

**Supporting Information of**  
**Advancing Bioactivity Prediction through Molecular**  
**Docking and Self-Attention: The Drug-Target Interaction**  
**Graph Neural Network (DTIGN)**

Yueming Yin<sup>1</sup>, Yuguang Mu<sup>2</sup>, Hoi Yeung Li<sup>2</sup>, and Adams Wai-Kin Kong <sup>\*1</sup>

<sup>1</sup>School of Computer Science and Engineering, Nanyang Technological University  
639798, Singapore.

<sup>2</sup>School of Biological Sciences, Nanyang Technological University 637551,  
Singapore.

January 13, 2024

---

<sup>\*</sup>Corresponding author: adamskong@ntu.edu.sg

## Contents

<b>A</b>	<b>Bioactivity Data Collection and Preprocessing</b>	<b>S-3</b>
<b>B</b>	<b>Comparing AutoDock Vina and QVina-W on the Learnability of their Generated Data</b>	<b>S-3</b>
<b>C</b>	<b>Experimental Settings</b>	<b>S-5</b>
<b>D</b>	<b>Benchmark Test on RMSE and Kendall's <math>\tau_B</math></b>	<b>S-8</b>
<b>E</b>	<b>Native and Docked structure on the pocket of attention</b>	<b>S-8</b>
<b>F</b>	<b>Correlation between attention values and docking accuracies</b>	<b>S-9</b>

## A Bioactivity Data Collection and Preprocessing

**Data Source.** There are numerous publicly available datasets containing recorded bioactivity information, with some of the larger ones being ChEMBL, PubChem, and ZINC. The data for this study was collected from ChEMBL<sup>1</sup> [1], a resource dedicated to compiling bioactive molecules possessing drug-like properties. ChEMBL encompasses over 2 million molecules and involves the analysis of 1.5 million biological/cellular and pharmacological assays across 15k protein targets. These protein targets span a diverse range of organisms, from humans to single-cell organisms, with around 6.778k proteins attributed to humans (as of April 2023). To extract bioactivity data, we filtered for Single Proteins (as distinct from Protein Compounds or Protein Families) that possess more than 1000 bioactive small molecules from the ChEMBL repository. Among the 6,778 human proteins cataloged in ChEMBL, only 900 fulfilled the criteria for further investigation. The collected information includes various bioactivity measurements (such as IC<sub>50</sub>, EC<sub>50</sub>, K<sub>d</sub>, K<sub>i</sub>, %inhibition, potency, etc.) and molecular representations (SMILES, ChEMBL Identifier), totaling 6.75 million records.

**Data Preprocessing.** To render the raw data into a format suitable for computer processing, we underwent the following preprocessing steps. Initially, molecules with “Null” assay values were removed. Then, assay types featuring a sample size of less than 6 were excluded, as the available data volume was deemed insufficient to adequately support model training, validation, and testing. Next, we approximated imprecise assay values with precise equivalents (for instance, from “> 100” to “= 100”), and standardized the assay values to the appropriate scale (for example, from “IC<sub>50</sub> = 1 nM” to “pIC<sub>50</sub> =  $-\log 10^{-9} \text{ M} = 9$ ”). Lastly, when confronted with the presence of the same assay type characterized by distinct units that couldn’t be interconverted, we separated them into distinct datasets (e.g., “Activity in percentage” and “Activity in nM”). Moreover, instances of duplicate assay values with identical units were averaged to ensure data consistency.

**Bioactivity Datasets Construction.** Each distinct assay category is treated as an individual dataset for each specific target, as different assay types hold varying biological implications. Within each dataset, data is partitioned into a 5-to-1 ratio, with 5 parts serving as training data and the remaining part as test data. This division enables a 5-fold cross-validation procedure conducted on the training data. During this progression, molecules are divided based on their Bemis-Murcko scaffolds and grouped with identical scaffolds within the same dataset whenever possible. This approach is employed to amplify the discrepancy among the training, validation, and testing data, thus facilitating the assessment of the models’ generalization capabilities.

