Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks

Yanqiao Zhu[1] Jeehyun Hwang[1] Keir Adams[2] Zhen Liu[3] Bozhao Nan[4] Brock A. Stenfors[4] Yuanqi Du[5]
Jatin Chauhan[1] Olaf Wiest[4] Olexandr Isayev[3] Connor W. Coley[2] Yizhou Sun[1] Wei Wang[1]

[1]University of California, Los Angeles [2]Massachusetts Institute of Technology [3]Carnegie Mellon University [4]University of Notre Dame [5]Cornell University

UCLA Samueli Computer Science

NSF Center for Computer Assisted Synthesis

Dataset Overview

Motivation

Molecular Representation Learning (MRL) is a promising approach for modeling molecules with machine learning.

Existing Graph Neural Network (GNN) models rely on a 2D molecular graph or a single 3D structure and thus overlook the *flexible nature* of molecules, which continuously interconvert across conformations via chemical bond rotations.

Problem Definition

For a given molecule or molecular complex, we assume that its geometry can be effectively characterized by a representative set of discrete, sampled conformers from the thermodynamically-accessible conformer distribution.

Formally, this set can be denoted as $\mathcal{C} = \{\boldsymbol{C}_i\}_{i=1}^{|\mathcal{C}|}$, where $\boldsymbol{C}_i \in \mathbb{R}^{|\mathcal{V}|\times 3}$ represents one conformer structure in 3D space. Each conformer is associated with a statistical weight corresponds to its probability under experimental conditions:

$$p_{\boldsymbol{C}_i} = \frac{\exp\left(-\frac{e_i}{k_B T}\right)}{\sum_j \exp\left(-\frac{e_j}{k_B T}\right)}$$

(e_i: conformer energy; T: temperature; k_B: Boltzmann constant)

Datasets and Tasks

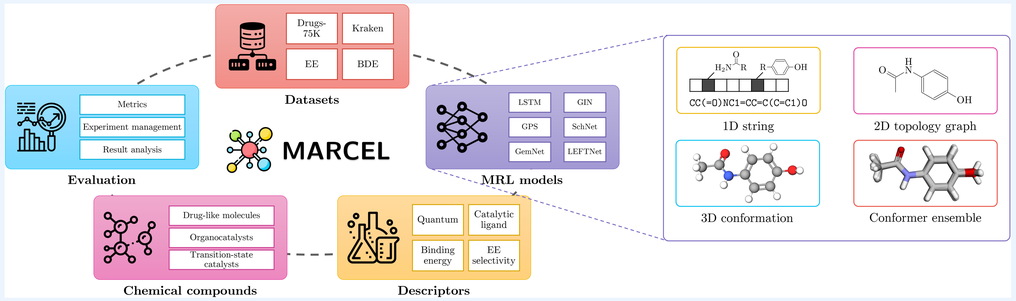
Drugs-75K is a subset of the GEOM-Drugs dataset. We aim to predict three DFT-based reactivity descriptors: ionization potential, electron affinity, and electronegativity.

Kraken is a dataset of monodentate organophosphorus (III) ligands. We consider four descriptors that quantify the steric size of a substituent: Sterimol B_5, Sterimol L, buried Sterimol B_5, and buried Sterimol L.

EE is a dataset of catalyst-substrate pairs with conformations of catalyst-substrate transition state complexes in two separate pro-S and pro-R configurations. The task is to predict the Enantiomeric Excess (EE) of the chemical reaction.

BDE is a dataset containing organometallic catalysts coordinated to two organic ligands with conformations of each unbound catalyst and the bound pose. The task is to predict the binding energy of the unbound and bound catalyst.

The MARCEL Benchmark



We present the MoleculAR Conformer Ensemble Learning (MARCEL) benchmark that comprehensively evaluates the potential of learning on conformer ensembles across a diverse set of molecules, datasets, and models.

Baseline Models

1D Models
- Random Forest
- LSTM
- Transformer

2D Graph Networks
- GIN
- GIN w/ Virtual Node (GIN-VN)
- ChemProp
- GraphGPS

3D Graph Networks
- SchNet
- DimeNet++
- GemNet
- PaiNN
- ClofNet
- LEFTNet

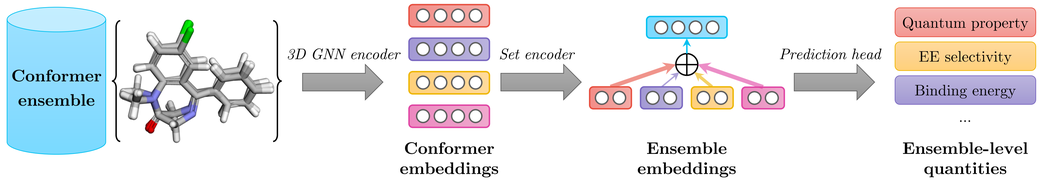
Ensemble Learning Strategies

Strategy 1: Training-Time Data Augmentation via Conformer Sampling

Enrich the training data by randomly sampling a conformer from the ensemble during each training epoch. Useful if conformer ensembles are only available at training time. During inference, the lowest-energy conformer is used to evaluate the model.

Strategy 2: Ensemble Learning with Explicit Set Encoders

First employ 3D GNNs to generate individual conformer embeddings and then aggregate them using a set encoder. Simultaneously encode the entire conformer ensemble at both training and inference time. Three simple set encoders considered: mean pooling, DeepSets, and self-attention.

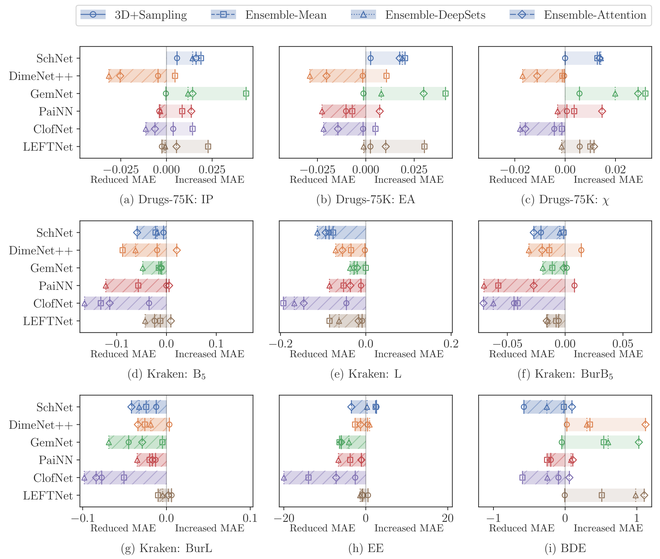


Results and Observations

The performance of the 1D, 2D, and 3D MRL models:

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

The *relative* improvement in test error for each 3D model when applying ensemble learning strategies:



Although performance varies across the datasets, tasks, and models, it is evident that ensemble learning strategies improve upon 3D models that only encode one conformer.

Observation 1: The conformer ensemble learning strategy with explicit set encoders frequently yields improved performance.

Observation 2: Sampling conformers at training time can improve performance, especially on imprecise conformer structures.

Observation 3: No model consistently outperforms the rest, with substantial task dependencies.