JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

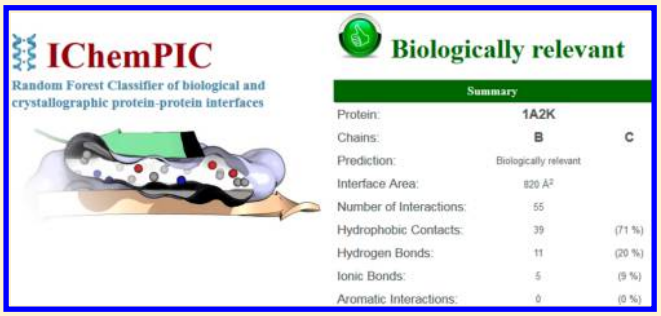
IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein−Protein Interfaces

Franck Da Silva, Jérémy Desaphy,[†] Guillaume Bret, and Didier Rognan*

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS−Université de Strasbourg, 67400 Illkirch, France

Supporting Information

ABSTRACT: Protein−protein interactions are becoming a major focus of academic and pharmaceutical research to identify low molecular weight compounds able to modulate oligomeric signaling complexes. As the number of protein complexes of known three-dimensional structure is constantly increasing, there is a need to discard biologically irrelevant interfaces and prioritize those of high value for potential druggability assessment. A Random Forest model has been trained on a set of 300 protein−protein interfaces using 45 molecular interaction descriptors as input. It is able to predict the nature of external test interfaces (crystallographic vs biological) with accuracy at least equal to that of the best state-of-the-art methods. However, our method presents unique advantages in the early prioritization of potentially ligandable protein−protein interfaces: (i) it is equally robust in predicting either crystallographic or biological contacts and (ii) it can be applied to a wide array of oligomeric complexes ranging from small-sized biological interfaces to large crystallographic contacts.



INTRODUCTION

Protein−protein interactions (PPI) stand at the heart of most pathophysiological situations in living cells and therefore have attracted more and more attention in drug discovery.[1−3] Among the many strategies to identify low molecular weight PPI modulators, rational structure-based approaches have historically played an important role, notably because of the possible integration of biophysical screening of fragment libraries (surface plasmon resonance, isothermal titration calorimetry, nuclear magnetic resonance spectroscopy, mass spectrometry) with X-ray structure determination.[4] To fully exploit the current structural knowledge on druggable targets, it is desirable to ascertain their true oligomeric state as well as their biological relevance. Throughout this article, we will consider as biological any protein−protein complex with a true biological relevance and function (e.g., cell adhesion, cell signaling, immune recognition, transcription). Homo- or hetero-oligomeric complexes resulting either from crystal packing or lacking any known biological function will be considered crystallographic. Unfortunately, inferring the quaternary structure and biological relevance from atomic coordinates in the Protein Data Bank (PDB)[5] is not straightforward. For example, the contents of the asymmetric unit (ASU) deposited in the PDB (the fraction of the crystallographic unit cell that has no crystallographic symmetry) can describe one or several copies of a macromolecule but with no particular indication on which oligomeric state (e.g., monomer, dimer) is the most relevant. Likewise, the ASU may need crystallographic symmetry operations to be applied before reconstituting the beforehand known biologically relevant macromolecular assembly (biological unit). Automated procedures to discriminate, from 3D structures, crystal from biologically relevant and stable interfaces are, therefore, needed to avoid long and costly biochemical experiments such as gel filtration, light scattering, or equilibrium sedimentation.

As a rule of thumb, crystallographic interfaces are generally much smaller (<1000 Å^2) than biologically relevant ones.[6] However, this simple rule suffers from many exceptions since some very important interfaces, like those involving α-helix recognition sites, may be quite small in size (e.g., 780 Å^2 for the p53−mdm2 complex). Many classification methods have been designed, therefore, to directly predict the oligomeric status of protein complexes from atomic coordinates.[7] The very first approach, reported in 1998 as PQS (protein quaternary structure file server),[8] used an empirical scoring function based on several contributions (interface contact area; number of interfacial buried residues, salt bridges, and disulfide bonds; solvation energy of quaternary structure formation). Although it is not perfect (at least 20% of misclassifications were reported by the authors themselves), the PQS server paved the way for many methods that can be grouped in two categories.

