doi:10.20944/preprints202201.0365.v1

Article

Effects of Data Quality and Quantity on Deep Learning for Protein-Ligand Binding Affinity Prediction

Frankie J Fan [1] and Yun Shi [2,*]

[1] School of Health, Medical and Applied Sciences, Central Queensland University, Bundaberg, Queensland 4670, Australia; f.fan@cqu.edu.au and hustakin@gmail.com
[2] Institute for Glycomics, Griffith University, Southport, Queensland 4222, Australia; y.shi@griffith.edu.au
* Correspondence: y.shi@griffith.edu.au

Abstract: Prediction of protein-ligand binding affinities is crucial for computational drug discovery. A number of deep learning approaches have been developed in recent years to improve the accuracy of such affinity prediction. While the predicting power of these models have advanced to some degrees depending on the dataset used for training and testing, the effects of the quality and quantity of the underlying data have not been thoroughly examined. In this study, we employed erroneous datasets and data subsets of different sizes, created from one of the largest databases of experimental binding affinities, to train and evaluate a deep learning system based on convolutional neural networks. Our results show that data quality and quantity do have significant impacts on the performance of trained models. Depending the variations in data quality and quantity, the performance differences could be comparable to or even larger than those observed among different deep learning approaches. This implies that continued accumulation of high-quality affinity data is important for improving deep learning models to better predict protein-ligand binding affinities.

Keywords: binding affinity prediction; machine learning; data quality; data quantity; deep learning

1. Introduction

Computational approaches using artificial intelligence techniques, especially machine learning (ML), have been increasingly utilized during various stages of pharmaceutical drug discovery and development in recent years [1–3]. Unlike the physics-centric “expert systems” traditionally used in computational drug discovery [4], ML-based methods are data-centric [5] and focus on learning from experience [6].

Numerous ML approaches, including deep learning (DL) ones, have recently been developed to predict protein-ligand interactions [7], as the affinity of such interaction is critical because it usually correlates with the activity of a drug (ligand) on its therapeutic target (protein). Additionally, several DL methods have shown great promises in accurate prediction of protein-ligand binding affinities [8–13].

However, most efforts in developing DL methods for affinity prediction have been focused on changes in the DL approach itself to improve performance, which may depend on the dataset used for model training and testing, while the effects of data quality and quantity have not been well characterized. In addition, it is generally believed that ML requires abundant, high-quality data and that data processing and cleaning constitutes at least 80% of ML practice while the application of algorithm only accounts for 20% [2].

In light of the pivotal role of data, we set out to investigate the extent to which data quality and quantity would influence the performance of ML approaches for protein-ligand affinity prediction. In this work, we employed a relatively user-friendly DL tool for protein-ligand interaction prediction, DeepPurpose [14], and one of the most comprehensive databases for protein-ligand binding affinities, BindingDB [15]. To provide insights into the possible effect of data quality, we introduce intentional errors of different degrees to the affinity label. To illustrate the potential effect of data quantity, we perform random selection of data subsets of varying sizes. Performance comparison of models trained on these manipulated datasets indicates that data quality and quantity do have a significant impact, and research efforts should be directed towards the continued collection and curation of high-quality affinity data.

2. Results

2.1. The datasets

Following data curation as detailed in Materials and Methods, we have obtained a dataset, herein referred to as the KDKI set, that consists of 365,021 unique entries. Among them are 199,138 unique ligand SMILES strings and 3,835 unique protein sequences. Most of these 365,021 entries have the length of ligand SMILES strings between 30 and 70 characters, and the length of protein amino-acid sequence between 300 and 500 residues (Figure 1a and b). Their binding affinities also show a reasonable distribution (Figure 1c), with most pK values between 5 and 9, i.e. K_d *or* K_i values between 1 nM and 10 μM. There are 33,728 protein-ligand pairs that each had multiple pK values prior to their inclusion in the KDKI dataset, and only the average of these multiple pK values were included in the KDKI dataset for each protein-ligand pair, ensuring each entry has a unique protein-ligand pair. Analysis of the range of multiple pK values illustrated that they are within a difference of 1, i.e. K_d *or* K_i values within one order of magnitude, for most of the 33,728 protein-ligand pairs (Figure 1d). This implies that data in the KDKI set are of good quality.

