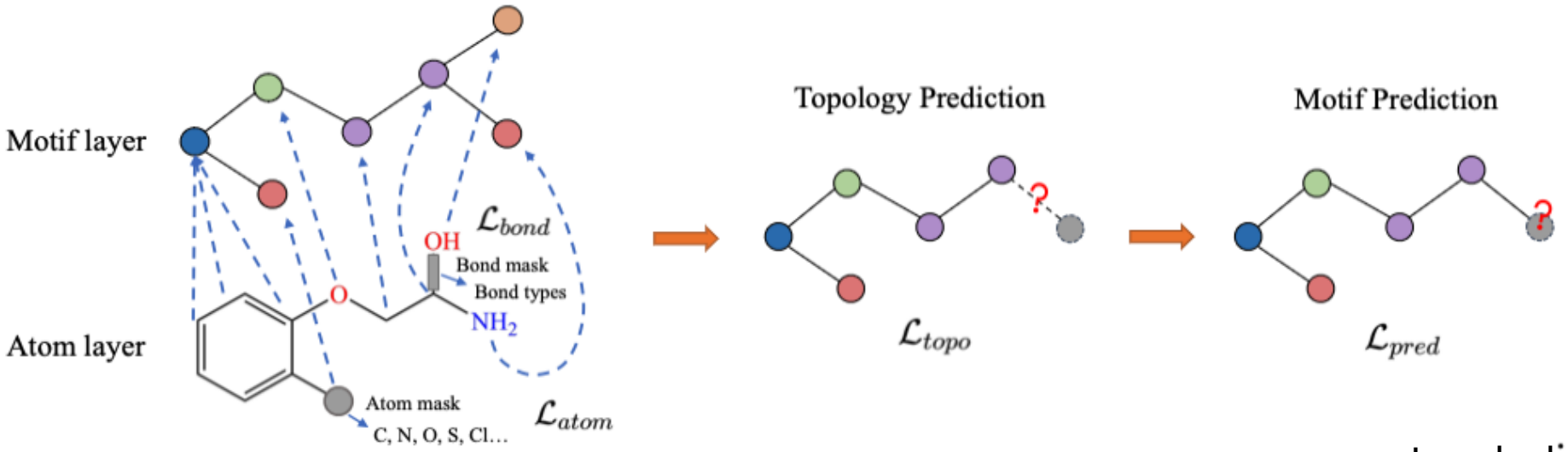
# Motif-based graph self-supervised learning for molecular property prediction

NIPS 2021




Lee deok joong

# Contents

# 01

# Introduction

# 01 Introduction

- Predicting molecular properties with data-driven methods has drawn much attention in recent years.
- Recently, some works applied Graph Neural Network (GNN) and some of its variants for molecular property prediction and obtained promising results
- GNNs can be pre-trained on unlabeled molecular data to first learn the general semantic and structural information before being finetuned for specific tasks.
- Similar issues have also been encountered in natural language processing and computer vision. Recent advances in NLP and CV address them by self-supervised learning (SSL) where a model is first pre-trained on a large unlabeled dataset and then transferred to downstream tasks with limited labels

# 01 Introduction

- However, we argue that existing self-supervised learning tasks on GNNs are sub-optimal since most of them fail to exploit the rich semantic information from graph motifs. Graph motifs can be defined as significant subgraph patterns that frequently occur
- We propose Motifbased Graph Self-supervised Learning (MGSSL) by introducing a novel self supervised motif generation framework for GNNs.

# 01 Introduction

- Firstly, MGSSL introduces a novel motif generation task that empowers GNNs to capture the rich structural and semantic information from graph motifs.
- Secondly, a general motif-based generative pre-training framework is designed to generate molecular graphs motif-by-motif.
- We introduce Multi-level self-supervised pre-training for GNNs where the weights of different SSL tasks are adaptively adjusted by the Frank-Wolfe algorithm.

