SUPPLEMENTARY MATERIAL OF THE UNREASONABLE EFFECTIVENESS OF PRETRAINING IN GRAPH OOD

TABLE OF CONTENTS

A DETAILED DISCUSSION ON FUTURE WORK

A.1 EXPLORATION OF MORE PRE-TRAINING METHODS AND OOD SCENARIOS

Our current work predominantly evaluates representative pre-training and OOD methods/scenarios. However, the field abounds with numerous other methodologies, as summarized in several surveys (Li et al., 2022c; Xia et al., 2022). Due to computational constraints, we could not explore each one exhaustively, leaving a potential avenue for future research.

A.2 DEVELOPMENT OF MODEL SELECTION APPROACHES

Our empirical evaluations, especially those concerning learning rate experiments, lead us to believe that developing pre-trained model selection strategies (e.g., (You et al., 2022)) for OOD generalization is a promising direction for future research.

A.3 Combination of Methods for Enhanced Performance

Future studies could potentially combine pre-trained models with invariant learning or data augmentation techniques to attain improved OOD generalization performance.

A.4 Potential Theoretical Understanding

Based on our current evaluations, there exists an opportunity to explore theoretical connections between graph pre-training and OOD, providing a richer, more in-depth understanding of the empirical performance. One potential direction is exploring some theoretical findings in self-supervised learning and pre-train models (Lee et al., 2021).

B Details on Datasets

B.1 Dataset Statistics

Table 1 summarizes the important key factors and statistics of the molecular datasets. Table A1 and A2 give the full dataset and graph statistics of molecular and general graph datasets used in the paper, respectively.

Table A1: Split statistics of general graph datasets.

Datasets	Domain	Shift	#. Graphs (training/validation/testing)	Avg. #. Node (training/validation/testing)	Avg. #. Edge (training/validation/testing)	#. Classes	Metrics
Motif	Basis	Covariate	18,000/3,000/3,000	17.1/15.8/14.9	48.9/33.0/31.5	3	Accuracy
		Concept	12,600/6,000/6,000	16.9/17.0/17/0	48.5/48.9/48.7		
	Size	Covariate	18,000/3,000/3,000	16.9/39.2/87.2	43.6/107.0/239.6		
		Concept	12,600/6,000/6,000	51.8/51.5/51.6	141.8/140.2/141.5		
CMNIST	Color	Covariate	42,000/7,000/7,000	75.0/75.0/75.0	1392.8/1393.7/1392.6	10	Accuracy
		Concept	29,400/14,000/14,000	75.0/75.0/75.0	1392.8/1393.5/1392.9		

Table A2: Split statistics of molecular datasets.

Datasets	Domain	Shift	#. Graphs (training/validation/testing)	Avg. #. Node (training/validation/testing)	Avg. #. Edge (training/validation/testing)	#. Classes / Task	#. Task	Metrics
DrugOOD	Scaffold	Covariate	21,519/19,041/19,048	39.4/26.8/22.5	85.8/58.4/47.7	2	1	ROC-AUC
	Assay		34,179/19,028/19,032	34.5/30.7/29.7	75.2/66.8/64.7	2		
	Size		36,597/17,660/16,415	38.0/25.6/20.0	82.8/56.0/43.3	2		
BBBP	Scaffold		1,631/204/204	22.5/33.4/27.5	48.4/72.3/59.8	2	1	
Tox21			6,264/783/784	16.5/26.8/26.6	33.7/58.1/57.8	2	12	
ToxCast			6,860/858/858	16.7/26.2/28.2	33.5/56.2/60.8	2	617	
SIDER			1,141/143/143	30.0/43.2/53.3	62.8/91.8/112.7	2	27	
ClinTox			1,181/148/148	25.5/32.6/24.6	54.2/71.0/53.4	2	2	
MUV			74,469/9,309/9,309	24.0/25.3/25.3	51.8/55.6/55.5	2	17	
HIV			32,901/4,113/4,113	25.3/27.8/25.3	54.1/61.1/55.6	2	1	
BACE			1,210/151/152	33.6/37.2/34.8	72.6/81.3/75.1	2	1	
OGBG-MolHIV	Scaffold	Covariate	24,682/4,113/4,108	26.2/24.9/19.8	56.7/54.5/40.6	2	1	
		Concept	15,274/9,382/9,927	24.6/26.5/26.6	53.1/56.9/57.1			
	Size	Covariate	26,169/2,773/3,961	27.8/15.5/12.1	60.1/32.8/24.9			
		Concept	14,483/9,676/10,762	31.3/20.0/19.4	67.7/42.8/41.5			
OGBG-MolPCBA	Scaffold	Covariate	262,764/44,019/43,562	26.9/23.7/20.9	58.2/51.6/44.6	2	128	AP
		Concept	159,158/90,740/119,821	25.5/26.4/26.7	55.2/57.0/57.7			
	Size	Covariate	269,990/48,430/31,925	27.9/19.1/15.0	60.5/40.9/31.5			
		Concept	150,121/108,267/115,205	27.6/24.5/24.4	59.8/53.0/52.6			
NCI1	Size	Covariate	1,942/215/412	20.8/20.7/61.1	44.6/44.6/132.9	2	1	MCC
NCI109			1,872/207/421	20.4/20.3/61.1	43.8/43.6/133.1	2	1	
PROTEINS			511/56/112	15.4/15.7/138.9	57.4/58.5/504.6	2	1	
DD			533/59/118	143.2/156.1/746.4	707.1/746.4/3814.7	2	1	

