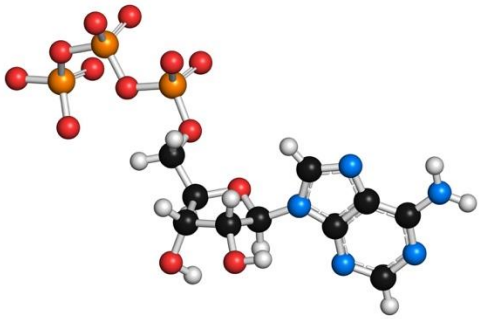
Bi-Level Contrastive Learning for Knowledge-Enhanced Molecule Representations

Pengcheng Jiang, Cao Xiao, Tianfan Fu, Taha Kass-Hout, Jimeng Sun, and Jiawei Han

Overview

- Background
- Motivation
- Methodology
- Experiment
- Conclusion

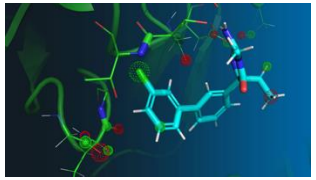
Background

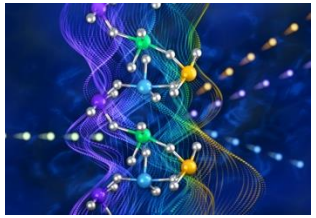


Task: Molecular Property Prediction

- Definition: Molecular property prediction involves using computational methods and machine learning techniques to estimate various properties of molecules, such as stability, reactivity, solubility, and biological activity.

- Importance:
  - Accelerated Drug Discovery: By predicting properties like toxicity and efficacy, researchers can identify promising drug candidates early, reducing the time and cost associated with experimental testing.
  - Material Science Advancement: Predicting properties such as conductivity or durability aids in designing materials with desired characteristics for various industrial applications.
  - Environmental Impact Assessment …

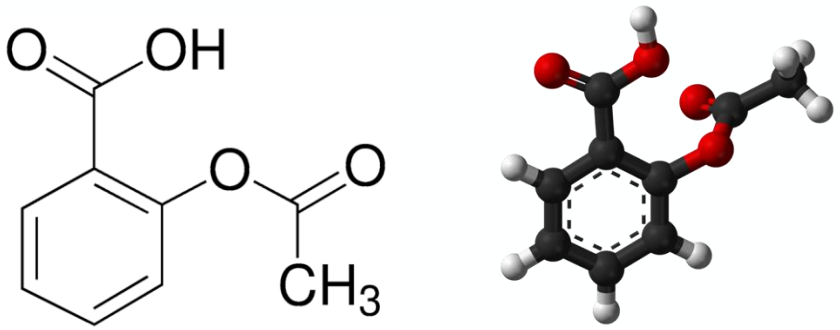




Background

Task: Molecular Property Prediction

Example:



Aspirin (Acetylsalicylic Acid)

Blood-Brain Barrier Penetration (BBBP) task:
- Can it cross the blood-brain barrier?

Toxicity prediction tasks:
- Is it toxic in high doses?

Solubility prediction:
- How soluble is it in water?

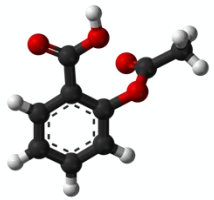
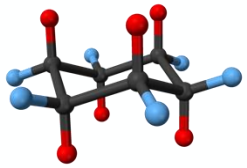
...

Background

Traditional Methods

Simplified Molecular Input Line Entry System (**SMILES**)

Example:

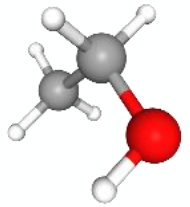
		SMILES String
Aspirin		O=C(C)Oc1ccccc1C(=O)O
Cyclohexane		C1CCCCC1

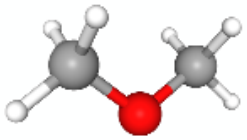
Early approach (linear modeling):

SMILES
➔ one-hot encoding
➔ feature vector

Main limitation:

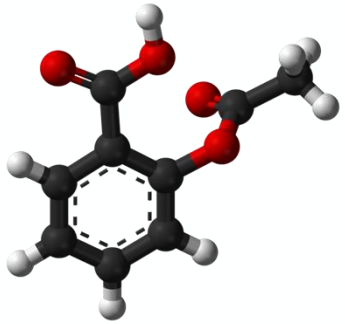
Do not capture structural information (e.g., isomers CH_3CH_2OH and CH_3OCH_3)



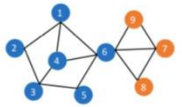
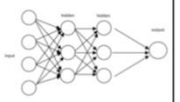
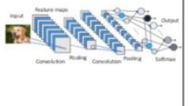
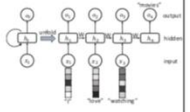
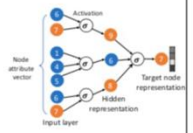


Background

GNN-Based Methods



Molecules are graphs
- Atoms as nodes and bonds as edges
- Capturing both the connectivity and spatial relationships

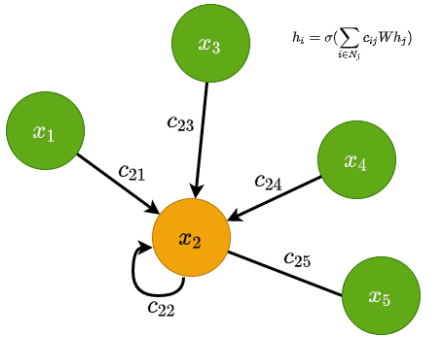
Data type	Multi-dimensional	Grid	Sequence	Graph
	Features: credit rating, account balance $x = (4.5, 500, 3, 5)$ #deposits, #withdraws		x = "I love watching movies."	
DL Architecture	Feed-forward Network	CNN	RNN	GNN
				

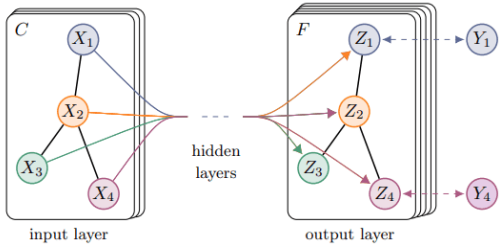
With advances in deep learning, **Graph Neural Networks (GNNs)** have become a prominent tool for learning molecular representations.

GNNs learn by aggregating information from a node's neighbors, making them well-suited to model interactions within molecular structures

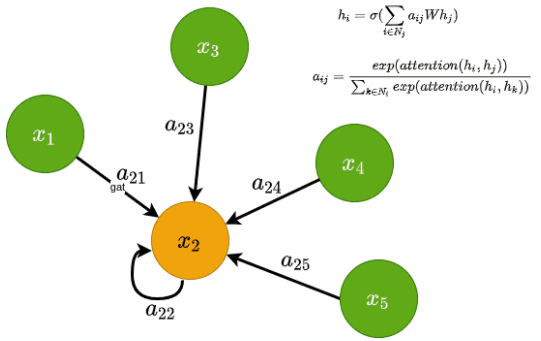
Background

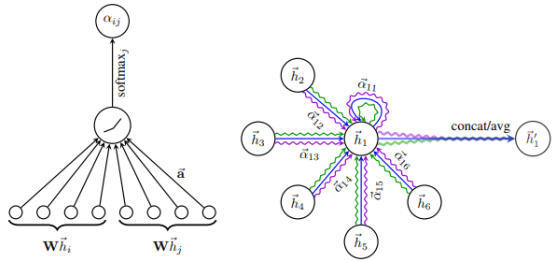
GNN-Based Methods



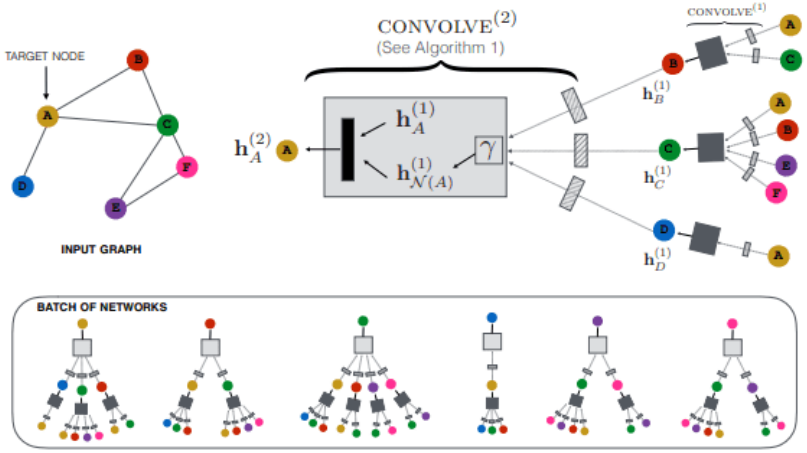


Graph Convolutional Networks (GCN)





Graph Attention Networks (GAT)

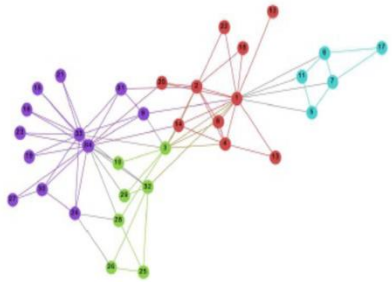


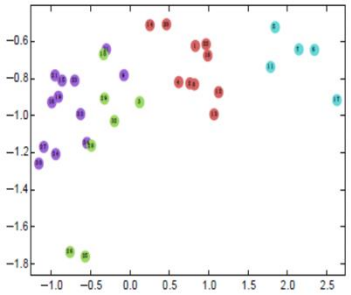
GraphSAGE

...

