International Conference on Learning Representations

ICLR

Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks

Presented by **Yanqiao ZHU**

✉ yzhu@cs.ucla.edu

@ https://web.cs.ucla.edu/~yzhu

The Scalable Analytics Institute (ScAi)

Department of Computer Science

University of California, Los Angeles (UCLA)

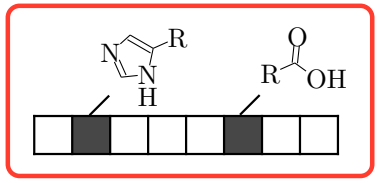
UCLA Samueli Computer Science

Joint work with Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Anton Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor W. Coley, Yizhou Sun, Wei Wang;

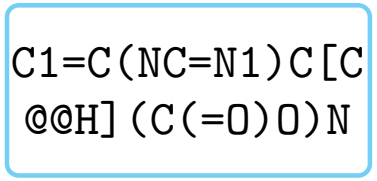
Supported by NSF Center for Computer Assisted Synthesis (C-CAS)

Molecular Representation Learning

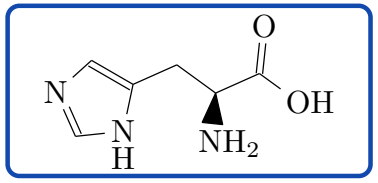
- Learning molecular representations is crucial for various biochemical applications like drug discovery, enzyme design
- 1D, 2D and 3D models are commonly used in MRL, with tradeoffs between efficiency and expressivity in capturing molecular structure
  - 1D/2D useful without 3D structural info, more efficient
  - 3D captures key geometric effects but more computationally expensive



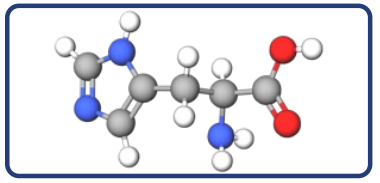
Fingerprint



SMILES String

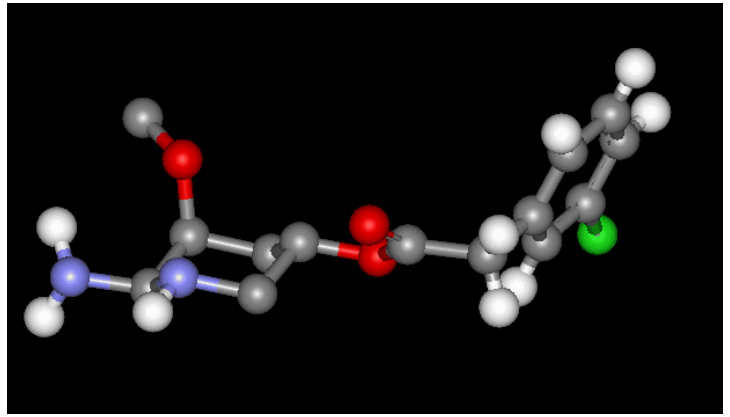


2D Topology


3D Geometry

Molecular Representation Learning (cont.)

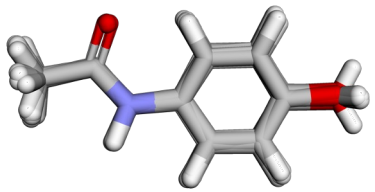
- Molecules are not rigid objects but are flexible structures
  - Molecules continuously interconvert between these conformers by rotations around bonds and vibrational motions
- The probability distribution over the thermodynamically-accessible conformer space (conformer ensemble) determines many chemical properties
  - Recognizing this flexibility and modeling the interconversion of conformers is important for accurate molecular modeling



Video from https://www.drugdesign.org/chapters/molecular-geometry/#conformers

Our Work: Modeling Conformer Ensembles

- Question: Can leveraging information from the entire conformer ensemble improve MRL models compared to encoding only a single conformation?