# 01 Introduction

❖ ***정리하자면..***

1. 최근 Molecular properties prediction에 deep learning같은 data driven methods들이 많이 주목받고 있고, 특히 GNN이 각광받고 있다.
2. 하지만 Molecular data를 labeling하는건 정말 어렵기 때문에, unlabeled data를 잘 다루는 Self-Supervised Learning방법론들이 연구되고 있음
3. 그럼에도 기존 Graph-based self-supervised manner들은 sub-optimal을 보여주는데, 이는 graph motif로 부터 rich한 sementic information을 capture하는데 어렵기 때문임
4. 그래서 우리는 graph motif를 잘 다룰 수 있는 self-supervised method를 제안함 [MGSSL]

   ➔ 3가지 모듈로 구성

   1) Chemistry-inspired Moleucule Fragmentation : BRICS 알고리즘에 기반해서 motif tree construction

   2) Motif Generation : DFS, BFS기반으로 graph motif distribution을 학습 [auto regression manner]

   3) Multi-level Self-supervised Pre-training

# 02

# Preliminaries

# 02 Preliminaries

## ❖ Molecular Property Prediction

- Prediction of molecular properties is a central research topic in physics, chemistry, and materials science [44]. Among the traditional methods, $density\ functional\ theory\ (DFT)$ is the most popular one and plays a vital role in advancing the field [21]. However, $DFT$ is very time-consuming and its complexity could be approximated as $O(N^3)$, where N denotes the number of particles.

➡ 어떤 분자가 존재할 수 있는지 그리고, 특정 분자의 모양, 성질 예측을 가능하게함

# 02 Preliminaries

❖ Preliminaries of Graph Neural Networks

$G = (V, E)$ $X_v\ (for\ v \in V) : node\ features$ $e_{uv}\ (for\ (u, v) \in E) : edge\ features$

- For example, in molecular property prediction, $G$ is a molecular graph, where nodes represent atoms and edges represent chemical bonds, and the label to be predicted can be toxicity or enzyme binding

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \mathcal{N}(v)\right\}\right)\right)$$

$$h_G = \text{READOUT}\left(\left\{h_v^{(K)} \mid v \in G\right\}\right).$$

# 02 Preliminaries

## ❖ Self-supervised Learning of Graphs

- Graph Self-supervised learning aims to learn the intermediate representations of unlabeled graph data that are useful for unknown downstream tasks.

1. Traditional graph embedding methods
2. Constrative models : multi view
3. Predictive models : node, edge attributes masking

# 03

## Motif-based Graph Self-supervised Learning

1. Chemistry-inspired Molecule Fragmentation
2. Motif Generation
3. Multi-level Self-supervised Pre-training

# 03 Motif-based Graph Self-supervised Learning

## ❖ Chemistry-inspired Molecule Fragmentation

- Given a dataset of molecules, the first step of our method is to decompose molecules into several fragments/motifs.
- There are many ways to fragment a given graph while the designed molecule fragmentation method should achieve the following goals:

  1) In a motif tree $T(G)$, the union of all motifs $M_i$ should equals $G$.

  2) In a motif tree $T(G)$, the motifs should have no intersections. That is $M_i \cap M_j = \emptyset$.

  3) The induced motifs should capture semantic meanings, e.g., similar to meaningful functional groups in the chemistry domain.

  4) The occurrence of motifs in the dataset should be frequent enough so that the pre-trained GNNs can learn semantic information of motifs that can be generalized to downstream tasks.

# 03 Motif-based Graph Self-supervised Learning

## ❖ Chemistry-inspired Molecule Fragmentation

- BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures algorithm)

  : 16 rules, breaks strategic bonds

  화학식에 따라 분자를 잘라서 분해하는 알고리즘 (화학적으로 맞게 분해해야함)

