西湖大學
WESTLAKE UNIVERSITY

Accelerate Drug Discovery with Pre-learned Potential

Jun Xia @ Westlake Univ.
Advisor: Stan Z. Li (IEEE Fellow)

CONTENTS

01 Myself

Myself

Profile

Hi there! I am Jun Xia, a Ph.D. student at Westlake University and Zhejiang University, advised by Chair Prof. Stan Z. Li. Before Joining Westlake, I received my B.E. degree with honors from Central South University in 2020. My primary research interests lie in Graph Machine Learning, with special emphasis on its applications in Drug Discovery and Computational Biochemistry.

Jun Xia 夏俊

Ph.D. Student
School of Engineering
Westlake University & Zhejiang University
Email: xiajun@westlake.edu.cn
Advisor: Stan Z. Li (IEEE Fellow)

Academic Service

- Program Committee Member:
  - Conferences: ICLR, ICML, NeurIPS, KDD, ACL, SDM, ECML, ICASSP, etc.
- Journal Reviewer: IEEE TIP, ACM TKDD, IEEE TNNLS, Neural Networks, etc.

Awards & Honors

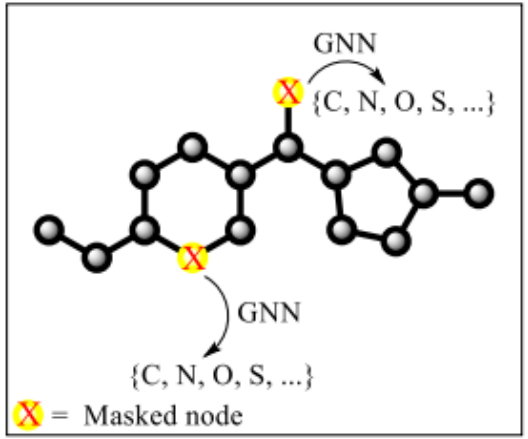
- 2023: Westlake Presidential Awards (The highest honor at Westlake Univ.).
- 2022: National Scholarship.
- 2022: ICML 2022 Participation Grant.
- 2021: Outstanding Student Cadre, Zhejiang University.
- 2021: Outstanding Student, Zhejiang University.

02

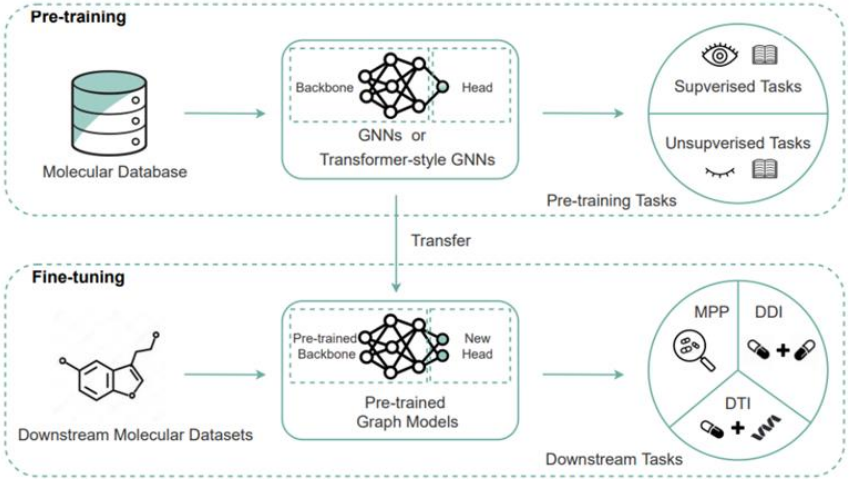
Mole-BERT (ICLR ’23)

Mole-BERT (ICLR ’23)

