Advancing Molecule Invariant Representation via Privileged Substructure Identification

Ruijia Wang
wangruijia@bupt.edu.cn
Beijing University of Posts and Telecommunications
China Telecom Cloud Computing Research Institute

Haoran Dai
stkm_dhr@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Cheng Yang
yangcheng@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Le Song
songle@biomap.com
BioMap Research
Mohamed bin Zayed University of Artificial Intelligence

Chuan Shi*
shichuan@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

ABSTRACT

Graph neural networks (GNNs) have revolutionized molecule representation learning by modeling molecules as graphs, with atoms represented as nodes and chemical bonds as edges. Despite their progress, they struggle with out-of-distribution scenarios, such as changes in size or scaffold of molecules with identical properties. Some studies attempt to mitigate this issue through graph invariant learning, which penalizes prediction variance across environments to learn invariant representations. But in the realm of molecules, core functional groups forming privileged substructures dominate molecular properties and remain invariant across distribution shifts. This highlights the need for integrating this prior knowledge and ensuring the environment split compatible with molecule invariant learning. To bridge this gap, we propose a novel framework named MILI. Specifically, we first formalize molecule invariant learning based on privileged substructure identification and introduce substructure invariance constraint. Building on this foundation, we theoretically establish two criteria for environment splits conducive to molecule invariant learning. Inspired by these criteria, we develop a dual-head graph neural network. A shared identifier identifies privileged substructures, while environment and task heads generate predictions based on variant and privileged substructures. Through the interaction of two heads, the environments are split and optimized to meet our criteria. The unified MILI guarantees that molecule invariant learning and environment split achieve mutual enhancement from theoretical analysis and network design. Extensive experiments across eight benchmarks validate the effectiveness of MILI compared to state-of-the-art baselines.

*Corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671886

CCS CONCEPTS

• **Applied computing** → **Computational biology**.

KEYWORDS

Molecule Representation Learning, Molecule Invariant Learning, Privileged Substructure Identification, Environment Split

ACM Reference Format:
Ruijia Wang, Haoran Dai, Cheng Yang, Le Song, and Chuan Shi. 2024. Advancing Molecule Invariant Representation via Privileged Substructure Identification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671886

1 INTRODUCTION

Molecules, the quintessential components of matter, hold a pivotal role in scientific exploration [2] and drug discovery [9, 53], where deciphering their properties can drive substantial innovation. Within this realm, Molecule Representation Learning (MRL) [4, 12, 18] becomes a vital field of study, which embeds complex molecules into computationally manageable vector representations. Recently, Graph Neural Networks (GNNs) [38, 43, 52, 54, 59, 61] have revolutionized MRL by leveraging molecule graphs to learn these representations, achieving state-of-the-art results in predicting molecular properties [17] and identifying potential drug candidates [69].

Despite their considerable achievements, they often rely on the fundamental assumption that molecules are independently and identically sampled from a consistent environment. In reality, the ever-changing landscape of real-world scenarios results in environmental changes and distribution shifts [25, 27, 58]. For example, in drug repurposing, molecules initially screened under certain conditions often require reassessment against entirely new diseases or biological targets. However, current GNN-based MRL methods exhibit notable performance degradation [22] in these out-of-distribution (OOD) scenarios, underscoring the pressing demand to enhance their generalization capabilities.

Recent research addressing the OOD challenge of GNNs mainly concentrates on graph invariant learning (GIL) [7, 14, 28, 46], assuming that the causal subgraph is invariant across environments while

environment subgraph varies. By penalizing prediction variance across environments, models capture causal factors rather than spurious correlations. Here, an open problem is effectively determining the environment split. Existing methods have explored various strategies, including predefined splits [3], graph augmentation [57], and additional models specifically for environment split [29].

Nevertheless, applying GIL to GNN-based MRL involves three key considerations: (1) Integration of domain knowledge. Privileged substructures [26, 36] are core functional groups determining molecular activity. For instance, the analgesic properties of Aspirin can be attributed to its ester functional group $-COO-$. This suggests that the invariant subgraph in MRL should be these chemically privileged substructures. But most current methods [35, 46] learn arbitrary subgraphs, overlooking this a priori knowledge. (2) Theoretical guidance for environment split facilitating GIL. Methods [31, 70] using graph augmentation might result in nonsensical molecules, losing the potential to provide insights to domain experts. On the other hand, predefined and learned splits from existing methods [29, 57] are independent of downstream GIL and do not theoretically ensure compatibility with GIL. (3) Unified model for environment split and downstream prediction. Current methods [62] often treat environment split and downstream prediction as a two-stage process, leading to a lack of mutual awareness and suboptimal performance.

To tackle the outlined key points, we propose a novel framework named MILI to advance **M**olecule **I**nvariant **L**earning via privileged substructure **I**dentification. To integrate domain knowledge, we formulate molecule invariant learning based on privileged substructure identification and introduce Substructure Invariance Constraint (SIC). We then theoretically establish two criteria for environment split to guarantee the enhancement of molecule invariant learning: the environments should be split based on the agreement between ground truth and downstream predictions from variant structures, aiming at (1) maximally violating SIC and (2) maintaining class distribution fairness. To fulfill these criteria, we design a dual-head graph neural network. A shared identifier identifies privileged substructures, followed by task and environment heads that make downstream predictions using privileged substructures and variant structures. In line with our criteria, the environments are split and optimized to violate SIC by maximizing invariant risk while enhance class distribution fairness by reweighting empirical risks. Ultimately, this unified framework allows for mutual reinforcement between environment split and molecule invariant learning. Extensive experiments across diverse datasets demonstrate the effectiveness of the proposed MILI.

In summary, our contributions are three-fold:

- We formalize molecule invariant learning based on privileged substructure identification and introduce substructure invariance constraint. Building upon this foundation, we propose criteria for environment split, theoretically ensuring their benefit to molecule invariant learning.
- To meet the criteria, we design a novel dual-head graph neural network with a shared identifier to identify privileged substructures. Subsequently, the interaction between environment and task heads mutually enhances the environment split and molecule invariant learning.
- Comprehensive experiments demonstrate the effectiveness of our MILI. Additionally, case studies of identified privileged substructures reflect its effective utilization of domain knowledge, offering valuable insights for drug design.

2 MOLECULE INVARIANT LEARNING

In this section, we define OOD generalization on molecules and then expand the invariant learning framework based on privileged substructure identification.

OOD Generalization on Molecules. The random variable of a molecule graph is denoted as $\mathbf{G}$, where nodes correspond to atoms and edges represent chemical bonds. Let $\mathcal{G}$ be the molecule graph space and $\mathcal{Y}$ the label space. We consider a dataset $D = \{(G_i, Y_i)\}_{i=1}^{N}$, where $G_i \in \mathcal{G}$ and $Y_i \in \mathcal{Y}$. In real-world applications, the dataset is often sourced from multiple environments $D = \{D^e\}_{e\in\mathcal{E}_{tr}}$. Here, $D^e = \{(G_i^e, Y_i^e)\}_{i=1}^{N_e}$ represents the dataset from environment e, and $\mathcal{E}_{tr}$ denotes the environment space in the training data.

Definition 1. *Let $\mathcal{E}$ represents the space of all possible environments and $\mathcal{H}$ denotes the molecule representation space. Suppose the predictor f can be decomposed into $f = \omega \circ \Phi$, where $\Phi : \mathcal{G} \to \mathcal{H}$ is an encoder mapping molecules into representations, and $\omega : \mathcal{H} \to \mathcal{Y}$ is a classifier that maps representations to the logit space of $\mathcal{Y}$ via a linear map. The goal of OOD generalization on molecules is to find an optimal predictor f^* that performs well across all environments*

$$f^*(\cdot) = \arg\min_{f} \sup_{e\in\mathcal{E}} \mathcal{R}^e(f). \tag{1}$$

Here, $\mathcal{R}^e(f) = \mathbb{E}_{P(\mathbf{G},\mathbf{Y}|e)}[\ell(f(\mathbf{G}), \mathbf{Y})]$ represents the empirical risk on environment e, and $\ell(\cdot,\cdot) : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ is a loss funcion.