A first type of approach, of which PISA[9] is representative, relies on first-principles physics to predict the stability of protein assemblies in solution. For example, PISA explicitly computes Gibbs dissociation free energies to predict the biological relevance of a macromolecular assembly. When applied to a dataset of 218 PDB structures, it achieved a

Received: April 6, 2015
Published: September 7, 2015

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

remarkable success rate of 90% in predicting true biological interfaces.[9] PISA can be considered to be a reference method, as it is currently used to predict quaternary structures of every entry of the RCSB PDB web site. A second group of methods[7,10−17] applies linear or nonlinear regression/classification models to predefined training sets (crystallographic, biological) in order to predict the quaternary structure of external test sets. Many geometrical and chemical complementarity descriptors of the interface can be used to discriminate, with comparable accuracies (ca. 85−90%), crystal from biological contacts. Very often, these methods (e.g., IPAC,[7] DiMoVo,[12] or NOXClass[13]) utilize a machine learning algorithm (support vector machine, decision trees, Bayesian inference) trained on atom or residue-based contact vectors to decide which parameter set is the most adequate for an optimal classification. Residue conservation of interface core residues[18,19] can be added to the above-cited descriptors, as, for example, in EPPIC,[20] to highlight the importance of highly buried core residues at biological interfaces.

To allow them to be compared, most studies have relied on a limited number of benchmarking datasets,[10,21,22] which turned out to be biased toward small crystal and high-affinity large biological interfaces.[7,12,20] As a consequence, most current classification methods have a much lower accuracy when applied to a set of interfaces (biological, crystal) with an equivalent distribution of interface areas. A recently designed dataset[20] paid attention to select true biological and crystallographic interfaces with a hard cutoff with respect to interface areas (mostly above 1000 Å^2). Since we finally aim at identifying potentially ligandable protein−protein interfaces that may be small in size,[23] none of the existing datasets appears to be satisfactory. We therefore designed a hand-curated dataset (FDS set) of 200 biologically relevant nonredundant protein−protein complexes of known X-ray structure, which was further supplemented by an equivalent number of 200 crystal interfaces filtered to span a comparable interface area range. We next used a machine learning algorithm (Random Forest) and 45 molecular interaction descriptors to train a model that, when applied to several external test sets, achieves good accuracy and robustness in distinguishing between crystallographic and biologically relevant interfaces, whatever their size.

COMPUTATIONAL METHODS

Datasets. *FDS Dataset.* Crystallographic interfaces were retrieved from two previously reported datasets.[20,21] First, 141 known monomeric proteins from the Bahadur dataset[21] with a crystalline interface area in the 400−1200 Å^2 range were retrieved as follows. Atomic coordinates of the asymmetric unit were retrieved from the RCSB Protein Data Bank, and one unit cell was reconstructed for each entry using AmberTools14.[24] For each structure and all possible pairs of chains, the interface area IA (eq 1) was measured with MSMS[25] after removing nonprotein atoms (solvent, ligands, ions) and using a probe radius and vertex density of 1.4 Å and 2.0/Å^2, respectively.

$$IA_{A,B} = \frac{(ASA_A + ASA_B) - ASA_{AB}}{2} \quad (1)$$

where $IA_{A,B}$ is the interface area between chains A and B, ASA_A is the solvent-accessible surface area of isolated chain A, ASA_B is the solvent-accessible surface area of isolated chain B, and ASA_{AB} is the solvent-accessible surface area of complex AB.

The largest interface area was kept for each PDB entry. The Bahadur set was then complemented by 82 interfaces from the DCxtal dataset[20] selected to present an interface area in the 1000−1500 Å^2 range for proteins reported to be monomeric in solution. The corresponding PDB files were directly retrieved from the EPPIC web site (http://www.eppic-web.org/ewui/#downloads). Protein redundancy was removed by keeping only one entry in cases where sequence identity between two different entries was above 70%, according to RCSB redundancy rules.[26] The final set of 200 nonredundant crystallographic interfaces (PDB identifier, protein name, chains, interface area, resolution, protein classification) is given in Table S1.