## B Comparing AutoDock Vina and QVina-W on the Learnability of their Generated Data

Taking ChEMBL202 as an example, we docked 957 ligands with known bioactivity onto 7 pockets, resulting in 6,446 molecules. They were divided into training (5,312) and testing

---

<sup>1</sup><https://www.ebi.ac.uk/chembl/>

(1,134) sets using a 5-fold cross-validation strategy. The geometric interaction graph neural network (GIGN) [2] was the backbone network, optimizing Mean Squared Error (MSE). Despite noise due to multiple pockets, we assessed docking methods generating training data. Graph Convolutional Network (GCN) [3] and Universal 3D Molecular Representation Learning Framework (Uni-Mol) [?] were used as controls. Metrics included Pearson correlation ( $r$ ), Root Mean Square Error (RMSE), and Kendall Tau-b Coefficient ( $\tau_B$ ) (see Appendix Section C). The performance of models trained on docking results using AutoDock-Vina and QVina-W was compared. Results in Table S2 highlighted their superior performance over ligand-based methods, particularly Vina-Docked with lower RMSE but higher  $r$  and  $\tau_B$ . Therefore, AutoDock Vina was selected for protein-ligand complex generation in this study.

Table S1: An example of pocket selection (ChEMBL202).

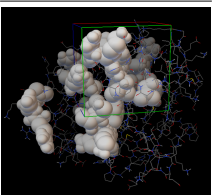
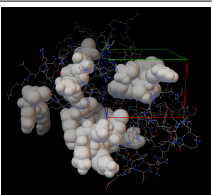
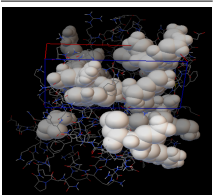
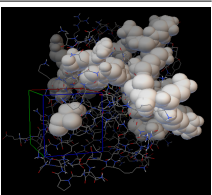
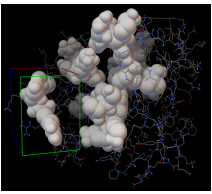
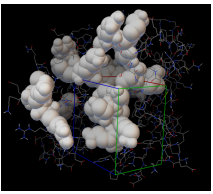
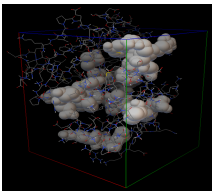
Pocket IDs	Pocket 1	Pocket 2	Pocket 3	Pocket 4
Binding Sites	10, 16-22	31-36	55-57, 65	71
				
Grid Boxes				
Pocket IDs	Pocket 5	Pocket 6	Pocket 7	
Binding Sites	77-79	117-124	ALL	
				
Grid Boxes				

Table S2: Comparison of bioactivity (pIC50 towards ChEMBL202) prediction performance on different data sources. ( $\uparrow$  or  $\downarrow$  denotes that larger or smaller values are better)

Data Sources	Data Types	Methods	Evaluation Metrics		
			$r$ ( $\uparrow$ )	RMSE ( $\downarrow$ )	$\tau_B$ ( $\uparrow$ )
Ligand Structure	2D Graph	GCN	0.2596	1.6651	0.1977
	3D Graph	Uni-Mol	0.3193	1.8373	0.2480
Protein-Ligand QVina-W-Docked Complexes	3D Graph	GIGN	0.4552	<b>1.4937</b>	0.3201
Protein-Ligand Vina-Docked Complexes			<b>0.5092</b>	1.5803	<b>0.3786</b>

Table S3: Benchmark Datasets of Bioactivity Prediction on Protein-Ligand Complexes.

IDs	Targets	Protein Names	Assay Types	PDB IDs <sup>1</sup>	# Binding Sites	# Molecules	# Pockets <sup>2</sup>	# Poses /Pocket
I1	CHEMBL202	Dihydrofolate reductase	IC50	1boz	8	957	7	1
I2	CHEMBL3976	Dipeptidyl peptidase 2	IC50	4ebb	3	1694	2	4
I3	CHEMBL333	72 kDa type IV collagenase	IC50	1ck7	24	3686	6	2
I4	CHEMBL2971	JAK2_HUMAN	IC50	3ugc	3	6207	3	4
I5	CHEMBL279	Vascular endothelial growth factor receptor 2	IC50	1ywn	3	9573	3	4
E1	CHEMBL3820	Hexokinase-4	EC50	3f9m	11	997	6	3
E2	CHEMBL4422	Free fatty acid receptor 1	EC50	5tzt	2	1693	3	3
E3	CHEMBL235	Peroxisome proliferator-activated receptor gamma	EC50	1zgy	4	3611	4	2

<sup>1</sup> The PDB IDs were selected with high-resolution and ligands that bound most binding sites. <sup>2</sup> The number of handpicked pockets based on binding site clusters.

