Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrated Structure-based Protein Interface Prediction

M. Walder
Yeshiva University
E. Edelstein
Yeshiva University
M. Carroll
Yeshiva University
S. Lazarev
Yeshiva University
J.E. Fajardo
Albert Einstein College of Medicine
A. Fiser
Albert Einstein College of Medicine
R. Viswanathan (raji@yu.edu)
Yeshiva University

Research Article

Keywords: Protein-protein interaction, interface prediction, structure-based method

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1551577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: Identifying protein interfaces can inform how proteins interact with their binding partners, uncover the regulatory mechanisms that control biological functions and guide the development of novel therapeutic agents. A variety of computational approaches have been developed for predicting a protein's interfacial residues from its known sequence and structure. Methods using the known three- dimensional structures of proteins can be template-based or template-free. Template-based methods have limited success in predicting interfaces when homologues with known complex structures are not available to use as templates. The prediction performance of template-free methods that only rely only upon proteins' intrinsic properties is limited by the amount of biologically relevant features that can be included in an interface prediction model.

Results: We describe the development of an integrated method, ISPIP, to explore the hypothesis that the efficacy of a computational prediction method of protein binding sites can be enhanced by using a combination of methods that rely on orthogonal structure-based properties of a query protein, combining and balancing both template-free and template-based features. ISPIP is a method that integrates these approaches through simple linear or logistic regression models and more complex decision tree models. On a diverse test set of 156 query proteins, ISPIP outperforms each of its individual classifiers in identifying protein binding interfaces.

Conclusions: The integrated method captures the best performance of individual classifiers and delivers an improved interface prediction. The method is robust and performs well even when one of the individual classifiers performs poorly on a particular query protein. This work demonstrates that integrating orthogonal methods that depend on different structural properties of proteins performs better at interface prediction than any individual classifier alone.

Background

Proteins execute their diverse range of biological functions through interactions with other proteins and small molecules, which lead to the formation of larger scale protein interaction networks (interactomes). A protein's function in the interactome is characterized by its interaction partners. Knowledge of a protein's binding interfacial residues is essential for elucidating the molecular mechanism by which it performs its function, for determining the functional effect of mutations, as well as for designing drugs to disrupt a biological network by targeting a specific protein-protein interaction (PPI) [1].

Experimental techniques commonly employed to determine the structure of protein complexes at atomic-scale resolution include X-ray crystallography [2, 3] nuclear magnetic resonance (NMR) spectroscopy [4], and cryo-electron microscopy (cryo-EM) [5]. Information about interface residues can also be obtained by alanine scanning mutagenesis experiments [6, 7] or various footprinting experiments, such as hydrogen/deuterium exchange or hydroxy radical footprinting [8]. Since X-ray crystallography requires crystallization of specimens, it can only be used to analyze non-dynamic complexes and often under non-

physiological conditions. While NMR does not require samples to be crystallized it is limited to determining the structure of smaller proteins with molecular weight around 20 kDa. Cryo-EM allows the structure of proteins to be visualized while they are in an aqueous environment, which resembles their native intracellular environment. However, cryo-EM experiments also require cryogenic temperatures, usually lower than − 135^{O}C, to maintain the sample in a vitrified state. More importantly, all these approaches require a prior knowledge of a cognate binding partner. Due to the limitations, low-throughput, and costly nature of experimental approaches, computational prediction methods are employed to streamline the process of identifying the interfacial residues of proteins.

Prediction methods can rely solely upon query proteins' sequence information (sequence-based), or they can also be based on query proteins' 3-dimensional structure (structure-based). Sequence-based methods can be implemented on almost any protein, whereas structure-based approaches are limited to proteins with known structures in the Protein Data Bank[9]. Sequence-based methods are based on finding relationships between the likelihood of a residue to be interfacial and its sequence-related properties like hydrophobicity distribution, interface propensity, and physico-chemical properties [10, 11]. In a typical sequence-based method, overlapping sequence segments of the query protein are obtained by using a sliding window of width ranging from 3 to 30 residues [12] with target residue at the center of these segments. Each segment is assigned a feature vector based on properties of amino acids. These feature vectors from a set of proteins with known interface residues are used to train machine learning algorithms like random forest [13] or support vector machine [13–19]. The trained models are then used in a binary classification problem to predict the interfacial residues of each query protein using its feature vectors as inputs.

Structure-based approaches depend upon the availability and quality of 3D structures, and most of these methods outperform sequence-based methods [20]. There are two main classes of structure-based methods, which are referred to here as "template-free" or "template-based" approaches. Template-free methods train machine learning algorithms on a dataset of experimentally determined protein complex structures to create a model that relates sequence and structural features with the likelihood for residues to be at the binding interface. These template-free methods may include sequence features such as hydrophobicity, propensity of amino acids to be at an interface, physico-chemical properties, evolutionary conservation, and structural features such as secondary structure, solvent-accessible surface area, and geometric shape [10, 14, 21, 22]. While template-free methods have been steadily enhanced over the past 20 years, their future improvement appears to be limited because further combination of existing features and classifiers has little impact on performance [10, 23]. In contrast, template-based approaches predict interfacial residues by mapping interface information onto the query protein from its homologues or structural neighbors with known complex structures [1]. The drawback of template-based methods is that their effectiveness is dependent upon the existence of homologues or structural neighbors that have had their complex structure experimentally determined [10].

Methods that require the structure of both proteins in a complex to make a prediction are called partner-specific, and methods that can make interface predictions on individual unbound proteins are referred to

as partner-independent. Some template-free methods, like ISPRED4 [24], are partner-independent, while other template-free approaches, like Daberdaku et al [25] 3D Zernike descriptor method, are partner-specific. Currently there are several template-based methods that depend on known structural neighbors for predicting interfaces. Some of these methods, like PS-HomPPI [26], are partner-specific. Other template-based methods, like PredUs 2.0 [1] and PriSE [27], do not need information about the binding partner. In order to be more generic, we focused on methods that can make predictions of interface residues without the knowledge of the cognate partner protein's structure.

A few meta-methods that integrate different interface predictors to generate a consensus prediction have also been developed. Meta-PPISP is one such meta-method that combines the predictors cons-PPISP [28], Promate [29], and PINUP [30] through linear regression analysis [31]. The success of a meta-method is contingent on the input predictors contributing orthogonal information to the consensus model [10]. The inputs for meta-PPISP have limited orthogonality because it combines three template-free approaches, and it does not consider inputs from template-based or docking-based approaches. Additionally, meta-PPISP employed linear regression analysis for method combination, which is likely less robust than using more complex tree-based regression models.

Both classes of structure-based methods described above, template-free and template-based, have strengths and limitations. To take advantage of the successes of both these types of methods, we aimed to create a meta-method that integrates the orthogonal template-based, template-free, and docking-based predictors. Among the available template-based methods that are not partner-specific, we chose PredUs 2.0, as the webserver was readily available and could be automated on a large dataset. For a similar reason, we chose ISPRED4 [24] as the template-free method. We have recently shown that protein interfaces can be predicted effectively using a docking-based approach without knowledge of the binding partner [32], and we refer to this method as DockPred. Our goal is to improve the PPI binding interface predictions made by DockPred by integrating this method with two other orthologous approaches, PredUs 2.0 (template-based) [1] and ISPRED4 (template-free) [24].

