Learning Substructure Invariance for Out-of-Distribution Molecular Representations

Nianzu Yang (杨念祖)

Department of Computer Science and Engineering

Shanghai Jiao Tong University

Background - OoD

- **Out-of-Distribution Generalization:** Assume that there is a potential environment variable $\mathbf{e}$ accounting for the distribution shift between the training and testing data. In general cases the goal is to predict the target label $\mathbf{y}$ given the associated input $\mathbf{x}$.

  Formulation:

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

  $\mathcal{E}$ denotes the support of environments, $f(\cdot)$ is the prediction model and $l(\cdot,\cdot)$ represents a loss function.

  The risk function under a given environment e:

$$\mathcal{R}_e(\mathbf{x}^e, \mathbf{y}^e) = \mathbb{E}_{(x,y)\sim p(\mathbf{x},\mathbf{y}|\mathbf{e}=e)}[l(f(x), y)|e]$$

Background - Invariant Learning

- Invariant Learning is an emerging line for solving the OOD generalization problem.
- These methods propose to find an invariant predictor that could uncover invariant relationships between inputs and targets across all environments.
- The invariant predictor aims to learn an invariant representation satisfying such a invariance principle.

> Invariance Principle:
> 1) sufficiency: shows sufficient predictive power for the target
> 2) invariance: contributes to equal performance for the downstream tasks across all environments

Background - MRL

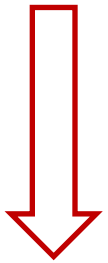
- Molecular Representation Learning (MRL) aims at embedding a molecule into a vector in latent space as a foundation model, on top of which the learned representations could be used for a variety of downstream tasks.
  - SMILES-based methods
  - Structure-based methods

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.

