

Supplementary Materials

How Data Quality and Quantity Affect a Deep Learning Model for Protein-Ligand Binding Affinity Prediction

Frankie J Fan¹ and Yun Shi^{2,*}

¹School of Health, Medical and Applied Sciences, Central Queensland University, Bundaberg, Queensland 4670, Australia; f.fan@cqu.edu.au and hustakin@gmail.com

²Institute for Glycomics, Griffith University, Southport, Queensland 4222, Australia; y.shi@griffith.edu.au

*Correspondence: y.shi@griffith.edu.au

Supplementary List

Included in a separate file

Supplementary Tables

Page 2 - 5

Table S1. Description of all datasets used in this study.

Dataset¹	Size	Description
KDKI set	365,021	Data curated from BindingDB
External test set	45,021	Random division of the KDKI set
Baseline set	320,000	
pK + [-1,1]	320,000	Adding a random value between -1 and 1 to each pK
pK + [-2,2]	320,000	Adding a random value between -2 and 2 to each pK
pK + [-3,3]	320,000	Adding a random value between -3 and 3 to each pK
pK +1/-1	320,000	Adding a random value of either 1 or -1 to each pK
pK +2/-2	320,000	Adding a random value of either 2 or -2 to each pK
pK +3/-3	320,000	Adding a random value of either 3 or -3 to each pK
160k set	160,000	Randomly selected from the Baseline set
80k set	80,000	
40k set	40,000	
20k set	20,000	
10k set	10,000	
5k set	5,000	
2.5k set	2,500	
1.25k set	1,250	
Ligand seen, protein seen	28,860~29,189	Subsets of the external test set based on the presence/absence (seen/unseen) of ligand/protein in the Baseline set, varying in size among triplicates
Ligand unseen, protein seen	15,759~16,059	
Ligand seen, protein unseen	46~77	
Ligand unseen, protein unseen	27~33	

¹ Except the KDKI set, all datasets were generated in triplicates.

Table S2. Performance comparison of models trained on datasets with different degrees of errors and tested on individual internal test sets.

Dataset	RMSE¹	MSE¹	PCC¹	Concordance¹
Baseline set	0.870 (0.002)	0.757 (0.004)	0.803 (0.001)	0.802 (0.001)
pK + [-1,1]	1.092 (0.003)	1.192 (0.007)	0.720 (0.002)	0.757 (0.001)
pK + [-2,2]	1.539 (0.008)	2.368 (0.024)	0.565 (0.006)	0.687 (0.002)
pK + [-3,3]	2.046 (0.005)	4.185 (0.019)	0.434 (0.003)	0.638 (0.001)
pK +1/-1	1.411 (0.001)	1.990 (0.002)	0.607 (0.002)	0.704 (0.001)
pK +2/-2	2.288 (0.006)	5.235 (0.026)	0.386 (0.006)	0.624 (0.002)
pK +3/-3	3.214 (0.003)	10.332 (0.17)	0.266 (0.005)	0.602 (0.002)

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

Table S3. Performance comparison of models trained on data subsets of different sizes and tested on the external test set.

Dataset	RMSE¹	MSE¹	PCC¹	Concordance¹
Baseline set	0.868 (0.008)	0.753 (0.014)	0.804 (0.004)	0.802 (0.002)
160k set	0.968 (0.008)	0.937 (0.015)	0.751 (0.006)	0.775 (0.003)
80k set	1.081 (0.009)	1.169 (0.020)	0.687 (0.003)	0.743 (0.002)
40k set	1.164 (0.011)	1.356 (0.026)	0.607 (0.007)	0.703 (0.004)
20k set	1.203 (0.017)	1.448 (0.026)	0.575 (0.004)	0.690 (0.003)
10k set	1.249 (0.002)	1.561 (0.041)	0.525 (0.007)	0.670 (0.003)
5k set	1.295 (0.035)	1.678 (0.005)	0.474 (0.012)	0.651 (0.005)
2.5k set	1.352 (0.018)	1.829 (0.096)	0.412 (0.026)	0.629 (0.009)
1.25k set	1.426 (0.046)	2.034 (0.050)	0.305 (0.035)	0.591 (0.015)

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

Table S4. Performance comparison of models trained on data subsets of different sizes and tested on individual internal test sets.

Dataset	RMSE¹	MSE¹	PCC¹	Concordance¹
Baseline set	0.870 (0.002)	0.757 (0.004)	0.803 (0.001)	0.802 (0.001)
160k set	0.969 (0.002)	0.939 (0.005)	0.751 (0.002)	0.774 (0.001)
80k set	1.083 (0.006)	1.173 (0.014)	0.688 (0.005)	0.742 (0.002)
40k set	1.173 (0.010)	1.376 (0.024)	0.597 (0.003)	0.698 (0.001)
20k set	1.197 (0.017)	1.433 (0.040)	0.577 (0.004)	0.690 (0.002)
10k set	1.263 (0.033)	1.595 (0.082)	0.528 (0.021)	0.671 (0.009)
5k set	1.259 (0.006)	1.584 (0.014)	0.461 (0.022)	0.645 (0.002)
2.5k set	1.304 (0.048)	1.701 (0.123)	0.449 (0.023)	0.643 (0.020)
1.25k set	1.407 (0.079)	1.984 (0.219)	0.296 (0.060)	0.593 (0.054)

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