CFD: LEARNING GENERALIZED MOLECULAR REPRESENTATION VIA CONCEPT-ENHANCED FEEDBACK DISENTANGLEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

To accelerate biochemical research, e.g., drug and protein discovery, molecular representation learning (MRL) has attracted much attention. However, most existing methods follow the closed-set assumption that training and testing data share identical distribution, which limits their generalization abilities in out-of-distribution (OOD) cases. In this paper, we explore designing a new disentangled mechanism for learning generalized molecular representation that exhibits robustness against distribution shifts. And an approach of Concept-Enhanced Feedback Disentanglement (CFD) is proposed, whose goal is to exploit the feedback mechanism to learn distribution-agnostic representation. Specifically, we first propose two dedicated variational encoders to separately decompose distribution-agnostic and spurious features. Then, a set of molecule-aware concepts are tapped to focus on invariant substructure characteristics. By fusing these concepts into the disentangled distribution-agnostic features, the generalization ability of the learned molecular representation could be further enhanced. Next, we execute iteratively the disentangled operations based on a feedback received from the previous output. Finally, based on the outputs of multiple feedback iterations, we construct a self-supervised objective to promote the variational encoders to possess the disentangled capability. In the experiments, our method is verified on multiple real-world molecular datasets. The significant performance gains over state-of-the-art baselines demonstrate that our method can effectively disentangle generalized molecular representation in the presence of various distribution shifts.

1 INTRODUCTION

With the rejuvenation of deep neural networks, to advance the development of biochemical research, e.g., drug discovery (Sanchez-Fernandez et al., 2022) and protein design (Ingraham et al., 2019), molecular representation learning (MRL) has attracted growing attention, which aims to transform molecules into low-dimensional and dense vectors (Gilmer et al., 2017; Yang et al., 2019; Rong et al., 2020; Fang et al., 2022; Yang et al., 2022; Zhuang et al., 2023). By means of the learned molecular representation, many downstream tasks could be addressed effectively, including drug property prediction (Wang et al., 2024b), binding affinity analysis (Karimi et al., 2019), and search of antibiotics (Stokes et al., 2020), etc.

Though MRL has achieved significant progress, most existing methods often follow the closed-set assumption, i.e., the training and testing data share the same distribution. However, in practical biochemical research, molecules may exhibit many variants and originate from different distributions (Ji et al., 2022; Mendez et al., 2019). Taking the virtual screening (Lim et al., 2019) as an example, the model may be affected by the distribution shift in the molecule itself, e.g., size (Ji et al., 2022) or scaffold (Wu et al., 2018) changes. Furthermore, some unpredictable events like COVID-19 may bring new target molecules from unknown distributions. Obviously, these scenarios pose significant challenges for closed-set assumption based MRL methods. To this end, improving the generalization of molecular representation is meaningful for reducing the impact of distribution shifts.

Existing methods (Ahuja et al., 2021; Yang et al., 2024) mainly focus on learning generalized representation for regular Euclidean data (e.g., images). Instead, to preserve rich structural information

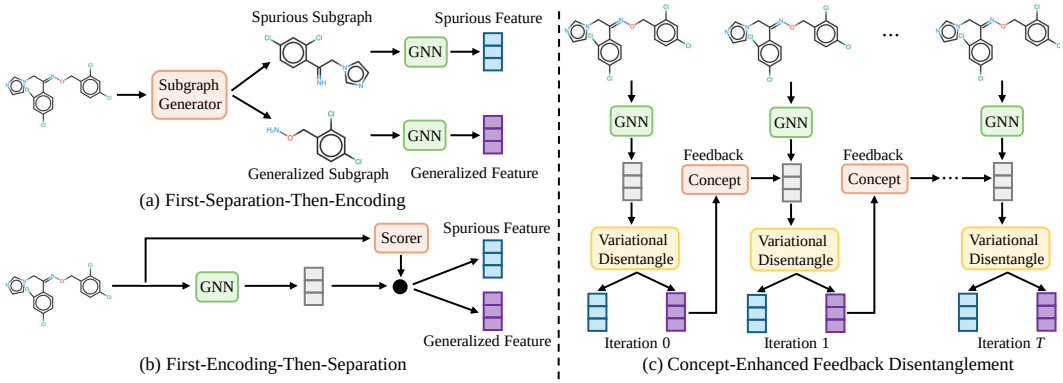

Figure 1: Three different mechanisms for disentangling generalized features. (a) and (b) are commonly used disentangled methods (Zhuang et al., 2023). (a) first uses a subgraph generator to separate the input, then each subgraph is encoded, respectively. (b) first encodes the input, then the representation is separated by a scorer (Zhuang et al., 2023). (c) is our proposed mechanism for learning generalized representation, i.e., Concept-Enhanced Feedback Disentanglement, which exploits the iteration manner to perform variational disentangled operations based on the feedback received from the previous iteration's output. The combination of all decomposed generalized features is taken as the final generalized representation for downstream tasks.

(Gilmer et al., 2017), molecules are often represented as graphs, a typical non-Euclidean data, where atoms and bonds are separately taken as nodes and edges. Thus, the methods (Ahuja et al., 2021; Yang et al., 2024) corresponding to Euclidean data could not be directly applied to molecular graph data. Particularly, the work (Ji et al., 2022) points out that existing OOD methods (Sagawa et al., 2019; Sun & Saenko, 2016; Zhang, 2017) fail to significantly improve the performance of MRL tasks against distribution shifts. In this paper, we explore designing a disentangled mechanism for learning generalized molecular representation.

Recently, some studies (Chen et al., 2022; Li et al., 2022b; Yang et al., 2022) have attempted to improve the generalization of molecular representation, which could be classified into two types, i.e., First-Separation-Then-Encoding and First-Encoding-Then-Separation (Zhuang et al., 2023). Specifically, as shown in Fig. 1 (a), the first type first divides the graph into generalized and spurious subgraphs and then encodes each part separately. However, this kind of method usually relies on the separation quality. For certain molecules involving complex structures, since there is no high-quality decomposition, the performance of these methods may be weakened severely. To this end, the work (Zhuang et al., 2023) introduces a "First-Encoding-Then-Separation" mechanism. As shown in Fig. 1 (b), it first employs a Graph Neural Network (GNN) (Kipf & Welling, 2017; Hamilton et al., 2017) to encode the molecule. Then, another GNN is utilized to calculate the score of molecule representation, which is taken as the critical factor for separating generalized and spurious features. However, for extremely complex and entangled molecules, e.g., biomacromolecules, only leveraging the calculated score could not effectively obtain expected separation results. Meanwhile, this method (Zhuang et al., 2023) fails to focus on certain important substructures containing invariant characteristics, affecting its performance in real-world molecules.

In Fig. 1 (c), we propose a new disentangled mechanism, i.e., Concept-Enhanced Feedback Disentanglement (CFD), which still follows the "First-Encoding-Then-Separation" strategy and exploits the feedback mechanism (Zamir et al., 2017) to learn generalized representation. Concretely, given an input molecule, we first exploit a GNN to encode its representation. Then, two dedicated variational encoders are designed to separately decompose distribution-agnostic generalized features and spurious features. Next, a series of concepts are learned to capture certain critical substructures involving invariant molecule characteristics. By fusing these concepts into the decomposed distribution-agnostic features, the generalization ability could be further strengthened. Later, we iteratively perform the above disentangled operations based on a feedback received from the previous iteration's output. Finally, based on the outputs of multiple feedback iterations, a self-supervised objective is constructed to encourage the variational encoders to own the disentangled capability. Extensive experimental results on multiple molecular datasets demonstrate the superiorities of our method.

The contributions are summarized as follows:

(1) We propose a paradigm of feedback disentanglement that utilizes the iteration manner to decompose generalized molecular representation, improving the robustness against distribution shifts.

(2) We propose a concept mining module focusing on certain critical substructures involving invariant characteristics. By fusing these concepts into the disentangled representation, the generalization ability could be further enhanced.

(3) We conduct extensive experiments on a diverse set of real-world molecule datasets (Zhuang et al., 2023) with various distribution shifts. The significant performance gains over multiple baselines demonstrate that our method can disentangle generalized molecular representation effectively. Moreover, to further verify the ability of capturing important substructure, our method is evaluated on a new task, i.e., Molecule's Ground-State Conformation Prediction (Xu et al., 2024). We observe that plugging our method into Transformer architecture (Xu et al., 2024) could still improve the prediction performance, indicating the superiorities of our method.

2 Related Work

OOD Generalization. Recent research (Wang et al., 2022a; Robey et al., 2021; Zhou et al., 2024) has indicated that existing machine learning methods usually suffer from significant performance degradation under distribution shifts, motivating a surge of works on OOD generalization. Particularly, most existing methods (Evuru et al., 2024; Ratner et al., 2017; Feder et al., 2024) attempt to improve the generalization from the data and representation perspectives. Particularly, an effective strategy is to utilize data augmentation to improve the generalization. Zhou et al. (Zhou et al., 2020) utilize domain information for creating an additive noise to increase the diversity of training data distribution while preserving the semantic information of data. Xu et al. (Xu et al., 2021a) exploit the classical Fourier transformation to obtain the augmented images with diverse styles, which enhances the generalization ability effectively. Besides, by simply mixing up different styles of input data, the work (Zhou et al., 2021) could strengthen the generalization performance significantly.