- **Pre-training on Molecules**



AttrMasking, ICLR’20

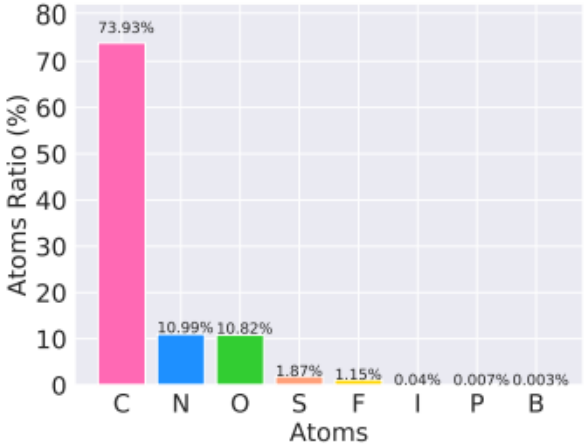


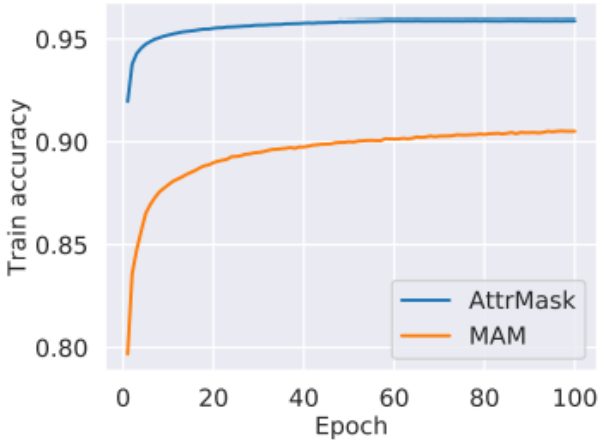
Pretraining-finetuning paradigm

Strategies for Pre-training Graph Neural Networks (Hu et al., ICLR 2020)

Mole-BERT (ICLR '23)

- Why the negative transfer issue would occur?



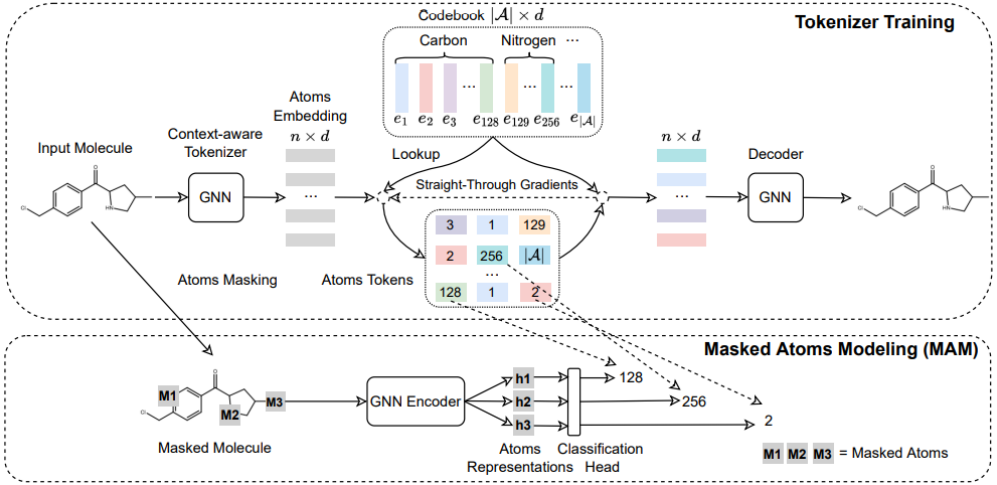


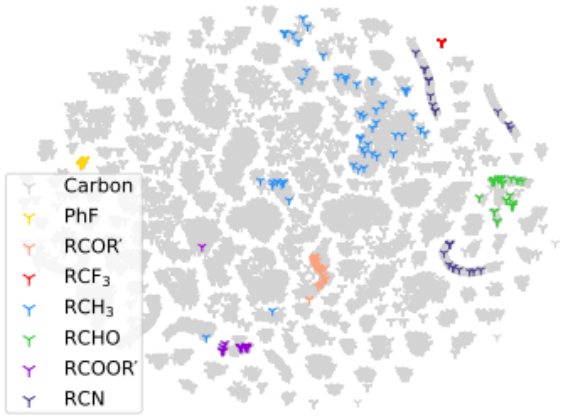
The atom vocabulary is extremely small and unbalanced

The pre-training task is too simple to learn informative representations

Mole-BERT (ICLR '23)