  : 중요한 구조적 및 기능적 성분을 유지하도록 절단

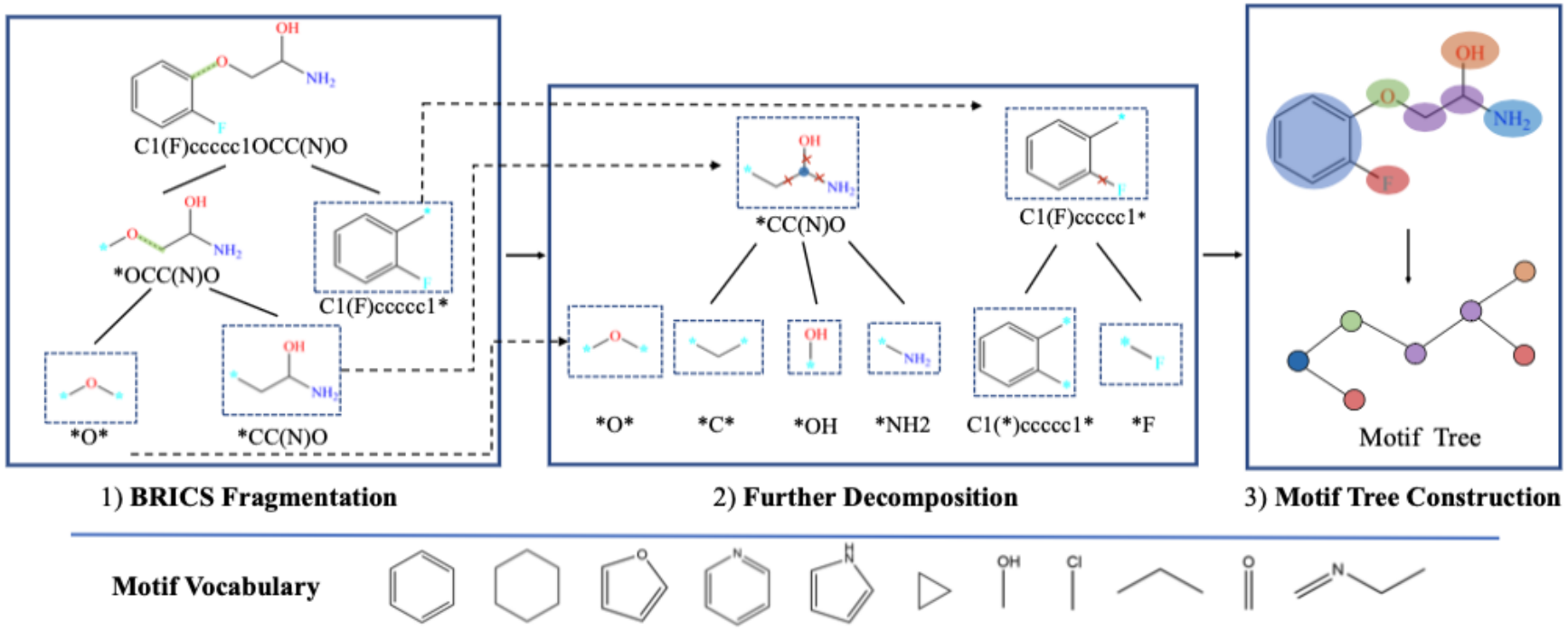
# 03 Motif-based Graph Self-supervised Learning

## ❖ Chemistry-inspired Molecule Fragmentation

- BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures algorithm)

  : post-processing procedure, to alleviate the combination explosion.

  ➔ effectively reduce the size of motif vocabulary and improve the occurrence frequency of motifs



