MIT license and DrugOOD dataset uses GPL3.0. Statistics of the adopted dataset are shown in Table 5.

E.2 Platform

All the experiments are conducted using PyTorch 1.10 and PyG 2.0.4, and the hardware is 128 AMD EPYC 7763 64-Core Processor, NVIDIA A800 80GB PCIe.

E.3 Details

In terms of the choice for GNN encoder, we adopt GIN and GIN with virtual nodes for DEROG and all the baselines. For those datasets with edge features, these features are embedded to the same dimensionality as node features at each graph convolutional layer.

Note that for initial features with the dimension of 768 in NLP datasets, graph rationales $\hat{G}$ are mapped to the same latent space by a linear layer before the calculation of $\text{Emb}(X) \odot \hat{G}$ in the final classifier.

E.4 Hyperparameter Settings

In the proposed DEROG, $\lambda_{\Phi_{\hat{G}}}$ and λ_{Φ_E} are set to 0.01, k and τ in contrastive loss is constrained to 2 and 0.1, respectively. The remaining hyperparameters are $\lambda_{\Phi_{\hat{Y}}}$, λ_{env} , and λ_{cl} . Other hyperparameters are searched within $\lambda_{\Phi_{\hat{Y}}} \in [0.1, 1]$, $\lambda_{env} \in [0.01, 0.1]$, $\lambda_{cl} \in [0.1, 1]$.

In addition to the parameters mentioned above, we keep the rest parameters the same for DEROG and all baselines for fair comparisons. We set the learning rate to 0.0001, the weight decay to 0.0001, the hidden dimension to 300, and the maximum number of epoch to 200.

F Addition Experiments

F.1 Synthetic Scenarios

In addition to real-world scenarios, we also compare performance of DEROG and all the baselines. Results are shown in Table 6. Though DEROG does not consistently outperform other baselines in all baselines, it manages to achieve the highest ranking on average.

In the scenario where DEROG ranks first, it exhibits a significant advantage over other methods. However, in several benchmarks, there is a notable gap between DEROG and methods that impose strong independence assumptions on the environment, invariant sub-graphs, and labels. Those with such assumptions can naturally fit for graph OOD generalization in synthetic scenarios since ‘‘Variation consistency’’ and ‘‘Variation sufficiency’’ proposed by GALA [19] are easily satisfied.

Table 6: Performance comparisons in synthetic scenarios.

Dataset	GOODMotif				GOODCMNIST				Avg. Rank
Domain	Basis		Size		Background		Color		
Shift Type	Covariate	Concept	Covariate	Concept	Covariate	Concept	Covariate	Concept	
ERM	69.97(2.54)	81.25(0.43)	56.02(6.93)	63.46(5.16)	18.02(1.63)	31.00(1.90)	25.61(4.33)	36.07(0.25)	<u>5.75</u>
IRM	40.39(0.82)	85.82(0.83)	58.10(5.35)	33.61(1.16)	19.53(2.01)	29.54(2.44)	27.74(3.56)	42.83(0.33)	6.25
VREx	42.04(2.19)	83.34(0.63)	58.95(4.96)	32.32(0.00)	13.89(1.76)	23.62(2.27)	26.17(1.20)	38.14(0.27)	9
DANN	41.20(0.69)	86.50(0.33)	60.19(3.80)	32.34(0.03)	15.76(2.05)	28.10(0.58)	28.28(1.59)	42.62(0.79)	7.375
DIR	39.44(2.55)	81.62(1.85)	45.89(4.38)	51.63(5.09)	14.59(4.66)	16.33(0.48)	32.39(4.14)	25.90(3.01)	11.375
GSAT	52.76(2.31)	51.98(1.62)	55.28(3.68)	62.95(1.47)	16.45(1.89)	28.23(0.40)	38.26(9.58)	46.72(0.93)	6.25
CIGA	43.07(3.95)	73.79(4.93)	53.22(3.14)	61.03(9.27)	16.44(4.11)	21.88(0.79)	25.10(5.29)	30.92(2.61)	10.25
GALA	45.25(2.25)	77.26(4.88)	42.98(2.19)	56.17(8.48)	19.29(2.02)	18.05(3.41)	19.68(2.69)	26.30(1.08)	10.75
AIA	70.62(2.41)	75.44(0.84)	49.53(7.08)	60.27(9.97)	15.95(2.82)	29.83(9.39)	25.98(7.13)	36.65(9.95)	7.875
iMoLD	71.30(1.42)	72.62(1.27)	52.01(4.25)	63.36(1.08)	10.22(0.93)	25.80(0.62)	25.67(8.37)	38.25(0.35)	8.5
EQuAD	65.25(4.70)	74.41(3.84)	58.26(2.01)	42.65(3.46)	18.68(1.79)	20.92(2.24)	25.21(6.44)	35.37(2.82)	9
IGM	72.60(2.76)	75.06(0.57)	52.64(5.69)	58.26(2.30)	17.27(0.93)	27.83(1.06)	26.17(2.16)	29.05(1.87)	7.875
LECI	85.81(9.44)	80.55(0.33)	50.37(9.01)	65.24(1.83)	16.78(2.59)	17.82(0.59)	56.52(9.61)	20.90(1.61)	7.5
DEROG-v1	44.94(0.63)	90.15(1.05)	75.52(1.90)	32.89(0.45)	24.39(0.56)	22.03(0.39)	33.71(0.04)	30.66(0.46)	6.75
DEROG-v2	48.14(2.31)	90.84(0.64)	75.81(1.66)	33.00(0.03)	24.47(0.77)	22.27(0.47)	34.03(0.38)	31.19(0.88)	5.5