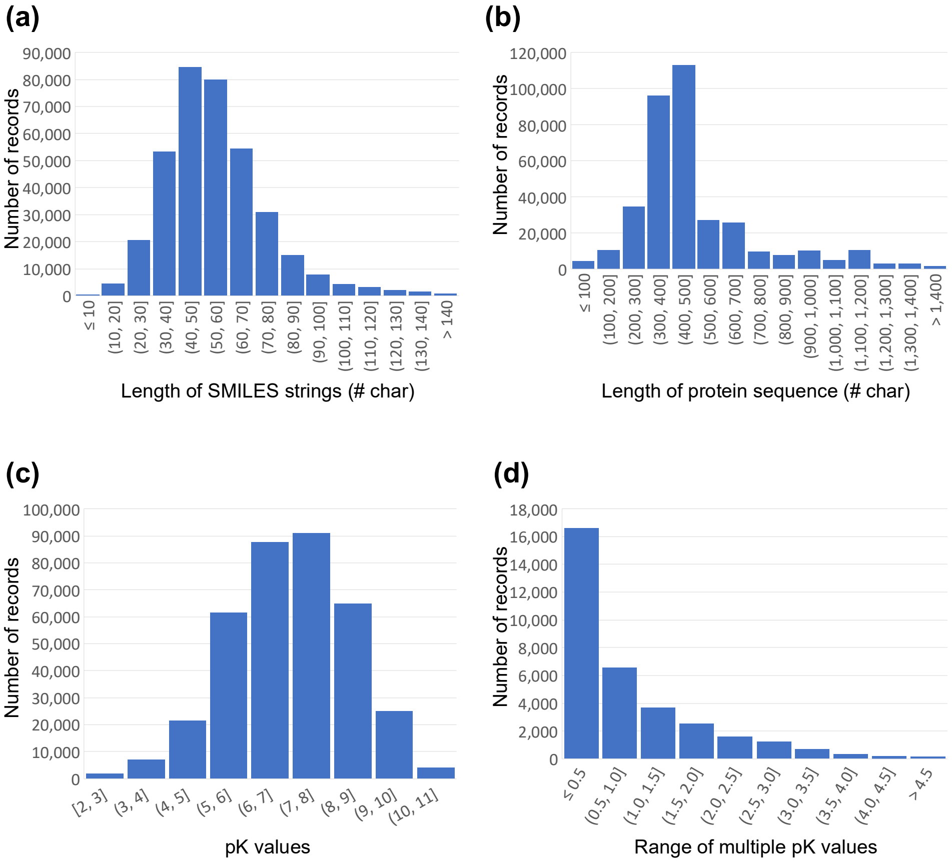

Figure 1. Histograms showing characteristics of the KDKI set: length distributions for (**a**) ligand SMILES strings and (**b**) protein sequence, as well as distributions of (**c**) pK values and (**d**) range of affinities for protein-ligand pairs with multiple pK values prior to inclusion in the KDKI set.

The KDKI set of 365,021 entries was subsequently divided into an "external" test set of 45,021 entries and a Baseline set of 320,000 entries, with no overlaps between these two sets. We then manipulated the Baseline set to introduce intentional errors to the pK value, mimicking possible experimental errors/variations and/or data curation errors as indicated by the range of multiple pK values (Figure 1d). In addition, subsets of different sizes were created from the Baseline set in order to probe the effect of data quantity, and the external test set was also divided into four subsets to examine the effect of the presence (or absence) of certain data contents (Table S1).

2.2. Effect of data quality on model performance

We implemented a DL model within DeepPurpose [14] that uses two convolutional neural network (CNN) blocks to learn representations for proteins and ligands based on their amino-acid sequences and SMILES strings, respectively, similar to DeepDTA [8]. Models trained on the Baseline set and datasets with incorrect pK values were tested on the external test set of 45,021 entries (Table 2). As expected, both the Pearson correlation coefficient (PCC) and the concordance index declined with increasing levels of errors in pK values, whereas discrepancies between predicted values and experimental values, as measured by root mean square error (RMSE) and mean square error (MSE), increased. More dramatic changes in performance were observed when each model was tested on its internal test set, likely due to the introduction of random errors to the internal test set in addition to its training and validation sets (Table S2).