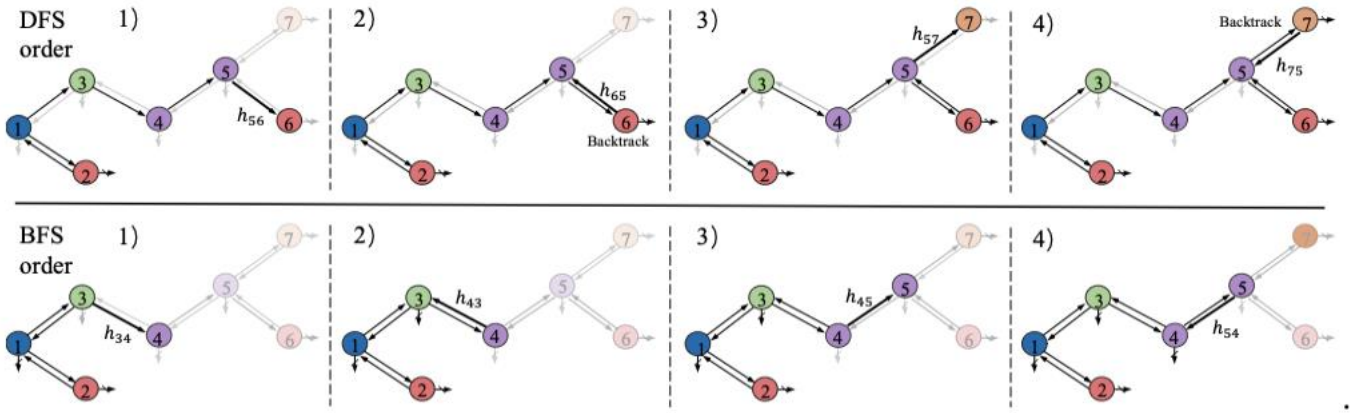
# 03 Motif-based Graph Self-supervised Learning

## ❖ Motif Generation

- The goal of the motif generation task is to let GNNs learn the data distribution of graph motifs so that the pre-trained GNNs can easily generalize to downstream tasks with several finetuning steps on graphs from similar domains.

$$p(\mathcal{T}(G);\theta) = \mathbb{E}_{\pi}\left[p_\theta(\mathcal{V}^\pi, \mathcal{E}^\pi)\right],$$

$$\log p_\theta(\mathcal{V}, \mathcal{E}) = \sum_{i=1}^{|\mathcal{V}|} \log p_\theta(\mathcal{V}_i, \mathcal{E}_i \mid \mathcal{V}_{<i}, \mathcal{E}_{<i}).$$

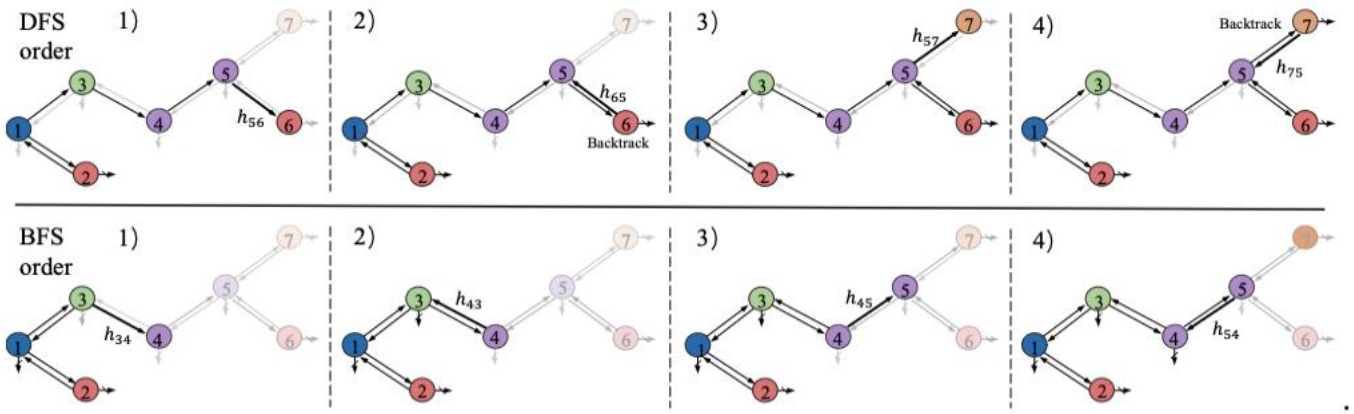


# 03 Motif-based Graph Self-supervised Learning

## ❖ Motif Generation

- Motif tree message passing

$$\mathrm{h}_{i,j} = \mathrm{GRU}\left(x_i,\ \{\mathrm{h}_{k,i}\}_{(k,i)\in\hat{\mathcal{E}}_t, k\neq j}\right),$$




