Supplementary Material

Experiment Details

Data Preprocessing We preprocessed the input SMILES strings to Structure-Data Files (SDFs). The processed SDFs can be found at https://figshare.com/articles/dataset/Well-curated_QSAR_datasets_for_diverse_protein_targets/20539893. The dataset is specified by its PubChem Assay ID (AID) (Wang et al. 2012). Prepossessing to the original data includes converting SMILES strings to 3D SDF files, generating 3D conformation, and filtering. Conversion from SMILES to SDF files is done using Open Babel (O'Boyle et al. 2011), version 2.4.1. Conformations are generated using Corina (Gasteiger, Rudolph, and Sadowski 1990), version 4.3. Molecules are further filtered with validity, duplicates with BioChemical Library (BCL) (Brown et al. 2022).

Data Split The datasets are randomly split into 80%/10%/10% for training, validation, and testing respectively. We then shrink the training set to contain only 10,000 inactive-labeled molecules, while keeping all active-labeled molecules. This shrinking technique was previously used by (Mendenhall and Meiler 2016) By shrinking the training data size, we can shorten the training time given the limited computational resources, while keeping most active signal because only those predicted active molecules will be experimentally validated and hence we focus on those. We did an empirical study on the shrinking effect on AID 2258 (302,402 molecules). Results are in Table 1. We can see there is indeed a decrease of performance in terms of logAUC$_{[0.001,0.1]}$. We leave the benchmarking of the full dataset in a future study.

Training Details To overcome the highly-imbalanced problem, we sample the training data in each batch according to the inverse frequency of the label occurence in the training set. For example, if the active label appear at 1% rate in the training set, it has a sampling weight of 1/0.01 =100, while if the inactive label appear 99% of time in the training set, it gets a sampling weight of 1/0.99 $\approx$ 1.01. The active-labeled data are thus roughly 100 times more likely to be sampled than inactive-labeled data in each batch.

Hyperparameters Details

The hyperparameters search space can be seen in Table 2.

For the actual hyperparameters used for training MolKGNN, see Tables 3 for details. For other benchmarking models except KerGNN, we use the same hyperparameters from their codes.

For KerGNN, we empirically observe that using the default hyperparameter setting achieves significantly low performance on our well-curated datasets and hence we further tune its hyperparamters as follows: batch size $\{64, 128\}$, the hidden unit of the linear layer $\{16, 32\}$.

Featurization

Different models have different ways of featurization. We use the original features reported in the original papers for each model used in the benchmarking. Our featurization is adapted from (Coley et al. 2017). Rdkit(version 2022.3.4) (Landrum et al. 2013) is used for the featurization. See Table 4 and 5 for details.

2.5D vs 3D

While many previous work have attempted to develop 3D models by including distance, angles, torsions into their model designs (Schütt et al. 2017; Klicpera, Groß, and Günnemann 2020; Liu et al. 2021), we demonstrate that 2.5D model can achieve comparable results in terms of AUC, or even better results in terms of logAUC$_[0.001, 0.1]$. We provide the explanation of why a model with seemly less information can accomplish this from a chemistry

Table 1: Results comparison between shrinked and full training set using MolKGNN over three runs for AID 2258.

Inactive Training Size	**logAUC**$_{[0.001,0.1]}$	**AUC**
10K Sample	0.296 ± 0.026	0.820 ± 0.021
Full	0.384 ± 0.003	0.816 ± 0.030

Table 2: Hyperparameter search space used for MolKGNN.

Hyperparameter	**Search Space**
Hidden Dimension	{32, 64}
Batch Size	{16, 32}
# Layers	{1, 2, 3, 4, 5}
Peak Learning Rate	{5e-1, 5e-2, 5e-3, 5e-4}
Dropout	{0.1, 0.2, 0.3}

Table 3: Hyperparameters used for MolKGNN.

Hyperparameter	**Value**
Node Feature Dimension	28
Edge Feature Dimension	7
Hidden Dimension	32
Batch Size	16
# Layers	4
# of Kernels of Degree 1	10
# of Kernels of Degree 2	20
# of Kernels of Degree 3	30
# of Kernels of Degree 4	50
Warmup Steps	300
Peak Learning Rate	5e-3
End Learning Rate	1e-10
Weight Decay	0.001
Epochs	20
Dropout	0.2

Table 4: Node features $\mathbf{X}_v$ for v

Indices	Description
0-11	One-hot encoding of element type: H, C, N, O, F, Si, P, S, Cl, Br, I, other
12-15	One-hot encoding of node degree: 1, 2, 3, 4
16	Formal charge
17	Is in a ring
18	Is aromatic
19	Explicit valence
20	Atom mass
21	Gasteiger charge
22	Gasteiger H charge
23	Crippen contribution to logP
24	Crippen contribution to molar refractivity
25	Total polar sufrace area contribution
26	Labute approximate surface area contribution
27	EState index

Table 5: Edge features $\mathbf{E}_{vu}$ for e_{vu}

Indices	Description
0	Is aromatic
1	Is conjugate
2	Is in a ring
3-6	One-hot encoding of bond type: 1, 1.5, 2, 3

perspective: Molecules can have different conformations as a result of the single bond rotation. The same molecule with different conformation consequently has different sets of torsions. However, the pharmacological activity is usually linked with few conformations (binding conformations) and hence related to certain sets torsions. It seems that knowing torsion could potentially help the activity prediction. Nevertheless, knowing which conformation is the binding conformation is a challenging task. A set of torsions related with a wrong predicted binding conformation is detrimental to the model performance. Hence we decide to build a conformation-invariant model and exclude torsion to circumvent this problem.

References

Brown, B.; Vu, O.; Geanes, A. R.; Kothiwale, S.; Butkiewicz, M.; Lowe, E. W.; Mueller, R.; Pape, R.; Mendenhall, J.; and Meiler, J. 2022. Introduction to the BioChemical Library (BCL): An application-based open-source toolkit for integrated cheminformatics and machine learning in computer-aided drug discovery. *Frontiers in pharmacology*, 341.

Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; and Jensen, K. F. 2017. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8): 1757–1772.

Gasteiger, J.; Rudolph, C.; and Sadowski, J. 1990. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Computer Methodology*, 3(6): 537–547.

Klicpera, J.; Groß, J.; and Günnemann, S. 2020. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*.

Landrum, G.; et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*.

Liu, Y.; Wang, L.; Liu, M.; Lin, Y.; Zhang, X.; Oztekin, B.; and Ji, S. 2021. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations*.

Mendenhall, J.; and Meiler, J. 2016. Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout. *Journal of computer-aided molecular design*, 30(2): 177–189.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; and Hutchison, G. R. 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1): 1–14.

Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; et al. 2012. PubChem's BioAssay database. *Nucleic acids research*, 40(D1): D400–D412.