Subgraph Aggregation for Out-of-Distribution Generalization on Graphs

Bowen Liu [1,2], Haoyang Li [3], Shuning Wang [2], Shuo Nie [2], Shanghang Zhang [1*]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]Harbin Institute of Technology
[3]Weill Cornell Medicine, Cornell University
2021211039@stu.hit.edu.cn, hal4032@med.cornell.edu, shuw65218@gmail.com, 2021211040@stu.hit.edu.cn, shanghang@pku.edu.cn

Abstract

Out-of-distribution (OOD) generalization in Graph Neural Networks (GNNs) has gained significant attention due to its critical importance in graph-based predictions in real-world scenarios. Existing methods primarily focus on extracting a single causal subgraph from the input graph to achieve generalizable predictions. However, relying on a *single* subgraph can lead to susceptibility to spurious correlations and is insufficient for learning invariant patterns behind graph data. Moreover, in many real-world applications, such as molecular property prediction, multiple critical subgraphs may influence the target label property. To address these challenges, we propose a novel framework, **Su**b**G**raph **A**gg**r**egation (`SuGAr`), designed to learn a diverse set of subgraphs that are crucial for OOD generalization on graphs. Specifically, `SuGAr` employs a tailored subgraph sampler and diversity regularizer to extract a diverse set of invariant subgraphs. These invariant subgraphs are then aggregated by averaging their representations, which enriches the subgraph signals and enhances coverage of the underlying causal structures, thereby improving OOD generalization. Extensive experiments on both synthetic and real-world datasets demonstrate that `SuGAr` outperforms state-of-the-art methods, achieving up to a 24% improvement in OOD generalization on graphs. To the best of our knowledge, this is the first work to study graph OOD generalization by learning multiple invariant subgraphs.

Introduction

Graph representation learning with graph neural networks (GNNs) has achieved significant success in addressing a variety of tasks involving structural data. However, most existing approaches operate under the in-distribution assumption, which posits that the training and testing graphs are sampled from the same distribution. This assumption is often unrealistic and frequently violated in many real-world scenarios. Consequently, GNN models often exhibit a marked decline in performance when confronted with shifts in graph distributions. Addressing the challenge of Out-of-Distribution (OOD) generalization in such cases is therefore both urgent and critical.

To mitigate the failure of OOD generalization, there has been a growing interest in incorporating the invariance principle from causality into GNNs (Peters, Bühlmann, and Meinshausen 2015; Chen et al. 2022; Fan et al. 2022; Li et al. 2022a; Wu et al. 2022b; Miao, Liu, and Li 2022a). The core idea of these graph invariant learning approaches is to identify the invariant subgraph within the input graph, which establishes an invariant relation with target labels across various graph distributions from different environments (Peters, Bühlmann, and Meinshausen 2015). Despite their success, a significant limitation of these methods is their focus on learning only a *single* causal subgraph, while in practical applications, multiple causal subgraphs widely exist. For instance, the activity of the molecule Aspirin (C9H8O4) is determined by both the carboxyl (-COOH) and ester (-COO-) functional groups. Concentrating solely on a single subgraph will inevitably result in the model capturing spurious subgraphs due to the expressivity of GNNs. This limitation underscores the necessity of learning multiple diverse causal subgraphs within the input graph. Therefore, the challenge remains: how to effectively capture multiple subgraphs to make OOD generalized predictions?

To address this issue, we propose a novel framework,

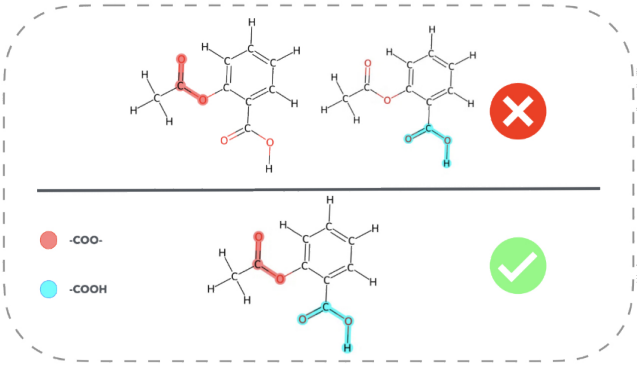


Figure 1: The molecule graph, Aspirin, contains two functional groups: -COOH (denoted with blue lines) and -COO- (denoted with red lines). If model can only capture one functional group as the invariant subgraph, either -COOH or -COO-, which will lead to suboptimal OOD generalization ability. However, `SuGAr` can capture all such functional groups in Aspirin molecule graph for promising performance.

*Corresponding author.

Sub**G**raph **A**gg**r**egation (SuGAr), which facilitates the learning and ensembling of diverse subgraphs, enabling the model to capture richer subgraph signals for enhanced OOD generalization on graphs. These multiple subgraphs are learned independently, with some capturing similar patterns while others identify distinct ones. This diversity, as a form of prior knowledge, enables SuGAr to encompass all critical subgraphs within a graph and to complement the missing patterns that a single model might fail to capture. Specifically, SuGAr first trains a collection of *invariant GNNs* in parallel, where each invariant GNN is designed to capture a distinct invariant subgraph. Then, we introduce a tailored method to promote subgraph diversity: subgraph sampling with a diversity regularizer, which can encourage each model to focus on different subgraphs within the input graph. Finally, we aggregate the subgraphs obtained from these independent training runs to enhance OOD generalization through ensembling and weight averaging. Notably, our proposed weight averaging method is capable of enhancing OOD performance on graphs without incurring additional inference costs. We conducted extensive experiments to validate the effectiveness of SuGAr across 15 datasets with various graph distribution shifts. The results not only demonstrate the efficacy of SuGAr in learning multiple subgraphs but also highlight its superiority in single-subgraph learning when only one causal subgraph is present. Remarkably, SuGAr demonstrates improvements across multiple graph datasets.

