Interpretable Chirality-Aware Graph Neural Network for Quantitative Structure Activity Relationship Modeling in Drug Discovery


Yunchao (Lance) Liu[1]


Yu Wang[1]


Oanh Vu[1]


Rocco Moretti[1]

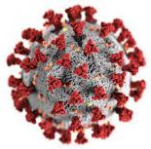

Bobby Bodenheimer[1]


Jens Meiler[1,2]


Tyler Derr[1]

[1]Vanderbilt University
[2] Leipzig University

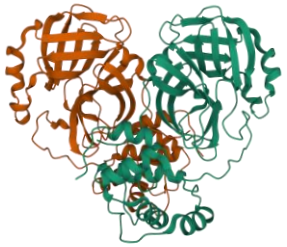
Drug Discovery Is The Process of Finding Target-Binding Molecules



SARS-COV-2

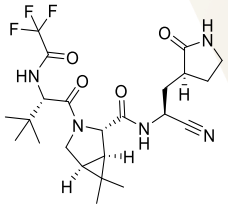
Essential for Viral Replication

Treat the Disease

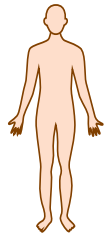


Target
3CL^{pro}

Bind and Inhibit Function



Small Molecule
Nirmatrelvir
(Used in Paxlovid^{TM})



Disease

Related to the Disease

Treat the Disease

Target

Bind and Alter Function

Lock-And-Key Model
Proposed by Emil Fischer in 1894
(Berg et al. 2008)

Small Molecule

* Rigid-body binding is assumed in this study

Berg, J. M., et al. (2008). Biochemistry

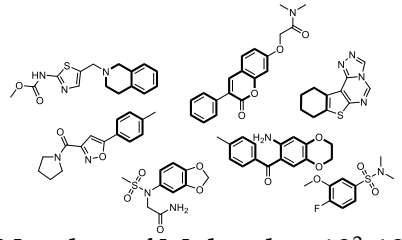
Images: https://www.apsf.org/article/an-update-on-the-perioperative-considerations-for-covid-19-severe-acute-respiratory-syndrome-coronavirus-2-sars-cov-2/; https://www.seekpng.com/png/detail/10-100352_lock-icon-lock-cartoon-png.png; https://upload.wikimedia.org/wikipedia/commons/thumb/6/65/Crypto_key.svg/2560px-Crypto_key.svg.png

Virtual Screening Reduces the Number of Compounds Necessary For Testing

How to find the small molecule?

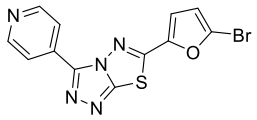
Experimental Screening

Chemical Libraries



Number of Molecules: 10^3-10^6

High Throughput Screening Equipment

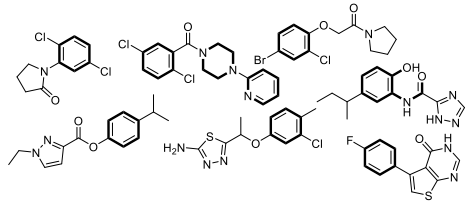


Hit Rate: 0.05%-0.5%

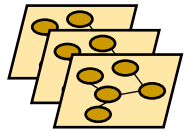
Experimental Data Training

Virtual Screening

Virtual Libraries



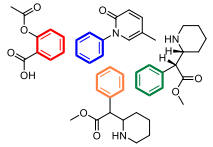
e.g., 10^{10} Virtual Molecules on the REAL database in Enamine Ltd.



Graph Neural Network(GNN)