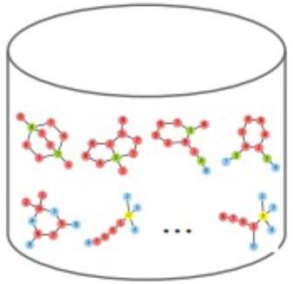
Background

GNN-Based Methods

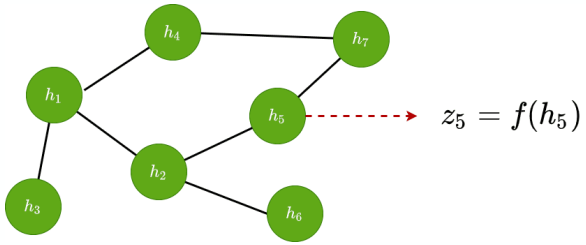




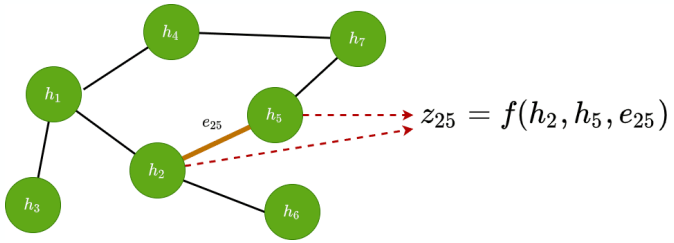
Node Representation



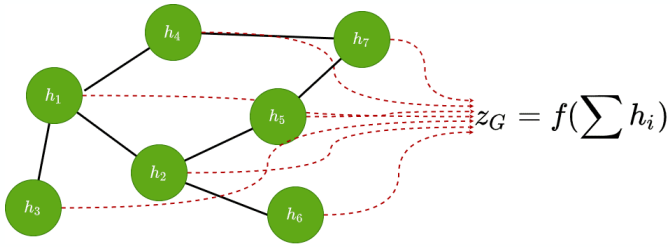
Graph Representation



Node Classification



Edge Classification



Graph Classification

Background

GNN-Based Methods – Pre-training

Researchers (Hu et al., ICLR 2020) found that GNN pre-training before fine-tuning effectively improves the performance of downstream tasks.

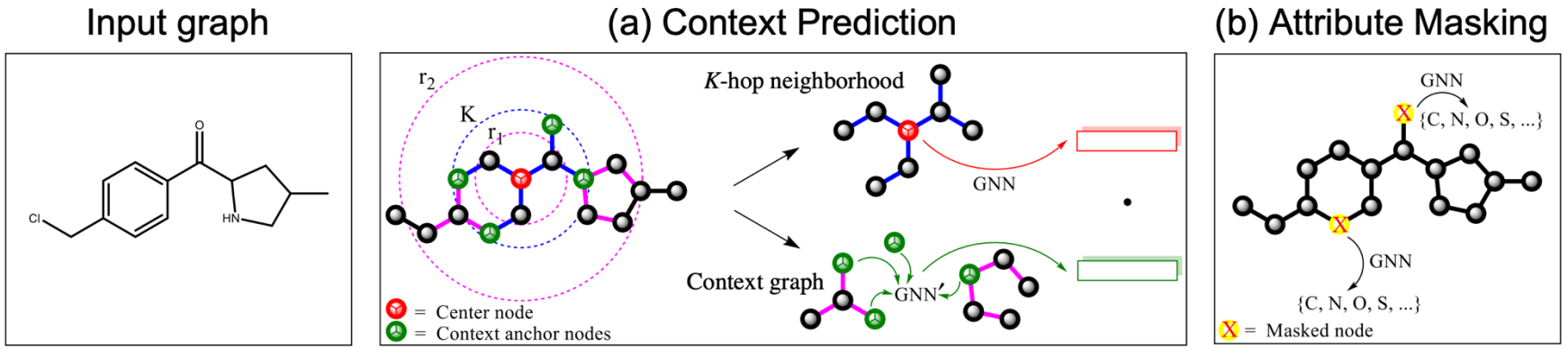
	Chemistry		
	Non-pre-trained	Pre-trained	Gain
GIN	67.0	**74.2**	**+7.2**
GCN	**68.9**	72.2	+3.4
GraphSAGE	68.3	70.3	+2.0
GAT	66.8	60.3	-6.5

Hu, Weihua, et al. "Strategies for pre-training graph neural networks." in ICLR 2020.

Background

GNN-Based Methods – Pre-training

Researchers (Hu et al., ICLR 2020) found that GNN pre-training before fine-tuning effectively improves the performance of downstream tasks.

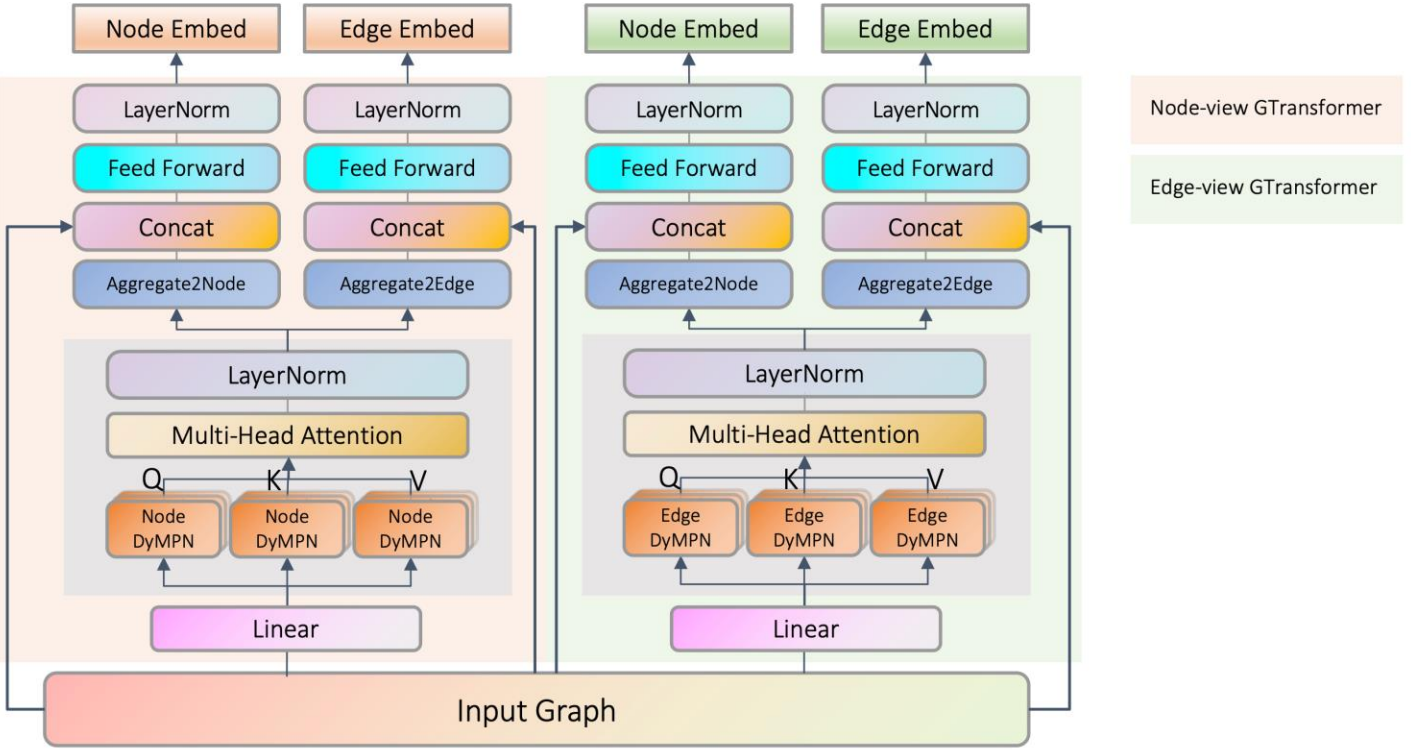


Node-level pre-training strategy

Hu, Weihua, et al. "Strategies for pre-training graph neural networks." in ICLR 2020.