The joint distribution of the molecule graph and its corresponding label is denoted as $P(\mathbf{G}, \mathbf{Y}) = P(\mathbf{G}, \mathbf{Y}|e)\ \forall e \in \mathcal{E}$. Distribution shifts refer to the scenario where the joint distribution in the training data $P_{tr}(\mathbf{G}, \mathbf{Y}) = P(\mathbf{G}, \mathbf{Y}|e)\ \forall e \in \mathcal{E}_{tr}$ differs from that in the test data $P_{test}(\mathbf{G}, \mathbf{Y}) = P(\mathbf{G}, \mathbf{Y}|e)\ \forall e \in \mathcal{E}\backslash\mathcal{E}_{tr}$.

Molecule Invariant Learning. Molecule graph $\mathbf{G}$ is characterized by privileged substructures $\{\mathbf{G}^p\}$ determining its properties. This indicates that the relationship between these privileged substructures and the corresponding label is invariant across all environments. The complement of $\{\mathbf{G}^p\}$ is denoted as $\mathbf{G}^v$, representing the structure that varies with environments. Following the invariant learning literature [8], we define the Substructure Invariance Constraint (SIC) for molecule invariant learning.

Definition 2. ***(Substructure Invariance Constraint).*** *Suppose the optimal identifier δ^* is learned to identify privileged substructures within a molecule graph $\mathbf{G}$. Then, the molecule invariant representation $\Phi^*(\delta^*(\mathbf{G}))$ needs to satisfy the following constraint*

$$P(\mathbf{Y}|\Phi^*(\delta^*(\mathbf{G})), e_1) = P(\mathbf{Y}|\Phi^*(\delta^*(\mathbf{G})), e_2),\ \forall e_1, e_2 \in \mathcal{E}. \tag{2}$$

To avoid trivial representations, this constraint is integrated as a regularization term in the training objective. Similar to Invariant Risk Minimization (IRM) [3], set ω as a constant scalar multiplier of 1.0 for each output dimension. The objective function for molecule invariant learning can be written as follows

$$\min_{\Phi,\delta} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(f\circ\delta) + \lambda \left\|\nabla_\omega \mathcal{R}^e(\omega\circ\Phi\circ\delta)\right\|. \tag{3}$$

Obviously, Eq. (3) necessitates predefined environments, the acquisition of which poses a practical challenge. Furthermore, the availability of environment labels does not imply their suitability for molecule invariant learning, leading to no guarantee of benefit [22].

3 CRITERIA FOR ENVIRONMENT SPLIT IN MOLECULE INVARIANT LEARNING

With the above formulation, we aim to derive criteria for environment split that facilitate molecule invariant learning. Intuitively, these environments should shed light on the variations of variant features [32, 48]. Therefore, if the environment split is based solely on the variant structure, it allows for an exposure of any variance. Here, we use identifier $\bar{\delta}(\mathbf{G}) = \mathbf{G}^v$ to represent the complement of the privileged substructure identification.

Suppose an alternative environment predictor $f^e \circ \bar{\delta}$ predicts the label using the variant structure $\mathbf{G}^v$. Contrasting with the molecule invariant representation that adheres to SIC, the results from the environment predictor $f^e(\bar{\delta}(\mathbf{G}))$ intentionally violate SIC.

Theorem 1. *For the optimal environment predictor $f^{e*}(\bar{\delta}^*(\mathbf{G}))$ that relies solely on the variant structure, denote the prediction as $\hat{\mathbf{Y}}^e$ and the ground truth as $\mathbf{Y}$. If the environments are split by*

$$\mathbf{e} = \mathbb{I}(\hat{\mathbf{Y}}^e = \mathbf{Y}), \tag{4}$$

where the function $\mathbb{I}$ determines the equality of two random variables, the substructure invariance constraint will be maximally violated.

Unfortunately, this ideal scenario requires that the environment predictor $f^e \circ \bar{\delta}$ exclusively utilizes the variant structure, a requirement complicated by no prior knowledge of its accurate extraction. In practical implementation, the learned environment split should maximally violate SIC to the fullest degree.

Criterion 1. *The environments $\mathbf{e}$ are split according to the agreement between the ground truth $\mathbf{Y}$ and the predictions $\hat{\mathbf{Y}}^e$ from a learnable environment predictor $f^e(\bar{\delta}(\mathbf{G}))$, i.e., $\mathbf{e} = \mathbb{I}(\hat{\mathbf{Y}}^e = \mathbf{Y})$. This environment split should be optimized to violate the substructure invariant constraint maximally.*

Furthermore, if the environment split $\mathbf{e}$ is determined by the optimal environment predictor $f^{e*} \circ \bar{\delta}^*$ and the ground truth $\mathbf{Y}$, we have molecule invariant representation $\Phi^*(\delta^*(\mathbf{G})) \perp \mathbf{e}|\mathbf{Y}$. Thus, a theorem can be deduced as follows.

Theorem 2. *For the environment split $\mathbf{e}$ determined by the optimal environment predictor $f^{e*} \circ \bar{\delta}^*$ and the ground truth $\mathbf{Y}$, the following equation*

$$\frac{P(\mathbf{Y} = y_1|e_1)}{P(\mathbf{Y} = y_2|e_1)} = \frac{P(\mathbf{Y} = y_1|e_2)}{P(\mathbf{Y} = y_2|e_2)} \tag{5}$$

holds for any $y_1, y_2 \in \mathcal{Y}$ and any $e_1, e_2 \in \mathcal{E}$.

This theorem establishes the relationship between the environment split and class distribution, indicating that the class distribution is fair to the environment split. Intending to reach the optimal scenario, we introduce the second optimization criterion for the environment split.

Criterion 2. *The environments $\mathbf{e}$ are split based on the ground truth $\mathbf{Y}$ and the predictions $\hat{\mathbf{Y}}^e$ from a learnable environment predictor $f^e \circ \bar{\delta}$. This environment split should be optimized to ensure the fairness of class distribution across diverse environments.*

We direct the readers to Appendix A for the proofs of all the above theorems.

4 MILI METHODOLOGY

Guided by the established criteria, we propose MILI, a molecule invariant learning model via privileged substructure identification. In this section, we first present the details of its neural network architecture. Following this, we illustrate the training procedure.

4.1 Dual-head Graph Neural Network

Revisiting the molecule invariant learning framework proposed in Eq. (3), the molecular property predictor consists of two parts $f \circ \delta$. The identifier δ is designated to identify privileged substructures, while f predicts molecular properties. In Sect. 3, the environment split is reliant on the environment predictor $f^e \circ \bar{\delta}$, where $\bar{\delta}$ acts as the complement of privileged substructure identification, and f^e predicts the label from the resultant variant structure. This gives rise to a dual-head graph neural network. Specifically, it utilizes a shared backbone as the identifier δ for privileged substructures, with f and f^e serving as the task and environment heads, respectively. The overall framework is depicted in Fig. 1, and the implementation of each module is introduced as follows.

Molecule Fragmentation. Initially, we fragmentize the molecule $\mathbf{G}$, provided in the SMILES format, into a collection of chemical substructures $\{\mathbf{G}_i^c\}_{i=1}^{N_s}$. This fragmentation is executed by Breaking Retrosynthetically Interesting Chemical Substructures (BRICS) [10], recognized for effectively isolating essential substructures from complex molecules.