B.2 Details on Dataset Introduction

DrugOOD (Ji et al., 2023). This benchmark supports AI-driven drug discovery with realistic molecular graph datasets. It automates OOD dataset curation using ChEMBL (Mendez et al., 2019) and offers diverse dataset splitting criteria, including scaffold, assay type and size, for tailored domain alignment. The task focus on drug target binding affinity prediction.

MoleculeNet (Wu et al., 2018). MoleculeNet stands as a comprehensive benchmark for molecular machine learning. It curates diverse public datasets, sets up evaluation standards, and offers open-source tools for different molecular learning methods, all accessible via the DeepChem open source library (Ramsundar et al., 2019).

The benchmark comprises multiple binary graph classification datasets, each designed to evaluate model performance across different facets of molecular interaction. Specifically, BBBP (Martins et al., 2012) evaluates the crucial measure of blood-brain barrier penetration, vital for understanding membrane permeability. Tox21 (Abdelaziz et al., 2016) offers toxicity data encompassing 12 biological targets, including nuclear receptors and stress response pathways. Toxcast (Richard et al., 2016) provides toxicology measurements based on over 600 in vitro high-throughput screenings, serving as a rich resource for understanding toxicity. SIDER (Kuhn et al., 2016) features a database detailing marketed drugs and adverse drug reactions, categorized into 27 system organ classes, offering insights into drug safety. ClinTox (Novick et al., 2013) (AAC) consists of qualitative data classifying drugs approved by the FDA and those that have failed clinical trials due to toxicity concerns. MUV (Gardiner et al., 2011) represents a subset of PubChem BioAssay (Kim et al., 2023), refined through nearest neighbor analysis, and tailored for validating virtual screening techniques. The HIV dataset originates from the Drug Therapeutics Program (DTP) AIDS Antiviral Screen (Riesen & Bunke, 2008), a comprehensive screening effort that evaluated the effectiveness of more than 40,000 compounds in inhibiting HIV replication. BACE (Subramanian et al., 2016) is a dataset that provides qualitative binding results for a collection of inhibitors targeting human β-secretase 1.

OGBG (Hu et al., 2020). OGBG is a specific subset within Open Graph Benchmark (OGB), containing representative datasets like OGBG-Molhiv, OGBG-Molpcba, and OGBG-PPA. OGBG-Molhiv and OGBG-Molpcba challenge graph property prediction with distribution shifts, specifically focusing on predicting molecular properties. They use a scaffold splitting approach, separating structurally distinct molecules into different subsets for a realistic evaluation of graph generalization. The dataset split follows GOOD benchmark (Gui et al., 2022). Specifically, for covariate shift with a distribution source of size, we arranged the molecules in descending order based on the number of nodes and split them into a ratio of $8:1:1$ for the training set, validation set, and testing set, respectively. Similarly, the entire dataset was ordered based on the Bemis-Murcko scaffold string of SMILES, maintaining the same ratio. For concept shift, exemplified by size, we categorized molecules into different groups based on different numbers of molecular nodes. Following this categorization, we selected samples from each group with different labels, forming the training set, validation set, and testing set, respectively, with a ratio of $3:1:1$. This grouping approach aligns with the scaffold-wise distribution, where molecules are categorized based on the Bemis-Murcko scaffold string of SMILES.

TU Datasets. (Morris et al., 2020) It is a collection of benchmark datasets for graph classification and regression. Among these datasets, NCI1, NCI109, PROTEINS, and DD stand out as important and representative graph classification datasets, each offering unique characteristics and complexities. NCI1 and NCI109 datasets are prominent in chemoinformatics. NCI1 is a binary graph classification dataset that focuses on anticancer compound classification. It comprises molecular graphs, with nodes representing atoms and edges indicating chemical bonds. NCI109 extends the challenge by expanding the number of classes and compounds. PROTEINS is a dataset focused on protein graphs, where each node represents a specific protein, and the edges signify various biologically relevant connections or associations between these proteins. The task is to predict the presence or absence of specific protein functions. DD is a real-world graph classification dataset, comprising $1,178$ protein network structures, each of which features 82 distinct node labels. The task is to classify each graph into one of two classes: an enzyme or a non-enzyme.

