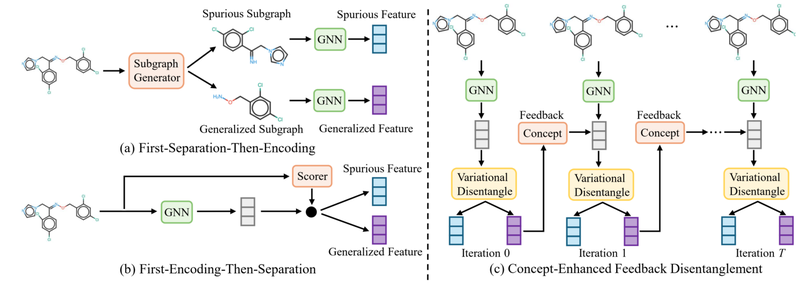
CFD: Learning Generalized Molecular Representation via Concept-Enhanced Feedback Disentanglement

Aming Wu Cheng Deng
Xidian University
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

Molecular Representation Learning

- In order to accelerate the development of biochemical research, e.g., drug discovery, **molecular representation learning (MRL)** has attracted growing attention, **aiming to transform molecules into low-dimensional and dense vectors**
- Though MRL has achieved significant progress, **most methods often follow the closed-set assumption, i.e., the training and testing data share the same distribution**, which limits the applications in open scenarios with unknown diverse distributions. Thus, **improving the generalization of molecular representation is meaningful for reducing the impact of distribution shifts**

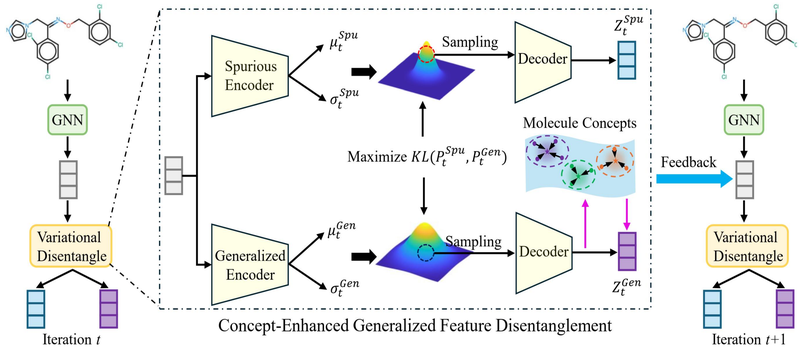
Generalized Representation Extraction



Currently, in order to extract generalized molecular representation, there exist two types of methods:

- The first type is First-Separation-Then-Encoding, i.e., first dividing the graph into generalized and spurious subgraphs and then encoding each part separately
- The second type is First-Encoding-Then-Separation, i.e., first using a GNN to encode the molecule and another GNN is utilized to calculate the score for separating generalized and spurious features
- Differently, we propose a new method, i.e., **Concept-Enhanced Feedback Disentanglement (CFD)**, which aims to **exploit the feedback mechanism to learn generalized representation**

Concept-Enhanced Feedback Disentanglement

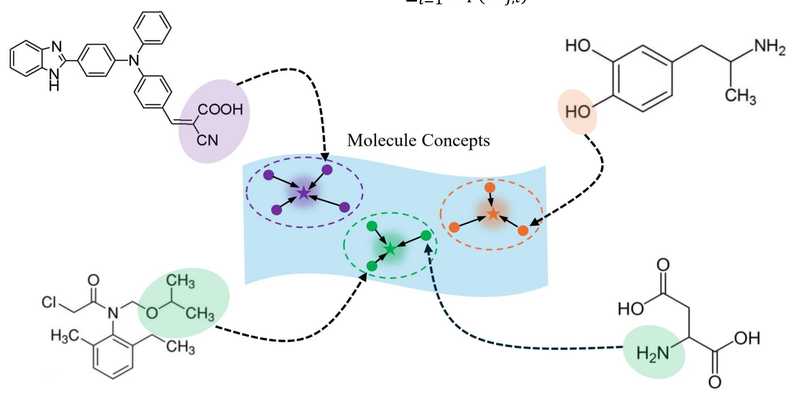


➢ Variational Disentanglement for Molecular Representation

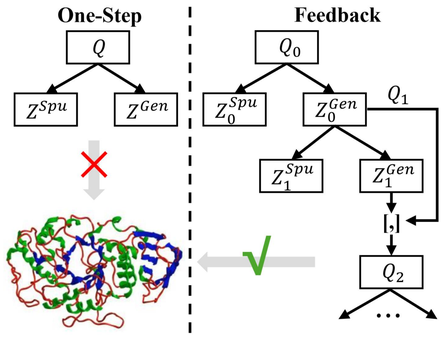
$$Z_t^{Spu} = D^{Spu}(\mu_t^{Spu} + \epsilon \cdot \exp\left(\sigma_t^{Spu}\right)),\ \tilde{Z}_t^{Gen} = D^{Gen}(\mu_t^{Gen} + \epsilon \cdot \exp(\sigma_t^{Gen})),$$

➢ Learning Molecule Concepts

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



Feedback Separating Generalized Representation



- We assume that the number of feedback iterations is T. Given the current feature Q_t that is the concatenation of the previous two iteration outputs, we first use the above operations to disentangle Q_t into Z_t^{Spu} and Z_t^{Gen}. Then, $Q_{t+1} = \emptyset([Z_{t-1}^{Gen}, Z_t^{Gen}])$ is taken as the input of the current step to repeat the above disentangled operations

Experiments

➢ Evaluation Performance on GOOD Benchmark

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**

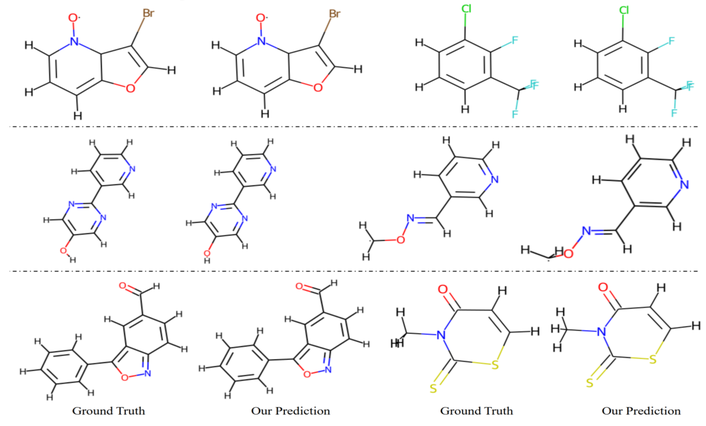
➢ Evaluation Performance on DrugOOD Dataset

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**

➢ Evaluation Performance on DrugOOD Dataset

	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**

➢ Visualization Analysis



Conclusion

Concept-Enhanced Feedback Disentanglement

- By performing multiple feedback iterations, our method progressively decompose expected features involving rich generalized information
- Meanwhile, fusing the molecule concepts that focus on substructures could further strengthen the generalization