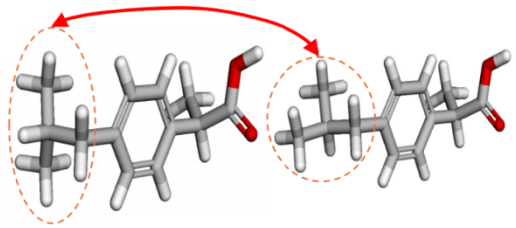
Structure-Aware Fusion with Progressive Injection for Multimodal Molecular Representation Learning

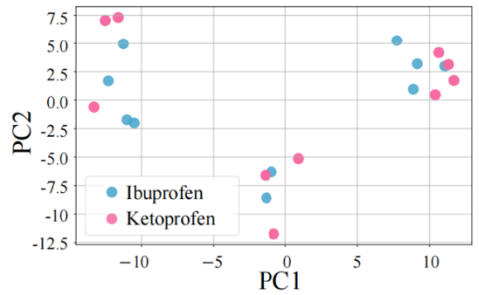
Authors: Zihao Jing, Yan Sun, Yan Yi Li , Sugitha Janarthanan, Alana Deng, Pingzhao Hu

Presenter: Zihao Jing

Problem & Challenges



(a) Two conformers show local 3D variation



(b) Conformers reveals PCA embedding instability

Conformer instability and modality collapse motivate SFP + PI.

- **3D conformers are unstable → unreliable fusion**

  Valid conformers vary by sampler/seed; early 3D-dependent fusion overfits these artifacts and yields inconsistent predictions.

- **Naive/symmetric fusion → modality collapse**

  Equal, everywhere fusion lets high-variance 3D dominate, drowning SMILES semantics and hurting robustness.

Key Ideas / Contributions

Robust fusion of 2D+3D structure with SMILES, injected progressively into a sequence backbone—strong gains across 29 tasks.

- **Structured Fusion Pipeline (SFP)**: align/encode 2D topology + 3D geometry into a stable structural prior.
- **Progressive Injection (PI):** asymmetrically inject that prior into the SMILES stream at later layers to avoid collapse.
- **Results:** avg +2.7% over best baseline across broad benchmarks; up to 27% gain on LD50.

