Self-Supervised Graph Transformer on Large-Scale Molecular Data

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

Presented by

Monjura Afrin Rumi

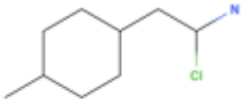
Motivation

- Use of deep learning in drug discovery, molecule property prediction
- Issues in using deep learning
  - insufficient labeled data for molecular tasks
    - Time-consuming and resource-costly
  - poor generalization capability of models in the enormous chemical space
- Pre-train using unlabeled data in self-supervised manner

Motivation

- Representation of molecules
  - SMILES – not topology aware
    - BERT
    - N-gram approach
  - Graph – preserves rich structural information
    - Context prediction
    - node-level self-supervised learning
    - graph property prediction for graph-level pre-training

Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(0)=0



Input molecule


Molecular graph

Proposed Method

- **GROVER**: **G**raph **R**epresentation fr**O**m self-super**V**ised m**E**ssage passing t**R**ansformer
- node/edge-level tasks
  - masks a local subgraph of the target node/edge
  - predicts this contextual property from node embeddings
- graph-level tasks
  - Extracts the semantic motifs existing in molecular graphs
  - predicts the occurrence of these motifs for a molecule from graph embeddings.

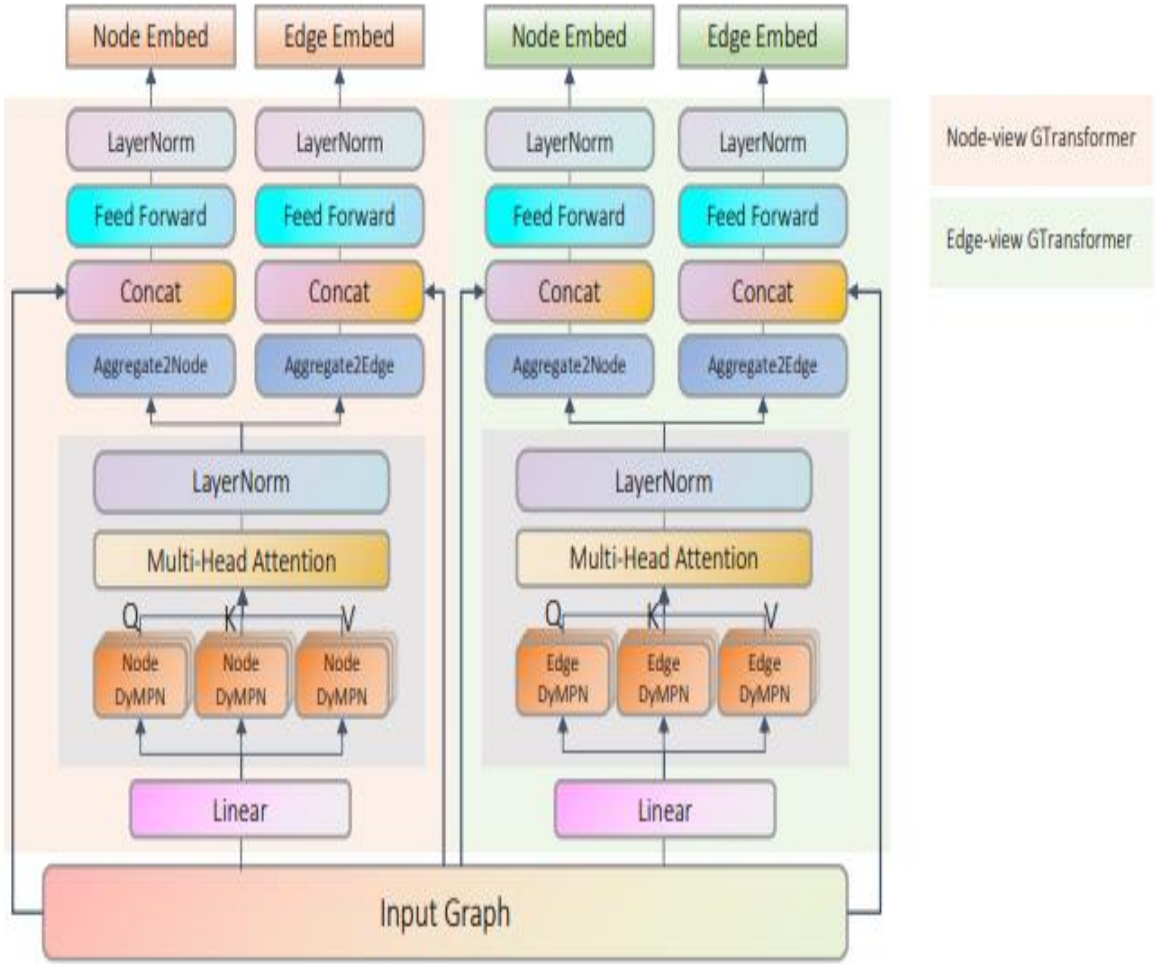
GROVER Architecture: GTransformer

- Attention mechanism

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d})\mathbf{V}$$

- Multi-head attention

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_k)\mathbf{W}^O,$$
$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^\mathbf{Q}, \mathbf{K}\mathbf{W}_i^\mathbf{K}, \mathbf{V}\mathbf{W}_i^\mathbf{V}),$$



GROVER Architecture: dyMPN

- Graph Neural Network
  - Message passing/ neighborhood aggregation
  - Update hidden state

$$\mathbf{m}_v^{(l,k)} = \text{AGGREGATE}^{(l)}(\{(\mathbf{h}_v^{(l,k-1)}, \mathbf{h}_u^{(l,k-1)}, \mathbf{e}_{uv}) \mid u \in \mathcal{N}_v\}),$$
$$\mathbf{h}_v^{(l,k)} = \sigma(\mathbf{W}^{(l)}\mathbf{m}_v^{(l,k)} + \mathbf{b}^{(l)}),$$

  - Randomized strategy for choosing KL
    I. Uniform distribution $K_l \sim U(a, b)$
    II. Truncated normal distribution $\phi(\mu, \sigma, a, b)$

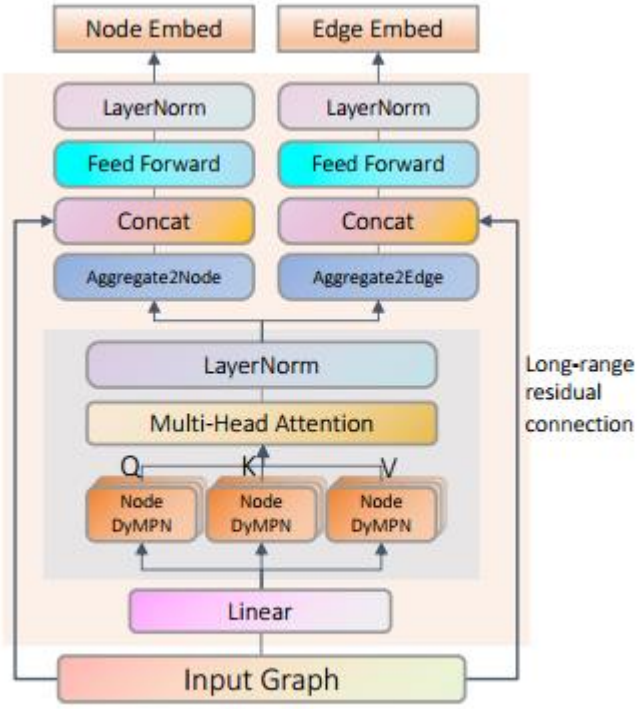


Figure 1: Overview of GTransformer.

GROVER Architecture - continued

- Aggregate hidden states of nodes

$$\mathbf{m}_v^{\text{node-embedding-from-node-states}} = \sum_{u \in \mathcal{N}_v} \bar{\mathbf{h}}_u$$

