CauVQ: Causal Vector Quantization for Graph OOD Generalization

Weihong Zhang[1], **Liang Bai**[1*], **Hangyuan Du**[1], **Xian Yang**[2]

[1] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006, China

[2] Alliance Manchester Business School, The University of Manchester, Manchester, M13 9PL, UK

zhangweihong1@sxu.edu.cn, bailiang@sxu.edu.cn, duhangyuan@sxu.edu.cn, xian.yang@manchester.ac.uk

Abstract

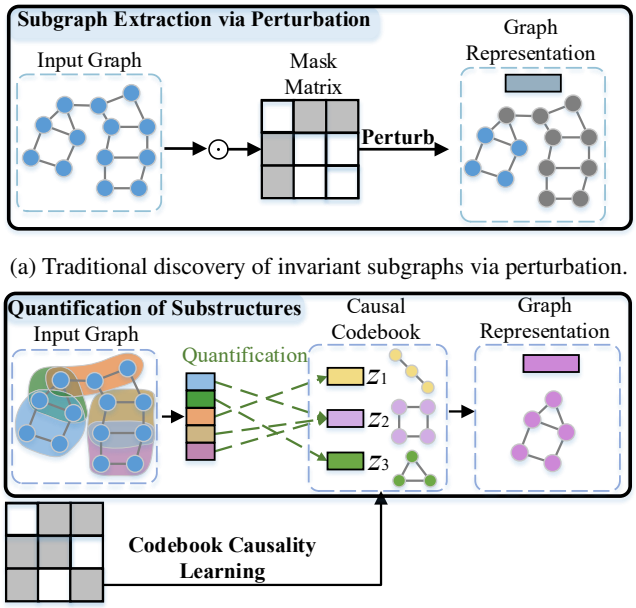
Graph Neural Networks (GNNs) perform well on in-distribution data but often fail under out-of-distribution (OOD) shifts due to reliance on spurious patterns. To address this, we propose CauVQ, a causal vector quantization framework that improves OOD generalization by identifying and leveraging invariant substructures that are causally predictive. To construct stable and symbolic graph representations, CauVQ decomposes each input into local substructures and maps them to a discrete codebook of prototypical motifs. This enables consistent and interpretable encoding across diverse graph domains. To isolate the causal substructures, we maximize their mutual information with graph labels and refine their representations using a learnable interaction matrix and a causal attention mechanism. Furthermore, we introduce a counterfactual regularization strategy to enforce prediction stability under substructure perturbations, encouraging the model to focus on truly causal patterns rather than superficial shortcuts. Extensive experiments across standard and OOD benchmarks demonstrate that CauVQ consistently outperforms state-of-the-art baselines in robustness and interpretability. Our framework offers a promising step toward reliable, explainable, and distribution-aware graph learning.

Code — https://github.com/astar-1/CauVQ

Introduction

Graph Neural Networks (GNNs) have become a cornerstone of graph representation learning (Guo et al. 2021; Wu et al. 2022c; Yu, Yin, and Zhu 2018). However, most existing approaches assume the training and test graphs are independently and identically distributed (IID)—an assumption that rarely holds in real-world scenarios. Due to sampling bias and limited data diversity, graph distributions often exhibit significant shifts between domains (Arjovsky et al. 2019; Fan et al. 2023; Li et al. 2022a). For instance, scaffold splits in molecular graphs can cause performance drops of up to 20% (Hu et al. 2020), severely limiting GNN reliability in high-stakes applications such as drug discovery. To address out-of-distribution (OOD) shifts, various strategies have been explored. Data augmentation methods (Yang et al. 2023; Pham et al. 2023) improve diversity but often rely on heuristics that may not preserve semantic invariance. Invariant learning (Yang et al. 2022; Li et al. 2022b) extracts domain-agnostic features but lacks interpretability. Contrastive learning (Zhu et al. 2024; Suresh et al. 2021) encourages consistency across augmented views but cannot disentangle causal substructures from confounders. Causality-based approaches (Sui et al. 2022; Fan et al. 2023; Wu et al. 2022a) begin to address spurious correlations but typically focus on global or node-level factors, overlooking substructure-level causality that plays a decisive role in many tasks.

Recent studies show that certain substructures critically



(a) Traditional discovery of invariant subgraphs via perturbation.

(b) Our proposed CauVQ framework.

Figure 1: Comparison of traditional perturbation-based methods (a) with CauVQ (b). Traditional methods perturb graphs to find invariant subgraphs but can be unstable. CauVQ decomposes graphs into discrete substructures and uses causal reasoning for robust identification.

*Liang Bai is the corresponding author: bailiang@sxu.edu.cn

determine graph labels, while others act as confounders (Miao, Liu, and Li 2022; Wang et al. 2022). Therefore, isolating invariant, causally-relevant substructures is essential for robust and explainable predictions. As illustrated in Figure 1a), traditional methods attempt to identify such substructures by perturbing the input graph, followed by selection or training strategies that assume invariance. However, these perturbation-based approaches often require strong connectivity constraints and may generate semantically incoherent subgraphs, leading to unstable learning and poor generalization.

To overcome these limitations, we propose CauVQ, a principled substructure representation framework based on discrete vector quantization (VQ) (Van Den Oord, Vinyals et al. 2017; Razavi, Van den Oord, and Vinyals 2019). Instead of relying on perturbation heuristics, we decompose graphs into local substructures and quantize them into a discrete codebook of prototypical motifs, forming symbolic and reusable representations. As shown in Figure 1b), CauVQ further incorporates a causal enhancement mechanism to refine these representations via a learnable interaction matrix and attention over codewords. This suppresses spurious correlations and emphasizes truly predictive substructures. In addition, we introduce a counterfactual regularization strategy to ensure prediction stability under substructure perturbations. CauVQ thus provides a unified, interpretable, and robust approach to graph classification under distribution shifts.