Code Link

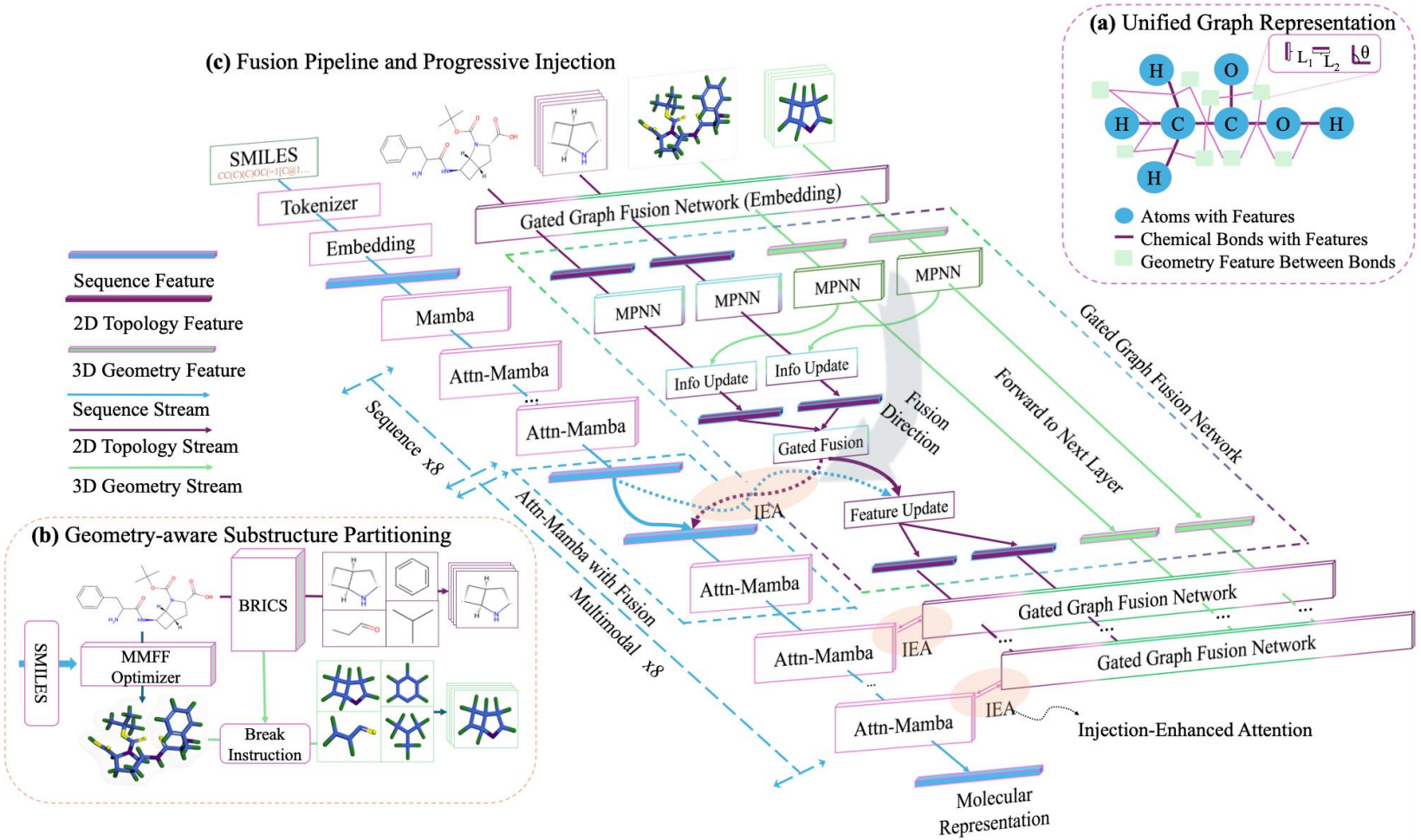
Paper Link

Contact Information

Zihao Jing, zjing29@uwo.ca

Pingzhao Hu, phu49@uwo.ca