- **Node-level: Masked Atoms Modeling (MAM)**





The learned carbons' embeddings

Group VQ-VAE's Objective:

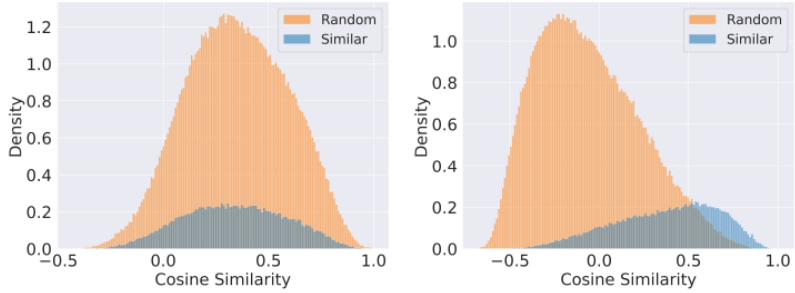
$$\mathcal{L}_{\mathtt{VQ}} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{v_i^T \widehat{v}_i}{\|v_i\| \cdot \|\widehat{v}_i\|}\right)^{\gamma} + \frac{1}{n}\sum_{i=1}^{n}\|\mathrm{sg}[h_i] - e_{z_i}\|_2^2 + \frac{\beta}{n}\sum_{i=1}^{n}\|\mathrm{sg}\left[e_{z_i}\right] - h_i\|_2^2$$

MAM's Objective:

$$\mathcal{L}_{\mathtt{MAM}} = -\sum_{\mathcal{G}\in\mathcal{D}}\sum_{i\in\mathcal{M}}\log p\left(z_i \mid \mathcal{G}^{\mathcal{M}}\right) \quad z_i = \mathrm{argmin}_j \left\|h_i - e_j\right\|_2$$

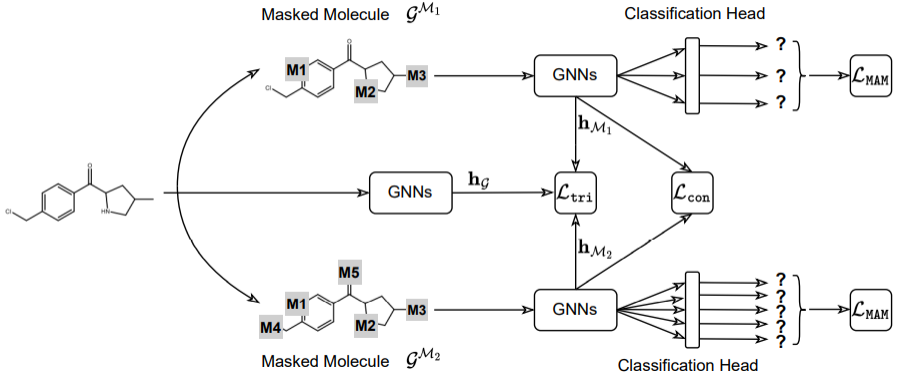
Mole-BERT (ICLR '23)

- **Triplet Masked Contrastive Learning (TMCL)**



(a) MAM (b) MAM + TMCL (Mole-BERT)

Similarity histograms of the learned representations



The general framework of Mole-BERT

$$\mathcal{L}_{\mathtt{tri}} = \sum_{\mathcal{G}\in\mathcal{D}} \max\left(sim\left(\mathbf{h}_{\mathcal{G}}, \mathbf{h}_{\mathcal{M}_2}\right) - sim\left(\mathbf{h}_{\mathcal{G}}, \mathbf{h}_{\mathcal{M}_1}\right) + m, 0\right)$$

TMCL's Objective: $\mathcal{L}_{\mathtt{TMCL}} = \mathcal{L}_{\mathtt{con}} + \lambda\mathcal{L}_{\mathtt{tri}}$, where $\mathcal{L}_{\mathtt{con}} = -\sum_{\mathcal{G}\in\mathcal{D}} \log \frac{e^{sim(\mathbf{h}_{\mathcal{M}_1}, \mathbf{h}_{\mathcal{M}_2})/\tau}}{\sum_{\mathcal{G}'\in\mathcal{B}} e^{sim(\mathbf{h}_{\mathcal{M}_1}, \mathbf{h}_{\mathcal{G}'})/\tau}}$