Focused Libraries

Predicted Active Molecules



Number of Molecules: 500-1000

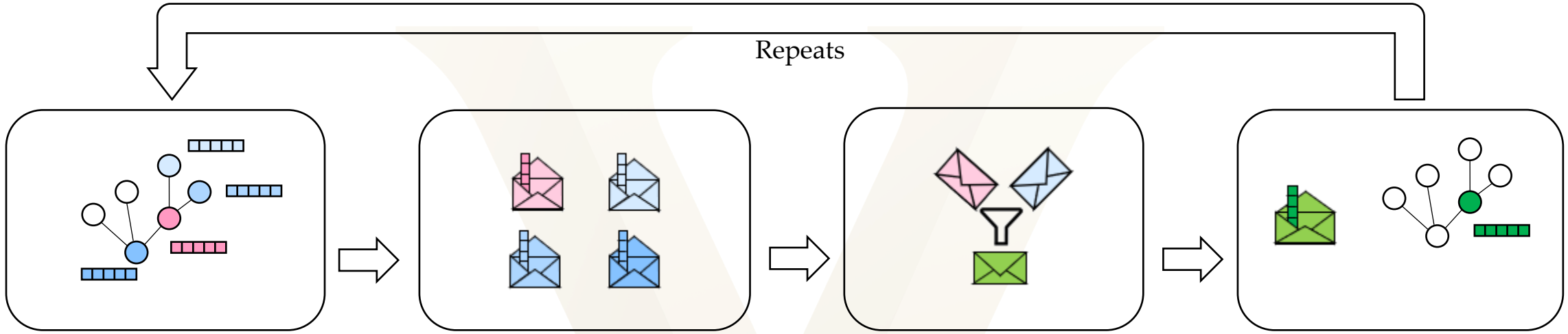
Experimental Validation

Make-On-Demand Synthesis

Images: https://web-material3.yokogawa.com/7aaf6ce1eb279bc69d63f9d7b02951bbe91b197a.png

Background: Message Passing Scheme

Neural Message Passing (Gilmer et al. 2017)



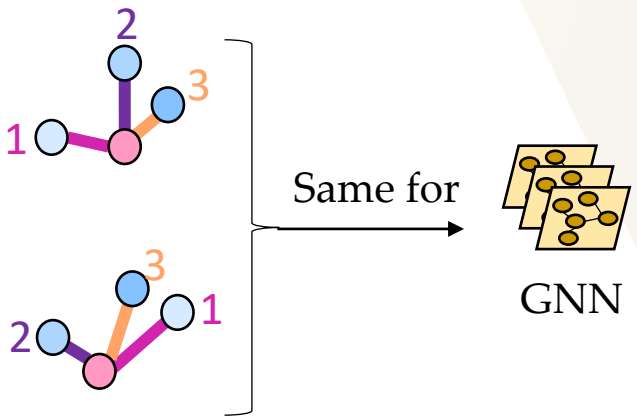
(A) State in Iteration $(k-1)$ **(B)** Message Preparation **(C)** Message Aggregation **(D)** New State in Iteration (k)

Key Idea: generate new node embedding by **aggregating attributes** of **neighboring nodes**

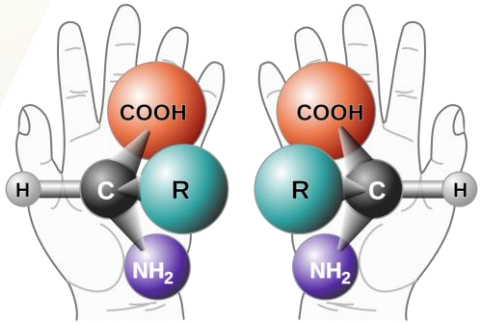
Gilmer, J., et al. (2017). Neural message passing for quantum chemistry

GNN Limitations

GNNs cannot distinguish different ordering of neighbors



Neighbor Ordering is important
Chiral Molecules



Chirality plays a significant role in binding (Sliwoski et al. 2012)

A new design of **Molecular Convolution** will solve this problem

Sliwoski, G., et al. (2012). BCL::EMAS--enantioselective molecular asymmetry descriptor for 3D-QSAR

Images: https://upload.wikimedia.org/wikipedia/commons/thumb/e/e8/Chirality_with_hands.svg/300px-Chirality_with_hands.svg.png

Intuition From Image Convolution

Image Convolution
In Edge Detection

10	10	10	0	0	0	0
10	10	10	0	0	0	0
10	10	10	0	0	0	0
10	10	10	0	0	0	0
10	10	10	0	0	0	0
10	10	10	0	0	0	0

10	0	*	1	0	=	10	0
10	0		1	0		10	0

= 20

Image Patch | Vertical Edge Kernel

10	0	*	1	1	=	10	0
10	0		0	0		0	0

= 10

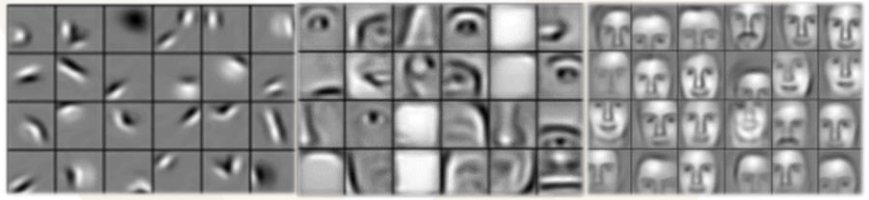
Image Patch | Horizontal Edge Kernel

Image Convolution
Multiply values in each pixel with corresponding values in the kernel and sum

