Self-Supervised Graph Transformer on Large-Scale Molecular Data

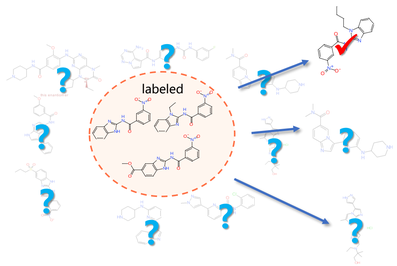
Yu Rong*, Yatao Bian*, Tingyang Xu,
Weiyang Xie, Ying Wei, Wenbing Huang,
Junzhou Huang

Tencent AI Lab

清華大學 Tsinghua University

Background & Contributions

GNNs are widely adopted for molecular tasks.



Challenges:

- ❑ Insufficient labeled molecules for supervised training
- ❑ Poor generalization capabilities to newly-synthesized molecules

GROVER

Integrating message passing networks with the Transformer-style architecture → delivers → More expressive encoders of molecules

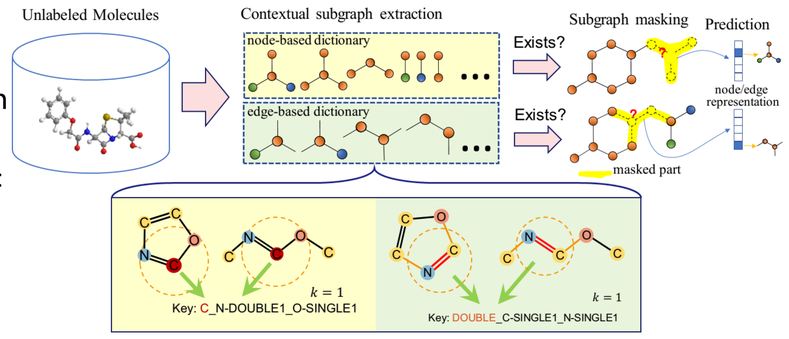
Carefully designed self-supervised tasks in node, edge & graph levels → enables → Rich structural and semantic information of molecules from enormous unlabeled data

Proposed: Self-supervised Tasks

Good self-supervision task shall have *reliable* and *cheap* prediction target

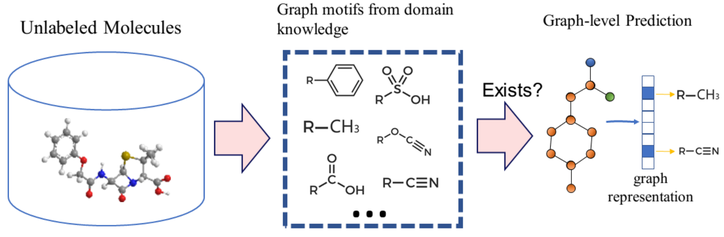
Node/edge level task: contextual property prediction

Target reflects contextual property: recurrent statistical properties of local subgraph → A multi-class classification problem



Graph level task: **motif** prediction

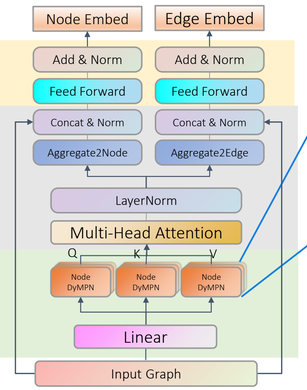
Motifs: recurrent sub-graphs, such as functional groups → A multi-label classification problem



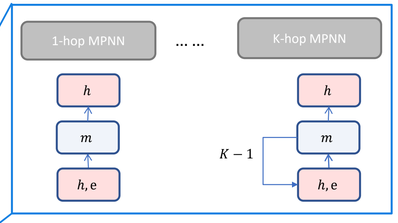
Proposed: GTransformer Architecture

A transformer based neural network with tailored GNNs as the self-attention building blocks.

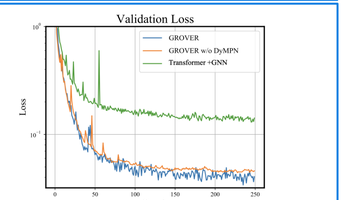
- Output for both node embedding and edge embeddings.
- **Multi-Head Attention**: model global interaction between nodes/edges.
- **Long-range Residual Connection**: alleviating the vanishing gradient and over-smoothing problem.
- **MPNN**: Extract local structural information of graphs.
- **dyMPN**: Randomize the message passing hops for the dynamic receptive field modeling.