$$s_{i,j} = \sum_{(k,i)\in\hat{\mathcal{E}}_t, k\neq j} \mathrm{h}_{k,i}$$

$$z_{i,j} = \sigma(\mathrm{W}^z x_i + \mathrm{U}^z s_{i,j} + b^z)$$

$$r_{k,i} = \sigma(\mathrm{W}^r x_i + \mathrm{U}^r \mathrm{h}_{k,i} + b^r)$$

$$\tilde{\mathrm{h}}_{i,j} = \tanh(\mathrm{W} x_i + U \sum_{k=\mathcal{N}(i)\setminus j} r_{k,i} \odot \mathrm{h}_{k,i})$$

$$\mathrm{h}_{i,j} = (1 - z_{ij}) \odot s_{ij} + z_{ij} \odot \tilde{\mathrm{h}}_{i,j}.$$

# 03 Motif-based Graph Self-supervised Learning

## ❖ Motif Generation

- Topology prediction
  - When MGSSL visits motif $i$, it needs to make binary predictions (1-layer NN with sigmoid)

$$p_t = \sigma\left(U^d \cdot \tau(W_1^d x_i + W_2^d \sum_{(k,i)\in\hat{\mathcal{E}}_t} h_{k,i})\right),$$

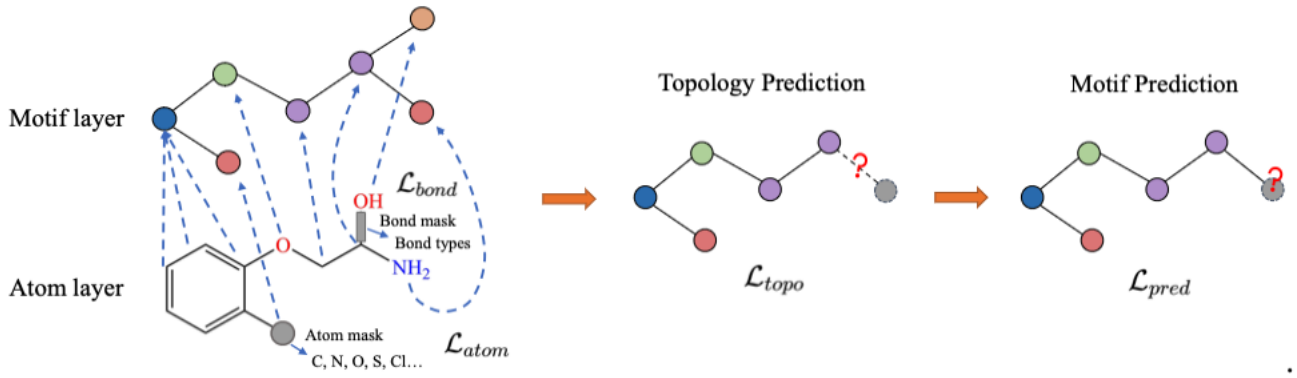
- Motif label prediction
  - When motif i generate a child motif j, we predict the label of child j

$$q_j = \text{softmax}(U^l \tau(W^l h_{ij})),$$

"Motif generation loss" ➡ $$\mathcal{L}_{motif} = \sum_t \mathcal{L}_{topo}(p_t, \hat{p}_t) + \sum_j \mathcal{L}_{pred}(q_j, \hat{q}_j).$$

# 03 Motif-based Graph Self-supervised Learning

## ❖ Multi-level Self-supervised Pre-training




- Atom level
  - Leverage attribute masking to let GNNs firstly learn the regularities of the node/edge attributes. [atom, bond attribute prediction]
- Motif level [motif generation]
  - Motif generations (topology prediction, motif prediction)

"Total loss" ➡ $$\mathcal{L}_{ssl} = \lambda_1 \mathcal{L}_{motif} + \lambda_2 \mathcal{L}_{atom} + \lambda_3 \mathcal{L}_{bond},$$

➔ MGDA-UB [Frank-Wolfe algorithm] : instead of grid search…

# 04

# Experimental results

# 04 Experimental results

## ❖ Datasets & baseline

- Datasets
  - ZINC15 dataset [250K unlabeled molecules sampled]
  - 8 binary classification benchmark datsets in “MoleculeNet”
  - “RDKit” to preprocess the SMILES strings from various datasets
  - “scaffold-split”