A collection of 200 biologically relevant nonredundant protein−protein interfaces (<70% pairwise sequence identity between any two chains) was manually assembled from the literature according to the following sources: (i) the recently described DCbio dataset[20] of homodimeric biological interfaces (74 PPIs); (ii) the 2P2I database[23] that archives heterodimers for which an X-ray structure exists for the complex, each individual monomer in the free state, and one partner is bound to a low molecular weight inhibitor (10 PPIs); (iii) existing small molecular weight inhibitors[27] for biologically relevant PPIs of known X-ray structure (five PPIs); (iv) cancer-related PPIs[1] of known X-ray structure (eight PPIs); (v) the PPI affinity database[28] of biologically relevant PPIs with available X-ray structures (complex, free state) and known experimental binding constant (54 PPIs); (vi) the dataset of "hot loops" mediated PPIs[29] of known X-ray structures (20 PPIs) and (vi) diverse biologically relevant heterodimeric proteins (18 PPIs). The biological relevance of all of these 200 complexes was checked manually according to known literature data.[20,23,27−29]

The corresponding structures were downloaded from the Protein Data Bank (PDB). Chains participating in the interface were selected manually according to the above-cited sources. After removing nonprotein atoms (solvent, ligands, ions), the buried interface area for the selected chains was computed as previously described in eq 1. The full set of 200 biological interfaces (PDB identifier, protein names, chains, interface area, selection mode, resolution, protein classification) is given Table S2.

The above-described 400 interfaces (crystallographic, biological) were randomly split into two sets (75% in the training set; 25% in the test set), maintaining an equal proportion of crystallographic and biological interfaces in each subset. Caution was also given to ascertain an equivalent distribution of interface area sizes in both sets. Following the above-described procedure, different random splits (75/25) do not influence the obtained results (best RF parameters, F-measure of the best RF models on the validation and external test sets; data not shown). Training or test set membership is given in Tables S1 and S2.

IPAC,[7] Ponstigl,[10] and Bahadur[21] datasets were retrieved from the PDB according to PDB identifiers and chain names described in the literature.

Atomic Coordinates. For each input PDB file, hydrogen atoms were added with Protoss,[30] a recently described method for the placement of hydrogen coordinates in protein−ligand complexes that takes tautomers and protonation states of both protein chains into account. The method generates the most probable hydrogen positions on the basis of an optimal hydrogen-bonding network using an empirical scoring function.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

Full atomic coordinates of FDS dataset entries can be downloaded at http://bioinfo-pharma.u-strasbg.fr/IChemPIC.

Protein−Protein Interface Descriptors. Interfaces between protein chains are detected following a three-step procedure as follows. First, the interface is broadly defined by counting pairwise distances between all atoms of different chains and keeping only patches for which at least 20 interatomic distances are shorter than 5 Å. In a second step, all intermolecular interactions (hydrophobic, aromatic, hydrogen bond, ionic bond) between the two selected chains are precisely defined using standard parameters of the in-house developed IChem toolkit.[31] The set of topological rules, used to define interactions based on atom pair-dependent distances and angles, is explicitly described in the previous report describing the IChem toolkit.[31] In the third step, an interaction pseudoatom (IPA) is placed at the mid-distance of each atom pair in an interaction according to IChem. Please note that hydrophobic IPAs are clustered if they are less than 1.0 Å apart.[31] If the total number of IPAs is higher than or equal to 5, then the interface is saved; otherwise, it is discarded. Last, a vector of 45 real numbers is generated for each remaining interface describing its size, chemical complementarity, and buriedness (Table S3). The final vector has the following form:

- the total number of IPAs (one parameter);
- the percentage of each of the four interaction types (four parameters);
- the distribution (in counts), for each interaction type, of the buriedness of the corresponding IPAs, binned in 10 intervals from 25 to 100% burial (4 × 10 = 40 parameters). Buriedness of each IPA was inferred as previously reported[32] by computing the proportion of 120 regularly spaced 8 Å long rays having their origin at the IPA 3D coordinates and intersecting the surrounding protein surface.

Altogether, the complete process comprising hydrogen atom addition, interaction detection, and descriptor generation is fast enough (a few seconds per PDB entry) to be applied to the entire PDB.

