Yuel: Compound-Protein Interaction Prediction with High Generalizability

Jian Wang[1] and Nikolay V. Dokholyan[1,2,3,4,*]

[1]Department of Pharmacology, Penn State College of Medicine, Hershey, PA 17033, USA.

[2]Department of Biochemistry & Molecular Biology, Penn State College of Medicine, Hershey, PA 17033, USA.

[3]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA.

[4]Department of Biomedical Engineering, Pennsylvania State University, University Park, PA 16802, USA.

[#] These authors contributed equally to this work.

* Correspondence: E-mail: dokh@psu.edu, Tel: (717) 531-5177.

Keywords: neural network, compound-protein interaction, attention

ABSTRACT

In recent years, numerous structure-free deep-learning-based neural networks have emerged aiming to predict compound-protein interactions for drug virtual screening. Although these methods show high prediction accuracy in their own tests, we find that they are not generalizable to predict interactions between unknown proteins and unknown small molecules, thus hindering the utilization of state-of-the-art deep learning techniques in the field of virtual screening. In our work, we develop a compound-protein interaction predictor, YueL, which can predict compound-protein interactions with high generalizability. Upon comprehensive tests on various data sets, we find that YueL has the ability to predict interactions between unknown compounds and unknown proteins. We anticipate our work can motivate broad application of deep learning techniques for drug virtual screening to supersede the traditional docking and cheminformatics methods.

INTRODUCTION

Drug discovery has strongly benefited from virtual drug screening[1] techniques based on either molecular docking of a library of compounds to a target protein, or cheminformatics-based approaches that utilize machine learning to derive knowledge from known binders to this target. Yet, molecular docking is usually time-consuming, while cheminformatics-based approaches suffer from low accuracy when no ligand for a given target is known. In recent years, numerous structure-free deep-learning-based neural networks (NNs)[2–5] have emerged aiming at predicting compound-protein interactions (CPI) for virtual screening. These NNs follow a common paradigm, which is to encode the compound and protein features first, and then concatenate them and subject to convolutional layers and fully-connected (FC) layers to predict CPI (Figure 2A). This *encoding-concatenation-FC* paradigm has been successful in many fields[6], but we find that for CPI these methods are not predictive. Conforming to this paradigm, NNs are inclined to memorize the protein sequence or the compound structure by adjusting the relative weights of protein features to compound features in FC layers, instead of identifying the binding patterns between the protein and the compound. This aspect significantly affects the generalizability of the NN, that is, the ability to predict the interaction between unknown compounds and unknown proteins, thus hindering the utilization of state-of-the-art deep learning techniques in the field of virtual drug screening. Here, we develop a new compound-protein interaction predictor *Yuel* (Figure 1A), which predicts CPI with high generalizability. In contrast to other NNs, FC layers in *Yuel* are only applied to individual atom and residue features (Figure 2B), and CPI is predicted by summing up the pairwise atom-residue interactions, each of which is evaluated by multiplying the feature of the corresponding compound atom and the feature of the corresponding protein residue (Figure 1A).

RESULTS

We train and test Yuel and two other typical structure-free neural networks, DeepDTA[2] and DeepConv-DTI[3], on two datasets (PDBbind[7] and Davis[8]). We divide each of the two datasets into a training set and a test set at a ratio of 8 to 2. We find the Spearman correlation between the experimental affinities and predicted affinities of Yuel, DeepDTA, and DeepConv-DTI on the two datasets are all above 0.6 (Figure 1B). This significant result for all NNs is not surprising as the training and test sets are sourced from the same dataset, thus many proteins and compounds are shared between the training set and the test set. To test the ability to predict the interactions between unknown proteins and unknown compounds, we train the models on Davis and test on PDBbind. Compounds and proteins are dissimilar in these two datasets (Figure 3), and Davis has

much smaller numbers of proteins and compounds than PDBbind (Methods). We find the Spearman correlation of *Yuel*, DeepDTA, and DeepConv-DTI are 0.489, 0.054 and -0.015, respectively, suggesting that DeepDTA, and DeepConv-DTI cannot predict the interaction between unknown proteins and unknown compounds, while *Yuel* has a significant predictability. We further trained the models on Davis and test on a third dataset Metz[9]. *Yuel* still outperforms DeepDTA and DeepConv-DTI in terms of the Spearman correlation (Figure 4).

