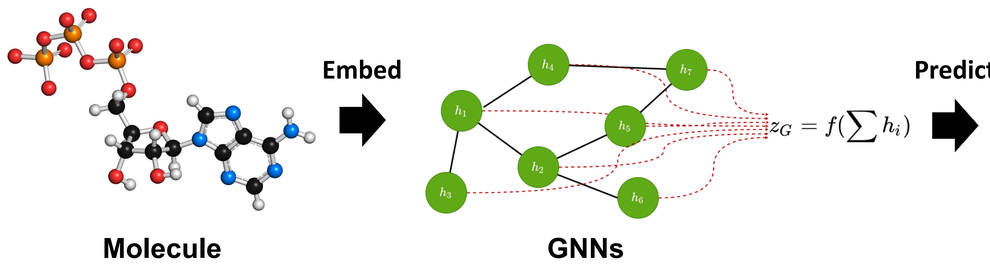
THRUST 1

Bi-Level Contrastive Learning for Knowledge-Enhanced Molecule Representations

Pengcheng Jiang,[1] Cao Xiao,[2] Tianfan Fu,[3] Jimeng Sun,[1] and Jiawei Han[1]
[1]Department of Computer Science, University of Illinois Urbana-Champaign; [2]GE HealthCare; [3]Rensselaer Polytechnic Institute

Motivation

Most previous approaches use graph neural networks (GNNs) to only learn the structure information in molecules, to predict molecule's property (e.g., BBBP, toxicity).



BBBP:
- "Can it cross the blood-brain barrier?"

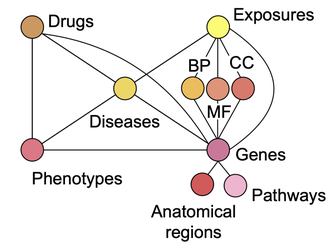
Toxicity:
- "Is it toxic in high doses?"

...

However, publicly available biochemical knowlegde bases remained largely unused (e.g., PubChem, PrimeKG) for this task.

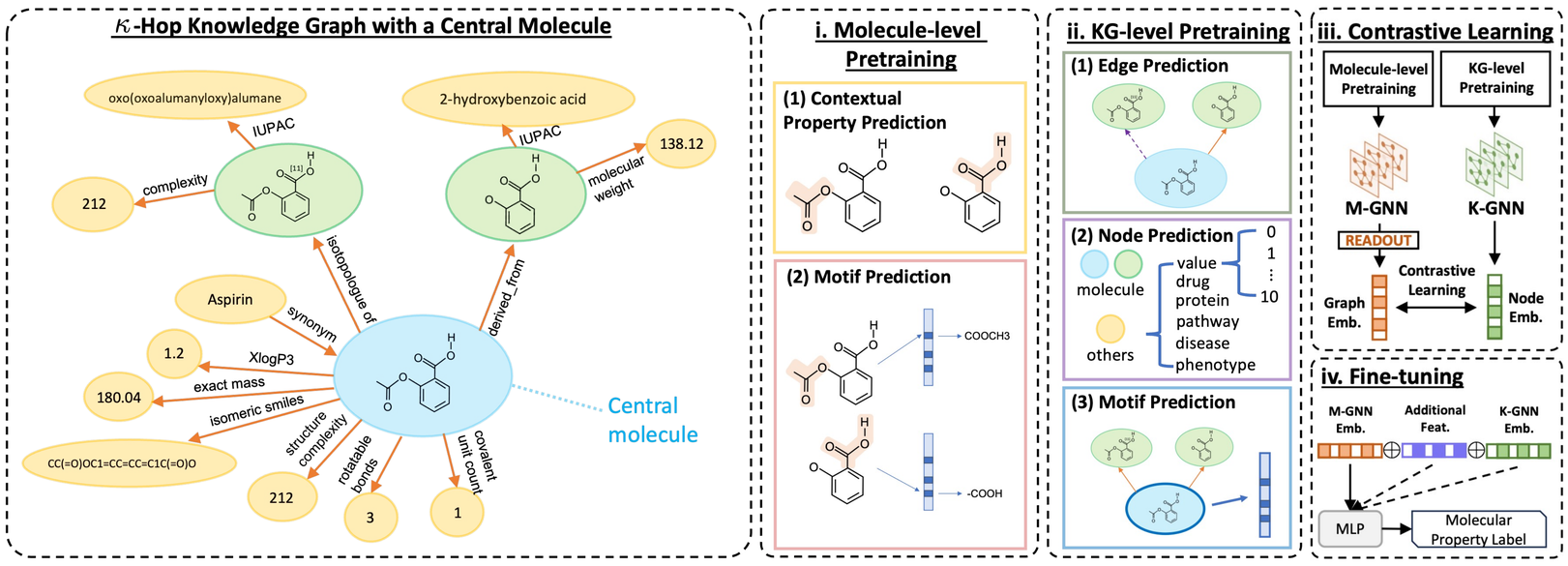
PubChem

PrimeKG



"Can we leverage the available biochemical knowledge for molecular property prediction?"