Method Overview

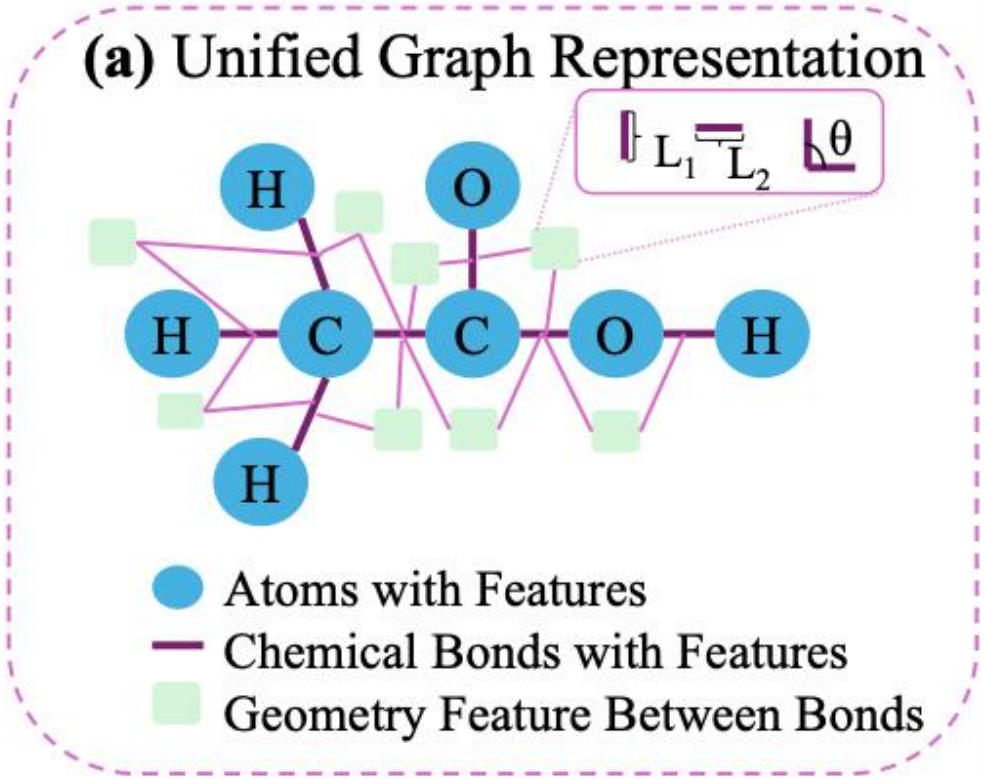


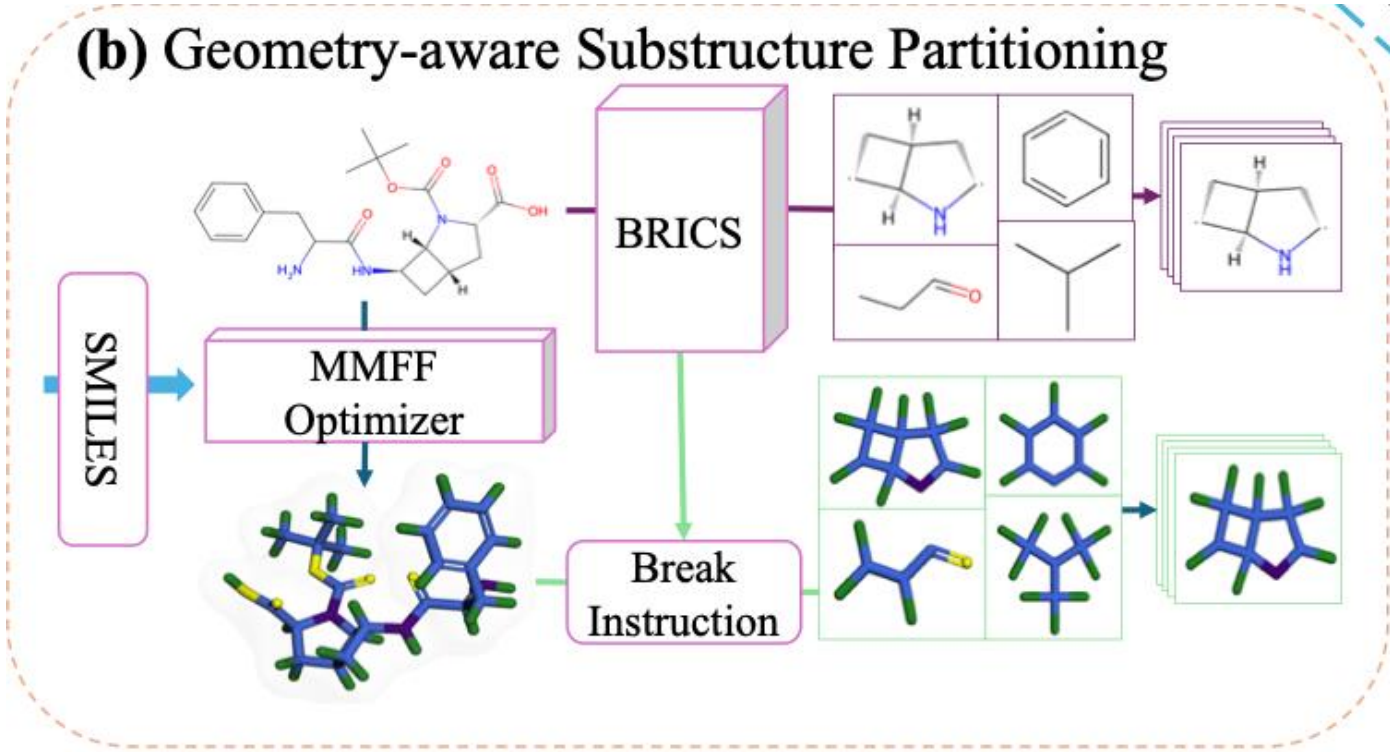
Two streams:

- Structural-fusion stream: 2D+3D fused → unified structural prior, propagated independently.
- Sequence stream: SMILES tokens modeled first; prior injected later via dedicated attention.

Novelty 1: Structured Fusion Pipeline (SFP)

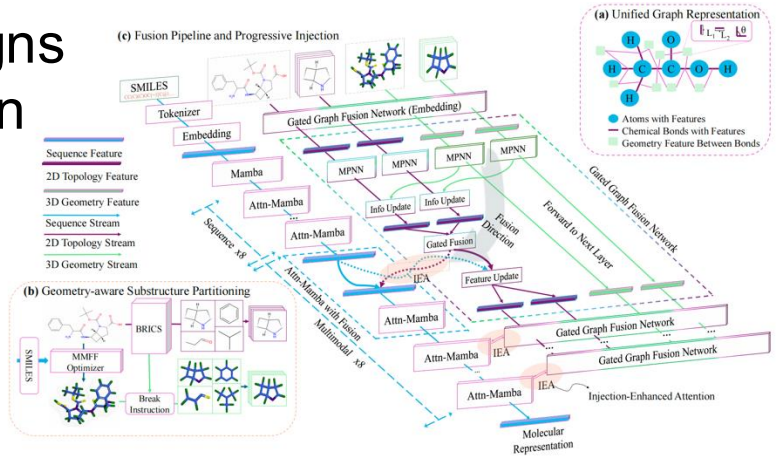
- **Structural Unified Graph:** nodes (atoms), bonds, and auxiliary geometric linkages (lengths & angles) create a rotation-invariant joint structure.
- **Geometry-Aware Substructure Partitioning:** extend BRICS to 3D; fuse global + local features with gated fusion to form the **structural prior**.

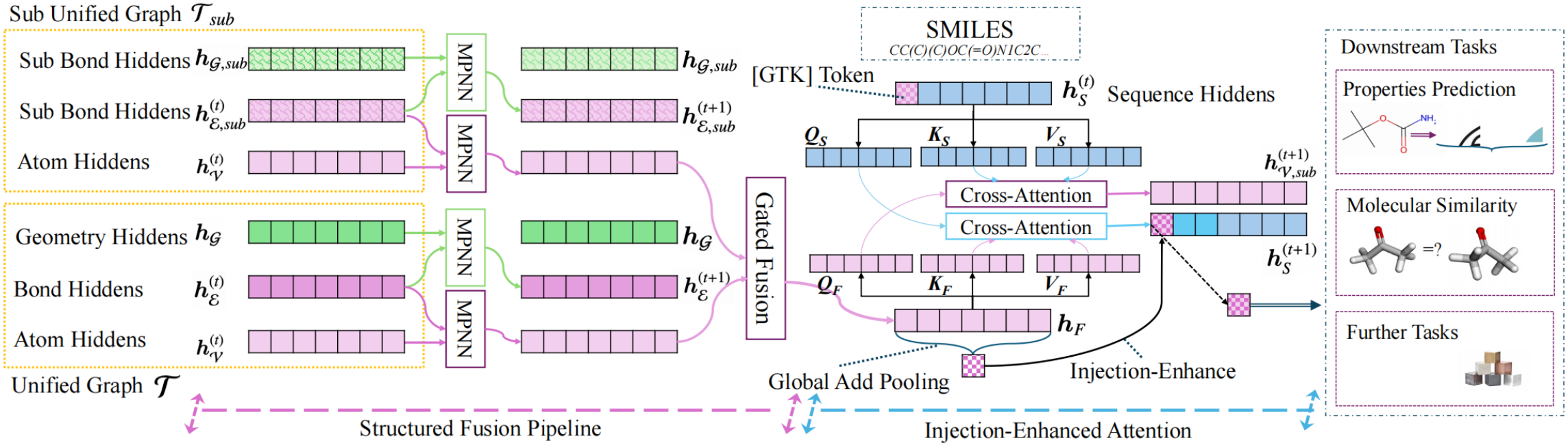




Novelty 2: Progressive Injection (PI)