Privileged Substructure Identifier. We use the representation of the complete molecule as a query, and substructure representations as keys, identifying privileged substructures based on the attention mechanism [49]. Specifically, a GIN encoder [60] is adopted to learn the representation $\vec{h}$ of the molecule $\mathbf{G}$

$$\vec{h} = \text{GIN}(\mathbf{G}). \tag{6}$$

We apply another GIN subencoder to chemical substructures $\{\mathbf{G}_i^c\}_{i=1}^{N_s}$ to obtain their representations $\{\vec{h}_i^c\}_{i=1}^{N_s}$,

$$\vec{h}_i^c = \text{GIN}(\mathbf{G}_i^c). \tag{7}$$

Treating the molecule representation $\vec{h}$ as a query and substructure representations $\{\vec{h}_i^c\}_{i=1}^{N_s}$ as keys, the attention coefficients $\vec{\alpha}$ can be calculated as

$$\alpha_i = \frac{\vec{h}W^q(\vec{h}_i^c W^k)^\top}{\sqrt{d}}, \tag{8}$$

where matrices $\{W^q, W^k\}$ are learnable linear transformations used to enhance expressive power, and d is the representation dimension. Considering that the task head f requires privileged substructures $\{\mathbf{G}^p\}$ as input, we can directly feed the representation $\vec{h}^p$ of them

$$\vec{h}^p = \text{softmax}(\vec{\alpha})H^c, \tag{9}$$

where H^c represents a matrix stacked with substructure representations $\{\vec{h}_i^c\}_{i=1}^{N_s}$. On the other hand, the environment head f^e necessitates the variant structure $\bar{\delta}(\mathbf{G})$ - the complement of privileged

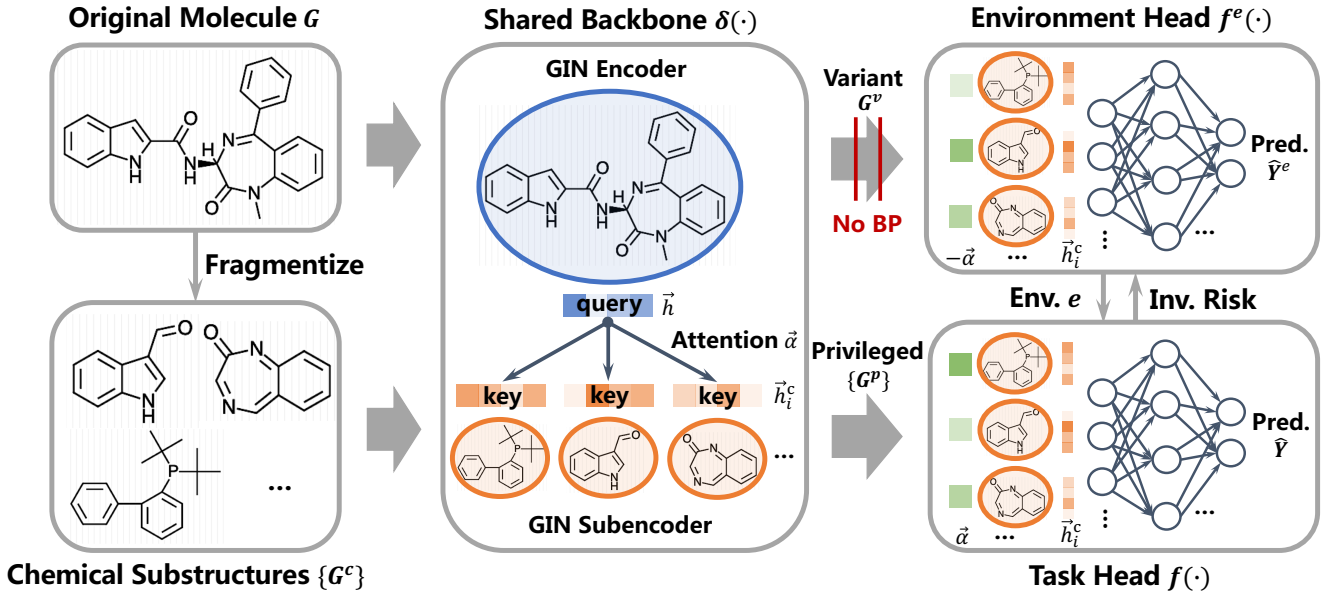


Figure 1: Overall framework of the proposed MILI. Fragmentized chemical substructures $\{G^c\}$ are fed into a dual-head graph neural network. (1) The shared backbone serves as identifier δ to identify privileged substructures. (2) The task head f and environment head f^e respectively utilize privileged substructures $\{G^p\}$ and variant structures G^v for molecular property prediction. Aligning with environment split criteria, the environment head f^e assigns environments e based on the agreement between predictions $\hat{Y}^e$ and ground truth Y. Meanwhile, the task head f considers class distribution fairness across environments and calculates invariant risk $\mathcal{R}_{inv}$ to refine the environment head. In each training iteration, the molecular property predictor $f \circ \delta$ and environment head f^e are optimized with awareness of each other.

substructures. To accomplish this, we utilize reverse attention

$$\vec{h}^v = \text{softmax}(-\vec{\alpha})H^c. \tag{10}$$

Here, $\vec{h}^v$ denotes the representation of the variant structure, serving as the input for the subsequent environment head.

Dual Heads. The task and environment heads employ the representations of privileged substructures and the variant structure for classification, respectively. Particularly, the task head f utilizes a multi-layer perceptron (MLP) to generate predictions $\hat{Y}$,

$$\hat{Y} = \text{softmax}(\text{MLP}(\vec{h}^p)). \tag{11}$$

Likewise, the environment head f^e obtains both a soft prediction $\vec{s}$ and the final prediction $\hat{Y}^e$,

$$\vec{s} = \text{softmax}(\text{MLP}(\vec{h}^v)), \tag{12}$$

$$\hat{Y}^e = \text{argmax}(\vec{s}). \tag{13}$$

According to Theorem 1, the principle of splitting environment **e** is the agreement between the ground truth Y and the predictions $\hat{Y}^e$. Therefore, $e_1 = \mathbb{I}(Y = \hat{Y}^e)$ and $e_2 = \mathbb{I}(Y \neq \hat{Y}^e)$.

Optimization Objective. Given the environment split, we can define per-environment risk $\mathcal{R}^e$,

$$\mathcal{R}^e(f \circ \delta) = \mathbb{E}_{P(G,Y|e)}\, \ell\left(\hat{Y}, Y\right), \tag{14}$$

where ℓ is the cross-entropy for classification. To fulfill Criterion 1 that the environment split maximally violates the substructure invariant constraint, we fix the molecular property predictor $f \circ \delta$ and optimize the environment head f^e by maximizing the invariant risk based on soft assignment

$$\tilde{\mathcal{R}}^e(f \circ \delta, \vec{s}) = \frac{1}{\sum_{i'} \mathbb{I}(e_{i'} = e)} \sum_i \mathbb{I}(e_i = e)\vec{s}_i[\hat{Y}_i^e]\, \ell\left(\hat{Y}_i, Y_i\right), \tag{15}$$

$$\mathcal{L}(f^e) = -\left\|\nabla_\omega \tilde{\mathcal{R}}^e(\omega \circ \Phi \circ \delta, \vec{s})\right\|. \tag{16}$$

Here, $\vec{s}_i[\hat{Y}_i^e]$ represents the element corresponding to the dimension of $\hat{Y}_i^e$. Please note that Eq. (16) does not include empirical risk. To prevent trivial solutions, we stop the backpropagation between the environment head f^e and the identifier δ, refining f^e to concentrate on variant information.