The first version of PredUs, developed in 2011, makes interface predictions for a query protein based on the known binding interfaces of the query's structural neighbors. An improved version (PredUs 2.0) was developed in 2015 by adding sequence information to the template-based prediction. Using a Bayesian approach, PredUs 2.0 combines an amino acid interface propensity score with the template-based score of PredUs [33]. The original PredUs program uses the structural alignment program Ska [34] to identify a query protein's structural neighbors and a structural alignment score is calculated [35]. Structural neighbors with a sequence similarity larger than 40% are identified using cd-hit [36] and retained. For every structural neighbor retained, PredUs calculates a contact frequency for each residue in the query protein by relating the structural neighbor's binding partner to the query protein. This is then weighted by the closeness of the structural neighbor to the query protein. PredUs uses a support vector machine (SVM) algorithm to generate its template-based prediction score [1]. PredUs 2.0 includes information on the interface propensity values of the residues to calculate an interface probability score for each query residue.

ISPRED4 is one of the best performing template-free protein binding interface predictors currently available. It was developed by training an SVM model on a dataset (DBv5Sel) of 314 different monomer chains with complex structures that had been resolved by X-ray crystallography. Interface residues are defined as those that lost at least 1Å^2 of Accessible Surface Area (computed with the DSSP program [37]) when transitioning from a protein's unbound to complex form. In the SVM model, each of the training proteins' surface residues are represented by a 46-dimensional feature vector consisting of 10 different groups of descriptors. The feature vector included 34 sequence-based features that formed 5 groups of descriptors that included evolutionary information. The feature vector also included 12 structure-based features that comprised 5 groups of descriptors. ISPRED4 combines its SVM model with a Grammatical-Restrained Hidden Conditional Random Field (GRHCRF) to account for possible correlations between neighboring surface residues. For a given query protein, ISPRED4 calculates interface prediction scores by plugging the query residues' feature vectors into its trained SVM/GRHCRF model [24].

DockPred demonstrated our previous hypothesis [38] that both substrate and non-substrate small organic molecules have a tendency to bind to similar, energetically favorable sites on a target protein ("sticky" sites) regardless of their biological relevance, also applies to the binding of proteins. The query protein is docked on 13 different non-cognate partner proteins that vary in size and represent different protein folds (immunoglobulin, and other small protein folds). The success of DockPred showed that non-cognate protein ligands preferentially bind to the cognate binding site of a target protein [32]. DockPred generates 2000 docked poses for each of the 13 binding partners using ZDOCK [39] or GRAMM [40]. The query protein residues are each assigned a probability to be at an interface by taking the average number of times a residue appears at the interface of 2000 docked poses for each of 13 different binding partners. A residue is considered to be at the interface of a docked pose if any atom of this residue is within 4.0 A of any atom of the binding partner and if the contact was considered legitimate according to the CSU program [41].

In this work, we present an **I**ntegrated **S**tructure-based **P**rotein **I**nterface **P**rediction (ISPIP) method that generates an enhanced consensus prediction by integrating the predictive strengths of orthogonal template-based (PredUs 2.0), template-free (ISPRED4), and docking-based (DockPred) predictors. To develop ISPIP, regression models of varying complexity were trained on the three input classifiers' interface scores for a training set of query proteins with known complex structures. Not only is ISPIP's consensus predictor significantly enhanced relative to DockPred and the other input predictors, it also outperforms a previous consensus predictor (meta-PPISP) and a complex structure-based method (VORFFIP).

Results

Enhanced interface prediction of ISPIP Model

When designing ISPIP, we aimed to develop a model that could predict query proteins' interfacial residues more effectively than the three orthogonal input predictors (DockPred, ISPRED4, PredUs2.0) alone. The

linear and logistic regression models that were initially constructed yielded enhanced interface classification of test set query proteins according to both single-threshold (Table 1 and threshold-free metrics (Fig. 1). The threshold employed for single-threshold evaluation was determined by a dynamic cutoff (see Methods section). In predicting the interfacial residues for the test set proteins from Set A, the logistic regression model generated an average F-score of 0.469, which is significantly higher than the F-scores generated by the input methods that range from 0.380 to 0.405. The one sample Kolmogorov-Smirnov test[42] showed that the F-scores from the different input methods and the integrated methods were not normally distributed. The two sample Kolmogorov-Smirnov test showed that the F-socre distributions obtained by the individual classifiers relative to the integrated method are statistically significantly different with > 95% confidence. All these two sample Kolmogorov-Smirnov tests resulting in a p-value < 0.05 show that the improvements in F-scores and MCC scores of ISPIP methods over the individual classifiers are statistically significant. MCC is generally considered to be more informative than F-scores because it captures model performance on both the positive and negative classes, whereas F-score is the harmonic mean between precision and recall, which are metrics that relate to the positive class. The linear and logistic ISPIP models yielded MCC values of 0.433, which is statistically significantly higher than the MCC values generated by the input methods, ranging from 0.324 to 0.355.

Table 1
ISPIP predictive enhancement by single-threshold metrics

Classifier	Average F-score	Average MCC
PredUs 2.0	0.400	0.351
ISPRED4	0.405	0.355
DockPred	0.380	0.324
Linear Regression	0.470	0.433
Logistic Regression	0.469	0.433

ISPIP's enhanced predictive capacity is also illustrated in the PR (Fig. 1B) and ROC (**Figure S1**) curves, and in their corresponding AUC metrics. The linear model had a PR-AUC of 0.441, which is significantly greater than the PR-AUC of DockPred, ISPRED4, and PredUs 2.0 (0.330, 0.346, and 0.349 respectively). Similar improved prediction results were obtained for Set B proteins shorter than 450 residues (**Figure S2).** A small percentage of a protein's residues appear at the interface, which makes interface prediction an imbalanced classification problem. PR curves are shown here because they are more informative than ROC curves for imbalanced datasets [43]. Nevertheless, the Supplementary Information (**Figures S1 and S2**) show the improvement of ROC-AUC from the input methods (0.726 to 0.823) to the ISPIP linear model (0.881). The statistical significance of ISPIP's enhanced ROC-AUC was confirmed using the STAR approach[44] **(Table S1)**.

Evolution of ISPIP model

After establishing the efficacy of combining orthogonal methods through simple regression models, we sought to further enhance ISPIP's predictive capability by using more complex decision tree algorithms, Random Forest (RF) and Gradient Boosted Trees (XGBOOST). After optimizing the parameters through a 5-fold Cross Validation process, the trained RF and XGBOOST models were implemented to predict the interface likelihood of query protein residues in the test sets from Set A and B. For Set A, the RF prediction achieved an MCC of 0.458 (Table 2) and PR-AUC of 0.476 (Fig. 1B), while the XGBOOST prediction achieved an MCC of 0.487 and PR-AUC of 0.516. A similar improvement was also observed in the average F-scores (Table 2) and AUC-ROC (**Figure S3**). The pattern of ISPIP's predictive capacity being initially enhanced from the regression models to RF, and further improved from RF to XGBOOST, is observed in both single-threshold and threshold-free metrics.

