Patterns, Volume *3*

Supplemental information

Quantitative evaluation of explainable graph neural networks for molecular property prediction

Jiahua Rao, Shuangjia Zheng, Yutong Lu, and Yuedong Yang

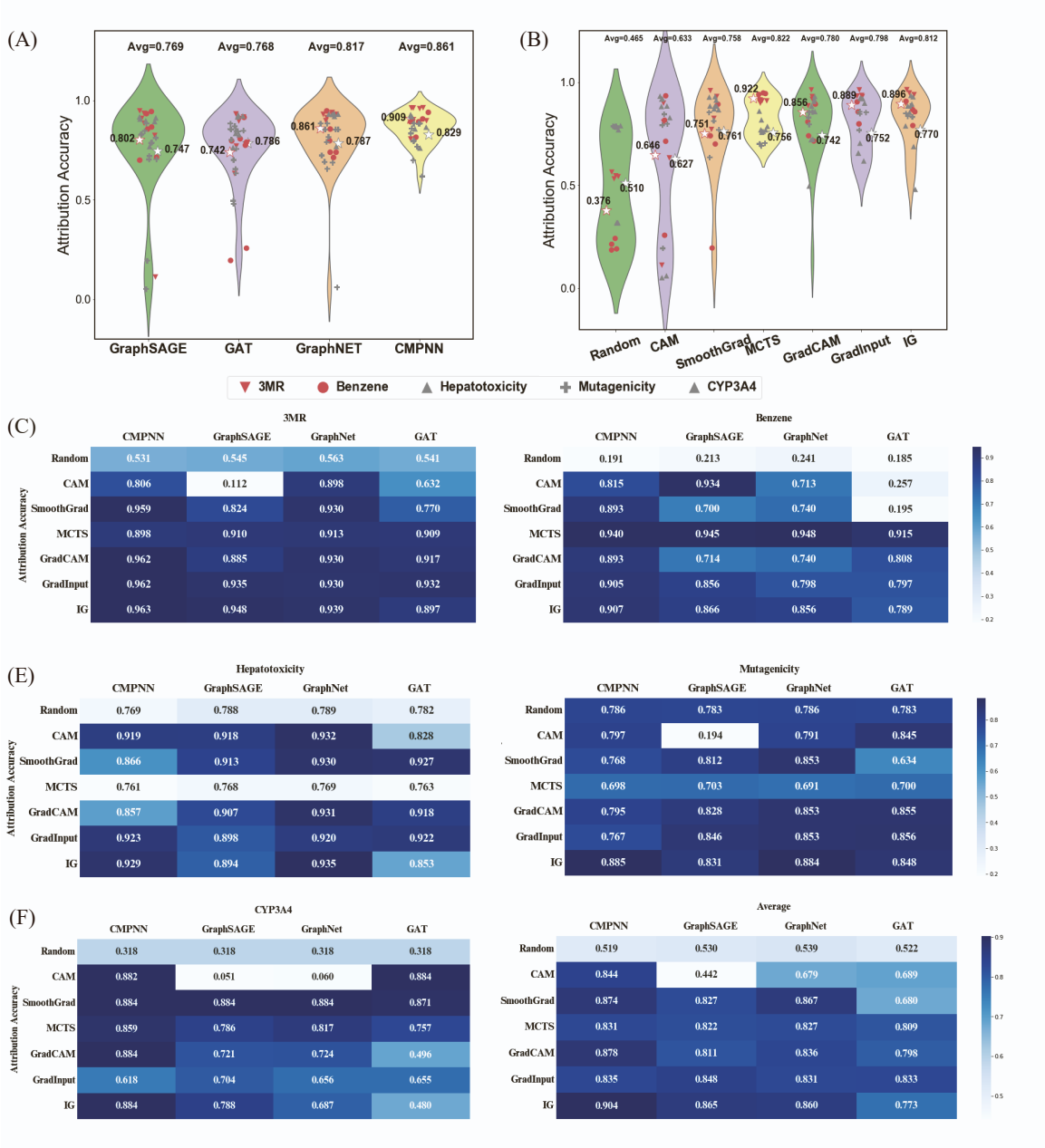


Figure S1. The average Attribution-Accuracy values. (A) four GNN models and (B) six XAI methods. (C-F) Detailed Attribution Accuracies for combinations of all XAI methods and GNN models for each dataset.