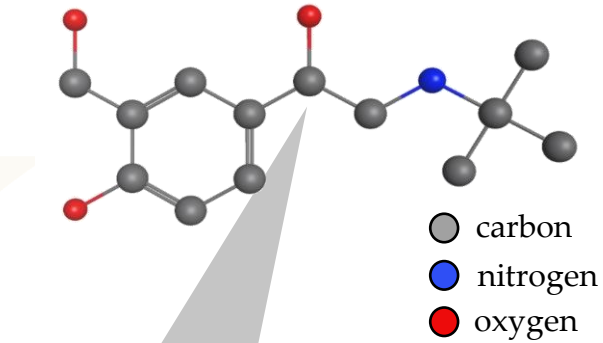
Observations

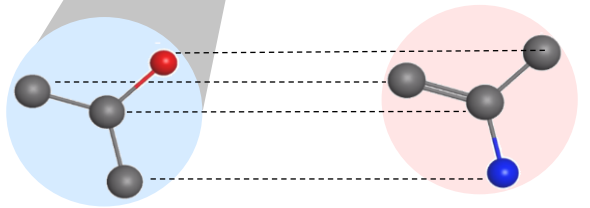
1. The higher the convolution value, the more similar a patch is to the kernel
2. The patterns in the kernels offers interpretability

e.g., kernels reveal patterns in a **face detection** task



Low-Level Patterns | Medium-Level Patterns | High-Level Patterns



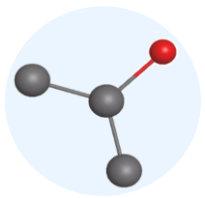


Molecular Neighborhood | Molecular Kernel

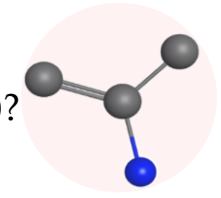
Similarity Score (SS) ⟸ outputs **Molecular Convolution**

1. The more similar, the higher the score
2. Molecular convolution should be rotation-invariant
3. Molecular kernels offer the benefit of interpretability