Table 2
Increased performance with ISPIP model evolution

ISPIP Model	Average F-score	Average MCC
Random Forest	0.490 ± .164	0.458 ± .169
XGBoost	0.516 ± .171	0.487 ± .176

Integration of orthogonal predictions to form a consensus prediction

Triosephosphate isomerase (1YPI.A) from *S. cerevisiae* is a query protein that illustrates how ISPIP integrates its three input methods to formulate an enhanced interface prediction. Out of the 23 annotated interfacial residues, PredUs 2.0 predicted 12 of them correctly (MCC = 0.402), DockPred predicted 13 of them correctly (MCC = 0.446), and ISPRED4 predicted 15 of them correctly (MCC = 0.523). The 3 predicted interfaces have 7 overlapping true-positive residues that are all present in the ISPIP interface prediction (Fig. 2). However, ISPIP is also able to retain additional, non-overlapping residues in its positive class to correctly predict a total of 19 of the 23 interfacial residues (MCC = 0.705).

Set A vs B: Including larger proteins in the dataset

Sulfite oxidase (1SOX.A) and *acetohydroxy acid isomeroreductase* (1YVE.I) were the two proteins larger than 450 residues that were added to Set B's test set to form Set A's test set. The optimized parameters for the 4 ISPIP regression models trained on Set B are included in the supplementary information **(Table S2)**. Only 1YVE.I was a significant outlier, as DockPred's prediction generated a negative MCC score (-0.061). The outlier slightly lowered the average MCC of DockPred from 0.336 in Set B to 0.324 in Set A, as well as ISPIP's average MCC from 0.495 to 0.487 (Table 3). Since Set A includes proteins of all size*s* and it has MCC scores very close to Set B scores, Set A results have been reported as the default in this paper.

Table 3
Set A vs Set B model performance

	Set A: Average MCC	Set B: Average MCC
DockPred	0.324 ± .219	0.336 ± .215
ISPIP XGBoost	0.487 ± .176	0.495 ± .170

Discussion

Linear vs. logistic regression

In general, linear regression models are suitable for cases where the target variable is continuous, and logistic models are appropriate for instances where the target variable is categorical. Since our target variable indicated whether a residue was experimentally determined to be at the interface or not (1 or 0), we expected the logistic model to have a greater predictive capacity than the linear model. Surprisingly, the logistic and linear models had almost identical F-scores (0.469 and 0.470, respectively) and PR-AUC values (0.437 and 0.441, respectively). This similarity in performance may indicate that three input variables are not enough to observe the difference in linear and logistic models, or perhaps it suggests that that the interfacial scores should be treated rather as a continuous variable in future studies.

Performance comparison to VORFFIP and meta-PPISP

We assessed the performance of ISPIP on the test set query proteins (Set A) relative to a top-performing structure-based method (VORFFIP) and the most recently available metamethod (meta-PPISP). Both methods base their predictions on the structure of the query protein and do not require structural information on a cognate partner. For this reason, as well as the availability and easy accessibility of their webservers, we chose to compare our results from ISPIP to these two methods. VORFFIP, developed by Segura et al. [45], uses a random forest method to integrate heterogeneous data including various residue level structural and energetic features, evolutionary sequence conservation, and crystallographic B-factor. Meta-PPISP is a metamethod that integrates the structure-based approaches cons-PPISP, Promate, and PINUP through linear regression. The most significant contrast in performance can be seen in the PR curves (Fig. 3), where ISPIP has as PR-AUC of 0.516 relative to VORFFIP's score of 0.313 and meta-PPIPSP's score of 0.293. Single-threshold metrics also confirm ISPIP's (MCC = 0.487) superior predictive capacity relative to meta-PPISP (0.295) and VORFFIP (0.301).

Driving factor of ISPIP's enhanced performance

One primary question about ISPIP was whether its enhanced performance was due to the best performing input method (ISPRED4) or a result of the classifier integration process. To investigate this, consensus models were trained on only 2 of the 3 input predictors on the Set A proteins. The PR-AUC of DockPred-PredUs 2.0 model was 0.421 (**Figure S4**), which was significantly greater than any of the 3 individual classifiers (PR-AUC = [0.330, 0.346]). This shows that the integration process itself, without the presence

of ISPRED4, already generates an enhanced predictor. Replacing PredUs 2.0 with ISPRED4 does improve the DockPred-ISPRED model (PR-AUC = 0.448), so the identity of the input predictors does have some effect on the final model. However, the main driver of ISPIP's enhanced prediction is the integration process, which can be seen when all 3 classifiers are combined to generate the complete ISPIP model with PR-AUC = 0.516.

Robustness of ISPIP

Nitrogenous iron protein (1CP2.A) from *C. pasteurianum* is a query protein that illustrates how ISPIP's predictive model is robust to one of its input methods performing poorly (Fig. 4). Out of the 13 annotated interfacial residues, DockPred predicted 11 of them correctly (MCC = 0.536) and ISPRED4 predicted 10 of them correctly (MCC = 0.481); however, PredUs 2.0 was only able to predict 4 interfacial residues correctly (MCC = 0.190). It is very likely that the underperformance of PredUs 2.0 is due to a dearth of structural neighbors with an experimentally determined complex structure for 1CP2.A. ISPIP is able to integrate the input predictors in a robust manner to correctly predict 10 of the 13 interfacial residues, despite the poor performance of PredUs 2.0.

Conclusions

There are currently several structure-based methods for predicting interfaces of proteins that can be categorized as either template-based or template-free. The performance of template-based methods is limited by the availability of template complex structures, and template-free approaches are limited by the number of biologically relevant features that can be included in the model. We show that these performance limitations can be overcome by using a method that integrates the orthogonal properties of a template-based, template-free, and docking-based approaches for consensus prediction of protein interfaces by using effective machine learning algorithms can improve performance. A dataset of 156 proteins chosen from the docking benchmark and NOX benchmark with less than 30% sequence similarity, ranging in size from < 100 residues to about 600 residues and representing various CATH superfamilies were used for training and testing ISPIP. We demonstrate that more complex machine learning algorithms like random forest and gradient boosted trees perform better than the simpler linear or logistic regression models. All of these integrated models perform better than the best performing individual classifier. The integrated method is robust even when one of the individual classifiers performs poorly on a query protein. Since it is often not possible to predict which individual classifier will perform better on a given query protein, using an integrated approach for interface prediction should have a greater chance of success at predicting protein interfaces.