Background

GNN-Based Methods – Pre-training

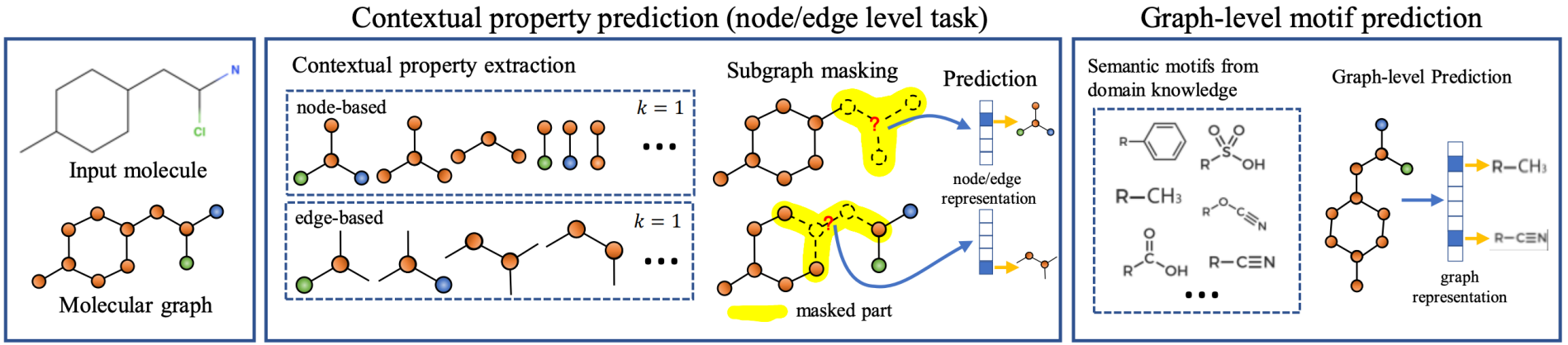


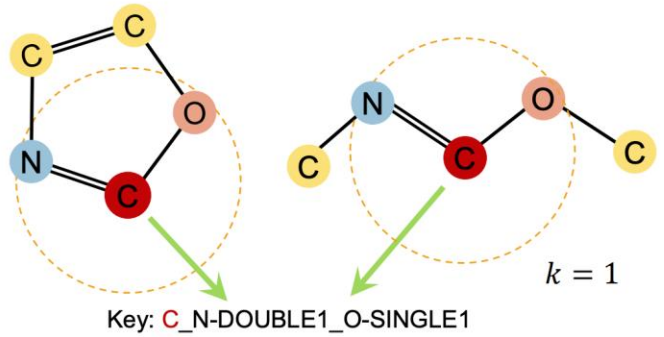
GTransformer by GROVER (Rong et al, NeurIPS 2020)

Rong, Yu, et al. "Self-supervised graph transformer on large-scale molecular data." in NeurIPS 2020

Background

GNN-Based Methods – Pre-training



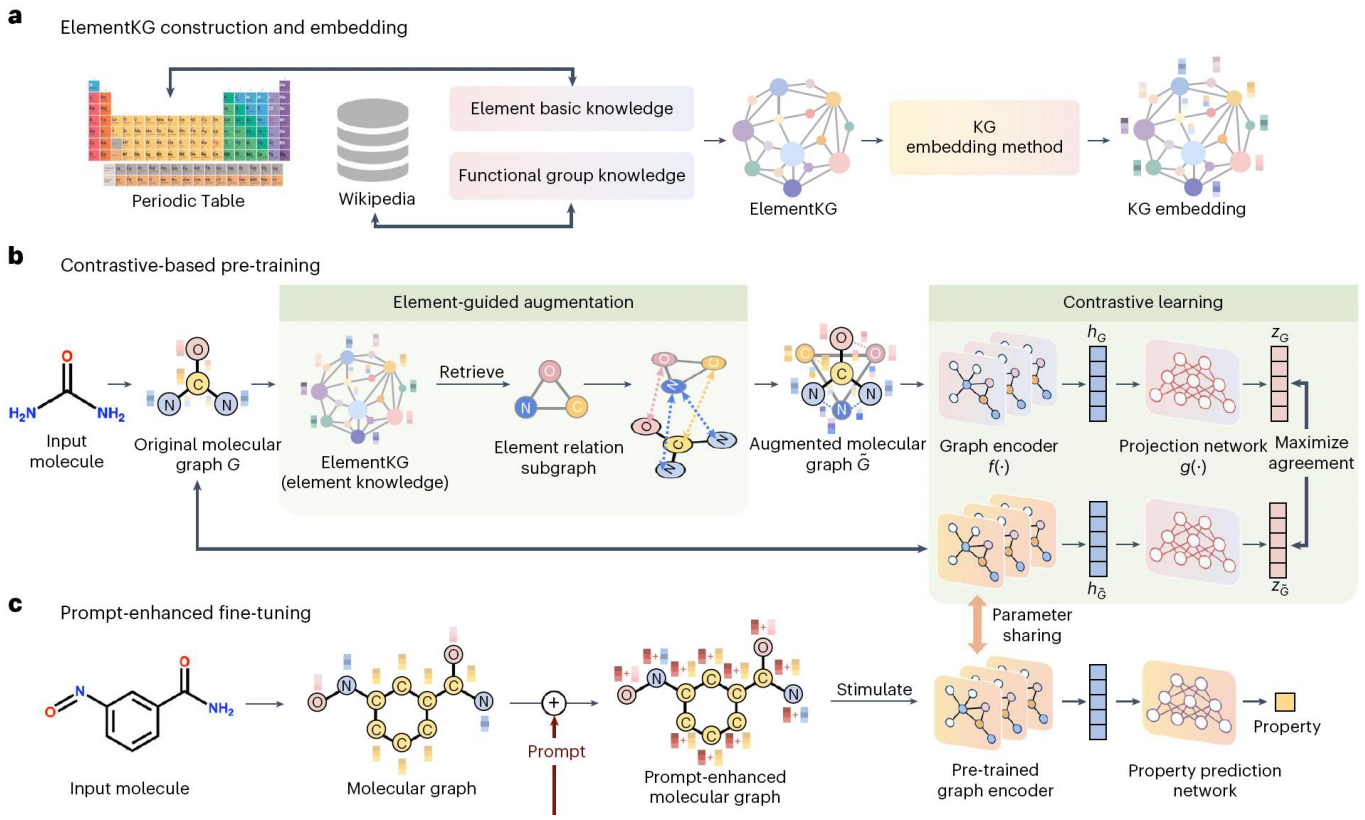


(Contextual property: the node-edge counts terms in alphabetical order)

Rong, Yu, et al. "Self-supervised graph transformer on large-scale molecular data." in NeurIPS 2020

Background

Knowledge-Enhanced Methods



KANO (Fang et al., 2023)

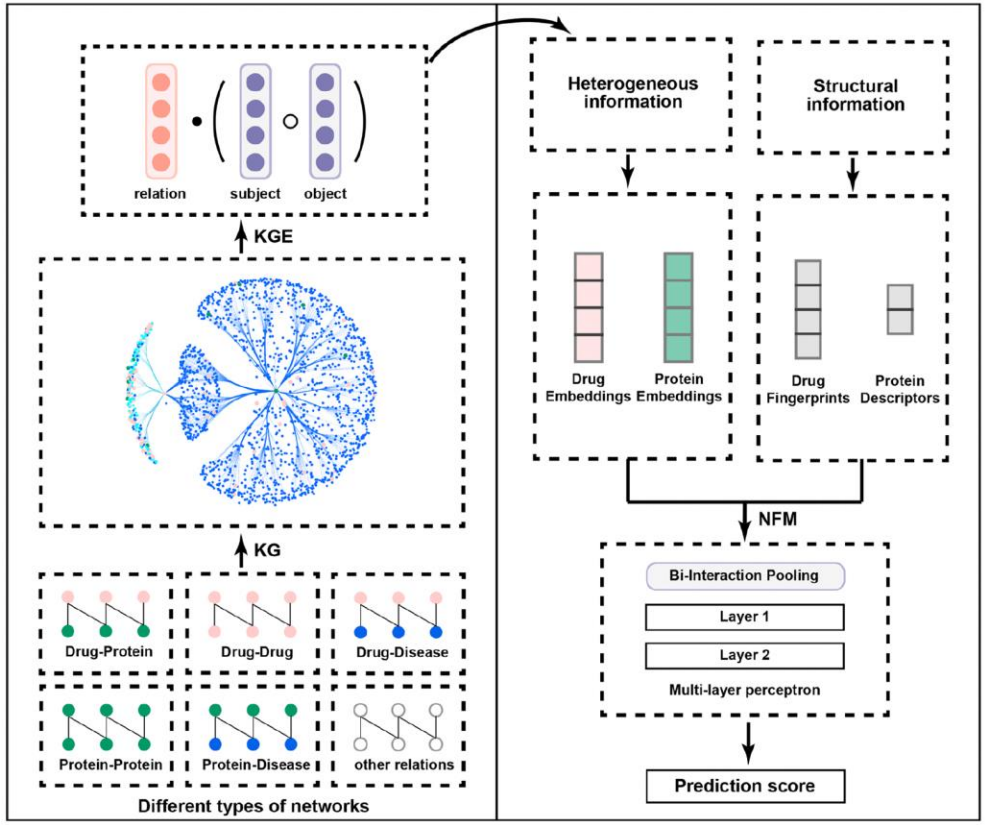
1. Build ElementKG, which encompasses the knowledge of chemical elements.

2. Augment the molecule with ElementKG, and transfer the knowledge to the molecule representation with contrastive learning.

Fang, Yin, et al. "Knowledge graph-enhanced molecular contrastive learning with functional prompt." *Nature Machine Intelligence*

Background

Knowledge-Enhanced Methods



KGE NFM (Ye et al., 2021)

Concatenate static KG embedding to the protein embedding to predict drug-target interaction.

(Also can be used for molecular property prediction tasks)

Ye, Qing, et al. "A unified drug–target interaction prediction framework based on knowledge graph and recommendation system." *Nature communications*.