Our contributions are summarized as follows:

- We introduce a novel framework, SuGAr, that identifies multiple invariant subgraphs for OOD generalization on graphs. This framework is well-suited for real-world scenarios where an arbitrary number of causal subgraphs may exist.
- We propose a novel ensemble method for graphs and a weight averaging method specifically tailored for GNNs, which effectively aggregate captured multiple invariant subgraphs into a single representation for OOD generalized predictions.
- We conduct comprehensive experiments on real-world graph benchmarks. The results verify the superiority of SuGAr against the state-of-the-art.
- To the best of our knowledge, this is the first work to enhance graph OOD generalization through multiple invariant subgraph learning.

Related Works

OOD generalization on graphs. Despite the success of graph machine learning, most existing approaches assume that training and testing graph data share the same distribution, a concept known as the in-distribution (I.D.) hypothesis. However, in real-world applications, this assumption is often violated due to uncontrollable data generation processes, resulting in distribution shifts between training and testing graphs. While significant efforts have been devoted to out-of-distribution (OOD) generalization for Euclidean data such as images and texts, these methods—ranging from Invariant Learning (Arjovsky et al. 2019; Ahuja et al. 2021; Creager, Jacobsen, and Zemel 2021a) to Domain Adaptation and Domain Generalization (Yu, Liang, and He 2022; Gulrajani and Lopez-Paz 2021; Cha et al. 2021; Arpit et al. 2021; Rame et al. 2022; Ganin et al. 2016; Sun and Saenko 2016; Li et al. 2018)—struggle to generalize to graph-structured data due to its non-Euclidean nature. Traditional graph machine learning methods also lack the capacity for OOD generalization, resulting in substantial performance degradation under distribution shifts. Recent advancements aim to enhance graph OOD generalization through two primary strategies: data-centric approaches (Li et al. 2024; Zhao et al. 2022; ju Park et al. 2022; Wu et al. 2022a; Feng et al. 2020) and invariant learning methods (Wu et al. 2022c; Gui et al. 2023; Chen et al. 2022; Miao, Liu, and Li 2022b; Yu, Liang, and He 2023; Liu et al. 2023). Data-centric methods manipulate graph data to bolster OOD generalization, exemplified by (Li et al. 2024), which introduces an environment-aware framework to extrapolate both structure and feature spaces, thereby generating OOD graph data. Conversely, invariant learning seeks to exploit consistent relationships between features and labels across diverse distributions while discarding spurious correlations that vary across environments. For instance, GSAT (Graph Stochastic Attention) (Miao, Liu, and Li 2022b) addresses graph-level OOD generalization by leveraging an attention mechanism to construct inherently interpretable GNNs, focusing on learning invariant subgraphs under distribution shifts. However, to the best of our knowledge, no existing work effectively addresses scenarios involving multiple invariant subgraphs within a single graph. In this work, we introduce a novel framework designed to more comprehensively capture the underlying invariances.

Weight average for OOD generalization. Recent studies (Izmailov et al. 2018; Cha et al. 2021; Arpit et al. 2021; Draxler et al. 2018) have demonstrated the efficacy of weight averaging in improving OOD generalization for Euclidean data. Techniques such as SWAD enhance OOD generalization by averaging multiple weights along the optimization trajectory, which, from the perspective of the loss landscape, helps identify flatter solutions. Building on the principles of linear mode connectivity (Frankle et al. 2019a; Nagarajan and Kolter 2019) and the observation that many independently trained models exhibit connectivity(Benton et al. 2021), DIWA (Rame et al. 2022) averages weights obtained from several independent training runs, utilizing shared initialization and mild hyperparameter tuning to increase functional diversity across the averaged models. However, the application of weight averaging strategies to non-Euclidean graph data, particularly in graph-structured neural networks, remains unexplored. Our work seeks to address this gap by promoting graph diversity, thereby discovering diverse invariant subgraphs through a appropriate weight averaging strategy.

Subgraph-based GNNs. Another relevant line of research pertains to GNN explainability, which seeks to identify a subgraph within the input as an explanation for a GNN prediction (Luo et al. 2020; Ying et al. 2019; Yuan et al. 2022). Some approaches employ causal reasoning to justify the generated explanations (Lin, Lan, and Li 2021), but they predominantly focus on interpreting GNN predictions rather than on

OOD generalization. While some works (Wu et al. 2022c; Yu et al. 2021) have attempted to bridge this gap by explicitly extracting subgraphs for both predictions and explanations, they typically address graphs and shifts generated under a specific Structural Causal Model (SCM) and therefore often fail to generalize to graphs produced under different SCMs (Chen et al. 2022). In contrast, our work is capable of learning multiple underlying subgraphs and generalizing effectively across different SCMs.

Problem Formulation

In addressing the challenge of OOD generalization on graphs, current approaches in invariant graph learning typically focus on recognizing a single underlying invariant subgraph G_c to predict the label Y (Wu et al. 2022b; Chen et al. 2022). The primary objective of OOD generalization on graphs is to train an *invariant Graph Neural Network (GNN)* $f := f_c \circ g$, which consists of two key components: a) a featurizer $g : G \to G_c$ that extracts the invariant subgraph G_c; and b) a classifier $f_c : G_c \to Y$ that predicts the label Y based on the extracted G_c. The learning objectives of f_c and g are formulated as

$$\begin{aligned} &\min_{f} \max_{e \in \mathcal{E}} \mathbb{E}_{(G,y)\sim(\mathbf{G},\mathbf{y}|\mathbf{e}=e)}[l(f_c(\widehat{G}_c), y) \mid e], \\ &\quad \text{s.t. } \widehat{G}_c = g(G). \end{aligned} \tag{1}$$

However, a single invariant GNN can only learn one subgraph, overlooking scenarios involving multiple subgraphs ($G_c^1 \dots G_c^n$). In this scenario, the objective is to learn a function $f^m := f_c^m \circ g^m$, where g^m extracts multiple underlying invariant subgraphs $\widehat{G}_c^m$ and f_c^m predicts the label based on $\widehat{G}_c^m$. This can be formulated as:

$$\begin{aligned} &\min_{f^m} \max_{e \in \mathcal{E}} \mathbb{E}_{(G,y)\sim(\mathbf{G},\mathbf{y}|\mathbf{e}=e)}[l(f_c^m(\widehat{G}_c), y) \mid e], \\ &\quad \text{s.t. } \widehat{G}_c^m = \widehat{G}_c^1 \cup \dots \cup \widehat{G}_c^n,\ \widehat{G}_c^m = g^m(G). \end{aligned} \tag{2}$$

Methodology

In this section, we present a detailed introduction to SuGAr. The proposed method is designed to build multiple invariant GNNs with identical initialization in parallel. These GNNs are trained using subgraph sampling and a diversity regularizer to promote diversity. This approach is essential because varying hyperparameters and data shuffling alone do not adequately foster the learning of diverse invariant subgraphs. Next, we introduce a novel ensemble and weight averaging method.