Motif. Motif is a synthetic dataset (Wu et al., 2022b). It has been created to address structural shifts in graph data. In this dataset, each graph is composed of a base and a motif. The bases are categorized into three distinct types: Tree (S = 0), Ladder (S = 1), and Wheel (S = 2). On the other hand, the motifs include Cycle (C = 0), House (C = 1), and Crane (C = 2), introducing various structural complexities into the dataset. The ground truth label Y for each graph is exclusively dictated by the motif (C). The primary objective in this dataset is to accurately classify the graphs into one of three classes: Cycle, House, or Crane.

CMNIST. CMNIST is a special dataset with graphs showcasing handwritten digits. These graphs are created from the MNIST dataset (Arjovsky et al., 2019) but preprocessed with superpixel (Monti et al., 2017). The goal is to classify each graph into one of the ten-digit categories, from 0 to 9.

C Details on Evaluated Methodologies

Fig. A1 gives the evaluation pipeline on pre-trained GNNs for graph OOD secenarios with a showing case on molecular graphs.

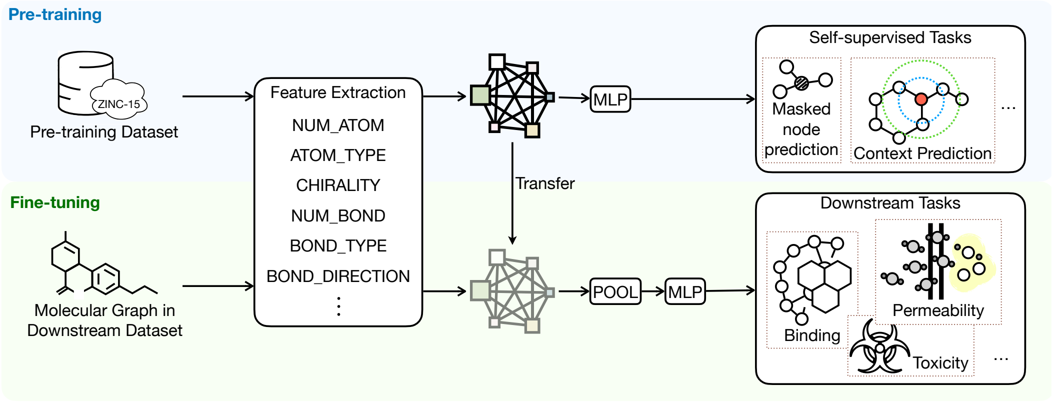

Figure A1: PodGenGraph pipeline for molecular graph pre-training and fine-tuning for downstream datasets.

C.1 Hyperparameter Details for Baseline Methods

CIGA. We used default hyperparameters as specified in the original paper for DrugOOD, TU datasets, Motif, and CMNIST. Specifically, in DrugOOD, the causal substructure size is set to 80% of each graph size for DrugOOD-Scaffold and DrugOOD-Assay, while it's 10% for DrugOOD-Size. The dropout rate is 0.5 for DrugOOD-Scaffold and DrugOOD-Assay, and 0.1 for DrugOOD-Size. For DrugOOD-Assay with CIGA-v1 and CIGA-v2, the coefficient for contrastive loss is set to 8 and 1, respectively. For DrugOOD-Scaffold with CIGA-v1 and CIGA-v2, it's 32 and 16, respectively. For DrugOOD-Size with CIGA-v1 and CIGA-v2, it's 16 and 2, respectively.

For TU datasets, we use a causal substructure size of 60% for NCI1, 70% for NCI109, and 30% for DD and PROTEINS. The coefficient for contrastive loss is 0.5 for NCI1 with CIGA-v1 and 1 for NCI1 with CIGA-v2. It's 2 for both NCI109 and DD with all CIGA versions. For PROTEINS, the coefficient for contrastive loss is 0.5 with both CIGA-v1 and CIGA-v2.

In Motif, the causal substructure ratio is 25%, and in CMNIST, it's 80%. For Motif, the coefficient of contrastive loss is chosen from $\{0.5, 1, 4, 8, 16, 32\}$, and for CMNIST, it's 32 with CIGA-v1 and 16 with CIGA-v2.

For datasets in MoleculeNet and scaffold distribution shift in OGBG datasets, we use hyperparameters similar to those in DrugOOD-Scaffold. For size distribution shift in OGBG datasets, the hyperparameters are aligned with those in DrugOOD-Size.

MoleOOD. We employed default hyperparameters as provided in the code release. Specifically, we selected the prior distribution from uniform, Gaussian distribution for all datasets. In DrugOOD, we utilized 20 domains for the domain prior across three datasets. For MoleculeNet and OGBG datasets, we varied the number of domains among $\{10, 15, 20\}$.