Mole-BERT's Objective: $\mathcal{L}_{\mathtt{Mole-BERT}} = \mathcal{L}_{\mathtt{MAM}} + \mathcal{L}_{\mathtt{TMCL}}$

Mole-BERT (ICLR '23)

- Results

1. ContextPred [1]

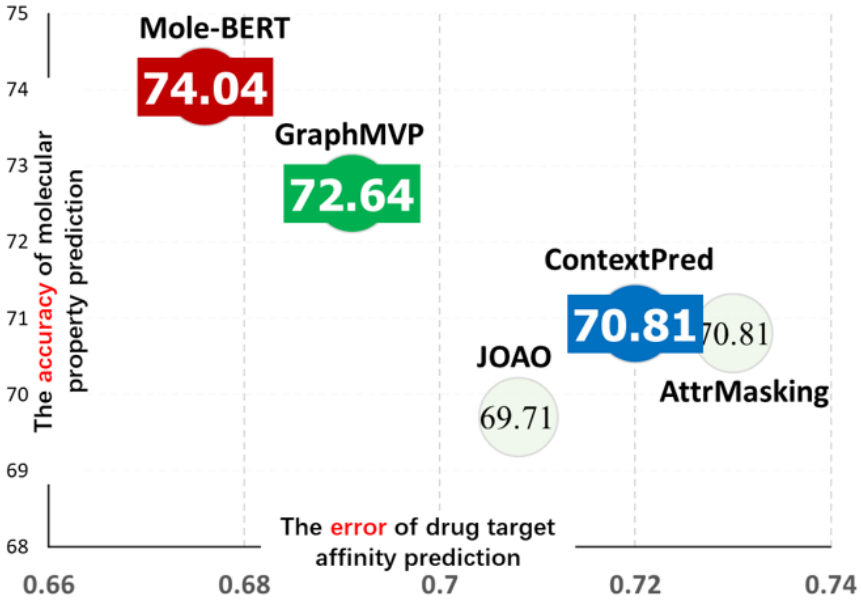
   (Stanford，2021.5)

2. GraphMVP [2]

   (Canada Mila Lab，2022.5)

3. Mole-BERT [3] (Ours)

   (Westlake University，2022.12)



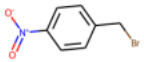
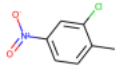
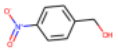
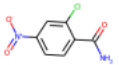
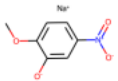
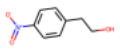
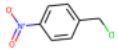
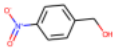
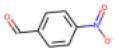
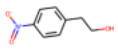
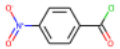
[1] W. Hu, B. Liu, and et al. Strategies for Pre-training Graph Neural Networks (ICLR 2020)

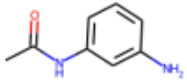
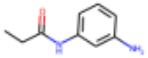
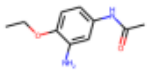
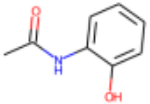
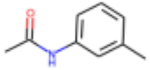
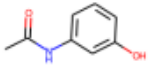
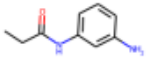
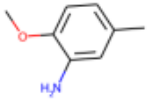
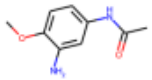
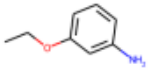
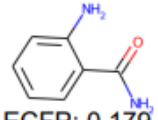
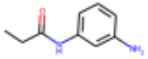
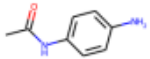
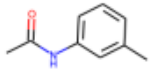
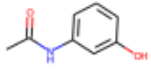
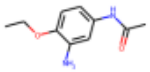
[2] S. Liu, H. Wang, and et al. Pre-training Molecular Graph Representation with 3D Geometry (ICLR 2022)

[3] J. Xia, C. Zhao, and **S. Z. Li**. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules (ICLR 2023)

Mole-BERT (ICLR '23)