F.2 Details of Ablation Studies

In Section 5.3, seven variants are implemented to study the importance of key components in DEROG. Here we provide implementation details of these variants.

The ‘‘w/ OBI’’ means DEROG with original Bayesian inference, where latent environments and graph rationales are optimized based on KL divergence. We adopt standard Gaussian distribution and uniform distribution as the prior distribution of latent variables, respectively. Results suggest that ‘‘w/o OBI’’ performs better when standard Gaussian distribution is employed.

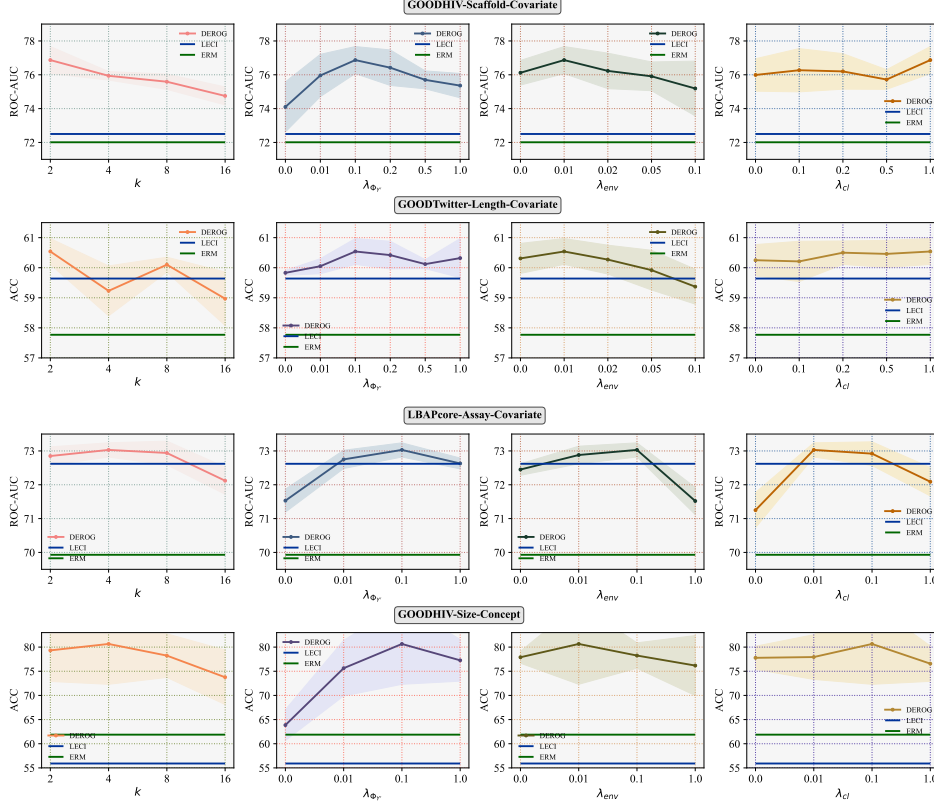


Figure 5: Performance of DEROG under various hyperparameter settings.

The “w/o EM” means parameters are updated all at once. The “w/o EM” and DEROG-2 have identical training objectives.

The “w/o $\mathcal{L}_{env}$ ” and “w/o $\mathcal{L}_{cl}$ ” each represents the variant which removes environment alignment or contrastive loss.