To ensure the fairness of class distribution stated in Criterion 2, the per-environment risk is reweighted based on the class proportion within environments

$$\ddot{\mathcal{R}}^e(f \circ \delta) = \mathbb{E}_{P(G,Y|e)} \frac{P(Y = y)}{P(Y = y|e)}\, \ell\left(\hat{Y} = y, Y = y\right). \tag{17}$$

Aligned with Theorem 2, the ideal value of $P(Y = y)/P(Y = y|e)$ is equal to 1. Practically, it serves to balance class distribution across environments dynamically. By applying increased penalization to larger values, the class distribution is refined through enhanced extraction of privileged substructures. Thus following Eq. (3), the loss function for the molecular property predictor $f \circ \delta$ is

$$\mathcal{L}(f, \delta) = \sum_{e \in \mathcal{E}_{tr}} \ddot{\mathcal{R}}^e(f \circ \delta) + \lambda \left\|\nabla_\omega \mathcal{R}^e(\omega \circ \Phi \circ \delta)\right\|, \tag{18}$$

Table 1: Quantitative OOD generalization performance measured by ROC-AUC (%±σ). The best is marked with boldface and the second best is with underline. (em dash: cannot run without environment labels)

Methods	OGB		IC50			EC50		
	BACE	*BBBP*	*Assay*	*Scaffold*	*Size*	*Assay*	*Scaffold*	*Size*
ERM [41]	78.10±1.30	68.77±0.85	70.87±0.99	68.96±0.26	68.03±1.96	67.87±2.22	66.03±1.18	62.49±1.17
IRM [3]	—	—	71.14±0.85	65.56±0.47	57.74±0.73	69.23±1.63	61.38±0.53	56.84±2.11
GroupDRO [40]	—	—	69.65±0.67	67.67±0.86	57.93±1.27	71.07±4.24	65.67±1.94	60.82±1.21
Mixup [66]	—	—	71.75±1.24	68.96±0.62	66.98±0.38	68.70±1.47	66.48±1.73	63.26±0.51
DIR [57]	76.49±2.80	66.52±1.33	67.16±2.00	64.45±1.39	59.03±1.67	67.07±2.22	63.14±1.64	59.64±1.20
GREA [31]	81.66±0.98	70.76±1.39	71.27±1.04	67.96±0.62	67.10±1.08	73.01±1.09	64.64±1.36	61.42±1.11
GSAT [35]	75.35±1.80	68.38±0.64	70.04±1.15	67.78±0.45	66.37±0.48	71.73±1.76	65.19±0.93	60.22±1.69
CAL [46]	77.29±1.60	68.33±1.27	69.42±1.64	64.64±0.80	64.44±1.51	70.54±2.30	64.96±0.83	60.56±1.26
CIGA [7]	76.29±2.50	68.06±1.37	70.80±1.13	68.37±1.88	66.25±0.47	69.37±1.81	67.53±1.07	65.54±0.67
MoleOOD [62]	77.61±4.90	64.77±2.44	71.60±1.00	67.68±1.12	66.47±1.67	70.77±1.93	65.71±1.45	64.21±1.02
iMoLD [70]	79.11±0.90	68.50±1.33	70.74±1.21	69.22±1.65	67.01±1.37	71.38±1.54	66.50±0.73	65.22±1.25
MILI	**85.16±1.65**	**72.56±0.65**	**72.67±0.52**	**69.58±1.01**	**68.40±0.57**	**77.11±1.37**	**68.07±1.27**	**65.97±0.96**

Algorithm 1 Model Training for MILI

Input: Dataset $D = \{(G_i, Y_i)\}_{i=1}^{N}$; Number of training epochs for environment head T_1; Number of training epochs for molecule property predictor T_2; Number of training iterations T
Output: Trained molecule property predictor $f \circ \delta$

```
Pretrain the molecule property predictor f ∘ δ using the cross-entropy empirical risk;
Initialize the environment head f^e ← f;
for i ← 1 to T do
    Fix the molecule property predictor f ∘ δ;
    Obtain the prediction Ŷ;
    for j ← 1 to T_1 do
        Compute the soft prediction s⃗ and the final prediction Ŷ^e of environment head f^e according to Eqs. (12) and (13);
        Compute the environment split e = 𝕀(Ŷ^e = Y);
        Optimize the environment head f^e according to Eq. (16);
    end for
    Fix the environment head f^e;
    Obtain the prediction Ŷ^e of environment head f^e;
    Obtain the environment split e = 𝕀(Ŷ^e = Y);
    for j ← 1 to T_2 do
        Compute the prediction Ŷ;
        Compute the reweighted risk according to Eq. (17);
        Optimize the molecule property predictor f ∘ δ according to Eq. (18);
    end for
end for
Output the molecule property predictor f ∘ δ;
```

where the first term represents the reweighted empirical risk and the second term is the invariant risk, with λ denoting the trade-off hyperparameter.

4.2 Training Procedure

We adopt an iterative training strategy between the molecule property predictor $f \circ \delta$ and the environment head f^e, allowing mutual benefit and enhancement. Alg. 1 presents the pseudocode.

The computational complexity of MILI primarily stems from iterative optimization. Each iteration mainly involves the updating of the environment head f^e and the molecule property predictor $f \circ \delta$. The complexity for the environment head f^e is $O(T_1 d^2)$, while $O(T_2(|\mathcal{E}|Dd + d^2))$ is for molecule property predictor $f \circ \delta$. Here, $|\mathcal{E}|$ represents the number of edges in the molecule graph, and D denotes the feature dimension. Consequently, the overall complexity of T-iteration MILI is around $T(T_1 d^2 + T_2|\mathcal{E}|Dd)$. Please note that the values of $\{T1, T2\}$ are significantly smaller than those in traditional Empirical Risk Minimization (ERM) training. Moreover, the number of iterations T is relatively modest. Therefore, the added computational overhead remains acceptable.

5 EXPERIMENTS

In this section, we assess the effectiveness of MILI via extensive experiments. Firstly, we compare MILI with state-of-the-art methods in OOD generalization for molecule representation learning. Following this, we present that the identified privileged substructures have substantial chemical significance. Subsequently, we analyze the mechanisms of MILI, validating the contributions of its modules. Lastly, we investigate the hyper-parameter sensitivity.

5.1 Experimental Setup

Datasets. We evaluated the proposed MILI on eight benchmark datasets. BACE and BBBP, from Open Graph Benchmark (OGB) [20], focus on binding affinity against human beta-secretase 1 and brain-blood barrier penetration, respectively. Their train-validation-test splits are determined by scaffold differences. The other six datasets are provided by DrugOOD [22], offering binary classification for

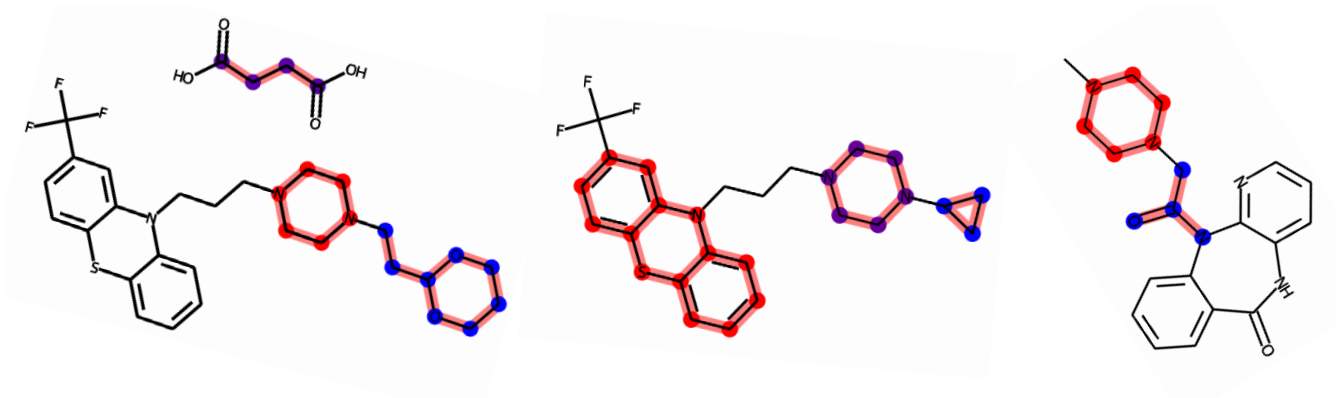


Figure 2: Visualization of three test cases from the BBBP dataset. The identified privileged substructures are highlighted using a color-coded scheme based on their learned weights. Functional group nodes are distinctly marked with the most significant group in red, followed by the second in blue, and the third in purple.

drug target binding affinity prediction. DrugOOD utilizes three split strategies (assay, scaffold, size) across IC50 and EC50 measurements, thus creating six datasets IC50/EC50-assay/scaffold/size. For all datasets, we adhere to the standard train-validation-test split. It is important to note that only the six DrugOOD datasets include manually specified environment labels. As the above property prediction tasks all relate to classification, we report the ROC-AUC score, consistent with previous studies [51, 62, 70]. The statistics of these datasets are summarized in Appendix B.

