

Supplementary Material for “MLT-LE: predicting drug–target binding affinity with multi-task residual neural networks”

Elizaveta Vinogradova^{1,2}, Karina Pats^{1,2}, Ferdinand Molnár^{2*}, Siamac Fazli^{3*},

1 Department of Biology, Nazarbayev University, Nur-Sultan, Kazakhstan

2 Computer Technologies Laboratory, ITMO University, Russia

3 Department of Computer Science, Nazarbayev University, Nur-Sultan, Kazakhstan

* ferdinand.molnar@nu.edu.kz; siamac.fazli@nu.edu.kz

Figures 1-2 below show a comparison of model performance on validation and test data when training for 100 epochs with a smaller batch size of 256. This experiment uses the same data partitions and predictive models as the main study. As can be seen in Figures 1-2 the performance of the models can be considered consistent for the two frameworks for two different batch sizes.

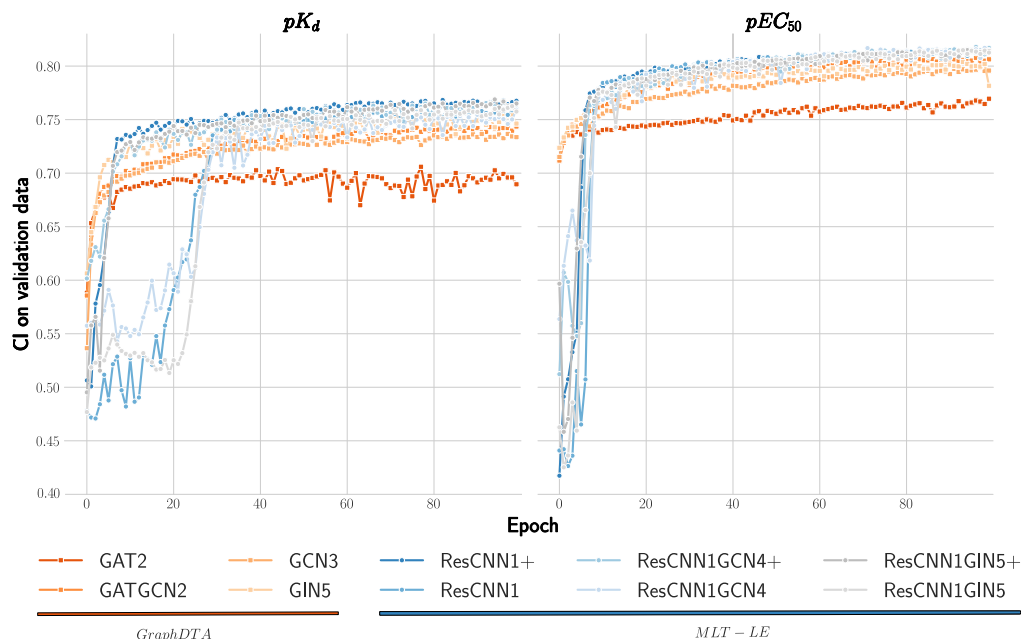


Figure 1. Comparison of model performance on the validation data for K_d and EC_{50} data when training in 100 epochs with a smaller batch size of 256. GraphDTA performance is shown in orange and MLT-LE performance is shown in blue.

Figure 3 shows the performance of the best MLT-LE model (ResCNN1GCN4) for all tasks EC_{50} subset test data.

Statistical significance test of performance on EC_{50} data

It was observed that the prediction accuracy of the GraphDTA models is lower than the prediction accuracy of the MLT-LE models in the test sets. The distribution of

		pK_d					pEC_{50}				
		RMSE	MSE	Pearson	Spearman	CI	RMSE	MSE	Pearson	Spearman	CI
$GraphDTA$	GAT2	1.213	1.471	0.576	0.531	0.689	1.034	1.070	0.658	0.644	0.733
	GATGCN2	1.055	1.114	0.703	0.657	0.743	0.823	0.678	0.803	0.802	0.806
	GCN3	1.083	1.172	0.685	0.631	0.733	0.850	0.723	0.791	0.788	0.799
	GIN5	1.065	1.134	0.709	0.667	0.749	0.832	0.692	0.798	0.794	0.802
$MLT - LE$	ResCNN1+	1.018	1.036	0.735	0.693	0.758	0.850	0.722	0.807	0.804	0.807
	ResCNN1	1.041	1.083	0.729	0.692	0.758	0.831	0.691	0.800	0.798	0.804
	ResCNN1GCN4+	1.036	1.074	0.726	0.687	0.756	0.811	0.658	0.814	0.810	0.811
	ResCNN1GCN4	1.060	1.123	0.716	0.673	0.751	0.794	0.631	0.819	0.815	0.814
	ResCNN1GIN5+	1.028	1.056	0.729	0.688	0.757	0.817	0.667	0.813	0.811	0.811
	ResCNN1GIN5	1.074	1.153	0.722	0.685	0.756	0.809	0.654	0.813	0.810	0.811

Figure 2. Comparison of the performances of the best models on the test set for K_d and EC_{50} data when training in 100 epochs with a smaller batch size of 256. The best scores obtained for each metric are shown in bold.

predictions between the two best models from each framework for the EC_{50} dataset can be seen in Fig.4.

In Table 1, we verify that the observed difference in predicted values between the models is statistically significant using a sign test.

Table 1. Significance testing

Given:	
N = 21154	Number of predicted values
Two-sided hypothesis:	
H₀: There is no difference between GraphDTA and MLT-LE values	
H₁: There is difference	
Number of positive and negative differences for sign-test:	
11999 - negative differences	Difference between matched pairs of predictions
9155 - positive differences	
Binomial test on pairs of differences:	
Number of trials = 21154	
Number of successes = 11999	
Probability of success = 50%	
P-value: 0.25E-84	
Conclusion	
The null hypothesis can be rejected with a p-value of 0.25E-84.	
The results suggest that the observation of such a difference between pairs of observations or more extreme is unlikely.	