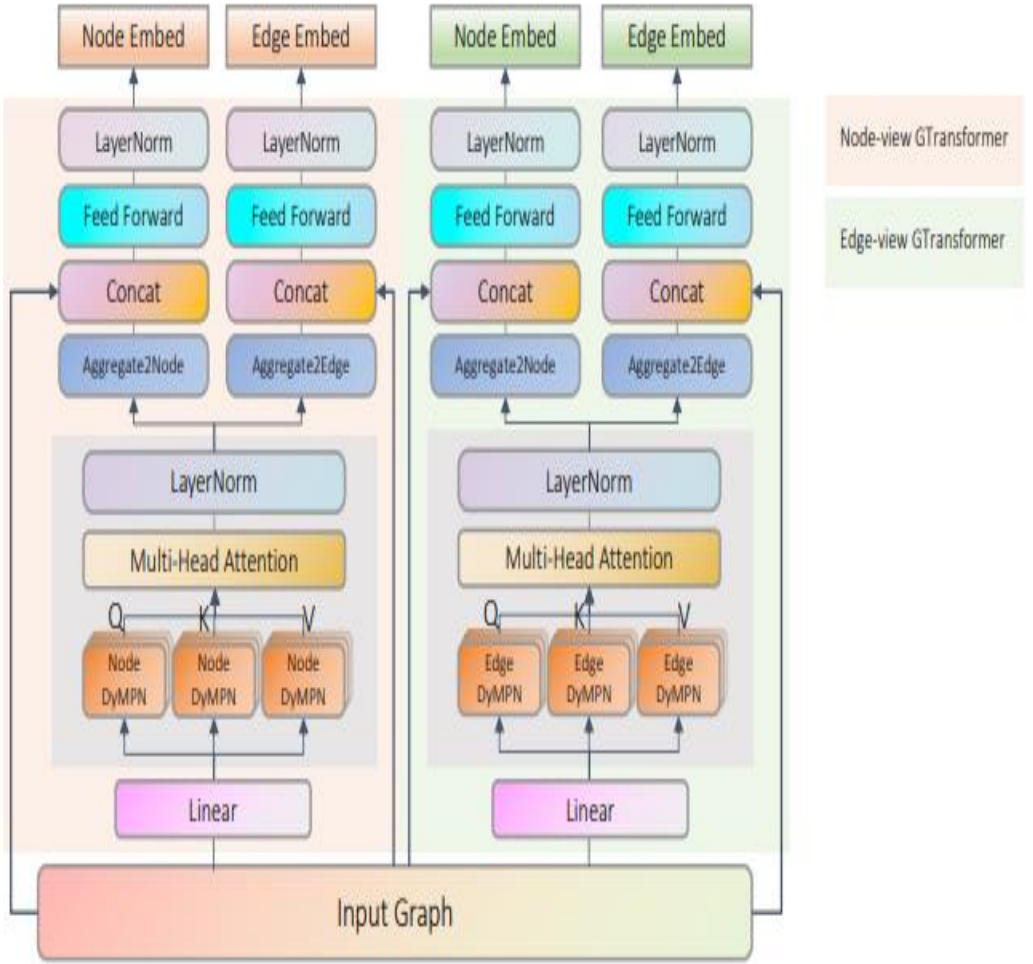
$$\mathbf{m}_{vw}^{\text{edge-embedding-from-node-states}} = \sum_{u \in \mathcal{N}_v \setminus w} \bar{\mathbf{h}}_u.$$

- Aggregate hidden states of edges

$$\mathbf{m}_v^{\text{node-embedding-from-edge-states}} = \sum_{u \in \mathcal{N}_v} \bar{\mathbf{h}}_{uv},$$

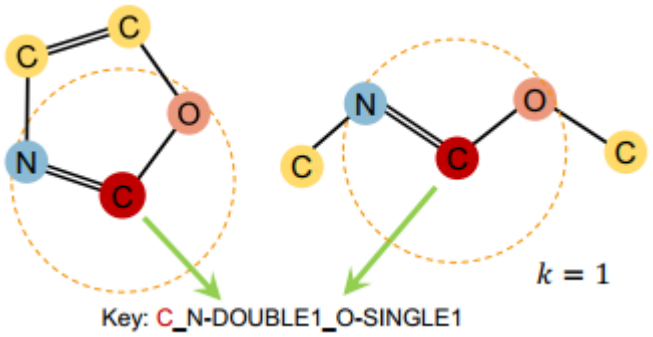
$$\mathbf{m}_{vw}^{\text{edge-embedding-from-edge-states}} = \sum_{u \in \mathcal{N}_v \setminus w} \bar{\mathbf{h}}_{uv}.$$