Motivation

1. Publicly available biochemical knowlegde bases remained largely unused for the molecule property prediction task.

PubChem

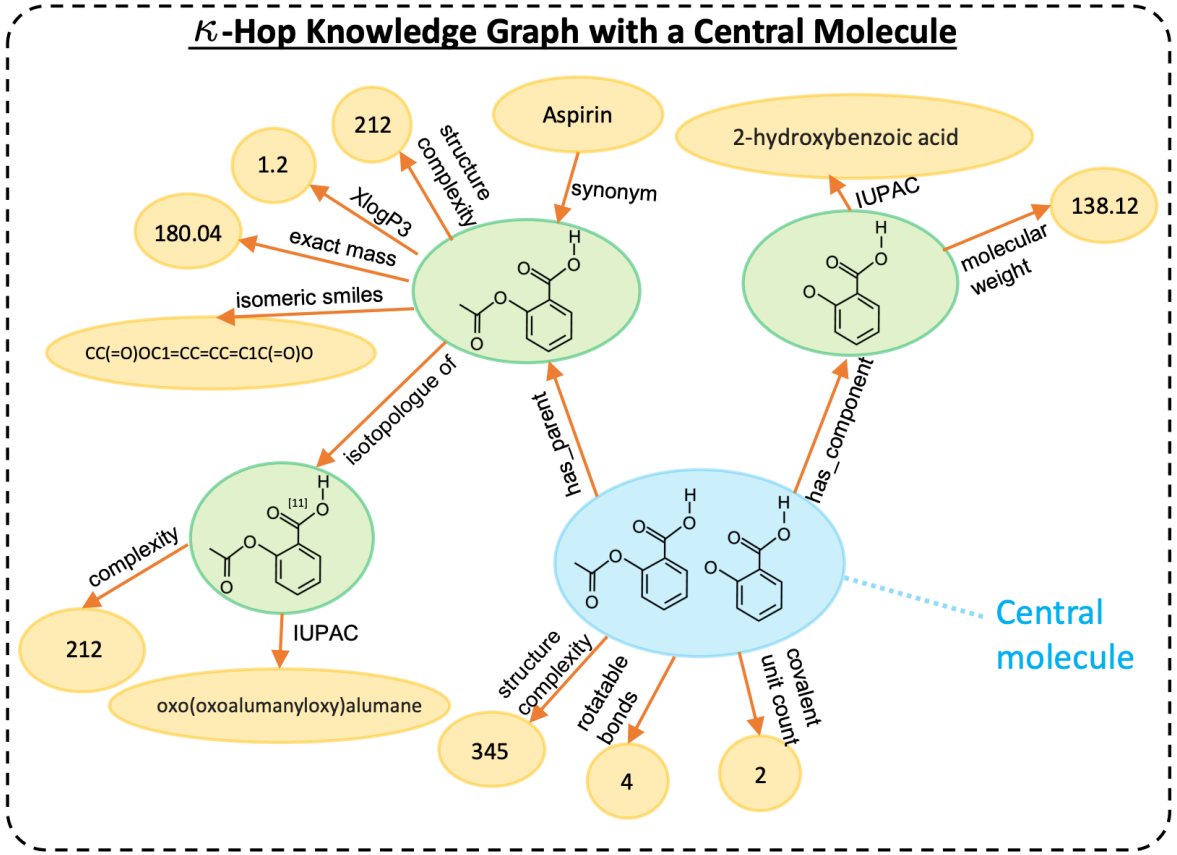
2. No framework/method is integrating such knowledge into molecular property task.

Motivation

In the biochemical databases, a molecule can be associated with many entities of different types.

For example:

1. Statistical properties (structure complexity, mass, rotatable bonds, covalent unit count …)
2. Drug (Medication)
3. Protein
4. Pathway
5. Disease
6. Phenotype
7. Other molecules

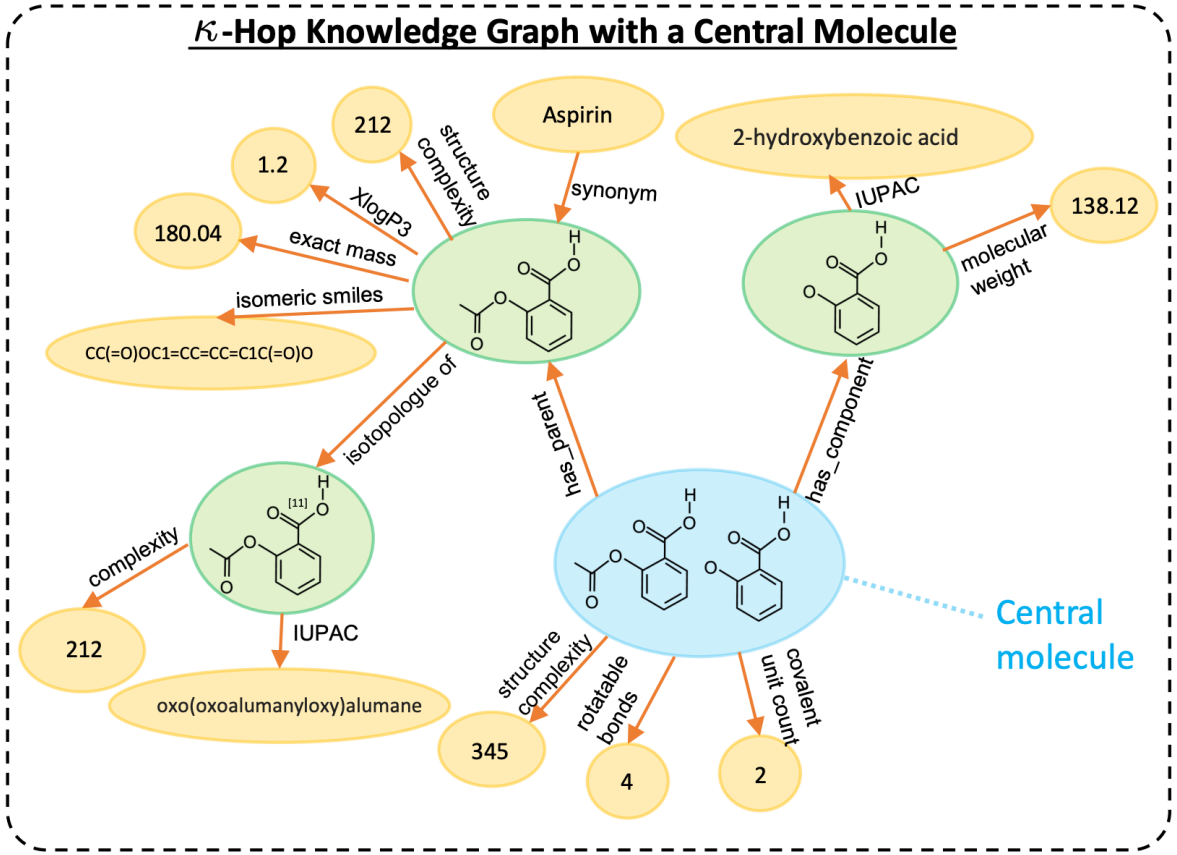


Motivation

In the biochemical databases, a molecule can be associated with many entities of different types.

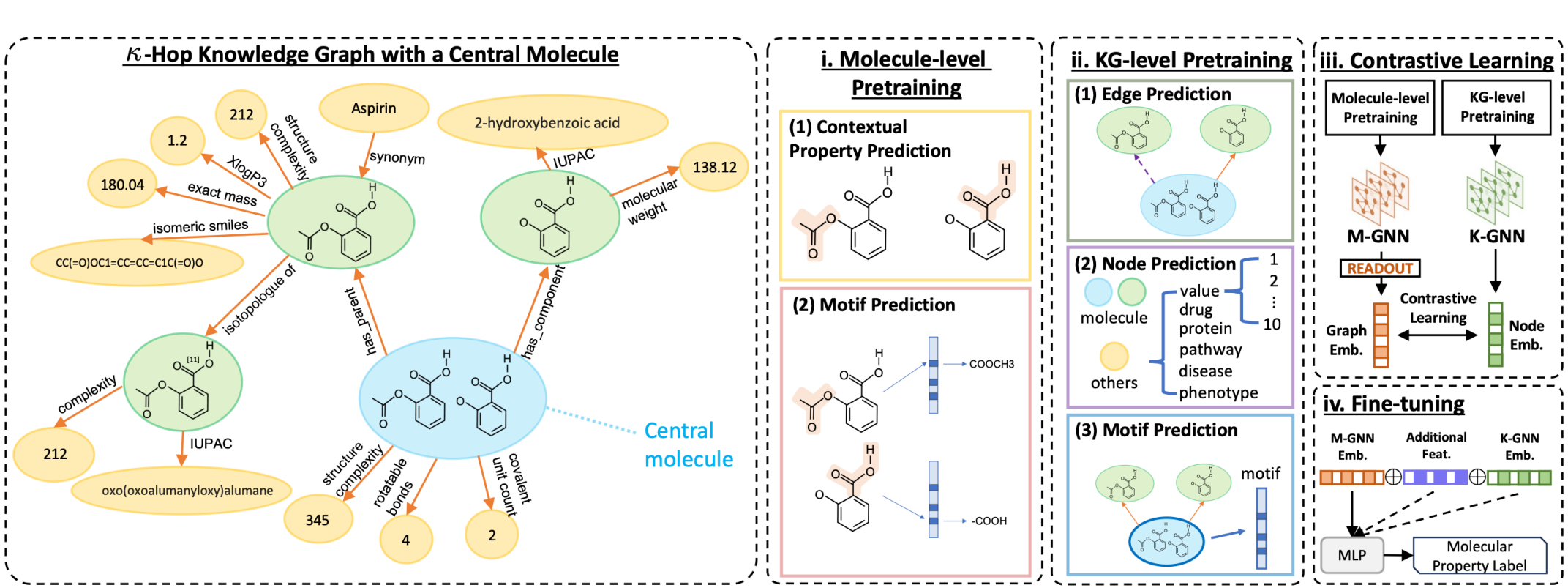
For example:

1. Statistical properties (structure complexity, mass, rotatable bonds, covalent unit count …)
2. Drug (Medication)
3. Protein
4. Pathway
5. Disease
6. Phenotype
7. Other molecules