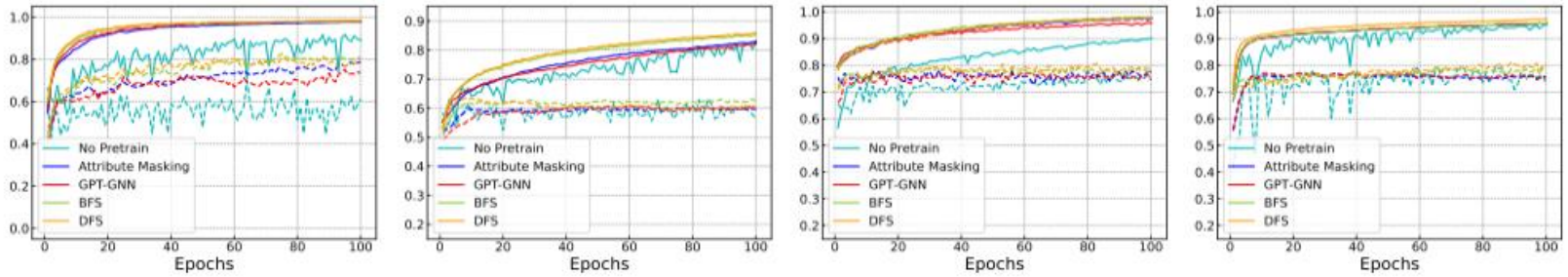
- Baseline

- **Deep Graph Infomax** [41] maximizes the mutual information between the representations of the whole graphs and the representations of its sampled subgraphs.
- **Attribute masking** [14] masks node/edge features and let GNNs predict these attributes.
- **GCC** [30] designs the pretraining task as discriminating ego-networks sampled from a certain node ego-networks sampled from other nodes.
- **Grover** [32] predicts the contextual properties based on atom embeddings to encode contextual information into node embeddings.
- **GPT-GNN** [15] is a generative pretraining task which predicts masked edges and node attributes.

# 04 Experimental results

Table 1: Test ROC-AUC (%) performance on molecular property prediction benchmarks using different pre-training strategies with GIN. The rightmost column averages the mean of test performance across the 8 datasets. The best result for each dataset are bolded.

| SSL methods | muv | clintox | sider | hiv | tox21 | bace | toxcast | bbbp | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| No pretrain | 71.7±2.3 | 58.2±2.8 | 57.2±0.7 | 75.4±1.5 | 74.3±0.5 | 70.0±2.5 | 63.3±1.5 | 65.5±1.8 | 67.0 |
| Infomax | 75.1±2.8 | 73.0±3.2 | 58.2±0.5 | 76.5±1.6 | 75.2±0.3 | 75.6±1.0 | 62.8±0.6 | 68.1±1.3 | 70.6 |
| Attribute masking | 74.7±1.9 | 77.5±3.1 | 59.6±0.7 | 77.9±1.2 | **77.2±0.4** | 78.3±1.1 | 63.3±0.8 | 65.6±0.9 | 71.8 |
| GCC | 74.1±1.4 | 73.2±2.6 | 58.0±0.9 | 75.5±0.8 | 76.6±0.5 | 75.0±1.5 | 63.5±0.4 | 66.9±0.7 | 70.4 |
| GPT-GNN | 75.0±2.5 | 74.9±2.7 | 59.3±0.8 | 77.0±1.7 | 76.1±0.4 | 78.5±0.9 | 63.1±0.5 | 67.5±1.3 | 71.4 |
| Grover | 75.8±1.7 | 76.9±1.9 | 60.7±0.5 | 77.8±1.4 | 76.3±0.6 | 79.5±1.1 | 63.4±0.6 | 68.0±1.5 | 72.3 |
| MGSSL (DFS) | 78.1±1.8 | 79.7±2.2 | 60.5±0.7 | **79.5±1.1** | 76.4±0.4 | **79.7±0.8** | 63.8±0.3 | **70.5±1.1** | 73.5 |
| MGSSL (BFS) | **78.7±1.5** | **80.7±2.1** | **61.8±0.8** | 78.8±1.2 | 76.5±0.3 | 79.1±0.9 | **64.1±0.7** | 69.7±0.9 | **73.7** |