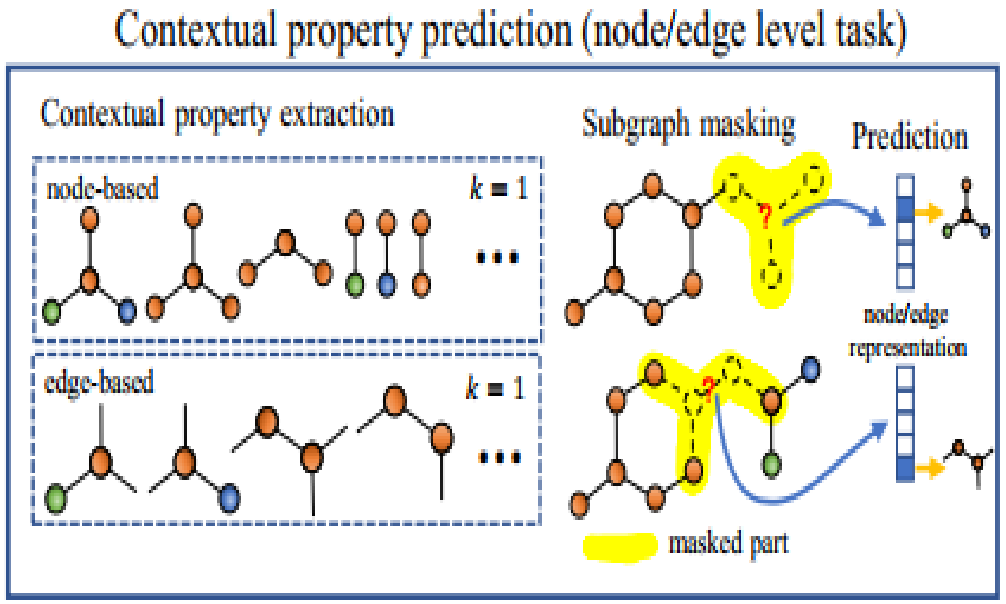
- long-range residual connection
  - Vanishing gradient
  - Over-smoothing



Self-supervised Task Construction

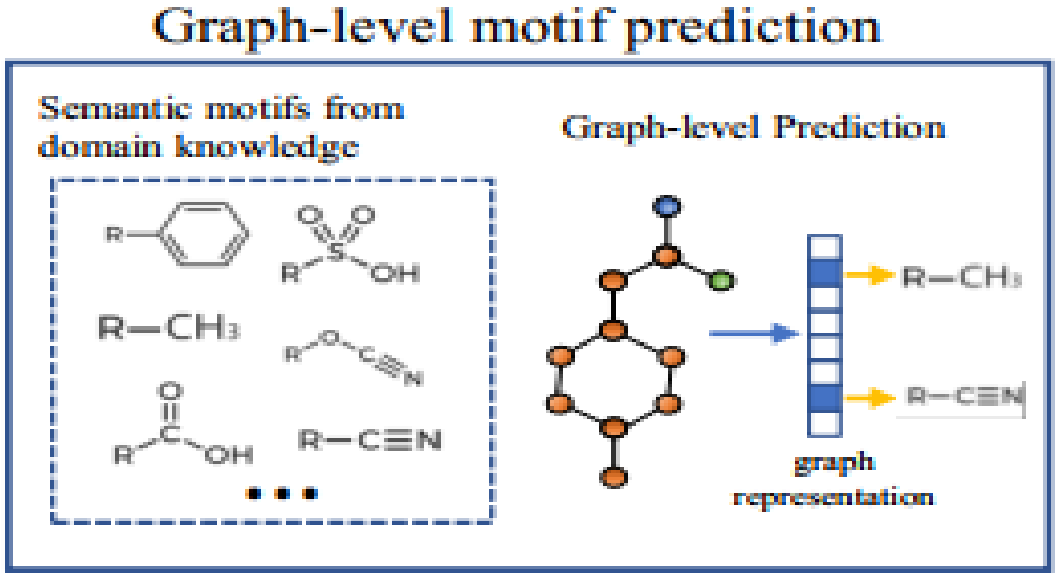
- Contextual Property Prediction





Self-supervised Task Construction

- **Graph-level Motif Prediction**
  - Motifs - recurrent sub-graphs among the input graph data
  - functional groups encodes the rich domain knowledge of molecules
  - Motif extraction tool: RDKit
  - Multi-label classification



Related Work

- **Molecular Representation Learning**
  - chemical fingerprint: represent molecules in the vector space
    - encode the neighbors of atoms in the molecule into a fix-length vector – ECFP
    - Neural fingerprints using convolutional layer - TF_Roubust
  - SMILES
    - RNN-based models to produce molecular representations
  - Graph representation
    - Graph Convolutional Network – GraphConv, Weave, SchNet
    - Graph Attention Network - AttentiveFP
    - GNN – MPNN, DMPNN
    - Hierarchical GNN - MGCN

Related Work

- **Self-supervised Learning on Graphs**
  - Learning objective – vertex proximity relationship
  - Vertex embedding - N-gram model
  - node/edge type prediction - Hu et.al.