Subgraph Diversity Injection

Sampler. As demonstrated in (Wei, Qiao, and Jadav 2023), randomness plays a crucial role in enhancing diversity by introducing variations in substructure selection, compelling GNNs to rely on different patterns within the input graph. This process leads to the creation of multiple GNN models that generalize classification in complementary ways. Inspired by this, we construct multiple GNN base models by applying GNNs to randomly selected substructures in the topological space, facilitating the learning of various subgraphs within the input graph. With a diverse set of learned subgraphs, a straightforward aggregation method can compensate for the limitations of a single invariant subgraph. However, randomness can sometimes disrupt the invariance within the training graph. To mitigate this issue, we introduce a diversity regularizer. This regularizer serves as an alternative or complementary approach to injecting diversity, particularly in worst-case scenarios where sampling impedes the learning of invariant subgraphs. However, this doesn't imply that sampling is ineffective; in fact, in most distribution shifts, the improvements achieved through sampling often exceed those provided by the diversity regularizer.

Diversity regularizer. To enhance the diversity of predicted subgraphs across independent training runs, a straightforward strategy is to diversify the activation values of any two given featurizers, specifically the predicted edge weights. We calculate their similarity on a training graph G with:

$$\delta\left\{g_{\phi_1}, g_{\phi_2}\right\}(G) = g_{\phi_1}(G) \cdot g_{\phi_2}(G) \tag{3}$$

where $g_{\phi_*}(x)$ represents activation values of the featurizer g. We incorporate equation 3 into the optimization objective to enhance diversity across a collection of models.

Subgraph Aggregation

Optimization objective. When training n base models in parallel, to leverage the power of diversity regularizer, we use the complete optimization objective:

$$\begin{aligned} \min_{\{\theta_i\},\{\phi_i\}} \sum_{i=1}^{n} &\left\{R(f_{c,\theta_i} \circ g_{\phi_i}) + \alpha \mathcal{L}_{con}^i\right\} \\ &+ \beta \sum_{i=1}^{n} \sum_{j \neq i} \delta\left\{g_{\phi_i}, g_{\phi_j}\right\}(G) \end{aligned} \tag{4}$$

where $R(f_{c,\theta_i} \circ g_{\phi_i})$ is the empirical risk of the ith base model, and $\mathcal{L}_{con}^i$ denotes the corresponding contrastive loss which minimizes the intra-class invariant subgraph similarity. These two terms encourage that each base model focuses on the invariant patterns of the input graph. The final term is the diversity regularizer, which promotes diversity among the invariant subgraphs learned by each base model. α and β are the weights assigned to the contrastive loss and the diversity regularizer, respectively. We leave more details about objective implementation in appendix.

After training all base GNNs, we introduce our ensemble and weight averaging methods to aggregate the invariant subgraphs they have learned.

Ensemble. Our ensemble approach, SuGAr (ENS), deviates from traditional ensemble methods by not directly combining individual predictions from the base GNN models via a consensus method. Instead, SuGAr (ENS) is implemented in three distinct stages:

- **Stage 1:** SuGAr (ENS) first utilizes the featurizer g from n base models to predict the edge weights of G. For each base model i, let $E^{(i)} = \{e_1^{(i)}, e_2^{(i)}, \dots, e_{|E|}^{(i)}\}$ be the set

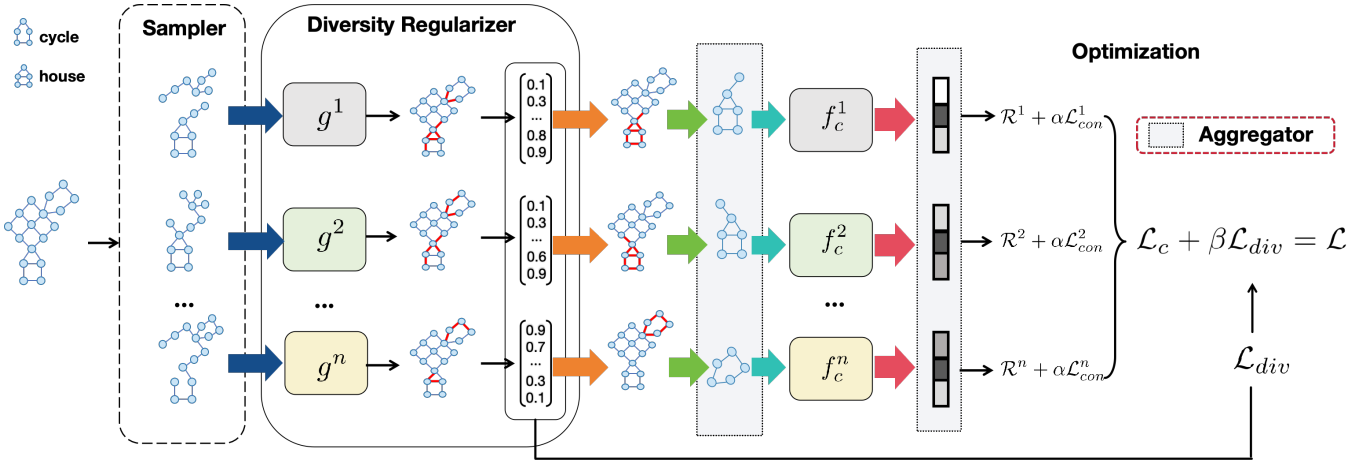


Figure 2: The illustration of **SubGraph Aggregation** (SuGAr). Our proposed method needs to classify graphs based on their motifs("House" and "Cycle") in two steps:(a) Training: A graph sampler randomly drops edges from the input graph to generate a set of different graphs. These graphs are then processed by their corresponding featurizers, g^i, to obtain predicted edge weights for each graph. The diversity regularizer encourages each g^i to extract different subgraphs $\widehat{G}_c^i$, and the corresponding classifier f_c^i makes predictions based on $\widehat{G}_c^i$.(b) Inference: During inference, a graph is fed into the well-trained g^is, which extract a diverse set of subgraphs. The aggregator then merges these subgraphs to form $\widehat{G}_c^m$, which serves as the input for the trained classifiers f_c. Finally, the aggregator combines the predictions from f_cs to make the final decision.