Random Forest Model. Random Forest (RF) models were built using the RandomForest 4.6-7 library[33] in the R statistical package.[34] A total of 500 decision trees (ntree parameter) were trained on all descriptors of the training set (n = 300), but the number of variables (mtry parameter) at each splitting node was varied. A 5-fold cross-validation procedure was used to randomly split the training set five times into an internal training (four-fifths of the dataset) and an internal test set (one-fifth of the dataset) and analyze the predictive power of RF models on the nonoverlapping five internal test sets. For each mtry value (integer between 2 and 10), the corresponding cross-validated model was assessed according to the following statistical parameters

$$\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{F-measure} = 2 \times (\text{sensitivity} \times \text{precision})/(\text{sensitivity} + \text{precision})$$

where TP are true positives (biological interfaces predicted biological), FP are false positives (crystallographic interfaces predicted biological), TN are true negatives (crystallographic interfaces predicted crystallographic), and FN are false negatives (biological interfaces predicted crystallographic)

The best mtry value was used (i) to derive 10 RF models from the full training set (300 complexes) by varying the random seed number and (ii) these 10 RF models were applied to predict the nature of interfaces in the 100 entries of the external test set.

Comparison to Other Methods. IChemPIC predictions were done using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) and rely on the majority of the 10 independent RF predictions (biological or crystallographic) to annotate an input protein−protein interface. In cases where there are an equal number of predictions for both types, the interface is predicted to be crystallographic. IChemPIC was compared to four state-of-the-art methods (NOXClass,[13] PISA,[9] DiMoVo,[12] and EPPIC)[20] on three external test sets. For each of these tools, standard parameters available at their online version were chosen, giving as input either the PDB code and chain letters (biological interfaces) or the above-prepared atomic coordinates for the two chains (crystallographic interfaces). For the NOXClass multistage SVM classification (http://noxclass.bioinf.mpi-inf.mpg.de/), the pairwise class probabilities (biological, crystallographic) were retained for each pair of protein chains, using three descriptors (interface area, interface area ratio, area-based amino acid composition). In PISA (http://www.ebi.ac.uk/pdbe/pisa/), the interface was defined as biological if the corresponding interface was predicted to be stable among all proposed assemblies. Otherwise, the interface was predicted to be crystallographic. Using the DiMoVo prediction method (http://albios.saclay.inria.fr/dimovo), a score above 0.50 was used to assign a potential biological function to an interface. Last, EPPIC predictions (Bio or Xtal) were done on a web server (http://www.eppic-web.org/ewui/) and based on the consensus voting scheme (final score) based on four descriptors (core sizes, geometry, core-rim, core-surface) of the input interface.

RESULTS AND DISCUSSION

Setting Up the FDS Dataset of Ligandable Protein−Protein Interfaces. None of the existing benchmark datasets is suitable for the purpose of discriminating crystallographic from biologically relevant protein−protein interfaces. On one hand, historical datasets[10,13,21,22] are biased by having a majority of crystallographic entries of much lower interface areas (500−1000 Å^2) than those of true biologically relevant entries (1000−3000 Å^2). On the other hand, the DC dataset[20] corrected this discrepancy by selecting entries with an homogeneous distribution of interface areas (1000−1500 Å^2) that, however, still falls outside the applicability domain of many biologically important PPIs (e.g., p53−mdm2 interface of 780 Å^2; PDB ID: 1YCR) modulated by low molecular weight inhibitors.[35] Both the Bahadur[21] and DC datasets, which do not overlap much with respect to the interface area range, were therefore merged to enlarge the applicability domain of our next predictions. Lastly, we manually collected an additional set of 115 biologically relevant PPIs to yield a final number of 400 interfaces, which was split into training (75% of data) and test (25% of data) sets. Inspecting the respective distribution of interface area sizes in both sets reveals no major bias, although biological interfaces remains, on average, slightly larger than

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

crystallographic ones (Figure 1). We will demonstrate later that the size of the interface has no major influence in discriminating crystallographic from biological interfaces.