In summary, our contributions are:

- **A causal vector quantization framework** that identifies invariant substructures, yielding stable and interpretable representations for OOD generalization.
- **A causal enhancement mechanism** based on learnable interaction and attention, promoting true causal reasoning.
- **A counterfactual regularization strategy** that enforces prediction stability under substructure perturbations.

Related Work

OOD Generalization in Graph Representation Learning

OOD generalization has become a central challenge in graph learning, leading to four main categories of solutions: data augmentation, invariant learning, disentanglement-based methods, and causality-based methods (Li et al. 2022a). *Data augmentation* techniques, such as MH-Aug (Park et al. 2021) and KDGA (Wu et al. 2022b), generate diverse topologies to enhance robustness, while others combine multiple augmentation strategies or feature perturbations (Lu et al. 2024; Kong et al. 2022; You et al. 2020). However, these heuristics often fail to capture invariant substructure-level patterns. *Invariant learning* methods seek domain-stable representations. For instance, GSAT (Miao, Liu, and Li 2022) uses the information bottleneck to mask irrelevant components, and IGM (Jia et al. 2024) trains on mixed environments to extract shared patterns. Despite their effectiveness, these approaches typically operate in latent spaces without explicitly identifying substructures, limiting interpretability. *Disentanglement-* and *causality-based* methods tackle OOD shifts at a more structural level. OOD-GCL (Li et al. 2024) and FactorGCN (Yang et al. 2020) learn factorized representations, while DisC (Fan et al. 2022) splits graphs into causal and bias subgraphs via edge masking. Causal methods like CAL+ (Sui et al. 2024) and StableGNN (Fan et al. 2023) leverage backdoor adjustment and causal reasoning to remove spurious correlations. However, most of these methods focus on node-level or global causal factors, overlooking fine-grained substructures essential for robust and interpretable OOD generalization.

Causal Substructure Learning and Interpretability

Substructure modeling offers a finer-grained view critical to understanding graph-level behaviors. Several methods aim to explain predictions via motifs or subgraphs, such as GNNExplainer (Ying et al. 2019), SubgraphX (Yuan et al. 2021), and PGExplainer (Luo et al. 2020). These methods improve interpretability post hoc but generally work in deterministic or continuous spaces. Others go beyond explanation—DIVE (Sun et al. 2024) adds regularization to reduce subgraph overlap, and SubDiff (Zhang et al. 2024) injects substructure signals into diffusion models for better molecular representations. While these works leverage substructures, they lack explicit causal modeling, limiting their ability to distinguish truly invariant components from confounders under distribution shifts. Recent studies show some substructures are determinative while others introduce spurious signals (Miao, Liu, and Li 2022; Wang et al. 2022). However, approaches that jointly model discrete substructures and apply causal reasoning remain rare. Addressing this gap is key to achieving both OOD robustness and mechanistic interpretability, motivating our proposed CauVQ framework.

Problem Formulation

We consider a graph classification task, where the dataset consists of N graphs $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^{N}$, with each graph $G_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i)$ comprising a set of nodes $\mathcal{V}_i$, edges $\mathcal{E}_i$, and node features $\mathbf{X}_i \in \mathbb{R}^{|\mathcal{V}_i| \times d}$. The label $y_i \in \mathcal{Y}$ denotes the graph-level class. The goal is to learn a function $f_\theta : G \to \mathcal{Y}$ that generalizes well to unseen graphs. GNNs are commonly adopted to encode such inputs by aggregating neighborhood information through message passing and applying a READOUT function to produce graph-level embeddings. In our work, we use standard GNN backbones as the base encoder without architectural modification.

In OOD scenarios, the test distribution $\mathbb{P}_{\text{test}}(G, y)$ differs from the training distribution $\mathbb{P}_{\text{train}}(G, y)$ due to various types of shifts, including structural shift ($\mathbb{P}_{\text{train}}(\mathcal{V}, \mathcal{E}) \neq \mathbb{P}_{\text{test}}(\mathcal{V}, \mathcal{E})$), feature shift ($\mathbb{P}_{\text{train}}(\mathbf{X}) \neq \mathbb{P}_{\text{test}}(\mathbf{X})$), and size shift (different distributions of $|\mathcal{V}|$). To ensure robustness under such shifts, the objective is to learn a model f_θ that minimizes the worst-case expected risk across a set of possible test distributions $\mathcal{P}$:

$$\min_{\theta} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(G,y)\sim\mathbb{P}} \left[\ell(f_\theta(G), y) \right]. \tag{1}$$