Methods

Development of ISPIP

In our previous work [32], we reported a docking-based interface classifier that is referred to here as DockPred. To enhance DockPred's predictive capability, we have developed a meta-classification model that integrates DockPred with orthogonal template-free (ISPRED4) and template-based (PredUs2.0) classifiers. This was accomplished using the three methods as inputs to train various regression models on the training set described below, and then using the trained models with optimized parameters to classify interface residues for the query proteins in the test set (Fig. 5).

Dataset

The dataset originally employed to test DockPred's performance consisted of 233 unbound protein structures from the Docking Benchmark version 5 [46] and NOX [47] databases. Each protein had an unbound structure, as well as a corresponding complex structure available from the Protein Data Bank (PDB) [9]. The CATH superfamily classification exists for 91 of the proteins in this dataset. The major CATH superfamilies represented are Immunoglobulin like (33 proteins), Rossman fold (20 proteins). The other superfamilies like TIM Barrel, four helix, OB fold and Jelly roll are also represented. The true interface residues, or annotated residues, were determined by the CSU program [41] to find the legitimate contacts between query proteins and their complex partner. Residues of interacting proteins were classified as part of the interface if at least one complementary atom contact was detected by CSU within 4.0 Å of the partner protein, and if the contact was legitimate as defined by Sobolev, et al.

Training and Test Sets

To eliminate redundancy of query proteins in the test and training sets, 156 proteins with less than 30% sequence similarity to all other proteins, as determined by Clustal Omega [48], were retained. The dataset contained proteins of sequence length ranging from 50 residues to 800 residues. Due to the challenges of docking larger proteins, DockPred's predictive capacity significantly declined for proteins larger than 450 residues. Hence, we split the dataset into two subsets. One subset (set A) contained all 156 proteins, while the other subset (set B) consisted of the 141 proteins with fewer than 450 residues. 31 of the proteins from set B were randomly assigned to the test set, and the remaining 110 were split into 5 cross-validation (CV) subsets of 22 proteins each. The 15 proteins that contain more than 450 residues were randomly added to these test and training sets to form the set A training and test sets. The CATH superfamilies described above are well represented by both training and test sets.

Machine Learning Algorithms

A. Regression Models

Using the three different classifiers, DOCKPRED, PredUS2.0 and ISPRED4, a normalized score between 0 and 1 was calculated for each residue for every protein in the training sets. This score represented the likelihood of a residue to be at the binding interface, as determined by each of the three classifiers. The three interface likelihood scores served as input variables (x_1, x_2, x_3) for the linear and logistic regression models. The annotated score for each residue (0 = non-interfacial, 1 = interfacial) served as the

target variable for the regression models. The logistic model parameters (b_1, b_2, b_3) were fit by maximum likelihood estimation according to the function:

$$P\left(i|\, x_1, x_2, x_3\right) = \frac{1}{1 + e^{-\left(b_0 + \sum_{j=1}^{3} b_j x_j\right)}} \quad (1)$$

Where P(i|x_1,x_2, x_3) is the probability that a residue, given the values of x_j, will be in one of two discrete categories (interfacial or non-interfacial).

The linear model parameters (b_1, b_2, b_3) were fit by ordinary least squares according to the function:

$$I\left(x_1, x_2, x_3\right) = b_0 + \sum_{j=1}^{3} b_j x_j \quad (2)$$

Where $I\left(x_1, x_2, x_3\right)$ is the continuous interfacial likelihood between 0 and 1. probability that a residue, given the values of x_j, will be in one of two discrete categories (interfacial or non-interfacial).

The regression parameters (b_1, b_2, b_3) were optimized using a 5-fold (CV) scheme. Of the 5 sets in the training set each with 22 proteins (Set B) or 24 proteins (Set A), 4 sets were used for training and the fifth set served as the test set for the CV process. The model parameters were determined as the ones that generated the highest average F-score on the test CV subset (Table 4**).** These optimized parameters were then used to generate the interface probabilities for the proteins in the test set.

Table 4
Optimized regression parameters for Set A proteins

Regression Model	PredUs 2.0 $\left(b_1\right)$	ISPRED4 $\left(b_2\right)$	DockPred $\left(b_3\right)$
Linear	0.196	0.313	0.313
Logistic	1.28	2.821	1.424

B. Ensemble Decision Tree Algorithms: Random Forest and XGBOOST

The first ensemble decision-tree model that integrated the interface likelihoods from the three input classifiers was the random forest (RF) algorithm [49, 50]. At each node of the tree in the RF model, the classifiers and the cutoff values, for each classifier and each level of the tree, were chosen to optimize the results. The parameters representing the ensemble of trees in the forest, the maximum number of levels for each tree, and a tree pruning parameter, α, which chooses the subtree that minimizes the cost complexity measure, were all optimized to find the best fitting model with the three classifiers. The values for the optimized parameters are shown in **Table S2** and these were optimized to yield the best values for the average F-score, calculated as described below. For all the RF calculations, the optimal values of 100 trees, 10 levels and a pruning parameter of zero were used. Once again, a five-fold CV was used to obtain

the optimized parameters. Once the trees are trained using the training set, the random forest model classifies the residues in the test set to one of the terminal nodes (leaves) in each tree in the forest. Based on the results from the training set, the probability of being an interface residue is calculated for each terminal node in each tree. A similar probability value was calculated for each terminal node for every tree in the random forest. Finally, the probability of being an interface residue was calculated as the average probability of all the terminal nodes in the forest into which the test residue is classified.

The other tree-based model employed for ISPIP involved gradient boosting. Like RF, gradient boosting constructs an ensemble of decision trees to generate an interface probability score. However, unlike RF, each successive tree produced by the gradient boosted algorithm “learns” from the previous trees in the forest by addition of a loss function and regularization parameter. Specifically, Histogram-based Gradient Boosting (XGBoost) was used to improve the speed and accuracy of the Decision Tree based regressor. A five-fold CV procedure was used to determine the optimal loss-function by maximizing the average F-score. The optimized parameters **(Table S2)** were used to generate the interface probabilities for every residue for proteins in the test set as was done with the other models.

Single-threshold evaluation of interface classifiers

The ISPIP method using regression, RF, and XGBOOST models generates interface likelihood scores (*p)*, between 0 and 1 for every residue of the proteins in the test set. To implement ISPIP as a classification model, each residue needs to be designated as interfacial (positive class) or non-interfacial (negative class). This can be based on a threshold value (p_{thr}) such that residues with $p > p_{thr}$ can be classified as interface residues. Another approach would be to choose a set number, N, of top-ranking residues (based on the *p* value) for every protein in the test set. This approach chooses a single threshold value, *N*, for each query protein. Zhang *et al.* [1] proposed a dynamic cutoff to determine N for each query protein according to the following equation:

$$N = 6.1R^{0.3} \quad (3)$$

where R referred to the number of the protein’s surface-exposed residues. Using this threshold value, the elements of the Confusion matrix, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) can be determined. The binary classifier evaluation metrics used in this work are shown in Table 5. We used the nonparametric Kolmogorov-Smirnov (KS) single sample test to determine if the F-scores were normally distributed. For non-normal distributions, we used the two sample KS test to determine if the null hypothesis, that the distributions from the individual classifiers and the integrated method are the same, is valid.