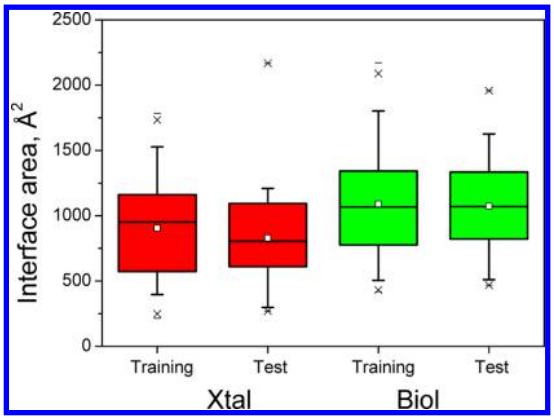


Figure 1. Distribution of interface areas in the designed FDS training and test sets (Xtal, crystallographic interface; Biol, biological interface). Boxes delimit the 25th and 75th percentiles; whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a square in the box. Crosses delimit the 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

About 80% of crystallographic complexes (both in the training and test sets) relate to enzymes with well-defined catalytic sites, with the rest being dominated by protein transporters. The proportion of enzymes in the biologically relevant set is lower (about 55% in both the training and test sets), with the biologically relevant set exhibiting more examples of recognition complexes involved in important biological processes (e.g., immune recognition, cell signaling, cell adhesion, transcription).

The average resolution of crystallographic and biologically relevant complexes was 1.84 ± 0.35 and 2.10 ± 0.58 Å, respectively. A large majority of structures (ca. 75%) in both sets was solved at high resolution (<2.5 Å). In the protein preparation step, we ascertained that all side chains participating in the interface were fully described. None of the herein described 400 PDB complexes present an incomplete side chain at the selected protein−protein interface. Along the same line, we checked that no small ions could stabilize the interface of the herein described complexes. Handling water molecules at the interface is tricky since water molecules are absent in 184 of the 400 complexes. We therefore decided to remove all water molecules, whatever their location, resulting in a unique preparation protocol and a fair comparison to other methods that do not take bound waters into consideration. Coming back to the 400 raw PDB files, it appears that bound waters are not frequently present at protein−protein interfaces. When this is the case (30% of 216 entries with explicit water molecules), bound waters typically engage in no more than a single hydrogen bond. We therefore do not believe that the decision to remove bound waters really affects the accuracy of our classifier.

PPI Detection and Descriptor Generation. We first detect the interface between two protein chains and then explicitly compute all nonbonded interactions (hydrophobic contacts, aromatic interactions, hydrogen, and ionic bonds) and generate a pseudoatom (IPA) for describing each interaction (Figure 2). A complex molecular assembly of thousands of

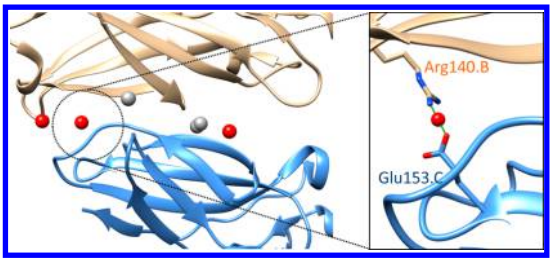


Figure 2. Interface (PDB ID: 4NNY) between interleukin-7 receptor subunit alpha (tan, chain C) and cytokine receptor-like factor 2 (blue, chain C). Six interaction pseudoatoms (spheres) are placed at the mid-distance of each pair of atoms in the interaction and assigned a property corresponding to the interaction (hydrophobic, aromatic, hydrogen bond, ionic bond). The right inset is a zoomed view of a single ionic bond that explicitly displays the interacting side chains.

atoms can be represented, therefore, by a much simpler set of a few IPAs (60−70 on average) describing both the nature and buriedness of the corresponding interaction. Since we explicitly consider hydrogen bonds, it is worth noting that all hydrogen atoms are added to the raw PDB files while optimizing the tautomeric and protonation states protein's amino acids.[30]

Although biological interfaces exhibit, on average, many more IPAs (86 ± 40) than crystal packing contacts (50 ± 30), the average percentages of interaction types are rather similar in both sets (Table 1). As expected, hydrophobic contacts are

Table 1. Average Percentage of Interaction Types at Crystallographic and Biological Interfaces

	protein−protein interface[a]	
interaction type	crystallographic	biological
hydrophobic	78.06 ± 15.70	83.32 ± 9.71
aromatic	0.24 ± 1.14	0.10 ± 0.32
hydrogen bond	17.97 ± 12.11	13.51 ± 7.24
ionic bond	3.65 ± 5.80	3.00 ± 3.87