The results of the performed test indicate that the observed difference in performance between the two frameworks in our experiment may be considered significant.

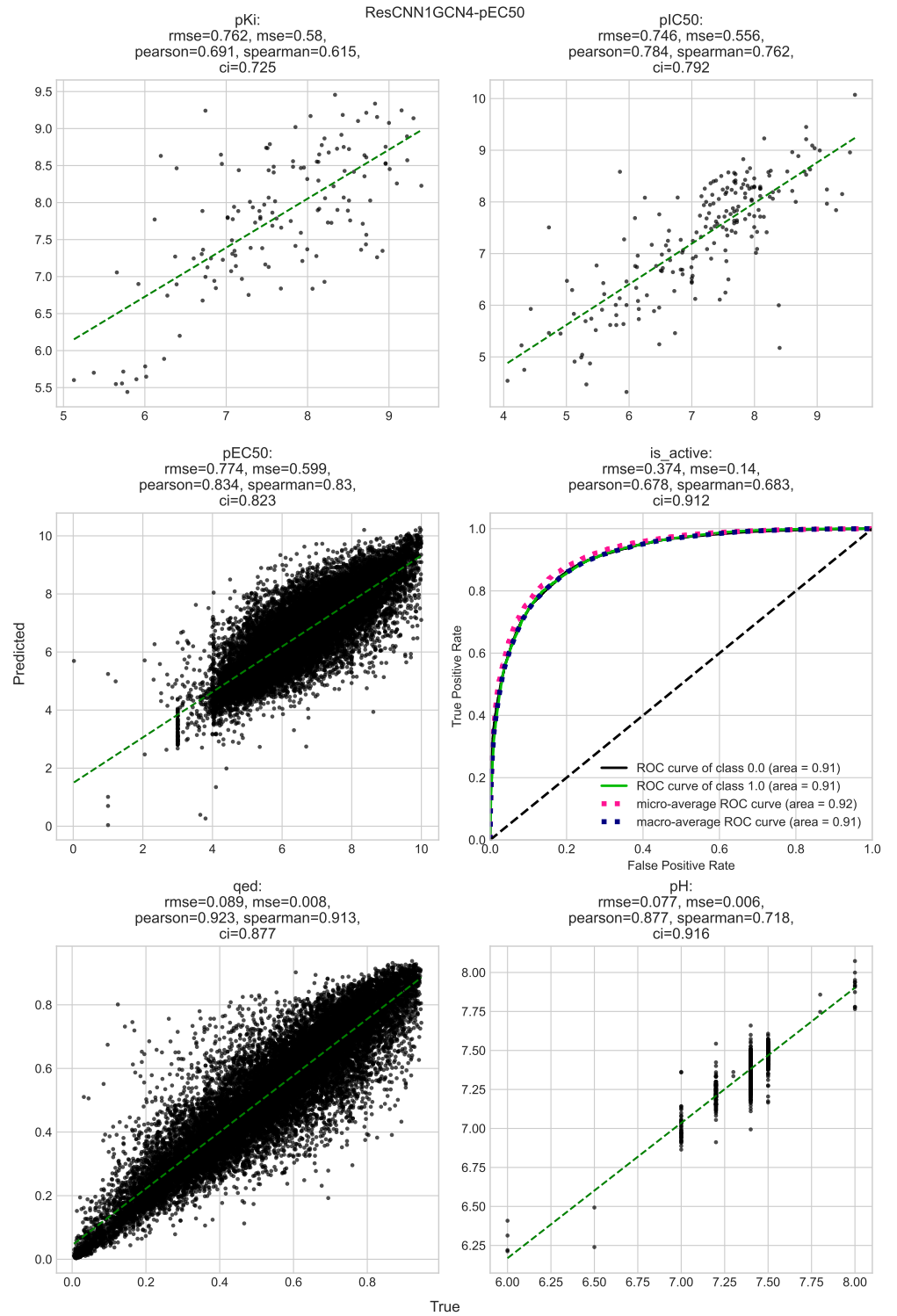


Figure 3. Performance of ResCNN1GCN4 model EC_{50} subset test data

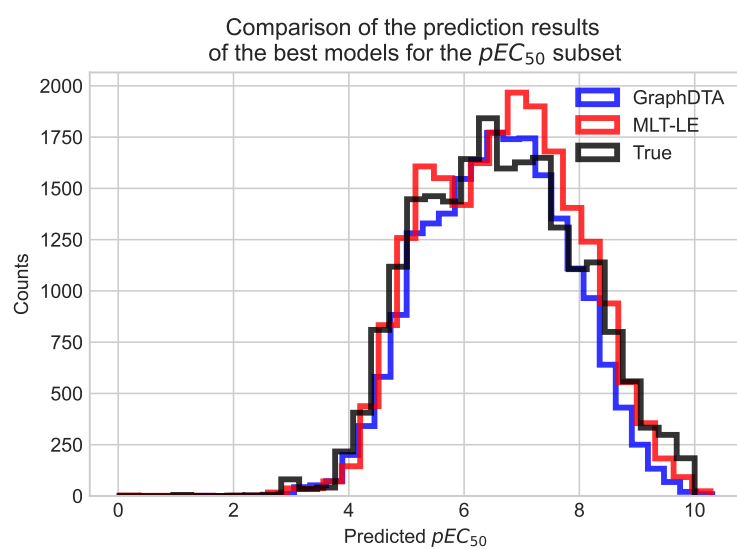


Figure 4. Distribution of predicted values compared to actual distribution