OoD Molecular Represention Learning

- **OOD General Formulation:**

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

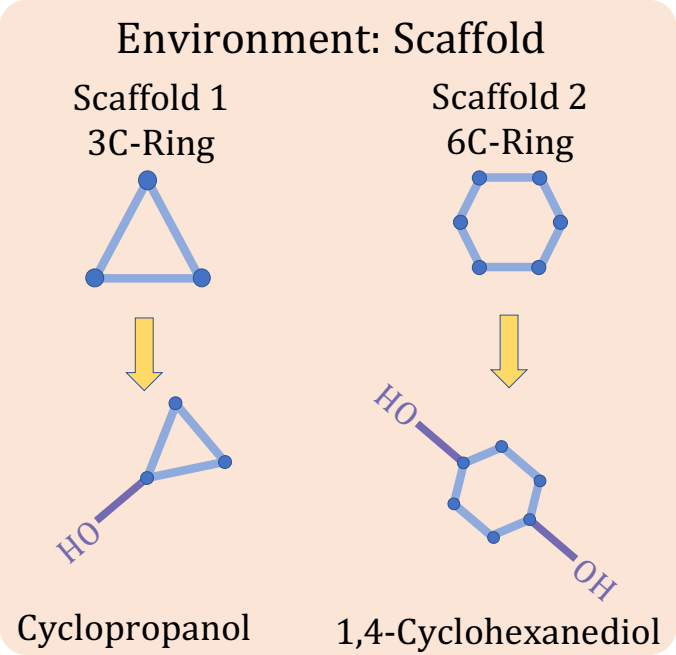


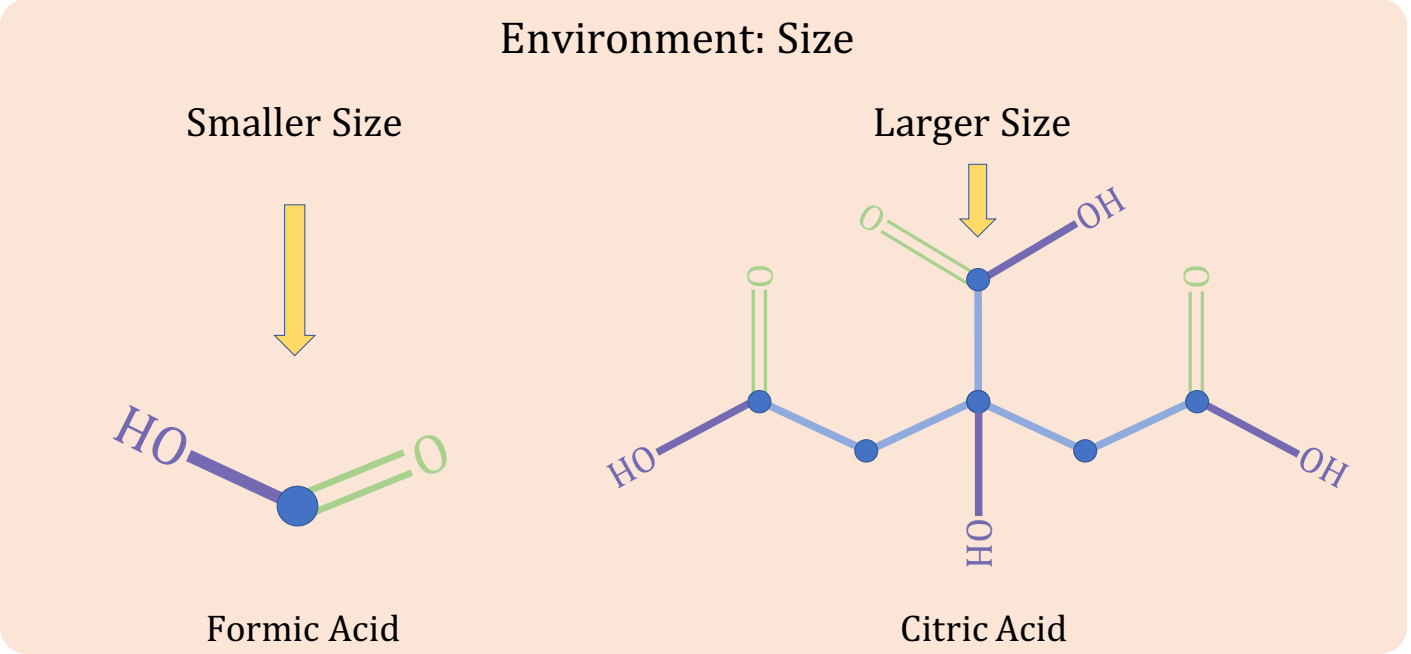
- **OoD on MRL:**

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

Motivating Examples

Key Observation: the (bio)chemical properties of a molecule are usually associated with a few privileged molecular substructures





the shared hydroxy (-OH)/ carboxy (-COOH) ➡ good water solubility

Environment Inference

- **Reasons for necessity**
  - Manual specifications of the environments may be unavailable
    - Labeling is time-consuming
  - Directly utilizing existing environment labels may be problematic
    - There is few molecules per environment on average.

- **A Variational Inference-based method**
  - Introduce a variational distribution $q_\kappa(\mathbf{e}|\mathbf{G},\mathbf{y})$ to approximate $p_\tau(\mathbf{e}|\mathbf{G},\mathbf{y})$
  - The learning objective:

$$\mathcal{L}_{elbo}(\tau,\kappa;\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum_{(G,y)\in\mathcal{G}}[\mathbb{E}_{q_\kappa}[\log p_\tau(y|G,e)] - D_{KL}(q_\kappa(e|G,y)\parallel p(e|G))]$$

Invariant Predictor

❑ Goal:

minimize the expectation of risks from different environments known in the training data:

$$\min_{\omega,\Phi} \mathbb{E}_{\mathbf{e}}[\mathcal{R}_{\mathbf{e}}(\mathbf{G}^{\mathbf{e}}, \mathbf{y}^{\mathbf{e}})], \text{ s.t. } \mathbf{y} \perp\!\!\!\perp \mathbf{e} \mid \Phi(\mathbf{G})$$

Φ : the molecule encoder
ω : the final predictor
$\mathbf{z}$: the denotation of $\Phi(\mathbf{G})$

from the perspective of information theory

$$\max_{\omega,\Phi} \mathrm{I}(\mathbf{z};\mathbf{y}), \text{ s.t. } \min_{\omega,\Phi} \mathrm{I}(\mathbf{y};\mathbf{e}|\mathbf{z})$$

Treating the outputs of ω and Φ as distribution $q_\theta(\mathbf{z}|\mathbf{G})$ and $q_\theta(\mathbf{y}|\mathbf{z})$

$$\max_{q_\theta(\mathbf{y}|\mathbf{z}),q_\theta(\mathbf{z}|\mathbf{G})} \mathrm{I}(\mathbf{z};\mathbf{y}), \text{ s.t. } \min_{q_\theta(\mathbf{y}|\mathbf{z}),q_\theta(\mathbf{z}|\mathbf{G})} \mathrm{I}(\mathbf{y};\mathbf{e}|\mathbf{z})$$