Methodology

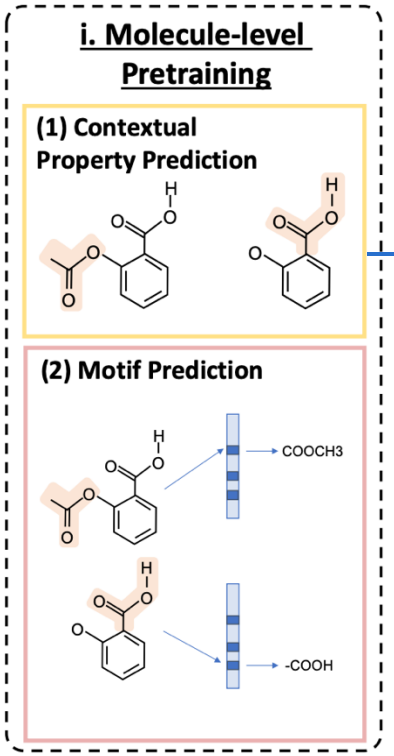
Proposed method: **Gode** (**G**raph as a N**ode**)



Methodology

Step 1: Molecule-Level Pre-training / M-GNN (Molecule Graph Neural Network) Pre-training

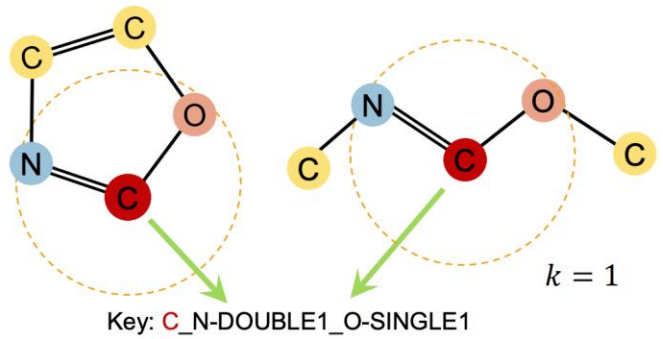
- M-GNN is a graph encoder encoding a molecule *MG* into a vector $\mathbf{h}_{MG}$
- We follow GROVER's pre-training strategy for M-GNN.



Node-level Contextual Property Prediction

- We randomly select a node (atom) v in the molecule, and use its embedding $\mathbf{h}_v$ to predict its contextual property. (Multi-class Classification)

$$\sum_{v}^{\mathcal{V}'_m} \log P(p_v|\mathbf{h}_v)$$

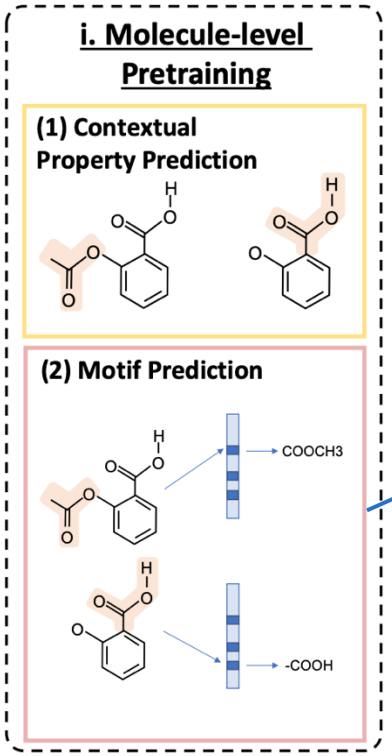


(Contextual property: the node-edge counts terms in alphabetical order)

Methodology

Step 1: Molecule-Level Pre-training / M-GNN (Molecule Graph Neural Network) Pre-training

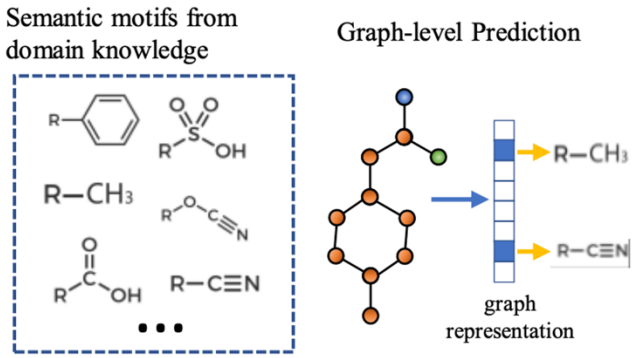
- M-GNN is a graph encoder encoding a molecule *MG* into a vector $\mathbf{h}_{MG}$
- We follow GROVER's pre-training strategy for M-GNN.



Graph-level Motif Prediction

- The molecule graph embedding $\mathbf{h}_{MG}$ is used to predict the presence of functional group motifs in the molecule. (Multi-label Classification)

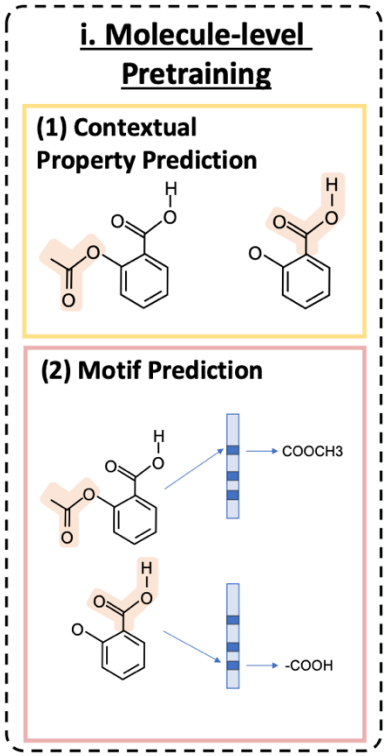
$$\sum_{j=1}^{n} y_j \log P(M_j|\mathbf{h}_{\mathrm{MG}}) + (1-y_j)\log(1-P(M_j|\mathbf{h}_{\mathrm{MG}}))$$



Methodology

Step 1: Molecule-Level Pre-training / M-GNN (Molecule Graph Neural Network) Pre-training

- M-GNN is a graph encoder encoding a molecule *MG* into a vector $\mathbf{h}_{MG}$
- We follow GROVER's pre-training strategy for M-GNN.



Node-level Contextual Property Prediction

- We randomly select a node (atom) v in the molecule, and use its embedding $\mathbf{h}_v$ to predict its contextual property. (Multi-class Classification)

Graph-level Motif Prediction

- The molecule graph embedding $\mathbf{h}_{MG}$ is used to predict the presence of functional group motifs in the molecule. (Multi-label Classification)

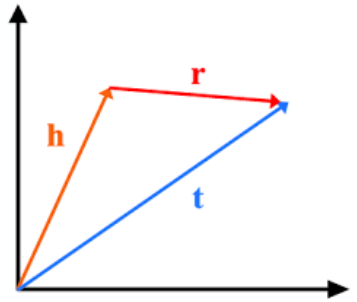
$$\mathcal{L}_{\mathrm{M}} = \sum_{v}^{\mathcal{V}'_m} \log P(p_v|\mathbf{h}_v) + \sum_{j=1}^{n} y_j \log P(M_j|\mathbf{h}_{\mathrm{MG}}) + (1-y_j)\log(1-P(M_j|\mathbf{h}_{\mathrm{MG}}))$$

Methodology

Step 2.1: KG Embedding Initialization & Molecule-Centric Knowledge Graph Extraction

Table 1: **Overview of MolKG**, a biochemical dataset we construct from PubChemRDF and PrimeKG.

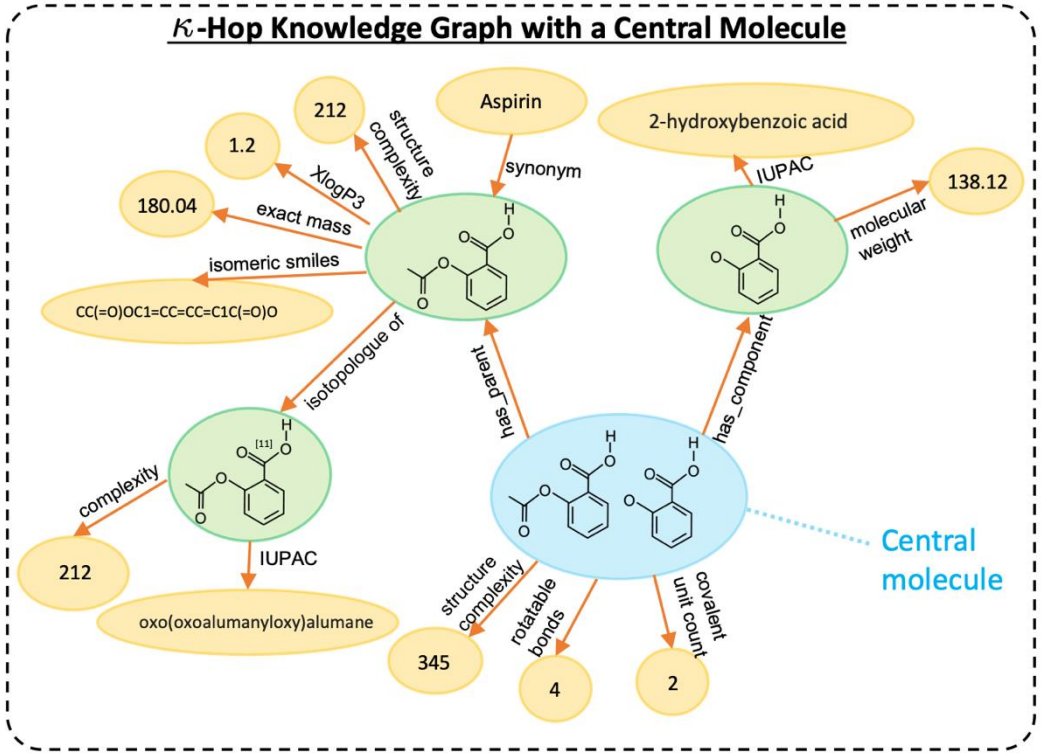
# Triples: 2,523,867 **# Entities**: 184,819 **# Relations**: 39 **# Entity Types**: 7 **# Molecules**: 65,454	
Entity Types	
molecule, gene/protein, disease, effect/phenotype, drug, pathway, value	
Relations	
drug_protein, contraindication, indication, off-label use, drug_drug, drug_effect, defined_bond_stereo_count, tpsa, rotatable_bond_count, xlogp3-aa, structure_complexity, covalent_unit_count, defined_atom_stereo_count, molecular_weight, hydrogen_bond_donor_count, undefined_bond_stereo_count, isotope_atom_count, exact_mass, mono_isotopic_weight, total_formal_charge, hydrogen_bond_acceptor_count, non-hydrogen_atom_count, tautomer_count, undefined_atom_stereo_count, xlogp3, cooccurence_molecule_molecule, cooccurence_molecule_disease, cooccurence_molecule_gene/protein, neighbor_2d, neighbor_3d, has_same_connectivity, has_component, has_isotopologue, has_parent, has_stereoisomer, to_drug, closematch, type, in_pathway	