Similarity Score Calculation

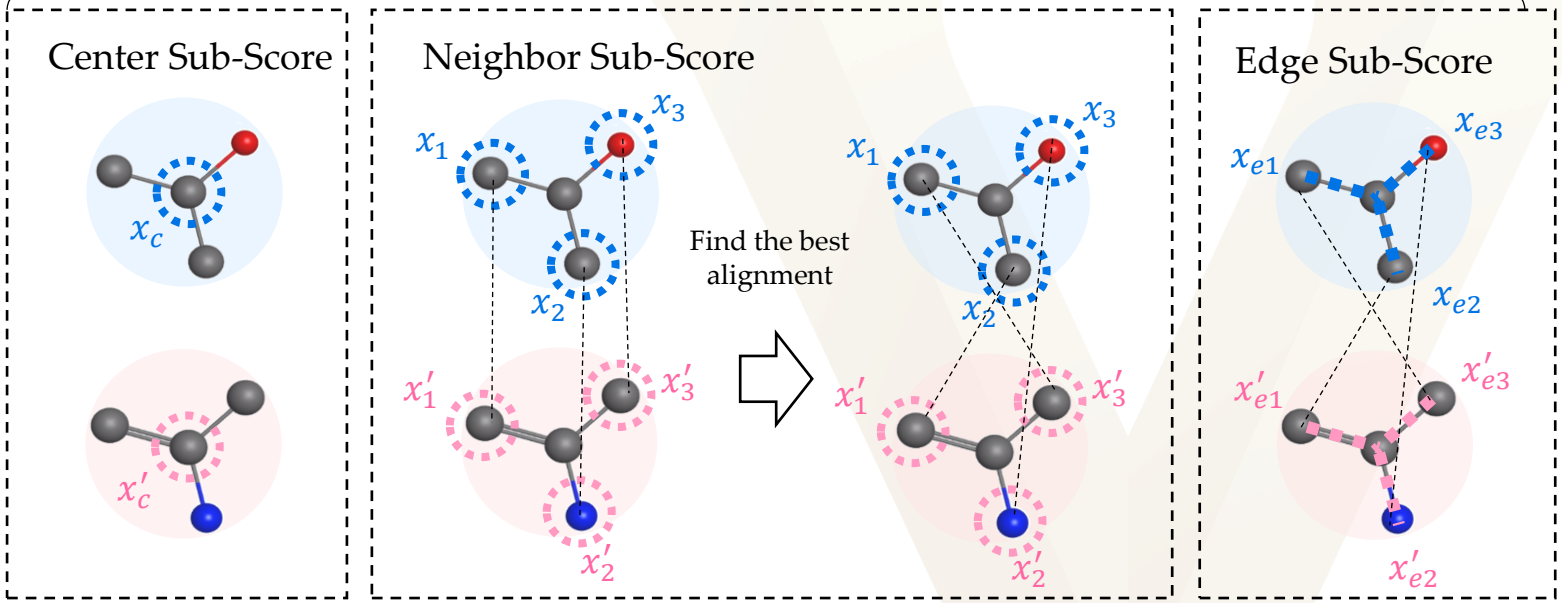


How to calculate **Similarity Score (SS)**?



Molecular Neighborhood

Molecular Kernel



x_1 aligns x'_1
x_2 aligns x'_2
x_3 aligns x'_3

x_1 aligns x'_3
x_2 aligns x'_1
x_3 aligns x'_2

x_{e1} aligns x'_{e3}
x_{e2} aligns x'_{e1}
x_{e3} aligns x'_{e2}

Is this the best alignment?

Similarity Score (SS)

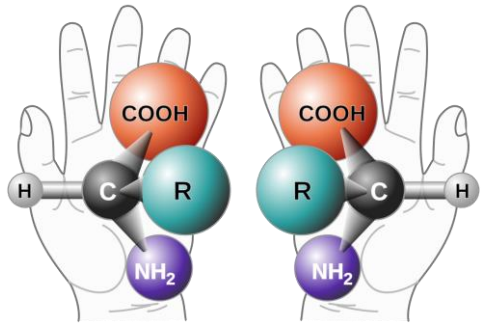
$$SS = \frac{1}{\sum w_{sub}} \sum_{s_{sub}\ in\ S} w_{sub} s_{sub}$$

S is the set of sub-scores
w_{sub} is a learned weight of the sub-score

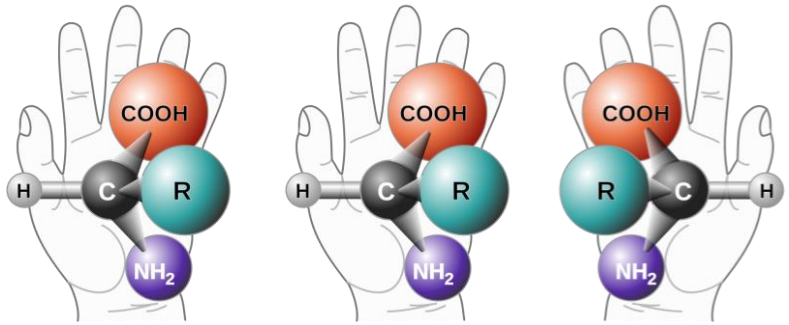
If $degree_{neighborhood} \neq degree_{kernel}$,

Then SS = 0

Signed Tetrahedral Volume For Chirality



How to distinguish chiral molecules?



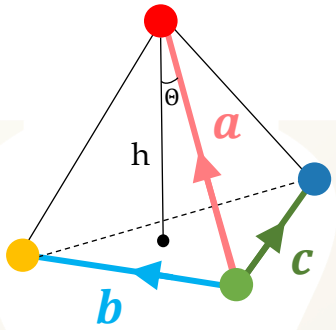
orientation1
indicator = 1

reference

orientation2
indicator = -1

SS = SS * indicator

Sliwoski, G., et al. (2012). BCL::EMAS--enantioselective molecular asymmetry descriptor for 3D-QSAR

