清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

UniCorn: A Unified Contrastive Learning Approach for Multi-view Molecular Representation Learning

Shikun Feng*, Yuyan Ni*, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, Yanyan Lan

AMSS

Background Problem

a) The relationship between different molecular pre-training methods is unexplored.

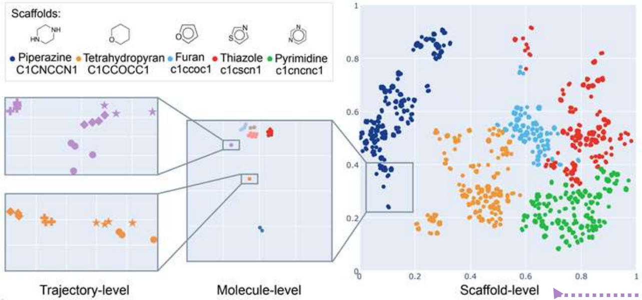
b) Existing prevalent molecular pre-train methods exhibit preference on specific types of downstream tasks, and the reason is unknow.

c) There lacks a universal model capable of effectively applying to various categories of molecular tasks for molecular pre-training.

c.

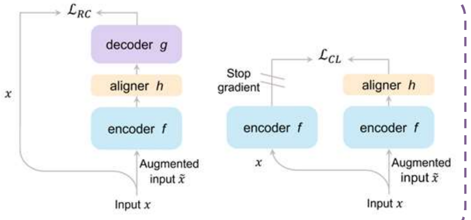
a. and b. demonstrate the *feasibility* and *necessity* to combine the strengths of existing methods by learning hierarchical molecular representations.

Thus we propose **UniCorn**, a unified pre-training framework, to learn multi-view molecular representations in the student encoder which is applicable to a wide array of downstream task s

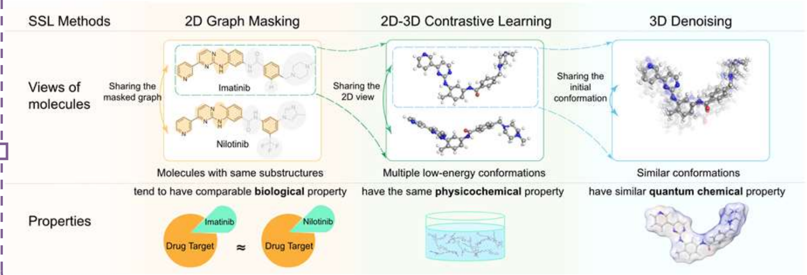


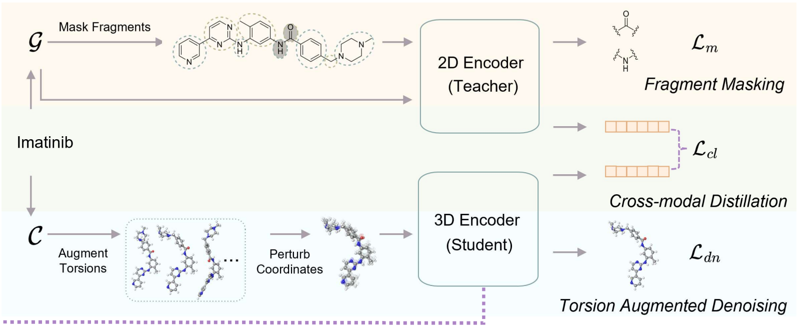
Analysis & Method

a. We theoretically unify *reconstructive* and *contrastive* methods, and comprehend them in a unified perspective by *contrastive learning* and *representation clustering*.



b. Based on the theoretical derivation, the three SSL methods lead to *clustering patterns* in molecular representation space at different granularity, thus benefiting specific tasks.


SSL Methods
2D Graph Masking
2D-3D Contrastive Learning
3D Denoising
Views of molecules
Sharing the masked graph
Imatinib
Nilotinib
Sharing the 2D view
Sharing the initial conformation
Molecules with same substructures tend to have comparable biological property
Multiple low-energy conformations have the same physicochemical property
Similar conformations have similar quantum chemical property
Properties
Imatinib
Drug Target
≈
Nilotinib
Drug Target



Results

- Quantum tasks (QM9)