of predicted edge weights. We select the top k edges from each set $E^{(i)}$ to form the invariant subgraph $\widehat{G}_c^i$, resulting in a set of invariant subgraphs $\{\widehat{G}_c^1, \widehat{G}_c^2, \ldots, \widehat{G}_c^n\}$, where $\widehat{G}_c^i$ is defined as:

$$\widehat{G}_c^i = \text{Top}_k(E^{(i)}) \tag{5}$$

- **Stage 2:** Similar to the way of obtaining a single invariant subgraph, we first average the edge weights predicted from all base models and then select the top k edges to obtain the underlying set of multiple invariant subgraphs $\widehat{G}_c^m$. This can be represented as:

$$\widehat{G}_c^m = \text{Top}_k\left(\frac{1}{n}\sum_{i=1}^{n} E^{(i)}\right) \tag{6}$$

Notably, in addition to the averaging mentioned above, we have also developed another approach to aggregate multiple invariant subgraphs. For each edge in G, we take the maximum edge weight from each set of edge weights $E^{(1)}, E^{(2)}, \ldots, E^{(n)}$ to obtain $\widehat{G}_c^m$. This can be represented as:

$$\begin{aligned} &\widehat{G}_c^m = \text{Top}_k(E^{(m)}), E^{(m)} = \{e_1^{(m)}, e_2^{(m)}, \ldots, e_{|E|}^{(m)}\}, \\ &e_k^{(m)} = \max_{i=1}^{n} e_k^{(i)}, k = 1, \ldots, |E|. \end{aligned} \tag{7}$$

- **Stage 3:** Finally, n classifiers make decisions for $\widehat{G}_c^m$ respectively, and these decisions are aggregated using a voting method, such as hard voting, soft voting, etc. We recommend soft voting as it offers better performance.

The soft voting can be represented as:

$$\text{FinalDecision} = \arg\max_k \frac{1}{n}\sum_{i=1}^{n} P_i(\widehat{Y} = k) \tag{8}$$

where $P_i(\widehat{Y} = k)$ is the probability assigned by classifier f_c^i using $\widehat{G}_c^m$, and k is the class label.

Weight Average. Another approach to aggregating subgraphs is to perform averaging in the weight space. To maintain linear mode connectivity in weight space (Frankle et al. 2019b), which ensures that the weights remain connected along a linear path where error is minimized, each base model is trained with a shared initialization (Rame et al. 2022). Let N individual weights $\theta_i{}_{i=1}^{N}$ be collected from independent learning procedures. The final invariant subgraphs are obtained through simple weight averaging (WA), which is defined as:

$$\theta_{\text{WA}} = \frac{1}{N}\sum_{i=1}^{N} \theta_i \tag{9}$$

Model Selection. The final step of our method is to select models that have learned useful information. Following (Rame et al. 2022), we explore two selection protocols:

- **Uniform:** This method selects all models, which is practical but may fail when some models have learned spurious subgraphs. As SuGAr aims to capture multiple diverse subgraphs, it may inadvertently learn spurious ones. The uniform selection, which aggregates all subgraphs, risks disrupting the invariant subgraphs by including these spurious subgraphs.
- **Greedy:** This method mitigates the shortcomings of the uniform approach by ranking models according to valida-

	SPMotif			SUMotif			Avg
	bias=0.33	*bias=0.60*	*bias=0.90*	*bias=0.33*	*bias=0.60*	*bias=0.90*	
ERM	59.26±5.19	49.33±3.16	36.89±0.89	57.72±8.40	56.72±7.83	40.34±4.26	50.04
IRM	63.98±8.51	61.58±12.85	47.14±12.13	61.39±13.10	58.48±15.47	48.35±14.91	46.82
V-Rex	69.18±7.34	58.76±11.51	43.81±13.21	63.24±15.63	65.23±14.18	44.03±12.40	57.38
IB-IRM	62.30±11.27	59.78±12.77	46.19±12.10	71.13±11.04	56.55±13.60	47.27±13.31	57.20
EIIL	63.82±12.43	59.42±13.16	42.51±11.89	67.66±10.27	55.64±13.11	41.22±8.56	55.04
GREA	74.53±8.89	68.26±10.53	48.45±13.96	68.17±17.64	54.39±12.37	51.83±15.03	60.94
GSAT	65.51±8.54	56.20±6.85	48.28±14.24	57.81±9.29	65.12±5.78	60.61±9.27	58.92
DISC	70.01±9.84	54.50±13.26	44.40±12.29	76.20±6.94	62.42±18.18	48.07±9.85	59.27
CAL	69.51±5.38	64.05±5.59	51.49±10.89	69.60±4.20	53.71±8.35	44.04±4.94	58.73
GIL	70.79±8.48	71.38±11.12	53.24±16.46	71.43±7.96	64.32±13.46	38.02±6.61	61.53
DIR	59.58±7.86	66.52±7.95	39.28±3.52	52.59±6.35	45.84±6.35	37.04±4.89	50.14
CIGA	63.70±8.47	64.42±12.69	53.20±19.19	64.17±12.10	53.20±18.48	48.28±14.24	57.82
DiWA(ERM)	49.07±5.93	51.72±4.53	36.65±1.72	60.19±3.35	67.26±7.22	51.09±7.25	52.66
DiWA	63.30±7.54	65.36±5.29	62.20±10.05	73.93±5.80	68.75±17.46	69.49±7.09	67.17
ENS(ERM)	54.15±2.15	49.75±1.75	35.95±0.41	65.51±6.24	68.25±7.25	42.86±2.30	52.75
ENSEMBLE	60.23±8.49	61.80±11.85	61.25±9.37	76.85±1.38	65.34±17.88	54.84±15.89	63.39
SuGAr (WA)	**82.90±8.98**	84.57±4.38	81.61±8.40	78.39±9.27	79.21±5.79	75.87±6.33	80.43
SuGAr (ENS)	82.82±8.90	**85.94±6.39**	**84.80±3.67**	**90.18±3.57**	**85.51±1.18**	**77.71±5.16**	**84.30**

