Supporting Information of
Advancing Bioactivity Modeling through Molecular Docking and Self-Attention

Yueming Yin[1], Yuguang Mu[2], Hoi Yeung Li[2], and Adams Wai-Kin Kong [*1]

[1]School of Computer Science and Engineering, Nanyang Technological University 639798, Singapore.
[2]School of Biological Sciences, Nanyang Technological University 637551, Singapore.

January 30, 2024

---
[*]Corresponding author: adamskong@ntu.edu.sg

Contents

A Comparing AutoDock Vina and QVina-W on the Prediction Accuracy 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 (1,134) sets using a 5-fold cross-validation strategy. The geometric interaction graph neural network (GIGN) [1] 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) [2] and Universal 3D Molecular Representation Learning Framework (Uni-Mol) [3] were used as controls. Metrics included Pearson correlation (r), Root Mean Square Error (RMSE), and Kendall Tau-b Coefficient (τ_B) (see Appendix Section B). 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 τ_B. Therefore, AutoDock Vina was selected for protein-ligand complex generation in this study.
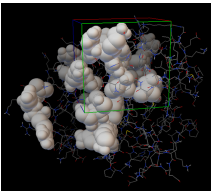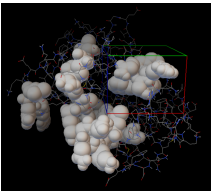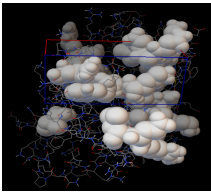
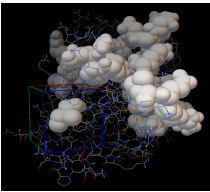Table S1: An example of pocket selection (CHEMBL202).
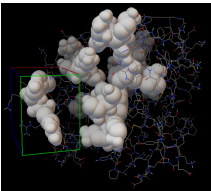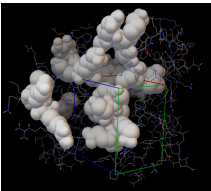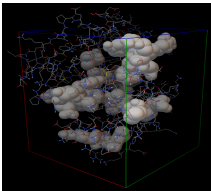
Pocket IDs	Pocket 1	Pocket 2	Pocket 3	Pocket 4
Binding Sites	10, 16-22	31-36	55-57, 65	71
Grid Boxes				

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. (↑ or ↓ denotes that larger or smaller values are better)

Data Sources	Data Types	Methods	Evaluation Metrics		
			r (↑)	RMSE (↓)	τ_B (↑)
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	**1.4937**	0.3201
Protein-Ligand Vina-Docked Complexes			**0.5092**	1.5803	**0.3786**

B 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 (τ_B) [4].

The r metric assesses the linear correlation between two variables and is defined as follows:

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

Here, y_i and $\hat{y}_i$ represent the true and predicted bioactivity values, respectively, and $\bar{y}$ and $\bar{\hat{y}}_i$ 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:

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

Here, yi 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.

τ_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:

$$\tau_B = \frac{C - D}{C + D}, \tag{3}$$

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. τ_B values range from -1 to 1, where 1 signifies complete agreement, 0 signifies no agreement, and -1 signifies complete disagreement between the rankings.

As for the 3D GNNs, they include the Spatial Graph Convolutional Networks (SGCN) [5] and the Universal 3D Molecular Representation Learning Framework (Uni-Mol) [3]. The source codes and usage instructions for these two 3D GNNs can be accessed on GitHub[1,2]. Notably, GINs and Uni-Mol offer pre-trained models on large-scale graph data, enabling us to fine-tune these models for enhanced performance.

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

---

[1] https://github.com/gmum/geo-gcn

[2] https://github.com/dptech-corp/Uni-Mol

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 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.