[a]Statistics from 27 186 protein−protein interactions (200 crystallographic and 200 biological interfaces from the FDS set) detected by IChem.[31]

dominant and represent about 80% of all interactions, although they are more populated in biological interfaces (Table 1). Aromatic interactions (edge-to-face and face-to-face) are quite rare, but they are more populated in the crystallographic set, thereby confirming previous observations.[21] Hydrogen bonds are more frequently found in crystallographic interfaces than in biological assemblies. However, the quality of these hydrogen bonds (e.g., strength, accessibility) is not taken into account in the current analysis. Lastly, the frequency of ionic bonds is rather constant for the two interface categories (3%). Since metal chelation is rarely found at protein−protein interfaces, this interaction type was discarded from the descriptor set to define the shortest possible input vector for RF modeling.

Random Forest Binary Classification Model. Random Forest (RF) is a highly versatile ensemble machine learning method for classification and regression that relies on many independent decision trees.[36] Each tree is created by bootstrap samples of the original training data using a randomly selected subset of features. Then individual trees are combined through a voting process to provide an unbiased prediction. In contrast with single-decision trees, random forests have a low variance and very few biases. Considering that random forests have few

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

parameters to tune (number of trees, number of variables at each split), the method is easy to use in order to produce a reasonable model fast and efficiently. Among its many potential applications, RF is increasingly used in life sciences as either a classifier or nonlinear regression method.[37]

In our application, the number of trees (ntree) was fixed to 500. Besides it having a clear influence on the overall computing time, variations of this parameter did not influence the herein presented results. The number of variables randomly sampled as candidates at each split (mtry parameter) was systematically varied from 2 to 10, and each model was repeated five times by varying a random seed number. Using a mtry value equal to 4, Random Forest modeling leads to a stable and robust 5-fold cross-validated model (F-measure = 0.776 ± 0.09) when applied to the FDS training set (Table 2).

Table 2. Statistics of the Best RF Model on the FDS Dataset

parameter	training set (n = 300)[a]	external set (n = 100)[b]
sensitivity	0.794 ± 0.017	0.728 ± 0.014
precision	0.759 ± 0.010	0.745 ± 0.018
specificity	0.747 ± 0.014	0.750 ± 0.025
accuracy	0.771 ± 0.009	0.739 ± 0.012
F-measure	0.776 ± 0.009	0.736 ± 0.010

[a]Mean and standard deviation of the best 5-fold cross-validated model (ntree = 500, mtry = 4), repeated five times using different random seed numbers. [b]Mean and standard deviation of the best RF model (ntree = 500, mtry = 4) at predicting the nature of 100 protein−protein interfaces, repeated 10 times using different random seed numbers.

The model is equally good at predicting either biological contacts (sensitivity) or crystallographic interfaces (specificity). When applied to the FDS external set of 100 PPIs, a moderate drop in accuracy (0.739 ± 0.012) and F-measure (0.736 ± 0.010) is observed, but the model is still robust at predicting the two categories of PPIs equally well (sensitivity = 0.728 ± 0.014; specificity = 0.750 ± 0.025; Table 2).

To be sure that observed data are neither the result of overtraining nor chance correlation, we first performed a y-scrambling test by randomly assigning the dependent variable (crystallographic or biological) to each of the 400 protein−protein interfaces of the FDS training and test sets. As expected, the F-measure of the corresponding RF models (same parameters as above) significantly dropped to mean values of 0.515 and 0.502 when applied to the training and external test sets, respectively. We next computed 45 RF models (ten runs/model) in which the values of the 45 descriptors were iteratively scrambled for each entry of the training set. For all 45 descriptors, the previously computed 300 values of descriptor d_i were just randomly assigned to the 300 objects (training interfaces). Analyzing the variations in the mean F-measure for the training set permits the identification of the most important parameters among our 45 descriptors (Figure 3).

Out of the 45 descriptors, 11 have a real contribution to the global model (>1% decrease in F-measure) when their respective values are scrambled. The most important parameters are clearly the number of interaction pseudoatoms (nPTS) and the percentage of fully buried hydrophobic contacts (hydro7−hydro10 descriptors; Table S3).

