Support Information

OnionNet: a multiple-layer inter-molecular contact based convolutional neural network for protein-ligand binding affinity prediction

*Liangzhen Zheng, Jingrong Fan and Yuguang Mu**

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive,

Singapore, 637551

*Corresponding author: ygmu@ntu.edu.sg

1. Model training

The deep neural network (DNN) model was trained on NVIDIA GTX 1070 GPUs with Tensorflow.keras module [1, 2]. An early-stopping strategy was adopted to avoid over-fitting with a patience as 20 epochs and a min_delta = 0.01 root mean square error (*RMSE*). Therefore, a best model is harnessed if the learning enters an overfitting stage. The best model containing weights was saved into an HDF5 file. After around 75 epochs, the loss is about to stay unchanged, and according the early stopping strategy, the model was terminated at epoch=171, and we adopted the model at epoch=151 and used it as the best model. The Pearson correlation coefficient (*PCC*, or *R*) is also monitored.

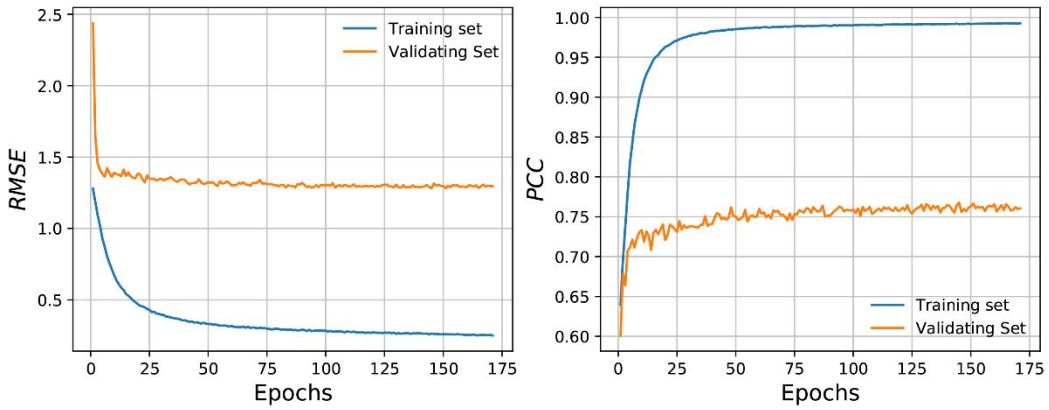

Figure S1 Performance of our model as a function of epoch number.

2. Customized loss function

During the training of our DNN models, instead of using the default mean squared error (*MSE*) as the loss function, we designed a customized loss function to optimize:

$$Loss = \alpha(1 - PCC) + (1 - \alpha)RMSE,$$

where *PCC* and *RMSE* are the correlation coefficient and root mean squared error respectively, while α is a positive and less than 1 weighting factor. In this study, α=0.7 is used. The customized loss function was designed to balance the correlation and error term.

3. Model training with missing features

To test the stability and robustness of the OnionNet model, we artificially created different datasets with missing features. For the multiple-layer inter-molecular features, we generated a *M*×*N* matrix, containing samples (by rows) horizontally and features (by columns) vertically. In detail, for each of the 3840 features, if we need to remove one feature, we replaced the original values (in the specific column) to zeroes to maintain the dimensions of the dataset. This way the artificially generated dataset could be reshaped and trained with the same DNN architecture.

Firstly, we examined the stability of the model by removing the features (n-1) * 64 to n*64 in shell n ($n \in [1, 64]$). For example, if n=1, we then remove features 1 to 64, and if n=6, we thus remove features 320 to 384. For all the 60 shells, we generated 60 datasets with values being replaced by zero in specific features. Then for each one of the datasets, a DNN model with the same structure as described in the method part of the main text was trained with the optimal batch size (64) and dropout rate (0.0) based on the same early stopping strategy. The $\Delta Loss$ of the last 20 steps was calculated based on the customized loss difference between the model and the best model without missing features.

$$\Delta Loss = Loss_{missing} - Loss_{best}$$

where the $Loss_{missing}$ and $Loss_{best}$=0.549 are the loss of the model with missing features and the best model we trained without missing features.

Similar, for models with missing specific element-type combination features, the values in the relevant features thus were replaced by zeroes. Since there are 64 different combinations, we generated 64 datasets and for each of them a model was trained using the same setting. The model performance was evaluated in the same strategy.

In crystal structures, the coordinates of hydrogen atoms are missing. However, in PDBbind database, the positions of the hydrogen atoms were modelled. Therefore, we generated another dataset without any features based on element-type combinations containing hydrogen element. The dataset was fed to a DNN model (called OnionNet_HFree hereafter) and the performance was evaluated. It seems like that without including hydrogen elements, the model shows equally well for the validating dataset (Figure S2). Although the inclusion of hydrogen atom coordinates may introduce bias to the model, from the training results, however the model is robust enough to tolerate the bias.