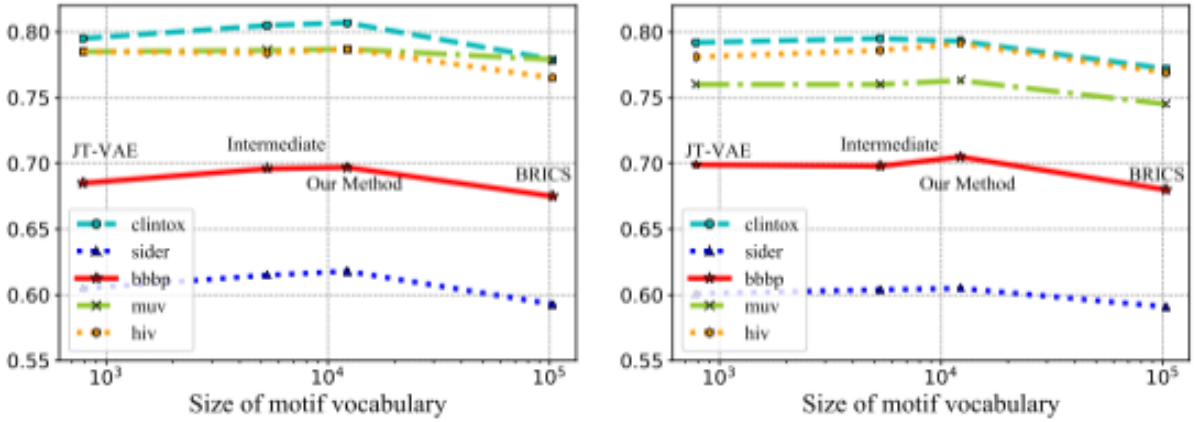



(a) clintox (b) sider (c) hiv (d) bace

# 04 Experimental results

Table 2: Compare pre-training gains with different GNN architectures, averaged ROC-AUC (%) on 8 benchmark datasets

| Model | GCN | GIN | RGCN | DAGNN | GraphSAGE |
|---|---|---|---|---|---|
| No pretrain | 68.8 | 67.0 | 68.3 | 67.1 | 68.3 |
| MGSSL (BFS) | 72.7 | 73.7 | 73.0 | 72.3 | 73.4 |
| Relative gain | 5.7% | 10.0% | 6.9% | 7.7 % | 7.5% |




(a) MGSSL (BFS) (b) MGSSL (DFS)

Figure 5: Influence of the size of motif vocabulary

# 04 Experimental results

| Methods | Avg. ROC-AUC |
|---|---|
| w/o atom-level | 73.0 |
| Sequential pre-training | 73.4 |
| Multi-level | **73.7** |

Table 3: Ablation studies on multi-level self-supervised pre-training.

1) The Atom-level pre-training tasks enables GNNs to first capture the atom-level information, which can benefit higher level
2) Since our multi-level pre-training unifies multi-scale pre-training tasks and adaptively assigns the weights for hierarchical tasks

# 05

# Conclusion and future work

# 05 Experiments & Conclusion

- We proposed Motif-based Graph Self-supervised Learning (MGSSL), which pre-trains GNNs with a novel motif generation task. Through pre-training, MGSSL empowers GNNs to capture the rich semantic and structural information in graph motifs.
- First, a retrosynthesis-based algorithm with two additional rules are leveraged to fragment molecule graphs and derive semantic meaningful motifs.
- Second, a motif generative pre-training framework is designed and two specific generation orders are considered (BFS and DFS). Furthermore, we designed a multi-level pre-training to unify hierarchical self-supervised tasks.
- Futrue work

  1) Designing more self-supervised pre-training tasks based on graph motifs.

  2) Exploring motif-based pre-training in other domains other than molecules.