Table S4: Graph Features of Protein-Ligand Complexes.

Atom Features	Types	Sizes	Descriptions
Atom type	One-hot	10	Heavy atom type [C, N, O, S, F, P, Cl, Br, I, others]
Degree	One-hot	7	Number of covalent bonds [0, 1, 2, 3, 4, 5, 6]
Implicit valence	One-hot	7	Implicit valence of the atom [0, 1, 2, 3, 4, 5, 6]
Hybridization	One-hot	5	[sp, sp2, sp3, sp3d, sp3d2]
Aromatic	Binary	1	Whether the atom is part of an aromatic system
Hydrogens	Integer	5	Number of connected hydrogens [0, 1, 2, 3, 4]
Intra-Molecular Bond Features	Types	Sizes	Descriptions
bond type	one-hot	4	[single, double, triple, aromatic]
conjugation	binary	1	whether the bond is conjugated [0/1]
ring	binary	1	whether the bond is in ring [0/1]
stereo	one-hot	4	[StereoNone, StereoAny, StereoZ, StereoE]
Intermolecular Interaction Features between atoms	Types	Sizes	Descriptions
Distance ( $d$ ) in Coulomb force	real-valued vector	64	$\left[ \exp\left(-\frac{1}{(\phi_1-\phi_0)^2}\ d^{-2}-\phi_0\ ^2\right), \dots, \exp\left(-\frac{1}{(\phi_{64}-\phi_{63})^2}\ d^{-2}-\phi_{63}\ ^2\right) \right], \phi_k = (6-5k/63)^{-2}$
Distance ( $d$ ) in London dispersion forces	real-valued vector	64	$\left[ \exp\left(-\frac{1}{(\phi_1-\phi_0)^2}\ d^{-6}-\phi_0\ ^2\right), \dots, \exp\left(-\frac{1}{(\phi_{64}-\phi_{63})^2}\ d^{-6}-\phi_{63}\ ^2\right) \right], \phi_k = (6-5k/63)^{-6}$

## C Experimental Settings

**Evaluation Metrics.** To comprehensively evaluate the model performance, this study employs two commonly used metrics for bioactivity prediction: the Pearson correlation coefficient ( $r$ ) and Root Mean Square Error (RMSE), along with a ranking metric, the Kendall Tau-b Coefficient ( $\tau_B$ ) [4].

The  $r$  metric assesses the linear correlation between two variables and is defined as fol-

lows:

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (1)$$

Here,  $y_i$  and  $\hat{y}_i$  represent the true and predicted bioactivity values, respectively, and  $\bar{y}$  and  $\bar{\hat{y}}$  are the means of the true and predicted bioactivity values, respectively.  $n$  stands for the total number of test samples. A larger  $r$  value indicates higher overall accuracy in the model predictions.

RMSE, a widely adopted metric for assessing regression model performance, is defined as the square root of the average of squared differences between predicted and actual values:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (2)$$

Here,  $y_i$  and  $\hat{y}_i$  represent the true and predicted bioactivity values, respectively, and  $n$  denotes the total number of test samples. Lower RMSE values indicate better performance of the regression model.

$\tau_B$  is a non-parametric statistic utilized to measure the agreement between two rankings, commonly used in fields like statistics and computational biology. It is calculated by counting concordant and discordant pairs in the rankings:

$$\tau_B = \frac{C - D}{C + D}, \quad (3)$$

Here,  $C$  refers to the number of concordant pairs, and  $D$  represents the number of discordant pairs. A concordant pair implies items ranked similarly in both rankings, while a discordant pair refers to differing rankings.  $\tau_B$  values range from -1 to 1, where 1 signifies complete agreement, 0 signifies no agreement, and -1 signifies complete disagreement between the rankings.

**Comparison Methods.** The prediction of ligand bioactivity based on deep learning primarily employs graph neural networks (GNNs), as demonstrated in the study by Bahi et al. [5]. This paper consequently compares two major categories of methods: 2D and 3D graph neural networks (GNNs). The 2D GNNs include graph convolutional networks (GCNs) [3], graph attention networks (GATs) [6], graph isomorphism networks (GINs) [7] pre-trained using supervised learning and context prediction [8], message passing neural networks (MPNNs) [9], Weave [10], neural fingerprint (Neural FP) [11], and Attentive FP [12]. The implementation and usage details of these 2D GNNs can be found on GitHub<sup>2</sup>.