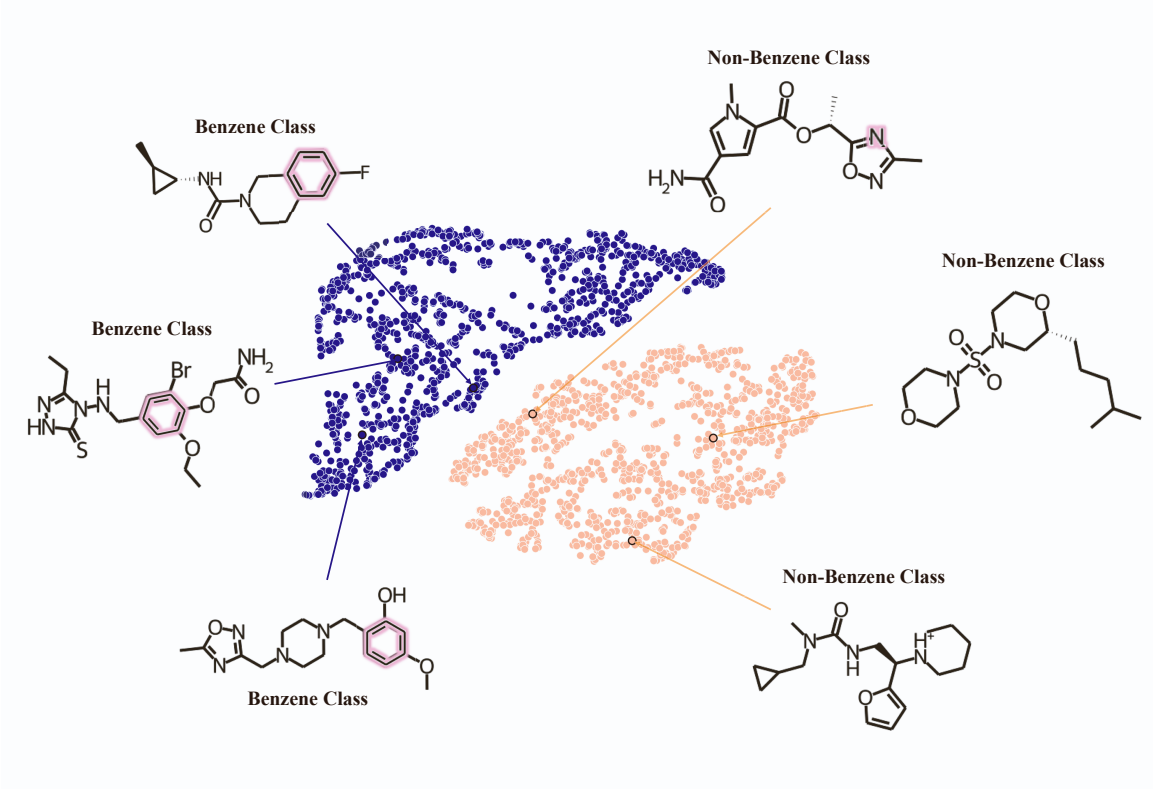


Figure S2. Visualization of Benzene dataset. The embeddings learned by CMPNN+IG over the Benzene dataset shown by t-SNE, together with representatively 3 benzene and 3 non-benzene molecules.