Table 1. Performance comparison of models trained on datasets with different degrees of errors and tested on the external test set.

Training set	RMSE[1]	MSE[1]	PCC[1]	Concordance[1]
Baseline set	0.868 (0.008)	0.753 (0.014)	0.804 (0.004)	0.802 (0.002)
pK + [-1,1]	0.925 (0.005)	0.855 (0.009)	0.774 (0.002)	0.785 (0.001)
pK + [-2,2]	1.012 (0.008)	1.024 (0.016)	0.721 (0.006)	0.756 (0.003)
pK + [-3,3]	1.076 (0.006)	1.158 (0.014)	0.678 (0.004)	0.736 (0.002)
pK +1/-1	0.989 (0.009)	0.978 (0.018)	0.737 (0.003)	0.765 (0.001)
pK +2/-2	1.110 (0.007)	1.232 (0.017)	0.652 (0.005)	0.723 (0.002)
pK +3/-3	1.163 (0.003)	1.352 (0.008)	0.606 (0.004)	0.702 (0.002)

[1] Reported values are mean (standard deviation) of triplicate runs.

2.3. Effect of data quantity on model performance

The performance of models trained on a series of subsets of the Baseline set exhibited a clear dependence on the size of the training set when tested on the external test set (Figure 2). The RMSE showed a significant decrease from over 1.4 to below 0.9, while the PCC rose from about 0.3 to about 0.8 with the size of the dataset increasing from 1,250 to 320,000 (Table S3). Similar changes in performance metrics were observed when each model was tested on its internal test set, indicating similar data qualities among these subsets of different sizes and the external test set (Table S4).

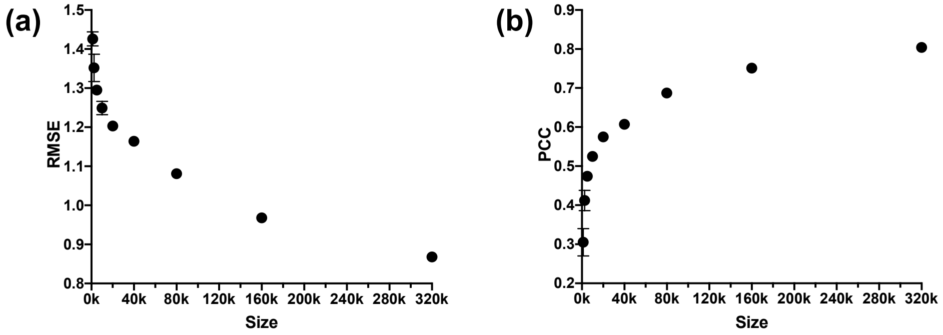

Figure 2. Performance of models trained on data subsets of different sizes and tested on the external test set, showing (**a**) root mean square error (RMSE) and (**b**) Pearson correlation coefficient (PCC) for different models. Data points represent mean values of triplicates and error bars indicate standard deviations, some of which are too small to show.

To further probe the effect of the presence (or absence) of ligand and/or the protein in the Baseline set, the model trained on the Baseline set was tested on corresponding subsets of the external test set. The results demonstrated that the presence of target (protein) in the training data has a substantial effect on the performance of the DL model, with >0.42 fall in RMSE and >0.24 rise in PCC, whereas the presence of drug (ligand) in the training data showed much less effect in improving the prediction performance (Table 2).

Table 2. Performance comparison of models trained on the baseline set and tested on different subsets of the external test set.

Testing set	RMSE[1]	MSE[1]	PCC[1]	Concordance[1]
External test set	0.868 (0.008)	0.753 (0.014)	0.804 (0.004)	0.802 (0.002)
Ligand seen, protein seen	0.831 (0.008)	0.691 (0.014)	0.807 (0.004)	0.804 (0.002)
Ligand unseen, protein seen	0.927 (0.008)	0.859 (0.015)	0.798 (0.005)	0.797 (0.002)
Ligand seen, protein unseen	1.413 (0.172)	2.015 (0.501)	0.501 (0.197)	0.699 (0.055)
Ligand unseen, protein unseen	1.348 (0.122)	1.827 (0.334)	0.553 (0.214)	0.701 (0.078)

[1] Reported values are mean (standard deviation) of triplicate runs.