Experiments

- Pre-training
  - 11 million (M) unlabelled molecules 10% for validation
  - Context radious k =1, node = 2518, edge = 2686
  - Randomly mask 15% of node and edge labels for prediction
  - RDKit - extract 85 functional groups as the motifs of molecules
  - $\text{GROVER}_{\text{base}}$ - ~48M parameters
  - $\text{GROVER}_{\text{large}}$ - ~100M parameters
- Fine-tuning tasks
  - train/validation/test - 8:1:1
  - scaffold splitting

Table 3: Dataset information

Type	Category	Dataset	# Tasks	# Compounds	Metric
Classification	Biophysics	BBBP	1	2039	ROC-AUC
	Physiology	SIDER	27	1427	ROC-AUC
		ClinTox	2	1478	ROC-AUC
		BACE	1	1513	ROC-AUC
		Tox21	12	7831	ROC-AUC
		ToxCast	617	8575	ROC-AUC
Regression	Physical chemistry	FreeSolv	1	642	RMSE
		ESOL	1	1128	RMSE
		Lipophilicity	1	4200	RMSE
	Quantum mechanics	QM7	1	6830	MAE
		QM8	12	21786	MAE

Results

- 6.1% relative improvement
- 2.2% classification
- 10.8% regression

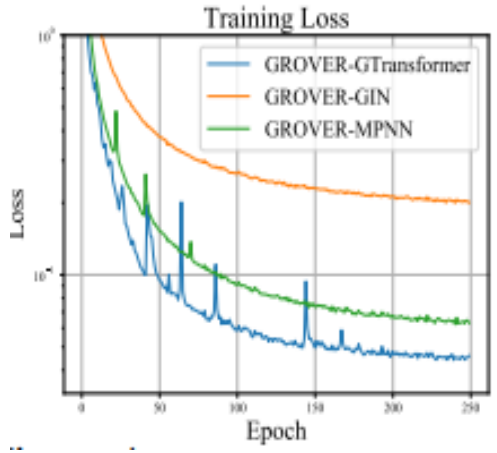
Classification (Higher is better)						
Dataset # Molecules	BBBP 2039	SIDER 1427	ClinTox 1478	BACE 1513	Tox21 7831	ToxCast 8575
TF_Robust [40]	$0.860_{(0.087)}$	$0.607_{(0.033)}$	$0.765_{(0.085)}$	$0.824_{(0.022)}$	$0.698_{(0.012)}$	$0.585_{(0.031)}$
GraphConv [24]	$0.877_{(0.036)}$	$0.593_{(0.035)}$	$0.845_{(0.051)}$	$0.854_{(0.011)}$	$0.772_{(0.041)}$	$0.650_{(0.025)}$
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]	$0.915_{(0.040)}$	$0.614_{(0.006)}$	$0.762_{(0.058)}$	$0.851_{(0.027)}$	$0.811_{(0.015)}$	$0.714_{(0.019)}$
$\text{GROVER}_{\text{base}}$	$0.936_{(0.008)}$	$0.656_{(0.006)}$	$0.925_{(0.013)}$	$0.878_{(0.016)}$	$0.819_{(0.020)}$	$0.723_{(0.010)}$
$\text{GROVER}_{\text{large}}$	$\mathbf{0.940}_{(0.019)}$	$\mathbf{0.658}_{(0.023)}$	$\mathbf{0.944}_{(0.021)}$	$\mathbf{0.894}_{(0.028)}$	$\mathbf{0.831}_{(0.025)}$	$\mathbf{0.737}_{(0.010)}$