As for the 3D GNNs, they include the Spatial Graph Convolutional Networks (SGCN) [13] and the Universal 3D Molecular Representation Learning Framework (Uni-Mol) [? ]. The source codes and usage instructions for these two 3D GNNs can be accessed on GitHub<sup>3,4</sup>. Notably, GINs and Uni-Mol offer pre-trained models on large-scale graph data, enabling us to fine-tune these models for enhanced performance.

<sup>2</sup>[https://github.com/awsml/dgl-lifesci/tree/master/examples/property\\_prediction/csv\\_data\\_configuration](https://github.com/awsml/dgl-lifesci/tree/master/examples/property_prediction/csv_data_configuration)

<sup>3</sup><https://github.com/gmum/geo-gcn>

<sup>4</sup><https://github.com/dptech-corp/Uni-Mol>

**Implementation Details.** The training data is divided into 5 subsets to perform 5-fold cross-validation, where the Root Mean Square Error (RMSE) is employed to assess the model generalization on the validation set. Training is stopped when the validation RMSE does not decrease over 100 epochs, and the model with the lowest validation RMSE will be evaluated for performance on the test set. The model selected with the lowest validation RMSE among the stopping points of 5 folds is considered the final result. The validation RMSE during the initial 40 epochs of training is not taken into account for the stopping criterion, as it is considered a warm-up phase for the model. The balance coefficient  $\gamma$  in Equation ?? is set to be 1 in all experiments.

During batch training, all poses of a single ligand docking with protein pockets are treated as a single entity to fully train the self-attention mechanism. Due to varying numbers of pockets and docking poses for different proteins, we set a uniform batch size of 128, which is divided by the number of pockets and poses to determine the number of unique ligands sampled per batch. For optimization, we employ the Adam optimizer with an initial learning rate of  $10^{-4}$  and a weight decay of  $10^{-4.3}$ . The learning rate is decayed by 95% every 10 training epochs. The graph convolutional module has 3 layers (i.e.,  $T = 3$ ). The hidden feature dimension of the model is consistently set to 256. The multi-head self-attention layer employs 8 heads, with a dropout rate of 0. The Multi-Layer Perceptron (MLP) consists of three linear layers with LeakyReLU activation and BatchNorm1d.

Data containing both bioactivity records and native protein-ligand complexes are highly limited. Consequently, apart from the ablation study, these rare data instances are exclusively reserved for validating the accuracy of the self-attention mechanism, without further utilization for model training.

## D Benchmark Test on RMSE and Kendall’s $\tau_B$

Table S5: Best cross-validated RMSE on the benchmark datasets.

Data	Data Type	Method	I1	I2	I3	I4	I5	E1	E2	E3	Average
Ligand-based	2D Graph	GCN	1.6651	1.9815	1.5389	1.2213	1.1334	0.9010	<b>0.8182</b>	1.2738	1.3167
		GAT	1.5935	1.7103	<b>1.5217</b>	1.1802	1.1715	0.9700	0.8278	1.2603	1.2794
		GIN (Pre-trained)	1.3344	1.9587	1.5386	1.2143	1.1844	0.9570	0.9457	1.3187	1.3065
		MPNN	1.4800	1.9113	1.6893	<b>1.1743</b>	1.1846	<b>0.8752</b>	0.8295	1.4227	1.3209
		Weave	1.6337	1.9423	1.6326	1.1762	1.1764	0.9990	0.8728	1.2861	1.3399
		Neural FP	1.8375	1.8250	1.5837	1.3622	1.2075	0.8863	1.1840	1.2962	1.3978
		Attentive FP	1.5786	1.8754	1.6390	1.1998	1.1485	0.9346	0.8544	1.2635	1.3117
	3D Graph	SGCN	1.6267	1.8401	1.8817	1.2668	1.1856	1.0230	0.9060	1.3370	1.3834
		UniMol (Pre-trained)	1.8373	1.6848	1.7296	1.1948	1.3928	1.8867	0.9067	1.2975	1.4913
Interaction-based	3D Graph	<b>DTIGN (Ours)</b>	<b>1.2823</b>	<b>1.6719</b>	1.5362	<b>1.1666</b>	<b>1.1251</b>	0.9546	0.8641	<b>1.2313</b>	<b>1.2290</b>

Table S6: Best cross-validated Kendall’s  $\tau_B$  on the benchmark datasets.