3. Discussion

Significant developments of DL models for protein-ligand affinity predictions have been achieved in recent years, with many of them outperforming physics-based molecular docking methods on certain datasets [10,16]. However, performance improvements among state-of-the-art DL approaches appeared to be small, with PCC approaching 1 (complete linearity) and RMSE values declining below 1, indicating most predicted K_d or K_i values are within one order of magnitude of experimental values [8,9,11–13]. Our default DL model that uses two CNN blocks without consideration of protein-ligand 3D structures also showed good performance with our Baseline set, as the > 0.8 PCC and < 0.9 RMSE values are comparable to the aforementioned state-of-the-art DL approaches. For example, DeepAffinity utilized a dataset of similar size to our Baseline set and achieved RMSE values between 0.73 to 0.94 and PCC between 0.76 and 0.86, depending on model setup [9], whereas we have corresponding RMSE of 0.870 and PCC of 0.803 (Table S2).

As the marginal performance improvements from novel DL approaches may not be very consequential for protein-ligand affinity prediction, we could benefit from analyzing the effects of data quality and quantity. Our results showed that even deviating from experimental pK values by up to one order of magnitude, a level of variations thought to be experimentally acceptable, could raise RMSE by 0.06 and reduce PCC by 0.03 (Table 1). These are already significant performance changes in comparison to differences among various state-of-the-art DL approaches. One of the most recently developed DL method, DeepDTAF, only improved the RMSE by 0.06 and the PCC by 0.01 over a previously-

doi:10.20944/preprints202201.0365.v1

developed DL model, Pafnucy, on a test set of only 290 protein–ligand pairs with known 3D structures [13,17]. Another state-of-the-art DL approach, MONN, only improved the RMSE by 0.03 and the PCC by 0.01 over DeepDTA, using a large dataset of over 260,000 training samples and over 110,000 test samples derived BindingDB database with IC_{50} values [11]. Therefore, it is possible that data quality could have a larger impact on the predicting power of the DL system on protein-ligand affinities than the DL system itself.

Data quantity also appears to have a pronounced effect on the prediction performance, as doubling the amount of data led to > 0.05 improvements for both RMSE and PCC in most cases (Table S3). The presence of ligand and/or protein structures in data used for model training showed an even larger effect, with the latter improving PCC by 0.24~0.31 (Table 2). Such a dramatic effect is consistent with previous studies showing that the absence of proteins and/or ligands in the training set could result in drastic reduction in DL model performance in predicting affinities, regardless of whether experimental 3D structures were used as training input [9,11]. This highlights the importance of collecting experimental data on new targets (proteins) without known binding affinities. Another study demonstrated that some DL models still provide good performance with up to 95% data missing from their original dataset, but only when they are predicting interactions (yes or no) rather than affinities (pK values) [18].

In summary, we have demonstrated the crucial roles that data can play in improving DL models for protein-ligand affinity prediction. Although the effects of data quality and quantity determined in this study are to some extent expected, our results do suggest that further collection and curation of quality data are as critical as improving the DL approach itself, if not more so, for more accurate prediction of protein-ligand binding affinities.