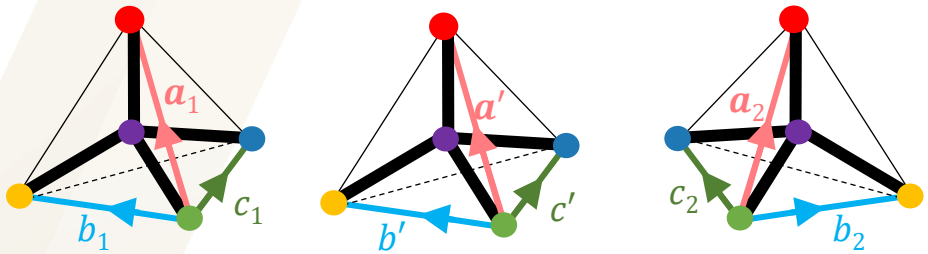


Vectors a, b, c are made from four neighbors

Signed Tetrahedral volume = $\frac{1}{6} * a \times b \bullet c$

This **volume** can be **positive** or **negative**

(Sliwoski et al. 2012)



Neighborhood 1 **Kernel** Neighborhood 2

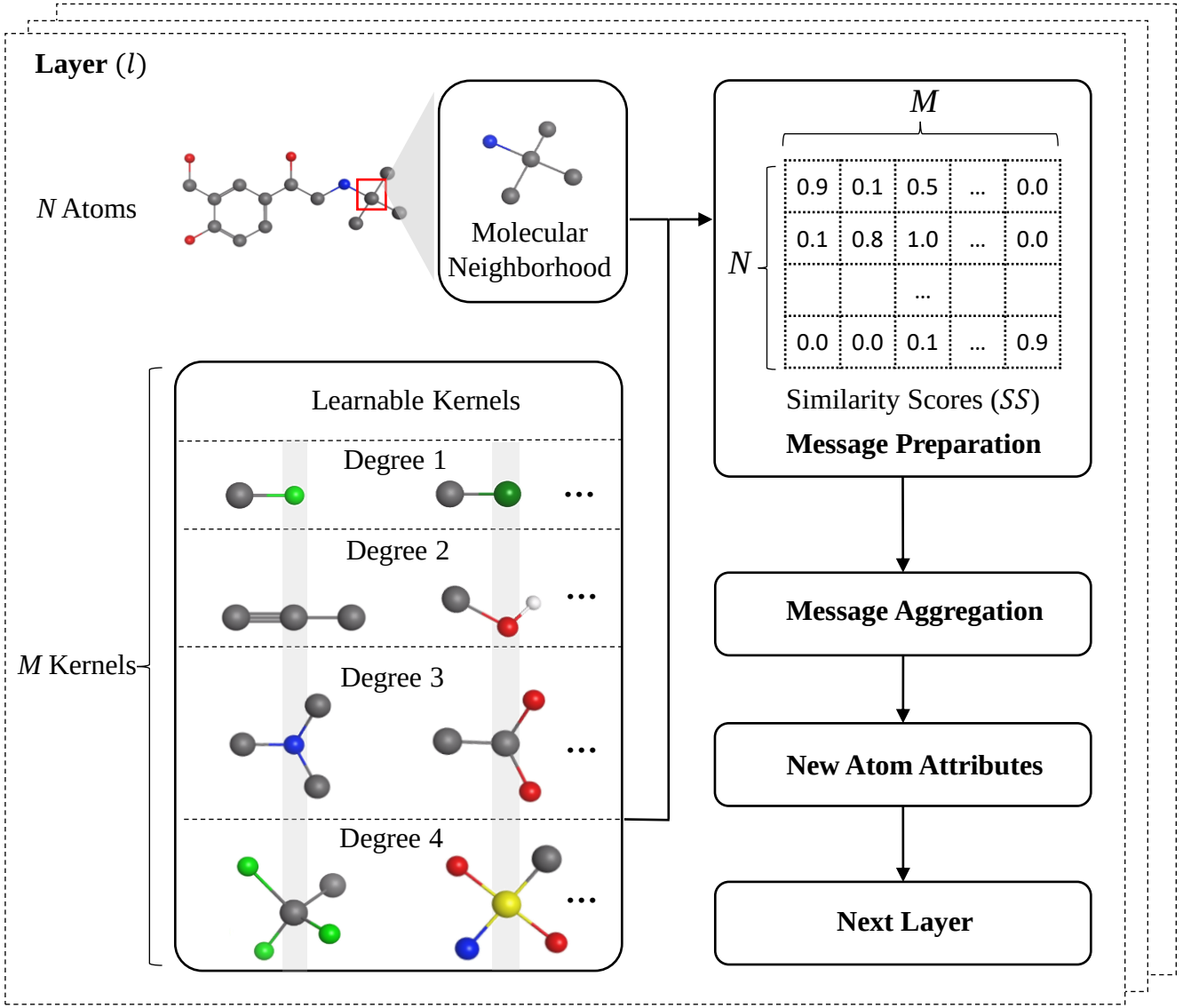
$$\text{Sign}(a_1 \times b_1 \bullet c_1) = \text{Sign}(a' \times b' \bullet c') \neq \text{Sign}(a_2 \times b_2 \bullet c_2)$$