Another effective strategy is to learn invariant representations (Chang et al., 2020; Creager et al., 2021). In general, the learned invariant representation should satisfy such a principle: 1) sufficiency: shows sufficient predictive power for the target, 2) invariance: contributes to optimal performance for the downstream tasks across all environments. Therefore, some works (Mahajan et al., 2021; Lv et al., 2022; Wang et al., 2022b) explore learning invariant representation from the causal perspective. Furthermore, feature disentanglement (Li et al., 2022a; Zhang et al., 2022; Wu et al., 2021) is also a commonly used mechanism to decompose distribution-agnostic representations. Though these methods have been demonstrated to be effective on many tasks, they often pay attention to regular Euclidean data (e.g., images). However, molecules are usually represented as graphs, which belong to a typical non-Euclidean data. At this time, directly employing the above methods to process graph data does not achieve satisfactory performance. Thus, it is meaningful to design a proper mechanism for learning generalized molecular representation.

OOD Generalization on Molecular Representation. Recently, in order to promote the application of biochemical research, learning generalized molecular representation has attracted growing attention. Some methods (Yang et al., 2022; Chen et al., 2022; Fan et al., 2022) follows the "first-separation-then-encoding" paradigm. And they first design a dedicated module to disentangle the original input graph into two different parts. Then, two independent GNN are utilized to encode these two parts, respectively. However, this kind of methods are prone to relying on the separation results. For certain molecules with complex structures, these methods (Yang et al., 2022; Chen et al., 2022; Fan et al., 2022) could not achieve the expected performance. To this end, Zhuang et al. (Zhuang et al., 2023) propose a new strategy, i.e., "first-encoding-then-separation". Concretely, they first employ a GNN to encode molecules. Then, a disentangled module is designed to perform representation decomposition. Though this method has been shown to be effective, for certain biomacromolecules, the proposed score module could not disentangle invariant representations successfully, limiting its application in real-world molecules. In this paper, we explore a new mechanism, i.e., Concept-Enhanced Feedback Disentanglement, which aims to utilize the iteration manner for learning generalized molecular representation. Meanwhile, a series of concepts are mined to focus on certain critical substructures, which further enhances the generalization ability. Extensive experimental results on multiple real-world molecule datasets demonstrate that our method could indeed disentangle distribution-agnostic invariant molecular representation.

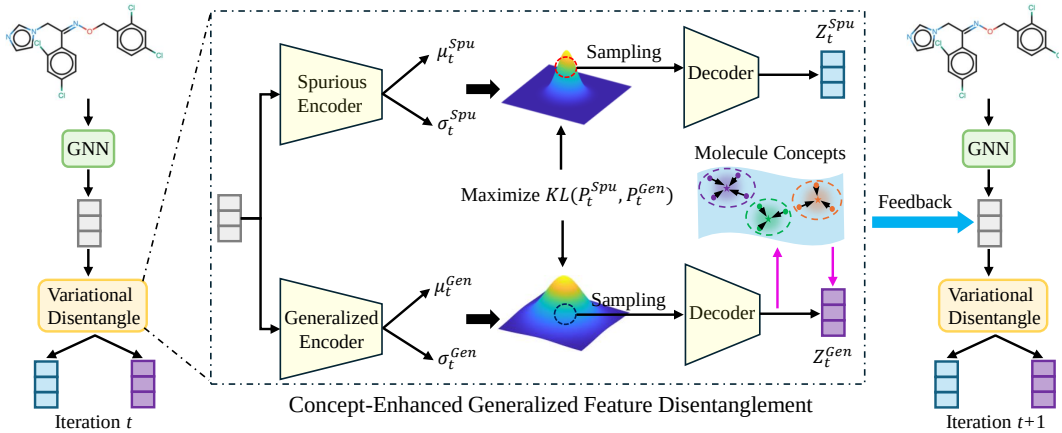

Figure 2: The details of our method for disentangling generalized molecule representation. Our method mainly contains two components: (1) Feedback Disentanglement, and (2) Molecule Concept Extraction. Particularly, based on the current extracted representation, two variational encoders are designed to separately obtain spurious features and generalized features. By maximizing the distance between two estimated distributions, the decomposition ability of two variational encoders could be enhanced. Next, a series of concepts are learned to focus on certain important substructures involving invariant characteristics. By fusing these concepts into the disentangled features, the generalization ability could be further enhanced. Finally, by combining the generalized features from multiple iterations, the output could be taken as the final generalized representation for downstream tasks.

3 Concept-Enhanced Feedback Disentanglement

In this paper, we focus on disentangling generalized molecular representation to reduce the impact of distribution shifts. Here, we first define the OOD generalization problem for MRL.

In general, a molecular graph can be represented as $G = (V, E)$, where V is the graph's node set corresponding to atoms and E denotes the graph's edge sets corresponding to bonds. The goal of MRL task is to predict the target label $\mathcal{Y}$ given the associated input molecule $\mathcal{G}$. The OOD problem on MRL is defined as follows:

$$\min_{f} \max_{e \in \xi} \mathbb{E}_{(\mathcal{G}_i, \mathcal{Y}_i) \backsim P(\mathcal{G}, \mathcal{Y}|e)} [\mathcal{L}(f(\mathcal{G}_i), \mathcal{Y}_i)|e], \tag{1}$$

where ξ represents the support of environments, $f(\cdot)$ is the prediction model and $\mathcal{L}(\cdot, \cdot)$ indicates a loss function. $\mathbb{E}_{(\mathcal{G}_i, \mathcal{Y}_i) \backsim P(\mathcal{G}, \mathcal{Y}|e)} [\mathcal{L}(f(\mathcal{G}_i), \mathcal{Y}_i)|e]$ is called the risk function under a given environment e (Krueger et al., 2021). The reason of resulting in the OOD problem on MRL lies in that the training data only cover very limited environments in ξ while the model is expected to perform well on all environments. Meanwhile, the OOD problem can be categorized into two cases, i.e., covariate and concept shift (Krueger et al., 2021; Widmer & Kubat, 1996). In covariate shift, the distribution of input differs. And the category space of input is the same. Instead, for concept shift, the distribution of input remains the same. But the category space between the training and testing data is inconsistent. In the experiments, we consider these two shift cases.

3.1 Variational Disentanglement for Molecular Representation

As shown in Fig. 2, our method follows "first-encoding-then-separation" mechanism (Zhuang et al., 2023) for learning generalized molecule representation. Different from this work (Zhuang et al., 2023), we leverage the feedback idea (Zamir et al., 2017) to iteratively extract distribution-agnostic representation, which is more applicable to molecules with highly complex structures. Moreover, considering that this work (Zhuang et al., 2023) fails to focus on critical molecular substructures, we propose to mine a series of concepts that involve invariant characteristics of certain molecule substructures, which could be used to further improve the generalization of the learned representation. Next, we will introduce the details of our method.

Specifically, given an input molecule $\mathcal{G}$, we first utilize a GNN to encode it. The output is F:

$$F = \text{GNN}(\mathcal{G}) \in \mathbb{R}^{|\mathcal{V}| \times d}, \tag{2}$$

where $|\mathcal{V}|$ is the number of atoms in $\mathcal{G}$, and d is the dimensionality of the encoded features. Then, a spurious encoder (i.e., $W_{\mu}^{Spu}(\cdot)$ and $W_{\sigma}^{Spu}(\cdot)$) and a generalized encoder (i.e., $W_{\mu}^{Gen}(\cdot)$ and $W_{\sigma}^{Gen}(\cdot)$), each consisting of three fully-connected layers, are designed to separately estimate the corresponding distribution. The processes are shown as follows:

$$\mu_t^{Spu} = W_{\mu}^{Spu}(Q_t), \;\; \sigma_t^{Spu} = W_{\sigma}^{Spu}(Q_t), \;\; \mu_t^{Gen} = W_{\mu}^{Gen}(Q_t), \;\; \sigma_t^{Gen} = W_{\sigma}^{Gen}(Q_t), \quad (3)$$

where $Q_t = \phi([Z_{t-1}^{Gen}, Z_{t-2}^{Gen}])$. $t = 0, \cdots, T$. And $Q_0 = F$, $Q_1 = Z_0^{Gen}$. '$[\cdot,\cdot]$' indicates the concatenation operation. '$\phi(\cdot)$' represents a fully-connected layer that transforms channels. $\mu_t^{Spu} \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\mu_t^{Gen} \in \mathbb{R}^{|\mathcal{V}| \times d}$ are the estimated means for spurious and generalized features at the current iteration t. $\sigma_t^{Spu} \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\sigma_t^{Gen} \in \mathbb{R}^{|\mathcal{V}| \times d}$ represent the estimated variances for spurious and generalized features. Next, we separately perform a sampling operation to obtain the disentangled features $Z_t^{Spu} \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\tilde{Z}_t^{Gen} \in \mathbb{R}^{|\mathcal{V}| \times d}$:

$$Z_t^{Spu} = D^{Spu}(\mu_t^{Spu} + \epsilon \cdot \exp(\sigma_t^{Spu})), \;\; \tilde{Z}_t^{Gen} = D^{Gen}(\mu_t^{Gen} + \epsilon \cdot \exp(\sigma_t^{Gen})), \quad (4)$$

where $D^{Spu}(\cdot)$ and $D^{Gen}(\cdot)$ are two defined decoders, each consisting of three fully-connected layers. ϵ denotes Gaussian noise sampled from $\mathcal{N}(0, I)$.