Table 1: OOD generalization performance on synthetic graphs. The best results are in bold and the second-best results are underlined.

tion accuracy and incrementally adding them only if they enhance SuGAr's validation performance. Compared to uniform selection, greedy selection effectively circumvents this limitation by prioritizing models based on their performance during aggregation, starting with the top-performing model. A new model is incorporated only if it improves validation performance, thereby excluding poorly performing models that may have learned spurious subgraphs.

Theoretical Analysis

Theorem *Given a set of graph datasets $\mathcal{D}$ that each sample contains multiple critical subgraphs with various environments, assuming that each training graph G has multiple invariant subgraphs, denoted as G_c^m, $G_c^m = G_c^1 \cup \cdots \cup G_c^n$, $G_c^1 \cap \cdots \cap G_c^n = \emptyset$,* then:
If $\forall G_c^i, |G_c^i| = s_c$, each solution to equation 10 identifies underlying multiple invariant subgraphs.

$$\begin{aligned} &\max I(\widehat{G}_c^m; Y), \\ &\text{s.t. } \widehat{G}_c \in \arg \max_{\widehat{G}_c^i = g^i(G), |\widehat{G}_c^i| \le s_c} \sum_{i=1}^{n} I(\widehat{G}_c^i; \widetilde{G}_c^i | Y), \\ &\arg\min \sum_{j \neq k} I(\widehat{G}_c^j; \widehat{G}_c^k | Y), \\ &\widehat{G}_c^m = \widehat{G}_c^1 \cup \cdots \cup \widehat{G}_c^n. \end{aligned} \tag{10}$$

The proof is shown in the appendix due to the page limit.

Experiments

We conducted experiments on 15 datasets to rigorously evaluate the effectiveness of SuGAr, encompassing both synthetic and real-world datasets that exhibit various distribution shifts. Specifically, we aim to address the following research questions: **RQ1:** Can SuGAr outperform state-of-the-art (SOTA) methods in OOD generalization on graphs? **RQ2:** Does SuGAr more accurately and comprehensively extract subgraphs compared to existing methods? **RQ3:** Do our proposed diversity injection methods yield a set of diverse and informative subgraphs? **RQ4:** Is SuGAr robust to variations in the weights of the contrastive loss α and the diversity regularizer β?

Datasets. We utilized the SPMotif datasets from DIR (Wu et al. 2022c), which involve artificial structural shifts and graph size shifts. Additionally, we developed SUMotif based on SPMotif to verify whether SuGAr generalizes effectively when multiple critical subgraphs are present. To evaluate SuGAr in real-world scenarios with more complex distribution shifts, we employed DrugOOD (Ji et al. 2022) from AI-aided Drug Discovery with Assay, Scaffold, and Size splits, the ColoredMNIST dataset injected with attribute shifts, and Graph-SST (Socher et al. 2013) with degree biases, following the methodology of CIGA (Chen et al. 2022). More dataset details are shown in the Appendix.

Baselines. We compared SuGAr with various baselines, including ERM (Vapnik 1991), as well as several SOTA OOD methods from both the Euclidean domain—such as IRM (Arjovsky et al. 2019), VREx (Krueger et al. 2021), EIIL (Creager, Jacobsen, and Zemel 2021b), and IB-IRM (Ahuja et al. 2021)—and the graph domain, including GREA (Liu et al. 2022), GSAT (Miao, Liu, and Li 2022b), CAL (Sui et al. 2022), GIL (Li et al. 2022b), DisC (Fan et al. 2022), and CIGA (Chen et al. 2022). Additionally, we included ENS(ERM), an ensemble method that uses ERM as the

Datasets	EC-Assay	EC-Scaffold	EC-Size	SST5	Twitter	CMNIST	Ki-Assay	Ki-Scaffold	Ki-Size	Avg
ERM	75.57±1.23	64.21±0.89	63.30±1.19	44.21±0.91	63.84±1.61	10.26±0.62	73.30±1.67	70.45±0.30	74.00±1.55	59.90
IRM	77.10±2.55	64.32±0.42	62.33±0.86	42.77±1.26	60.42±1.06	15.15±3.66	75.10±3.38	69.32±1.84	76.25±0.73	61.21
V-Rex	75.57±2.17	64.73±0.53	62.80±0.89	42.48±1.67	60.50±2.05	17.12±5.68	74.16±1.46	71.40±2.77	76.68±1.35	61.44
IB-IRM	64.70±2.50	62.62±2.05	58.28±0.99	43.02±1.94	60.80±2.50	13.06±1.97	71.98±3.26	69.55±1.66	70.71±1.95	57.19
EIIL	64.20±5.40	62.88±2.75	59.58±0.96	43.79±1.19	60.15±1.44	11.80±0.42	74.24±2.48	69.63±1.46	76.56±1.37	59.51
GREA	66.87±7.53	63.14±2.19	59.20±1.42	43.29±0.85	59.92±1.48	13.92±3.43	73.17±1.80	67.82±4.67	73.52±2.75	58.40
GSAT	76.07±1.95	63.58±1.36	61.12±0.66	43.24±0.61	60.13±1.51	10.51±0.53	72.26±1.76	70.16±0.80	75.78±2.60	59.46
DISC	61.94±7.76	54.10±5.69	57.64±1.57	40.67±1.19	57.89±2.02	15.08±0.21	54.12±8.53	55.35±10.5	50.83±9.30	54.07
CAL	75.10±2.71	64.79±1.58	63.38±0.88	39.60±1.80	55.36±2.67	11.46±1.82	75.10±0.73	60.35±11.3	73.69±2.29	57.65
GIL	70.56±4.46	61.59±3.16	60.46±1.91	43.30±1.24	61.78±1.66	13.19±2.25	75.22±1.73	71.08±4.83	72.93±1.79	58.90
CIGA	77.52±0.97	61.76±1.13	63.74±1.43	44.20±1.89	60.94±1.04	10.44±0.39	71.98±2.65	73.98±2.37	77.00±2.36	60.17
DiWA(ERM)	77.26±1.61	65.03±0.65	62.83±0.30	43.63±0.97	62.56±2.07	10.09±0.30	73.29±1.99	72.61±2.42	75.61±1.98	60.32
DiWA	75.62±2.88	64.36±1.23	64.23±1.06	44.09±0.49	62.26±2.00	27.60±13.2	74.38±1.92	73.53±2.51	77.68±2.37	62.64
ENS(ERM)	77.10±1.57	65.02±0.72	64.05±0.86	44.98±1.13	62.73±0.64	10.33±0.34	73.74±1.79	70.99±2.07	75.98±1.98	60.55
ENSEMBLE	78.27±1.26	64.80±0.43	64.29±0.54	46.20±0.58	63.37±1.83	11.84±2.36	76.94±2.08	75.63±1.06	80.48±0.87	62.42
SuGAr (WA)	76.25±1.43	65.27±1.17	64.53±2.72	43.38±1.46	63.72±1.80	28.49±17.61	75.03±4.14	77.12±1.43	79.19±2.45	63.66
SuGAr (ENS)	**78.62±0.84**	**65.56±0.43**	**66.36±1.06**	**46.88±0.69**	**66.44±1.20**	**28.63±17.63**	**77.57±1.85**	**77.18±1.45**	**81.47±0.37**	**65.41**