LiSA. We utilized the default hyperparameters provided in the code release. The inner loop was set to 20 for all datasets. We employed 3 subgraph generators and a coefficient loss regularization term of 0.1 across all datasets.

D Full Results

D.1 Results on Different Datasets.

Appendix Table A3-A9 give the full results on the OOD performances of all evaluated methods sperated by datasets.

Table A3: Testing ROC-AUC on Drug-OOD datasets (Ji et al., 2023) with covariate shift. Blue shaded rows indicate pre-training strategies. The first and second best-performing methods (except the ID training) are in **bold** and $\underline{\textbf{bold}}$, respectively.

	DrugOOD-Scaffold	DrugOOD-Assay	DrugOOD-Size	Avg
CIGA-v1	69.27$_{\pm0.81}$	72.36$_{\pm0.60}$	67.08$_{\pm0.82}$	69.57
CIGA-v2	69.68$_{\pm0.21}$	**73.28**$_{\pm0.35}$	68.02$_{\pm0.51}$	$\underline{\mathbf{70.32}}$
MoleOOD	68.01$_{\pm0.39}$	71.18$_{\pm0.63}$	66.61$_{\pm0.36}$	68.60
LiSA	65.71$_{\pm0.25}$	67.66$_{\pm0.63}$	65.78$_{\pm0.46}$	66.38
ContextPred	70.01$_{\pm0.13}$	$\underline{\mathbf{72.80}}_{\pm0.55}$	**68.42**$_{\pm0.10}$	**70.41**
AttrMask	**70.68**$_{\pm0.31}$	71.56$_{\pm0.43}$	$\underline{\mathbf{68.22}}_{\pm0.15}$	70.15
Mole-BERT	$\underline{\mathbf{70.04}}_{\pm0.25}$	71.19$_{\pm0.09}$	67.92$_{\pm0.19}$	69.60
GIN-OOD	67.31$_{\pm0.50}$	71.20$_{\pm0.29}$	66.67$_{\pm0.26}$	68.39
GIN-ID	84.36$_{\pm0.15}$	87.07$_{\pm0.62}$	87.69$_{\pm0.77}$	86.37

Table A4: Testing ROC-AUC on MoleculeNet datasets (Wu et al., 2018) with covariate shift. Blue shaded rows indicate pre-training strategies.

	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	Avg
CIGA-v1	65.50$_{\pm1.62}$	73.87$_{\pm0.54}$	62.81$_{\pm0.55}$	57.40$_{\pm4.40}$	55.00$_{\pm1.60}$	68.10$_{\pm1.30}$	75.79$_{\pm1.09}$	73.60$_{\pm4.30}$	67.75
CIGA-v2	68.69$_{\pm1.37}$	72.25$_{\pm1.46}$	58.53$_{\pm1.85}$	54.90$_{\pm2.13}$	66.37$_{\pm3.22}$	70.99$_{\pm1.34}$	73.19$_{\pm4.22}$	78.56$_{\pm2.34}$	68.16
MoleOOD	$\underline{\mathbf{69.71}}_{\pm1.56}$	73.65$_{\pm0.85}$	62.90$_{\pm0.96}$	$\underline{\mathbf{62.01}}_{\pm0.58}$	**89.93**$_{\pm3.90}$	67.79$_{\pm2.46}$	**78.29**$_{\pm\mathbf{0.51}}$	**81.10**$_{\pm1.97}$	$\underline{\mathbf{73.36}}$
LiSA	65.26$_{\pm2.01}$	66.32$_{\pm0.76}$	59.56$_{\pm0.57}$	57.28$_{\pm0.66}$	65.00$_{\pm2.60}$	67.91$_{\pm1.13}$	62.57$_{\pm1.30}$	69.97$_{\pm3.06}$	64.92
ContextPred	69.32$_{\pm1.03}$	74.47$_{\pm0.36}$	$\underline{\mathbf{63.43}}_{\pm0.40}$	60.45$_{\pm0.60}$	57.40$_{\pm3.16}$	$\underline{\mathbf{77.36}}_{\pm1.11}$	77.56$_{\pm0.95}$	79.41$_{\pm1.96}$	68.38
AttrMask	64.95$_{\pm3.40}$	$\underline{\mathbf{76.22}}_{\pm0.41}$	63.36$_{\pm0.50}$	60.15$_{\pm0.57}$	70.47$_{\pm3.43}$	74.93$_{\pm2.07}$	76.41$_{\pm0.70}$	$\underline{\mathbf{79.88}}_{\pm0.61}$	71.37
Mole-BERT	**71.88**$_{\pm1.12}$	**76.90**$_{\pm0.33}$	**64.18**$_{\pm0.31}$	**62.74**$_{\pm0.89}$	$\underline{\mathbf{78.88}}_{\pm2.24}$	**78.62**$_{\pm\mathbf{1.51}}$	$\underline{\mathbf{78.10}}_{\pm0.65}$	80.88$_{\pm1.45}$	**74.62**
GIN-OOD	65.78$_{\pm4.90}$	73.95$_{\pm0.28}$	62.13$_{\pm0.71}$	57.38$_{\pm1.65}$	57.29$_{\pm5.91}$	70.40$_{\pm1.80}$	75.06$_{\pm2.06}$	70.78$_{\pm5.29}$	66.70
GIN-ID	93.13$_{\pm0.58}$	82.60$_{\pm0.20}$	70.93$_{\pm0.28}$	62.57$_{\pm0.81}$	84.91$_{\pm2.10}$	79.49$_{\pm1.44}$	80.86$_{\pm1.11}$	86.73$_{\pm1.72}$	80.55