To promote Z_t^{Spu} and $\tilde{Z}_t^{Gen}$ to separately contain rich spurious and distribution-agnostic information, we first enlarge the distance between the two estimated distributions. Meanwhile, to encourage $\tilde{Z}_t^{Gen}$ to involve plentiful invariant characteristics related to the input molecule, we further define a contrast loss. Finally, the loss $\mathcal{L}_{vd}^t$ for variational disentanglement is defined as follows:

$$\mathcal{L}_{vd}^t = -\log \frac{\exp(\mathrm{sim}(\tilde{Z}_t^{Gen}, F)/\tau)}{\exp(\mathrm{sim}(\tilde{Z}_t^{Gen}, F)/\tau) + \exp(\mathrm{sim}(Z_t^{Spu}, F)/\tau)} - \lambda \cdot KL(P_t^{Spu}, P_t^{Gen}), \quad (5)$$

where $\mathrm{sim}(\tilde{Z}_t^{Gen}, F)$ denotes the average of the cosine similarity between all corresponding elements of $\tilde{Z}_t^{Gen}$ and F. τ and λ are two hyper-parameters, which are separately set to 1.0 and 0.01 in the experiments. $P_t^{Spu} = \mathcal{N}(\mu_t^{Spu}, \sigma_t^{Spu})$ and $P_t^{Gen} = \mathcal{N}(\mu_t^{Gen}, \sigma_t^{Gen})$. Obviously, by minimizing the loss $\mathcal{L}_{vd}^t$, the gap between $\tilde{Z}_t^{Gen}$ and Z_t^{Spu} could be enlarged from both semantic and distribution perspectives, which enhances the disentangled ability of our method.

3.2 Learning Molecule Concepts

In general, the critical substructures involving invariant characteristics usually play an important role in downstream tasks. To this end, we explore mining a series of concepts that do not specialize to one particular type or class of molecule, which is instrumental in strengthening the generalization of the learned representation (Locatello et al., 2020).

Concretely, the concepts are defined as $C \backsim \mathcal{N}(\mu_c, \sigma_c) \in \mathbb{R}^{K \times d}$, where μ_c and σ_c are learnable parameters. K indicates the number of concepts. Next, based on the disentangled representation $\tilde{Z}_t^{Gen} \in \mathbb{R}^{|\mathcal{V}| \times d}$, we calculate descriptors that represent molecule-level information:

$$\mathcal{M} = \mathrm{KAN}(\tilde{Z}_t^{Gen}), \;\; \mathcal{H}_i = \sum_{j=1}^{|\mathcal{V}|} \frac{\exp(\mathcal{M}_{j,i})}{\sum_{i=1}^{K} \exp(\mathcal{M}_{j,i})} (\tilde{Z}_{t,j}^{Gen} - C_i), \quad (6)$$

where $i = 1, \cdots, K$. Here, we employ one KAN layer (Liu et al., 2024) to transform $\tilde{Z}_t^{Gen}$, which outputs $\mathcal{M} \in \mathbb{R}^{|\mathcal{V}| \times K}$. Different from MLPs having fixed activation functions on neurons, KANs have learnable activation functions (Liu et al., 2024), which is beneficial for improving the flexibility of the learned concepts. $\mathcal{H} \in \mathbb{R}^{K \times d}$ represents the output descriptors. '$\tilde{Z}_{t,j}^{Gen} - C_i$' denotes the residual operation. By means of this operation, we can assign molecule substructure features to the corresponding concept $C_i \in \mathbb{R}^d$. Finally, the learned concepts are utilized to enhance the current representation $\tilde{Z}_t^{Gen}$. The corresponding processes are shown as follows:

$$\tilde{\mathcal{H}} = \mathrm{KAN}(\mathcal{H}_r), \;\; Z_t^{Gen} = \mathrm{KAN}([\tilde{\mathcal{H}}, \tilde{Z}_t^{Gen}]), \quad (7)$$

where $H_r \in \mathbb{R}^{1 \times Kd}$ is the reshaped result of $\mathcal{H}$. $\tilde{\mathcal{H}} \in \mathbb{R}^{1 \times d}$. $Z_t^{Gen} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the output generalized representation at the step t. By the concatenation operation, the learned concepts could be fused into the current decomposed representation, which is conducive to enhancing the generalization of the output representation Z_t^{Gen}. In the experiments, Z_t^{Gen} is utilized as the representation for computing the corresponding task loss, e.g., molecule property prediction, which promotes the learned concepts to capture plentiful task-specific invariant molecule characteristics.

3.3 FEEDBACK SEPARATING GENERALIZED MOLECULAR REPRESENTATION

Existing methods (Zhuang et al., 2023; Yang et al., 2022) for learning generalized molecular representation generally employ a one-step disentangled strategy, i.e., directly separating the input into generalized and spurious parts. However, in practice, we may encounter some unknown biomacromolecules containing more atoms and highly complex structures. At this time, using the original one-step mechanism could not obtain satisfactory disentangled results. To this end, we exploit the feedback idea (Zamir et al., 2017) to iteratively separate generalized molecular representation.

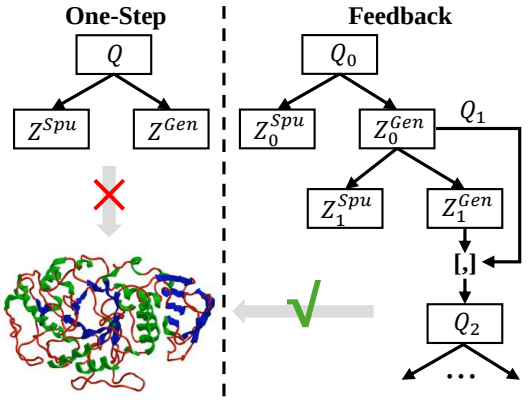

Figure 3: Computation graph of One-Step vs Feedback Disentanglement. For the molecules with complex structures, using one-step disentanglement could not obtain the expected separation results. Instead, the feedback mechanism with multiple iteration steps is beneficial for decomposing complex structures.

As shown in Fig. 3, we assume that the number of feedback iterations is T. Given the current feature $Q_t \in \mathbb{R}^{|\mathcal{V}| \times d}$ that is the concatenation of the previous two iteration outputs, we first leverage the above operations to disentangle Q_t into Z_t^{Spu} and Z_t^{Gen}, where $t = 0, \cdots, T$. Then, $Q_{t+1} = \phi([Z_{t-1}^{Gen}, Z_t^{Gen}])$ is taken as the input of the current step to repeat the above disentangled operations. In the computation graph of feedback disentanglement (as shown in Fig. 3), we can observe that compared with One-Step disentanglement, the feedback strategy follows a hierarchical disentangled mechanism, which is instrumental in improving the ability of decomposing the molecules with complex structures. Finally, through multiple feedback iterations, we can obtain a series of generalized outputs, i.e., (Z_0^{Gen}, Z_1^{Gen}, $\cdots$, Z_T^{Gen}), and spurious outputs, i.e., (Z_0^{Spu}, Z_1^{Spu}, $\cdots$, Z_T^{Spu}). A self-supervised objective is constructed to close the distance between the output generalized features at each step and enlarge the distance between the decomposed generalized and spurious features at the same step, which strengthens the disentangled ability effectively:

$$\mathcal{L}_{fd} = -\log(\frac{1}{T \times T} \sum_{i=0}^{T} \sum_{j=0}^{T} \text{sim}(Z_i^{Gen}, Z_j^{Gen})_{i \neq j} - \beta \cdot \frac{1}{T} \sum_{i=0}^{T} \text{sim}(Z_i^{Gen}, Z_i^{Spu})), \tag{8}$$

where β is a hyper-parameter, which is set to 0.5 in the experiments. Finally, we take the concatenation result, i.e., $Z^{Gen} = \psi([Z_0^{Gen}, Z_1^{Gen}, \cdots, Z_T^{Gen}]) \in \mathbb{R}^{|\mathcal{V}| \times d}$, as the last generalized representation for downstream tasks, where $\psi(\cdot)$ indicates a fully-connected layer that transforms the channels.

For specific downstream tasks, their goals are to provide the generalized representation Z^{Gen} with predictive capabilities. Since our method is to learn generalized molecule representations, it could be applied to different molecular tasks. During training, the choice of prediction loss function depends on the type of task. For classification loss, we employ the cross-entropy loss, while for regression tasks, we utilize the mean squared error loss. Taking the binary classification task as an example, the cross-entropy loss is computed between the predicted probability $\hat{\mathcal{Y}} = Prob(Z^{Gen})$ and the ground-truth label $\mathcal{Y}$:

$$\mathcal{L}_{pred} = \mathcal{Y} \log \hat{\mathcal{Y}} + (1 - \mathcal{Y}) \log(1 - \hat{\mathcal{Y}}), \tag{9}$$

where $Prob(\cdot)$ represents a function that computes probability. Finally, the learning objective can be defined as the weighted sum of the above losses:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{pred} + \alpha_1 \cdot \mathcal{L}_{fd} + \alpha_2 \cdot \sum_{t=0}^{T} \mathcal{L}_{vd}^t, \tag{10}$$

where α_1 and α_2 are hyper-parameters, which are separately set to 0.1 and 0.01 in the experiments.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the research questions. (1) Can our method CFD achieve better OOD generalization performance against baselines? (2) Does our method

Table 1: Evaluation performance on GOOD benchmark. - denotes abnormal results caused by under-fitting declared in the leaderboard, and / denotes that the method cannot be applied to this dataset. For different distribution shifts, our method obtains superior generalization performance.