- **Injection Enhanced Attention:** bidirectional cross-attention aligns structure ↔ sequence, then **pooled prior** updates a global token [GTK]; prior continues evolving in state-space layers (Mamba) across depth.
- **Why progressive?** Early symmetric fusion distorts semantics; delayed injection yields better stability and convergence.





Benchmarks

Models	BBB	HIA	Pgp	Bioav.	Tox-Avg.	CYP-Avg.	Top2Cnt/10
TDC Datasets - Classification - AUROC ↑							
AttentiveFP	0.855±0.011	0.974±0.007	0.892±0.012	0.632±0.039	0.842±0.010	0.749±0.008	0
FPGNN	0.888±0.018	0.958±0.012	0.930±0.007	0.666±0.035	0.860±0.017	0.866±0.004	4
DMPNN	0.864±0.010	0.976±0.004	0.889±0.005	0.617±0.050	0.821±0.019	0.819±0.004	2
AttrMasking	0.892±0.012	0.978±0.006	0.929±0.006	0.577±0.087	0.846±0.021	0.817±0.005	4
ContextPred	0.897±0.004	0.975±0.004	0.923±0.005	0.671±0.026	0.818±0.017	0.827±0.003	1
TranFoxMol	0.868±0.019	0.951±0.036	0.875±0.011	0.619±0.019	0.837±0.017	0.860±0.006	0
DeepMol	0.774±0.023	0.880±0.012	0.821±0.007	0.509±0.026	0.735±0.015	0.770±0.008	0
MuMo	0.899±0.014	0.979±0.013	0.942±0.019	0.714±0.021	0.840±0.015	0.880±0.017	7

Models	BACE-R	BACE-S	BBBP-R	BBBP-S	CLINTOX	SIDER	TOX21
MoleculeNet - Classification - AUROC ↑							
FPGNN	0.831±0.011	0.831±0.011	0.904±0.020	0.892±0.019	0.732±0.068	0.661±0.014	0.833±0.004
TransFoxMol	0.780±0.032	0.780±0.032	0.907±0.024	0.881±0.015	0.830±0.047	0.636±0.022	0.816±0.011
ChemBERTa-2	0.848±0.037	0.848±0.037	0.932±0.037	0.892±0.019	0.933±0.054	0.708±0.090	0.809±0.029
MoLFormer	0.873±0.009	0.833±0.009	0.889±0.028	0.868±0.013	0.888±0.044	0.651±0.016	0.804±0.013
MolBERT	0.882±0.015	0.832±0.015	0.955±0.008	0.949±0.013	0.875±0.041	-	-
GROVER	0.779±0.059	0.779±0.059	0.849±0.008	0.823±0.020	0.685±0.066	0.635±0.034	0.808±0.014
Uni-Mol	0.840±0.031	0.84±0.031	0.889±0.025	0.886±0.016	0.818±0.065	0.666±0.021	0.812±0.007
MuMo	0.878±0.046	0.849±0.014	0.962±0.007	0.957±0.011	0.985±0.011	0.677±0.009	0.834±0.009