C   Benchmark Test on RMSE and Kendall's τ_B

Table S3: 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	**0.8182**	1.2738	1.3167
		GAT	1.5935	1.7103	**1.5217**	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	**1.1743**	1.1846	**0.8752**	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	**DTIGN (Ours)**	**1.2823**	**1.6719**	1.5362	**1.1666**	**1.1251**	0.9546	0.8641	**1.2313**	**1.2290**

Table S4: Best cross-validated Kendall's τ_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	**0.3118**	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	**0.2976**	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	**DTIGN (Ours)**	**0.4165**	**0.3926**	0.2899	0.2725	**0.2261**	**0.1095**	**0.3296**	**0.2714**	**0.2885**

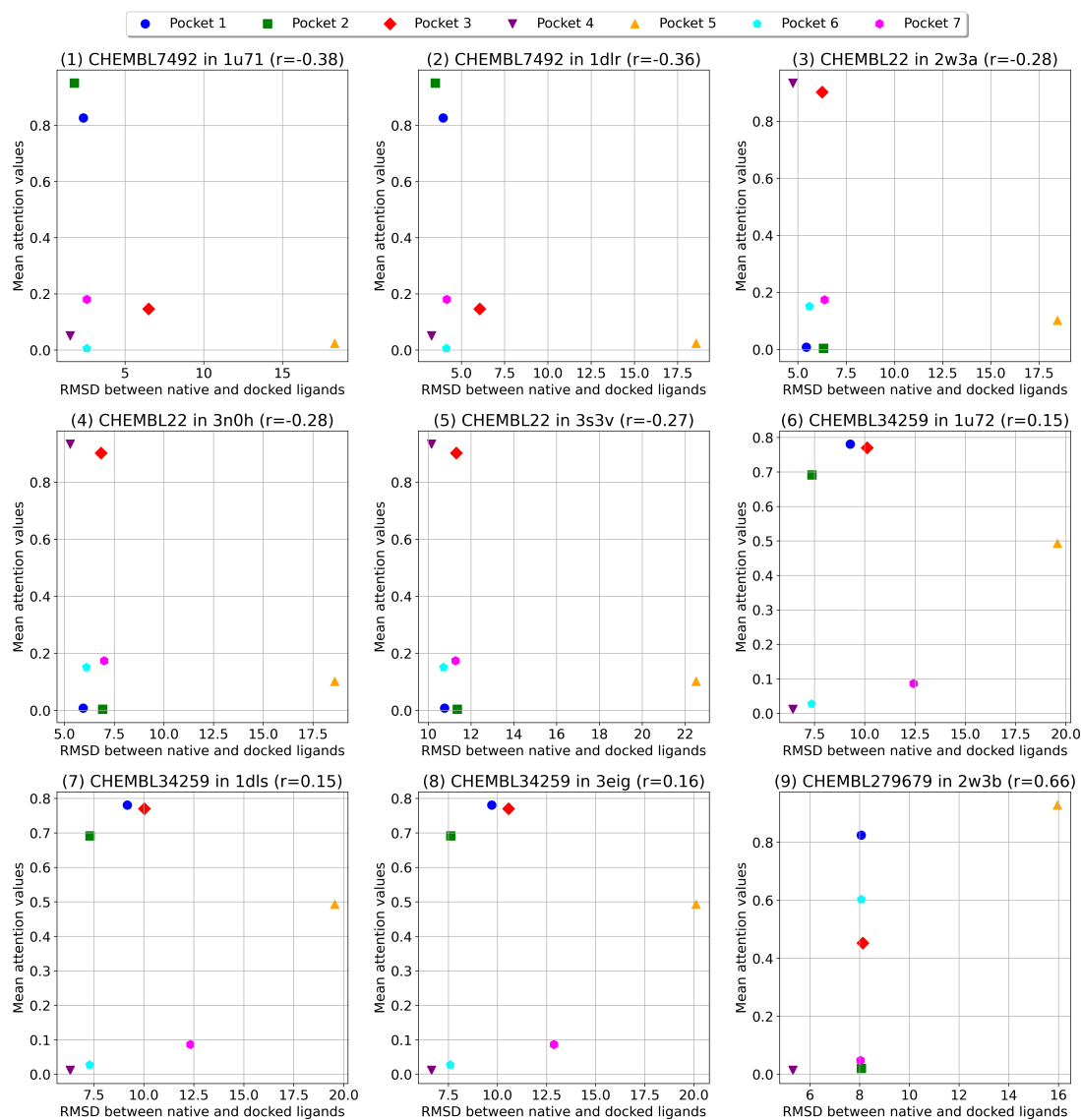# D   Correlation between attention values and docking accuracies

Figure S1: 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] Ziduo Yang, Weihe Zhong, Qiujie 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.

[2] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

[3] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: a universal 3d molecular representation learning framework. 2023.

[4] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[5] 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.