Same tetrahedral volume sign

Different tetrahedral volume sign

MolKGNN Overview

Molecular-Kernel Graph Neural Network (MolKGNN)



The Datasets Used Were Carefully Curated

Commonly used benchmark small molecule datasets in the GNN community

- **Limited number of data**
  - MUTAG (188 molecules) (Debnath et al. 1991)
- **No processing to remove experimental artifacts**
  - Ogbg-molpcba (>400 k molecules) (Hu et al. 2020)
- **Not relevant to molecular activity against therapeutic targets**
  - PCQM4Mv2 (quantum chemistry) (Nakata et al. 2017)

Datasets Used In Our Work
(Butkiewicz et al. 2013)

Original Source
PubChem database (Kim et al. 2019)

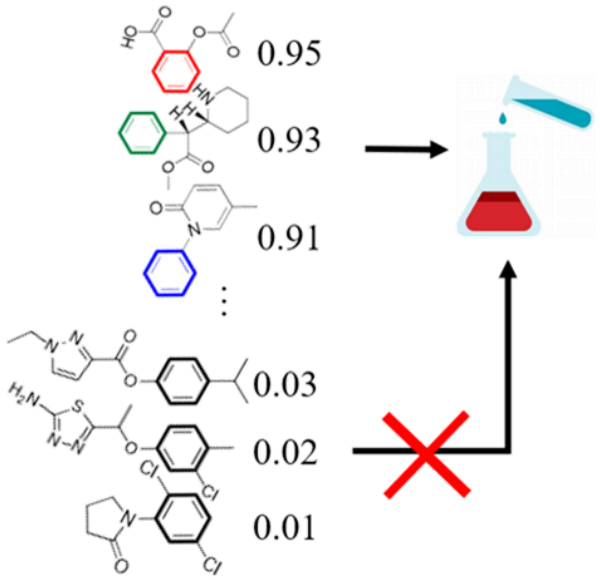
Reasons For Choosing This Dataset
1. At least 150 confirmed active compounds present
2. Diverse target classes
3. Realistic (large number and imbalanced label)

Protein Target Class	Protein Target (PubChem AID)	Total # of Graphs	# of Active Labels	Per Graph Avg. # of Nodes (Edges)
GPCR	Orexin1 Receptor (435008)	218,156	233	45.14 (94.37)
	M1 Muscarinic Receptor Agonists (1798)	61,832	187	43.60 (91.37)
	M1 Muscarinic Receptor Antagonists (435034)	61,755	362	43.61 (91.41)
Ion Channel	Potassium Ion Channel Kir2.1 (1843)	301,490	172	44.41 (92.81)
	KCNQ2 Potassium Channel (2258)	302,402	213	44.44 (92.88)
	Cav3 T-type Calcium Channels (463087)	100,874	703	43.75 (91.57)
Transporter	Choline Transporter (488997)	302,303	252	44.46 (92.90)
Kinase	Serine/Threonine Kinase 33 (2689)	319,789	172	44.85 (93.70)
Enzyme	Tyrosyl-DNA Phosphodiesterase (485290)	341,304	281	46.13 (96.50)

Debnath, A. K., et al. (1991). Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity
Hu, W., et al. (2020). Open graph benchmark: Datasets for machine learning on graphs
Nakata, M., et al. (2017). PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry
Butkiewicz, M., et al. (2013). Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database
Kim, S., et al. (2019). PubChem 2019 update: improved access to chemical data