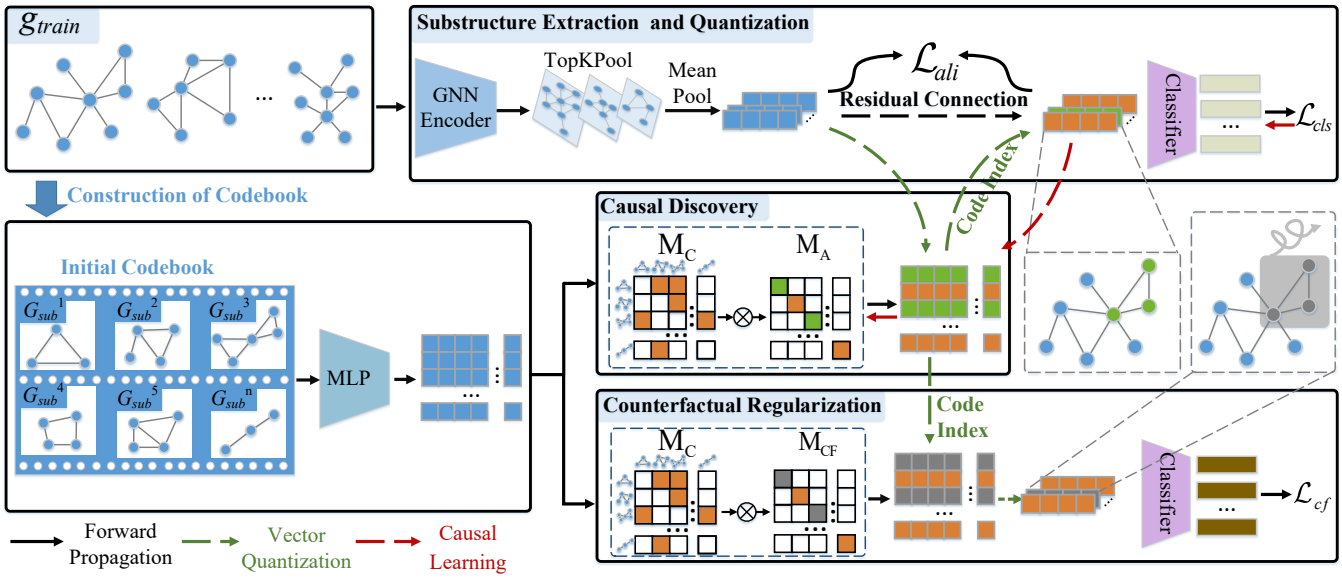


Figure 2: The CauVQ framework consists of three main modules. First, it extracts meaningful subgraphs from the input graph and quantizes them into discrete codewords for stable, interpretable representations. Next, it refines these embeddings by emphasizing causal substructures and suppressing spurious signals through residual optimization and learnable interactions. Finally, the framework applies counterfactual perturbations on substructures to enforce prediction stability, improving robustness under distribution shifts.

To achieve this, the model should focus on invariant substructures $G_{\text{inv}} \subseteq G$ that preserve conditional label distributions across domains:

$$\mathbb{P}_{\text{train}}(y \mid G_{\text{inv}}) = \mathbb{P}_{\text{test}}(y \mid G_{\text{inv}}), \tag{2}$$

while avoiding reliance on variant substructures G_{var} that spuriously correlate with labels in the training set but not in the test set. The core challenge thus lies in accurately identifying G_{inv} and ensuring the model focuses on them while suppressing the misleading influence of G_{var}.

Methodology

Overview

In this work, we propose CauVQ, a novel method designed to tackle the OOD generalization challenge in graph learning. The core insight behind our approach is that graph properties are often determined by a small set of critical substructures that remain invariant across distribution shifts. Unlike existing methods that implicitly encode such invariance in continuous latent spaces, we explicitly identify and model these crucial substructures to ensure robustness against OOD perturbations.

An overview of the CauVQ is depicted in Figure 2. To enable strong OOD generalization, we address two key challenges: how to effectively identify and extract meaningful substructures from a graph, and how to determine which substructures are most significant for predicting graph-level properties among the many extracted candidates.

To tackle these challenges, CauVQ integrates three key components in a unified design. First, we proposed the *Substructure Extraction and Quantization*, where we use a local pooling strategy to partition each graph into substructures and introduce a discrete codebook to quantize their representations, providing stable and interpretable substructure embeddings. Second, we perform *Causal Discovery*, investigating both substructure-substructure interactions and substructure-label relationships through causal inference principles, using backdoor adjustment to highlight truly causal substructures while mitigating the influence of confounders. Finally, we incorporate *Counterfactual Regularization*, simulating interventions on substructures to guide optimization toward more robust and causally consistent solutions, thereby remitting the potential issue where gradient-based optimization may incorrectly associate spurious substructures with labels and trap the model in local optima.

Substructure Extraction and Quantization

To obtain representations of local substructures within graphs, we adopt *TopK Pooling* (Gao and Ji 2019) to select high-importance nodes based on learned scores. Unlike simple node sampling, TopK Pooling preserves the connectivity among the selected nodes, effectively forming a valid subgraph that encapsulates important local patterns. Subsequently, we used MeanPool to produces a local substructure embeddings $\mathbf{Z} \in \mathbb{R}^{\hat{n} \times d}$, where $\hat{n}$ denotes the number of selected nodes and d is the embedding dimension. The operation is defined as:

$$\mathbf{Z} = \text{MeanPool}(\text{TopKPool}(\text{GNN}(\mathbf{A}, \mathbf{X}))), \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ are the adjacency and feature matrices, respectively.

After obtaining the local substructure representation, we emphasize the crucial step of quantization, which forms the core of our CauVQ framework. Rather than operating directly on continuous substructure embeddings, we transform them into a discrete, symbolic space to enable more robust, interpretable, and causally consistent learning. Importantly, the local substructure embeddings obtained via TopK pooling are semantically meaningful substructures selected based on learned importance scores, capturing rich semantic information but not strictly corresponding to canonical topological motifs.

To bridge this gap, we construct a codebook $\mathbf{C} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\} \in \mathbb{R}^{N \times d}$, which serves as a dictionary of discrete, topologically grounded substructure motifs. Specifically, we extract a comprehensive set of candidate substructures from the entire training set using community detection algorithm, and employ Weisfeiler-Lehman (WL) subtree hashing to identify and merge isomorphic substructures. This process ensures that each codeword corresponds to a unique, canonical substructure motif, providing strong topological semantics. In particular, since there are a large number of meaningless motifs in the codebook, we adopted multiple strategies to reduce its size. The specific results are presented in the *Appendix*. Candidate codewords are then initialized and further refined through an MLP:

$$\mathbf{h}_i = \text{MLP}(\text{MeanPool}(G^i_{sub})), \tag{4}$$

where G^i_{sub} denotes the i-th canonical substructure.