- **Molecule Retrieval**

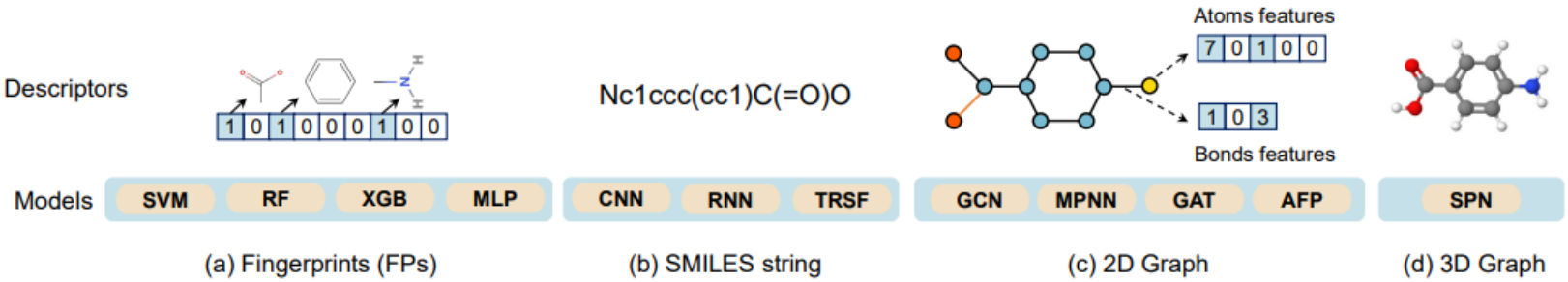
Query Molecule						
	MAM	 ECFP: 0.525	 ECFP: 0.692	 ECFP: 0.380	 ECFP: 0.328	 ECFP: 0.647
	Mole-BERT	 ECFP: 0.691	 ECFP: 0.692	 ECFP: 0.688	 ECFP: 0.647	 ECFP: 0.564

Query Molecule						
	Mole-BERT w/o $\mathcal{L}_{\mathtt{tri}}$	 ECFP: 0.863	 ECFP: 0.511	 ECFP: 0.464	 ECFP: 0.555	 ECFP: 0.511
	MAM	 ECFP: 0.863	 ECFP: 0.156	 ECFP: 0.561	 ECFP: 0.154	 ECFP: 0.179
	Mole-BERT	 ECFP: 0.863	 ECFP: 0.691	 ECFP: 0.555	 ECFP: 0.511	 ECFP: 0.511

03 Independent Feature Embedding (NeurIPS'23)

Independent Feature Embedding (NeurIPS'23)

Molecular Descriptors & Machine Learning Models



Independent Feature Embedding (NeurIPS'23)

西湖大學 WESTLAKE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Results & Observations

Dataset (No.)	Metric	SVM	XGB	RF	CNN	RNN	TRSF	MLP	GCN	MPNN	GAT	AFP	SPN
BACE (1,513)	AUC_ROC	0.886	**0.896**	0.890	0.815	0.559	0.835	0.887	0.880	0.846	0.886	0.879	0.882
HIV (40,748)	AUC_ROC	0.817	0.823	0.826	0.733	0.639	0.748	0.791	**0.834**	0.814	0.812	0.819	0.818
BBBP (2,035)	AUC_ROC	0.913	**0.926**	0.923	0.760	0.693	0.897	0.918	0.915	0.872	0.902	0.893	0.905
ClinTox (1,475)	AUC_ROC	0.879	0.919	0.933	0.685	0.813	**0.963**	0.890	0.889	0.868	0.891	0.907	0.912
SIDER (1,366)	AUC_ROC	0.626	0.638	**0.644**	0.591	0.515	0.641	0.617	0.633	0.603	0.614	0.620	0.613
Tox21 (7,811)	AUC_ROC	0.820	0.837	0.838	0.766	0.734	0.817	0.834	0.830	0.816	0.829	**0.845**	0.827
ToxCast (8,539)	AUC_ROC	0.725	0.785	0.778	0.735	0.74	0.780	0.781	0.767	0.736	0.768	**0.788**	0.772
MUV (93,087)	AUC_PRC	**0.093**	0.072	0.069	0.045	0.094	0.059	0.018	0.056	0.019	0.055	0.044	0.058
SARS-CoV-2 (14,332)	AUC_ROC	0.599	**0.700**	0.686	0.688	0.649	0.643	0.638	0.646	0.640	0.683	0.651	0.663
ESOL (1,127)	RMSE	0.676	**0.583**	0.647	2.569	1.511	0.718	0.653	0.773	0.695	0.661	0.594	0.671
Lipop (4,200)	RMSE	0.683	**0.585**	0.626	1.016	1.207	0.947	0.633	0.665	0.669	0.680	0.664	0.630
FreeSolv (639)	RMSE	1.063	**0.715**	1.014	2.275	2.205	1.504	1.046	1.316	1.327	1.304	1.139	1.159
QM7 (6,830)	MAE	**42.814**	52.726	51.403	81.165	158.160	64.363	86.060	64.530	107.013	78.217	59.973	55.727
QM8 (21,786)	MAE	0.0364	0.0126	**0.0098**	0.0205	0.0295	0.0232	0.0104	0.0154	0.0109	0.0187	**0.0098**	0.0103