The equivalent tractable objective in practical instantiation:

$$\mathcal{L}_{inv}(\theta;\mathcal{G},\tau) = \frac{1}{|\mathcal{G}|} \sum_{(G,y)\in\mathcal{G}} \left|\log q_\theta(y|G) - \mathbb{E}_{p(\mathbf{e}|\mathbf{G})}[\log p_\tau(y|G,e)]\right| + \beta\mathbb{E}_{\mathbf{e}}\left[\frac{1}{|\mathcal{G}^e|} \sum_{(G,y)\in\mathcal{G}^e} [-\log q_\theta(y|G)]\right]$$

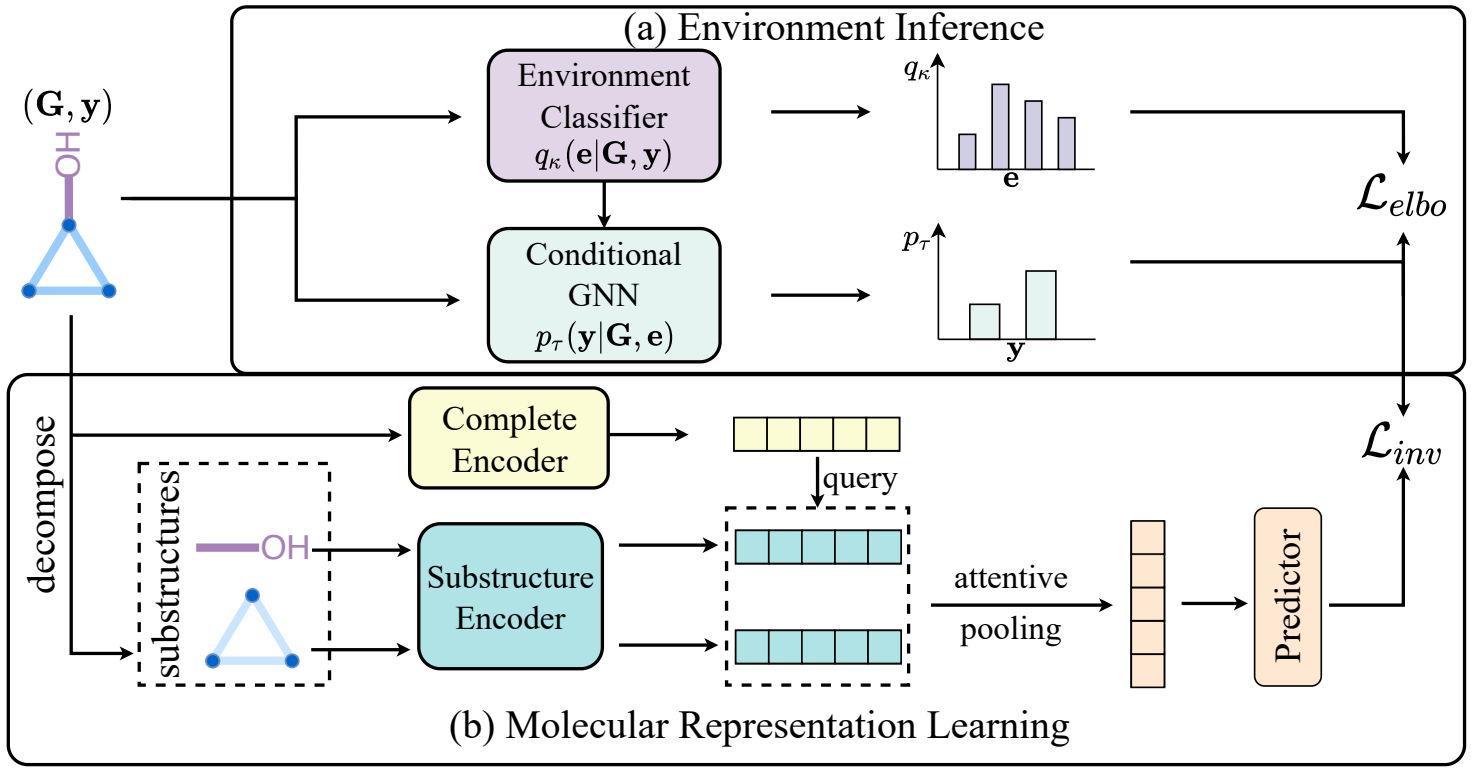
Theoretical Justification

$$\mathcal{L}_{inv}(\theta;\mathcal{G},\tau)=\underbrace{\frac{1}{|\mathcal{G}|}\sum_{(G,y)\in\mathcal{G}}\left|\log q_\theta(y|G)-\mathbb{E}_{p(\mathbf{e}|\mathbf{G})}[\log p_\tau(y|G,e)]\right|}_{①}+\underbrace{\beta\mathbb{E}_{\mathbf{e}}\left[\frac{1}{|\mathcal{G}^e|}\sum_{(G,y)\in\mathcal{G}^e}[-\log q_\theta(y|G)]\right]}_{②}$$

- **Theorem 1**. With $q_\theta(\mathbf{y}|\mathbf{z})$ treated as a variational distribution, minimizing term ① contributes to $\min_{q_\theta(\mathbf{y}|\mathbf{z}),q_\theta(\mathbf{z}|\mathbf{G})} \mathrm{I}(\mathbf{y};\mathbf{e}|\mathbf{z})$, letting show equal performance for the downstream tasks across all environments, i.e. $p(\mathbf{y}|\mathbf{z},\mathbf{e})=p(\mathbf{y}|\mathbf{z})$.

- **Theorem 2.** Regarding $q_\theta(\mathbf{y}|\mathbf{z})$ as a variational distribution, minimizing term ② equals to $\max_{q_\theta(\mathbf{y}|\mathbf{z}),q_\theta(\mathbf{z}|\mathbf{G})} \mathrm{I}(\mathbf{z};\mathbf{y})$, letting $\mathbf{z}$ show sufficient predictive power for downstream tasks.

Overview of MoleOOD



- two-stage training strategy to search for optimal parameters
  - 1) optimizing the environment-inference model: $\kappa^*, \tau^* \leftarrow \arg\max_{\kappa,\tau} \mathcal{L}_{elbo}(\tau, \kappa; \mathcal{G}^{train})$
  - 2) optimizing the molecule encoder and the predictor: $\theta^* \leftarrow \arg\min_{\theta} \mathcal{L}_{inv}(\theta; \mathcal{G}^{train}, \tau)$