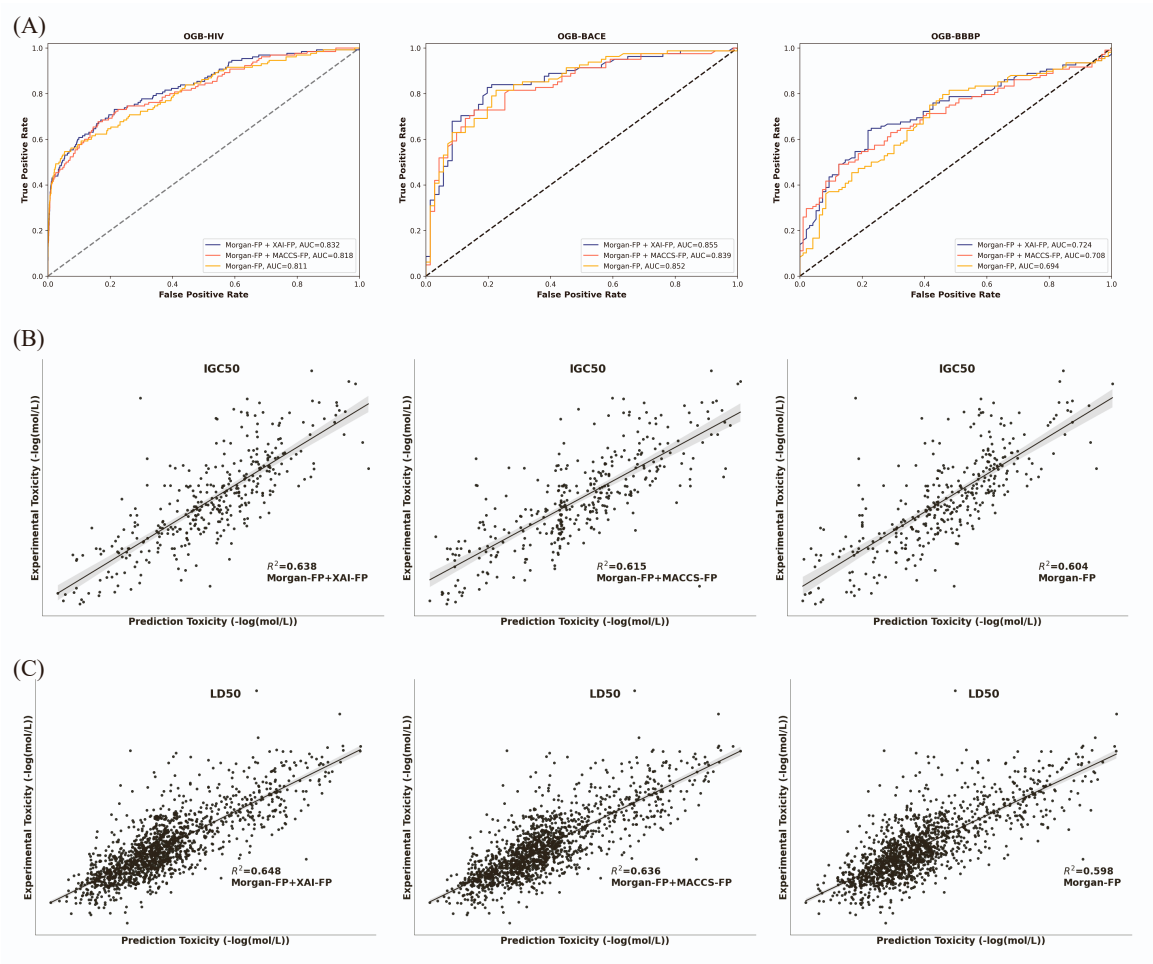


Figure S3. Results of XAI-assisted FPs experiments with Random Forest. (A) The ROC curves of Morgan-FP, Morgan-FP + MACCS-FP, and Morgan-FP + XAI-FP with Random Forest on the OGB-HIV, OGB-BACE and OGB-BBBP dataset. (B) The predicted results of Morgan-FP + XAI-FPs, Morgan-FP + MACCS-FP and Morgan-FP with Random Forest model for the IGC50 dataset. (C) The predicted results of Morgan-FP + XAI-FPs, Morgan-FP + MACCS-FP and Morgan-FP with Random Forest model for the LD50 dataset.

Table S1. Predictive Performance comparison of baseline methods (from previous literature) with our GNN models on FreeSolv, Lipophilicity, ESOL, ClinTox, BBBP benchmarks.