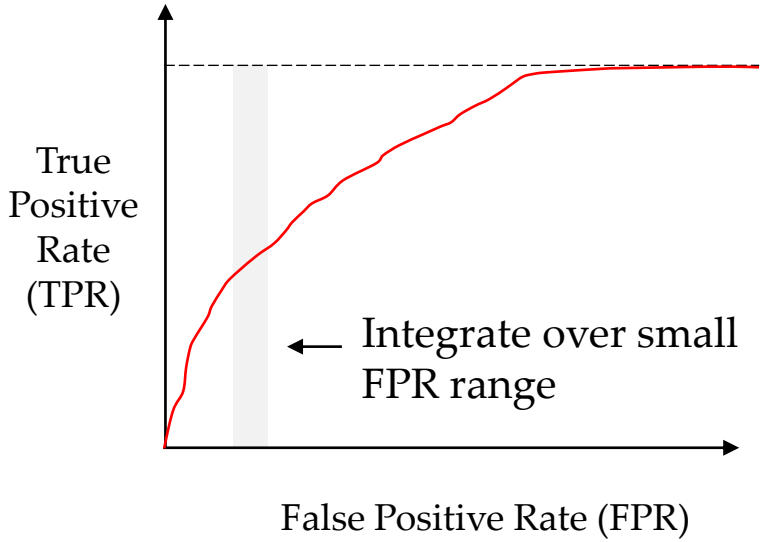
Domain Relevant Metric Is Used For Evaluation

Predicted Activity Score



Only top-ranked predictions will be **experimentally validated** due to the cost.

Ranged logAUC
(Mysinger et al. 2010)



Ranged logAUC is used to **bias toward** the performance of these **top-ranked predictions**

$\text{logAUC}_{[0.001,\ 0.1]}$
(Vu et al. 2019; Mendenhall et al. 2016)

For A Perfect Classifier:

$$\text{logAUC}_{[0.001,\ 0.1]} = 1$$

For A Random Classifier:

$$\text{logAUC}_{[0.001,\ 0.1]} = \frac{\int_{0.001}^{0.1} x \, \mathrm{d} \log_{10} x}{\int_{0.001}^{0.1} 1 \, \mathrm{d} \log_{10} x} = \frac{\int_{-3}^{-1} 10^{u} \mathrm{d}u}{\int_{-3}^{-1} 1 \mathrm{d}\, u} \approx 0.0215$$

Mysinger, M. M., et al. (2010). Rapid context-dependent ligand desolvation in molecular docking
Vu, O., et al. (2019). BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization
Mendenhall, J., et al. (2016). Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout

Quantitative Result

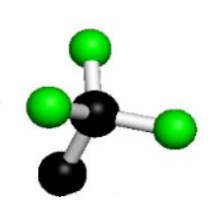
Domain Relevant Metric: $LogAUC_{[0.001, 0.1]}$

PubChem AID	Our Model	3D GNN			Chirality-Sensitive	Similar Architecture
	MolKGNN	SchNet	SchereNet	DimeNet++	ChIRo	KerGNN
435008	**0.255±0.014**	0.187±0.027	0.215±0.024	0.203±0.047	0.168±0.019	0.147±0.015
1798	0.174±0.029	0.195±0.025	0.196±0.035	**0.208±0.035**	0.165±0.040	0.078±0.042
435034	0.227±0.022	**0.246±0.020**	0.230±0.034	0.235±0.044	0.211±0.023	0.179±0.045
1843	**0.362±0.033**	0.358±0.037	0.258±0.048	0.284±0.034	0.326±0.010	0.292±0.027
2258	0.301±0.028	0.240±0.037	**0.380±0.037**	0.340±0.032	0.251±0.010	0.195±0.020
463087	0.390±0.056	0.332±0.022	**0.399±0.011**	0.389±0.026	0.258±0.019	0.150±0.011
488997	0.303±0.027	**0.319±0.017**	0.309±0.029	0.315±0.011	0.193±0.029	0.081±0.023
2689	**0.415±0.020**	0.324±0.020	0.401±0.016	0.367±0.049	0.351±0.048	0.264±0.017
485290	**0.498±0.015**	0.333±0.047	0.450±0.039	0.463±0.040	0.295±0.068	0.223±0.026
Average $LogAUC_{[0.001, 0.1]}$ (↑)	**0.325**	0.282	0.315	0.312	0.247	0.179
Average Rank (↓)	**2.333**	3.222	2.556	2.556	4.556	5.778