Table 5
Binary classifier evaluation metrics

Precision = $\frac{TP}{TP+FP}$
Recall = True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
False Positive Rate (TPR) = $\frac{FP}{FP+TN}$
F-Score = $\frac{2*Precision*Recall}{Precision+Recall}$
MCC = $\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Threshold Free Evaluation Metrics - AUC under ROC and PR Curves:

Receiver Operator Characteristic (ROC) curves were generated using python's scikit package[49]by plotting the true positive rate (TPR) vs the false positive rate (FPR) for different threshold values of *p*, ranging from 0 to 1. Precision-recall (PR) curves were generated with the same package by plotting the precision vs recall for different threshold values of *p*, ranging from 0 to 1. The area under the curve (AUC-ROC and AUC-PR) was calculated using the trapezoidal method.

We assessed the statistical significance of the differences in AUC-ROC for the different methods using the STAR software[44] uses the chi-squared distribution to test the null hypothesis that there is no difference between the AUC-ROC curves originating from the different methods. The difference between any two methods is assessed at a significance level of 0.05.

Declarations

Ethics Approval and consent to participate:

Not Applicable

Consent for Publication

Not Applicable

Availability of Data and Materials

ISPIP is implemented in python and the code and sample data can be freely accessed from GitHub repository: https://github.com/eved1018/ISPIP.

Competing Interests

The authors declare that they have no competing interests.

Funding

Office of Extramural Research, National Institutes of Health, GM136357 and AI141816

Authors' Contributions

A.F, E.J.F.,R.V. and M.W wrote the main manuscript text.

S.L. and M.C. prepared Figure 4 and performed individual method calculations.

E.E. integrated the individual methods and uploaded the code to GitHub.

M.W. prepared all the tables and Figures.

All authors reviewed the manuscript.

Acknowledgment

Not applicable

References

1. Zhang, Q.C., et al., *PredUs: a web server for predicting protein interfaces using structural neighbors*. Nucleic Acids Res, 2011. **39**(Web Server issue): p. W283-7.
2. Kobe, B., et al., *Crystallography and protein-protein interactions: biological interfaces and crystal contacts*. Biochem Soc Trans, 2008. **36**(Pt 6): p. 1438–41.
3. Shi, Y., *A glimpse of structural biology through X-ray crystallography*. Cell, 2014. **159**(5): p. 995–1014.
4. O'Connell, M.R., R. Gamsjaeger, and J.P. Mackay, *The structural analysis of protein-protein interactions by NMR spectroscopy*. Proteomics, 2009. **9**(23): p. 5224–32.
5. Callaway, E., *The revolution will not be crystallized: a new method sweeps through structural biology*. Nature, 2015. **525**(7568): p. 172–4.
6. Morrison, K.L. and G.A. Weiss, *Combinatorial alanine-scanning*. Curr Opin Chem Biol, 2001. **5**(3): p. 302–7.
7. Simoes, I.C., et al., *New Parameters for Higher Accuracy in the Computation of Binding Free Energy Differences upon Alanine Scanning Mutagenesis on Protein-Protein Interfaces*. J Chem Inf Model, 2017. **57**(1): p. 60–72.
8. Li, K.S., et al., *Hydrogen-Deuterium Exchange and Hydroxyl Radical Footprinting for Mapping Hydrophobic Interactions of Human Bromodomain with a Small Molecule Inhibitor*. J Am Soc Mass Spectrom, 2019. **30**(12): p. 2795–2804.

9. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235–42.
10. Esmaielbeiki, R., et al., *Progress and challenges in predicting protein interfaces*. Brief Bioinform, 2016. **17**(1): p. 117–31.
11. Esmaielbeiki, R. and J.C. Nebel, *Scoring docking conformations using predicted protein interfaces*. BMC Bioinformatics, 2014. **15**: p. 171.
12. Maheshwari, S. and M. Brylinski, *Template-based identification of protein-protein interfaces using eFindSitePPI*. Methods, 2016. **93**: p. 64–71.
13. Sikic, M., S. Tomic, and K. Vlahovicek, *Prediction of protein-protein interaction sites in sequences and 3D structures by random forests*. PLoS Comput Biol, 2009. **5**(1): p. e1000278.
14. Yan, C., D. Dobbs, and V. Honavar, *A two-stage classifier for identification of protein-protein interface residues*. Bioinformatics, 2004. **20 Suppl 1**: p. i371-8.
15. Murakami, Y. and K. Mizuguchi, *Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites*. Bioinformatics, 2010. **26**(15): p. 1841–8.
16. Ahmad, S. and K. Mizuguchi, *Partner-aware prediction of interacting residues in protein-protein complexes from sequence data*. PLoS One, 2011. **6**(12): p. e29104.
17. Sriwastava, B.K., et al., *PPIcons: identification of protein-protein interaction sites in selected organisms*. J Mol Model, 2013. **19**(9): p. 4059–70.
18. Chen, X.W. and J.C. Jeong, *Sequence-based prediction of protein interaction sites with an integrative method*. Bioinformatics, 2009. **25**(5): p. 585–91.
19. Garcia-Garcia, J., et al., *iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments*. J Mol Biol, 2017. **429**(3): p. 382–389.
20. Xue, L.C., et al., *Computational prediction of protein interfaces: A review of data driven methods*. FEBS Lett, 2015. **589**(23): p. 3516–26.
21. Gallet, X., et al., *A fast method to predict protein interaction sites from sequences*. J Mol Biol, 2000. **302**(4): p. 917–26.
22. Ofran, Y. and B. Rost, *Predicted protein-protein interaction sites from local sequence information*. FEBS Lett, 2003. **544**(1–3): p. 236–9.
23. Gil, N. and A. Fiser, *The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis*. Bioinformatics, 2019. **35**(1): p. 12–19.
24. Savojardo, C., et al., *ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model*. Bioinformatics, 2017. **33**(11): p. 1656–1663.
25. Daberdaku, S. and C. Ferrari, *Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction*. BMC Bioinformatics, 2018. **19**(1): p. 35.
26. Xue, L.C., D. Dobbs, and V. Honavar, *HomPPI: a class of sequence homology based protein-protein interface prediction methods*. BMC Bioinformatics, 2011. **12**: p. 244.
27. Jordan, R.A., et al., *Predicting protein-protein interface residues using local surface structural similarity*. BMC Bioinformatics, 2012. **13**: p. 41.

28. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data*. Proteins, 2005. **61**(1): p. 21–35.
29. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites*. J Mol Biol, 2004. **338**(1): p. 181–99.
30. Liang, S., et al., *Protein binding site prediction using an empirical scoring function*. Nucleic Acids Res, 2006. **34**(13): p. 3698–707.
31. Qin, S. and H.X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction*. Bioinformatics, 2007. **23**(24): p. 3386–7.
32. Viswanathan, R., et al., *Protein-protein binding supersites*. PLoS Comput Biol, 2019. **15**(1): p. e1006704.
33. Hwang, H., D. Petrey, and B. Honig, *A hybrid method for protein-protein interface prediction*. Protein Sci, 2016. **25**(1): p. 159–65.
34. Petrey, D. and B. Honig, *GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences*. Methods Enzymol, 2003. **374**: p. 492–509.
35. Yang, A.S. and B. Honig, *An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance*. J Mol Biol, 2000. **301**(3): p. 665–78.
36. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*. Bioinformatics, 2006. **22**(13): p. 1658–9.
37. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers, 1983. **22**(12): p. 2577–637.
38. Hajduk, P.J., J.R. Huth, and S.W. Fesik, *Druggability indices for protein targets derived from NMR-based screening data*. J Med Chem, 2005. **48**(7): p. 2518–25.
39. Pierce, B.G., Y. Hourai, and Z. Weng, *Accelerating protein docking in ZDOCK using an advanced 3D convolution library*. PLoS One, 2011. **6**(9): p. e24657.
40. Vakser, I.A., *Main-chain complementarity in protein-protein recognition*. Protein Eng., 1996. **9**(9): p. 741–1049.
41. Sobolev, V., et al., *Automated analysis of interatomic contacts in proteins*. Bioinformatics, 1999. **15**(4): p. 327–32.
42. Riffenburgh, R.H., *Tests on the Distribution Shape of Continuous Data*, in *Statistics in Medicine (Second Edition)*. 2006, ScienceDirect. p. 369–386.
43. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. PLoS One, 2015. **10**(3): p. e0118432.
44. Vergara, I.A., et al., *StAR: a simple tool for the statistical comparison of ROC curves*. BMC Bioinformatics, 2008. **9**: p. 265.
45. Segura, J., P.F. Jones, and N. Fernandez-Fuentes, *Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams*. BMC Bioinformatics, 2011. **12**: p. 352.

46. Vreven, T., et al., *Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2*. J Mol Biol, 2015. **427**(19): p. 3031–41.
47. Zhu, H., et al., *NOXclass: prediction of protein-protein interaction types*. BMC Bioinformatics, 2006. **7**: p. 27.
48. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019*. Nucleic Acids Res, 2019. **47**(W1): p. W636-W641.
49. Pedregosa, V., Gramfort et al, *Scikit-learn: Machine Learning in Python*, in *JMLR*. 2011. p. 2825–2830.
50. Breiman, L., *Machine Learning*. Vol. 45. 2001.

Figures

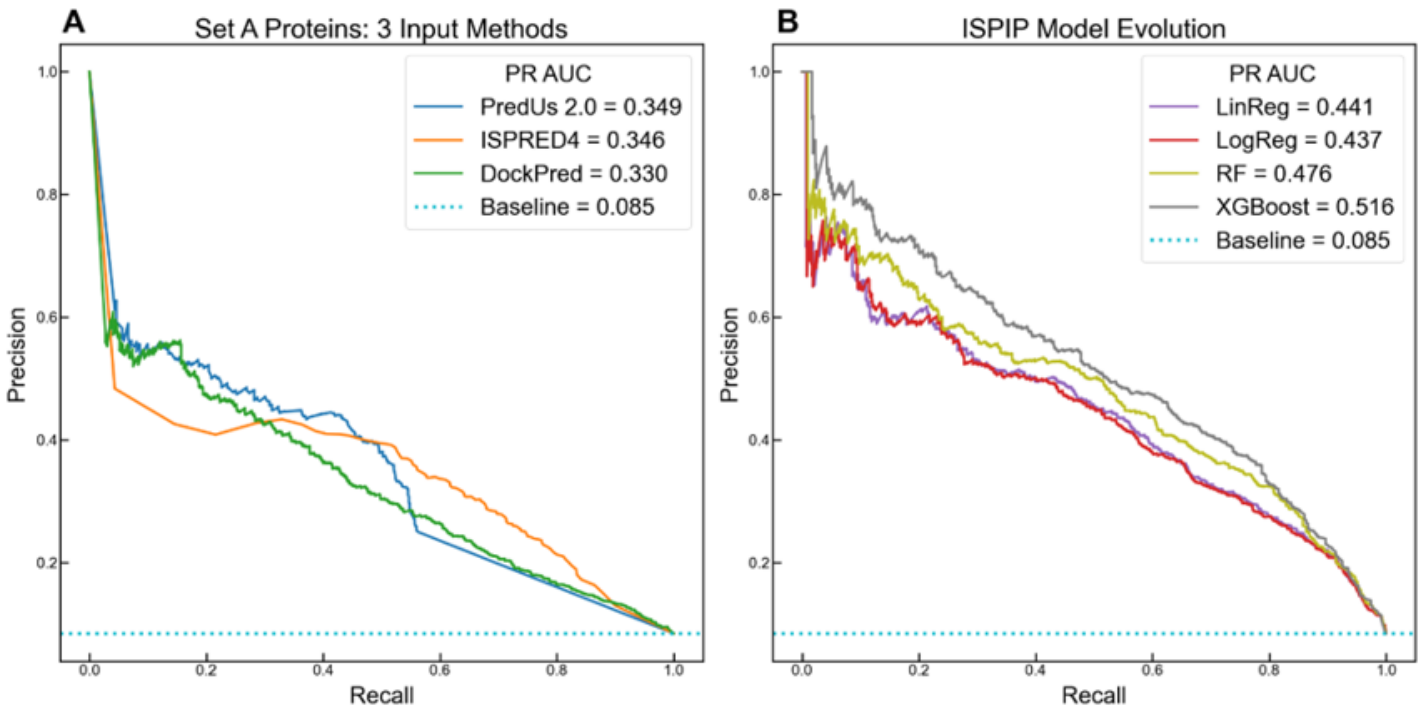

Figure 1

Enhanced prediction as ISPIP model evolves: (A) The PR curves of the 3 input methods indicate that PredUs 2.0 and ISPRED4 perform slightly better than DockPred. **(B)** All the ISPIP models significantly outperform the input predictors, and PR-AUC is boosted as the model evolves from simple linear regression to more complex ensemble decision tree algorithms.