Baselines. We thoroughly compare our MILI against ERM [41] and three categories of OOD baselines. Specifically, (1) three general OOD methods for Euclidean data, including the invariant learning method IRM [3], the group distributionally robust optimization method GroupDRO [40], and the data augmentation method Mixup [66]. Notably, these methods necessitate manual specification of environments, we limited this comparison to DrugOOD datasets. (2) Two interpretable graph learning methods, DIR [57] and GREA [31]. (3) Five graph OOD methods, namely GSAT [35], CAL [46], CIGA [7], MoleOOD [62], and iMoLD [70]. All methods use GIN [60] backbones and are configured using parameters reported in the original papers or selected via grid search.

Implementation. We implement MILI using the PyTorch deep learning library [1]. For the encoder and subencoder in the identifier, we adopt GIN implementations from the Open Graph Benchmark [20]. As outlined in Alg. 1, during the pretraining of the molecule property predictor using cross-entropy empirical risk, we save the model parameters that exhibit the largest performance gap between the training and validation sets. For hyper-parameter tuning, we employ grid search on the validation set, adjusting the learning rate from $\{1e-2, 5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5\}$, the number of GIN layers from 2 to 6, the dropout rate from $\{0.1, 0.3, 0.5, 0.7\}$, and the trade-off parameter λ in Eq. (18) from $\{0.1, 1, 10, 50, 100, 150, 200, 250\}$. The Adam optimizer [23] is used for efficient gradient-based optimization.

[1] https://pytorch.org/

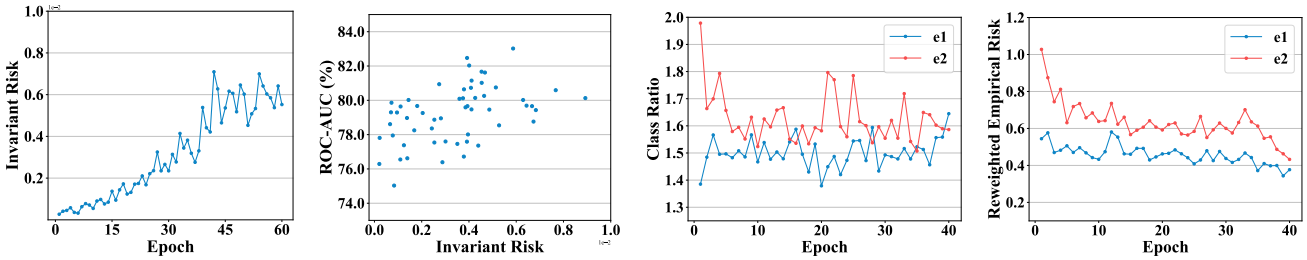
5.2 Main Results

Performance Comparison. In Table 1, we report the mean and standard deviation results over 5 independent trials with different random seeds.

From these results, we draw several conclusions: (1) The proposed MILI consistently outperforms all baselines on the datasets, demonstrating that our unified molecule invariant learning framework substantially enhances the generalizability of molecule representation learning. (2) The performance advantage of MILI over MoleOOD can be attributed to the environment split criteria and the inspired dual-head graph neural network architecture, which ensures that the environment split and molecule invariant learning reinforce each other. (3) The improvements of OOD baselines over ERM highlight the importance of considering the OOD scenario in molecule representation learning. Without specialized design, neural networks are prone to adopting spurious correlations. Furthermore, the superior performance of graph OOD baselines relative to those for Euclidean data underscores the inherent suitability of graph-based learning for capturing intricate patterns in molecules. (4) The distribution shifts of IC50/EC50-scaffold/size datasets have a relatively small impact on ERM models, resulting in limited advantages for OOD methods on these datasets. A large portion of OOD methods fail to surpass ERM, whereas the proposed MILI still achieves the best performance in this scenario.

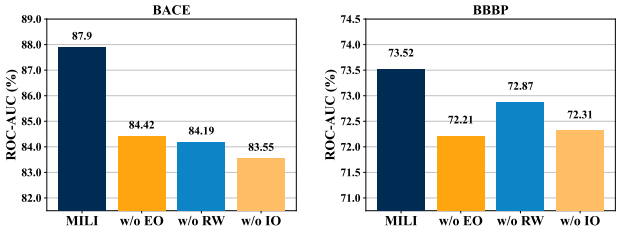
Privileged Substructure Identification. To enhance the understanding of MILI, we detail three cases of identified privileged substructures on the BBBP dataset, as shown in Fig. 2. These substructures are ranked by the learned weights, highlighted in red, blue, and purple, respectively.

Focusing on the brain-blood barrier (BBB) penetration, the BBBP dataset classifies molecules based on their ability to permeate the brain cell membrane. Notably, the most significant functional group $N1CCN(CC1)$ (piperazine), marked in red, is pivotal in aiding molecular penetration through the BBB. Piperazine's presence enhances BBB crossing, beneficial for central nervous system entry [45]. Additionally, the blue-highlighted $CC(=O)N$ (acetamide) moderately supports BBB penetration, leveraging its amide group



(a) Loss changes of environment head (b) Impact of invariant risk on model (c) Changes of class ratio (d) Changes of empirical risk

Figure 3: Optimization analysis on BACE dataset.



(a) BACE (b) BBBP

Figure 4: Impact of environment optimization, empirical risk reweighting, and iterative optimization on (a) BACE and (b) BBBP datasets.

for hydrogen bonding and polarity. Despite this, piperazine structures outperform acetamide in BBB penetration due to their nitrogen-rich ring structure, which receives higher importance in the ranking. These findings underscore the proficiency of MILI in identifying and analyzing substructures crucial for drug efficacy and safety. Meanwhile, its capability to adapt to distribution shifts ensures that the designed drugs are robust and reliable, making it a useful tool in modern pharmaceutical research and development.

5.3 Model Analysis

Ablation Studies. Recall that MILI optimizes the environment head by maximizing invariant risk to satisfy Criterion 1, addresses Criterion 2 by reweighting empirical risks according to class proportions, and establishes mutual reinforcement between molecule invariant learning and environment split via iterative optimization. To ascertain the effectiveness of these modules, we conduct ablation studies. Specifically, we compare MILI with three variants: w/o EO (MILI without environment optimization), w/o RW (with empirical risk reweighting), and w/o IO (without iterative optimization). The results on the BACE and BBBP datasets are depicted in Fig. 4.

From the plots, we observe that MILI consistently outperforms the other variants. Such a phenomenon is not surprising and underscores that the integration of environment optimization, empirical risk reweighting, and iterative optimization is critical for the robust performance of MILI. Each module contributes to the effectiveness of the overall framework.

Optimization Analysis. To elucidate the learning process of MILI, We present invariant risk changes of the environment head and their impact on model performance, the class ratio and the reweighted empirical risk changes within environments on the BACE dataset, as illustrated in Fig. 3.