Method: Gode (Graph as a Node) Framework



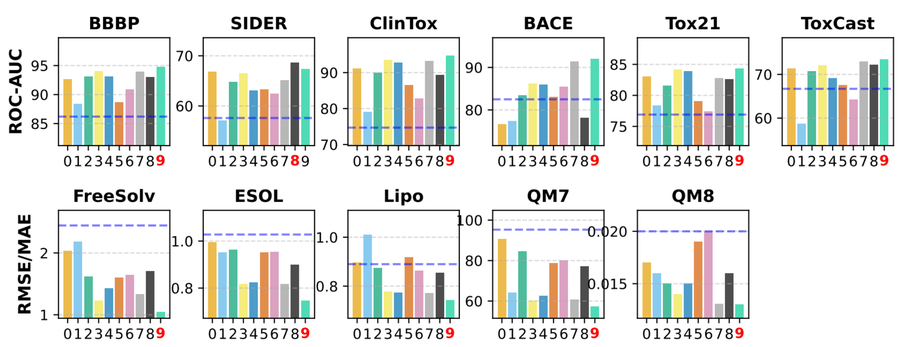
- ❑ We construct MolKG – a molecule-centric biochemical knowledge graph dataset
  - ❖ Each molecule corresponds to its central sub-graph in MolKG
- ❑ We conduct GNN pre-training at two levels: molecule-level and KG-level
  - ❖ At molecule-level, a molecule is a graph
  - ❖ At KG-level, a molecule is a node
- ❑ We align bi-level molecule representations via contrastive learning
- ❑ The pre-trained and aligned molecule embedding can then be fine-tuned to any downstream tasks

Results across 11 Molecular Property Prediction Tasks:

	Classification (Higher is Better)						Regression (Lower is Better)				
Dataset	**BBBP**	**SIDER**	**ClinTox**	**BACE**	**Tox21**	**ToxCast**	**FreeSolv**	**ESOL**	**Lipophilicity**	**QM7**	**QM8**
# Molecules	2039	1427	1478	1513	7831	8575	642	1128	4200	6830	21786
# Tasks	1	27	2	1	12	617	1	1	1	1	12
GCN	71.8 ± 0.9	53.6 ± 0.3	62.5 ± 2.8	71.6 ± 2.0	70.9 ± 0.3	65.0 ± 6.1	2.870 ± 0.140	1.430 ± 0.050	0.712 ± 0.049	122.9 ± 2.2	0.037 ± 0.001
GIN	65.8 ± 4.5	57.3 ± 1.6	58.0 ± 4.4	70.1 ± 5.4	74.0 ± 0.8	66.7 ± 1.5	2.765 ± 0.180	1.452 ± 0.020	0.850 ± 0.071	124.8 ± 0.7	0.037 ± 0.001
SchNet	84.8 ± 2.2	54.5 ± 3.8	71.7 ± 4.2	76.6 ± 1.1	76.6 ± 2.5	67.9 ± 2.1	3.215 ± 0.755	1.045 ± 0.064	0.909 ± 0.098	74.2 ± 6.0	0.020 ± 0.002
MPNN	91.3 ± 4.1	59.5 ± 3.0	87.9 ± 5.4	81.5 ± 4.4	80.8 ± 2.4	69.1 ± 1.3	1.621 ± 0.952	1.167 ± 0.430	**0.672 ± 0.051**	111.4 ± 0.9	**0.015 ± 0.001**
DMPNN	91.9 ± 3.0	63.2 ± 2.3	89.7 ± 4.0	85.2 ± 5.3	**82.6 ± 2.3**	71.8 ± 1.1	1.673 ± 0.082	1.050 ± 0.008	0.683 ± 0.016	103.5 ± 8.6	**0.016 ± 0.001**
MGCN	85.0 ± 6.4	55.2 ± 1.8	63.4 ± 4.2	73.4 ± 3.0	70.7 ± 1.6	66.3 ± 0.9	3.349 ± 0.097	1.266 ± 0.147	1.113 ± 0.041	77.6 ± 4.7	0.022 ± 0.002
N-GRAM	91.2 ± 1.3	63.2 ± 0.5	85.5 ± 3.7	87.6 ± 3.5	76.9 ± 2.7	-	2.512 ± 0.190	1.100 ± 0.160	0.876 ± 0.033	125.6 ± 1.5	0.032 ± 0.003
HU. et.al	70.8 ± 1.5	62.7 ± 0.8	72.6 ± 1.5	84.5 ± 0.7	78.7 ± 0.4	65.7 ± 0.6	2.764 ± 0.002	1.100 ± 0.006	0.739 ± 0.003	113.2 ± 0.6	0.022 ± 0.001
$\text{GROVER}_{\text{Large, GTrans}}$	86.2 ± 3.9	57.6 ± 1.6	74.7 ± 4.4	82.5 ± 4.4	76.9 ± 2.3	66.7 ± 2.6	2.445 ± 0.761	1.028 ± 0.145	0.890 ± 0.050	95.3 ± 5.6	0.020 ± 0.003
MGSSL	70.5 ± 1.1	64.1 ± 0.7	80.7 ± 2.1	79.7 ± 0.8	76.4 ± 0.4	64.1 ± 0.7	-	-	-	-	-
MolCLR	73.3 ± 1.0	61.2 ± 3.6	89.8 ± 2.7	82.8 ± 0.7	74.1 ± 5.3	65.9 ± 2.1	2.301 ± 0.247	1.113 ± 0.023	0.789 ± 0.009	90.0 ± 1.7	0.019 ± 0.013
$\text{MolCLR}_{\text{GTrans}}$	76.7 ± 2.2	63.3 ± 2.5	89.3 ± 3.1	87.7 ± 1.8	80.2 ± 3.2	70.4 ± 2.1	2.124 ± 0.223	0.982 ± 0.109	0.767 ± 0.064	88.9 ± 4.8	0.018 ± 0.002
$\text{KGE\_NFM}_{\text{w/ MolKG}}$	92.4 ± 2.4	**65.3 ± 1.4**	87.3 ± 2.0	78.1 ± 2.1	79.8 ± 3.3	**72.6 ± 1.8**	1.942 ± 0.441	1.027 ± 0.201	0.877 ± 0.071	87.6 ± 3.2	**0.016 ± 0.001**
$\text{KANO}_{\text{CMPNN}}$	**92.6 ± 1.8**	**65.5 ± 1.6**	**92.9 ± 1.1**	**90.7 ± 3.1**	81.8 ± 1.1	**72.5 ± 1.9**	**1.320 ± 0.244**	**0.902 ± 0.104**	**0.641 ± 0.012**	**66.5 ± 3.7**	**0.013 ± 0.001**
$\text{KANO}_{\text{GTrans}}$	**93.7 ± 2.3**	63.8 ± 1.2	**93.6 ± 0.7**	**90.4 ± 1.5**	**81.2 ± 1.8**	**72.5 ± 1.5**	**1.443 ± 0.315**	**0.914 ± 0.092**	**0.651 ± 0.018**	**63.6 ± 4.1**	**0.013 ± 0.002**
GODE (ours)	**94.8 ± 1.9**	**67.4 ± 1.4**	**94.7 ± 2.9**	**92.0 ± 2.2**	**84.3 ± 1.2**	**73.4 ± 0.9**	**1.048 ± 0.314**	**0.746 ± 0.128**	0.743 ± 0.043	**57.2 ± 3.0**	**0.013 ± 0.001**