Benchmarking results

Key observations:

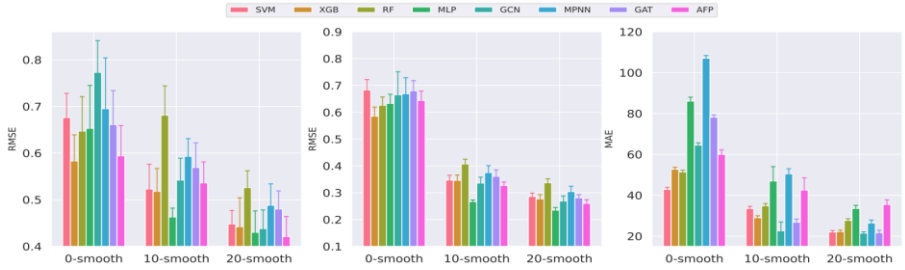
- Deep models underperform non-deep ones in most cases.
- It is irregular data patterns, NOT solely the small size of molecular datasets to blame for the failure of deep models!
- Tree models (XGB and RF) exhibit a particular advantage over other models.

Independent Feature Embedding (NeurIPS'23)

西湖大學 WESTLAKE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Explanation 1: **Deep models struggle to learn non-smooth target functions that map molecules to properties.**

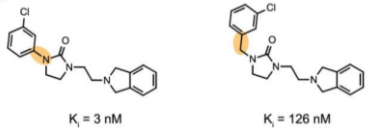


Results on smoothed datasets

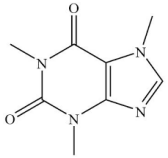
Target name (Response type)	Metric	SVM	XGB	RF	CNN	RNN	TRSF	MLP	GCN	MPNN	GAT	AFP
CB1 (Agonism EC_{50})	$RMSE_{nc}$	0.652	0.623	**0.619**	0.934	0.712	0.785	0.707	0.932	0.938	0.960	0.909
	$RMSE_{c}$	0.773	**0.767**	0.770	0.944	0.823	0.888	0.807	0.992	0.989	0.975	0.967
	$\Delta\mathcal{R}$	18.55%	23.11%	24.39%	1.15%	15.59%	13.12%	14.1%	6.37%	5.47%	1.55%	6.35%
DAT (Inhibition K_i)	$RMSE_{nc}$	0.589	0.579	**0.577**	0.871	0.692	0.801	0.664	0.927	0.820	0.995	0.865
	$RMSE_{c}$	0.744	**0.696**	0.730	0.894	0.783	0.934	0.792	1.003	0.921	1.042	0.995
	$\Delta\mathcal{R}$	26.30%	20.18%	26.64%	2.48%	13.15%	16.70%	19.40%	8.23%	12.38%	4.74%	15.11%
PPARα (Agonism EC_{50})	$RMSE_{nc}$	**0.535**	0.552	0.561	0.854	0.696	0.799	0.606	0.856	0.833	0.892	0.749
	$RMSE_{c}$	**0.671**	0.678	0.685	0.962	0.825	0.968	0.713	0.870	0.872	0.929	0.823
	$\Delta\mathcal{R}$	25.42%	22.83%	22.10%	12.69%	15.64%	21.26%	17.77%	1.72%	4.78%	4.21%	9.90%
DOR (Inhibition K_i)	$RMSE_{nc}$	0.598	0.592	**0.591**	0.938	0.893	0.873	0.663	1.095	0.958	1.102	1.018
	$RMSE_{c}$	0.861	0.854	**0.836**	1.098	1.036	1.032	0.874	1.259	1.152	1.281	1.179
	$\Delta\mathcal{R}$	43.98%	44.14%	41.46%	17.06%	16.01 %	18.26%	31.85%	14.93%	20.27%	16.26%	15.83%