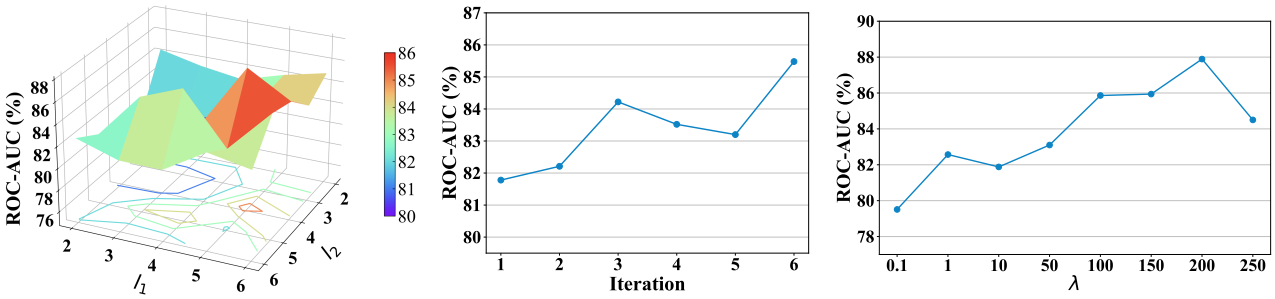
Fig. 3 (a) reveals an upward trend in the invariant risk as the optimization progresses, which aligns with the optimization objective in Eq. (16). Here, the environment head maximizes invariant risk to ensure that the environment split maximally violates SIC and exploits variant information to the utmost. Furthermore, it can be observed that the loss gradually converges, which is consistent with theoretical findings in related work [3, 39] demonstrating the existence of an upper bound on invariant risk. To investigate how the loss of the environment head impacts the model's effectiveness, we vary the number of training epochs for the environment head on the BACE dataset while setting the iteration number to one. We plot the invariant risk and ROC-AUC as a scatter plot in Fig. 3 (b). It shows that the invariant risk and the final model performance are generally positively correlated, indicating that optimizing the environment head more effectively benefits downstream molecule invariant learning.

According to Theorem 2, class ratios between different environments should ideally equalize. Fig. 3 (c) shows that initial disparities in class ratios $P(y_1|\mathbf{e})/P(y_2|\mathbf{e})$ between environments e_1 and e_2 gradually narrow. This indicates that the reweighting in Eq. (17) effectively refines the shared identifier, moving the class ratios closer to the ideal state. Fig. 3 (d) depicts a consistent decline in empirical risks across different environments during optimization, suggesting that the proposed molecule invariant learning framework indeed facilitates downstream tasks.

5.4 Hyper-parameter Sensitivity

In this subsection, we investigate the hyper-parameter sensitivity of MILI, focusing on GIN layers, the number of iterations, and the trade-off hyperparameter λ. Specifically, we adjust the number of GIN layers in the encoder (l_1) and subencoder (l_2) from 2 to 6. As for the number of iterations, the value ranges from 1 to 6. The trade-off hyperparameter λ is explored in $\{0.1, 1, 10, 50, 100, 150, 200, 250\}$. We report the results on the BACE dataset in Fig. 5.

Effect of GIN Layers. As observed in Fig. 5 (a), both the encoder and subencoder exhibit relatively poor performance when the number of GIN layers is too low or too high. We infer that fewer layers may limit the representation capacity to encapsulate the intricate patterns in molecule structures, whereas too many layers risk over-parameterization.



(a) GIN layers (b) Number of iterations (c) Trade-off hyperparameter λ

Figure 5: The hyper-parameter sensitivity on BACE dataset.

Effect of Iterations. Improvements in performance with additional iterations indicate that MILI benefits from iterative optimization. Each iteration might enable environment split and molecule invariant learning to promote each other. Nevertheless, more iterations escalate time complexity. Thus, we should balance the trade-off between performance and complexity.

Effect of Trade-off Hyperparameter. Considering the relatively small value of invariant risk, too small λ causes the optimization to neglect the associated penalty for invariant risk. As depicted in Fig. 5 (c), this may lead the model to learn more spurious correlations, subsequently degrading its performance on the OOD test set. On the other hand, setting λ too high may result in the inadequate optimization of empirical risk, affecting its predictive performance.

6 RELATED WORK

In line with the focus of our work, we briefly review the most related work on molecule representation learning and OOD generalization.

6.1 Molecule Representation Learning

Existing molecule representation learning methods can be classified into three categories. The first is fingerprint-based methods [4, 13, 21], which utilize handcrafted representations [12, 37] to encode topological substructures. While effectively capturing substructural presence, these methods suffer from bit collisions and vector sparsity, limiting their representation capacity. The second is sequence-based methods [18, 21] that leverage SMILES (Simplified Molecular Input Line Entry System) [55] strings to represent molecules. These methods employ sequence-based models such as recurrent neural networks [65] and Transformer [49] to learn molecule representations. However, their main challenge lies in comprehending SMILES syntax. For example, spatially distant atoms may appear adjacent in the sequence. The final category is graph-based methods [38, 43, 59], which model molecules by treating each atom as a node and each chemical bond as an edge. Many works have showcased the profound potential of graph neural networks [17, 44] in analyzing and predicting molecular behavior, significantly advancing the field of molecule representation learning.

Despite their remarkable achievements, these methods predominantly assume that training and testing molecules are independently sampled from an identical environment. However, this assumption often falls short in real-world scenarios, underscoring the urgency for OOD generalization.

6.2 OOD Generalization

The vulnerability of deep neural networks to significant performance drops under distribution shifts has spurred extensive research in OOD generalization [42]. Three lines of methods have emerged for OOD generalization in Euclidean data: group distributionally robust optimization [27, 67], domain adaptation [15, 30], and invariant learning [1, 3, 6]. Group distributionally robust optimization considers groups of distributions and optimizes across all groups. Domain adaptation [11, 34, 47] strives to align data distributions with some additional assumptions [68]. Invariant learning [5, 8] seeks an invariant predictor that upholds invariant relationships across all environments. It does this by learning invariant representations that meet the invariant principle: sufficiency for predictive accuracy and invariance to environmental changes.

However, most existing methods require explicitly multiple environments within the training dataset. This requirement for detailed annotation is not only exceedingly expensive for non-Euclidean data [33, 39] but also inherently problematic due to potential inaccuracies in the predefined split. Furthermore, some studies have indicated that the direct application of these methods to complex molecule graphs [7] frequently fails to yield promising results [22].

6.3 OOD Generalization on Graphs

Recently, there has been a surge of interest regarding OOD generalization on non-Euclidean graphs. Some methods [7, 14, 28, 46] adopt the "first-separation-then-encoding" paradigm to identify invariant substructures in the explicit structural space. Moreover, MoleOOD [62] and GIL [29] utilize inferred environmental labels to learn invariant representations based on the invariant principle; GREA [31] and iMoLD [70] learn disentangled invariant representations within the latent space. OOD generalization on graphs can also be enhanced by related works [35, 57] in the explainability [63, 64] of graph neural networks (GNNs) [16, 19, 24, 50, 56],

which aim to pinpoint a subgraph as the rationale behind a GNN prediction. Although some methods incorporate causality to justify the generated explanations, their primary focus remains on the explainability of GNN predictions rather than OOD generalization.

Methods learning invariant representations in the latent space lack interpretability, while those in the explicit structural space usually use arbitrary subgraphs as basic units. In molecule representation learning, the properties of molecules are frequently determined by chemical substructures [26, 36]. Injecting this prior knowledge is crucial for identifying invariant substructures in molecules and providing new insights to experts. Notably, the most relevant work [62] incorporates this knowledge, whose core idea of environment split is to use variational inference to approximate the posterior $p_\tau(e|G, y)$. Specifically, two GINs are employed to model $q_\kappa(e|G, y)$ and $p_\tau(y|G, e)$, and environment split is achieved by maximizing the ELBO. The separation between environment split and molecule representation learning as independent models leads to a lack of awareness that cannot guarantee mutual benefits.