Data	Data Type	Method	I1	I2	I3	I4	I5	E1	E2	E3	Average
Ligand-based	2D Graph	GCN	0.1977	0.0488	<b>0.3118</b>	0.2908	0.2039	-0.1224	0.2431	0.1663	0.1675
		GAT	0.2316	0.3690	0.2818	0.1505	0.1683	-0.0596	0.3198	0.1849	0.2058
		GIN (Pre-trained)	0.2820	-0.1099	0.2415	0.2048	0.1375	-0.1099	0.1548	0.1621	0.1204
		MPNN	0.2393	0.2283	0.2182	0.1180	0.0476	0.0688	0.2152	0.0683	0.1505
		Weave	0.0411	-0.2845	0.2407	0.1565	0.1241	-0.1022	0.2327	0.2004	0.0761
		Neural FP	-0.1367	-0.0346	0.2853	0.2255	0.1178	-0.0638	0.2519	0.0282	0.0842
		Attentive FP	0.1600	0.0892	0.2570	<b>0.2976</b>	0.1927	-0.0340	0.3068	0.1713	0.1801
	3D Graph	SGCN	0.0841	-0.3600	0.0689	-0.1456	0.1769	-0.1456	0.1769	0.0566	-0.0110
		UniMol (Pre-trained)	0.2480	0.0628	0.1361	0.0152	0.1362	0.0152	0.2661	0.0983	0.1222
Interaction-based	3D Graph	<b>DTIGN (Ours)</b>	<b>0.4165</b>	<b>0.3926</b>	0.2899	0.2725	<b>0.2261</b>	<b>0.1095</b>	<b>0.3296</b>	<b>0.2714</b>	<b>0.2885</b>

## E Native and Docked structure on the pocket of attention

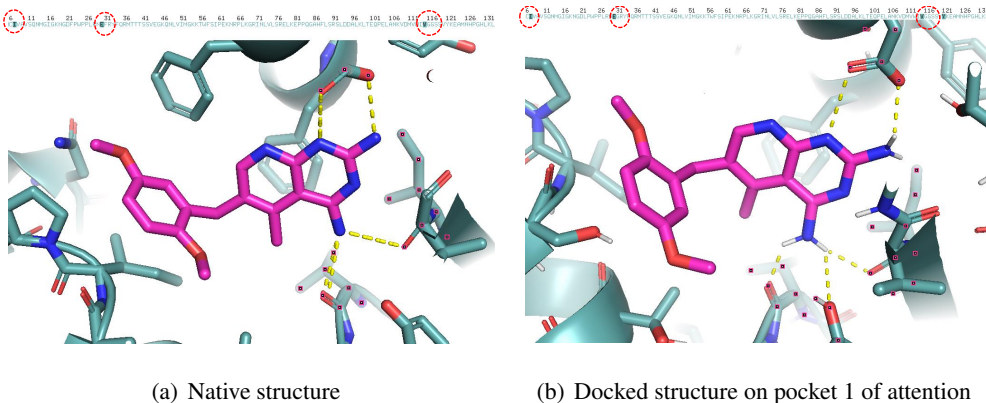


Figure S1: Binding poses for CHEMBL202 target protein and test ligand CHEMBL7492: (a) native pose and (b) AutoDock Vina-docked structure in the pocket 1 of model attention. Polar interactions between ligand and protein are marked by yellow dashed lines, with matching protein sequence residues above. Overlapping binding residues in native and docking structures encircled in red dashed lines.