Table A5: Performance evaluation on OGBG datasets (Hu et al., 2020) with covariate shift. OGBG-MolPCBA is evaluated by AP, while OGBG-MolHIV is evaluated by ROC-AUC. Blue shaded rows indicate pre-training strategies. The first and second best-performing methods (except the ID training) are in **bold** and $\underline{\textbf{bold}}$, respectively.

	OGBG-MolPCBA		OGBG-MolHIV	
	Size	Scafflod	Size	Scafflod
CIGA-v1	10.51$_{\pm0.17}$	10.24$_{\pm1.98}$	61.81$_{\pm1.68}$	69.40$_{\pm2.39}$
CIGA-v2	9.65$_{\pm0.12}$	10.62$_{\pm1.04}$	59.55$_{\pm2.56}$	69.40$_{\pm1.97}$
LiSA	6.52$_{\pm0.20}$	8.67$_{\pm0.24}$	59.65$_{\pm1.44}$	68.92$_{\pm0.92}$
ContextPred	13.30$_{\pm0.37}$	**22.14**$_{\pm0.43}$	60.47$_{\pm0.88}$	**70.69**$_{\pm1.12}$
AttrMask	$\underline{\mathbf{13.50}}_{\pm0.38}$	$\underline{\mathbf{21.89}}_{\pm0.27}$	$\underline{\mathbf{62.29}}_{\pm0.91}$	$\underline{\mathbf{70.29}}_{\pm1.57}$
Mole-BERT	**16.19**$_{\pm0.24}$	17.33$_{\pm0.12}$	**66.95**$_{\pm0.93}$	69.63$_{\pm0.96}$
GIN-OOD	12.85$_{\pm0.34}$	13.03$_{\pm0.43}$	60.06$_{\pm1.63}$	65.41$_{\pm1.70}$
GIN-ID	28.10$_{\pm0.69}$	30.80$_{\pm0.54}$	79.49$_{\pm0.55}$	80.86$_{\pm1.11}$

Table A6: Testing Matthews correlation coefficient on TU datasets with covariate shift. Blue shaded rows indicate pre-training strategies. The first and second best-performing numbers (except the ID training) are in **bold** and **bold**, respectively.

	NCI1	NCI109	PROTEINS	DD
CIGA-v1	$0.22_{\pm 0.07}$	$0.23_{\pm 0.09}$	$0.40_{\pm 0.06}$	$0.29_{\pm 0.08}$
CIGA-v2	$\mathbf{0.27}_{\pm 0.07}$	$0.22_{\pm 0.05}$	$0.31_{\pm 0.12}$	$\mathbf{0.26}_{\pm 0.08}$
LiSA	$0.24_{\pm 0.01}$	$0.26_{\pm 0.02}$	$\mathbf{0.43}_{\pm 0.05}$	$\mathbf{0.37}_{\pm 0.07}$
InfoGraph	$\mathbf{0.39}_{\pm 0.01}$	$\mathbf{0.38}_{\pm 0.01}$	$\mathbf{0.53}_{\pm 0.07}$	$\mathbf{0.35}_{\pm 0.04}$
GIN-OOD	$0.21_{\pm 0.06}$	$0.16_{\pm 0.05}$	0.23 ± 0.05	$0.25_{\pm 0.09}$
GIN-ID	$0.45_{\pm 0.03}$	$0.44_{\pm 0.02}$	$0.46_{\pm 0.03}$	$0.40_{\pm 0.04}$

Table A7: Testing accuracy on general graph datasets with covariate shift. Blue shaded rows indicate pre-training strategies. The first and second best-performing numbers (except the ID training) are in **bold** and **bold**, respectively.

