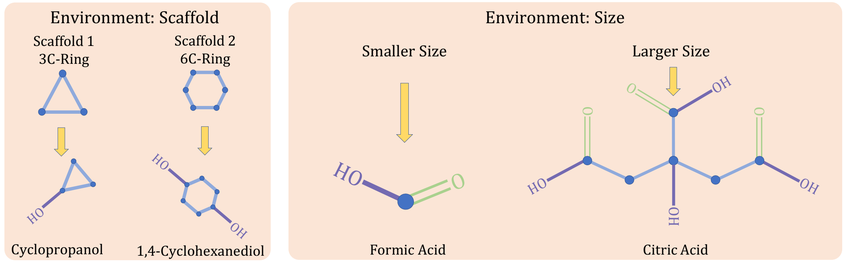
Learning Substructure Invariance for Out-of-Distribution Molecular Representations

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, Junchi Yan

Motivation Examples



Left: the shared substructure hydroxy ***(-OH)*** invariantly contributes to the water solubility of the two molecules which contain different scaffolds, i.e. sampled from different environments by definition.

Right: the water solubility of the two molecules with different sizes can be attributed to the shared substructure carboxy ***(-COOH)*** invariantly, where different sizes are regarded as indicators to define different environments.

The (bio)chemical properties of a molecule are usually associated with a few privileged molecular substructures!

Invariance Principle

Sufficiency:

shows sufficient predictive power for the target

Invariance:

contributes to equal (optimal) performance for the downstream tasks across all environments

Introduction

Existing MRL methods do not differentiate invariant and spurious substructures. Hence, the spurious correlations between irrelevant substructures and the target label will be encoded to learned molecular representations. When tested on unseen environments, the downstream classifier will be easily misled by these spurious correlations.

With the knowledge that (bio)chemical properties of a molecule are usually associated with a few privileged substructures, we aim to suppress such spurious correlations and leverage environment-invariant substructures that more stably relate with the labels across environments to learn invariant molecular representations. Notice that the learned invariant molecular representations should satisfy the invariance principle.

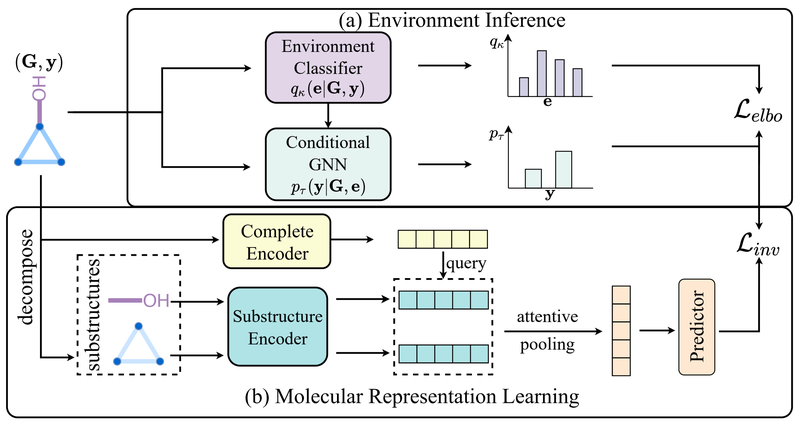
Problem Formulation

A molecular graph can be represented as $G = (V, E)$, where V is the graph's node set corresponding to atoms constituting the molecule and E denotes the graph's edge sets corresponding to chemical bonds. The training and testing molecule graph datasets are denoted as $\mathcal{G}^{train} = \{(G_i, y_i)\}_{i=1}^{N^{train}}$ and $\mathcal{G}^{test} = \{(G_i, y_i)\}_{i=1}^{N^{test}}$. Notice that the test dataset is drawn outside the distribution of the training dataset. The goal of molecule representation learning task is to predict the target label given $\mathbf{y}$ the associated input molecule $\mathbf{G}$.

We can formulate the OOD problem on MRL tasks as:

$$\min_{f} \max_{e \in \mathcal{E}} \mathbb{E}_{(G_i, y_i) \sim p(\mathbf{G}, \mathbf{y} | \mathbf{e}=e)} [l(f(G_i), y_i) | e].$$

Methodology



Overview. The whole training procedure is divided into two stages:

1) Optimize the environment-inference model. Given an input molecule (**G**,**y**), we first infer the latent environment variable **e**. This model is made up of the Environment Classifier and the Conditional GNN. They are responsible for modeling $q_\kappa(\mathbf{e}|\mathbf{G},\mathbf{y})$ and $p_\tau(\mathbf{y}|\mathbf{G},\mathbf{e})$, respectively.