7 CONCLUSION

In this paper, we formalize molecule invariant learning based on privileged substructure identification and propose substructure invariance constraint. On this foundation, we theoretically derive criteria for environment split and implement them through a dual-head graph neural network. Therefore, our framework ensures mutual enhancement between environment split and molecule invariant learning from theoretical and network design perspectives.

Limitations and Broader Impact. One limitation of MILI is its current focus on classification tasks, presenting an opportunity for future work to broaden its application across more diverse downstream tasks. Our work delves into the OOD problems in molecule representation learning, a prevalent and inevitable scenario in real-world applications. Applying machine learning to molecules still faces numerous practical challenges, such as accurately predicting chemical reactivity. We expect our work will inspire further research combining domain knowledge with machine learning techniques, contributing to the realization of AI4Science.

ACKNOWLEDGMENTS

This work is partly supported by the National Natural Science Foundation of China (No. U20B2045, 61772082, 62192784, 62172052, U1936104) and Young Elite Scientists Sponsorship Program (No. 2023QNRC001) by CAST.

REFERENCES

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. In *NeurIPS*, Vol. 34. 3438–3450.
[2] Bruce Alberts. 2017. *Molecular biology of the cell.* Garland science.
[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).
[4] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* 71 (2015), 58–63.
[5] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *ICML*. 1448–1458.
[6] Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. 2022. When Does Group Invariant Learning Survive Spurious Correlations?. In *NeurIPS*, Vol. 35. 7038–7051.
[7] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. 2022. Learning causally invariant representations for out-of-distribution generalization on graphs. In *NeurIPS*, Vol. 35. 22131–22148.
[8] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *ICML*. 2189–2200.
[9] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* 12, 1 (2020), 1–22.
[10] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. 2008. On the Art of Compiling and Using'Drug-Like'Chemical Fragment Spaces. *ChemMedChem: Chemistry Enabling Drug Discovery* 3, 10 (2008), 1503–1507.
[11] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, Vol. 32.
[12] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* 42, 6 (2002), 1273–1280.
[13] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*, Vol. 28.
[14] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing graph neural networks via learning disentangled causal substructure. In *NeurIPS*, Vol. 35. 24934–24946.
[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research* 17, 59 (2016), 1–35.
[16] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. 2018. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *ICLR*.
[17] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *ICML*. 1263–1272.
[18] Garrett B Goh, Nathan O Hodas, Charles Siegel, and Abhinav Vishnu. 2017. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034* (2017).
[19] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*, Vol. 30.
[20] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, Vol. 33. 22118–22133.
[21] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. 2020. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* 36, 22-23 (2020), 5545–5547.
[22] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. 2023. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *AAAI*, Vol. 37. 8023–8031.
[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[24] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
[25] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery* 3, 11 (2004), 935–949.
[26] Justin Klekota and Frederick P Roth. 2008. Chemical substructures that enrich for biological activity. *Bioinformatics* 24, 21 (2008), 2518–2525.
[27] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*. 5815–5826.
[28] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[29] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. 2022. Learning invariant graph representations for out-of-distribution generalization. In *NeurIPS*, Vol. 35. 11828–11841.
[30] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*. 624–639.
[31] Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. 2022. Graph rationalization with environment-based augmentations. In *KDD*. 1069–1078.
[32] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Heterogeneous risk minimization. In *ICML*. 6804–6814.
[33] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *ICML*. 7313–7324.
[34] Toshihiko Matsuura and Tatsuya Harada. 2020. Domain generalization using a mixture of multiple latent domains. In *AAAI*, Vol. 34. 11749–11756.

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

[36] Chuleeporn Phanus-Umporn, Watshara Shoombuatong, Veda Prachayasittikul, Nuttapat Anuwongcharoen, and Chanin Nantasenamat. 2018. Privileged substructures for anti-sickling activity via cheminformatic analysis. *RSC advances* 8, 11 (2018), 5920–5935.

[37] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

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

[39] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. 2020. The Risks of Invariant Risk Minimization. In *ICLR*.

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally Robust Neural Networks. In *ICLR*.

[41] Stephan R Sain. 1996. The nature of statistical learning theory.

[42] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* (2021).

[43] Hiroyuki Shindo and Yuji Matsumoto. 2019. Gated graph recursive neural networks for molecular property prediction. *arXiv preprint arXiv:1909.00259* (2019).

[44] Zeren Shui and George Karypis. 2020. Heterogeneous molecular graph neural networks for predicting molecule properties. In *ICDM*. 492–500.

[45] Manvi Singh, Reshmi Divakaran, Leela Sarath Kumar Konda, and Rajendra Kristam. 2020. A classification model for blood brain barrier penetration. *Journal of Molecular Graphics and Modelling* 96 (2020), 107516.

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

[47] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*. 443–450.

[48] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *ICCV*. 1417–1427.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, Vol. 30.

[50] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[51] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. 2021. Chemical-Reaction-Aware Molecule Representation Learning. In *ICLR*.

[52] Ruijia Wang, Shuai Mou, Xiao Wang, Wanpeng Xiao, Qi Ju, Chuan Shi, and Xing Xie. 2021. Graph structure estimation neural networks. In *WWW*. 342–353.

[53] Ruijia Wang, Yiwu Sun, Yujie Luo, Shaochuan Li, Cheng Yang, Xingyi Cheng, Hui Li, Chuan Shi, and Le Song. 2024. Injecting Multimodal Information into Rigid Protein Docking via Bi-level Optimization. *NeurIPS* 36.

[54] Xiao Wang, Ruijia Wang, Chuan Shi, Guojie Song, and Qingyong Li. 2020. Multi-component graph convolutional collaborative filtering. In *AAAI*, Vol. 34. 6267–6274.

[55] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28, 1 (1988), 31–36.

[56] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *ICML*. 6861–6871.

[57] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2021. Discovering Invariant Rationales for Graph Neural Networks. In *ICLR*.

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

[59] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* 63, 16 (2019), 8749–8760.

[60] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How Powerful are Graph Neural Networks?. In *ICLR*.

[61] Cheng Yang, Maosong Sun, Wayne Xin Zhao, Zhiyuan Liu, and Edward Y Chang. 2017. A neural network approach to jointly modeling social networks and mobile trajectories. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 1–28.

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

[63] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, Vol. 32.

[64] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 5 (2022), 5782–5799.

[65] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

[67] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. 2022. Correct-N-Contrast: a Contrastive Approach for Improving Robustness to Spurious Correlations. In *ICML*. 26484–26516.

[68] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. 2019. On learning invariant representations for domain adaptation. In *ICML*. 7523–7532.

[69] Tianyi Zhao, Yang Hu, Linda R Valsdottir, Tianyi Zang, and Jiajie Peng. 2021. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in bioinformatics* 22, 2 (2021), 2141–2150.

[70] Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. 2023. Learning Invariant Molecular Representation in Latent Discrete Space. In *NeurIPS*.

Table 2: Statistics of datasets.

	Dataset	# Train	# Validation	# Test	# Total	Split Scheme
OGB	**BACE**	1,210	151	152	1,513	Scaffold
	BBBP	1,631	204	204	2,039	Scaffold
DrugOOD	**IC50-assay**	34,179	19,028	19,028	72,235	Assay
	IC50-scaffold	21,519	19,041	19,048	59,608	Scaffold
	IC50-size	36,597	17,660	16,415	70,672	Size
	EC50-assay	4,540	2,572	2,490	9,602	Assay
	EC50-scaffold	2,570	2,532	2,533	7,635	Scaffold
	EC50-size	4,684	2,313	2,398	9,395	Size

A PROOFS

Theorem. *For the optimal environment predictor $f^{e*}(\bar{\delta}^*(\mathbf{G}))$ that relies solely on the variant structure, denote the prediction as $\hat{\mathbf{Y}}^e$ and the ground truth as $\mathbf{Y}$. If the environments are split by*

$$\mathbf{e} = \mathbb{I}(\hat{\mathbf{Y}}^e = \mathbf{Y}), \tag{19}$$

where the function $\mathbb{I}$ determines the equality of two random variables, the substructure invariance constraint will be maximally violated.