Datasets	FreeSolv	Lipophilicity	ESOL	ClinTox	BBBP
Methods	R^2	R^2	R^2	AUROC	AUROC
ECFP4	0.640[a]	0.577[a]	0.778[a]	0.640[a]	0.955[a]
AGBT-FP	0.935[b]	0.776[b]	-	0.935[b]	0.763[b]
MPNN	0.923[a]	0.697[a]	**0.939[a]**	0.923[a]	0.960[a]
SSL-FP	0.678[c]	0.740[c]	0.925[c]	0.688[c]	0.953[c]
GraphSAGE	0.922	0.688	0.929	0.933	0.962
GAT	0.919	0.685	0.924	0.889	0.959
GraphNET	**0.941**	0.798	0.919	0.938	0.961
CMPNN	0.933	**0.823**	0.932	**0.941**	**0.963**

[a]The prediction results are derived from Wu et al[1].
[b]The prediction results are derived from Chen et al[2].
[c]The prediction results are derived from Chen et al[3].

Table S2. Predictive Performance comparison of baseline methods (from previous literature) with our GNN models on LD50, IGC50, LC50, LC50DM, LogP datasets.

Datasets	LD50	IGC50	LC50	LC50DM	LogP
Methods	R^2	R^2	R^2	R^2	R^2
ECFP4	0.586[a]	0.647[a]	0.573[a]	0.452[a]	0.857[a]
MACCS	0.643[a]	0.643[a]	0.608[a]	0.434[a]	0.867[a]
HybridModel	0.629[a]	0.810[a]	0.678[a]	0.616[a]	0.893[a]
AGBT-FP	0.671[a]	**0.842[a]**	0.783[a]	**0.830[a]**	0.905[a]
GraphSAGE	0.654	0.778	0.731	0.724	0.904
GAT	0.662	0.809	0.742	0.798	0.899
GraphNET	**0.683**	0.834	**0.793**	0.810	0.923
CMPNN	0.679	0.841	0.776	0.813	**0.939**

[a]The prediction results are derived from Chen et al[2].

Table S3. Explainability Performance comparison of baseline methods (RF and XGBoost) with GNN models (explained by IG).

Datasets	3MR	Benzene	Hepatotoxicity	Mutagenicity
Methods	Attribution-ACC			
RF + Feature Importances	0.516	0.621	0.523	0.607
XGBoost + SHAP Values	0.678	0.740	0.701	0.688
GraphSAGE + IG	0.948	0.866	0.894	0.831
GAT + IG	0.897	0.789	0.853	0.848
GraphNET + IG	0.939	0.856	**0.935**	0.884
CMPNN + IG	**0.963**	**0.907**	0.929	**0.885**

Table S4. The predicted substructures by representative combinations of GNN models and XAI methods.