Each semantic substructure embedding $\mathbf{Z}$ is then assigned to its nearest topologically grounded codeword through a nearest neighbor search:

$$\hat{\mathbf{Z}} = \text{VQ}(\mathbf{Z}, \mathbf{C}) = \mathbf{h}_{k^*}, \quad \text{where } k^* = \arg\min_k \|\mathbf{Z} - \mathbf{h}_k\|^2. \tag{5}$$

Here, we adopted the "straight-through gradient" to ensure the backpropagation of the gradient. This mapping step anchors semantically rich but potentially noisy substructure embeddings to discrete, interpretable, and topologically meaningful substructure motifs. It provides a consistent foundation for downstream causal enhancement and counterfactual analysis, fully embodying the vector quantization spirit central to CauVQ.

Causal Discovery

Causal Substructure Identification. To explicitly identify the causal substructures, we aim to identify and enhance substructures that exhibit robust, domain-invariant associations with graph-level labels Y, allowing the model to focus on truly causal signals rather than shortcut patterns. Inspired by causal inference, we introduce a strategy similar to backdoor adjustment, encouraging the model to emphasize genuinely causal substructures while suppressing spurious (see *Appendix* for theoretical details).

To this end, we maximize the mutual information (MI) between substructure representations and labels, aiming to reduce the conditional uncertainty of labels given the substructure set. This objective is formulated as:

$$\max MI(\hat{\mathbf{Z}}; Y) = \sum_{c=1}^{C} H(Y = c) - H(Y = c | \hat{\mathbf{Z}} \in S_c), \tag{6}$$

where S_c denotes the set of substructure embeddings associated with label c. Since $H(Y = c)$ is constant, maximizing MI reduces to minimizing conditional entropy:

$$\min -\mathbb{E}_{Y|S_c}[\log \mathbb{P}_\Phi(Y = c | \hat{\mathbf{Z}} \in S_c)], \tag{7}$$

which encourages S_c to provide informative and discriminative support for classification. This formulation ensures that substructure representations retain task-relevant information while filtering out non-causal noise.

However, directly selecting the optimal S_c from the codebook $\mathbf{C}$ is computationally intractable due to its combinatorial nature. To address this, we introduce two *learnable matrices* $\mathbf{M_A} \in \mathbb{R}^{N \times N}$ and $\mathbf{M_C} \in \mathbb{R}^{N \times N}$, which are used to represents the relative importance of each codeword and models relational influence among codewords, respectively. These matrices are jointly optimized during training, allowing the model to softly reweight and aggregate codewords without explicit discrete combinatorial search.

At this point, we can define a refined causal codebook $\hat{\mathbf{C}} = \{\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \ldots, \hat{\mathbf{h}}_N\}$ by aggregating and reweighting substructures:

$$\hat{\mathbf{h}}_i = \mathbf{M_A}[i, i] \cdot \sum_{k=1}^{N} \mathbf{M_C}[k, i] \cdot \mathbf{h}_k. \tag{8}$$

This formulation produces softened, structure-aware substructure motifs, forming a robust foundation for downstream causal reasoning and prediction.

Codebook Refinement and Residual Optimization. Since $\hat{\mathbf{h}}_i$ in $\hat{\mathbf{C}}$ is finite and discrete, important fine-grained semantic details from the original continuous substructure embeddings may be lost during vector quantization. Such information loss can lead to degraded representation expressiveness and hinder the accurate identification of critical substructures. To mitigate this, we adopt a residual strategy inspired by *Residual Vector Quantization (RVQ)* (Zhuang et al. 2023), which refines codebook-based representations and preserves additional semantic information.

Following the same nearest neighbor assignment mechanism as in Eq. 5, each subgraph embedding $\mathbf{Z}$ is first quantized to its nearest refined causal codeword $\hat{\mathbf{h}}_{k^*}$ in $\hat{\mathbf{C}}$. We denote this assigned embedding as:

$$\hat{\mathbf{Z}} = \text{VQ}(\mathbf{Z}, \hat{\mathbf{C}}) = \hat{\mathbf{h}}_{k^*}, \quad \text{where } k^* = \arg\min_k \|\mathbf{Z} - \hat{\mathbf{h}}_k\|^2. \tag{9}$$

Instead of fully replacing $\mathbf{Z}$ with its quantized version $\hat{\mathbf{Z}}$, we introduce a residual connection that explicitly combines them:

$$\tilde{\mathbf{Z}} = \hat{\mathbf{Z}} + \mathbf{Z}. \tag{10}$$

This residual addition retains complementary fine-grained semantic information from the original embedding while benefiting from the discrete, causal-aware structure of $\hat{\mathbf{Z}}$.

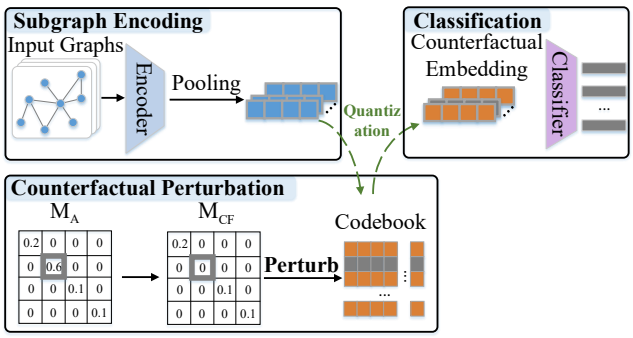


Figure 3: Counterfactual regularization in CauVQ. Perturb substructures to create counterfactual graphs, and reduce reliance on spurious features.

This design ensures that important subtle details are not lost and enhances downstream robustness.

To encourage alignment between the refined representation and the original subgraph embedding, we define an alignment loss:

$$\mathcal{L}_{ali} = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} \|\tilde{\mathbf{Z}}_i - \mathbf{Z}_i\|^2. \tag{11}$$

Here, $\mathcal{T}$ represents the number of graphs in the training set. The final classification loss is computed based on the refined substructure representations:

$$\mathcal{L}_{cls} = -\frac{1}{\mathcal{T}C} \sum_{i=1}^{\mathcal{T}} \sum_{c=1}^{C} \mathbb{P}(Y=c) \log \mathbb{P}_\Phi(Y=c|\{\tilde{\mathbf{Z}}_i\}). \tag{12}$$

Overall, the residual optimization strategy allows our model to preserve crucial semantic information during quantization, resulting in more robust and causally consistent graph-level predictions.