Method	GOOD-HIV ↑				GOOD-ZINC ↓				GOOD-PCBA ↑			
	scaffold		size		scaffold		size		scaffold		size	
	covariate	concept	covariate	concept	covariate	concept	covariate	concept	covariate	concept	covariate	concept
ERM	69.55	72.48	59.19	61.91	0.1802	0.1301	0.2319	0.1325	17.11	21.93	17.75	15.60
IRM	70.17	71.78	59.94	-(-)	0.2164	0.1339	0.6984	0.1336	16.89	22.37	17.68	15.82
VREx	69.34	72.21	58.49	61.21	0.1815	0.1287	0.2270	0.1311	17.10	21.65	17.80	15.85
GroupDRO	68.15	71.48	57.75	59.77	0.1870	0.1323	0.2377	0.1333	16.55	21.91	16.74	15.21
Coral	70.69	72.96	59.39	60.29	0.1769	0.1303	0.2292	0.1261	17.00	22.00	17.83	16.88
DANN	69.43	71.70	62.38	65.15	0.1746	0.1269	0.2326	0.1348	17.20	22.03	17.71	15.78
Mixup	70.65	71.89	59.11	62.80	0.2066	0.1391	0.2531	0.1547	16.52	20.52	17.42	13.71
DIR	68.44	71.40	57.67	74.39	0.3682	0.2543	0.4578	0.3146	16.33	23.82	16.04	16.80
GSAT	70.07	72.51	60.73	56.96	0.1418	0.1066	0.2101	0.1038	16.45	20.18	17.57	13.52
GREA	71.98	70.76	60.11	60.96	0.1691	0.1157	0.2100	0.1273	16.28	20.23	17.12	13.82
CAL	69.12	72.49	59.34	56.16	/	/	/	/	15.87	18.62	16.92	13.01
DisC	58.85	64.82	49.33	74.11	/	/	/	/	/	/	/	/
MoleOOD	69.39	69.08	58.63	55.90	0.2752	0.1996	0.3468	0.2275	12.90	12.92	12.64	10.30
CIGA	69.40	71.65	61.81	73.62	/	/	/	/	/	/	/	/
iMoLD	72.93	74.32	62.86	77.43	0.1410	0.1014	0.1863	0.1029	17.32	22.58	18.02	18.21
Ours (CFD)	**76.42**	**77.83**	**64.14**	**79.28**	**0.1187**	**0.0765**	**0.1421**	**0.0852**	**19.78**	**25.64**	**19.18**	**20.03**

possess the ability to capture important substructures and improve the performance of molecular substructure prediction? (3) How does each component contribute to the final performance?

4.1 Experimental Setup

Datasets. For OOD molecular representation learning, we follow the settings of the work (Zhuang et al., 2023) and employ two real-world datasets, i.e., GOOD (Gui et al., 2022) that is a systematic benchmark tailored specifically for graph OOD problems, and DrugOOD (Ji et al., 2022) that is a OOD benchmark for AI-aided drug discovery. Besides, for molecule's ground-state prediction, we follow the settings of the work (Xu et al., 2024) and utilize Molecule3D (Xu et al., 2021b) and QM9 (Ramakrishnan et al., 2014) to evaluate the ability of our method for focusing on substructures. In Appendix, we will introduce more details about these datasets.

Implementation Details. Our method is used for learning specific molecule representation and could be plugged into existing methods (Zhuang et al., 2023; Xu et al., 2024). The number of the learned concepts is set to 12. The number of feedback iterations is set to 8. It is worth noting that during feedback, we do not introduce new parameters.

Baselines. For OOD molecular representation learning, we mainly compare our method with some common approaches for non-Euclidean data, i.e., ERM (Vapnik, 2013), IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2019), Coral (Sun & Saenko, 2016), DANN (Ganin et al., 2016), Mixup (Zhang, 2017), and iMoLD (Zhuang et al., 2023). Meanwhile, our method is compared with some graph-specific algorithms, e.g., graph OOD algorithms such as CAL (Sui et al., 2022), DisC (Fan et al., 2022), MoleOOD (Yang et al., 2022), and CIGA (Chen et al., 2022), as well as interpretable graph learning methods, e.g., DIR (Wu et al., 2022), GSAT (Miao et al., 2022), and GREA (Liu et al., 2022).

4.2 Analysis of OOD Molecular Representation Learning

To evaluate the generalization ability of our method, our method is evaluated on GOOD benchmark (Gui et al., 2022). Table 1 shows the performance. We can observe that some graph-specific OOD methods could achieve effective performance on some synthetic datasets (e.g., to predict whether a specific motif is present in a synthetic graph). However, they do not perform well on molecules with complex data structures and distributions, which shows that our method could apply to molecules with complex structures and improve the generalization performance. Besides, the performance of iMoLD (Zhuang et al., 2023) outperforms MoleOOD (Yang et al., 2022), which shows the effectiveness of the strategy "first-encoding-then-separation". However, iMoLD (Zhuang et al., 2023) fails to capture the critical substructures that involve certain invariant characteristics, which may affect its performance. We can see that our method improves the performance significantly. This indicates that the feedback disentanglement could indeed decompose generalized features. Meanwhile, our method could be applied to complex molecules. This also shows that the learned concepts are beneficial for capturing important substructures, which further strengthens the generalization ability.

Table 2 further evaluates our method on DrugOOD dataset (Ji et al., 2022). Particularly, the recent work MILI (Wang et al., 2024a) follows "first-separation-then-encoding" mechanism. And it first identifies privileged substructure and develops a dual-head GNN to achieve invariant representations of substructures. When the molecular substructures are accurately identified, the generalization performance of OOD molecule representation could be improved effectively. We can see that our method still outperforms iMoLD (Zhuang et al., 2023) and MILI (Wang et al., 2024a). This demonstrates that using the feedback iteration idea is beneficial for extracting generalized molecular characteristics. By fusing the concepts capturing plentiful substructure content, the generalization ability could be further enhanced.

Table 2: Evaluation performance on DrugOOD dataset.

Method	IC50 ↑			EC50 ↑		
	Assay	Scaffold	Size	Assay	Scaffold	Size
ERM	71.63	68.79	67.50	67.39	64.98	65.10
IRM	71.15	67.22	61.58	67.77	63.86	59.19
Coral	71.28	68.36	64.53	72.08	64.83	58.47
MixUp	71.49	68.59	67.79	67.81	65.77	65.77
DIR	69.84	66.33	62.92	65.81	63.76	61.56
GSAT	70.59	66.45	66.70	73.82	64.25	62.65
GREA	70.23	67.02	66.59	74.17	64.50	62.81
CAL	70.09	65.90	66.42	74.54	65.19	61.21
DisC	61.40	62.70	61.43	63.71	60.57	57.38
MoleOOD	71.62	68.58	65.62	72.69	65.74	65.51
CIGA	71.86	69.14	66.92	69.15	67.32	65.65
iMoLD	72.11	68.84	67.92	77.48	67.79	67.09
MILI	72.67	69.58	68.40	77.11	68.07	65.97
Ours (CFD)	**73.86**	**70.02**	**69.73**	**78.32**	**69.13**	**67.62**

Table 3: Performance of molecule's ground-state prediction based on Molecule3D and QM9 datasets.

	Validation			Test		
	D-MAE↓	D-RMSE↓	C-RMSD↓	D-MAE↓	D-RMSE↓	C-RMSD↓
	(a) Molecule3D Random Split					
RDKit DG	0.581	0.930	1.054	0.582	0.932	1.055
RDKit ETKDG	0.575	0.941	0.998	0.576	0.942	0.999
DeeperGCN-DAGNN (Xu et al., 2021b)	0.509	0.849	-	0.571	0.961	-
GINE (Hu et al., 2019)	0.590	1.014	1.116	0.592	1.018	1.116
GATv2 (Brody et al., 2021)	0.563	0.983	1.082	0.564	0.986	1.083
GPS (Rampášek et al., 2022)	0.528	0.909	1.036	0.529	0.911	1.038
GTMGC (Xu et al., 2024)	0.432	0.719	0.712	0.433	0.721	0.713
GTMGC + Ours	**0.397**	**0.682**	**0.684**	**0.407**	**0.695**	**0.688**
	(b) QM9					
RDKit DG	0.358	0.616	0.722	0.358	0.615	0.722
RDKit ETKDG	0.355	0.621	0.691	0.355	0.621	0.689
GINE (Hu et al., 2019)	0.357	0.673	0.685	0.357	0.669	0.693
GATv2 (Brody et al., 2021)	0.339	0.663	0.661	0.339	0.659	0.666
GPS (Rampášek et al., 2022)	0.326	0.644	0.662	0.326	0.640	0.666
GTMGC (Xu et al., 2024)	0.262	0.468	0.362	0.264	0.470	0.367
GTMGC + Ours	**0.223**	**0.434**	**0.305**	**0.218**	**0.442**	**0.309**

4.3 Analysis of Molecule's Ground-State Prediction

The molecular ground-state conformation refers to the lowest energy state of a molecule on its potential energy surface (Xu et al., 2024). It belongs to substructures of original molecules, which play a crucial role in determining the physical, chemical, and biological properties of molecules. By evaluating this task, we can directly observe CFD's ability for capturing critical substructures. Here, our method is directly plugged into GTMGC (Xu et al., 2024) that is a transformer architecture. The training processes and optimization objectives are kept unchanged. In Table 3, we can observe that plugging our method could improve GTMGC's prediction performance effectively. This indicates that our method is able to capture critical substructure characteristics. Meanwhile, our method could be plugged into Transformer architecture, showing our method's flexibility.

