Supporting Information

A Deep Scoring Neural Network replacing the scoring function components to improve the performance of structure-based molecular docking

Lijuan Yang[1,2,3,4], Guanghui Yang[1,4], Xiaolong Chen[1,4], Qiong Yang[1,4], Xiaojun Yao[5], Zhitong Bing[1,4], Yuzhen Niu[6*], Liang Huang[2] and Lei Yang[1,4*]

1. Institute of modern physics, Chinese Academy of Science, Lanzhou 730000, China
2. School of Physics and Technology, Lanzhou University, Lanzhou 730000, China
3. School of Physics, University of Chinese Academy of Science, Beijing 100049, China
4. Advanced Energy Science and Technology Guangdong Laboratory, Huizhou 516000, China
5. College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou 730000, China
6. Shandong Provincial Research Center for Bioinformatic Engineering and Technique, School of Life Sciences, Shandong University of Technology, Zibo 255049, China

[*] Corresponding to: Yuzhen Niu, Email address: niuyzh12@lzu.edu.cn

[*] Corresponding to: Lei Yang, Email address: lyang_imp@outlook.com

1. Evaluation method

1.1 Evaluation protocol

According to different input information, we built three VS models, Fingerprint model, Deep Scoring-MC and Deep Scoring-PLC. For each model, we carried out the evaluation process shown in Fig. S1 on DUD-E and AD respectively. We created our own clusters of proteins using the hierarchical clustering module of scipy and ensured that proteins with greater than 80% sequence identity were removed from the training set. The training and testing process is shown in Fig. S1. Among the 101 proteins, one protein was selected as the test set, the proteins similar to the selected protein were deleted from the other 100 proteins, and the remaining proteins were used as the training set to train the model. Using the above method, 101 proteins were tested one by one. Finally, the average of the test results on all proteins indicated the performance of the algorithm.

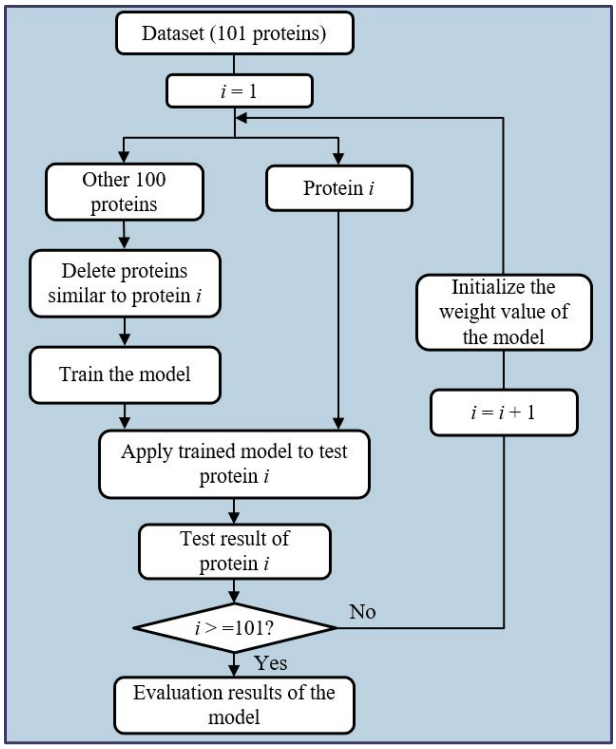

Fig. S1. The leave-one-out cross-validation process.

1.2 Evaluation indicators

The indicators widely used to evaluate the virtual screening are: enrichment factor (EF) and area under the ROC curve (AUC). A higher enrichment factor means better

prediction performance. In our experiments, we reported $EF_{0.1\%}$, $EF_{0.5\%}$, $EF_{1\%}$, $EF_{2\%}$, $EF_{5\%}$ and $EF_{10\%}$ for evaluating Deep Scoring performance. The higher the value of AUC, the better the classification performance of the model. But when the value of AUC is equal to 0.5, it means that the model has no predictive ability, which is equivalent to guessing. In our experiment, we added the area under the Percision Recall Curve (AUPR) as an evaluation indicator to ensure the accuracy of the model predicting positive samples. To evaluate the model based on the above three indicators, a very fair evaluation result can be obtained by us.