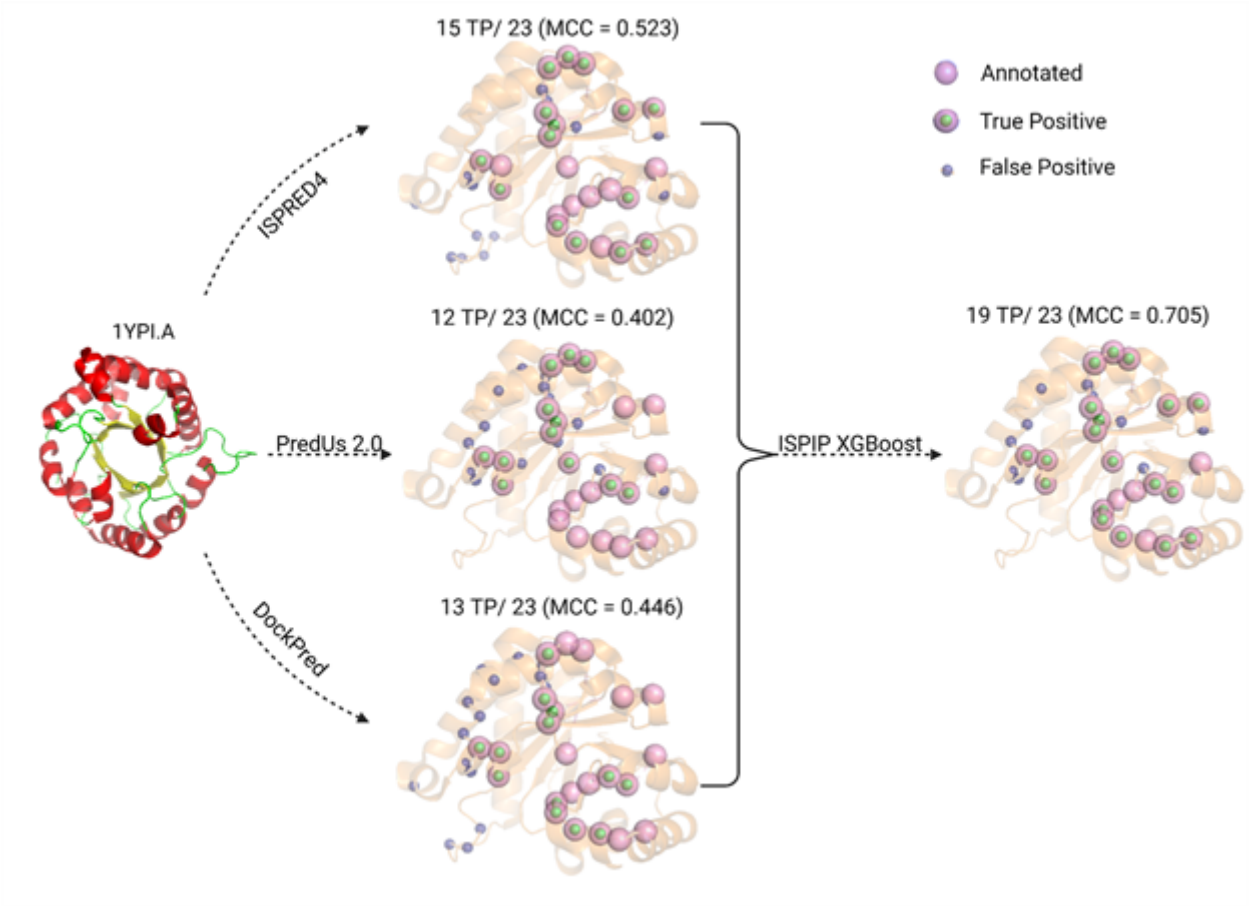

Figure 2

ISPIP consensus prediction of interface residues: On the left, the structure (1YPI.A) is shown. In the middle, the interface prediction of the 3 input classifiers is displayed. On the right, the ISPIP consensus prediction includes overlapping and unique TP residues of the input classifiers to yield an improved interface prediction of 19 TP out of the 23 annotated residues.

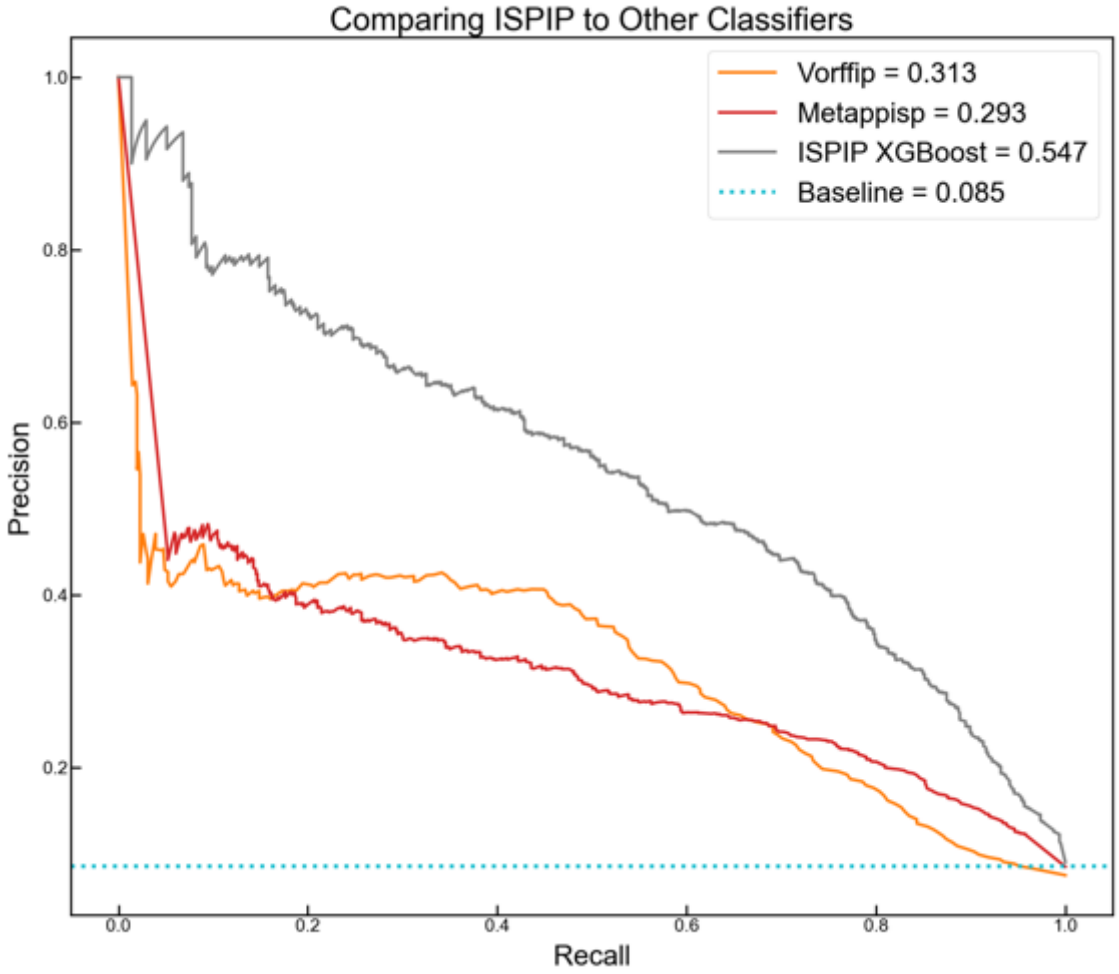

Figure 3

ISPIP outperforms other structure-based classifiers and meta-predictors: The PR curves highlight ISPIP's improved performance of a complex structure-based classifier (VORFFIP) and previous meta-predictor (meta-PPISP).