We train a TransE model on the entire MolKG for the initialization of the node embeddings

Methodology

Step 2.1: KG Embedding Initialization & Molecule-Centric Knowledge Graph Extraction



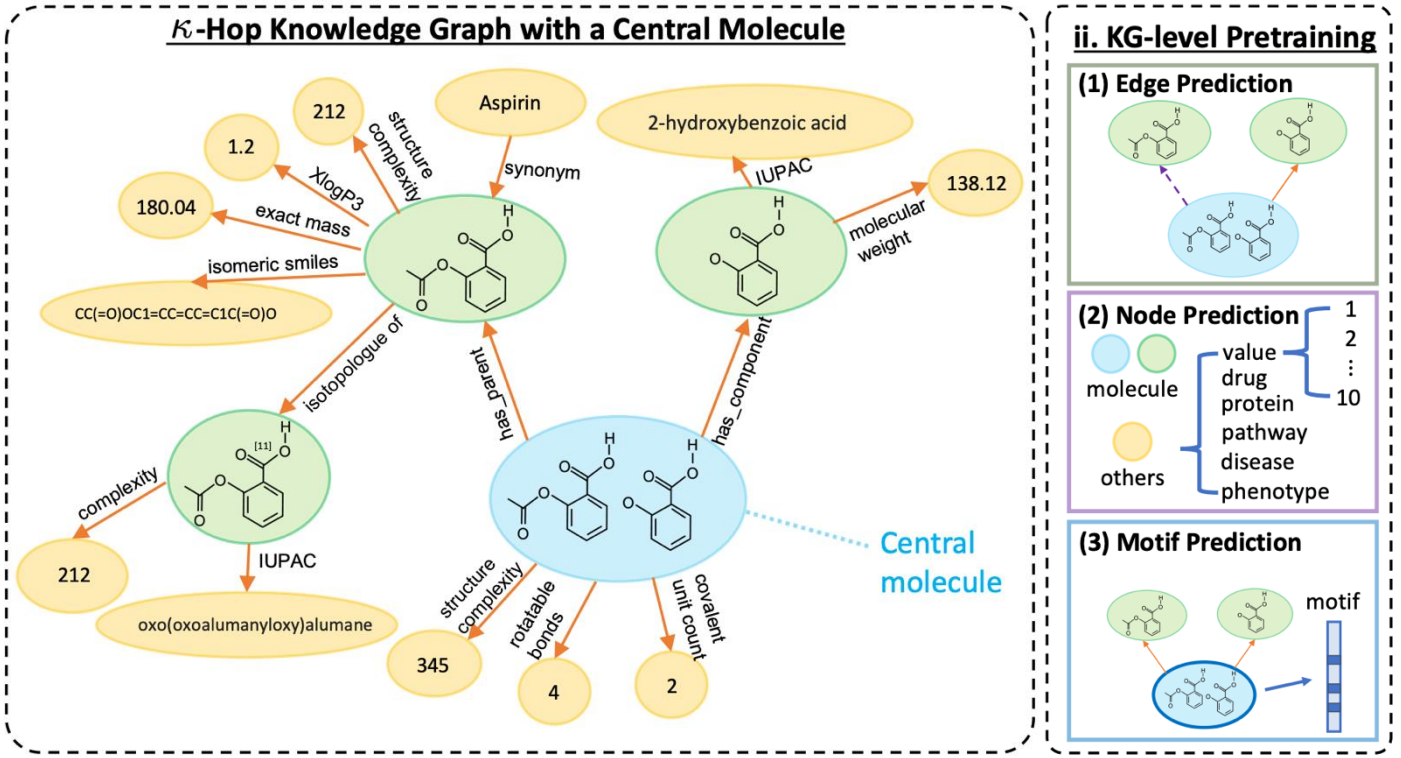
To get the KG data for each molecule, we extract a κ-hop subgraph from the entire KG to capture a its local neighborhood information.

To avoid over-smoothing, we terminate the expansion of a graph branch upon reaching a non-molecule node

Methodology

Step 2.2: Molecule-Centric Knowledge Graph Pretraining / K-GNN Pre-training

- K-GNN is a graph encoder encoding a molecule node into a vector $\mathbf{h}_{KG}$



We conduct the following task to pre-train K-GNN:

(1) Edge Prediction, a multi-class classification task aiming at correctly predicting the edge type between two nodes.

(2) Node Prediction, a multi-class classification task predicting the category of a node in the molecule-centric subgraph.

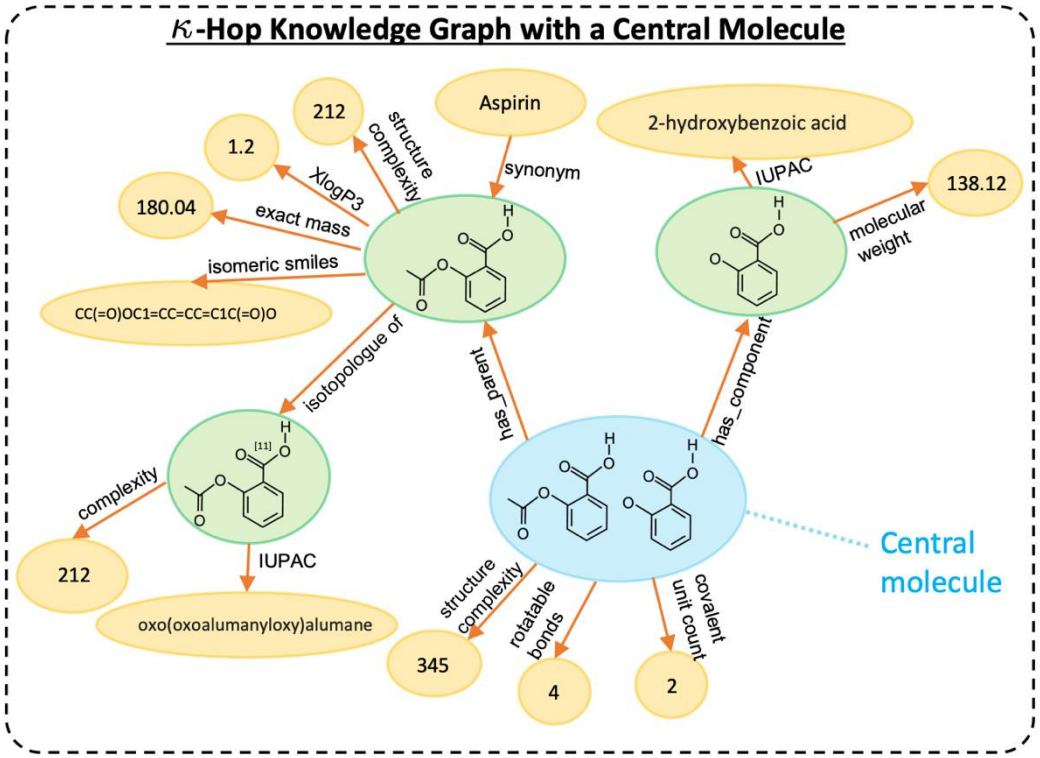
(3) Node-level Motif Prediction, a multi-label classification task predicting the motif of the central molecule node.