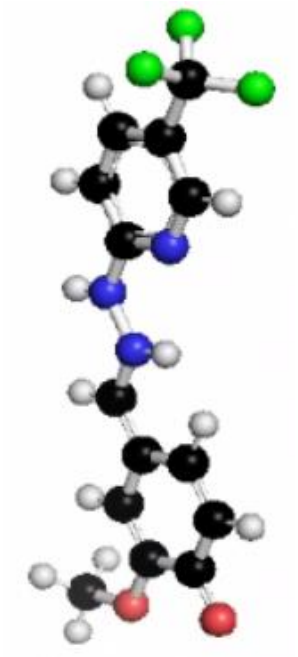
General Metric: AUC

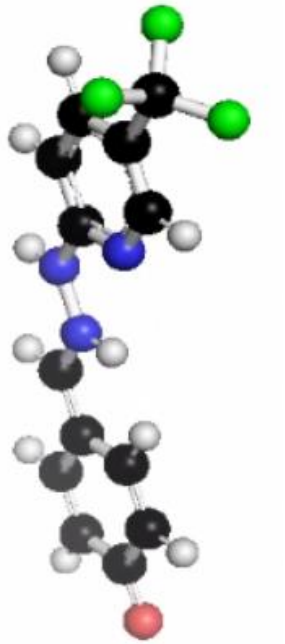
Average AUC (↑)	0.843	**0.844**	0.826	0.823	0.823	0.816
Average Rank (↓)	2.889	**2.111**	3.778	3.889	4.000	4.222

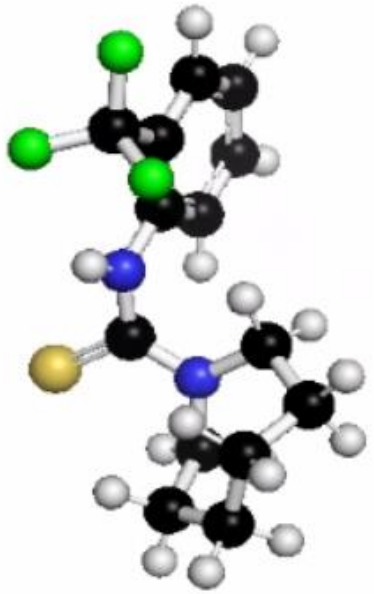
Intepretability Result



Translated Kernel Pattern







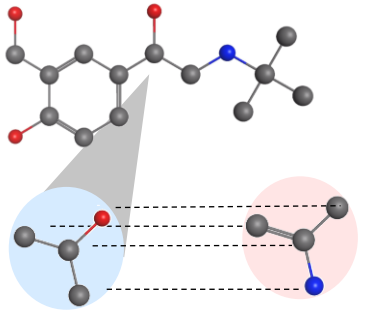
- Hydrogen (H)
- Carbon (C)
- Nitrogen (N)
- Oxygen (O)
- Fluorine (F)
- Sulfur (S)

- highlights the **interpretability of MolKGNN**
- provides an example of a learned kernel **aligning with domain knowledge**
- this is a known **important substructure** in medicinal chemistry, i.e., the **trifluoromethyl group**

Conclusion

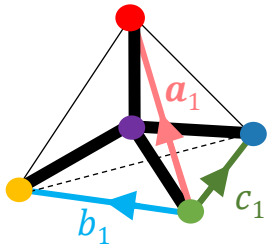
1. Novel Interpretable Molecular Convolution

We proposed a new framework named MolKGNN that uses a new convolution operation



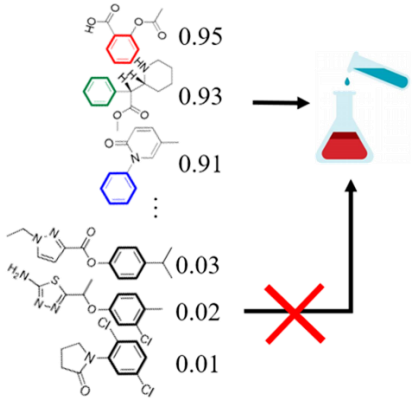
2. Better Chirality Characterization

the chirality calculation module in our design only needs a lightweight linear algebra calculation.



3. Comprehensive Evaluation in Computer-Aided Drug Discovery

realistic datasets and metric are used to demonstrate the superior performance of MolKGNN



Acknowledgements

LiVE Lab
Learning in Virtual Environments

Dr. Bobby Bodenheimer
Dr. Lauren Buck

Dr. Jens Meiler
Dr. Benjamin Brown
Dr. Oanh Vu
Dr. Rocco Moretti
Dr. Shannon Smith
Tracy Tang

Dr. Tyler Derr
Yi Zhang
Yu Wang
Yuying Zhao

Thank You!

Yunchao (Lance) Liu 刘运超

Ph.D Candidate in Computer Science at Vanderbilt University

Research Interests:

- AI Drug Design
- Geometric Deep Learning
- Self-supervised Learning

Future Directions:

- Better 3D Models
- More Interpretable Models
- Higher Data Efficient Models

Website: www.LiuYunchao.com