The “w/o $H(\tilde{Y})$ ” stands for the variant where $\tilde{Y}$ is not optimized by its negative entropy. Parameters in the pseudo-label classifier is entirely updated by the loss term of latent environment generator and graph rationale extractor.

The “w/o GRL” is DEROG with no gradient reverse layer attached after latent environment generator and graph rationale extractor.

The “w/o $E/\hat{G}/E\&\hat{G}$ ” represent the variants where latent environments, graph rationales, or both are replaced with standard Gaussian noises.

G Hyperparameter Sensitivity Analysis

We further conduct a hyperparameter sensitivity analysis on DEROG. We evaluate the model performance w.r.t. varying values of k , $\lambda_{\Phi_{\tilde{Y}}}$, λ_{env} , and λ_{cl} on GOODHIV, LBAPcore, and GOODTwitter. Here, k is the number of negative samples in contrastive loss; $\lambda_{\Phi_{\tilde{Y}}}$, λ_{env} , and λ_{cl} represent the weighting factor of $H(Q_{\Phi_{\tilde{Y}}}(\tilde{Y}|G))$ in Eq. 8, the terms $\mathcal{L}_{env}$ and $\mathcal{L}_{cl}$ in Eq. 11, respectively. In our experiments, we vary one hyperparameter with others fixed. For better comparison, two representative baselines LECI and ERM are included as reference. Specifically, LECI is the runner-up method while ERM is a widely adopted method for OOD problem. The results are shown in Fig. 5. Despite the existence of some fluctuations, our method still outperforms runner-up baselines on a wide

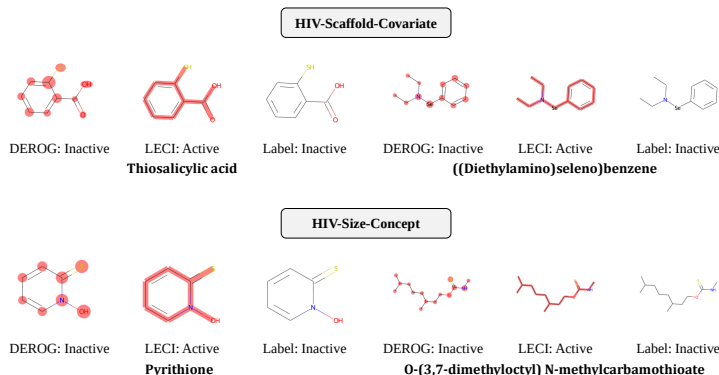


Figure 6: Case analysis of DEROG and LECI on GOODHIV benchmark.

range of hyperparameter values. These results demonstrate the insensitivity of our method towards hyperparameters.

H Case Analysis

Unlike previous methods that show examples of the segmentation of causal and non-causal sub-graphs on handcrafted datasets, our approach cannot similarly visualize the invariance within specific molecules due to their complexity. However, we selected cases where predictions of DEROG are correct but those of LECI failed in two molecular datasets for analysis. There is a tendency for LECI to misclassify ineffective drugs as effective. In Figure 6, we colored the nodes in the samples predicted by DEROG based on the L1 norm of graph rationale, and we colored the edges in the samples predicted by LECI according to the edge weight masks. After double-checking all four samples on PubChem, we find out that all of them are unable to inhibit the HIV virus, while Thiosalicylic acid and Pyrithione failed the test because of their toxicity. It can be observed that our model is capable of distinguishing certain functional groups while ensuring that other parts of the molecule still have an impact on label prediction.

I Limitations

A potential limitation of this paper is that a thorough understanding of the example presented in Section 1 and the case analysis provided in Appendix H may require a certain level of domain knowledge, particularly within the field of computational chemistry. This could potentially restrict the accessibility of this paper to readers who are not familiar with the underlying application context. We have made our best efforts to explain that some previous assumptions may not hold in realistic OOD scenarios. We sincerely appreciate the reader’s patience and understanding in this regard.

Another potential limitation is that the proposed $\mathcal{L}_{env}$ requires environment labels during training, which may not be available in certain practical scenarios. While the removal of the term could slightly degrade the model performance, our method can still perform well compared to SOTA baselines.

J Broader Impact

This paper focuses on the generalization capabilities of graph learning algorithms on real-world out-of-distribution graphs. Findings of this study and the proposed method have broader impacts and potential applications in various real-world scenarios, such as AI-assisted drug discovery, text sentiment analysis, and other graph-level learning tasks. The improvement of the generalization capability of graph learning methods could contribute to advancements in fields that rely on graph-based analysis and decision-making.