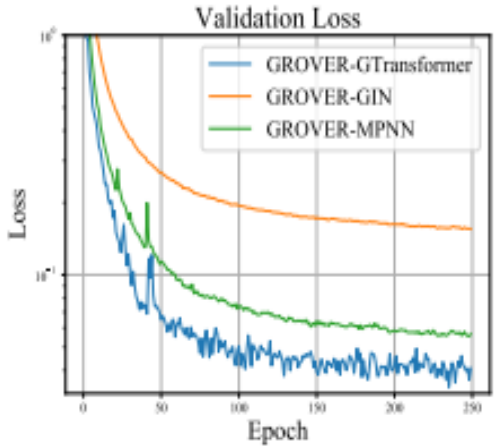
Regression (Lower is better)					
Dataset # Molecules	FreeSolv 642	ESOL 1128	Lipo 4200	QM7 6830	QM8 21786
TF_Robust [40]	$4.122_{(0.085)}$	$1.722_{(0.038)}$	$0.909_{(0.060)}$	$120.6_{(9.6)}$	$0.024_{(0.001)}$
GraphConv [24]	$2.900_{(0.135)}$	$1.068_{(0.050)}$	$0.712_{(0.049)}$	$118.9_{(20.2)}$	$0.021_{(0.001)}$
Weave [23]	$2.398_{(0.250)}$	$1.158_{(0.055)}$	$0.813_{(0.042)}$	$94.7_{(2.7)}$	$0.022_{(0.001)}$
SchNet [45]	$3.215_{(0.755)}$	$1.045_{(0.064)}$	$0.909_{(0.098)}$	$74.2_{(6.0)}$	$0.020_{(0.002)}$
MPNN [13]	$2.185_{(0.952)}$	$1.167_{(0.430)}$	$0.672_{(0.051)}$	$113.0_{(17.2)}$	$0.015_{(0.002)}$
DMPNN [63]	$2.177_{(0.914)}$	$0.980_{(0.258)}$	$0.653_{(0.046)}$	$105.8_{(13.2)}$	$0.0143_{(0.002)}$
MGCN [30]	$3.349_{(0.097)}$	$1.266_{(0.147)}$	$1.113_{(0.041)}$	$77.6_{(4.7)}$	$0.022_{(0.002)}$
AttentiveFP [61]	$2.030_{(0.420)}$	$0.853_{(0.060)}$	$0.650_{(0.030)}$	$126.7_{(4.0)}$	$0.0282_{(0.001)}$
N-GRAM [29]	$2.512_{(0.190)}$	$1.100_{(0.160)}$	$0.876_{(0.033)}$	$125.6_{(1.5)}$	$0.0320_{(0.003)}$
$\text{GROVER}_{\text{base}}$	$1.592_{(0.072)}$	$0.888_{(0.116)}$	$0.563_{(0.030)}$	$\mathbf{72.5}_{(5.9)}$	$0.0172_{(0.002)}$
$\text{GROVER}_{\text{large}}$	$\mathbf{1.544}_{(0.397)}$	$\mathbf{0.831}_{(0.120)}$	$\mathbf{0.560}_{(0.035)}$	$72.6_{(3.8)}$	$\mathbf{0.0125}_{(0.002)}$

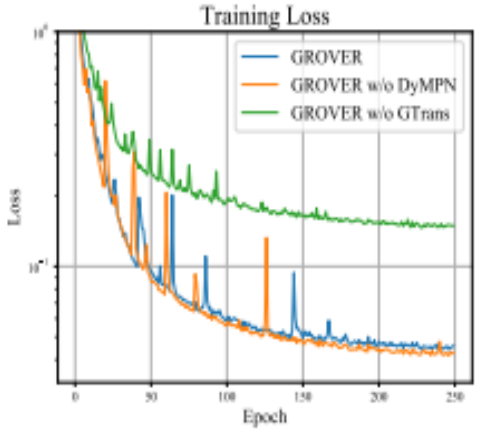
Ablation study

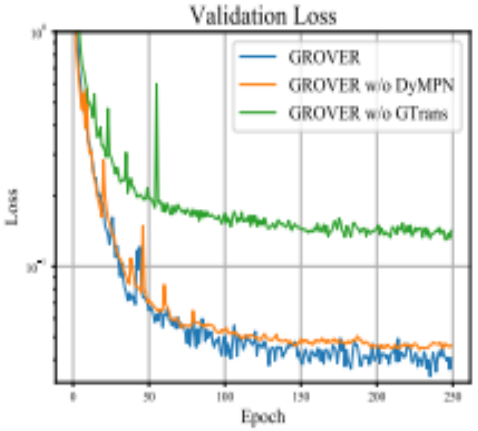
- Pre-training
  - an average AUC increase of 3.8%
- GTransformer Backbone
  - GIN and MPNN
  - toy data set with 600K molecules
- Effect of dyMPN and GTransformer
  - GROVER w/o dyMPN
  - GROVER w/o GTrans

	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









Thank You