Models	LD50	Caco-2	PPBR	LIPO	Models	ESOL	Freesolv
TDC Datasets - Regression - MAE ↓					MoleculeNet - Regression - RMSE ↓		
AttentiveFP	0.678±0.012	0.401±0.032	9.373±0.335	0.572±0.007	ChemBERTa-2	0.633±0.132	1.219±0.206
FPGNN	0.638±0.024	0.326±0.040	8.465±1.709	0.544±0.011	FPGNN	0.658±0.006	1.106±0.195
DMPNN	0.607±0.022	0.388±0.077	8.158±0.314	0.448±0.014	GROVER	0.617±0.077	1.901±0.459
AttrMasking	0.685±0.025	0.546±0.052	10.075±0.202	0.547±0.024	MoLFormer	0.653±0.029	1.190±0.046
ContextPred	0.669±0.030	0.502±0.036	9.445±0.224	0.535±0.012	MolBERT	0.617±0.091	1.311±0.257
TranFoxMol	0.645±0.036	0.487±0.068	9.055±0.523	0.525±0.024	TranFoxMol	0.930±0.261	1.225±0.155
DeepMol	0.589±0.006	0.327±0.012	9.533±0.162	0.660±0.004	Uni-Mol	0.769±0.153	1.598±0.153
MuMo	0.426±0.031	0.315±0.055	7.324±0.323	0.448±0.007	**MuMo**	0.536±0.061	1.082±0.088

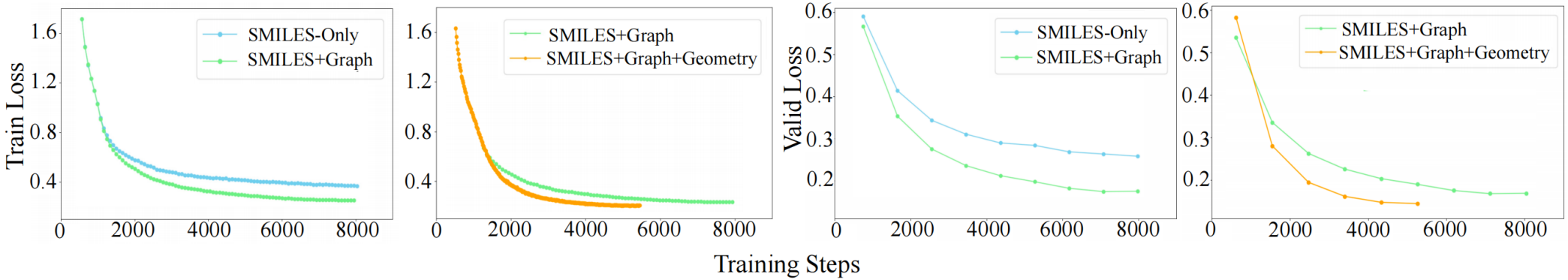
- Benchmarks: 29 tasks (TDC, MoleculeNet, Reaxtica).
- Overall: MuMo averages +2.7% over best baseline; rank-1 in 17/21 (TDC + MoleculeNet) and +27% on LD50 from TDC.
- QM datasets: wins 7/10 tasks, robust to conformer sensitivity.
- Broader chemical reaction benchmarks.

Model	HOMO/LUMO/GAP ↓	α ↓	C_v ↓	μ ↓↓	R^2 ↓	ZPVE ↓
GROVER-base	0.0079±3e-04	2.365±0.302	1.103±0.339	0.618±0.002	113.01±4.206	0.0035±3e-04
GROVER-large	0.0083±6e-04	2.240±0.385	0.853±0.186	0.623±0.006	85.85±6.816	0.0038±5e-04
GEM	0.0067±4e-05	0.589±0.0042	0.237±0.0137	0.444±0.0015	25.67±0.743	0.0011±2e-05
Uni-Mol	0.0043±2e-05	0.363±0.009	0.183±0.002	0.155±0.0015	4.805±0.055	0.0011±3e-05
Uni-Mol2 310M	0.0036±1e-05	0.315±0.003	0.143±0.002	0.092±0.0013	4.672±0.245	0.0005±1e-05
Uni-Mol2 570M	0.0036±2e-05	0.315±0.004	0.147±0.007	0.089±0.0015	4.523±0.080	0.0005±3e-05
Uni-Mol2 1.1B	0.0035±1e-05	0.305±0.003	0.144±0.002	0.089±0.0004	4.265±0.067	0.0005±8e-05
MuMo 505M	0.0030±1e-05	0.283±0.003	0.126±0.003	0.400±0.0018	18.08±0.533	0.0005±1e-05