In Fig. 4, we show two prediction examples from our method. We can see that compared with ground truth, our method could accurately capture those critical substructures, further demonstrating the effectiveness of our method.

4.4 Ablation Analysis

To evaluate the effectiveness of each component, we take GOOD-ZINC (Gui et al., 2022) as the dataset to make an ablation analysis.

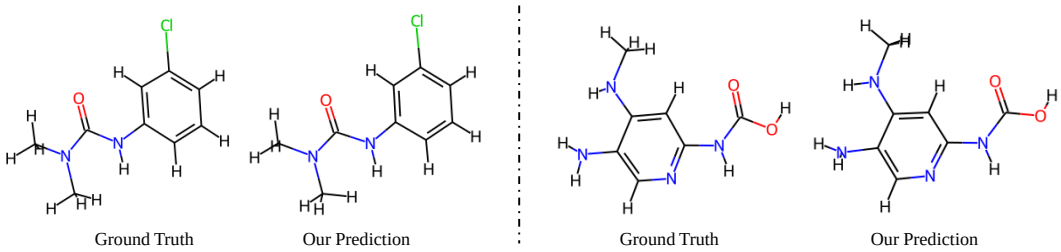

Figure 4: The prediction results of the molecule's ground-state conformation. Compared with ground truth, our method could localize the critical substructures, improving the prediction accuracy.

Analysis of our method. As shown in Fig. 2, our method mainly consists of two components, i.e., feedback disentanglement and molecule concept mining. Meanwhile, these two modules could be taken as the independent module. In Table 4, we make an ablation experiment. 'Concepts' represents that the concept learning module is plugged into ERM (Vapnik, 2013). We can see that for different shifts, using feedback disentanglement and learning molecule concepts could improve the performance significantly. This shows that our method could effectively separate generalized features from spurious features. Meanwhile, the mined concepts could capture critical substructure characteristics, further enhancing the disentangled features' generalization.

Table 4: Analysis of our CFD method. 'Feedback' and 'Concepts' separately indicate feedback disentanglement and molecule concept learning. 'Cov-Sold', 'Cet-Sold', 'Cov-Size', and 'Cet-Size' denote scaffold-covariate, scaffold-concept, size-covariate, and size-concept shifts, respectively.

Feedback	Concepts	Cov-Sold	Cet-Sold	Cov-Size	Cet-Size
✓		0.1391	0.0923	0.1754	0.0983
	✓	0.1685	0.1172	0.2046	0.1268
✓	✓	**0.1187**	**0.0765**	**0.1421**	**0.0852**

Number of feedback iterations. To achieve the expected generalized representation, we employ multiple feedback iterations. In Table 5, we can see that the iteration number affects the generalization performance. For molecules with complex structures, using fewer iterations could not decompose the expected generalized representation effectively, which weakens the performance. Meanwhile, we observe that the performance change is very small when the iteration number is larger than 12 for the given dataset. For our method, the performance of using 8 feedback iterations is the best.

Table 5: Analysis of feedback iteration number.

Iterations	Cov-Sold	Cet-Sold	Cov-Size	Cet-Size
1	0.1378	0.1012	0.1789	0.1021
4	0.1285	0.1008	0.1573	0.0963
8	0.1187	**0.0765**	**0.1421**	0.0852
12	0.1191	0.0799	0.1453	**0.0844**
16	**0.1172**	0.0805	0.1479	0.0867

Number of molecule concepts. As shown in Fig. 2, we explore mining a series of molecule concepts to focus on critical substructures. In Table 6, we can observe that the concept number affects the performance. Using fewer concepts could not capture sufficient substructure characteristics effectively, which reduces the performance. For our method, when the number is set to 12, the generalization performance is the best. Besides, during mining molecule concepts (equation 6 and equation 7), we employ KAN (Liu et al., 2024) to improve the concept accuracy. Here, we find that replacing KAN with MLP results in performance degradation. For example, for Cov-Sold case, the performance is decreased by around 0.6%, indicating KAN's effectiveness.

Table 6: Analysis of molecule concept number. Here, we only change the concept number. Other operations and settings are kept unchanged.

Number	Cov-Sold	Cet-Sold	Cov-Size	Cet-Size
8	0.1233	0.0892	0.1541	0.0927
12	**0.1187**	0.0765	**0.1421**	**0.0852**
16	0.1201	**0.0743**	0.1478	0.0874

5 Conclusion

In this paper, we propose a new method, i.e., Concept-Enhanced Feedback Disentanglement, for learning generalized molecular representation against distribution shifts. Particularly, by performing multiple feedback iterations, our method could progressively decompose expected features involving rich generalized information. Meanwhile, fusing the concepts that focus on substructures could further strengthen the generalization. Extensive experimental results on multiple datasets and tasks demonstrate that our method could indeed disentangle generalized molecular representations.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *NeurIPS*, 34:3438–3450, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, 2024.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *ICML*, pp. 1448–1458. PMLR, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *NeurIPS*, 35:22131–22148, 2022.

Yongqiang Chen, Yatao Bian, Kaiwen Zhou, Binghui Xie, Bo Han, and James Cheng. Does invariant graph learning via environment augmentation learn invariance? *NeurIPS*, 36, 2024.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, pp. 2189–2200. PMLR, 2021.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Utkarsh Tyagi, Dinesh Manocha, et al. Coda: Constrained generation based data augmentation for low-resource nlp. *arXiv preprint arXiv:2404.00415*, 2024.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. *NeurIPS*, 35:24934–24946, 2022.

Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *AAAI*, volume 36, pp. 3968–3976, 2022.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. Causal-structure driven augmentations for text ood generalization. *NeurIPS*, 36, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, pp. 1263–1272. PMLR, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *NeurIPS*, 35:2059–2073, 2022.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *NeurIPS*, 30, 2017.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *NeurIPS*, 32, 2019.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Andreĭ Nikolaevich Kolmogorov. *On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables*. American Mathematical Society, 1961.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pp. 5815–5826. PMLR, 2021.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *AAAI*, volume 36, pp. 7399–7407, 2022a.

Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. *NeurIPS*, 35:11828–11841, 2022b.

Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of chemical information and modeling*, 59(9):3981–3988, 2019.

Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph rationalization with environment-based augmentations. In *KDD*, pp. 1069–1078, 2022.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 33:11525–11538, 2020.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pp. 8046–8056, 2022.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pp. 7313–7324. PMLR, 2021.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *ICML*, pp. 15524–15543. PMLR, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *NeurIPS*, 35: 14501–14515, 2022.

Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *NeurIPS*, 30, 2017.

Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *NeurIPS*, 34:20210–20229, 2021.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *NeurIPS*, 33:12559–12571, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Contrastive learning of image-and structure-based representations in drug discovery. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *KDD*, pp. 1696–1705, 2022.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022a.

Ruijia Wang, Haoran Dai, Cheng Yang, Le Song, and Chuan Shi. Advancing molecule invariant representation via privileged substructure identification. In *KDD*, pp. 3188–3199, 2024a.

Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. *arXiv preprint arXiv:2206.05263*, 2022b.

Yili Wang, Kaixiong Zhou, Ninghao Liu, Ying Wang, and Xin Wang. Efficient sharpness-aware minimization for molecular graph transformer models. *ICLR*, 2024b.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4178–4193, 2021.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. *ICLR*, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Guikun Xu, Yongquan Jiang, PengChuan Lei, Yan Yang, and Jim Chen. Gtmgc: Using graph transformer to predict molecule's ground-state conformation. In *ICLR*, 2024.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pp. 14383–14392, 2021a.

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs. *arXiv preprint arXiv:2110.01717*, 2021b.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pp. 1–28, 2024.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388, 2019.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *NeurIPS*, 35:12964–12978, 2022.

Amir R Zamir, Te-Lin Wu, Lin Sun, William B Shen, Bertram E Shi, Jitendra Malik, and Silvio Savarese. Feedback networks. In *CVPR*, pp. 1308–1317, 2017.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, pp. 8024–8034, 2022.

Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, volume 34, pp. 13025–13032, 2020.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning invariant molecular representation in latent discrete space. *NeurIPS*, 36:78435–78452, 2023.

A APPENDIX

To learn generalized molecular representation, this paper proposes a new method, i.e., Concept-Enhanced Feedback Disentanglement, aiming to use feedback decomposition and concepts involving critical substructures to improve the generalization of the learned representation. In the appendix, we provide more experimental details, additional analyses, and more prediction results.