Experiments on OGB benchmark

Table. ROC-AUC results on four datasets from OGB benchmark

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**

- MoleOOD achieves consistent significant improvements across four read-world datasets with different backbones (GCN, GIN and GraphSAGE)
  - our method can achieve up to 5.9% improvement

Experiments on DrugOOD benchmark

Table. ROC-AUC results for six datasets from DrugOOD benchmark

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**

- DrugOOD provides more diverse splitting indicators than OGB, including assay, scaffold and size
- Except on IC50-size, our method outperforms all baselines across all datasets
  - our method can achieve up to 3.9% improvement

Ablation Study

Table. Ablation study on EC50-Assay/Scaffold/Size datasets

Method	**Assay**	**Scaffold**	**Size**
ERM (GIN + ERM loss)	69.35 ± 7.38	63.92 ± 2.09	60.94 ± 1.95
MixUp	70.62 ± 2.12	64.53 ± 1.66	62.67 ± 1.41
DANN	66.97 ± 7.19	64.33 ± 1.82	61.11 ± 0.64
Our architecture + ERM loss	71.44 ± 2.02	65.99 ± 0.42	64.23 ± 0.71
GIN + new learning objective	72.07 ± 1.14	66.33 ± 1.38	64.43 ± 1.10
DANN using our inferred environment label	68.83 ± 2.44	64.95 ± 1.07	62.56 ± 1.54
Our model using given environment label	71.94 ± 2.77	66.29 ± 0.85	63.38 ± 1.20
Our full model	**73.25 ± 1.24**	**66.69 ± 0.34**	**65.09 ± 0.90**

We analyze the contributions of different model components to the final performance.

Conclusion

- Proposes to leverage the invariance principle which opens a new perspective for handling substructure-aware distribution shifts.
- Practical applicability for molecular OOD learning where the manual specifications of the environments are often unavailable.
- Extensive experiments on ten public datasets demonstrate our model yields consistent and significant improvements.

Thanks