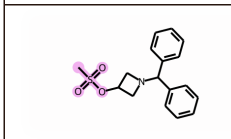
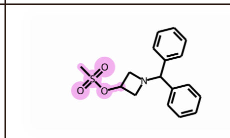
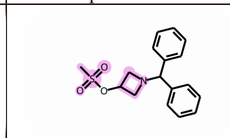
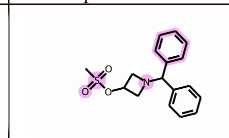
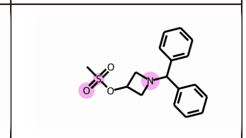
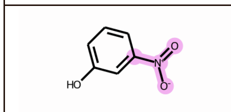
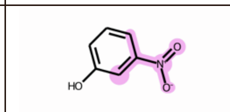
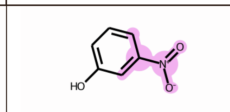
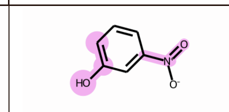
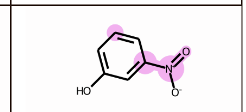
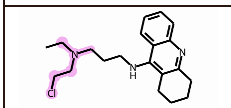
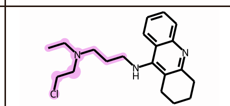
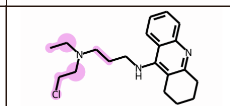
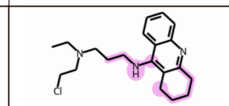
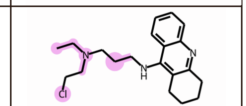
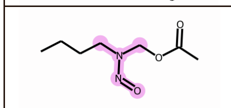
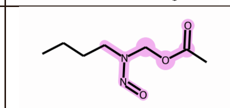
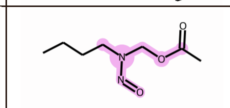
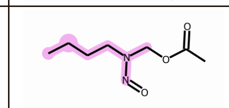
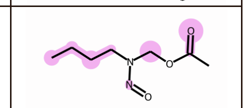
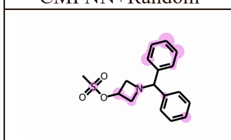
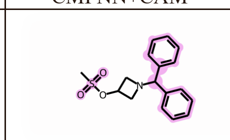
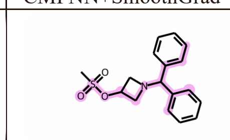
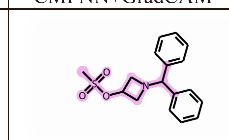
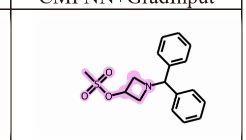
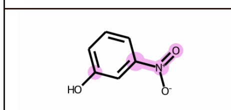
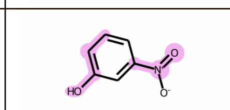
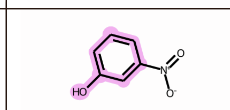
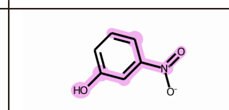
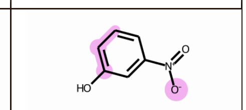
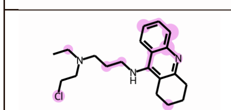
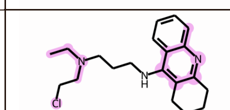
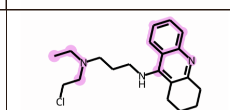
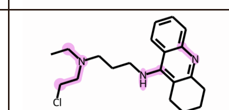
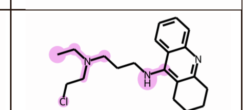
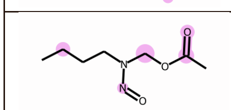
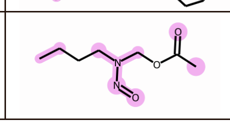
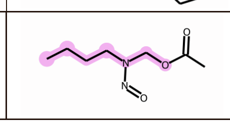
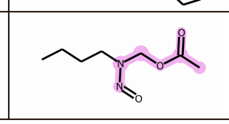
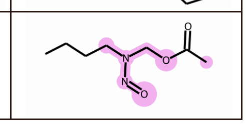
Ground-truth	CMPNN+IG	GraphNET+IG	GraphSAGE+IG	GAT+IG
				
				
				
				
CMPNN+Random	CMPNN+CAM	CMPNN+SmoothGrad	CMPNN+GradCAM	CMPNN+GradInput
				
				
				
				

Table S5. The Performance comparison of XAI method and medicinal chemists.

	Prediction Performance		**Explanation Performance**	
	Accuracy	Precision	Attribution-ACC	Attribution-Precision
Random	-	-	0.568	0.123
XAI	**0.92**	0.800	0.839	0.431
MC1	0.68	**0.842**	0.820	0.352
MC2	0.60	0.596	**0.852**	**0.471**
MC3	0.62	0.667	0.801	0.232
MC4	0.50	0.559	0.804	0.180
MC5	0.58	0.595	0.794	0.294
MC6	0.54	0.594	0.759	0.197
MC7	0.62	0.656	0.804	0.220

Table S6. Statistics of XAI-Benchmarks.

Benchmark	Task	Dataset Name	Compounds	Substructure
Synthetic Benchmarks	Classification	3MR	3152	Three-membered ring
	Classification	Benzene	12000	Benzene ring
Experimental Benchmarks	Classification	Mutagenicity	6506	Mutagenicity alerts
	Classification	Hepatotoxicity	587	Hepatotoxic alerts
	Classification	CYP450	9122	Substructure differences

Table S7. Atom features and Bond features.