To demonstrate that other NNs are inclined to memorize protein sequences or compound structures, we train the models on the original PDBbind training set, but then shuffle the residues of each protein sequence of the PDBbind test set for testing. The sequence shuffling does not significantly affect DeepDTA and DeepConv-DTI predictions of CPI (Spearman correlation is 0.7 and 0.6, respectively). On the contrary, *Yuel*'s prediction falls drastically to Spearman correlation of 0.2 (Figure 5), in line with expectations (Figure 5). Thus, DeepDTA and DeepConv-DTI are memorizing the compound structures to predict CPI in PDBbind, which is attributed to 1) the inherent defect of the *encoding-concatenation-FC* paradigm and 2) the extremely low completeness (Methods) of the PDBbind dataset.

Finally, Y*uel* can also identify compound atoms that interact with proteins (hotspot atoms) and residues that interact with compounds (hotspot residues). Since different compound atoms have different predicted interactions with different residues (Figure 1C), we adopt a strategy of adding up the interactions of an atom to all residues as the possibility that this atom is a hotspot atom. This strategy results in an average AUC of 0.58 (Figure 1D) for PDBbind. Similarly, the predicted interactions of most hotspot residues are ranked within 30 (Figure 1E). Particularly, for 2IHQ, Yuel can predicted 9 out of the 11 hotspot atoms (Figure 1F&1G) and 11 out of the 18 hotspot residues (Figure 1H).

METHODS

The Architecture of Yuel

Yuel requires two inputs to estimate interactions between a specific compound and the protein target: (i) compound's SMILES[10] code and (ii) protein's amino acid sequence. Given the protein sequence, *Yuel* first encodes it using BLOSUM62 matrix[11]. Each residue corresponds to a column in the BLOSUM62 matrix. The features of non-standard amino acids are zero-initialized. Then, the protein features are updated through 3 1-D convolution layers. The sequence is zero-padded to a fixed length of 2048. Given the SMILES of the compound, *Yuel* first employs rdkit[12] to

represent it by a graph (*N, V, E*), where *N* is the number of nodes, *V* is the feature vector of each atom, and *E* is the feature vector of each bond. The feature of each atom is the concatenation of the one-hot encoding of atom type, number of bonds, bond type, mass, and charge vectors. The feature of each bond is the bond order. The graph is then subject to two graph convolution layers to update the features. The protein features and the compound features are then subject to 5 fully-connected layers, separately. Following the fully-connected layers, the protein feature matrix and the compound feature matrix are then multiplied to obtain an attention matrix. Each row of the attention matrix corresponds to a compound atom, and each column corresponds to a protein residue. The maximum value of each column is then selected and added up to get the final predicted compound-protein interaction value. We implement this network using TensorFlow 2.3.1[13].

Fully-connected layer

The existence of fully-connected layers in other NN-based CPI predictors[2,3,5] makes them inclined to memorize the sequence of proteins and the 2D structure of small molecules, instead of identifying the binding patterns between them. All current CPI prediction NNs concatenate the features obtained from small molecules with the features obtained from protein sequences, and then flatten all the features and concatenate them to subject to fully connected layers (Figure 2A). By contrast, in *Yuel*, we do not concatenate the features of compounds with the features of proteins, but only fully connect the features of each residue and each atom individually, and then multiply the features together (Figure 2B). Since the feature of each atom in the compound is obtained through GNN, it actually contains the information of its neighboring atoms; similarly, since the feature of each residue in the protein is obtained through 3 CNN layers, it actually contains information of its sequence-neighboring residues. In this way, after the fully connected layer, if a large value is obtained by multiplying the feature of a compound's atom by the feature of a residue, it means that the group of atoms centered at this atom may interact with the fragment centered at this residue. Therefore, the essence of *Yuel* is to predict the interaction between all possible combinations of atomic groups in the compound and amino acid fragments in the protein. In stark contrast, other NN-based predictors concatenate the features of all compound atoms with the features of all protein residues and then flatten and subject them to fully connected layers. This strategy may also identify atom-residue interactions, but it is more likely to cause the neural network to memorize the 2D structure of the small molecule and the sequence of the protein by assigning unique weights to compound atoms or protein residues in the fully-connected layers, thus losing the ability to predict the interaction between unknown proteins and the unknown small

molecules. In the worst case, this type of neural networks may even predict the compound-protein interaction based solely on small molecules or only on proteins.