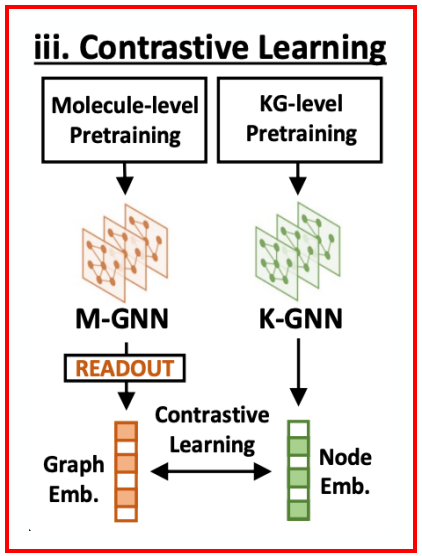
$$\mathcal{L}_{\mathrm{K}} = \lambda_{\mathrm{m}} \underbrace{\sum_{j=1}^{n} \mathrm{BCE}(y_j, P(M_j|\mathbf{h}_c))}_{\text{motif prediction}} + \lambda_{\mathrm{n}} \underbrace{\mathrm{CE}(v', P(v'|\mathbf{h}_v))}_{\text{node prediction}}$$
$$+ \lambda_{\mathrm{e}} \underbrace{\mathrm{CE}((u,v)', P((u,v)'|\mathbf{h}_u \oplus \mathbf{h}_v))}_{\text{edge prediction}} \quad (3)$$

Methodology

Step 3: Contrastive Learning

- We align the same molecule encoded by M-GNN and K-GNN with contrastive learning





To make the task challenging, we divide the negative samples into two groups :

(1) Randomly sampled from all negative molecule-centric KG subgraphs;

(2) Sampled from the sub-graphs of the neighbor molecule nodes connected to the positive molecule node

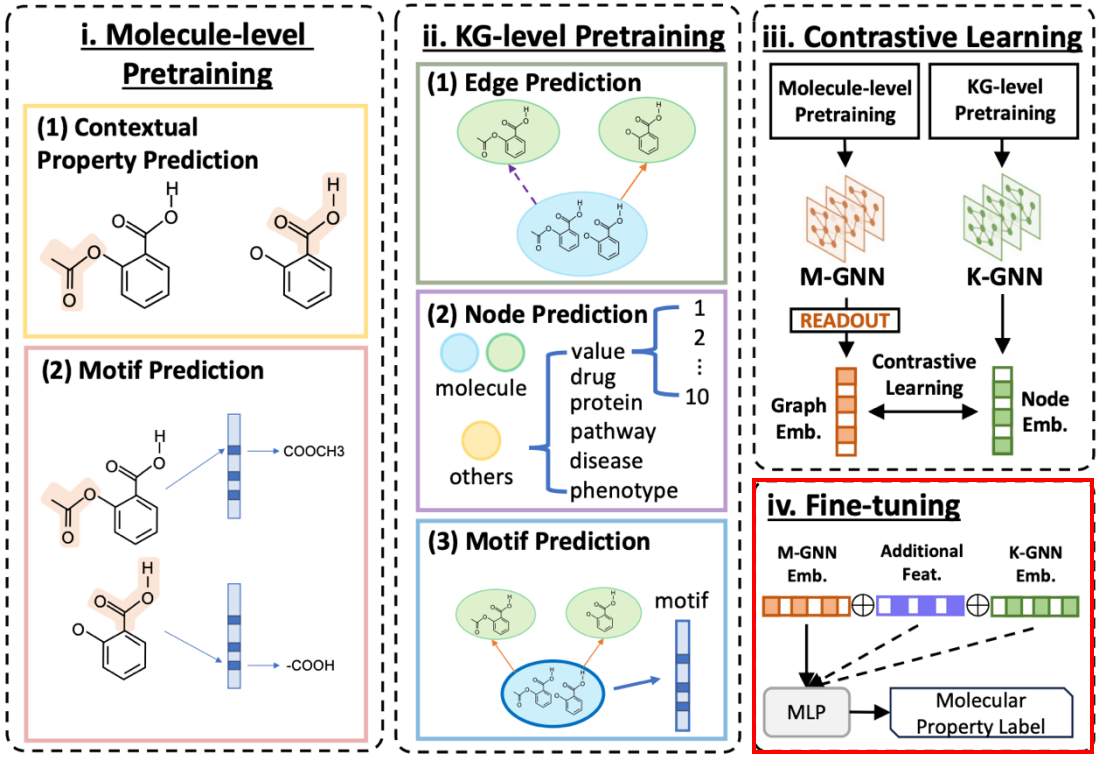
$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\text{sim}(f(m_i), g(s_i))) + (1-y_i)\log(1-\text{sim}(f(m_i), g(s_i)))\Big],$$

$$\text{sim}(f(m_i), g(s_i))) = \frac{\exp(\tau^{-1}\mathbf{h}_{\text{MG}(i)}^{\text{T}}\mathbf{h}_{\text{KG}(i)})}{\exp(\tau^{-1}\mathbf{h}_{\text{MG}(i)}^{\text{T}}\mathbf{h}_{\text{KG}(i)})+1},$$

Methodology

Step 4: Fine-Tuning

- Fine-tune the pre-trained molecule embedding on the downstream tasks!



For each molecule:

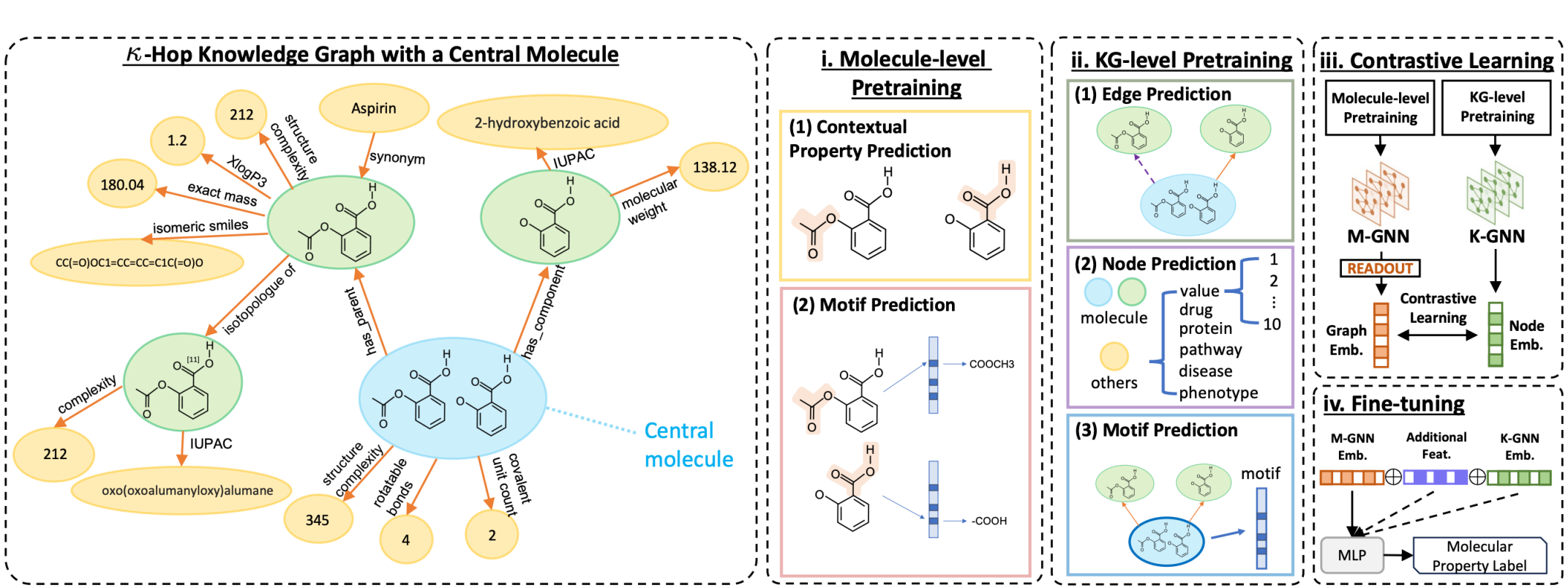
(1) We encode it with the pre-trained M-GNN and K-GNN.

(2) We extract additional molecule-level features (as a multi-hot vector) with RDKit (following previous works).

(3) We concatenate three embeddings into one, and fine-tune it on downstream classification/regression tasks.

Methodology

Our method: **Gode** (**G**raph as a N**ode**)



Experiment - Datasets

Six classification tasks:

Dataset	# Molecules	# Tasks	Description
BBBP (Martins et al. 2012)	2039	1	The Blood-Brain Barrier Penetration (BBBP) dataset aids drug discovery, especially for neurological disorders. It characterizes a compound's ability to cross the blood-brain barrier, influencing treatment efficacy for brain disorders.
SIDER (Kuhn et al. 2016)	1427	27	The Side Effect Resource (SIDER) provides adverse effects data of marketed medications. This is crucial for pharmacovigilance, enabling potential side effects predictions of new compounds based on molecular properties.
ClinTox (Gayvert, Madhukar, and Elemento 2016)	1478	2	ClinTox compares drugs that gained FDA approval versus those rejected due to toxic concerns. This assists researchers in anticipating toxicological profiles of new compounds.
BACE (Subramanian et al. 2016)	1513	1	The BACE dataset offers insights into potential inhibitors for human β-secretase 1 (BACE-1), an enzyme linked to Alzheimer's. It's vital for neurological drug discovery targeting Alzheimer's treatments.
Tox21 (Huang and Xia 2017)	7831	12	Tox21 offers a comprehensive toxicity profile of compounds. Central to the 2014 Tox21 Data Challenge, it aims at enhancing predictions for toxic responses to ensure safer drug design.
ToxCast (Richard et al. 2016)	8575	617	ToxCast provides toxicity labels from high-throughput screenings, enabling swift evaluations and guiding early drug development stages.

MoleculeNet dataset: https://moleculenet.org/datasets-1

Experiment - Datasets

Five regression tasks:

Dataset	# Molecules	# Tasks	Description
FreeSolv (Mobley and Guthrie 2014)	642	1	A dataset that brings together information on the hydration free energy of molecules in water. The dual presence of experimental data and alchemical free energy calculations offers researchers a robust platform to understand solvation processes and predict such properties for novel molecules.
ESOL (Delaney 2004)	1128	1	Understanding the solubility of compounds is fundamental in drug formulation and delivery. The ESOL dataset chronicles solubility attributes, providing a structured framework to predict and modify solubility properties in drug design.
Lipophilicity (Gaulton et al. 2012)	4200	1	Extracted from the ChEMBL database, this dataset focuses on a compound's affinity for lipid bilayers—a key factor in drug absorption and permeability. It provides valuable insights derived from octanol/water distribution coefficient experiments.
QM7 (Blum and Reymond 2009)	6830	1	A curated subset of GDB-13, the QM7 dataset houses details on computed atomization energies of stable, potentially synthesizable organic molecules. It provides an arena for validating quantum mechanical methods against empirical data, bridging computational studies with experimental chemistry.
QM8 (Ramakrishnan et al. 2015)	21786	12	A more extensive dataset, QM8 encompasses computer-generated quantum mechanical properties. It details aspects like electronic spectra and the excited state energy of molecules, offering a robust resource for computational chemists aiming to predict or understand such attributes.