2. AUC and AUPR distribution on DUD-E and AD

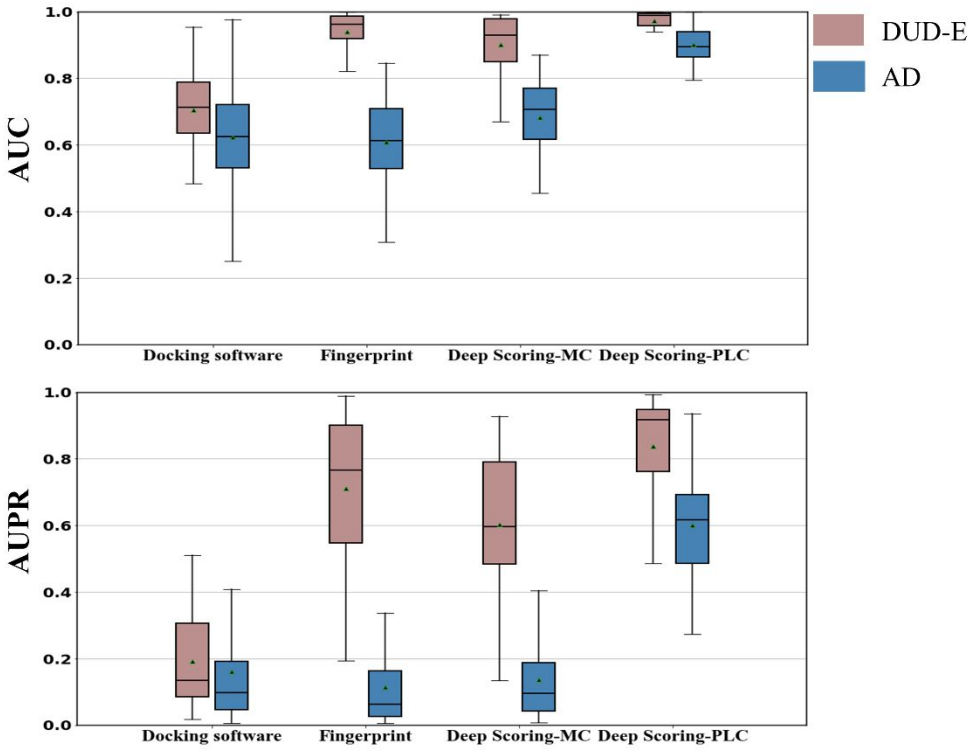

Fig. S2. Box plots of AUC and AUPR results of different input information in DUD-E and AD datasets. The horizontal lines indicate median, and the triangle represent the mean value.

3. PDBbind data preparation

The PDBbind dataset provides experimentally verified binding structures of protein-

ligand complexes, which are used to train our model to distinguish the most favorable binding posture for a given protein and ligand pair. In our experiments, we re-docked ligand from the PDBbind data set with settings –exhaustiveness = 50 –num_modes = 20 to generate compound conformations. The poses with a root mean square deviation (RMSD) less than 2Å was labeled as positive samples, and postures with an RMSD greater than 4Å as negative samples, where the RMSD is obtained according to the experimentally verified structure. The conformations whose RMSD is between 2 Å and 4 Å were omitted. After redocking, 16,031 proteins in PDBbind were used to evaluate the ranking ability of Deep Scoring, including 33,817 positive samples and 211,464 negative samples, with a total of 245,281 poses.

In this work, we assessed the pose prediction performance based on the intra-target ranking and the cross-target ranking. With intra-target ranking, the set of poses belonging to a single target is split into training and test sets, all test poses are ranked to generate a ROC curve. When training the intra-target ranking model, we shuffled all the data and selected 13,817 positive samples and 35244 negative samples as the test set, and the remaining samples as the training set. The cross-target ranking is to divide the data set into training set and test set according to protein, so that training set and test set do not share the same protein. When training the cross-target ranking model, 3031 proteins were randomly selected from 16061 proteins for testing, and the rest were used to train the model. Finally, used AUC and AUPR as indicators to evaluate Deep Scoring's ranking ability.

In this ranking method, all the conformations of a given compound-protein can be sorted to find the pose with the lowest RMSD. For the same compound, a scoring function can compare all conformations and give a reasonable RMSD ranking, even though the pose with low RMSD has a low score as long as the other poses with high RMSD have worse scores.

4. Independent test result

Table S1. The AUC value of different models when testing on MUV.

Proteins	ADV	Fingerprint	Deep Scoring-MC	Deep Scoring-PLC
466	0.524	0.528	0.566	0.607
548	0.394	0.721	0.609	0.758
600	0.679	0.524	0.627	0.602
689	0.52	0.520	0.548	0.657
692	0.502	0.502	0.638	0.678
832	0.568	0.706	0.625	0.765
846	0.492	0.625	0.618	0.844
852	0.519	0.704	0.546	0.792
859	0.644	0.620	0.635	0.607
Aver	0.538	0.606	0.601	0.701

Table S2. The AUC value of different models when testing on CHEMBL.

Proteins	ADV	Fingerprint	Deep Scoring-MC	Deep Scoring-PLC
4daj	0.480	0.743	0.681	0.788
3ks9	0.490	0.806	0.750	0.852
1ms6	0.558	0.767	0.764	0.721
2xul	0.403	0.677	0.596	0.697

4s0v	0.594	0.638	0.747	0.717
1mkd	0.678	0.85	0.766	0.84
4xuf	0.607	0.754	0.778	0.785
4kik	0.529	0.697	0.746	0.857
5ek0	0.728	0.758	0.726	0.806
1hvy	0.771	0.741	0.700	0.853
1mq4	0.644	0.821	0.793	0.867
2qyk	0.779	0.763	0.795	0.835
3v2y	0.461	0.703	0.714	0.824
Aver	0.605	0.748	0.735	0.803

5. Docking parameters

5.1 AutoDockVina

The docking process is based on the default settings defined in AutoDockVina (ADV). We modify the "exhaustiveness" parameter in "conf.txt" to 16, and set other parameters to default values. Although ADV can output multiple docking postures, in our experiment, we only considered the first output result, which is consistent with the best posture by ADV. Through the above process, we used ADV to obtain the three-dimensional structure information of the PLC.

5.2 Schrodinger-Glide

The protein is processed by the Protein Preparation Wizard, in which all parameters are set to default values. Use the binding site and receptor grid box parameters collected for each protein in the DUD-E dataset for Receptor Grid Generation. The preprocess method for small molecules is to add hydrogen atoms. Finally, the saved structures from

the previous step are docked and scored by the Glide standard precision (SP) scoring mode.

6. Deep Scoring sensitivity to different hyperparameter values

Table S3. Deep Scoring sensitivity to different hyperparameter values.

Hyperparameter	**Value**	**AUC**	**AUPR**	$\boldsymbol{EF_{5\%}}$
Learning rate	0.0001	0.879	0.493	9.264
	0.0005	0.901	0.601	10.947
	0.00075	0.893	0.598	10.795
Embedding size	24	0.900	0.594	10.675
	32	0.901	0.601	10.947
	64	0.900	0.604	10.737
k_c, k_p	6, 0	0.682	0.137	3.322
	5, 2	0.897	0.599	10.377
	6, 2	0.901	0.601	10.947
	6, 4	0.892	0.593	10.604