Datasets

Our benchmark datasets includes the PDBbind dataset[7], Davis dataset[8], and Metz dataset[9]. The PDBbind dataset is compiled from PDBbind database (version 2018, the general set), which contains a high-quality set of protein-ligand complexes with available structural data and corresponding binding affinities. Each complex was provided with an affinity value of certain measurement type. The Davis dataset is obtained from the Supplementary Information of Davis et al. work[8]. The Metz dataset is obtained from the Supplementary Information of Metz et al. work[9]. The PDBbind dataset contains 2509 small molecules and 10251 proteins, with a total of 13,311 interactions. On average, each small molecule corresponds to only 5.3 interactions and each protein corresponds to only 1.3 interactions. The Davis dataset contains 68 small molecules and 379 proteins, with a total of 9125 interactions. On average, each small molecule corresponds to 134.2 interactions, and each protein corresponds to 24.1 interactions. The Metz dataset contains 1471 small molecules and 170 proteins, with a total of 35307 interactions. On average, each small molecule corresponds to 24.0 interactions, and each protein corresponds to 207.7 interactions. We divide each dataset into a training set and a test set at a ratio of 8 to 2.

Dataset completeness

We propose a metric, the completeness of a data set $C_{DS} = N_{interaction}/(N_{protein} \cdot N_{compound})$, to evaluate whether a data set is suitable for training a neural network for compound-protein interaction prediction. $N_{protein}$ is the number of proteins; $N_{compound}$ is the number of compounds; $N_{interaction}$ is the number of interactions in the data set. If the completeness of the dataset set is 1, it means that the interaction between each small molecule and each protein in the dataset is included. The completeness of PDBbind, Davis, and Metz is 0.00052, 0.35, and 0.14, respectively. Training DeepDTA or DeepConv-DTI on a dataset with an extremely low completeness (e.g. PDBbind) will make the network inclined to predict CPI by memorizing the protein or the compound structure because the $N_{compound}/N_{interaction}$ ratio or the $N_{protein}/N_{interaction}$ is close to 1. Unlike other NN-based methods, the reliability of *Yuel* does not depend on the completeness of the training set.

DATA AVAILABILITY

Source codes and test data are in the Supplementary Data. They are also deposited in: *https://bitbucket.org/dokhlab/yuel*. We provide a website to use Yuel: *https://yuel.dokhlab.org* or *https://dokhlab.med.psu.edu/cpi*.

ACKNOWLEDGEMENTS

We acknowledge support from the National Institutes for Health (R35 GM134864), National Science Foundation (2040667) and the Passan Foundation.

DECLARATION OF INTERESTS

The authors declare no competing financial interest.

REFERENCES

1. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
2. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
3. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **15**, e1007129 (2019).
4. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329–3338 (2019).
5. Li, S. *et al.* MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst.* **10**, 308–322 (2020).
6. Webb, S. Deep learning for biology. *Nature* **554**, (2018).
7. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405 (2015).
8. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*

29, 1046–1051 (2011).

9. Metz, J. T. *et al.* Navigating the kinome. *Nat. Chem. Biol.* **7**, 200–202 (2011).

10. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

11. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22**, 1035–1036 (2004).

12. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. (2013).