Counterfactual Regularization

While the causal refinement mechanism enhances the model's focus on informative substructures, it remains uncertain whether these attended substructures are truly causal. To further validate the causal impact of these substructures and strengthen model robustness under distribution shifts, we introduce a counterfactual regularization inspired by intervention semantics in causal inference. The overall process is illustrated in Figure 3.

We simulate explicit interventions on the refined codebook $\hat{\mathbf{C}}$ to test whether the model's predictions rely on the attended substructures. Based on the learned causal attention mask $\mathbf{M_A}$, we identify high-attention substructures and construct a counterfactual codebook $\bar{\mathbf{C}} = \{\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \ldots, \bar{\mathbf{h}}_N\}$ by replacing or perturbing these high-attention components.

$$\bar{\mathbf{h}}_i = \mathbf{M_{CF}}[i,i] \cdot \sum_{k=1}^{N} \mathbf{M_C}[k,i] \cdot \mathbf{h}_k, \tag{13}$$

where $\mathbf{M_{CF}}$ is a binary or soft intervention mask indicating which codewords to perturb. Specifically, we determine a threshold B via a KS test on attention scores and intervene on substructures whose attention exceeds B, effectively simulating "removal" of high-confidence causal candidates.

Following the same nearest neighbor assignment as in Eq. 9, each original subgraph embedding $\mathbf{Z}$ is mapped to its counterfactual codeword:

$$\bar{\mathbf{Z}} = \mathrm{VQ}(\mathbf{Z}, \bar{\mathbf{C}}) = \bar{\mathbf{h}}_{k^*}, \quad \text{where } k^* = \arg\min_k \|\mathbf{Z} - \bar{\mathbf{h}}_k\|^2. \tag{14}$$

To enforce that model predictions are not spuriously dependent on these intervened substructures, we require that the classification results of these counterfactual samples to be collapsed. This encourages the model to focus on truly stable, causally reliable signals rather than shortcut correlations.

To achieve this, we define the following counterfactual loss:

$$\mathcal{L}_{cf} = \underbrace{\frac{1}{\hat{\mathcal{T}}C} \sum_{i=1}^{\hat{\mathcal{T}}} \sum_{c=1}^{C} \left(\mathbb{P}_\Phi(Y=c|\bar{\mathbf{Z}}_i) - \bar{p}_c\right)^2}_{\mathcal{L}_{var}} - \underbrace{\sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c}_{\mathcal{L}_{ent}}, \tag{15}$$

$$\bar{p}_c = \frac{1}{\hat{\mathcal{T}}} \sum_{i=1}^{\hat{\mathcal{T}}} \mathbb{P}_\Phi(Y=c|\bar{\mathbf{Z}}_i). \tag{16}$$

Here, $\hat{\mathcal{T}}$ represents the number of counterfactual samples. The first term $\mathcal{L}_{var}$ minimizes prediction variance across counterfactual samples, while the second term $\mathcal{L}_{ent}$ reduces entropy to encourage prediction to be collapsed.

Finally, the total training objective integrates classification, alignment, and counterfactual regularization:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{ali} + \lambda_2 \cdot \mathcal{L}_{cls} + (1-\lambda_2) \cdot \mathcal{L}_{cf}, \tag{17}$$

where λ_1 controls the trade-off between substructure alignment and classification, and λ_2 adjusts the emphasis on counterfactual consistency.

Experiment

Datasets and Baselines

To comprehensively evaluate the OOD generalization performance of CauVQ, we conduct experiments on eight benchmark datasets. We first validate our method on standard graph classification datasets, including *PROTEINS*, *MUTAG*, and *NCI1* (Morris et al. 2020), which cover various biological and chemical molecular structures. To explicitly assess OOD performance, we further adopt the synthetic dataset *Spurious-Motif*, the image-derived graph dataset *MNIST-75sp*, and the large-scale public benchmark *OGB*. These datasets simulate distribution shifts such as structural changes, semantic changes, and size variations. Detailed statistics and split settings for all datasets are summarized in Table 1. The diverse shift types present strong challenges for substructure-level causal generalization.

For baseline comparison, we include representative methods grouped into three categories: *(1) Standard Graph Neural Networks.* GCN and GIN serve as strong standard baselines for graph classification tasks; *(2) Invariant Representation Learning Methods.* GIL (Li et al. 2022b), C2R (Yue

Dataset	Data division	Classes	Avg. Nodes	Avg. Edges
PROTEINS	10-fold	2	39.1	72.8
MUTAG	10-fold	2	17.9	39.6
NCI1	10-fold	2	29.8	27.5
Spurious-Motif (0.5)	Structural Shift	3	29.6	42.0
Spurious-Motif (0.7)	Structural Shift	3	30.8	45.9
Spurious-Motif (0.9)	Structural Shift	3	29.4	42.5
MolHIV	Scaffold/Size	2	25.5	27.5
MolBBBP	Scaffold/Size	2	24.1	26.0
MolSIDER	Scaffold/Size	27	33.6	35.4
MNIST-75sp	Feature Shift	10	66.8	600.2

Table 1: Datasets statistics.

Method	PROTEINS (%)	MUTAG (%)	NCI1 (%)
GK	71.67±0.55	81.58±2.11	62.49±0.27
WL	74.68±0.50	82.05±0.36	76.22±0.18
DGK	75.68±0.54	87.44±2.72	78.81±0.46
Sub2Vec	63.11±1.90	75.94±2.46	58.67±1.34
SortPool	75.48±1.62	86.17±7.53	**79.00±1.68**
DiffPool	76.25±4.21	85.61±6.22	75.06±3.66
CauVQ	**77.90±3.07**	**88.80±6.05**	77.40±1.14

Table 2: Classification performance on ordinary graph classification datasets (Accuracy). The bold denotes the best, underlined the second-best.

et al. 2024), CAL (Sui et al. 2022), InfoIGL (Mao et al. 2024); *(3) Causality- or Rationale-Based Methods.* DIR (Wu et al. 2022a), DisC (Fan et al. 2022), DARE (Yue et al. 2022), GREA (Liu et al. 2022), GSAT (Miao, Liu, and Li 2022). Additional implementation details are provided in *Appendix*.