	Motif		CMNIST
	Basis	Size	
CIGA-v1	$66.43_{\pm 11.31}$	$49.14_{\pm 8.34}$	$\mathbf{32.22}_{\pm 2.67}$
CIGA-v2	$67.15_{\pm 8.19}$	$\mathbf{54.42}_{\pm 3.11}$	$32.11_{\pm 2.53}$
LiSA	$\mathbf{82.55}_{\pm 7.18}$	$\mathbf{62.90}_{\pm 8.30}$	$\mathbf{33.21}_{\pm 13.43}$
InfoGraph	$\mathbf{86.85}_{\pm 2.43}$	$53.43_{\pm 8.09}$	$24.39_{\pm 2.09}$
GIN-OOD	$62.01_{\pm 3.92}$	$52.94_{\pm 2.93}$	$26.28_{\pm 5.95}$
GIN-ID	$92.15_{\pm 0.04}$	$92.16_{\pm 0.07}$	$77.80_{\pm 0.20}$

D.2 Different Statistical Metrics

Appendix Fig. A2-A3 show the additional statistical evaluation on the performances of all approaches on Drug-OOD and Molecule-Net datasets. The metrics include median, IQM, mean, and the optimality gap. Results also reveal that the pre-trained models achieve well-performance results compared with baseline approaches.

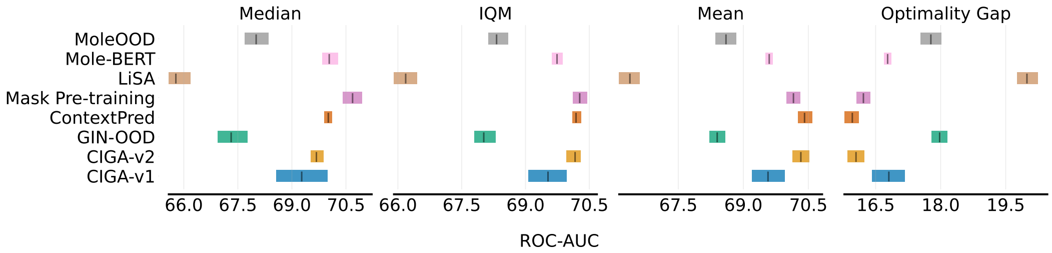

Figure A2: Aggregate performance on DrugOOD averaged across three datasets: `DrugOOD-lbap-core-ic50-assay`, `DrugOOD-lbap-core-ic50-scaffold`, and `DrugOOD-lbap-core-ic50-size`. Better results are indicated by higher mean, median, and IQM scores, along with a lower optimality gap.

Table A8: Performance evaluation on OGBG datasets (Hu et al., 2020) with concept shift. OGBG-MolPCBA is evaluated by AP, while OGBG-MolHIV is evaluated by ROC-AUC. Blue shaded rows indicate pre-training strategies. The first and second best-performing numbers (except the ID training) are in **bold** and **bold**, respectively.

	OGBG-MolPCBA		OGBG-HIV	
	Size	Scafflod	Size	Scafflod
CIGA-v1	$9.22_{\pm0.09}$	$8.33_{\pm0.06}$	$72.80_{\pm1.35}$	$70.79_{\pm1.55}$
CIGA-v2	$8.31_{\pm0.12}$	$8.71_{\pm0.12}$	$\underline{\mathbf{73.62}}_{\pm1.33}$	$\underline{\mathbf{71.65}}_{\pm1.33}$
LiSA	5.05 ± 0.32	$8.55_{\pm0.63}$	$72.36_{\pm4.75}$	$69.46_{\pm0.83}$
ContextPred	$11.39_{\pm0.21}$	$15.71_{\pm0.38}$	$70.41_{\pm0.38}$	$68.77_{\pm0.90}$
AttrMask	$\underline{\mathbf{11.87}}_{\pm0.24}$	$\underline{\mathbf{16.14}}_{\pm0.49}$	$70.59_{\pm0.58}$	$71.50_{\pm0.55}$
Mole-BERT	$\mathbf{15.71}_{\pm0.26}$	$\mathbf{21.29}_{\pm0.53}$	$\mathbf{75.94}_{\pm0.91}$	$\mathbf{76.13}_{\pm0.39}$
GIN-OOD	$12.76_{\pm0.62}$	$17.27_{\pm0.63}$	$70.20_{\pm1.12}$	$62.36_{\pm2.20}$
GIN-ID	$28.10_{\pm0.69}$	$30.80_{\pm0.54}$	$79.49_{\pm0.55}$	$80.86_{\pm1.11}$

Table A9: Testing accuracy on general graph datasets with concept shift. Blue shaded rows indicate pre-training strategies. The first and second best-performing numbers (except the ID training) are in **bold** and **bold**, respectively.