13. Abadi, M. *et al.* Tensorflow: a system for large-scale machine learning. in *OSDI* vol. 16 265–283 (2016).

FIGURES

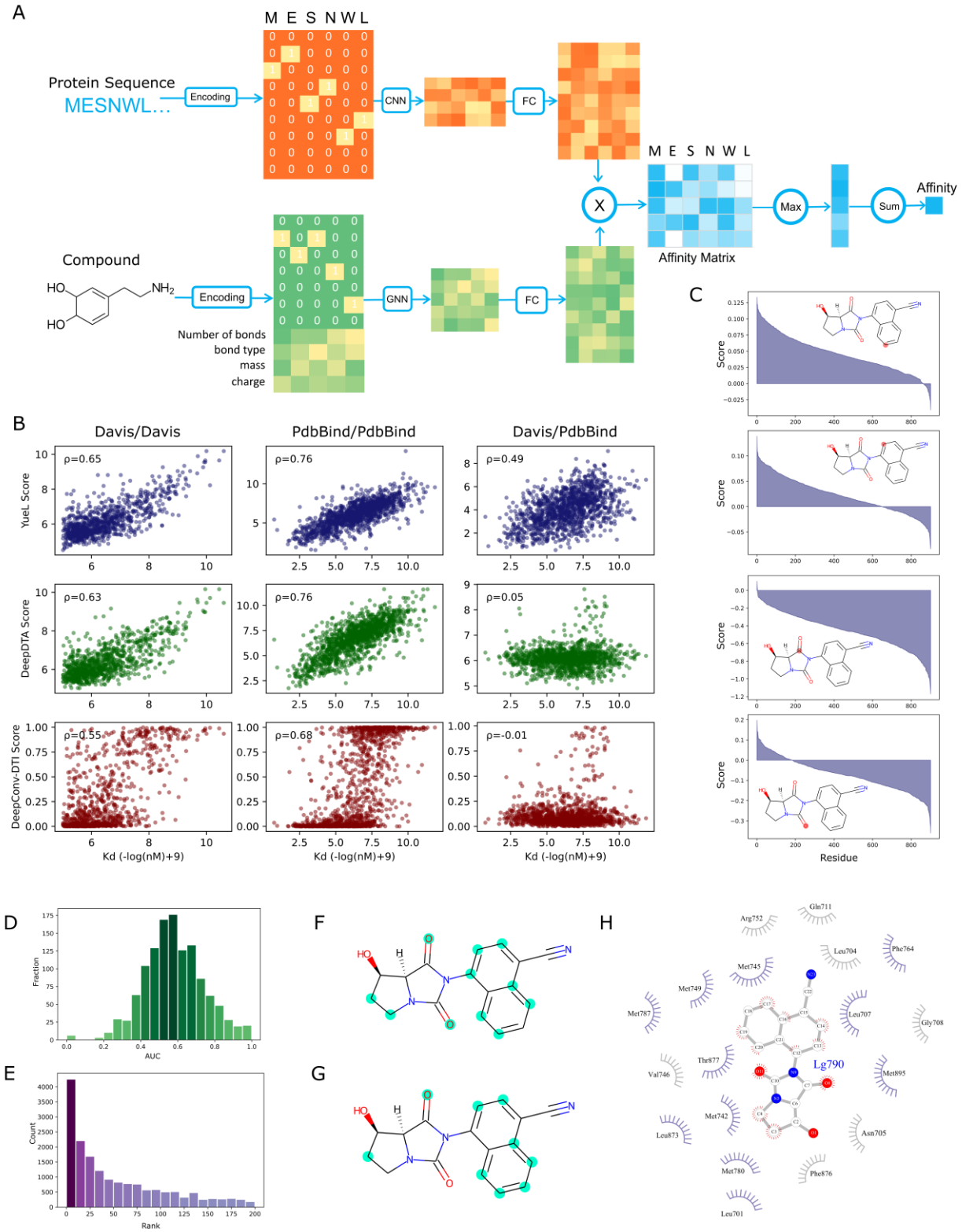

Figure 1. Test of the ability of Yuel to predict compound-protein interaction. (A) Workflow of Yuel.

(B) Test of the performance of Yuel, DeepDTA, and DeepConv-DTI on the Davis and the PDBbind data set. (C) The predicted protein-binding interactions of fours atoms in N-Aryl-Hydroxybicyclohydantoin (LG790), the ligand of the rat androgen receptor (PDB ID: 2IHQ). (D) The histogram of AUC of using Yuel to predict hotspot atoms in LG790. (E) The ranks of hotspot residues in the protein of 2IHQ. (F) The ground-truth hotspot atoms in LG790. (G) The predicted hotspot atoms in LG790. (H) The predicted hotspot residues in the protein of 2IHQ.

