Supplementary Information for:

Machine learning-based chemical binding similarity using evolutionary relationships of target genes

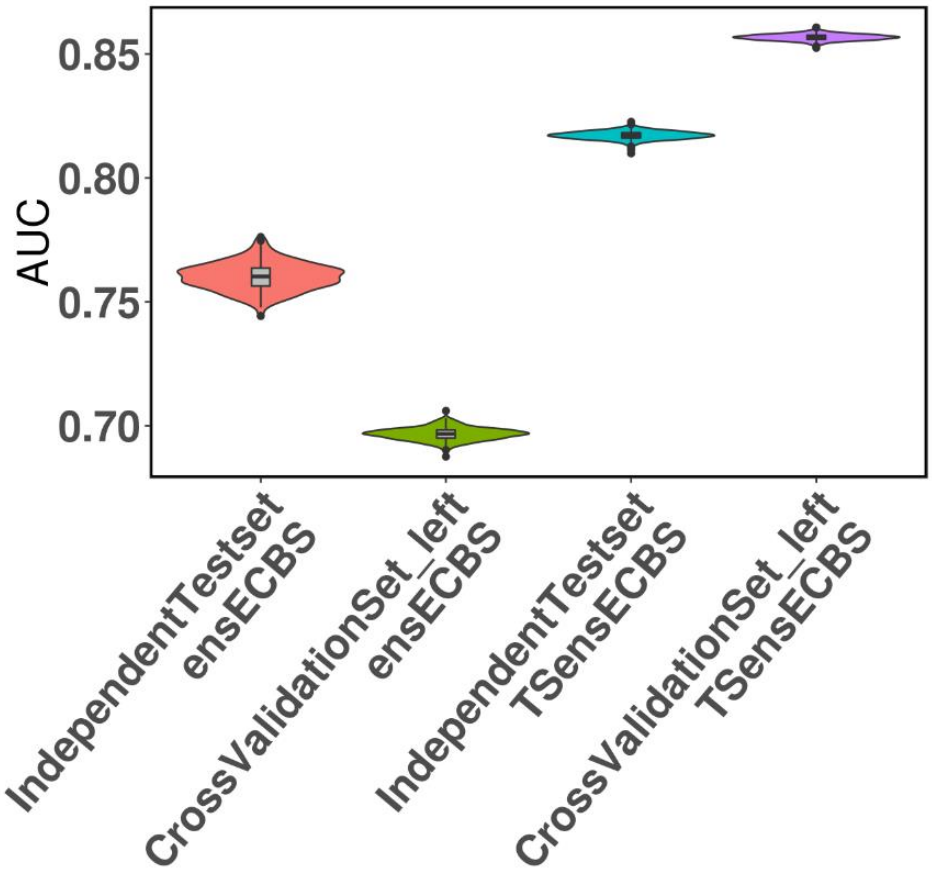
Keunwan Park[1,*], Young-Joon Ko[1], Prasannavenkatesh Durai[1], Cheol-Ho Pan[1]

[1] Natural Product Informatics Research Center, KIST Gangneung Institute of Natural Products, Gangneung, 25451, Republic of Korea

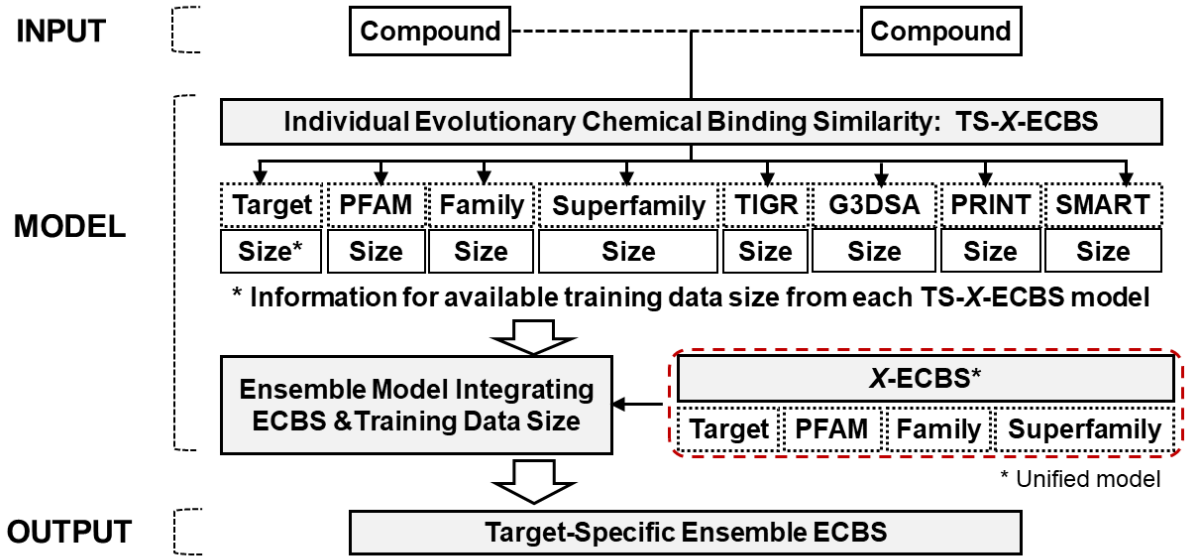
* To whom correspondence should be addressed. Tel: [+81-33-650-3663]; Fax: [+82-33-650-3629]; Email: [keunwan@kist.re.kr]