	Motif		CMNIST
	basis	size	
CIGA-v1	$72.50_{\pm4.02}$	$58.63_{\pm6.66}$	$34.80_{\pm3.33}$
CIGA-v2	$77.48_{\pm2.54}$	$\mathbf{70.65}_{\pm4.81}$	$\mathbf{39.39}_{\pm3.30}$
LiSA	$\mathbf{87.89}_{\pm1.61}$	$\underline{\mathbf{70.36}}_{\pm2.61}$	$\underline{\mathbf{36.56}}_{\pm0.40}$
InfoGraph	$\underline{\mathbf{79.36}}_{\pm1.12}$	$64.79_{\pm1.68}$	$19.19_{\pm2.17}$
GIN-OOD	$72.12_{\pm1.89}$	$58.23_{\pm1.73}$	$29.53_{\pm0.50}$
GIN-ID	$92.15_{\pm0.04}$	$92.16_{\pm0.07}$	$77.80_{\pm0.20}$

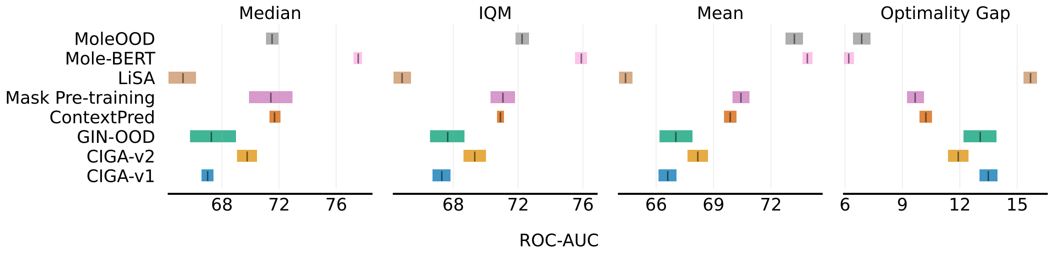

Figure A3: Aggregate performance on MoleculeNet averaged across eight datasets: BBBP, Tox21, ToxCast, SIDER, ClinTox, MUV, HIV, BACE. Better results are indicated by higher mean, median, and IQM scores, along with a lower optimality gap.

D.3 Different Backbones

Appendix Fig. A4-A7 show the performance on molecular prediction with different GNN architectures (GIN and GAT).

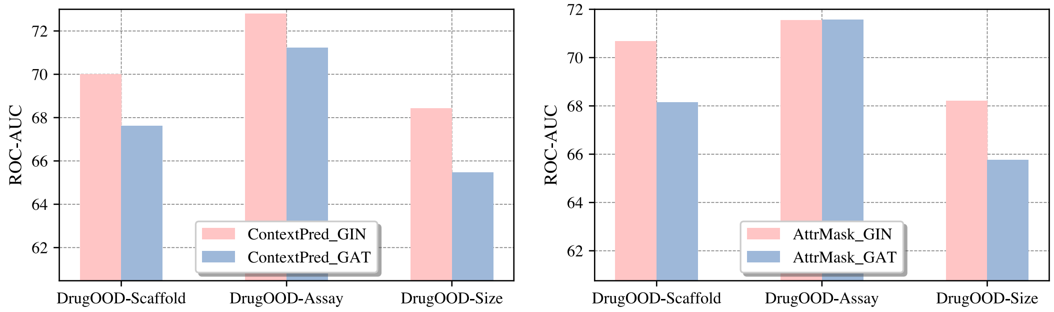

Figure A4: Comparison of ROC-AUC performance (%) on the DrugOOD dataset using the GIN and GAT backbones, respectively.

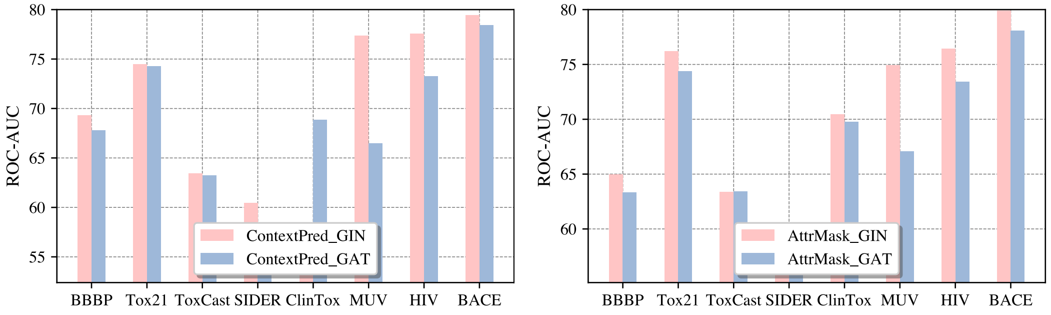

Figure A5: Comparison of ROC-AUC performance (%) on the MoleculeNet dataset using the GIN and GAT backbones, respectively.