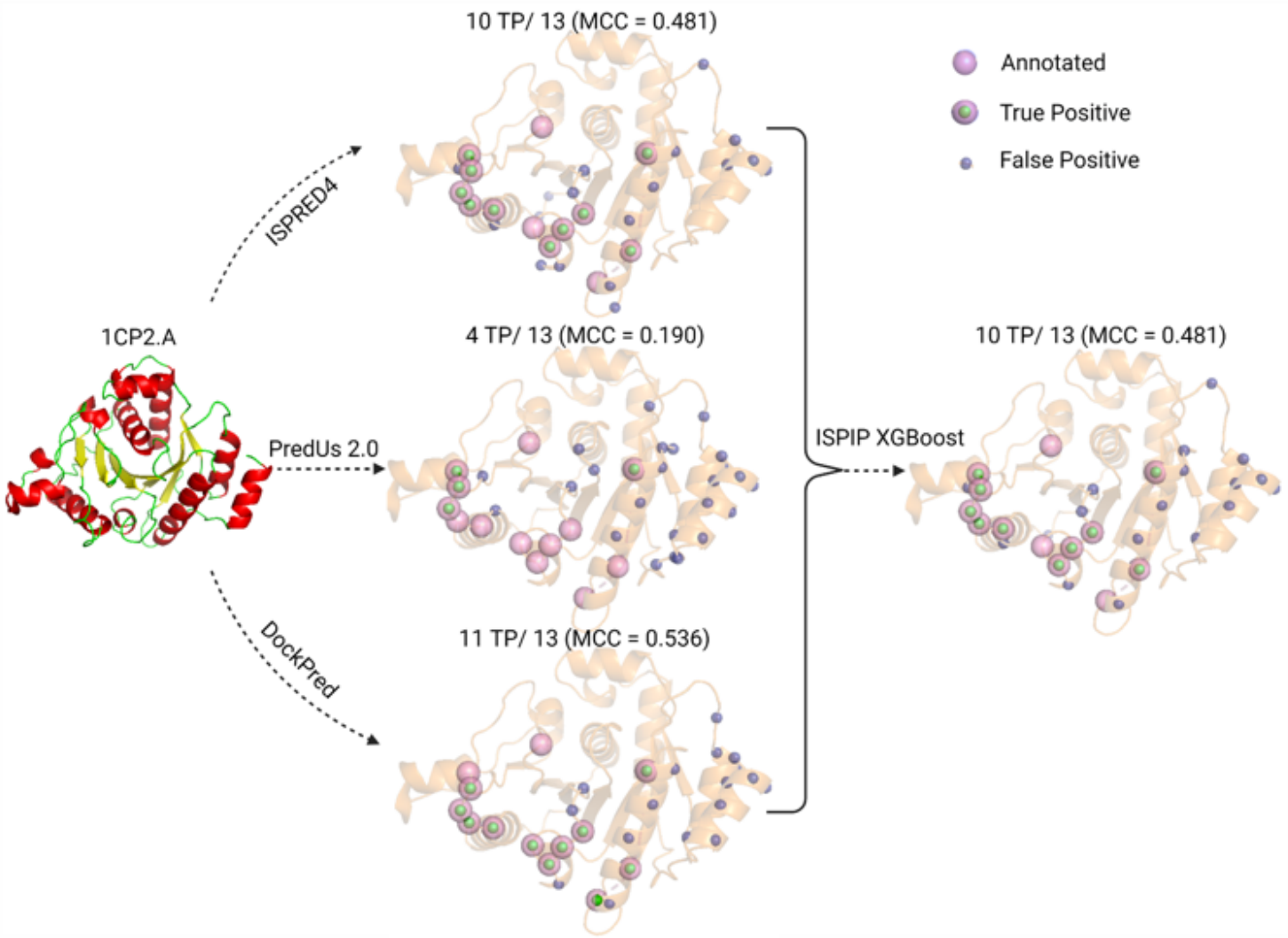

Figure 4

ISPIP is robust to poor performance of input classifier: On the left, the structure of 1CP2.A) is shown. In the middle, the interface prediction of the 3 input classifiers is displayed. PredUs 2.0 has an especially poor prediction relative to the other 2 input classifiers. On the right, the ISPIP has a robust consensus prediction with 10 TP out of the 13 annotated residues, despite the poor performance of the PredUs 2.0 input classifier.

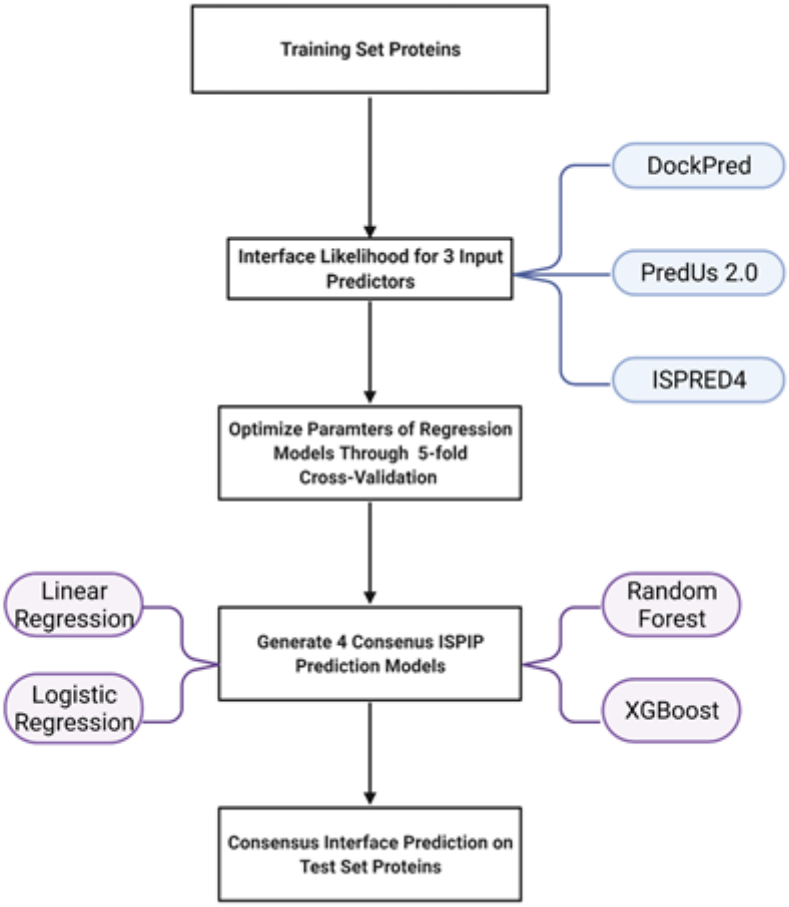

Figure 5

Flowchart of ISPIP Methodology: ISPIP's classification models are generated through training on the interface likelihoods of the three input predictors. (Created with BioRender.com)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigs.docx
- SupplementaryTables.docx