Table 2: OOD generalization performance under realistic graph distribution shifts. The best results are in bold and the second-best results are underlined.

base model. For all methods utilizing CIGA, we maintained the same selection ratio (i.e., s_c) for the base models and employed soft voting for all ensemble methods. Each base model in the ENSEMBLE configuration was a CIGA model trained on the full graph. For weight averaging (WA), we report results using the SOTA WA method DiWA (Rame et al. 2022), excluding MA (Arpit et al. 2021) and SWAD (Cha et al. 2021) due to their poor performance on graphs. We applied the Greedy selection strategy for all WA methods, as it consistently outperformed the Uniform strategy across all datasets. DiWA(ERM) used ERM for weight averaging, while DiWA employed CIGA. Both methods followed a shared initialization and mild hyperparameter search setup as described in (Rame et al. 2022). All ensemble and WA methods, as well as SuGAr, employed 10 base models.

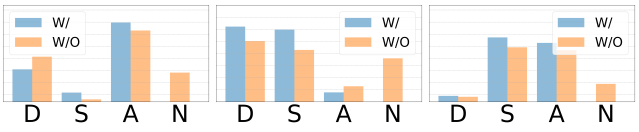
Evaluation. For all datasets, except DrugOOD, we reported classification accuracy, while for DrugOOD we followed the protocol in (Ji et al. 2022) and reported the ROC-AUC. Evaluations were repeated multiple times, with models selected based on validation performance. We report the mean and standard deviation with **5** times runs of the respective metrics.

Main Results(RQ1). To address RQ1, we conducted a comparative analysis of SuGAr against a range of baseline methods. The performance of SuGAr, relative to current SOTA methods, is presented in tables 1 and 2. SuGAr not only achieved the best results but also consistently secured the majority of second-best positions across 15 datasets, including both synthetic and real-world scenarios. Notably, SuGAr outperformed other SOTA baselines by up to 20% on synthetic datasets and exceeded their performance across all real-world datasets. In 6 out of 9 real-world datasets, the **mean-1*std** of SuGAr surpassed the mean of the best baseline results. Unlike existing methods that perform well under limited distribution shifts but suffer significant performance drops under certain conditions, SuGAr consistently demonstrated robust performance across a wide range of distribution shifts. In contrast, other baselines, whether from the Euclidean or graph domains, frequently underperformed compared to ERM. This indicates that these methods fail to extract invariant features or subgraphs and may even rely on spurious correlations.

SuGAr vs. WA & ENS. We compared SuGAr with SOTA WA and ensemble methods to highlight its superiority on graph data. To the best of our knowledge, there are no existing methods specifically designed for weight averaging on graphs; current techniques are primarily developed for Euclidean data. Due to their poor performance on graphs, we excluded SOTA weight averaging methods such as MA and SWAD, focusing instead on DiWA, which applies to graph data. As shown in tables 1 and 2, while DiWA performed well on most datasets, it significantly underperformed its base model on some, highlighting the limitations of current weight averaging methods in identifying flatter solutions in the loss landscape for generalizable predictions. In contrast, SuGAr (WA), a WA method tailored for graph data, consistently outperformed DiWA on 14 out of 15 datasets. Notably, SuGAr (WA) achieved performance comparable to or better than ensemble methods, without incurring additional inference time. For ensemble methods, both ENS(ERM) and ENSEMBLE underperformed their respective base models, ERM and CIGA, on certain datasets. This suggests that existing ensemble methods, which simply promote diversity among base models, often fail to aggregate useful information and may even disrupt it. However, our proposed ensemble method, specifically designed for graph data, effectively aggregates genuinely useful information, leading to more generalizable predictions. As shown in

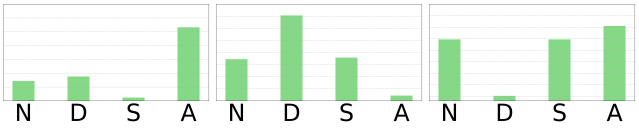
tables 1 and 2, SuGAr (ENS) consistently secured top-tier positions and exhibited lower variance compared to other methods.