B DATASETS AND METRICS

Datasets. For molecular representation learning, we employ two benchmarks, i.e., GOOD (Gui et al., 2022) and DrugOOD (Ji et al., 2022). Particularly, we follow the work (Zhuang et al., 2023) to utilize three molecular datasets for the prediction task: (1) GOOD-HIV (Wu et al., 2018), where the purpose is binary classification to predict whether a molecule can inhibit HIV; (2) GOOD-ZINC (Gómez-Bombarelli et al., 2018), which is a regression dataset aimed at predicting molecular solubility; and (3) GOOD-PCBA (Wu et al., 2018), which includes 128 bioassays and forms 128 binary classification tasks. Each dataset contains two environment-splitting strategies (scaffold and size), and two shift types (covariate and concept) are applied per splitting outcome, resulting in a total of 12 distinct datasets (as shown in Table 1). Furthermore, DrugOOD (Ji et al., 2022) provides three environment-splitting strategies, including assay, scaffold, and size, and applies these three splitting to two measurements (IC50 and EC50). As a result, we obtain 6 datasets (as shown in Table 2), and each dataset contains a binary classification task for drug target binding affinity prediction

For molecule's ground-state prediction, we follow the work (Xu et al., 2024) to utilize Molecule3D (Xu et al., 2021b) and QM9 (Wu et al., 2018) as benchmarks. Particularly, Molecule3D aims to use DNNs to predict the ground-state 3D geometries of molecules based solely on their 2D graph structure. And this large-scale dataset comprises approximately 4 million molecules, each with its own 2D molecular graph, ground-state 3D geometric structure, and four additional quantum properties. In this paper, we employ the random splitting according to the same distribution based on the molecule's core component. Besides, QM9 is a quantum chemistry dataset that provides geometry, energy, electronic, and thermodynamic properties for nearly 130,000 organic molecules with 9 heavy atoms, containing molecular most stable conformation.

Metrics. For molecular representation learning, we report the ROC-AUC score for GOOD-HIV and DrugOOD datasets as the task is binary classification. For GOOD-ZINC, we use the Mean Average Error (MAE) since the task is regression. While for GOOD-PCBA, we use Average Precision (AP) averaged over all tasks as the evaluation metric due to extremely imbalanced classes.

For molecule's ground-state prediction, we follow the work (Xu et al., 2024) to utilize three metrics, i.e., D-MAE, D-RMSE, and C-RMSD. Concretely, given a dataset with N interatomic distances, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between the prediction $\{d_i^*\}_{i=1}^N$ and the ground truth $\{d_i\}_{i=1}^N$ are used to evaluate the performance at node-pair level:

$$\mathrm{D-MAE}(\{d_i\}_{i=1}^N, \{d_i^*\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^*|, \tag{11}$$

$$\mathrm{D-RMSE}(\{d_i\}_{i=1}^N, \{d_i^*\}_{i=1}^N) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - d_i^*)^2}. \tag{12}$$

Additionally, C-RMSD between the ground truth $\mathbf{G}$ and the prediction $\mathbf{G}^*$ is computed as follows:

$$\mathrm{C-RMSD}(\mathbf{G}, \mathbf{G}^*) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_i - g_i^*\|_2^2}. \tag{13}$$

C THE CAUSE OF MOLECULAR DISTRIBUTION SHIFT

For molecule data, there exist many kinds of distribution shifts. In this paper, we investigate three types of shift strategies, i.e., scaffold, size, and assay. Here, we give more explanations.

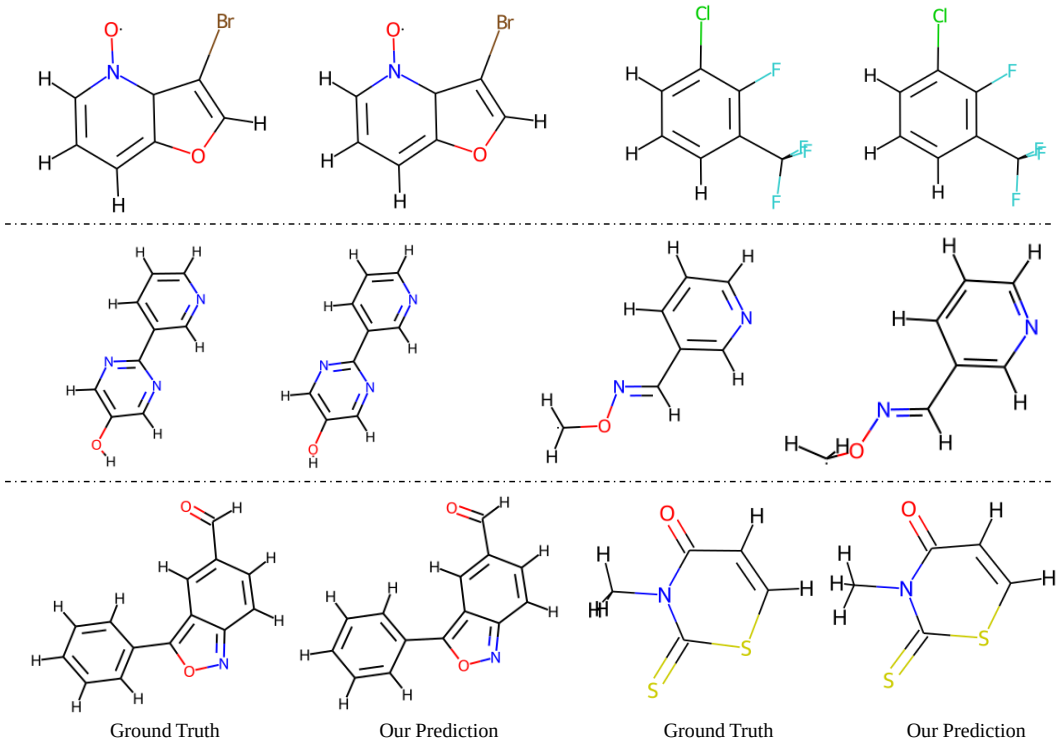

Figure 5: More prediction results of molecule's ground-state conformation. We can observe that our method could accurately localize critical ground-state parts that determine the physical, chemical, and biological properties of the molecule in most cases, which demonstrates the effectiveness of our method for learning generalized molecule representation and capturing substructure characteristics.

(1) **Scaffold Environment**: molecular scaffold is the fundamental structure of a molecule with desirable bioactive properties. Molecules with the same scaffold belong to the same environment. Distribution shift arises when there is a change in the molecular scaffold.

(2) **Size Environment**: the size of a molecule refers to the total number of atoms in the molecule. Molecular size is also an inherent structural characteristics of molecular graphs. The distribution shift occurs when size changes.

(3) **Assay Environment**: assay is an experimental method as an examination or determination for molecular characteristics. Due to variations in assay environments and targets, activity values measured by different assays often differ significantly. Samples tested within the same assay belong to a single environment, while a change in the assay leads to a distribution shift.

D More Ablation Experiments of Hyper-Parameters

In equation 10, we utilize two hyper-parameters, i.e., α_1 and α_2, to balance the two loss terms, i.e., $\mathcal{L}_{fd}$ and $\mathcal{L}_{vd}$. Here, based on GOOD-ZINC with Covariate-Scaffold shift, we make an ablation analysis. We only change the settings of hyper-parameters.

Analysis of hyper-parameter α_1. In equation 10, the goal of the loss term $\mathcal{L}_{fd}$ is to encourage the feedback disentangled process to decompose generalized representation effectively. Thus, the setting of hyper-parameter α_1 is important. Here, we make an experiment. When α_1 is separately set to 0.5, 0.1, and 0.05, the corresponding performance is 0.1251, **0.1187**, and 0.1239.

Analysis of hyper-parameter α_2. In equation 10, the loss term $\mathcal{L}_{vd}^t$ is to promote the decomposed features to involve plentiful invariant characteristics related to the input molecule while enlarging the estimated distribution gap between the generalized and spurious features. Thus, the hyper-parameter α_2 is still an important hyper-parameter. Here, we make an experiment. When α_2 is separately set to 0.05, 0.01, and 0.005, the performance is 0.1248, **0.1187**, and 0.1225.

Algorithm 1 Training Process of Concept-Enhanced Feedback Disentanglement

Input: Molecule Dataset $\{G, Y\}$, randomly initialized GNN, randomly initialized spurious encoder and generalized encoder, randomly initialized concept learning module, weight α_1 for the loss $\mathcal{L}_{fd}$, weight α_2 for the loss $\mathcal{L}_{vd}^{t}$.
Output: Trained parameters.
while *train* **do**
  Sample molecules from the given dataset $\{G, Y\}$.
  Utilize a GNN to extract the representation F.
  for t = 0 **do**
    $Q_0 = F$.
    Use equation 3 to estimate corresponding means (i.e., μ_0^{Spu} and μ_0^{Gen}) and variances (i.e., σ_0^{Spu} and σ_0^{Gen}) for spurious and generalized features.
    Perform sampling and encoding operations to obtain Z_0^{Spu} and $\tilde{Z}_0^{Gen}$ based on equation 4.
    Learn molecule concepts C and calculate the generalized representation Z_0^{Gen} based on equation 6 and equation 7.
  end
  for t = 1 **do**
    $Q_1 = Z_0^{Gen}$.
    Use equation 3 to estimate corresponding means (i.e., μ_1^{Spu} and μ_1^{Gen}) and variances (i.e., σ_1^{Spu} and σ_1^{Gen}) for spurious and generalized features.
    Perform sampling and encoding operations to obtain Z_1^{Spu} and $\tilde{Z}_1^{Gen}$ based on equation 4.
    Learn molecule concepts C and calculate the generalized representation Z_1^{Gen} based on equation 6 and equation 7.
  end
  for t = 2, ..., T **do**
    $Q_t = \phi([Z_{t-1}^{Gen}, Z_{t-2}^{Gen}])$.
    Use equation 3 to estimate corresponding means (i.e., μ_t^{Spu} and μ_t^{Gen}) and variances (i.e., σ_t^{Spu} and σ_t^{Gen}) for spurious and generalized features.
    Perform sampling and encoding operations to obtain Z_t^{Spu} and $\tilde{Z}_t^{Gen}$ based on equation 4.
    Learn molecule concepts C and calculate the generalized representation Z_t^{Gen} based on equation 6 and equation 7.
  end
  Calculate the overall training objective $\mathcal{L}_{\text{task}}$ using equation 5, equation 8, equation 9.
  Update the trained parameters based on equation 10.
end

Finally, Algorithm 1 shows the training processes of our method. For step $t >= 2$, the feedback is from the outputs of the previous two steps. By means of feedback disentanglement, our method could decompose generalized molecular representation effectively, which improves the performance of downstream tasks. In Fig. 5, we show more prediction results of molecule's ground-state conformation. We can see that our method could localize critical ground-state molecular substructures, which further demonstrates the superiorities of our method.