Overall Performance on Standard Classification

We evaluate the performance of CauVQ in standard in-distribution (IID) settings using three widely adopted molecular graph benchmarks: PROTEINS, MUTAG, and NCI1. The results, summarized in Table 2, show that CauVQ consistently achieves strong performance across all datasets.

In particular, CauVQ achieves the highest accuracy on PROTEINS and MUTAG, and delivers competitive results on NCI1. These results demonstrate that, even in the absence of explicit distribution shifts, our approach benefits from robust substructure representations and causal refinements. This enables the model to effectively capture meaningful graph-level patterns and make accurate predictions.

Overall, these findings confirm that the discrete substructure quantization and causal enhancement mechanisms introduced in CauVQ not only improve robustness under OOD scenarios but also maintain excellent performance in standard IID graph classification tasks.

OOD Generalization: Performance under Distribution Shifts

To evaluate CauVQ under various OOD shift scenarios, we introduced shifts during data partitioning in both real-world and synthetic datasets. The results are summarized in Table 3. On real datasets OGB, we evaluated under scaffold and size shifts, where molecule graphs are split by chemical scaffolds or molecular sizes to simulate real-world distribution gaps. Particularly, CauVQ achieves strong ROC-AUC scores in scaffold shift scenario, such as 79.13% on MolHIV, 71.61% on MolBBBP and 61.64% on MolSIDER, outperforming most baselines and demonstrating robustness to unseen molecular structures. Meanwhile, in the size shift scenario, our method also demonstrates strong performance, particularly on the multi-classification dataset MolSIDER, where it achieves an improvement of up to 2% compared to the baseline methods. On synthetic datasets and feature shift scenarios, including Spurious-Motif with varying bias levels and MNIST-75sp with feature shifts, CauVQ consistently outperforms existing methods. This indicates strong resistance to spurious correlations and adaptability to shifts in node attributes or motifs. Notably, due to CauVQ's stronger emphasis on graph structure offsets, its performance on the MNIST-75sp dataset is slightly lower. Additionally, in robustness experiments on MolBBBP with varying drop ratios from 0% to 70% (Table 4), simulating training data perturbations, CauVQ maintains the highest average ROC-AUC (68.36%) and low variance (3.71). This confirms its stability and reliability even under severe data reduction.

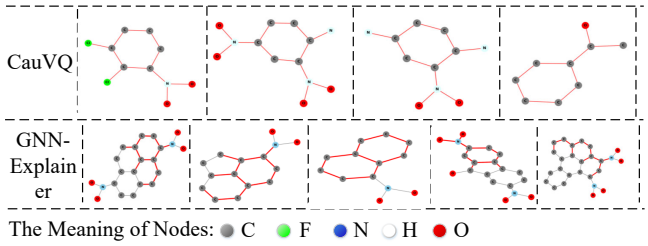


Figure 4: Visualizations of crucial substructures identified on the MUTAG dataset. The top row shows substructures extracted by CauVQ, which successfully capture aromatic amine groups known to be causally linked to the attribute of graphs. The bottom row presents substructures highlighted by GNNExplainer, which mainly focus on carbon ring structures without identifying the true causal motifs.

Interpretability

To demonstrate that our method effectively identifies true causal substructures within graphs, we selected the real-world dataset MUTAG, which provides strong chemical interpretability. We visualized the causal substructures identified by CauVQ on MUTAG, as shown in Figure 4. In this figure, the substructures identified by CauVQ correspond to aromatic amines, which have been extensively verified as carcinogenic in both biological and chemical studies.

In contrast, GNNExplainer primarily highlights edges on carbon rings, indicating a relative insufficiency in pinpointing the aromatic amines that are the true causal factors for molecular carcinogenicity. These results demonstrate the powerful capability of our method to accurately identify meaningful causal substructures, showcasing its superior interpretability and causal faithfulness compared to existing explanation methods.

Method	MolHIV (AUC)		MolBBBP (AUC)		MolSIDER (AUC)		Spurious-Motif (ACC)			MNIST-75sp
	Size	Scaffold	Size	Scaffold	Size	Scaffold	bias=0.5	bias=0.7	bias=0.9	(ACC)
GCN	64.77±1.23	71.28±1.88	71.22±2.74	66.65±2.42	53.55±1.56	61.08±0.75	40.91±3.98	37.72±7.63	35.66±3.23	11.95±1.49
GIN	67.54±1.56	74.47±2.93	78.34±3.21	65.84±2.24	53.45±2.31	59.77±1.76	39.50±4.71	38.72±5.31	37.68±4.47	12.01±0.42
DIR	64.35±6.45	63.03±6.07	76.41±4.43	66.60±1.39	53.90±0.50	49.89±1.15	44.44±6.21	48.91±7.61	41.31±6.52	18.93±4.58
DISC	56.51±10.09	77.31±1.01	75.66±3.16	69.63±2.06	–	–	45.85±6.60	48.85±11.54	38.59±4.00	12.62±1.13
GSAT	**69.86±2.34**	75.24±1.01	72.47±1.89	69.53±2.29	51.75±2.45	60.41±0.96	45.17±4.22	55.67±4.58	47.32±3.67	23.81±1.86
DARE	–	78.19±0.06	–	68.79±1.25	–	60.20±2.87	48.43±10.8	40.02±4.04	43.31±6.31	13.11±0.54
GREA	66.44±4.13	77.14±1.53	77.30±3.52	67.22±1.97	52.53±1.78	58.64±0.52	42.51±4.58	53.31±15.09	45.68±7.79	11.72±0.21
CAL	62.38±1.42	73.39±0.77	79.54±4.81	65.82±3.97	48.30±0.87	59.65±1.16	47.34±6.81	55.41±3.23	44.74±1.28	12.58±1.23
InfoIGL	64.72±2.40	75.37±1.63	61.24±11.27	63.54±1.35	–	–	**67.03±2.71**	63.49±2.95	45.06±0.59	22.10±1.95
C2R	69.58±2.12	79.10±0.06	**81.45±3.67**	69.99±1.22	51.68±2.89	61.31±1.17	52.03±14.37	59.13±4.13	62.03±3.04	**24.33±3.11**
CauVQ	69.73±2.56	**79.13±0.37**	81.10±3.45	**71.61±1.97**	**55.26±2.78**	**61.64±0.77**	60.46±0.72	**64.54±9.79**	**71.38±4.9**	21.62±2.98