Contents:

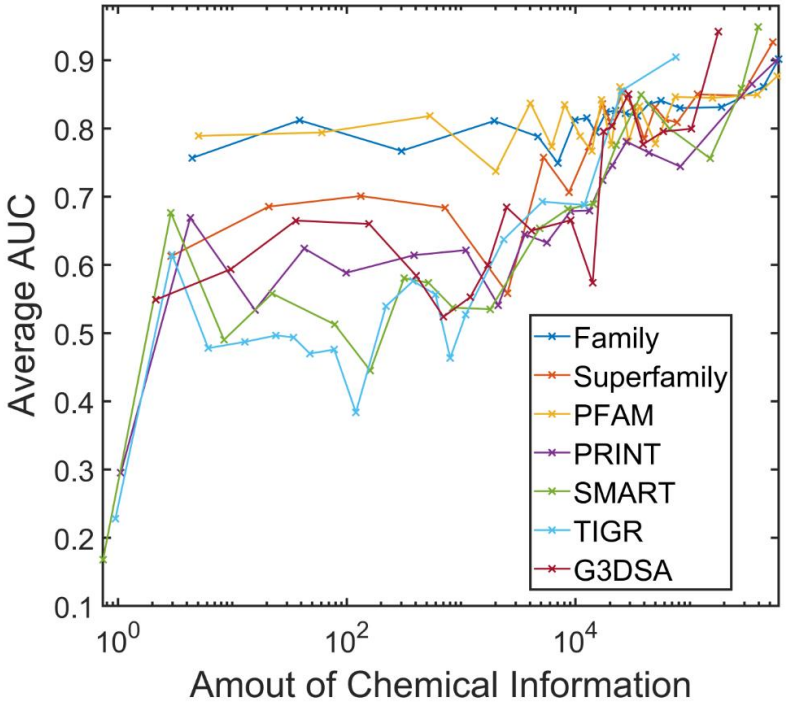
Supplementary Figures S1-S6



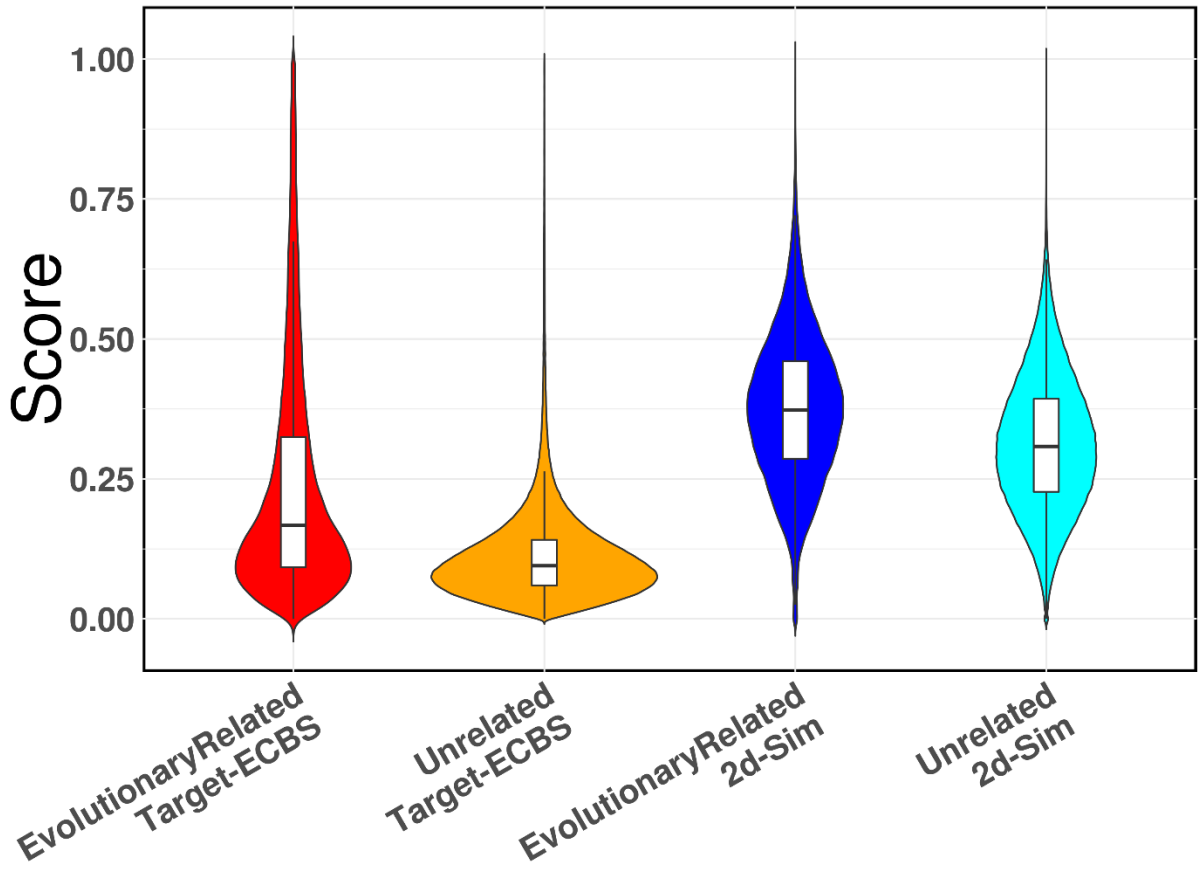
Supplementary Figure S1. Performance stability of the unified and target-specific ensemble model (ensECBS and TS-ensECBS) is shown for the cross-validation set and the independent set, respectively. The model evaluation procedure is repeated 100 times because it is based on the random selection of test set (1:11 where 1 corresponds to training set and 11 to test set). All the test results (AUC values) in the evaluation procedure are combined and summarized by the violin plot. Details for the model evaluation procedure is described in the Material & Methods section.