Methods	Models	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$<R^2>$ (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{molK}$)
Multimodal	3D InfoMax	0.0280	0.057	25.9	21.6	42.1	0.141	1.67	13.30	13.81	13.62	13.73	0.030
	GraphMVP	0.0270	0.056	25.8	21.6	42.0	0.136	1.61	13.07	13.03	13.31	13.43	0.029
	MoleculeSDE	0.0260	0.054	25.7	21.4	41.8	0.151	1.59	12.04	12.54	12.05	13.07	0.028
	MoleculeJAE	0.0270	0.056	26.0	21.6	42.7	0.141	1.56	10.70	10.81	10.70	11.22	0.029
	MoleBLEND	0.0370	0.060	21.5	19.2	34.8	0.417	1.58	11.82	12.02	11.97	12.44	0.031
3D Denoising	Transformer-M	0.0370	0.041	17.5	16.2	27.4	0.075	**1.18**	9.37	9.41	9.39	9.63	0.022
	SE(3)-DDM	0.0150	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
	3D-EMGP	0.0200	0.057	21.3	18.2	37.1	**0.092**	1.38	8.60	8.60	8.70	9.30	0.026
	Frad	0.0100	0.037	15.3	13.7	27.8	0.342	1.42	5.33	5.62	5.55	6.19	0.020
	UniCorn	**0.0085**	**0.036**	**13.0**	**11.9**	**24.9**	0.326	1.40	**3.99**	**3.95**	**3.94**	**5.09**	**0.019**

- Biological tasks (MoleculeNet)

Methods	Models	BBBP	Tox21	MUV	BACE	ToxCast	SIDER	ClinTox	HIV	Avg.
Graph Masking	AttrMask	65.0±2.3	74.8±0.2	73.4±2.0	79.7±0.3	62.9±0.1	61.2±0.1	87.7±1.1	76.8±0.5	72.7
	GROVER	70.0±0.1	74.3±0.1	67.3±1.8	82.6±0.7	65.4±0.4	64.8±0.6	81.2±3.0	62.5±0.9	71.0
	GraphMAE	72.0±0.6	75.5±0.6	76.3±2.4	83.1±0.9	64.1±0.3	60.3±1.1	82.3±1.2	77.2±1.0	73.9
	Mole-BERT	71.9±1.6	76.8±0.5	78.6±1.8	80.8±1.4	64.3±0.2	62.8±1.1	78.9±3.0	78.2±0.8	74.0
Multimodal	3D InfoMax	69.1±1.0	74.5±0.7	74.4±2.4	79.7±1.5	64.4±0.8	60.6±0.7	79.9±3.4	76.1±1.3	72.3
	GraphMVP	68.5±0.2	74.5±0.4	75.0±1.4	76.8±1.1	62.7±0.1	62.3±1.6	79.0±2.5	74.8±1.4	71.7
	MoleculeSDE	71.8±0.7	76.8±0.3	80.9±0.3	79.5±2.1	65.0±0.2	60.8±0.3	87.0±0.5	78.8±0.9	75.1
	MoleBLEND	73.0±0.8	77.8±0.8	77.2±2.3	83.7±1.4	66.1±0.0	**64.9±0.3**	87.6±0.7	79.0±0.8	76.2
	UniCorn	**74.2±1.1**	**79.3±0.5**	**82.6±1.0**	**85.8±1.2**	**69.4±1.1**	64.0±1.8	**92.1±0.4**	**79.8±0.9**	**78.4**

- Quantum tasks (MD17)

Models	Aspirin	Benzene	Ethanol	Malonal -dehyde	Naphtha -lene	Salicy -lic Acid	Toluene	Uracil
MoleculeJAE	1.289	0.345	0.365	0.613	0.498	0.712	0.480	0.463
MoleculeSDE	1.112	0.304	0.282	0.520	0.455	0.725	0.515	0.447
SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
Coord	0.211	0.169	0.096	**0.139**	0.053	0.109	0.058	**0.074**
Frad	0.209	0.199	0.091	0.142	0.053	0.108	0.054	0.076
UniCorn	**0.168**	**0.165**	**0.086**	0.152	**0.046**	**0.098**	**0.052**	0.084

- Physicochemical tasks (MoleculeNet)

Models	ESOL	FreeSolv	Lipo
AttrMask	1.112±0.048	-	0.730±0.004
GROVER	0.983±0.090	2.176±0.052	0.817±0.008
3D InfoMax	0.894±0.028	2.337±0.227	0.695±0.012
GraphMVP	1.029±0.033	-	0.681±0.010
MoleBLEND	0.831±0.026	1.910±0.163	0.638±0.004
UniCorn	**0.817±0.034**	**1.555±0.075**	**0.591±0.016**

UniCorn achieves optimal results across 33 out of 38 molecular tasks that span a wide range of quantum, physicochemical, and biological domains.

*Equal contribution.

Contact us: lanyanyan@air.tsinghua.edu.cn;
fsk21@mails.tsinghua.edu.cn;
niyuyan17@mails.ucas.ac.cn.

Previous work: Fractional denoising(ICML2023)
Sliced denoising(ICLR2024)