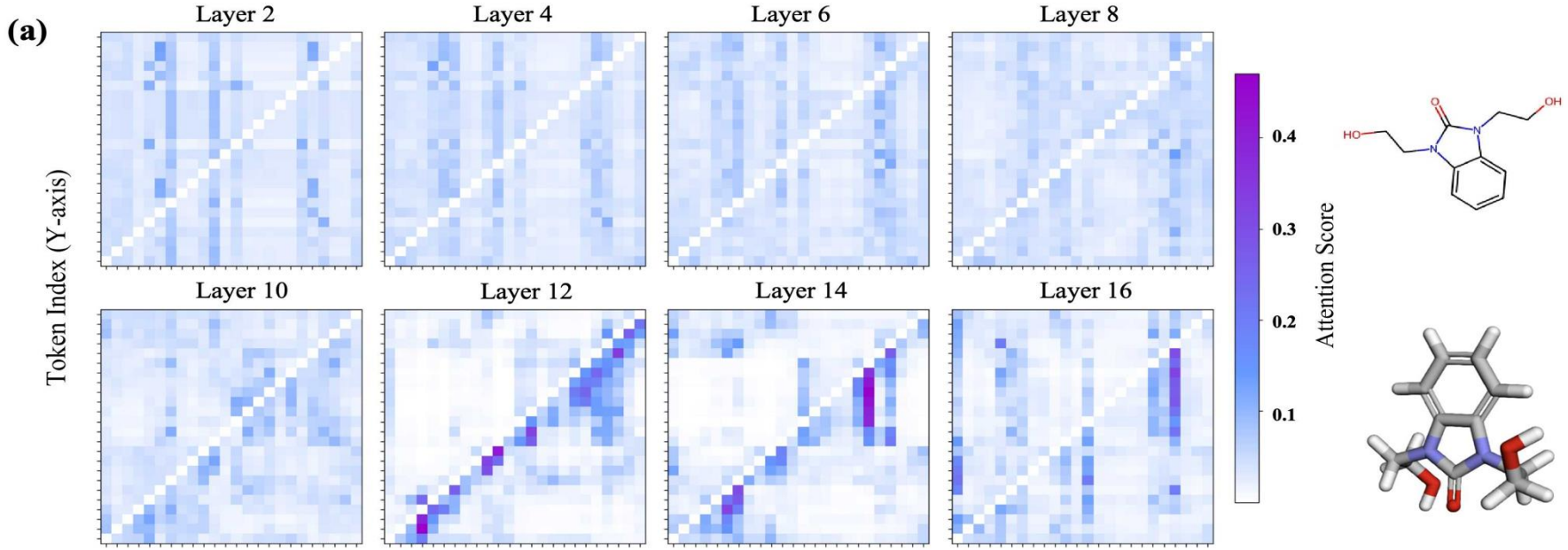
BHC (R^2 ↑, Reaction Yield)		CPA (MAE ↓, Catalytic Activity)		HTE (R^2 ↑, Reaction Yield)	
Models	Value	Models	Value	Models	Value
Reaxtica	0.94	Reaxtica	0.144	Reaxtica	0.87
MFF	0.92	MFF	0.144	rxnfp	0.81
rxnfp	0.95	Denmark et al.	0.152	DRFP	0.85
MuMo	**0.952**±0.002	**MuMo**	**0.144**±0.000	**MuMo**	**0.873**±0.002

Insights

- Pretraining losses show the effectiveness of modality fusion.



- Attention visualization. From Layer 9, the graph and geometry feature start interacting with the SMILES stream.



Welcome to connect and collaborate!