E More Chemical Inspiration of Our Method

This method is mainly based on the following chemical observations:

(1) A chemical molecule is a complex composed of multiple atoms and bonds. The reason for the OOD problem is the structural change of molecules. For example, changes in molecular chirality can lead to deviations in molecular properties. Thus, learning robust molecule representation is important for molecule-based tasks, e.g., drug discovery.

(2) Lines 082-095 have indicated that existing methods could be classified into two types, i.e., First-Separation-Then-Encoding and First-Encoding-Then-Separation (as shown in Fig. 1 (a) and (b)). As indicated in Lines 272-283 and Fig. 3, these methods for learning generalized molecular representation generally employ a one-step disentangled strategy, i.e., directly separating the input into

generalized and spurious parts. However, in practice, we may encounter some unknown biomacromolecules containing more atoms and highly complex structures, e.g., entanglement of multiple substructures. At this time, using the original one-step mechanism could not obtain satisfactory disentangled results. To this end, we exploit the feedback idea to iteratively and progressively separate generalized molecular representation.

(3) In general, the molecule's ground-state conformation belongs to the substructures of molecules and determines their properties. Therefore, it is important to capture substructure-related characteristics. To this end, we define a series of concepts to build connections between molecule substructures and corresponding concepts, which further improves the performance.

F More Experiments to Evaluate Invariance

Table 7: To further demonstrate the ability of disentangling invariance, we further evaluate our method on Two-Piece Graph Datasets Chen et al. (2024).

Datasets	{0.8, 0.9}	{0.7, 0.9}	Avg.
DisC Fan et al. (2022)	45.06	39.42	42.24
CIGA Chen et al. (2022)	57.87	43.62	50.75
GALA Chen et al. (2024)	76.42	72.50	74.46
Ours	**78.26**	**74.32**	**76.29**

We follow the work Chen et al. (2024) and further evaluate our method on Two-Piece Graph Datasets Chen et al. (2024). Particularly, each dataset is generated from a variation of two-piece graph model, denoted as a, b, where a refers to the invariant correlation strength and b refers to the spurious correlation strength. The results are shown in Table 7.

We can observe that our method indeed improves the performance. Particularly, when the invariant correlation strength is weaker than the spurious correlation strength, the performance of our method outperforms state-of-the-art methods, further demonstrating that our method is indeed beneficial for disentangling invariant characteristics.

G Comparison between RVQ and Molecule Concept Learning

Our molecule concept learning module is significantly different from Residual Vector Quantization (RVQ) Zhuang et al. (2023). In essence, RVQ is to perform feature replacement. Our method is to build the connection to capture critical molecule substructures.:

(1) **The motivation is different**. RVQ aims to leverage the discretized continuous representation to improve the model generalization. As shown in Fig. 6, our method is to learn a series of concepts that could capture critical molecule structure information involving plentiful task-specific invariant molecule characteristics.

(2) **The operations are different**. RVQ first introduces a shared learnable codebook as a discrete latent space. For each node representation in molecular graph, RVQ looks up and fetches the nearest neighbor in the codebook and outputs it as the result. Meanwhile, a sum operation is further used to strengthen the representation ability. During training, similar to VQ-VAE Van Den Oord et al. (2017), RVQ employs the exponential moving average updates for the codebook.

Differently, our concepts are first sampled from a learnable Gaussian distribution. And we employ KAN operation to promote the learning of concepts. Importantly, as shown in equation 6, the weighted residual operation is to align critical molecule substructure features to the corresponding concept, building the connection between molecule substructures and corresponding concepts. During training, we only utilize the task loss, e.g., molecule property prediction, to promote the learned concepts to capture plentiful task-specific invariant molecule characteristics.

(3) **Experimental results show the effectiveness of our method.** Taking GOOD-ZINC as the example, for iMoLD Zhuang et al. (2023), we first remove the RVQ operation and keep other operations unchanged. We observe that removing RVQ does not affect the performance. Instead, our concept mining module could be plugged into iMoLD to further improve its performance. For the

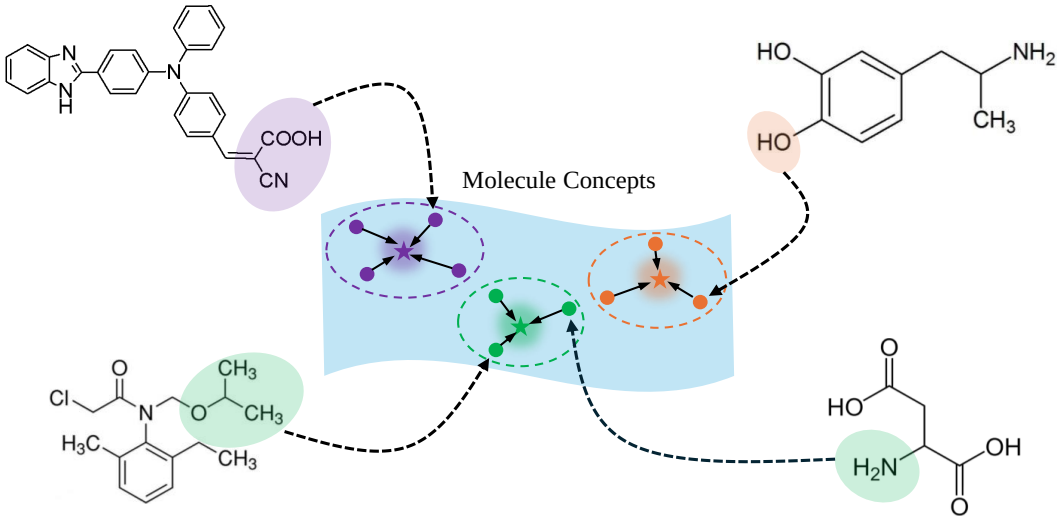

Figure 6: Molecule concepts (colorful stars) are learned from all molecule structures, which are not specific to certain molecule structures. By means of the residual operation (as shown in solid arrows), we can build the connection between molecule substructures and corresponding concepts, which is beneficial for aligning critical molecule substructure features to the corresponding concept.

covariate case of scaffold, the performance is improved by 0.0142. Furthermore, the prediction results of the molecule's ground-state conformation in Fig. 4 and 5 also show that our learned concepts are beneficial for capturing critical substructures.

H Further Analysis of the Loss Function

equation 10 is the optimization objective, consisting of three terms, i.e., $\mathcal{L}_{pred}$, $\mathcal{L}_{fd}$ (as shown in equation 8), and $\mathcal{L}_{vd}^{t}$ (as shown in equation 5).

Proposition 1. $\mathcal{L}_{vd}^{t}$ mainly is a contrastive loss. Minimizing $\mathcal{L}_{vd}^{t}$ could enlarge the gap between the disentangled $\tilde{Z}_t^{Gen}$ and Z_t^{Spu}. Meanwhile, $\tilde{Z}_t^{Gen}$ is promoted to retain plentiful input-related content.

Analysis. The first term of $\mathcal{L}_{vd}^{t}$ is related to the InfoNCE loss Oord et al. (2018). Concretely, we set $\mathcal{B}_1 = \exp(\mathrm{sim}(\tilde{Z}_t^{Gen}, F))$ and $\mathcal{B}_2 = \exp(\mathrm{sim}(Z_t^{Spu}, F))$:

$$\begin{aligned} \mathbb{E}[\log \frac{\mathcal{B}_1}{\mathcal{B}_1 + \mathcal{B}_2}] &= \mathbb{E}[\log \mathcal{B}_1] - \mathbb{E}[\log(\mathcal{B}_1 + \mathcal{B}_2)] \\ &\leq \mathbb{E}[\log \mathcal{B}_1] - \mathbb{E}[\log \mathcal{B}_2] \end{aligned} \tag{14}$$

Obviously, maximizing equation 14 narrows the semantic distance between $\tilde{Z}_t^{Gen}$ and the input F and enlarges the semantic gap between Z_t^{Spu} and F. Equally, this operation enlarges the gap between $\tilde{Z}_t^{Gen}$ and Z_t^{Spu}, which is beneficial for strengthening the disentangled ability.

Proposition 2. Minimizing $\mathcal{L}_{fd}$ is instrumental in achieving the invariant representation Z^{Gen}.

