Fragment-based Pretraining and Finetuning on Molecular Graphs

Kha-Dinh Luong, Ambuj Singh

{vluong,ambuj}@cs.ucsb.edu

Department of Computer Science

University of California, Santa Barbara, USA

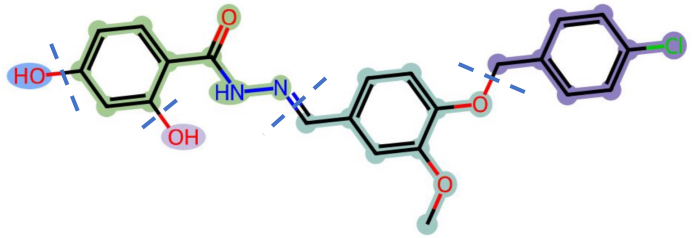
Motivation

Molecular graphs exhibit structural patterns at multiple resolutions.

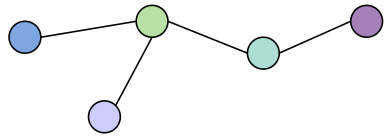
Existing pretraining strategies on Graph Neural Networks include:

- Node-level tasks: only encode local structural patterns.
- Graph-level tasks: may miss granular details.

Few works have explored pretraining at the fragment level, a promising middle ground.