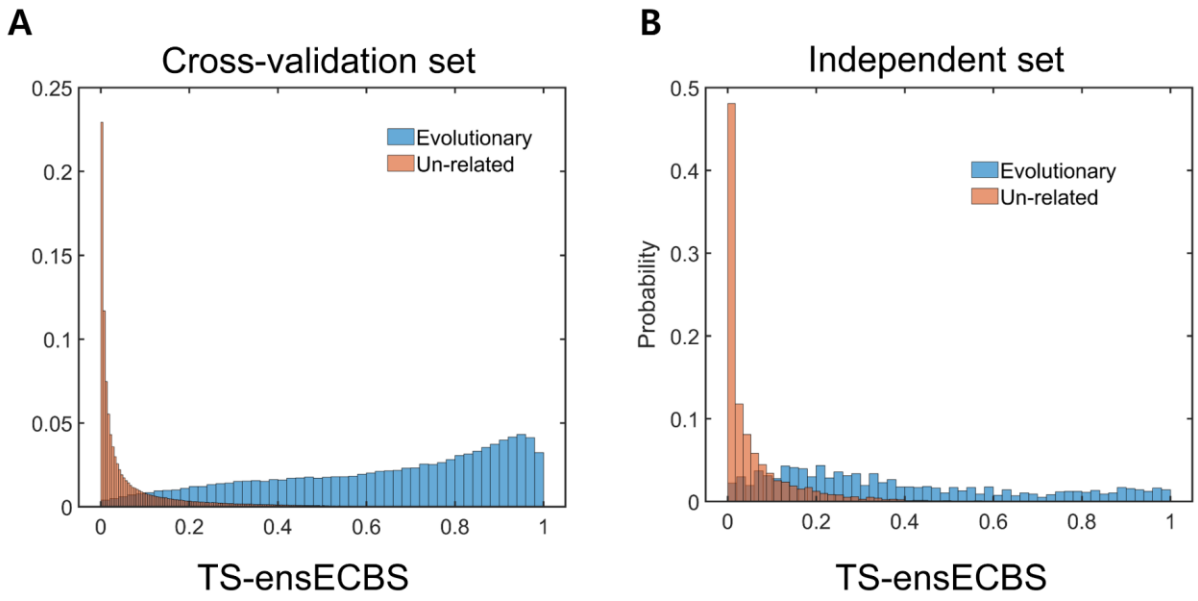
Supplementary Figure S2. Schematic overview of the TS-ensECBS model structure. The size information for training data of each evolutionary annotation and the scores from the unified *X*-ECBS models (shown in the red box) are additionally included in the ensemble procedure.



