Text S3. Interface Prediction Performance Evaluation

To evaluate the performance of our interface residue predictor PS-HomPPI, we used measures including Matthews correlation coefficient (CC), F1 score (also called F-measure), Sensitivity (recall), and Specificity (precision) as discussed in [73, 74]. The Sensitivity is the proportion of all true interface residues that are correctly predicted as interface residue. The Specificity is the proportion of all predicted interface residues that are in fact interface residues. The CC is a measure of how predictions correlate with true interface residues and non-interface residues. CC ranges from -1 to 1. When predictions match actual interface and non-interface labels perfectly, CC is 1. When predictions totally disagree with actual labels, CC is -1. Random predictions yield a CC of 0. F1 score is a weighted average of Sensitivity and Specificity. F1 gives 1 and 0 for perfect and all-false predictions, respectively.

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TP}{TP+FP}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}}$$

$$F1 = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity}$$

where TP stands for true positive (i.e., the number of correctly predicted interface residues), TN stands for true negative (i.e., the number of correctly predicted non-interface residues) , FP stands for false positive (i.e., the number of falsely predicted interface residues), and FN stands for false negative (i.e., the number of falsely predicted non-interface residues).
These measures describe different aspects of interface prediction performance. Often it is possible to trade off one performance measure (e.g., Specificity) against another (e.g., Sensitivity) by varying the score cutoff that is used to transform prediction score to binary predictions (0 for non-interface residues, and 1s for interface residues). Compared with the rest of the measurements, CC and F1 score are more balanced measurements that measure the overall performance of a binary predictor. Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) is not an appropriate measurement to evaluate a highly unbalanced dataset (the number of interface residues is highly outnumbered by non-interface residues). A trivial interface predictor that always predicts a residue as non-interface residue can have a high Accuracy (dominated by high TN, low FN, and zero FP and TP). Therefore, we choose not to report Accuracy in this study.