Table 3: Results on the datasets OGB (size shift and scaffold shift), Spurious-Motif (structure shift), and MNIST-75sp (feature shift). The "–" indicates unavailable or unreported results.

Method	Drop Ratio (%)						Avg	Var
	0	10	20	30	50	70		
GCN	66.65	64.27	63.91	62.31	61.68	59.75	63.09	4.74
GIN	65.84	66.13	64.01	63.06	63.89	59.15	63.81	5.96
DIR	66.60	65.37	66.32	65.53	65.64	65.37	65.80	**0.22**
CAL	66.13	66.26	66.18	66.66	64.84	60.12	65.03	5.14
GREA	69.50	70.04	68.24	66.85	**66.57**	63.61	67.46	4.56
GSAT	67.22	66.13	67.06	65.06	61.53	62.91	64.98	4.47
InfoIGL	63.42	62.17	65.00	66.45	62.81	61.30	63.52	3.00
C2R	69.99	68.69	67.06	67.15	66.08	65.61	67.43	2.24
CouVQ	**70.41**	**70.55**	**69.58**	**67.71**	65.99	**65.95**	**68.36**	3.71

Table 4: The robustness performance of various methods under the condition that the training set is disturbed.

Setting	MolBBBP Scaffold	MolSIDER Scaffold	Spurious-Motif bias-0.7
w/o VQ	68.19±1.69	58.50±1.75	46.08 ± 1.35
w/o Causal Module	68.07±0.81	58.51±0.61	62.80 ± 6.67
w/o Regularization	67.65±2.39	57.44±0.95	60.74 ± 8.04
CauVQ	**71.61±1.97**	**61.64±0.77**	**64.54 ± 9.79**

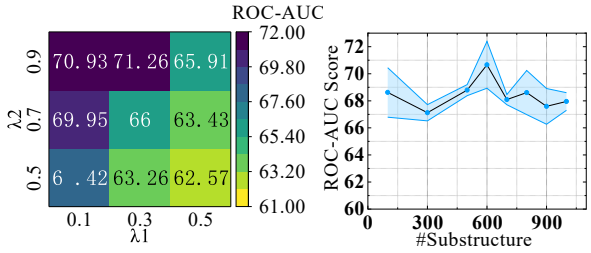
Table 5: The ablation experiment of CauVQ was conducted to test the functions of the three modules.

Ablation Study

To assess the contributions of each component in CauVQ, we conduct an ablation study by removing key modules, as summarized in Table 5. Firstly, we remove VQ results in performance drops across all datasets, most notably on Spurious-Motif (64.54 ± 9.79% → 46.08 ± 1.35%), underscoring its critical role in capturing robust topological information. Subsequently, excluding the causal module leads to consistent declines, which highlighting its importance in identifying causal patterns and mitigating spurious correlations. Finally, the removal of regularization causes significant performance degradation, especially on MolBBBP (71.61 ± 1.97% → 67.65 ± 2.39%).

Hyper-parameter Sensitivity

we analyze the effects of key hyper-parameters introduced in this paper on the model's performance using the MolBBBP dataset. The joint influence of hyper-parameters λ_1 and λ_2 is



(a) (b)

Figure 5: Hyper-parameter sensitivity analysis on the MolBBBP dataset. (a) Joint influence of λ_1 and λ_2 on ROC-AUC performance. (b) Effect of the number of substructures in the codebook on ROC-AUC.

illustrated in Fig. 5a, where λ_1 takes values in [0.1, 0.3, 0.5] and λ_2 in [0.5, 0.7, 0.9]. Overall, when λ_1 is set to 0.1, the model exhibits low sensitivity to changes in λ_2 and performs consistently well across different settings.

For the hyper-parameter representing the number of substructures in the codebook, the results are presented in Fig. 5b. From these results, we observe that an excessive number of substructures reduces the distinctiveness among substructures, negatively impacting the quantization of molecular graphs. Conversely, too few motifs fail to capture important but rare substructures, thereby reducing the model's expressiveness and fitting ability.

Conclusion

In this work, we propose Causal Vector Quantization (CauVQ), a novel method that extends vector quantization with causal inference to tackle the OOD generalization problem in graph classification tasks. Unlike traditional OOD generalizaiton methods, CauVQ explicitly identifies causally relevant substructures, suppresses spurious correlations, and enhances robustness through a causality-aware design. Extensive experiments on synthetic and real-world datasets demonstrate that CauVQ achieves superior performance and offers improved generalization under distribution shifts.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62432006, U21A20473, 62276159, 62576198) and the Fundamental Research Program of Shanxi Province (No.202303021223004).

References

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

Fan, S.; Wang, X.; Mo, Y.; Shi, C.; and Tang, J. 2022. Debiasing graph neural networks via learning disentangled causal substructure. *Advances in Neural Information Processing Systems*, 35: 24934–24946.