Features	Description	Size
Atom		
Atom type	Type of atom (ex.C,N,O), by atomic number.	100
# Bonds	Number of bonds the atom is involved in.	6
Formal charge	Integer electronic charge assigned to atom.	5
Chirality	Unspecified, tetrahedral CW/CCW, or other.	4
# Hs	Number of bonded Hydrogen atom.	5
Hybridization	sp, sp^2, sp^3, sp^3d, or sp^3d^2	5
Aromaticity	Whether this atom is part of an aromatic system.	1
Atomic mass	Mass of the atom, divided by 100.	1
Bond		
Bond type	Single, double, triple, or aromatic.	4
Conjugated	Whether the bond is conjugated.	1
In ring	Whether the bond is part of a ring.	1
Stereo	None, any, E/Z or cis/trans.	6

Table S8. Hyperparameter ranges for four types of GNN models.

Model	Hyperparameters	Range
GraphSAGE	Batch size	{32, 64}
	Learning rate	{0.0001, 0.001, 0.01}
	Embedding size	{32, 64, 128}
	Num of layer	{3, 5}
	Aggregation type	{Add, Mean, Max}
	Num of epoch	{50, 100, 300}
GAT	Batch size	{50, 64}
	Learning rate	{0.0001, 0.001, 0.01}
	L2 regularization	{0.0, 0.1}
	Embedding size	{128, 256}
	Num of layer	{2, 4}
	Attention head	{4, 8}
	Dropout	{0.0, 0.1}
	Num of epoch	{50, 100, 300}
GraphNET	Batch size	{32, 64}
	Learning rate	{0.001, 0.01}
	L2 regularization	{0.0, 0.1}
	Node embedding size	{64, 256}
	Edge embedding size	{32, 64}
	Num of layer	{2, 4}
	Num of Set2Set layer	{2, 3}
	Aggregation type	{Add, Mean, Max}
	Num of epoch	{30, 100, 300}
CMPNN	Batch size	{50, 64}
	Learning rate	{0.0001, 0.001}
	L2 regularization	{0.0, 0.1}
	Embedding size	{256, 300}
	Dropout	{0.0, 0.1}
	Num of layer	{2, 4}
	Undirected	{False, True}
	Num of FFN layer	{2, 5}
	Embedding size of FFN layer	{256, 512}
	Num of epoch	{30, 50}

Table S9. The summary of public benchmark for XAI-assisted Fingerprints Experiments.

Task	Dataset	Compounds	Train	Test
Classification	OGB-BBBP	1835	1631	204
	OGB-BACE	1362	1210	152
	OGB-HIV	37014	32901	4113
Regression	IGC50	1792	1434	358
	LD50	7413	5931	1482

Table S10. Experiment results of XAI-assisted Fingerprints on Hepatotoxicity and Ames Mutagenicity datasets.

	Hepatotoxicity		Ames Mutagenicity	
Metric	ACC	%increase	AUROC	%increase
Morgan-FP	0.517	-	0.871	-
Morgan-FP + MACCS-FP	0.526	1.74%	0.893	2.52%
Morgan-FP + XAI-FP	**0.548**	**5.99%**	**0.904**	**3.79%**

Table S11. Hyperparameter ranges for Random Forest in XAI-assisted Experiments.

Model	Hyperparameters	Range
Random Forest	n_estimators	{100, 200, 500, 800, 1000}
	max_depth	{5, 6, 7, 8}
	min_samples_split	{2, 3, 4}
	min_samples_leaf	{1, 3, 5}
	min_weight_fraction_leaf	{0.0, 0.1, 0.2}
	max_leaf_nodes	{None, 1, 5}

Supplemental References

[1]. Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., and Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. Chem. Sci. *9*, 513–530. 10.1039/c7sc02664a.

[2]. Chen, D., Gao, K., Nguyen, D.D., Chen, X., Jiang, Y., Wei, G.W., and Pan, F. (2021). Algebraic graph-assisted bidirectional transformers for molecular property prediction. Nat. Commun. *12*. 10.1038/s41467-021-23720-w.

[3]. Chen, D., Zheng, J., Wei, G.W., and Pan, F. (2021). Extracting Predictive Representations from Hundreds of Millions of Molecules. J. Phys. Chem. Lett. *12*, 10793–10801. 10.1021/acs.jpclett.1c03058.