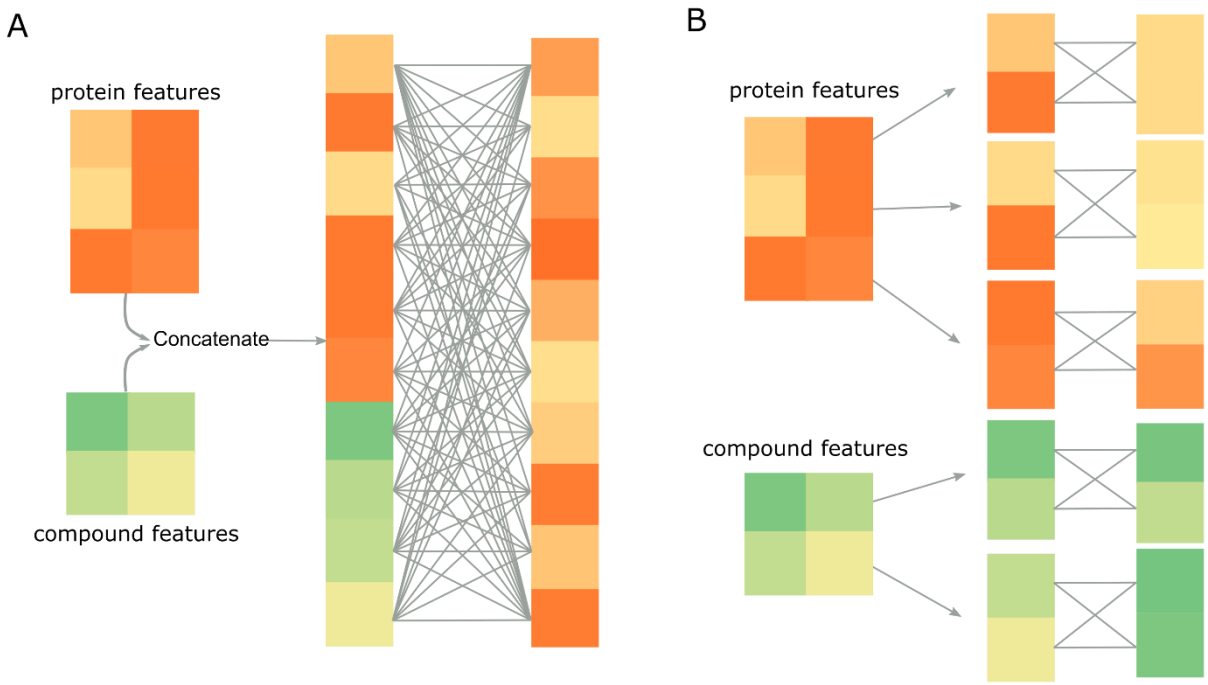

Figure 2. Comparison of the fully-connected layers in other networks (A) and Yuel (B). (A) The protein features and the compound features are first concatenated and flattened to a 1-D feature vector. The 1-D feature vector is then subject to full-connected layers. (B) The protein features and the compound features are first split to individual residue features and atom features. Each residue feature and each atom feature is subject to fully-connected layers, individually. Finally, the residue features and the atom features are multiplied to obtain an attention matrix. Each row of the attention matrix corresponds to a specific compound atom and each column of the attention matrix corresponds to a specific protein residue.

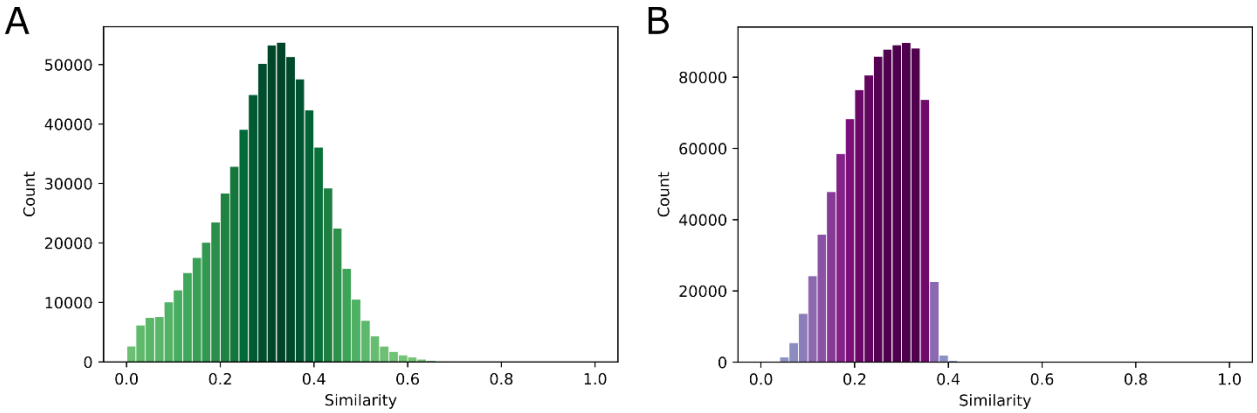

Figure 3. Compounds and proteins in the Davis dataset and the PDBbind dataset are dissimilar. (A) The histogram of the similarities of compounds between the David dataset and the PDBbind dataset. (B) The histogram of the similarities of proteins between the David dataset and the PDBbind dataset.