Fan, S.; Wang, X.; Shi, C.; Cui, P.; and Wang, B. 2023. Generalizing graph neural networks on out-of-distribution graphs. *IEEE transactions on pattern analysis and machine intelligence*, 46(1): 322–337.

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Guo, Z.; Zhang, C.; Yu, W.; Herr, J.; Wiest, O.; Jiang, M.; and Chawla, N. V. 2021. Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021*, 2559–2567.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Jia, T.; Li, H.; Yang, C.; Tao, T.; and Shi, C. 2024. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8562–8570.

Kong, K.; Li, G.; Ding, M.; Wu, Z.; Zhu, C.; Ghanem, B.; Taylor, G.; and Goldstein, T. 2022. Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 60–69.

Li, H.; Wang, X.; Zhang, Z.; Chen, H.; Zhang, Z.; and Zhu, W. 2024. Disentangled graph self-supervised learning for out-of-distribution generalization. In *Forty-first International Conference on Machine Learning*.

Li, H.; Wang, X.; Zhang, Z.; and Zhu, W. 2022a. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*.

Li, H.; Zhang, Z.; Wang, X.; and Zhu, W. 2022b. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35: 11828–11841.

Liu, G.; Zhao, T.; Xu, J.; Luo, T.; and Jiang, M. 2022. Graph rationalization with environment-based augmentations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1069–1078.

Lu, B.; Zhao, Z.; Gan, X.; Liang, S.; Fu, L.; Wang, X.; and Zhou, C. 2024. Graph out-of-distribution generalization with controllable data augmentation. *IEEE Transactions on Knowledge and Data Engineering*.

Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; and Zhang, X. 2020. Parameterized Explainer for Graph Neural Network. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 19620–19631. Curran Associates, Inc.

Mao, W.; Wu, J.; Liu, H.; Sui, Y.; and Wang, X. 2024. Invariant graph learning meets information bottleneck for out-of-distribution generalization. *arXiv preprint arXiv:2408.01697*.

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

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

Park, H.; Lee, S.; Kim, S.; Park, J.; Jeong, J.; Kim, K.-M.; Ha, J.-W.; and Kim, H. J. 2021. Metropolis-hastings data augmentation for graph neural networks. *Advances in Neural Information Processing Systems*, 34: 19010–19020.

Pham, K.; Le, H.; Ngo, M.; and Tran, T. 2023. Improving out-of-distribution generalization with indirection representations. In *The Eleventh International Conference on Learning Representations*.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Sui, Y.; Mao, W.; Wang, S.; Wang, X.; Wu, J.; He, X.; and Chua, T.-S. 2024. Enhancing out-of-distribution generalization on graphs via causal attention learning. *ACM Transactions on Knowledge Discovery from Data*, 18(5): 1–24.

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*, 1696–1705.

Sun, X.; Wang, L.; Liu, Q.; Wu, S.; Wang, Z.; and Wang, L. 2024. DIVE: subgraph disagreement for graph out-of-distribution generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2794–2805.

Suresh, S.; Li, P.; Hao, C.; and Neville, J. 2021. Adversarial graph augmentation to improve graph contrastive learning. *Advances in Neural Information Processing Systems*, 34: 15920–15933.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, Y.; Magar, R.; Liang, C.; and Barati Farimani, A. 2022. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 62(11): 2713–2725.

Wu, A. P.; Markovich, T.; Berger, B.; Hammerla, N.; and Singh, R. 2022a. Causally-guided regularization of graph attention improves generalizability. *arXiv preprint arXiv:2210.10946*.

Wu, L.; Lin, H.; Huang, Y.; and Li, S. Z. 2022b. Knowledge distillation improves graph structure augmentation for graph neural networks. *Advances in Neural Information Processing Systems*, 35: 11815–11827.

Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2022c. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5): 1–37.

Yang, L.; Zheng, J.; Wang, H.; Liu, Z.; Huang, Z.; Hong, S.; Zhang, W.; and Cui, B. 2023. Individual and structural graph information bottlenecks for out-of-distribution generalization. *IEEE Transactions on Knowledge and Data Engineering*, 36(2): 682–693.

Yang, N.; Zeng, K.; Wu, Q.; Jia, X.; and Yan, J. 2022. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35: 12964–12978.

Yang, Y.; Feng, Z.; Song, M.; and Wang, X. 2020. Factorizable graph convolutional networks. *Advances in Neural Information Processing Systems*, 33: 20286–20296.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3634–3640. International Joint Conferences on Artificial Intelligence Organization.

Yuan, H.; Yu, H.; Wang, J.; Li, K.; and Ji, S. 2021. On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning*, 12241–12252. PMLR.

Yue, L.; Liu, Q.; Du, Y.; An, Y.; Wang, L.; and Chen, E. 2022. DARE: Disentanglement-Augmented Rationale Extraction. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 26603–26617. Curran Associates, Inc.

Yue, L.; Liu, Q.; Liu, Y.; Gao, W.; Yao, F.; and Li, W. 2024. Cooperative classification and rationalization for graph generalization. In *Proceedings of the ACM Web Conference 2024*, 344–352.

Zhang, J.; Liu, Z.; Wang, Y.; Feng, B.; and Li, Y. 2024. Subgdiff: A subgraph diffusion model to improve molecular representation learning. *Advances in Neural Information Processing Systems*, 37: 29620–29656.

Zhu, Y.; Shi, H.; Zhang, Z.; and Tang, S. 2024. Mario: Model agnostic recipe for improving ood generalization of graph contrastive learning. In *Proceedings of the ACM Web Conference 2024*, 300–311.

Zhuang, X.; Zhang, Q.; Ding, K.; Bian, Y.; Wang, X.; Lv, J.; Chen, H.; and Chen, H. 2023. Learning invariant molecular representation in latent discrete space. *Advances in Neural Information Processing Systems*, 36: 78435–78452.