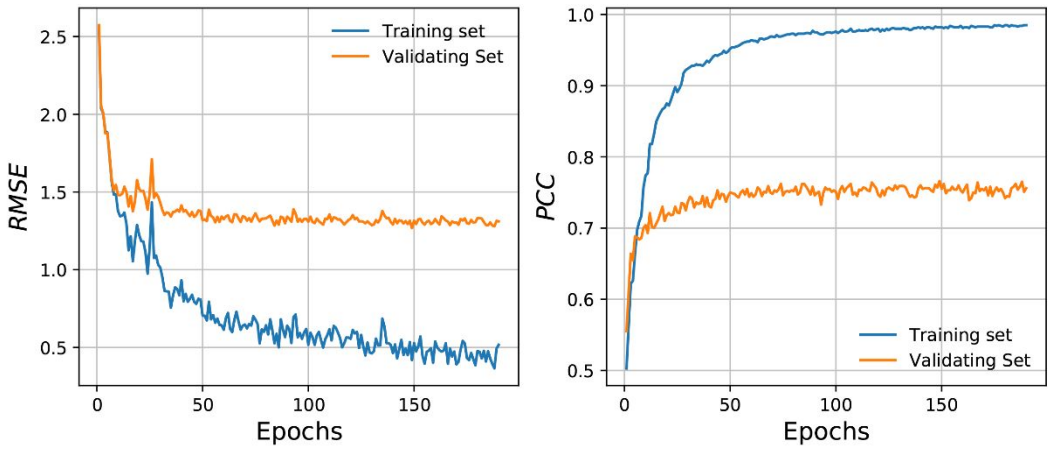

Figure S2. The decay curves of *RMSE* and *PCC* for the model training with a hydrogen-free dataset.

4. Predicting pK_a of the "native-like" docking poses

The ability to predict the pKa of a docking pose would enable the model a more powerful and practical tool as an alternative scoring function to enable large-scale virtual screening. Here, using AutoDock Vina

[3], we docked a ligand into its native target binding pocket. And we selected the "good" protein-ligand docked complexes for pKa prediction. If the root-mean-square-deviation (*RMSD*) between the docking pose and the native ligand conformation (in a crystal or NMR structure) is less than a cutoff = 3 Å, the complex of the target (protein) and the docking pose of the ligand thus is a "good" complex. For one protein-ligand complex, Vina was required to output 20 docking poses, out of which the best docking pose is selected based on minimum *RMSD*, and if the minimum *RMSD* is less than the cutoff, then the "good" complex is passed to generate multiple-layer element-type based inter-molecular features. This way a dataset was generated based on the docking poses and *RMSD*.

We randomly selected 318 "good" complexes from the training set after docking. The full list was provided in the supplementary table. For the 181 complexes in the testing set, only 149 "good" complexes were generated by docking with Vina, and then were submitted for inter-molecular featurization. An exhaustiveness= 32 and the box sizes=40 Å were adopted for docking, using the centroid of the native conformation of the ligand as pocket center.

The OnionNet_HFree model thus was used to predict the pK_a for the "good" complexes generated from docking simulations.

5. HIV protease docking and rescoring

The receptor conformation of HIV protease (PDB ID: 1A30) [4] was downloaded from RCSB Protein Data Bank. The crystal water molecules and its native ligand were removed. The SMILES codes of the active and decoys molecules were downloaded from the Deepchem project (https://deepchem.io/), which provides well defined and cleaned datasets for machine learning model constructions in computer-aided drug discovery (CADD). The full list of the SMILES codes of the molecules were downloaded from Deepchem github repository (https://github.com/deepchem/deepchem/tree/master/examples/hiv). With Rdkit package [5], we generated the 3D conformations of the molecules, added polar hydrogen atoms and optimized the conformations with MMFF force-field. The file format conversion was completed with Open Babel package [6]. The molecules were docked into HIV protease ligand binding pocket using AutoDock Vina [3], with an exhaustiveness as 32 and box sizes as 40, using the centroid of the original ligand (in PDB ID: 1A30) in the pocket as the pocket center. The best docking pose per molecule according to the lowest docking score of Vina thus was selected and was adopted for inter-molecular contacts feature generation and pK_a prediction with the OnionNet_HFree model.

6. Table S1: Training sets of the SFs in main text Table 2

Scoring functions	Training sets
OnionNet	PDBBind v2016 removing the refine sets and core set
KDeep	PDBBind v2016 refine set removing the core set
RF-score-v3	PDBBind v2013 and/or 2016 refine set removing the core set
Pafnucy	PDBBind v2016 removing the refine sets and core set
AGL	Refine sets of CASF2007, CASF2013, CASF2016 after removing core sets
kNN-score	PDBBind v2013 removing the core set and low-quality complexes
X-score	PDBBind v2013 removing the core set and low-quality complexes
ChemScore	PDBBind v2013 removing the core set and low-quality complexes
ChemPLP	PDBBind v2013 removing the core set and low-quality complexes

AutoDock Vina	CASF2013 complexes
AutoDock	CASF2013 complexes

References

1. Chollet, F., *Keras*. 2015.
2. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
3. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
4. Louis, J.M., et al., *Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease.* Biochemistry, 1998. **37**(8): p. 2105-2110.
5. Landrum, G., *RDKit: Open-source cheminformatics*. 2006.
6. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* Journal of cheminformatics, 2011. **3**(1): p. 33.