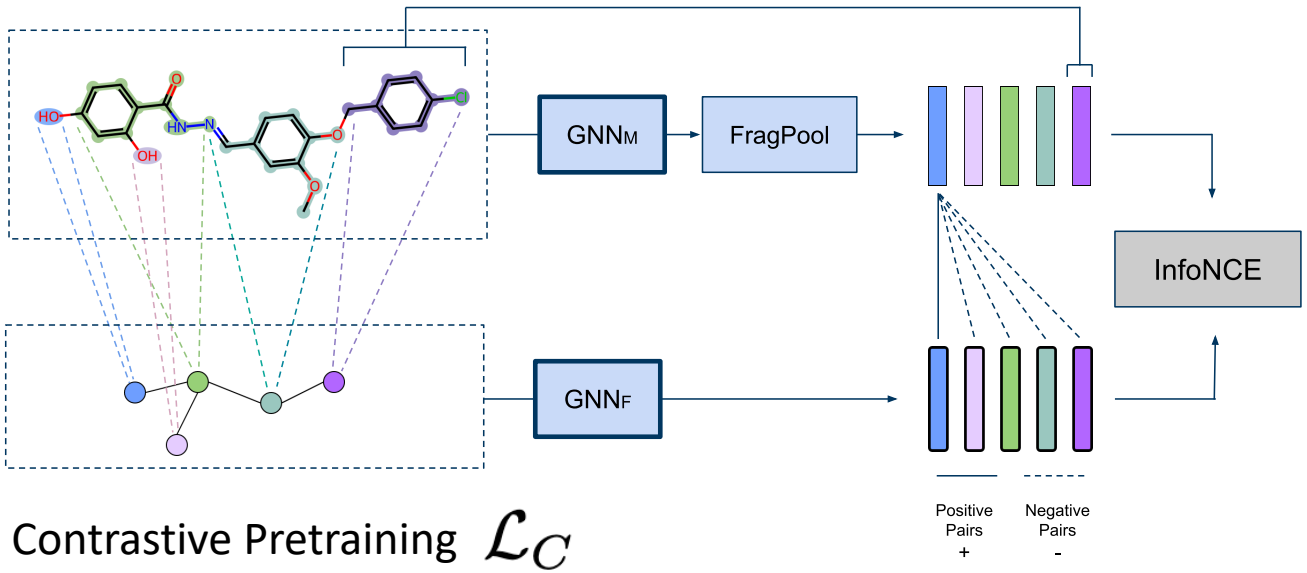
Molecular Graph



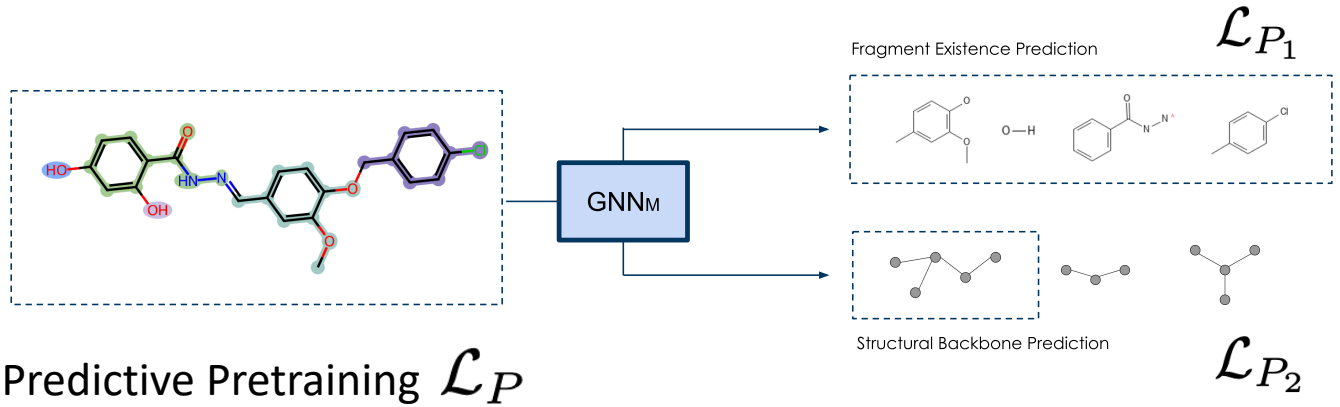
Fragment Graph

We prepare molecular graphs and fragment graphs and utilizing them in both pretraining and finetuning.
Molecules are fragmented according to a vocabulary extracted via existing frequency-based methods.

Fragment-based Pretraining



Contrastive Pretraining $\mathcal{L}_C$



Predictive Pretraining $\mathcal{L}_P$

GNNM Process molecular graphs, encoding local patterns

GNNF Process fragment graphs, encoding global patterns

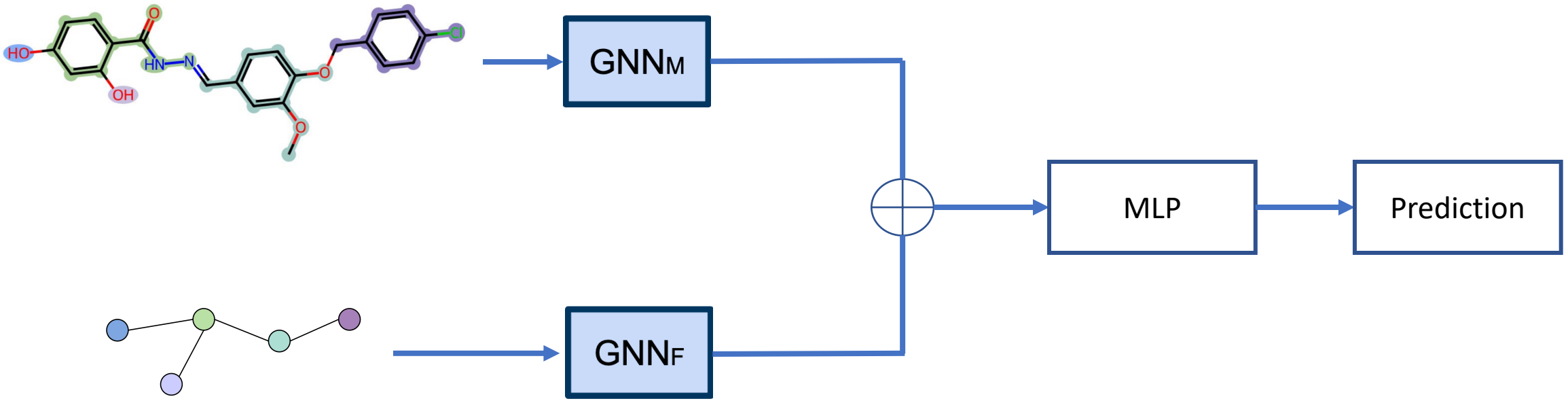
Contrastively enforces both local and global structural patterns into node embeddings

Combined Pretraining

$$\mathcal{L}_P = \mathcal{L}_{P_1} + \mathcal{L}_{P_2}$$

$$\mathcal{L} = \alpha\mathcal{L}_P + (1-\alpha)\mathcal{L}_C$$

Fragment-based Finetuning



Both pretrained $\mathrm{GNN_M}$ and $\mathrm{GNN_F}$ can be utilized in downstream finetuning and prediction.