Multi-scenario Analysis(RQ2). Next, we analyze the OOD generalization performance in multi-subgraph scenarios. In tables 1 and 2, we evaluated SuGAr on SUMotif and DrugOOD to demonstrate its superiority in learning multiple invariant subgraphs. For the synthetic dataset SUMotif, which comprises a combination of two motif graphs directly determining the label and a base graph providing spurious correlations, we observed that CIGA failed and suffered significant performance drops when multiple subgraphs were present. In contrast, SuGAr consistently achieved high performance and low variance across both cases. Notably, SuGAr improved CIGA's performance by 30% under a bias of 0.6. For real-world datasets, we assessed the effectiveness of SuGAr on DrugOOD, which contains drug molecules with multiple functional groups (subgraphs). Previous methods, focusing on capturing a single subgraph, failed to consistently improve upon ERM. Conversely, SuGAr provided the most comprehensive coverage, irrespective of the number of functional groups (subgraphs) in the molecules, consistently outperforming other baselines and delivering constant improvements over ERM.



(a) SPMotif-0.9 (b) EC50-Scaffold (c) EC50-Size

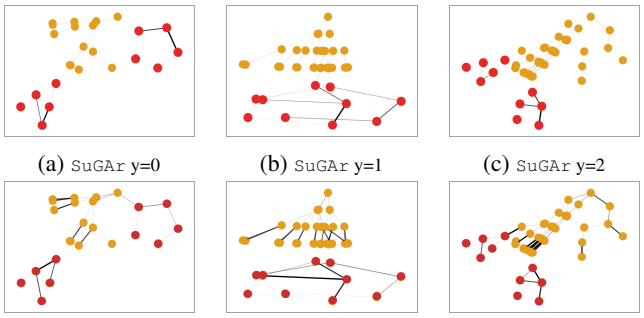
Figure 3: Ablation Studies for SuGAr (ENS).



(a) SPMotif-0.9 (b) EC50-Scaffold (c) EC50-Size

Figure 4: Ablation Studies for SuGAr (WA).

Ablation Studies(RQ3). To assess the importance of diversity injection, we designed the following variants and conducted experiments on three challenging datasets: SPMotif-0.9, EC50-Scaffold, and EC50-Size: (1) **D**: Removed the diversity regularizer, retaining only sampler to learn diverse subgraphs. (2) **S**: Removed sampler, retaining only the diversity regularizer to learn diverse subgraphs. (3) **A**: Retained both sampler and the diversity regularizer. (4) **N**: Removed all components of the proposed method. For SuGAr (ENS), we also compared the performance with (**w/**) and without (**w/o**) our aggregation method to verify its effectiveness. As shown in figures 3 and 4, the SU-None variant, which lacked diversity injection, failed to outperform the best performers among the other three variants that incorporated diversity. This suggests that our diversity injection approach effectively identifies a diverse set of useful subgraphs. These subgraphs are then aggregated to form a comprehensive subgraph that contains valuable information for making robust predictions.

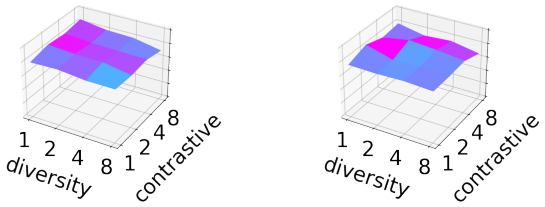


(a) SuGAr y=0 (b) SuGAr y=1 (c) SuGAr y=2

(d) CIGA y=0 (e) CIGA y=1 (f) CIGA y=2

Figure 5: Comparison of SuGAr and CIGA.

Case Study. We present a visualization of the subgraphs predicted by SuGAr on SUMotif. Each edge is displayed according to its predicted edge weight, with edges becoming thicker and darker as their weights increase. As shown in figure 5, CIGA struggled in multi-subgraph scenarios, identifying both invariant and spurious subgraphs. In contrast, SuGAr successfully identified only the invariant subgraphs, effectively ignoring all parts of the spurious subgraphs.



(a) SuGAr (ENS) (b) SuGAr (WA)

Figure 6: Hyperparameter sensitivity.

Hyperparameter Sensitivity(RQ4). We also examined the hyperparameter sensitivity of SuGAr on SPMotif-0.6. As shown in Figure 6, SuGAr maintains strong OOD generalization performance across a wide range of hyperparameter settings, demonstrating its robustness to different hyperparameter choices.

Conclusions

This paper introduces **Su**b**G**raph **A**gg**r**egation (SuGAr), a framework for improving OOD generalization on graphs by learning diverse invariant subgraphs, reducing spurious correlations. Tested on 15 datasets with various shifts, SuGAr demonstrates significant generalization improvements, providing a generalized solution for real-world scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62476011).

References

Ahuja, K.; Caballero, E.; Zhang, D.; Gagnon-Audet, J.-C.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. In *Advances in Neural Information Processing Systems*.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*.

Arpit, D.; Wang, H.; Zhou, Y.; and Xiong, C. 2021. Ensemble of Averages: Improving Model Selection and Boosting Performance in Domain Generalization. *Cornell University - arXiv,Cornell University - arXiv*.

Benton, G.; Maddox, W.; Lotfi, S.; and Wilson, A. 2021. Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. *arXiv: Learning,arXiv: Learning*.

Cha, J.; Chun, S.; Lee, K.-J.; Cho, H.-C.; Park, S.-H.; Lee, Y.; and Park, S. 2021. SWAD: Domain Generalization by Seeking Flat Minima. *Cornell University - arXiv,Cornell University - arXiv*.

Chen, Y.; Zhang, Y.; Bian, Y.; Yang, H.; Ma, K.; Xie, B.; Liu, T.; Han, B.; and Cheng, J. 2022. Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs. In *Advances in Neural Information Processing Systems*.

Creager, E.; Jacobsen, J.; and Zemel, R. S. 2021a. Environment Inference for Invariant Learning. In *International Conference on Machine Learning*, volume 139, 2189–2200.

Creager, E.; Jacobsen, J.-H.; and Zemel, R. 2021b. Environment Inference for Invariant Learning. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2189–2200. PMLR.

Draxler, F.; Veschgini, K.; Salmhofer, M.; and Hamprecht, F. 2018. Essentially No Barriers in Neural Network Energy Landscape. *International Conference on Machine Learning,International Conference on Machine Learning*.