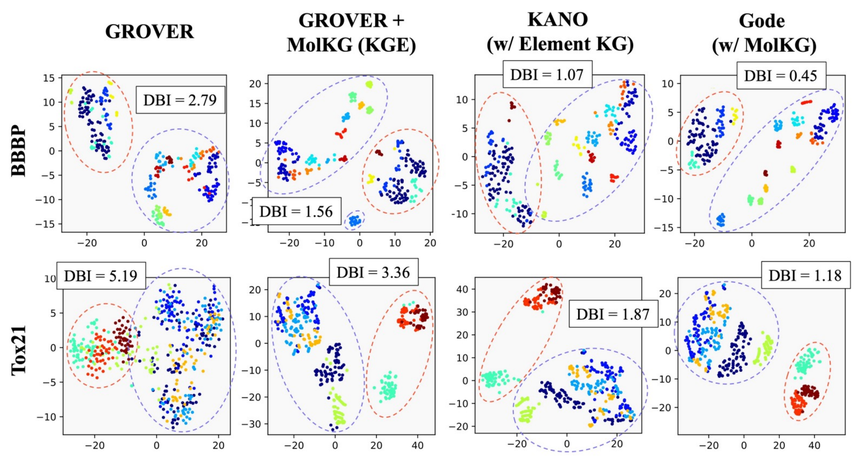
→ **Gode** achieves state-of-the-art performance on 10/11 tasks, outperforming all the previous baselines (including KANO, which leverages chemical element knowledge)

Case	KGE	κ	Pret.	Cont.	Embedding
0	✓	–	✗	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KGE}}$
1	✗	2	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$
2	✓	2	✓	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$
3	✓	2	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$
4	✓	2	✓	✓	$\mathbf{h}_{\text{MG}}$
5	✓	3	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{KG}}$
6	✓	3	✓	✓	$\mathbf{h}_{\text{MG}}$
7	✓	2	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{f}}$
8	✓	–	✗	✗	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{f}} \oplus \mathbf{h}_{\text{KGE}}$
9	✓	2	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{f}} \oplus \mathbf{h}_{\text{KG}}$



→ We conduct extensive ablation studies to show the effectiveness of each component in Gode.

Visualization



The learned **Gode** embedding has the most distinct clusters for different scaffolds (shown as different colors).
→ Highest representational power

Future Directions

- ❑ Applications in Drug Discovery
- ❑ Inclusion of Generative AI's power

Paper: https://pat-jj.github.io/assets/pdf/gode.pdf
Code: https://github.com/pat-jj/Gode

This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No.2019897.

MOLECULE MAKER LAB INSTITUTE

Collaborating Institutions

ILLINOIS