- We present the first Molecular Conformer Ensemble Learning (MARCEL) benchmark
  - Covers diverse molecules, broader than typical GNN benchmark datasets
  - Implements benchmarking suite with representative 1D, 2D, 3D MRL models
  - Explores 2 strategies to incorporate conformer ensembles into 3D models

Dataset Overview

Drugs-75K	Kraken	EE (Enantiomeric Excess)	BDE (Binding Energy)
• 75K drug-like molecules with 550K conformers • Task: Predicting quantum chemical properties	• 1.5K organophosphorus ligands with 20K conformers • Task: Predicting steric descriptors	• 1K catalyst-substrate pairs with 28K pro-R/S configurations • Task: Predicting enantioselectivity	• 6K organometallic catalysts with 110K unbound and bound poses • Task: Predicting binding energies

- Diverse molecules: drugs, organocatalysts, transition-metal catalysts
- Includes four chemically-relevant tasks for both molecules and reactions
- Emphasis on Boltzmann-averaged properties of conformer ensembles computed at the DFT level

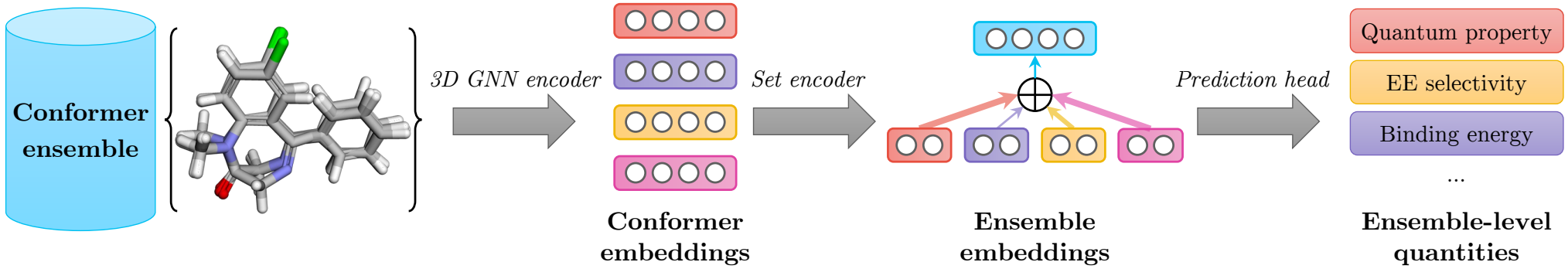
Benchmarked MRL Models

- 1D Sequence Models
  - Random forest based on fingerprints
  - Neural networks (RNNs/Transformers) on SMILES strings
- 2D Graph Neural Networks
  - Model molecular connectivity (atoms as nodes, bonds as edges)
- 3D Graph Neural Networks
  - Model molecular spatial conformations

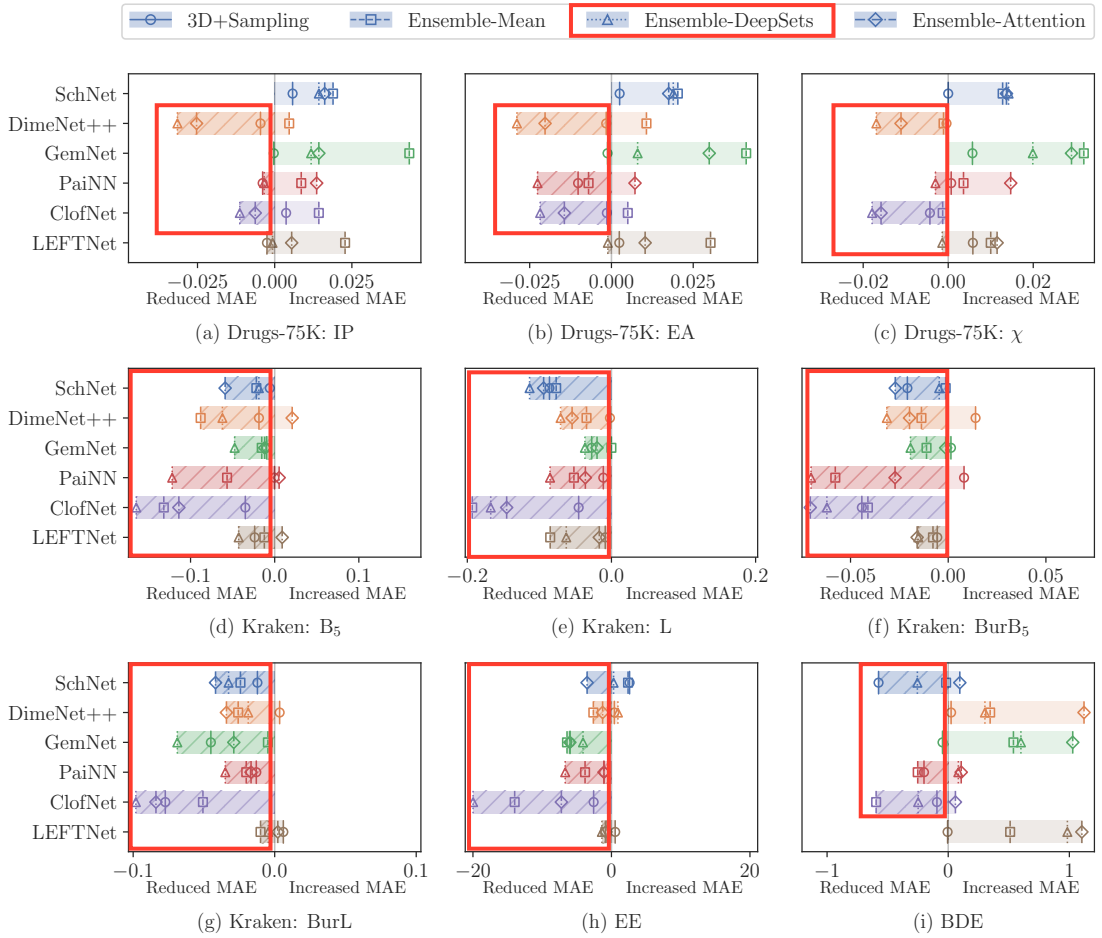
1D Models	2D Graph Networks		3D Graph Networks	
• Random Forest	• GIN	• ChemProp	• SchNet	• PaiNN
• LSTM	• GIN w/ Virtual Node (GIN-VN)	• GraphGPS	• DimeNet++	• ClofNet
• Transformer			• GemNet	• LEFTNet