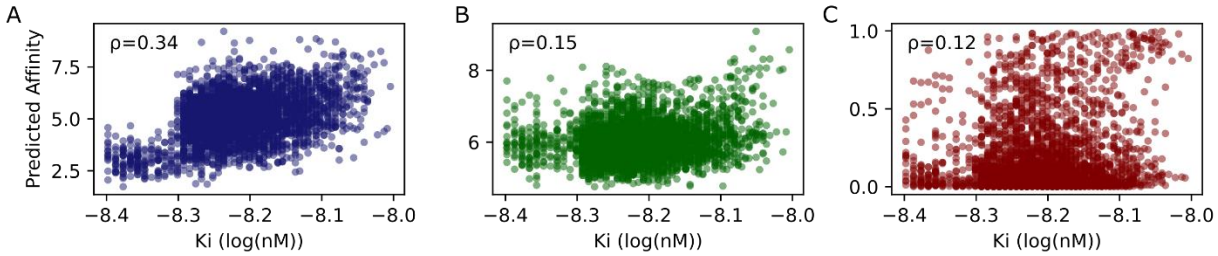

Figure 4. Comparison of the ability of Yuel, DeepDTA, and DeepConv-DTI to predict compound-protein interactions in the Metz dataset. All models are trained on the Davis dataset. (A) The relationships between the compound-protein interaction predicted by Yuel and the experimental affinity data of the Metz test set. (B) The relationships between the compound-protein interaction predicted by DeepDTA and the experimental affinity data of the Metz test set. (C) The relationships between the compound-protein interaction predicted by DeepConv-DTI and the experimental affinity data of the Metz test set.

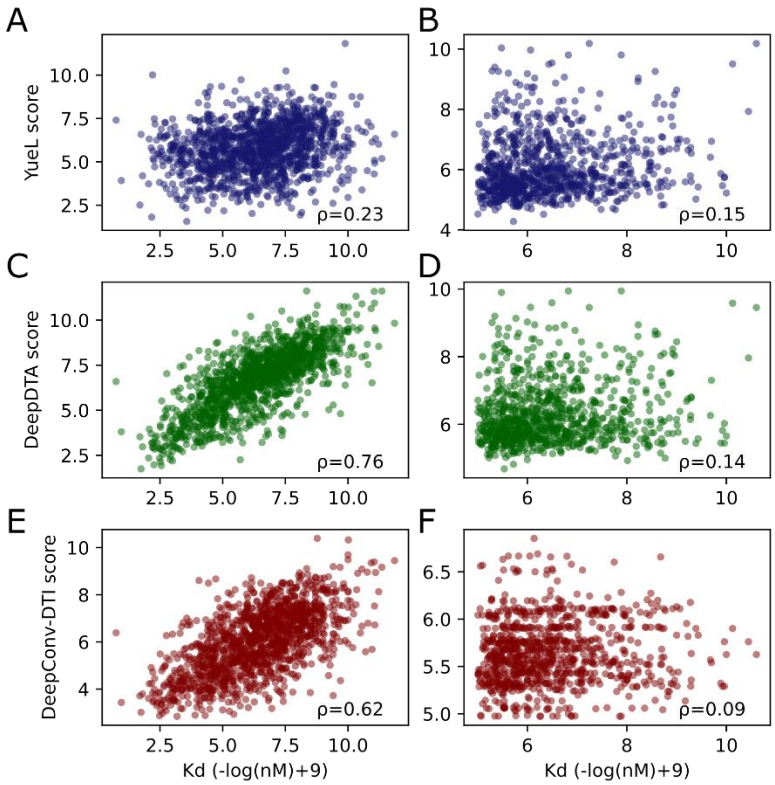

Figure 5. Comparison of Yuel, DeepDTA, and DeepConv-DTI in the shuffled PDBbind and Davis datasets. The models are trained on the normal training set, but the amino acids in each sequence of the test set are shuffled. (A) The relationships between the compound-protein interaction predicted by Yuel and the experimental affinity data of the PDBbind test set. (B) The relationships between the compound-protein interaction predicted by Yuel and the experimental affinity data of the Davis test set. (C) The relationships between the compound-protein interaction predicted by DeepDTA and the experimental affinity data of the PDBbind test set. (D) The relationships between the compound-protein interaction predicted by DeepDTA and the experimental affinity data of the Davis test set. (E) The relationships between the compound-protein interaction predicted by DeepConv-DTI and the experimental affinity data of the PDBbind test set. (F) The relationships between the compound-protein interaction predicted by DeepConv-DTI and the experimental affinity data of the Davis test set.