Permutating the values taken by the total number of IPAs (nPTS) decreases the overall F-measure of the model by 1.6%

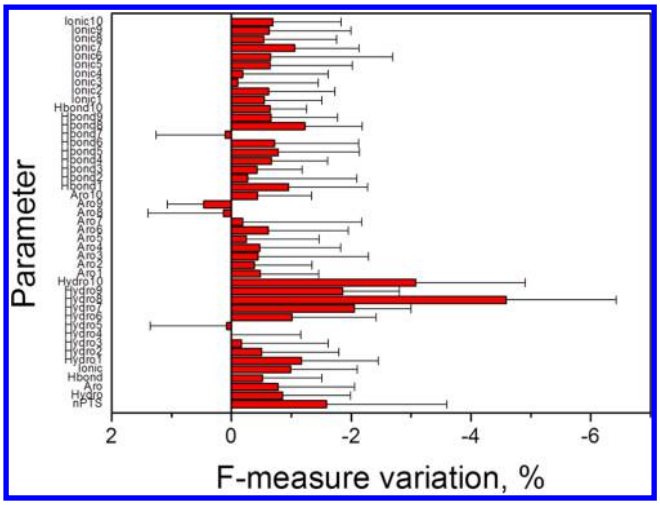


Figure 3. Influence of the permutation of descriptor values on the mean F-measure of 10 RF models obtained with the best cross-validated parameters (ntree = 500, mtry = 4) and trained on the FDS training set.

(Figure 3). Although accessible hydrophobic contacts (hydro1−hydro6 parameters) do not really contribute to the overall F-measure, the more buried hydrophobic interactions (hydro7−hydro10 parameters) are truly critical. Remarkably, permutating the value of the hydro10 parameter (percentage of 100% buried hydrophobic contacts) decreases the F-measure of the RF model by almost 3% (Figure 3). Accordingly, hydrophobic core interface residues, defined as buried by at least 95%, have recently been described as key determinants of biological interfaces.[20] Of less importance, but still significant, is the percentage of other interactions (hydrogen bonds, ionic bonds) and their buriedness, which tends to be higher in biological interfaces than in crystal packing contacts (Figure 3). Scrambling the values of four out of the 45 parameters (hydro5, Aro8, Aro9, Hbond7) leads to slightly better RF models. The largest observed decrease in F-measure (scrambling values of hydro8 parameter) is only by 5% and is probably explained by compensatory effects upon removal of the most critical descriptor. To demonstrate this assumption, we suppressed the hydro8 descriptor from the vector, recomputed a RF model on the set of $n - 1$ descriptors (F-measure of 0.705 on the training set), and iteratively scrambled again the $n - 1$ descriptor values. This time, the most critical descriptor is hydro10 (the former second most important descriptor starting from the full set of descriptors), with a much higher mean decrease in the F-measure (11.3 ± 3.3%). This observation perfectly illustrates our hypothesis and the compensatory effect of the hydro10 parameter upon removing the influence of the hydro8 descriptor.

The weaker contribution of the hydro9 parameter (count of hydrophobic IPAs buried is between 91.6 and 100%) with respect to that of the related hydro8 (count of hydrophobic IPAs buried is between 83.3 and 91.6%) and hydro10 (count of 100% buried hydrophobic IPAs) parameters is intriguing and explained by a peculiar distribution of these parameter values when comparing crystallographic and biologically relevant interfaces (Figure 4). Hence, the distributions of hydro8 and hydro10 counts are clearly different when examining the two subsets of interfaces (shift to higher hydro8 values in crystallographic contacts; shift to higher hydro10 values in biological interfaces). Intriguingly, the hydro9 parameter values