Incorporating Conformer Ensembles

- Strategy 1: Training data augmentation
  - Randomly sample a conformer from ensemble during each training epoch
  - Useful if ensembles only available at training time
- Strategy 2: Ensemble learning with explicit set encoders
  - Generate individual conformer embeddings with 3D GNNs
  - Aggregate embeddings using set encoders (e.g., DeepSets)
  - Encode entire ensemble at both training and inference time

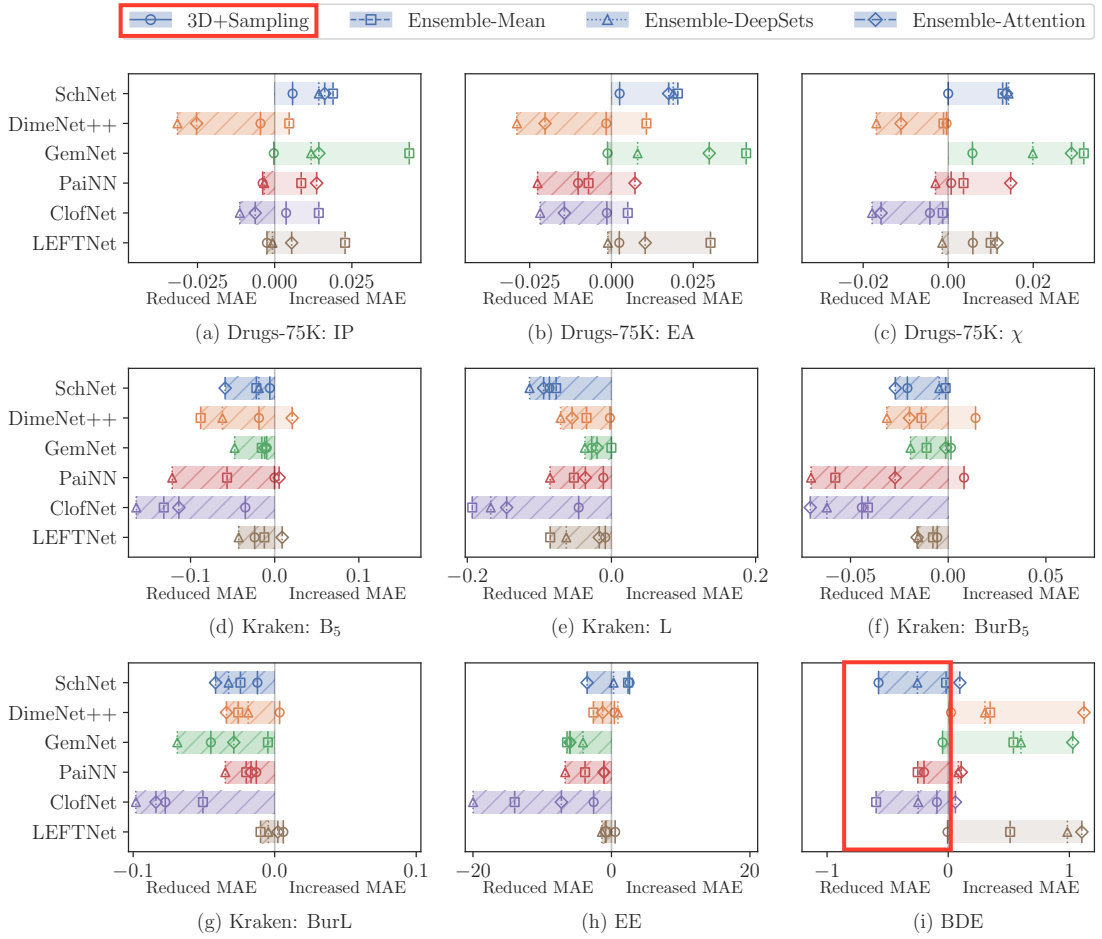


Key Benchmark Results (1/3)



- Explicit set encoders improve single-conformer 3D GNNs
  - DeepSets demonstrates improvements in 42 out of 54 experiments across 9 tasks and 6 base 3D models
  - Mean pooling loses discriminative power
  - Self-attention shows mixed results

Key Benchmark Results (2/3)



- Training data augmentation improves robustness
  - Sampling conformers during training improves robustness, especially on imprecise structures (BDE)
  - Improvement not consistent, possibly due to uniform sampling

Key Benchmark Results (3/3)

Category	Model	Drugs-75K			Kraken				EE	BDE
		IP	EA	χ	B_5	L	$BurB_5$	BurL		
1D	Random forest	0.4987	0.4747	0.2732	0.4760	0.4303	0.2758	0.1521	61.2963	3.0335
	LSTM	0.4788	0.4648	0.2505	0.4879	0.5142	0.2813	0.1924	64.0088	2.8279
	Transformer	0.6617	0.5850	0.4073	0.9611	0.8389	0.4929	0.2781	62.0816	10.0771
2D	GIN	0.4354	0.4169	0.2260	0.3128	0.4003	**0.1719**	**0.1200**	62.3065	2.6368
	GIN+VN	0.4361	0.4169	0.2267	0.3567	0.4344	0.2422	0.1741	62.3815	2.7417
	ChemProp	0.4595	0.4417	0.2441	0.4850	0.5452	0.3002	0.1948	61.0336	2.6616
	GraphGPS	0.4351	0.4085	0.2212	0.3450	0.4363	0.2066	0.1500	61.6251	2.4827
