Supplementary Materials for

BASE: a web service for providing compound-protein binding affinity prediction datasets with reduced similarity bias

Hyojin Son[1], Sechan Lee[1], Jaeuk Kim[1], Haangik Park[1], Myeong-Ha Hwang[1] and Gwan-Su Yi[1*]

1 Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

[*] **Corresponding author:** Gwan-Su Yi, Ph.D. **E-mail**: gwansuyi@kaist.ac.kr

This PDF file includes:

Figures. S1 to S2

Evaluation of binding affinity prediction for high CV compounds based on feature types

We evaluated the performance of various feature types and their combinations for the high CV group, focusing on binding affinity prediction (Fig. S1). As expected, ECFP4 alone was difficult to predict due to the variability in binding affinities across compounds for each protein in the high CV group. However, the combination of different features significantly improved predictive performance. Specifically, the combination of protein sequence and ECFP4 features showed a Pearson correlation coefficient of 0.845, matching the performance of models that included all three feature types (ECFP4, sequence, and interaction). The overall predictive performance for the high CV group was slightly lower than for the low CV group, but the trend remained consistent: combining features resulted in better predictions.

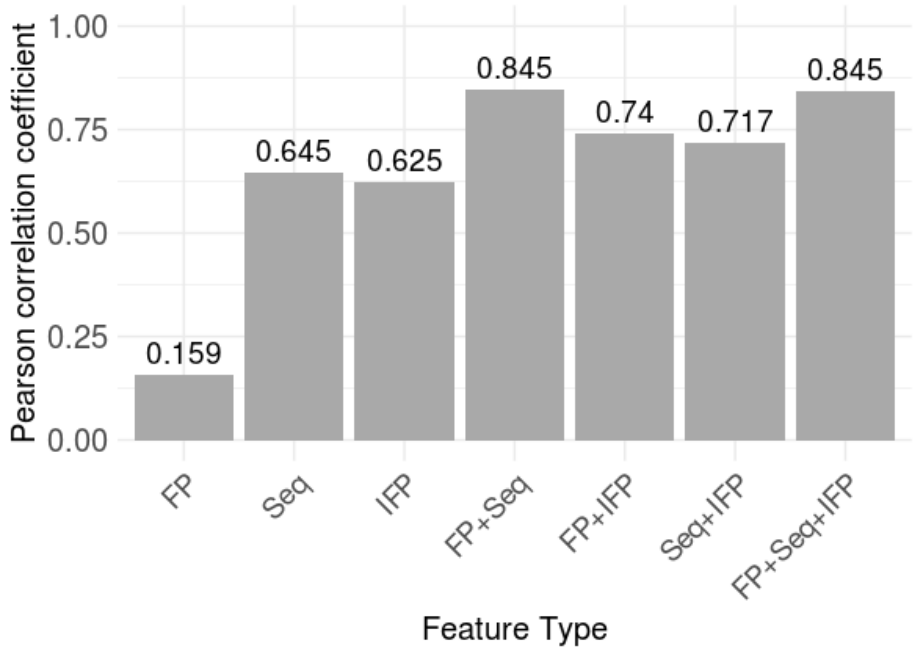

Fig. S1. Pearson correlation coefficient for binding affinity prediction based on feature types and their combinations for the high CV group.

SHAP analysis for feature importance in high CV compounds

Next, we used SHAP values to analyze feature importance in a model incorporating all three feature types (ECFP4, sequence, and interaction) (Fig. S2). This plot displays the top 20 most influential features in the combined model, ranked by their mean absolute SHAP values. Unlike the low CV group, where compound features dominated, protein features played a more significant role in the high CV group. Similar to the low CV group, no ECIF features were among the top 20. The prominence of protein features in the high CV group highlights their key role in predicting binding affinity, especially when compound features alone are insufficient to capture the variability.

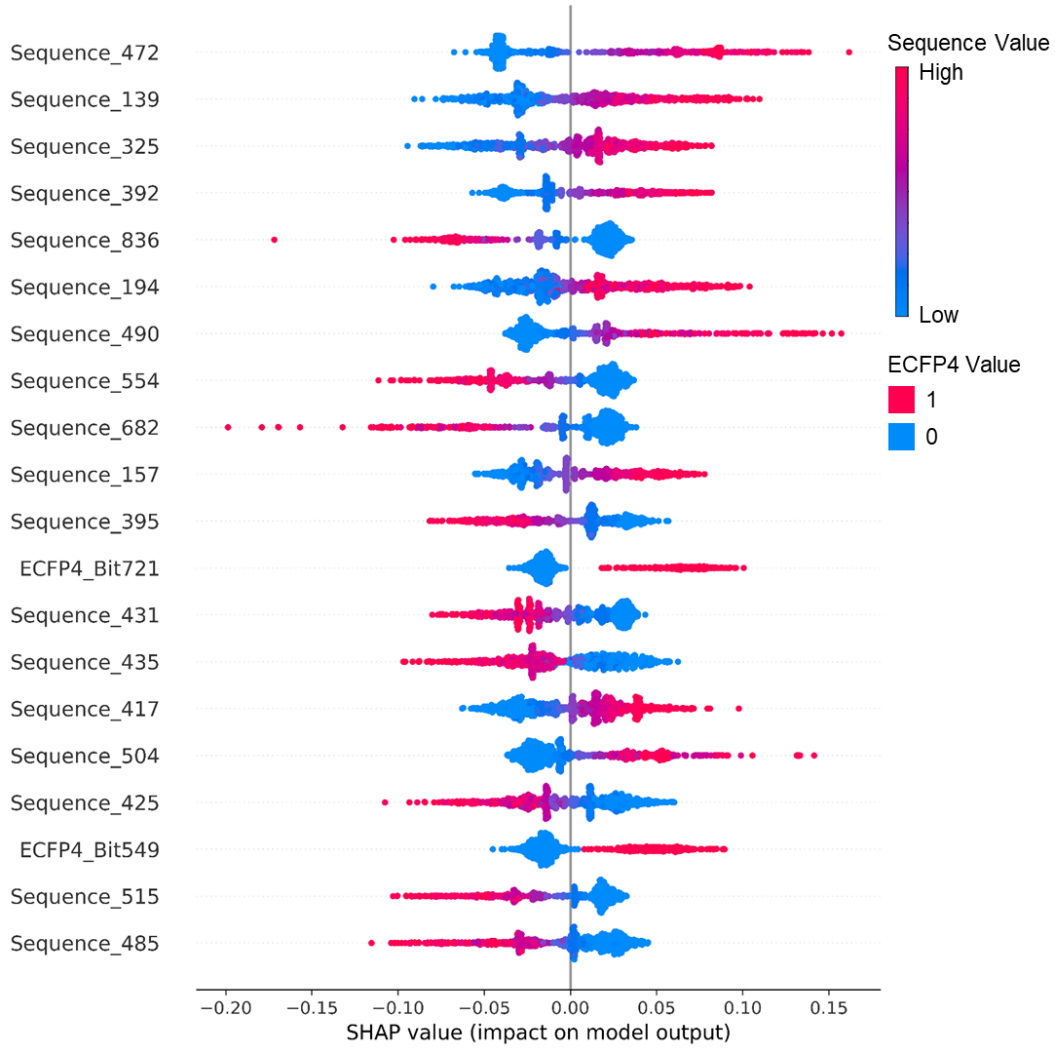

Fig. S2. SHAP analysis of feature importance in binding affinity predictions for the high CV group.