4. Materials and Methods

4.1. Data curation and manipulation

The raw BindingDB dataset was retrieved from https://www.bindingdb.org/ by downloading the file BindingDB_All_2021m11.tsv.zip. This expanded file contains over 2 million entries with affinity measurements in equilibrium dissociation constant (K_d), inhibition constant (K_i), which is essentially the dissociation constant for an inhibitor, half maximal inhibitory concentration (IC_{50}), and half maximal effective concentration (EC_{50}). Since IC_{50} and EC_{50} values are dependent on the concentration of proteins, we only selected entries with K_d and K_i values. We removed entries without SMILES strings and entries without the number of protein chains being 1. We also removed samples with incomplete information to determine protein sequences (such as those containing the letter “X", lower-case letters, or Arabic numerals). Imprecise affinity values with < or > prefixes as well as extreme values outside of range of (0.01 nM, 10 mM) were also deleted. As a result, we found 95% of remaining entries have 5 to 150 SMILES characters and 50 to 1500 amino acids in the protein sequence, and thereby removed entries outside of such length ranges. We next converted each affinity value K to log space pK so that the new label equals $-\log_{10}(K_d \text{ or } K_i)$ for easier regression. Moreover, for each protein-ligand pair with multiple affinity measurements, we used the geometric mean of all its affinity values, i.e. the arithmetic mean of its pK values, as its only affinity label after merging. Finally, we ended up with a dataset, namely the KDKI set, of 365,021 records stored in a csv file, and each entry comprises a ligand SMILES string, a protein amino-acid sequence, and an affinity label pK calculated from $-\log_{10}(K_d \text{ or } K_i)$.

We randomly selected 45,021 entries from the KDKI set as an independent “external” test set. To ensure reproducibility, we performed the selection in triplicates, producing three distinct testing sets, each with 45,021 entries, and three different remaining sets, each with 320,000 entries. The remaining set of 320,000 entries was referred to as the Baseline set.

The external test set of 45,021 records was divided into four subsets: 1) with ligand found (seen) and protein found (seen) in the Baseline set; 2) with ligand not found (unseen) and protein found (seen) in the Baseline set; 3) with ligand found (seen) and protein not found (unseen) in the Baseline set; 4) with ligand not found (unseen) and protein not found (unseen) in the Baseline set. The Baseline set of 320,000 entries was subject to the following manipulations. To test the effect of data quality, we deliberately introduced errors to the affinity values in two different ways. On one hand, we added a random float number from the uniform distribution in the ranges of [-1,1], [-2,2], and [-3,3], respectively, to the pK value of each entry, resulting in three datasets named pK + [-1,1], pK + [-2,2], and pK + [-3,3], respectively. On the other hand, we added a random number of either 1 or -1, either 2 or -2, and either 3 or -3, to the pK value of each entry, leading to another three datasets known as pK +3/-3, pK +3/-3, pK +3/-3, respectively. Furthermore, we randomly selected subsets with 1,250, 2,500, 5,000, 10,000, 20,000, 40,000, 80,000, and 160,000 entries, respectively, resulting in 8 subsets with different amounts of missing data (Table S1).

4.2. Model training and testing

Training of models with different datasets were performed using two CNN blocks to encode ligand SMILES strings and protein amino-acid sequences, respectively. Default values were kept for other parameters as provided by DeepPurpose, as this setting has been shown to reproduce similar results from DeepDTA [14]. A random splitting of data was carried out prior to training with a ratio of 7:1:2 for training, validation, and testing sets, respectively. The same numbers of the filters were used for both target and drug CNN blocks, i.e. 32, 64, and 96 for the first, second, and third layers, respectively. The corresponding lengths of the filter size for drugs and targets were [4, 6, 8] and [4, 8, 12], respectively. Dimensions of the hidden neurons were set to [1024, 1024, 512]. The training was conducted with 100 epochs and mini-batch size of 256 to update the weights of the network. The default learning rate was set to 0.001. To accelerate the training calculation, a Tesla P100-PCIE-16GB GPU on CQUniversity Marie Curie HPC Cluster was utilized.

Internal testing of the trained model was automatically performed with the aforementioned randomly split testing set by DeepPurpose [14]. We also tested all trained models on our “external” test set. Four evaluation metrics were used for testing, namely root mean squared error (MSE), mean squared error (MSE), Pearson correlation coefficient (PCC) [19], and concordance index [20], to measure the differences and correlations between predicted pK values and experimental pK values stored in the test set.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, List S1: The KDKI set in a csv file; Table S1: Description of all datasets used in this study. Table S2: Performance comparison of models trained on datasets with different degrees of errors and tested on individual internal test sets; Table S3: Performance comparison of models trained on data subsets of different sizes and tested on the external test set; Table S4: Performance comparison of models trained on data subsets of different sizes and tested on individual internal test sets.