Proof. To measure how well the substructure invariant constraint (SIC) is maintained, we introduce an environment sufficiency gap metric

$$\Delta(f^e \circ \bar{\delta}, \mathbf{e}) = \mathbb{E}\Big[\Big|\mathbb{E}\big[(\mathrm{Y}|f^e(\bar{\delta}(\mathbf{G})), e_1)\big] - \mathbb{E}\big[(\mathrm{Y}|f^e(\bar{\delta}(\mathbf{G})), e_2)\big]\Big|\Big]. \tag{20}$$

By substituting the definition of environment split $\mathbf{e} = \mathbb{I}(\hat{\mathbf{Y}}^e = \mathbf{Y})$, we obtain

$$\Delta(f^e \circ \bar{\delta}, \mathbf{e}) = \mathbb{E}\Big[\Big|\mathbb{E}\big[(\mathrm{Y}|\hat{\mathbf{Y}}^e, \mathbb{I}(\hat{\mathbf{Y}}^e = \mathrm{Y}))\big] - \mathbb{E}\big[(\mathrm{Y}|\hat{\mathbf{Y}}^e, \mathbb{I}(\hat{\mathbf{Y}}^e \neq \mathrm{Y}))\big]\Big|\Big]. \tag{21}$$

Given that the prediction $\hat{\mathbf{Y}}^e$ relies solely on the variant structure, it follows that $\mathbf{Y} \perp \hat{\mathbf{Y}}^e$. Consequently,

$$\Delta(f^e \circ \bar{\delta}, \mathbf{e}) = \mathbb{E}\Big[\Big|\mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e = \mathrm{Y}))\big] - \mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e \neq \mathrm{Y}))\big]\Big|\Big]. \tag{22}$$

Considering the downstream binary classification, we have $\hat{\mathbf{Y}}^e \in \{0, 1\}$. When $\hat{\mathbf{Y}}^e = 0$,

$$\Big|\mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e = \mathrm{Y}))\big] - \mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e \neq \mathrm{Y}))\big]\Big| = |0 - 1| = 1. \tag{23}$$

Similarly, when $\hat{\mathbf{Y}}^e = 1$,

$$\Big|\mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e = \mathrm{Y}))\big] - \mathbb{E}\big[(\mathrm{Y}|\mathbb{I}(\hat{\mathbf{Y}}^e \neq \mathrm{Y}))\big]\Big| = |1 - 0| = 1. \tag{24}$$

Therefore, for each instance, the absolute difference is 1. This leads to an overall environment sufficiency gap $\Delta(f^e \circ \bar{\delta}, \mathbf{e})$ equating to the maximum value 1. □

Theorem. *For the environment split $\mathbf{e}$ determined by the optimal environment predictor $f^{e*} \circ \bar{\delta}^*$ and the ground truth $\mathbf{Y}$, the following equation*

$$\frac{P(\mathrm{Y} = y_1|e_1)}{P(\mathrm{Y} = y_2|e_1)} = \frac{P(\mathrm{Y} = y_1|e_2)}{P(\mathrm{Y} = y_2|e_2)} \tag{25}$$

holds for any $y_1, y_2 \in \mathcal{Y}$ and any $e_1, e_2 \in \mathcal{E}$.

Proof. Considering that the molecule invariant representation $\Phi^*(\delta^*(\mathbf{G}))$ satisfies the substructure invariant constraint (SIC),

$$P(\mathrm{Y} = y_1|\Phi^*(\delta^*(\mathbf{G})), e_1) = P(\mathrm{Y} = y_1|\Phi^*(\delta^*(\mathbf{G})), e_2),\ \forall e_1, e_2 \in \mathcal{E} \tag{26}$$

holds for any $y_1 \in \mathcal{Y}$. Since the environment split $\mathbf{e}$ is determined by the optimal environment predictor $f^{e*} \circ \bar{\delta}^*$ and the ground truth $\mathbf{Y}$, we have $\Phi^*(\delta^*(\mathbf{G})) \perp \mathbf{e}|\mathbf{Y}$. Therefore,

$$P(\Phi^*(\delta^*(\mathbf{G}))|e_1, \mathrm{Y} = y_1) = P(\Phi^*(\delta^*(\mathbf{G}))|e_2, \mathrm{Y} = y_1),\ \forall e_1, e_2 \in \mathcal{E}. \tag{27}$$

Combining Eq. (26), we can obtain

$$\begin{aligned}
\frac{P(\mathrm{Y} = y_1|\Phi^*(\delta^*(\mathbf{G})), e_1)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_1, \mathrm{Y} = y_1)} &= \frac{P(\mathrm{Y} = y_1|\Phi^*(\delta^*(\mathbf{G})), e_2)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_2, \mathrm{Y} = y_1)} \iff \\
\frac{P(\mathrm{Y} = y_1, e_1)}{P(\Phi^*(\delta^*(\mathbf{G})), e_1)} &= \frac{P(\mathrm{Y} = y_1, e_2)}{P(\Phi^*(\delta^*(\mathbf{G})), e_2)} \iff \\
\frac{P(\mathrm{Y} = y_1|e_1)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_1)} &= \frac{P(\mathrm{Y} = y_1|e_2)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_2)} \iff \\
\frac{P(\mathrm{Y} = y_1|e_1)}{P(\mathrm{Y} = y_1|e_2)} &= \frac{P(\Phi^*(\delta^*(\mathbf{G}))|e_1)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_2)}.
\end{aligned} \tag{28}$$

For any $y_2 \in \mathcal{Y}$, a similar conclusion can be drawn

$$\frac{P(\mathrm{Y} = y_2|e_1)}{P(\mathrm{Y} = y_2|e_2)} = \frac{P(\Phi^*(\delta^*(\mathbf{G}))|e_1)}{P(\Phi^*(\delta^*(\mathbf{G}))|e_2)}. \tag{29}$$

Noting that the RHS of Eq. (28) and Eq. (29) are equal, thus

$$\begin{aligned}
\frac{P(\mathrm{Y} = y_1|e_1)}{P(\mathrm{Y} = y_1|e_2)} &= \frac{P(\mathrm{Y} = y_2|e_1)}{P(\mathrm{Y} = y_2|e_2)} \iff \\
\frac{P(\mathrm{Y} = y_1|e_1)}{P(\mathrm{Y} = y_2|e_1)} &= \frac{P(\mathrm{Y} = y_1|e_2)}{P(\mathrm{Y} = y_2|e_2)}.
\end{aligned} \tag{30}$$

□

B DETAILS OF DATASETS

In this work, we leverage eight public benchmarks to evaluate our model. Specifically,

- Open Graph Benchmark (OGB) [20] provides two notable datasets: BACE and BBBP. BACE focuses on binding affinity to human beta-secretase 1, with each molecule labeled according to its binding interaction with this enzyme, a crucial target in Alzheimer's disease research. BBBP assesses Brain-Blood Barrier Penetration, an essential factor for the effectiveness of neuroactive drugs. The labels indicate the ability of molecules to permeate the brain cell membrane and enter the central nervous system.

- The remaining six datasets originate from DrugOOD [22]: IC50-assay, IC50-scaffold, IC50-size, EC50-assay, EC50-scaffold, and EC50-size. The suffixes appended to these dataset names delineate the methodology for their respective train-validation-test splits. These six datasets are primarily concerned with ligand-based affinity prediction, a critical measure in pharmacology, where each molecule is labeled as either active or inactive based on bioassay results.

The statistics of datasets are summarized in Table 2.