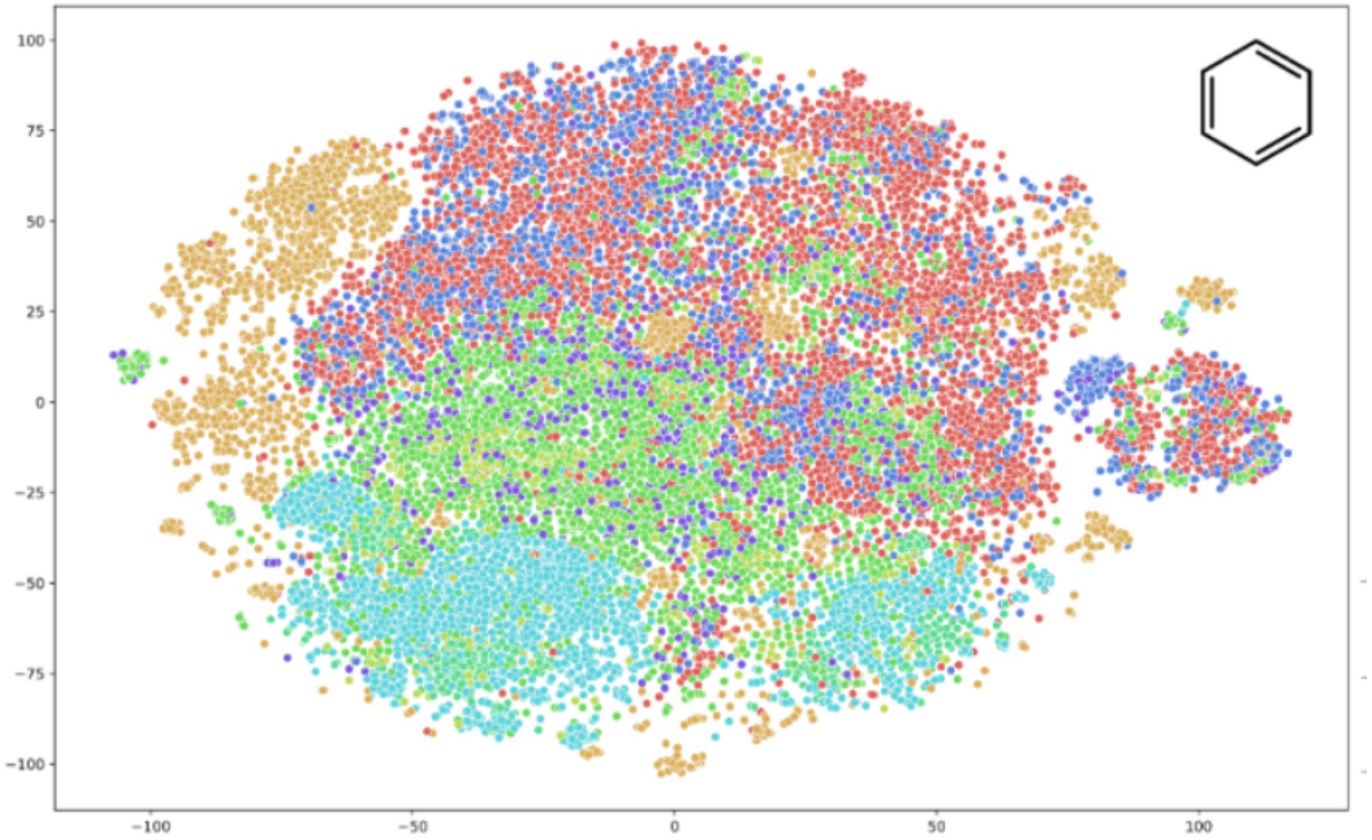
Evaluation: Common Benchmarks

Pretraining Strategies	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	Avg. Rank	Avg. AUC
AttrMasking	64.3 ± 2.8	**76.7 ± 0.4**	64.2 ± 0.5	61.0 ± 0.7	71.8 ± 4.1	74.7 ± 1.4	77.2 ± 1.1	79.3 ± 1.6	7.88	71.15
ContextPred	68.0 ± 2.0	75.7 ± 0.7	63.9 ± 0.6	60.9 ± 0.6	65.9 ± 3.8	75.8 ± 1.7	77.3 ± 1.0	79.6 ± 1.2	7.56	70.89
G-Motif	66.9 ± 3.1	73.6 ± 0.7	62.3 ± 0.6	61.0 ± 1.5	77.7 ± 2.7	73.0 ± 1.8	73.8 ± 1.2	73.0 ± 3.3	14.25	70.16
G-Contextual	69.9 ± 2.1	75.0 ± 0.6	62.8 ± 0.7	58.7 ± 1.0	60.6 ± 5.2	72.1 ± 0.7	76.3 ± 1.5	79.3 ± 1.1	11.88	69.34
GPT-GNN	64.5 ± 1.4	74.9 ± 0.3	62.5 ± 0.4	58.1 ± 0.3	58.3 ± 5.2	75.9 ± 2.3	65.2 ± 2.1	77.9 ± 3.2	13.63	67.16
GraphLoG	67.8 ± 1.9	75.1 ± 1.0	62.4 ± 0.2	59.5 ± 1.5	65.3 ± 3.2	73.6 ± 1.2	73.7 ± 0.9	80.2 ± 3.5	12.56	69.70
GraphCL	69.7 ± 0.7	73.9 ± 0.7	62.4 ± 0.6	60.5 ± 0.9	76.0 ± 2.7	69.8 ± 2.7	**78.5 ± 1.2**	75.4 ± 1.4	12.13	70.78
JOAO	70.2 ± 1.0	75.0 ± 0.3	62.9 ± 0.5	60.0 ± 0.8	81.3 ± 2.5	71.7 ± 1.4	76.7 ± 1.2	77.3 ± 0.5	9.56	71.89
JOAOv2	71.4 ± 0.9	74.3 ± 0.6	63.2 ± 0.5	60.5 ± 0.7	81.0 ± 1.6	73.7 ± 1.0	77.5 ± 1.2	75.5 ± 1.3	8.94	72.14
GraphMVP	68.5 ± 0.2	74.5 ± 0.4	62.7 ± 0.1	62.3 ± 1.6	79.0 ± 2.5	75.0 ± 1.4	74.8 ± 1.4	76.8 ± 1.1	10.00	71.70
MGSSL	68.9 ± 2.5	74.9 ± 0.6	63.3 ± 0.5	57.7 ± 0.7	67.5 ± 5.5	73.2 ± 1.9	75.7 ± 1.3	**82.1 ± 2.7**	10.94	70.41
GraphFP-JT$_C$	71.5 ± 0.9	75.2 ± 0.5	63.6 ± 0.5	62.0 ± 1.0	77.7 ± 4.5	76.0 ± 2.2	75.6 ± 1.0	79.7 ± 1.3	6.13	72.66