Results on activity cliffs

The superiority of deep models



Activity cliffs



Chemical molecule

Image

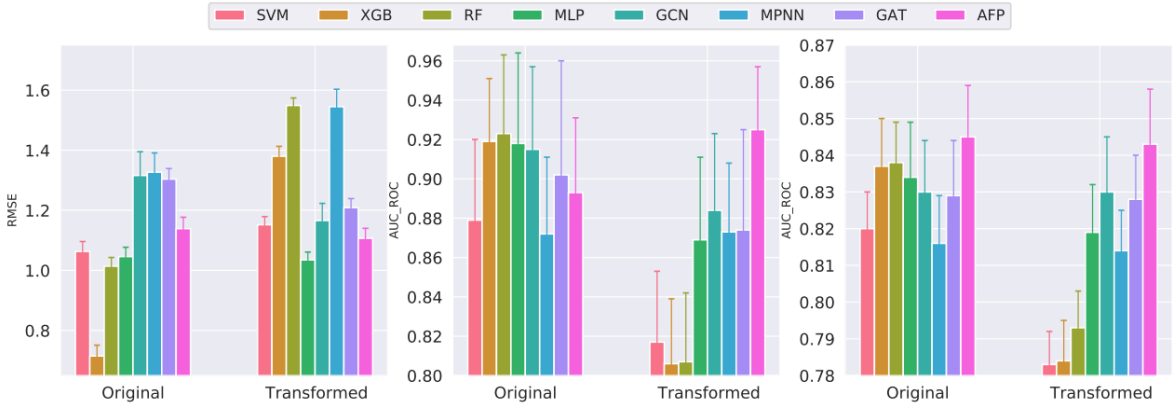
Data smoothing level

Independent Feature Embedding (NeurIPS'23)

西湖大學 WESTLAKE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Explanation 2: **Deep models mix different dimensions of molecular features, whereas tree models make decisions based on each dimension of the features separately.**



Results on orthogonally transformed datasets

$$\widehat{x}_i = \boxed{\mathcal{Q}} x_i, \quad y_i = W^\top x_i + b, \quad \mathcal{Q}^{-1} = \mathcal{Q}^\top \quad y_i = \widehat{W}^\top \widehat{x}_i + b = \widehat{W}^\top \mathcal{Q} x_i + \hat{b}, \ \widehat{W} = \mathcal{Q} W$$

Orthogonal matrix

Independent Feature Embedding (NeurIPS'23)

西湖大學 WESTLAKE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Independent Feature Embedding

$$f_x = [\sin(v) || \cos(v)], \quad v = [2\pi c_1 x, \ldots, 2\pi c_k x]$$

Deep models can be approximated as Neural Tangent Kernel

$$h_{\mathrm{NTK}}\left(f_{x_i}^{\mathrm{T}} f_{x_j}\right) = h_{\mathrm{NTK}}\left(g_{\boldsymbol{c}}\left(x_i - x_j\right)\right)$$

$$k_{\mathrm{NTK}}(x_i, x_j) = \mathbb{E}_{\theta \sim \mathcal{N}} \left\langle \frac{\partial f(x_i;\theta)}{\partial \theta}, \frac{\partial f(x_j;\theta)}{\partial \theta} \right\rangle$$

IFM creates a tunable stationary NTK

$$f_{x_i} \cdot f_{x_j} = \sum_{i=1}^{k} \cos(2\pi c_i (x_i - x_j)) := g_{\boldsymbol{c}}(x_i - x_j),$$

