Supplement to: GraphDTA: Predicting drug--target binding anity with graph neural networks

1 Error analysis for drugs

We examined the latent space with regard to the prediction error by drugs. Figure 1 shows the median errors plotted against the first six principal components of the latent space, where we see that the hard-to-predict drugs usually appear close to the origin.

2 Error analysis for proteins

We now analyze the effect of homologous proteins on test set performance. To do this, we first cluster the target proteins in the **test set** using CLANS sequence clustering algorithm [1], which results in 5 major clusters (see Figure 2). Second, we calculate the average absolute error of each cluster (see Table 1 and Figure 3). Third, we estimate how well each cluster is represented by the **training set**. Here, we see that errors vary across the clusters. For example, cluster 2 has high error, while cluster 4 has low error. Yet, we find that the training set represents all 5 clusters equally well, by a 5:1 ratio. This suggests that the variation in test set performance is not simply explained by asymmetrical representation of protein families within the training set.

Cluster	#testing pairs	#training pairs	#proteins	Average absolute error
1	2159	10952	192	0.2372
2	1519	7455	132	0.3193
3	202	1022	18	0.2475
4	188	900	16	0.0867
5	117	563	10	0.2177

Table 1: The average absolute error of each cluster, and their representation in training and test sets.

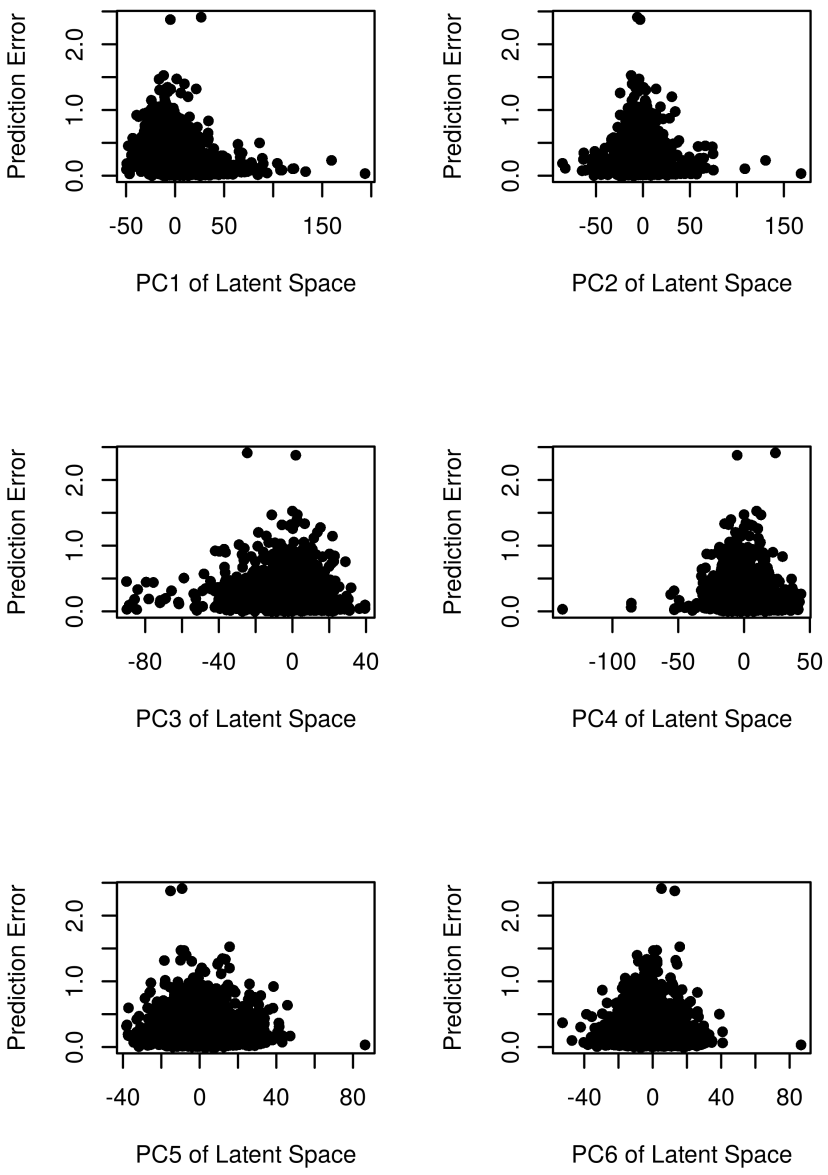

Figure 1: This figure shows the per-drug median errors plotted against the first 6 principal components, where we see that the hard-to-predict drugs usually appear close to the origin. We interpret this to mean that drugs with unique molecular sub-structures are always easy to predict.

Figure 2: Cluster distribution based on sequence similarity in 2D space.

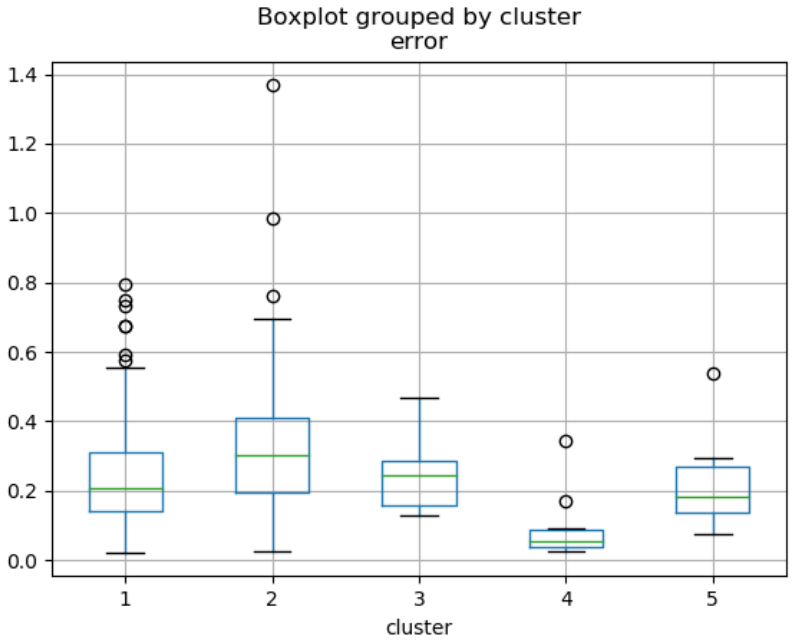

Figure 3: The absolute error of each cluster.

References

[1] Tancred Frickey and Andrei Lupas. CLANS: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*,

20(18):3702–3704, 2004.