Analysis. In equation 8, $\mathcal{L}_{fd}$ mainly consists of two terms, i.e., $\sum_{i=0}^{T}\sum_{j=0}^{T} \mathrm{sim}(Z_i^{Gen}, Z_j^{Gen})_{i \neq j}$ and $\sum_{i=0}^{T} \mathrm{sim}(Z_i^{Gen}, Z_i^{Spu})$:

$$\sum_{i=0}^{T}\sum_{j=0}^{T} \mathrm{sim}(Z_i^{Gen}, Z_j^{Gen})_{i \neq j} - \sum_{i=0}^{T} \mathrm{sim}(Z_i^{Gen}, Z_i^{Spu}) \tag{15}$$

Maximizing equation 15 first enlarges the gap between Z_i^{Gen} and Z_i^{Spu} at the i-th iteration, which is beneficial for enhancing the disentangled ability. Secondly, Z_i^{Gen} and Z_j^{Gen} separately represent disentangled generalized outputs at different iterations. Here, we assume that when taking a representation involving plentiful invariant information as the input, the disentangled output should contain less spurious information, i.e., the output is equal to the input. Therefore, narrowing the semantic distance is conducive to learning invariant representations across various environments.

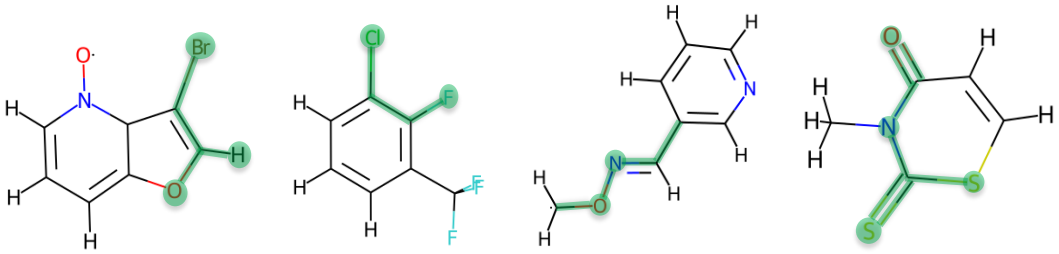

Figure 7: Visualization examples of molecule concepts. Here, we directly project the captured concepts into the predicted molecules. The green regions represent the focused content.

I Examining Importance of Concepts Across Iterations

Table 5 analyzes the impact of iteration number. We can observe that when the iteration number increases, inserting concepts could improve the performance effectively, which indicates the effectiveness of the learned concepts. To further analyze the impact of iteration number on concepts, we also make a new ablation experiment, i.e., removing the concepts at the 4th, and 8th iteration. The results are shown as follows:

Table 8: Examining importance of concepts across iterations.

Iterations	Concepts	CovSold	CetSold	CovSize	CetSize
4	✓	0.1285	0.1008	0.1573	0.0963
4	–	0.1402	0.1117	0.1738	0.1102
8	✓	0.1187	0.0765	0.1421	0.0852
8	–	0.1189	0.0783	0.1449	0.0897

We can observe that when the iteration number is small, inserting concepts could significantly affect the performance. Instead, when the iteration number is large, the impact of concepts is somewhat weak. The reason may be that when the iteration number is small, leveraging concepts is beneficial for capturing critical substructures, improving the performance. As the iteration number increases, concepts progressively pay attention to meaningful substructure characteristics, reducing the iteration impact on performance.

J More Analysis of Molecule Concepts

We perform an ablation experiment of the decoder that calculates task-related representation. We observe that for classification and regression tasks, using decoders with more layers leads to the performance degradation. The reason may be that this operation introduces more parameters, resulting in the overfitting risk.

In Fig. 7, we show some visualization examples of molecule concepts. Since these concepts are learned from all molecule data, they are not specific to certain kinds of molecules, which is beneficial for improving the generalization of molecule concepts. For different predictions, we can observe that the learned concepts could indeed focus on critical molecule substructures, which strengthens the performance of molecule prediction.

K Analysis of Computational Costs

We agree with you that introducing feedback iterations indeed increases computational costs. During training, based on the same batch size, the memory is increased by around 0.9GB. However, in Fig. 8, we observe that the convergence speed is significantly faster than the baseline Zhuang et al. (2023).

During inference, compared with one-step disentanglement, the inference time of using 8 iterations is increased by around 0.06s. Although the inference time is longer, the disentangled invariant representations are more accurate, which is instrumental in improving the generalization ability. Particularly, for molecules with multiple atoms and various structures, we observe that employing

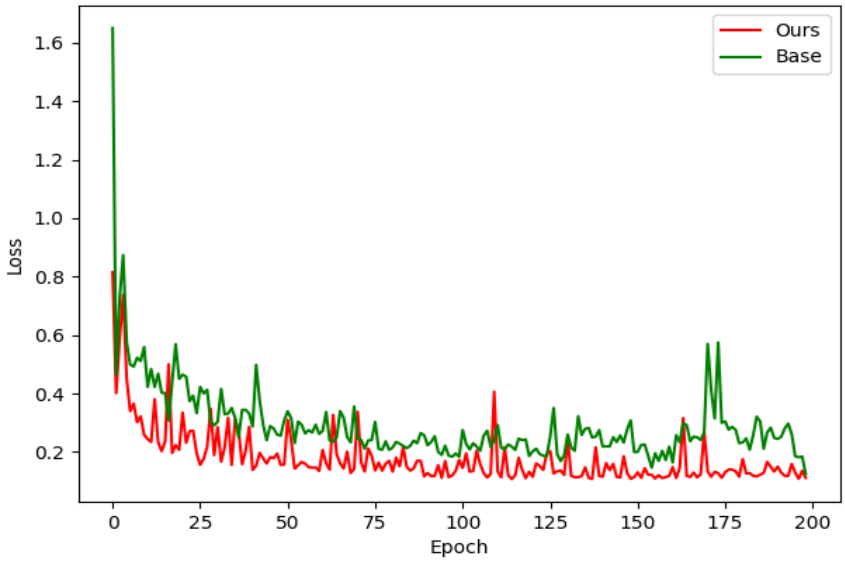

Figure 8: The training loss of the baseline Zhuang et al. (2023) and our method.

our method could significantly improve the performance of downstream tasks, further demonstrating the effectiveness of our method.

L Quantify Contribution of Feedback Mechanism

To quantify the contribution of feedback iterations, we make an ablation experiment of the iteration number. Here, we only change the iteration number. Other operations are kept unchanged. The results are shown in Table 5.

Table 9: Quantify contribution of feedback mechanism on highly complex molecules.

1	4	8	12	16
0.3014	0.1867	0.1341	0.1279	0.1254

We can observe that compared with one-step disentanglement, using multiple feedback iterations indeed improves the performance, indicating that our method is beneficial for decomposing invariant molecule representations. Besides, we also observe that when the iteration number is larger than 12, the performance improvement is limited. The reason is that for the current dataset, using 12 iterations is sufficient for disentangling highly complex molecules.

Finally, we observe that complex molecules usually contain more atoms. Therefore, to further demonstrate the effectiveness of our method on highly complex molecules, we select those molecules with more than 12 atoms for additional evaluation. The results are shown in Table 9.

We can see that for complex molecules, using one-step disentanglement could not decompose invariant characteristics effectively. Using multiple feedback iterations could indeed decouple the representations involving plentiful invariant information, which strengthens the generalization ability.

M Protein-Ligand Binding-Affinity Prediction

For OOD molecular representation learning, in Table 1 and 2, we show the corresponding results. Furthermore, in Table 3, our method is evaluated on an additional task, i.e., molecule's ground-state prediction. These results demonstrate that our method could effectively disentangle invariant molecule characteristics, which strengthens the generalization ability.

Finally, to further demonstrate the superiorities of our method, we evaluate our method on protein-ligand binding-affinity prediction. Binding affinity is an important metric in biology and medicinal chemistry that measures the strength of a reversible association between biological macromolecules,

such as proteins or DNA, and small-molecule ligands, such as drugs Boulougouri et al. (2024). Here, we follow the work Boulougouri et al. (2024) and employ the same datasets and metrics to verify our feedback disentanglement method. The results are shown in Table 10.

Table 10: Analysis of protein-ligand binding-affinity prediction.

Method	**RMSE**	**MAE**	**R**
MAT	1.457	1.154	0.747
DimeNet	1.453	1.138	0.752
CMPNN	1.408	1.117	0.765
SR-BIND Boulougouri et al. (2024)	1.383	1.122	0.780
CFD	**1.356**	**1.073**	**0.794**

Here, MAT, DimeNet, CMPNN are three GNN-based methods. **RMSE**, **MAE**, and **R** are three predefined metrics. We can observe that for the challenging protein-ligand binding-affinity prediction, employing our method still improves the performance for the given three metrics. This further demonstrates that for biological macromolecules, using feedback disentanglement is meaningful, which improves the generalization of the learned molecule representations.

N More Interpretations of KAN

In general, molecules consist of multiple atoms and bonds. For certain biomacromolecules, their structures are entanglement of various substructures, which poses the challenge for learning molecule concepts that focus on critical substructures.

Inspired by the Kolmogorov-Arnold representation theorem Kolmogorov (1961), KAN allows learning custom activations of the network. In this way, it is possible to analyze the contribution of individual components of the input, then providing a more transparent view of the network's decision-making process.

Thus, using KAN could sufficiently convert the analyzed components to the corresponding concepts that involve plentiful molecule-aware substructure information, which is beneficial for strengthening the generalization of molecule representations.