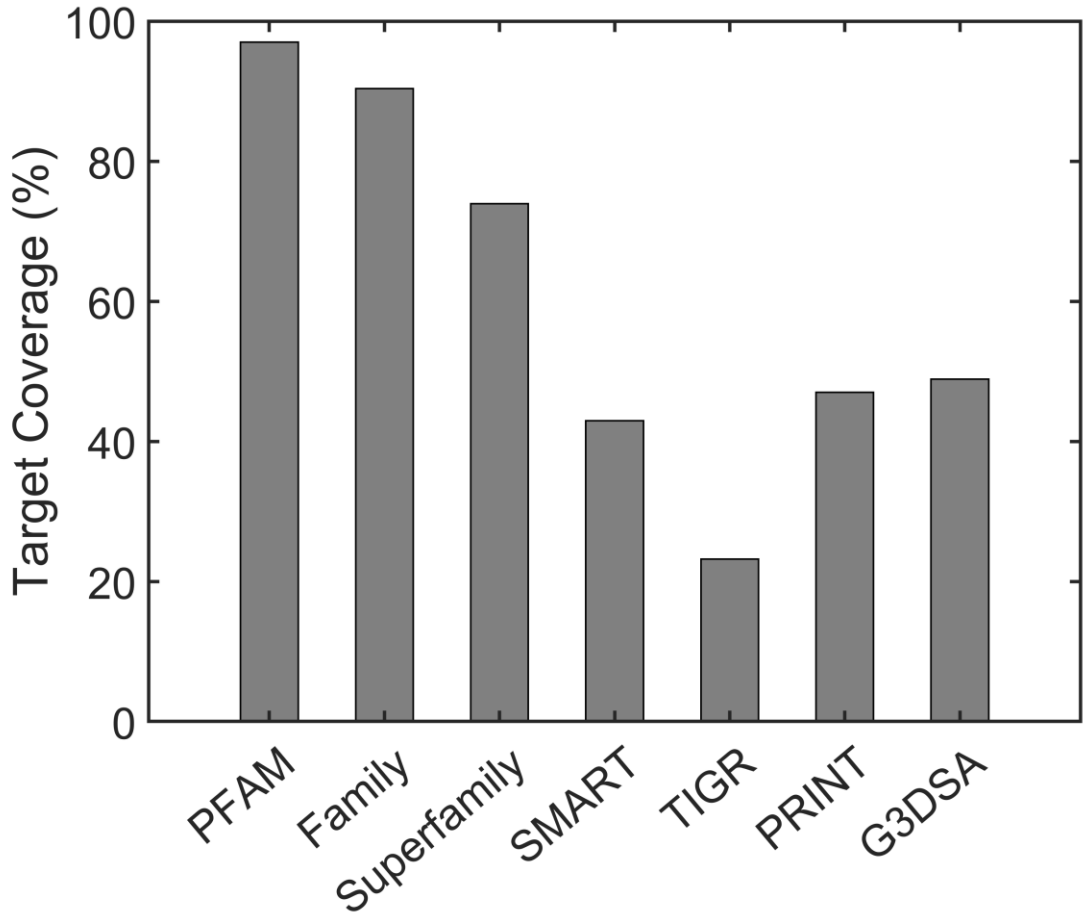
Supplementary Figure S3. The AUC values by the TS-*X*-ECBS models for each target are averaged and plotted according to the amount of chemical pair information in the training dataset. The test results for the cross-validation set are used to calculate the AUC values. The clear positive correlations suggest that inclusion of more chemical binding data in the training set increases the prediction accuracy.



Supplementary Figure S4. All pairwise chemical similarity scores between the drugs not binding to common targets. The score distributions by Target-ECBS and 2-D chemical structure similarity are shown separately for the ERCPs (whose targets are not identical but evolutionarily related) and unrelated chemical pairs. It is assumed that the ERCPs are potential target-binding candidates even though it is not experimentally validated for now. The average scores by Target-ECBS for ERCPs and unrelated drug pairs were 0.24 and 0.11, respectively, and the corresponding average scores by 2-D structure similarity were 0.38 and 0.31.



Supplementary Figure S5. Score distributions for evolutionarily related chemical pairs (ERCPs) and unrelated chemical pairs are compared by the TS-ensECBS model. The predicted scores from (A) cross-validation set and (B) independent set are used to check the cut-off values for a clear separation of ERCPs and unrelated chemical pairs. For both test sets, the ECBS scores above 0.5 represent high evolutionary relatedness of chemical compounds.



Supplementary Figure S6. The target coverage percentage (7,774 targets) is shown for each evolutionary annotation. It shows that 74% of the targets are annotated by Superfamily information and PFAM, Family annotated more than 90% of targets. However, SMART, TIGRFAM, PRINT, and Gene3D showed less than 50% target annotation coverages.