3D	SchNet	0.4394	0.4207	0.2243	0.3293	0.5458	0.2295	0.1861	17.7421	2.5488
	DimeNet++	0.4441	0.4233	0.2436	0.3510	0.4174	0.2097	0.1526	**14.6414**	**1.4503**
	GemNet	**0.4069**	**0.3922**	**0.1970**	**0.2789**	**0.3754**	0.1782	0.1635	18.0338	1.6530
	PaiNN	0.4505	0.4495	0.2324	0.3443	0.4471	0.2395	0.1673	20.2359	2.1261
	ClofNet	0.4393	0.4251	0.2378	0.4873	0.6417	0.2884	0.2529	33.9473	2.6057
	LEFTNet	0.4174	0.3964	0.2083	0.3072	0.4493	0.2176	0.1486	19.7974	1.5328

- Model performance depends on dataset and task
  - No single model consistently outperforms others
  - 1D/2D models perform well on small molecular datasets
  - 3D models excel on large datasets and reaction tasks
  - Model selection should consider dataset size, task, and efficiency-expressivity trade-off

Discussions

- Limitations:
  - Ensemble strategies do not always improve performance
  - High computational cost for encoding all conformers
  - Limited task and chemical space coverage
- Future directions:
  - Develop task-specific approaches integrating domain knowledge
  - Explore model efficiency-complexity trade-offs
  - Expand task and chemical space diversity
  - Investigate efficient ensemble encoding architectures
  - Explore physically-informed conformer sampling

Open-Source Library

MARCEL

- Datasets, model implementations, training pipeline available at https://github.com/SXKDZ/MARCEL
- Enables the community to reproduce results and extend the benchmark
- Aims to stimulate collaborative research at the intersection of machine learning and chemistry

THANKS

Code Repo

Paper

Slides