MoleculeNet dataset: https://moleculenet.org/datasets-1

Experiment - Results

Main Result

Table 2: **Performance on six classification benchmarks (ROC-AUC, higher is better) and five regression benchmarks (RMSE for FreeSolv, ESOL, Lipophilicity and MAE for QM7/8, lower is better).** We report the mean and standard deviation. Top-3 and top-1 results are highlighted in **bold** and **bold red**, respectively. We highlight the backbone model, and the models that apply the backbone. Table split: Non-KG methods, other KG-based methods, and our method.

	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**

Experiment - Results

Ablation Study

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	✓	✓	$\mathbf{h}_{\text{MG}} \oplus \mathbf{h}_{\text{f}} \oplus \mathbf{h}_{\text{KG}}$

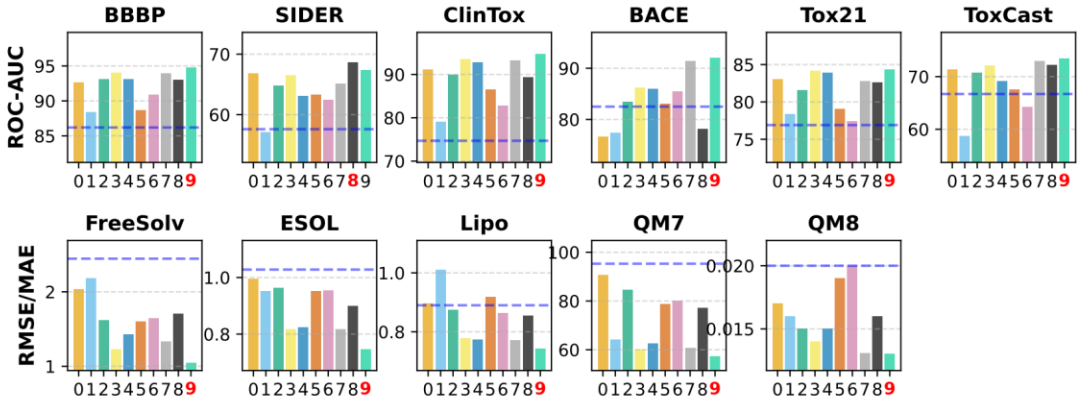


Figure 2: **Ablation study configurations and results.** (Left) Variants. "KGE": KG embedding initialization. "κ": κ-hop KG subgraph. "Pret.": KG-level pre-training. "Cont.": contrastive learning. "Embedding": input to MLP for fine-tuning. (Right) Performance comparison across different datasets and configurations. We highlight the **best** configuration for each dataset. The dotted blue lines denote the performance achieved by the backbone model (GROVER).

Findings:

1.[C1 vs C3]: KGE initialization is important.

2.[C0 vs C3] and (C8 vs C9): K-GNN training is important.

3.[C2 vs C3]: Contrastive learning is important.

4.[Backbone (blue line) vs C4]: Information can be transferred from K-GNN to M-GNN through contrastive learning.

5.[C3 vs C5] and [C4 vs C6]: 2-hop outperforms 3-hop.

6.[C4 vs C7] and [C0 vs C8] and [C3 vs C9]: Functional group features improve results.

Experiment - Results

Ablation Study

Table 3: Study the Effects of (top) Bi-level Self-supervised Pre-training and (below) Relationship Exclusion from MolKG.

(a) Effect of Pre-training on M-GNN and K-GNN

M-GNN Pret.	K-GNN Pret.	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast	FreeSolv	ESOL	Lipo	QM7	QM8
✓	✓	**94.8**	**67.4**	**94.7**	**92.0**	**84.3**	**73.4**	**1.048**	**0.746**	0.743	**57.2**	**0.013**
✗	✓	92.2	62.6	89.4	89.8	80.6	70.8	1.313	0.834	**0.708**	64.6	0.016
✓	✗	93.2	66.7	90.7	81.6	83.1	71.9	1.563	0.841	0.876	74.4	0.017
✗	✗	88.9	62.1	88.4	84.1	81.6	69.4	1.944	0.978	0.845	77.9	0.017

(b) Effect of Relationship Exclusion from MolKG

Knowledge Graph	BBBP	SIDER	ClinTox	BACE	Tox21	ToxCast	FreeSolv	ESOL	Lipo	QM7	QM8
MolKG	94.8	**67.4**	**94.7**	**92.0**	**84.3**	**73.4**	**1.048**	**0.746**	**0.743**	**57.2**	0.013
w/o *indication*	93.8	65.7	93.4	91.6	84.2	73.0	1.063	0.754	0.751	58.1	0.013
w/o *xlogp3 & xlogp3-aa*	93.7	66.0	94.2	91.1	83.0	72.8	1.189	0.789	0.782	57.8	**0.012**
w/o *tautomer_cnt & covalent_unit_cnt*	94.3	66.5	93.1	90.9	83.5	72.5	1.272	0.761	0.759	61.7	0.014
w/o *nbr_2d & nbr_3d & has_same_conn*	**95.0**	67.3	93.6	91.3	**84.3**	72.7	1.058	0.749	0.748	57.6	0.013

Findings:

1. Both M-GNN and K-GNN pre-training are crucial for the performance gain.

2. Different relation types have different impacts on tasks. (e.g., Removing "tautomer count" and "covalent unit count" notably impacted FreeSolv and QM7, suggesting their importance for predicting solvation and quantum properties.)

Experiment - Results

Visualization

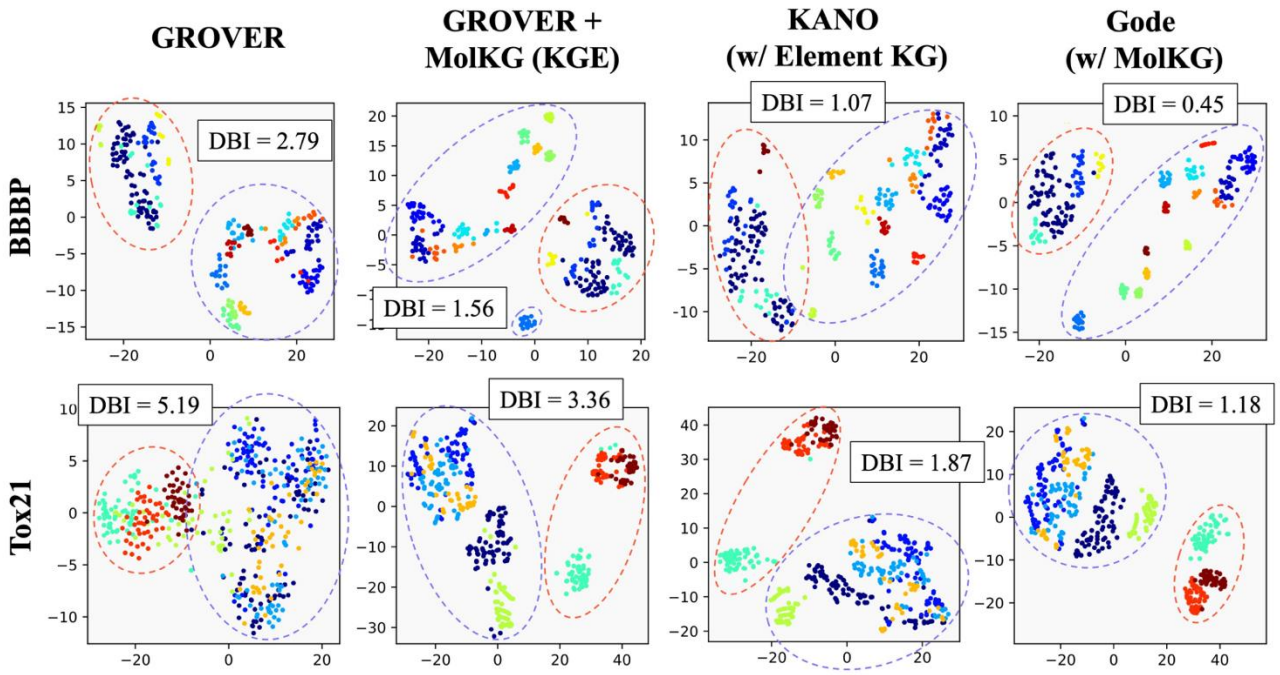


Figure 4: **t-SNE visualization of molecule embeddings across two tasks**. Each color represents a unique scaffold (molecule substructure). We compare the embeddings from GROVER, GROVER augmented with static KG embeddings from our MolKG, KANO, and GODE. The clustering quality is assessed using the DB index.

- GROVER shows poor separation with intermingled molecule scaffolds
- GROVER + MolKG static embeddings show improved but still overlapping clusters
- GODE achieves:
  - Clearest cluster separation
  - Minimal scaffold overlap
  - Lowest Davies-Bouldin Index, outperforming KANO and other approaches

Conclusion

ILLINOIS

Thank you!

Patrick Jiang