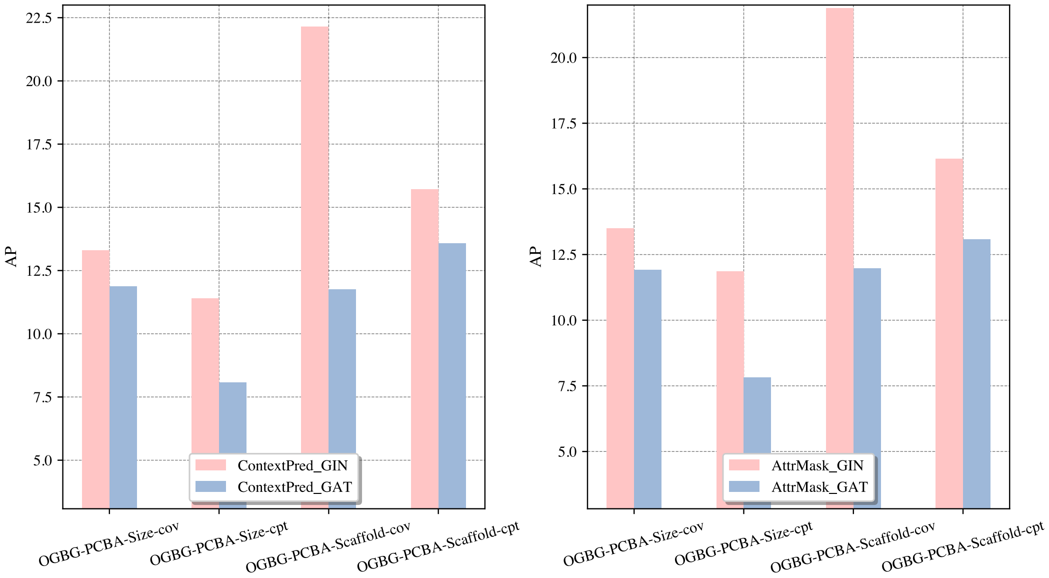

Figure A6: Comparison of AP on the OGBG-PCBA dataset using the GIN and GAT backbones, respectively.

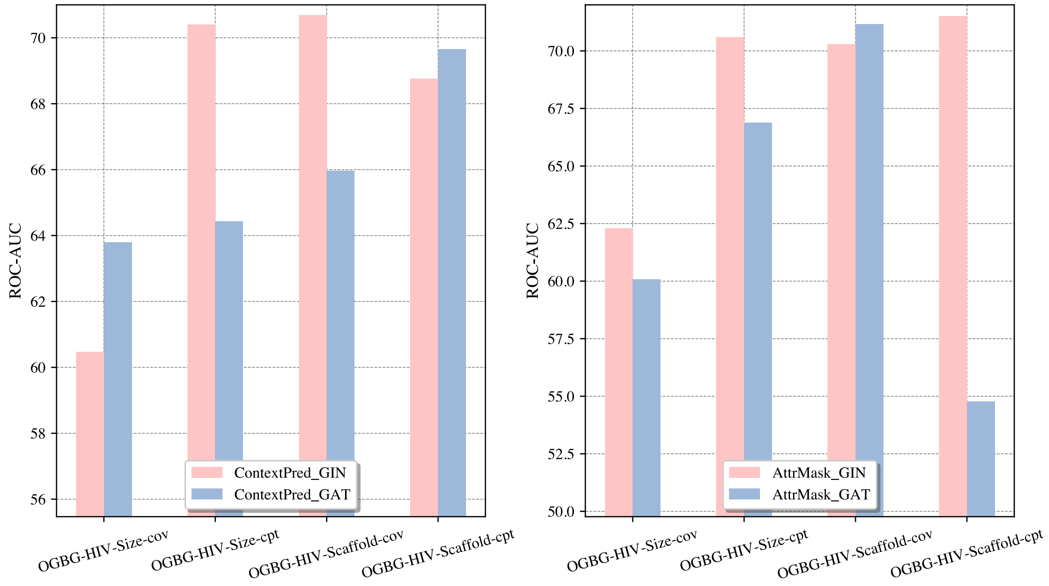

Figure A7: Comparison of ROC-AUC performance (%) on the OGBG-HIV dataset using the GIN and GAT backbones, respectively.

E Reproducibilty Statement

E.1 Details

The experiments are implemented on an 8 Intel Xeon Gold 5220R and 4 NVidia A100 GPUs. We use the publicly accessible code libraries of all evaluated methods. The detailed implementation can be found through this anonymous link: https://sites.google.com/view/podgengraph/.

E.2 Used Libraries and Licenses

In our implementation, we have used the following libraries which are covered by the corresponding licenses:

- Tensorflow (Apache License 2.0)
- Pytorch (BSD 3-Clause ”New” or ”Revised” License)
- Numpy (BSD 3-Clause ”New” or ”Revised” License)
- RDKit (BSD 3-Clause ”New” or ”Revised” License)
- scikit-image (BSD 3-Clause ”New” or ”Revised” License)
- wilds (MIT License)
- Codebase of CIGA: link, (MIT license)
- Mole-OOD: link, (MIT license)
- Codebase of LiSA: link
- Codebase of Mask pretraining and context prediction: link, (MIT Liecense)
- Codebase of InfoGraph: link
- Codebase of Molecule-BERT: link