Author Contributions: Conceptualization, Y.S.; methodology, Y.S. and F.J.F.; software, F.J.F.; validation, Y.S. and F.J.F.; formal analysis, Y.S. and F.J.F.; investigation, F.J.F. and Y.S.; resources, F.J.F.; data curation, F.J.F. and Y.S..; writing—original draft preparation, Y.S.; writing—review and editing, Y.S. and F.J.F.; supervision, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Our source codes for data manipulation, model training, and model testing were adapted from DeepPurpose and are stored at GitHub (https://github.com/hustakin/DeepPurpose/). The KDKI set was provided as a compressed csv file in Supplementary Materials.

Acknowledgments: We gratefully acknowledge CQUniversity's eResearch support and the use of the High Performance Computing facility (www.cqu.edu.au/hpc) in the completion of this work. This research was also supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smith, J.S.; Roitberg, A.E.; Isayev, O. Transforming Computational Drug Discovery with Machine Learning and AI. *ACS Med. Chem. Lett.* **2018**, *9*, 1065–1069, doi:10.1021/acsmedchemlett.8b00437.
2. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477, doi:10.1038/s41573-019-0024-5.
3. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial Intelligence in Drug Discovery and Development. *Drug Discov. Today* **2021**, *26*, 80–93, doi:10.1016/j.drudis.2020.10.010.
4. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395, doi:10.1124/pr.112.007336.
5. Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* **2009**, *24*, 8–12, doi:10.1109/MIS.2009.36.
6. Mitchell, T.M. *Machine Learning*; 1st ed.; McGraw-Hill Education: New York, United States, 1997; ISBN 0-07-042807-7.
7. Dhakal, A.; McKay, C.; Tanner, J.J.; Cheng, J. Artificial Intelligence in the Prediction of Protein–Ligand Interactions: Recent Advances and Future Directions. *Brief. Bioinform.* **2021**, bbab476, doi:10.1093/bib/bbab476.
8. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug–Target Binding Affinity Prediction. *Bioinformatics* **2018**, *34*, i821–i829, doi:10.1093/bioinformatics/bty593.
9. Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: Interpretable Deep Learning of Compound–Protein Affinity through Unified Recurrent and Convolutional Neural Networks. *Bioinformatics* **2019**, *35*, 3329–3338, doi:10.1093/bioinformatics/btz111.
10. Zheng, L.; Fan, J.; Mu, Y. OnionNet: A Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. *ACS Omega* **2019**, *4*, 15956–15965, doi:10.1021/acsomega.9b01997.
11. Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: A Multi-Objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Syst.* **2020**, *10*, 308-322.e11, doi:10.1016/j.cels.2020.03.002.
12. Nguyen, T.; Le, H.; Quinn, T.P.; Nguyen, T.; Le, T.D.; Venkatesh, S. GraphDTA: Predicting Drug–Target Binding Affinity with Graph Neural Networks. *Bioinformatics* **2021**, *37*, 1140–1147, doi:10.1093/bioinformatics/btaa921.
13. Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: A Deep Learning Method to Predict Protein–Ligand Binding Affinity. *Brief. Bioinform.* **2021**, *22*, bbab072, doi:10.1093/bib/bbab072.
14. Huang, K.; Fu, T.; Glass, L.M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: A Deep Learning Library for Drug–Target Interaction Prediction. *Bioinformatics* **2020**, *36*, 5545–5547, doi:10.1093/bioinformatics/btaa1005.
15. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein–Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201, doi:10.1093/nar/gkl999.
16. Zhu, F.; Zhang, X.; Allen, J.E.; Jones, D.; Lightstone, F.C. Binding Affinity Prediction by Pairwise Function Based on Neural Network. *J. Chem. Inf. Model.* **2020**, *60*, 2766–2772, doi:10.1021/acs.jcim.0c00026.
17. Stepniewska-Dziubinska, M.M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction. *Bioinformatics* **2018**, *34*, 3666–3674, doi:10.1093/bioinformatics/bty374.
18. Huang, K. *MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction*; 2021;
19. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242.
20. Gönen, M.; Heller, G. Concordance Probability and Discriminatory Power in Proportional Hazards Regression. *Biometrika* **2005**, *92*, 965–970, doi:10.1093/biomet/92.4.965.