Sample a random-hop MPNN at each iteration



Better generalization ability



Experimental Results

GROVER$_{base}$: 48M parameters

GROVER$_{large}$: 100M parameters

Pre-training GROVER models on 10M unlabelled molecules → Verifying on down-stream tasks with fine-tuning → Significant improvement over SOTA models on 11 challenging benchmarks

Performance comparison for classification problems (regression results see appendix)

250 Nvidia V100 GPUs used

Pre-trained methods

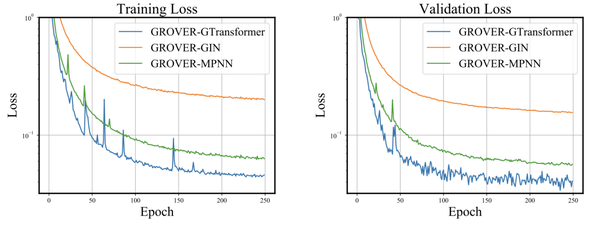
Dataset	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast
# Molecules	2039	1427	1478	1513	7831	8575
	Classification (Higher is better)					
TF_Robust [40]	$0.860_{(0.087)}$	$0.607_{(0.033)}$	$0.765_{(0.085)}$	$0.824_{(0.022)}$	$0.698_{(0.012)}$	$0.585_{(0.031)}$
GraphConv [24]	$0.877_{(0.036)}$	$0.593_{(0.035)}$	$0.845_{(0.051)}$	$0.854_{(0.011)}$	$0.772_{(0.041)}$	$0.650_{(0.025)}$
Weave [23]	$0.837_{(0.065)}$	$0.543_{(0.034)}$	$0.823_{(0.023)}$	$0.791_{(0.008)}$	$0.741_{(0.044)}$	$0.678_{(0.024)}$
SchNet [45]	$0.847_{(0.024)}$	$0.545_{(0.038)}$	$0.717_{(0.042)}$	$0.750_{(0.033)}$	$0.767_{(0.025)}$	$0.679_{(0.021)}$
MPNN [13]	$0.913_{(0.041)}$	$0.595_{(0.030)}$	$0.879_{(0.054)}$	$0.815_{(0.044)}$	$0.808_{(0.024)}$	$0.691_{(0.013)}$
DMPNN [63]	$0.919_{(0.030)}$	$0.632_{(0.023)}$	$0.897_{(0.040)}$	$0.852_{(0.053)}$	$0.826_{(0.023)}$	$0.718_{(0.011)}$
MGCN [30]	$0.850_{(0.064)}$	$0.552_{(0.018)}$	$0.634_{(0.042)}$	$0.734_{(0.030)}$	$0.707_{(0.016)}$	$0.663_{(0.009)}$
AttentiveFP [61]	$0.908_{(0.050)}$	$0.605_{(0.060)}$	$0.933_{(0.020)}$	$0.863_{(0.015)}$	$0.807_{(0.020)}$	$0.579_{(0.001)}$
N-GRAM [29]	$0.912_{(0.013)}$	$0.632_{(0.005)}$	$0.855_{(0.037)}$	$0.876_{(0.035)}$	$0.769_{(0.027)}$	-[4]
HU. et.al[18]	$0.915_{(0.040)}$	$0.614_{(0.006)}$	$0.762_{(0.058)}$	$0.851_{(0.027)}$	$0.811_{(0.015)}$	$0.714_{(0.019)}$
GROVER$_{base}$	$0.936_{(0.008)}$	$0.656_{(0.006)}$	$0.925_{(0.013)}$	$0.878_{(0.016)}$	$0.819_{(0.020)}$	$0.723_{(0.010)}$
GROVER$_{large}$	$\mathbf{0.940}_{(0.019)}$	$\mathbf{0.658}_{(0.023)}$	$\mathbf{0.944}_{(0.021)}$	$\mathbf{0.894}_{(0.028)}$	$\mathbf{0.831}_{(0.025)}$	$\mathbf{0.737}_{(0.010)}$

- ❑ GROVER models achieve the best performance
- ❑ Large model (GROVER$_{large}$) enjoys high expressive power
- ❑ Larger improvement achieved for dataset with less label info.

Self-supervised pre-training is useful

	GROVER	No Pretrain	Abs. Imp.
BBBP (2039)	**0.940**	0.911	+0.029
SIDER (1427)	**0.658**	0.624	+0.034
ClinTox (1478)	**0.944**	0.884	+0.060
BACE (1513)	**0.894**	0.858	+0.036
Tox21 (7831)	**0.831**	0.803	+0.028
ToxCast (8575)	**0.737**	0.721	+0.016
Average	**0.834**	0.803	+0.038

GTransformer backbone matters



References

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019