Fan, S.; Wang, X.; Mo, Y.; Shi, C.; and Tang, J. 2022. Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure. In *Advances in Neural Information Processing Systems*.

Feng, W.; Zhang, J.; Dong, Y.; Han, Y.; Luan, H.; Xu, Q.; Yang, Q.; Kharlamov, E.; and Tang, J. 2020. Graph Random Neural Networks for Semi-Supervised Learning on Graphs. *Neural Information Processing Systems,Neural Information Processing Systems*.

Frankle, J.; Dziugaite, G.; Roy, D.; and Carbin, M. 2019a. Linear Mode Connectivity and the Lottery Ticket Hypothesis. *arXiv: Learning,arXiv: Learning*.

Frankle, J.; Dziugaite, G.; Roy, D.; and Carbin, M. 2019b. Linear Mode Connectivity and the Lottery Ticket Hypothesis. *arXiv: Learning,arXiv: Learning*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1): 2096–2030.

Gui, S.; Liu, M.; Li, X.; Luo, Y.; and Ji, S. 2023. Joint Learning of Label and Environment Causal Independence for Graph Out-of-Distribution Generalization. arXiv:2306.01103.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. *Cornell University - arXiv,Cornell University - arXiv*.

Ji, Y.; Zhang, L.; Wu, J.; Wu, B.; Huang, L.-K.; Xu, T.; Rong, Y.; Li, L.; Ren, J.; Xue, D.; Lai, H.; Xu, S.; Feng, J.; Liu, W.; Luo, P.; Zhou, S.; Huang, J.; Zhao, P.; and Bian, Y. 2022. 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.

ju Park, H.; Lee, S.; Kim, S. H.; Park, J.; Jeong, J.; Kim, K.; Ha, J.-W.; and Kim, H. J. 2022. Metropolis-Hastings Data Augmentation for Graph Neural Networks. In *Neural Information Processing Systems*.

Krueger, D.; Caballero, E.; Jacobsen, J.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. C. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *International Conference on Machine Learning*, 5815–5826.

Li, H.; Zhang, Z.; Wang, X.; and Zhu, W. 2022a. Learning Invariant Graph Representations for Out-of-Distribution Generalization. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Li, H.; Zhang, Z.; Wang, X.; and Zhu, W. 2022b. Learning Invariant Graph Representations for Out-of-Distribution Generalization. In *Advances in Neural Information Processing Systems*.

Li, X.; Gui, S.; Luo, Y.; and Ji, S. 2024. Graph Structure Extrapolation for Out-of-Distribution Generalization. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 27846–27874. PMLR.

Li, Y.; Gong, M.; Tian, X.; Liu, T.; and Tao, D. 2018. Domain Generalization via Conditional Invariant Representation. arXiv:1807.08479.

Lin, W.-Y.; Lan, H.; and Li, B. 2021. Generative Causal Explanations for Graph Neural Networks. *Cornell University - arXiv,Cornell University - arXiv*.

Liu, G.; Zhao, T.; Xu, J.; Luo, T.; and Jiang, M. 2022. Graph Rationalization with Environment-based Augmentations. *arXiv preprint arXiv:2206.02886*.

Liu, Y.; Ao, X.; Feng, F.; Ma, Y.; Li, K.; Chua, T.-S.; and He, Q. 2023. FLOOD: A Flexible Invariant Learning Framework for Out-of-Distribution Generalization on Graphs. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 1548–1558. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; and Zhang, X. 2020. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems*, 19620–19631.

Miao, S.; Liu, M.; and Li, P. 2022a. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. arXiv:2201.12987.

Miao, S.; Liu, M.; and Li, P. 2022b. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. *International Conference on Machine Learning*.

Nagarajan, V.; and Kolter, J. 2019. Uniform convergence may be unable to explain generalization in deep learning. *Cornell University - arXiv,Cornell University - arXiv*.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2015. Causal inference using invariant prediction: identification and confidence intervals. *Cornell University - arXiv,Cornell University - arXiv*.

Rame, A.; Kirchmeyer, M.; Rahier, T.; Rakotomamonjy, A.; patrick gallinari; and Cord, M. 2022. Diverse Weight Averaging for Out-of-Distribution Generalization. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 2022. Causal Attention for Interpretable and Generalizable Graph Classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 1696–1705. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. arXiv:1607.01719.

Vapnik, V. 1991. Principles of Risk Minimization for Learning Theory. In *Advances in Neural Information Processing Systems*, 831–838.

Wei, W.; Qiao, M.; and Jadav, D. 2023. GNN-Ensemble: Towards Random Decision Graph Neural Networks. arXiv:2303.11376.

Wu, L.; Lin, H.; Huang, Y.; and Li, S. Z. 2022a. Knowledge Distillation Improves Graph Structure Augmentation for Graph Neural Networks. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 11815–11827. Curran Associates, Inc.

Wu, Q.; Zhang, H.; Yan, J.; and Wipf, D. 2022b. Handling Distribution Shifts on Graphs: An Invariance Perspective. In *International Conference on Learning Representations*.

Wu, Y.; Wang, X.; Zhang, A.; He, X.; and Chua, T.-S. 2022c. Discovering Invariant Rationales for Graph Neural Networks. In *International Conference on Learning Representations*.

Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. *Cornell University - arXiv,Cornell University - arXiv*.

Yu, J.; Liang, J.; and He, R. 2023. Mind the Label Shift of Augmentation-Based Graph OOD Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11620–11630.

Yu, J.; Xu, T.; Rong, Y.; Bian, Y.; Huang, J.; and He, R. 2021. Graph Information Bottleneck for Subgraph Recognition. In *International Conference on Learning Representations*.

Yu, J. C.; Liang, J.; and He, R. 2022. Finding Diverse and Predictable Subgraphs for Graph Domain Generalization. *arXiv preprint arXiv:2206.09345*.

Yuan, H.; Yu, H.; Gui, S.; and Ji, S. 2022. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–19.

Zhao, T.; Liu, Y.; Neves, L.; Woodford, O.; Jiang, M.; and Shah, N. 2022. Data Augmentation for Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 11015–11023.