GraphFP-JT$_{CF}$	70.2 ± 1.7	72.7 ± 0.8	62.5 ± 0.9	59.3 ± 1.3	75.9 ± 5.6	73.9 ± 1.3	73.0 ± 1.9	74.2 ± 2.8	13.56	70.21
GraphFP$_C$	71.5 ± 1.6	75.5 ± 0.4	63.8 ± 0.6	61.4 ± 0.9	78.6 ± 2.7	**77.2 ± 1.5**	76.3 ± 1.0	78.2 ± 3.4	5.50	72.81
GraphFP$_P$	68.2 ± 1.2	**76.0 ± 0.5**	63.2 ± 0.7	59.3 ± 1.0	53.8 ± 3.8	74.5 ± 2.1	76.7 ± 1.0	80.7 ± 4.8	9.50	69.05
GraphFP$_{CP}$	71.3 ± 1.7	75.5 ± 0.5	64.7 ± 0.2	61.3 ± 0.6	73.7 ± 3.9	76.6 ± 1.8	76.3 ± 1.0	81.3 ± 2.2	5.19	72.59
GraphFP$_{CF}$	70.1 ± 1.8	74.3 ± 0.3	**65.3 ± 0.8**	**64.7 ± 1.0**	**87.7 ± 5.8**	74.5 ± 1.8	76.1 ± 2.0	77.1 ± 2.1	7.25	73.73
GraphFP$_{CPF}$	**72.0 ± 1.7**	74.0 ± 0.7	63.9 ± 0.9	63.6 ± 1.2	84.7 ± 5.8	75.4 ± 1.9	78.0 ± 1.5	80.5 ± 1.8	**4.56**	**74.01**

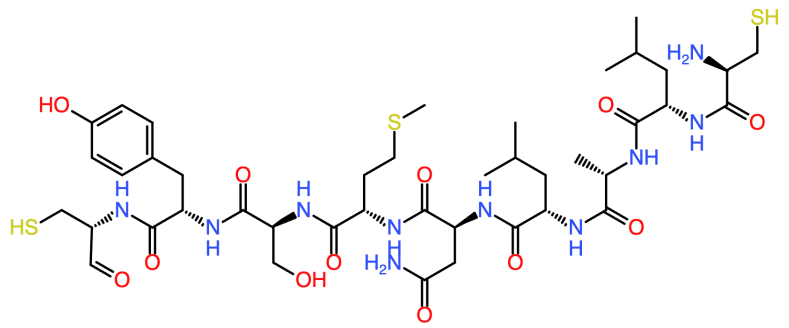
C: Contrastive Pretraining
P: Predictive Pretraining
F: Fragment-based Finetuning

Red: Best result
Red: Second-best result
Red: Third-best result

Evaluation: Capturing Global Structures



t-SNE: embeddings of the same fragment from different graphs. Colors indicate distinct structural backbones.



Methods	Peptide-func Test AP	Peptide-struct Test MAE
GCN	0.5930 ± 0.0023	0.3496 ± 0.0013
GCNII	0.5543 ± 0.0078	0.3471 ± 0.0010
GIN	0.5498 ± 0.0079	0.3547 ± 0.0045
GatedGCN	0.5864 ± 0.0077	0.3420 ± 0.0013
GatedGCN+RWSE	0.6069 ± 0.0035	0.3357 ± 0.0006
GraphFP$_{CF}$	0.6267 ± 0.0073	0.3137 ± 0.0019

Long-range Graph Benchmarks (peptides)