Independent Feature Embedding (NeurIPS'23)

西湖大學 WESTLAKE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

- **IFM improves deep models' performance on molecules**

Dataset (No.)	Metric	MLP	GCN	MPNN	GAT	AFP	P-Best (Model)	IFM-MLP	IFM-GCN	IFM-MPNN	IFM-GAT	IFM-AFP
BACE (1,513)	AUC_ROC	0.887	0.880	0.846	0.886	0.879	0.896 (XGB)	**0.915**	0.903	0.866	0.894	0.907
HIV (4,0748)	AUC_ROC	0.791	0.834	0.814	0.812	0.819	0.834 (GCN)	0.816	**0.862**	0.846	0.838	0.849
BBBP (2,035)	AUC_ROC	0.918	0.915	0.872	0.902	0.893	0.926 (XGB)	0.937	**0.945**	0.908	0.933	0.940
ClinTox (1,475)	AUC_ROC	0.890	0.889	0.868	0.891	0.907	**0.963** (TRSF)	0.941	0.938	0.929	0.953	0.959
SIDER (1,366)	AUC_ROC	0.617	0.633	0.603	0.614	0.620	0.644 (RF)	0.646	0.649	0.638	0.647	**0.652**
Tox21 (7,811)	AUC_ROC	0.834	0.830	0.816	0.829	0.845	0.845 (AFP)	0.842	0.839	0.837	0.849	**0.853**
ToxCast (8,539)	AUC_ROC	0.781	0.767	0.736	0.768	0.788	0.788 (AFP)	0.795	0.790	0.772	0.797	**0.806**
MUV (93,087)	AUC_PRC	0.018	0.056	0.019	0.055	0.044	0.093 (SVM)	0.052	0.113	0.068	**0.124**	0.097
SARS-CoV-2 (14,332)	AUC_ROC	0.638	0.646	0.640	0.683	0.651	0.700 (XGB)	0.675	0.682	0.686	**0.716**	0.704
ESOL (1,127)	RMSE	0.653	0.773	0.695	0.661	0.594	0.583 (XGB)	0.587	0.728	0.673	0.566	**0.561**
Lipop (4,200)	RMSE	0.633	0.665	0.669	0.680	0.664	0.585 (XGB)	**0.556**	0.577	0.568	0.584	0.578
FreeSolv (639)	RMSE	1.046	1.316	1.327	1.304	1.139	**0.715** (XGB)	0.862	0.916	0.911	0.908	0.883
QM7 (6,830)	MAE	86.060	64.530	107.013	78.217	59.973	42.814 (SVM)	66.570	38.793	84.918	59.595	**33.775**
QM8 (21,786)	MAE	0.0104	0.0154	0.0109	0.0187	0.0098	0.0098 (AFP)	0.0091	0.0114	0.0085	0.0139	**0.0079**

Results on Normal Molecular Datasets

Target name (Response type)	Metric	MLP	GCN	MPNN	GAT	AFP	P-Best (Model)	IFM-MLP	IFM-GCN	IFM-MPNN	IFM-GAT	IFM-AFP
CB1 (Agonism EC_{50})	$RMSE_c$	0.807	0.992	0.989	0.975	0.967	0.767 (XGB)	**0.715**	0.748	0.756	0.741	0.746
DAT (Inhibition K_i)	$RMSE_c$	0.792	1.003	0.921	1.042	0.995	0.696 (XGB)	**0.646**	0.682	0.673	0.665	0.670
PPARα (Agonism EC_{50})	$RMSE_c$	0.713	0.870	0.872	0.929	0.823	0.671 (SVM)	0.623	0.634	0.649	0.661	**0.616**
DOR (Inhibition K_i)	$RMSE_c$	0.874	1.259	1.152	1.281	1.179	0.836 (RF)	**0.787**	0.813	0.796	0.799	0.810

Results on Activity Cliff Cases

西湖大學
WESTLAKE UNIVERSITY

THANKS

Jun Xia @ Westlake Univ.
Homepage: https://junxia97.github.io/