## F Correlation between attention values and docking accuracies

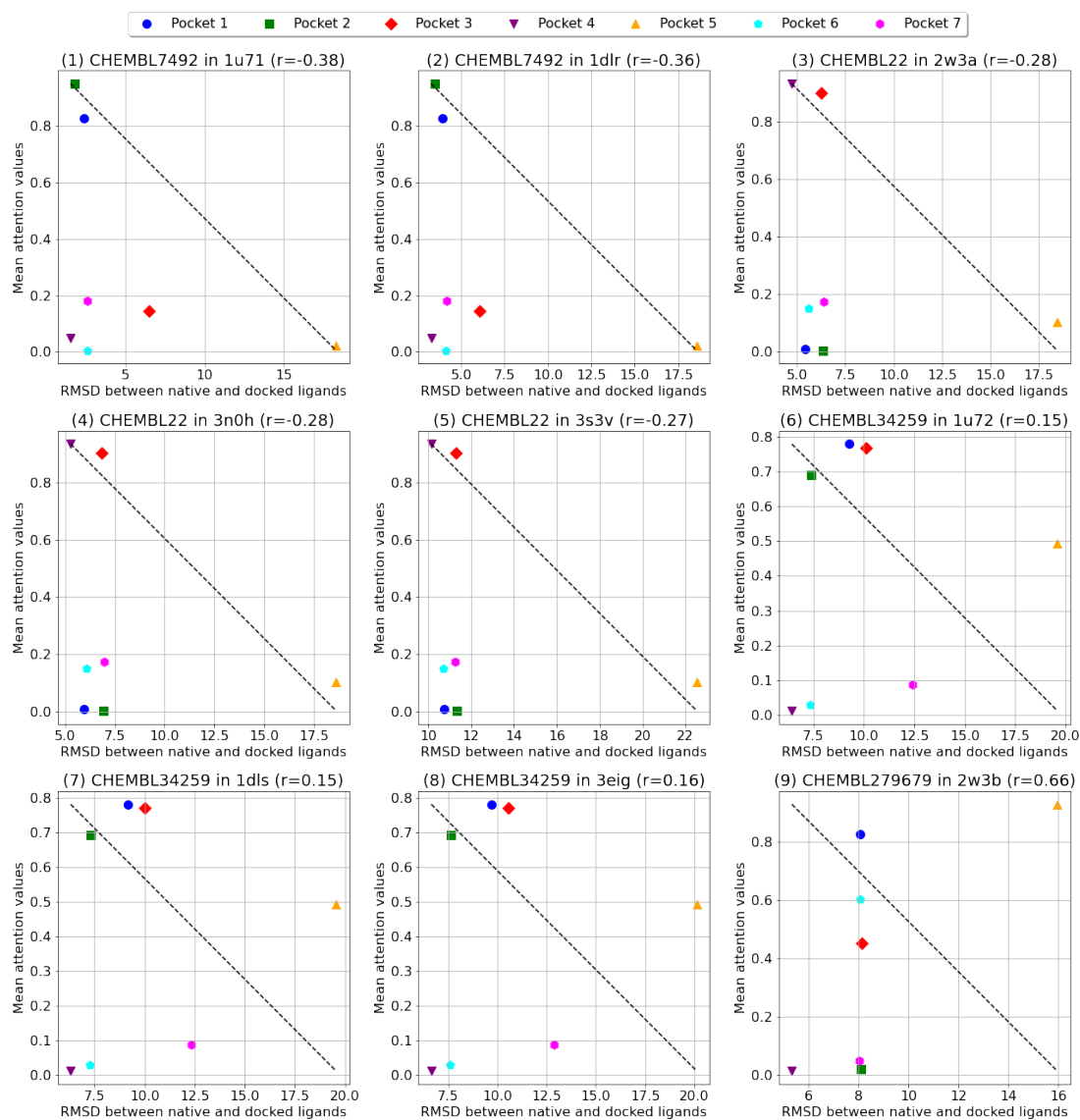


Figure S2: The docking accuracy (RMSD) in the pockets of model attention. RMSD was calculated between docked and native poses of test ligands and the protein (CHEMBL202) onto 7 pockets. The correlation between RMSD and average attention value during model training was evaluated using the Pearson correlation coefficient ( $r$ ). Subfigure titles contain “CHEMBL7492” for the test ligand ID and “1u72” for the PDB ID of its native structure with the protein.

## References

- [1] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.
- [2] Ziduo Yang, Weihe Zhong, Qiuji Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The Journal of Physical Chemistry Letters*, 14(8):2020–2033, 2023.
- [3] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.
- [4] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.
- [5] Meriem Bahi and Mohamed Batouche. Deep learning for ligand-based virtual screening in drug discovery. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE, 2018.
- [6] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- [7] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.
- [8] W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.
- [9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.
- [10] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.
- [11] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.
- [12] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, and Hualiang Jiang. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

- [13] Tomasz Danel, Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, and Łukasz Maziarka. Spatial graph convolutional networks. In *International Conference on Neural Information Processing*, pages 668–675. Springer, 2020.