2) Optimize the molecule encoder and the final predictor. We can use any existing MRL method as the Complete Encoder to learn a molecular representation. Meanwhile, we decompose the molecule into substructures and use a simple GNN model as the Substructure Encoder to learn a representation for each substructure. We next use the representation as query and then attentively aggregate substructures' representations to obtain the final representation, which is then fed to the Predictor.

Results

Methods	BACE	BBBP	SIDER	HIV
GCN	80.01 ± 3.49	67.92 ± 1.07	58.90 ± 1.30	76.35 ± 2.01
GCN + virtual node	77.51 ± 3.07	68.19 ± 1.86	60.71 ± 1.34	75.76 ± 2.21
GCN + ours.	**84.33 ± 1.07**	**70.62 ± 0.99**	**63.38 ± 0.67**	**77.73 ± 0.76**
GIN	77.83 ± 3.15	66.93 ± 2.31	59.05 ± 1.47	76.58 ± 1.02
GIN + virtual node	79.64 ± 2.02	66.77 ± 0.95	59.12 ± 0.95	77.11 ± 0.96
GIN + ours.	**81.09 ± 2.03**	**69.84 ± 1.84**	**61.63 ± 1.08**	**78.31 ± 0.24**
GraphSAGE	77.41 ± 1.19	70.58 ± 0.58	58.00 ± 0.95	76.98 ± 1.13
GraphSAGE + virtual node	78.34 ± 2.08	69.29 ± 0.99	59.48 ± 1.37	77.28 ± 1.53
GraphSAGE + ours.	**82.95 ± 0.85**	**71.02 ± 0.75**	**61.09 ± 0.28**	**79.39 ± 0.51**

Performance comparison with baselines on 4 out-of-distribution molecular property prediction datasets from Open Graph Benchmark in terms of ROC-AUC (%), namely, **BACE**, **BBBP**, **SIDER** and **HIV**. The best and the runner-up results are highlighted in **bolded** and underlined respectively. We emphasize the comparison against `∗ + **virtual node**`, a variant of the original method augmented by an additional node connecting to all nodes in the raw graphs.

Dataset	IC50			EC50		
Environment	Assay	Scaffold	Size	Assay	Scaffold	Size
ERM	70.93 ± 2.10	67.31 ± 1.72	67.40 ± 0.56	69.35 ± 7.38	63.92 ± 2.09	60.94 ± 1.95
IRM	70.85 ± 2.41	66.06 ± 1.23	58.46 ± 2.11	69.94 ± 1.03	63.74 ± 2.15	58.30 ± 1.51
DeepCoral	69.82 ± 4.23	66.36 ± 2.57	59.21 ± 2.09	69.42 ± 3.35	63.66 ± 1.87	56.13 ± 1.77
DANN	70.00 ± 1.03	63.61 ± 2.32	65.77 ± 0.47	66.97 ± 7.19	64.33 ± 1.82	61.11 ± 0.64
MixUp	70.22 ± 3.66	66.43 ± 1.08	**67.77 ± 0.23**	70.62 ± 2.12	64.53 ± 1.66	62.67 ± 1.41
GroupDro	69.98 ± 1.74	64.09 ± 2.05	58.46 ± 2.69	70.52 ± 3.38	64.13 ± 1.81	59.06 ± 1.50
Ours.	**71.38 ± 0.68**	**68.02 ± 0.55**	66.51 ± 0.55	**73.25 ± 1.24**	**66.69 ± 0.34**	**65.09 ± 0.90**

Evaluation with other OOD generalization methods on 6 out-of-distribution datasets from DrugOOD in terms of ROC-AUC (%). The best and the runner-up in each columns are highlighted in **bolded** and underlined respectively. Note the baselines except ERM and MixUp all require environment labels. All methods including ours use GIN as backbones.

Our method achieves up to 5.9% and 3.9% improvement over the strongest baselines on OGB and DrugOOD benchmarks, respectively!

Contributions

- To our best knowledge, this is the first work that formulates the OOD problem in MRL background and proposes to leverage the invariance principle which opens a new perspective for handling substructure-aware distribution shifts.

- Under the environment-invariance principle with specific substructure invariance priors, we propose a new learning objective to learn robust representations. In particular, our model does not require environment labels which in fact can be noisy and unreliable, but instead achieve environment inference in an unsupervised manner. This design endows our model with practical applicability for molecular OOD learning where the manual specifications of the environments are often unavailable.

- Results demonstrate that our model yields consistent and significant improvements over various existing MRL methods as backbones and also achieves competitive or even superior prediction compared to state-of-the-art models tailored to OOD learning with environment labels used as extra inputs in both training and testing.