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

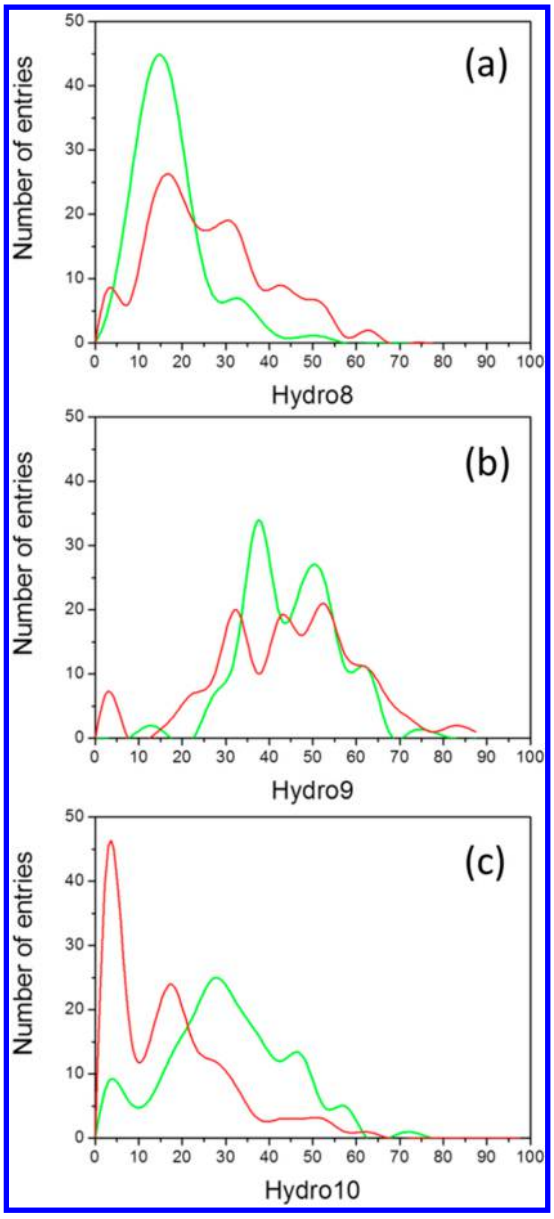


Figure 4. Distribution of the hydro8–hydro10 parameter counts across the FDS training set (green: biological interfaces; red: crystallographic interfaces).

are similarly distributed (Figure 4), thereby explaining why it contributes less to the RF cross-validated model.

To confirm the above-suggested importance of some interface parameters (nPTS, hydro7–hydro10), we ranked the 300 training interfaces by decreasing value of each descriptor (45 lists of PDB entries ranked by decreasing counts for the current descriptor under investigation). We next performed a binary classification of the 300 entries (crystallographic, biological) from the ranks obtained in these 45 lists. A perfect descriptor would lead to a classification (ROC AUC = 1) in which all 150 biological interfaces are ranked higher than the first ranked crystallographic interface. Using the ROC classification, we can estimate the relative importance of each descriptor in discriminating the two categories. Any single descriptor-based classification with an AUC higher than 0.7 (Figure 5) indicates that this descriptor is particularly efficient. This analysis confirms the crucial role of two parameters

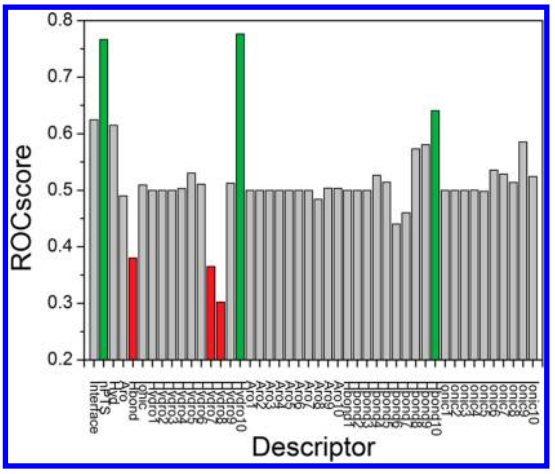


Figure 5. Area under the ROC curve for a binary classification (crystallographic or biological) of 300 interfaces (FDS training set) based on decreasing counts of each of the 45 IChemPIC descriptors.

(nPTS, hydro10) in the discrimination of the two interface subsets. This complementary analysis also shows that counts observed for three descriptors (Hbond, hydro7, and hydro8) are, indeed, higher for crystallographic contacts (ROCscore < 0.50) and therefore also helps to discriminate the two sets of PDB entries. Importantly, using the interface area as a descriptor does not lead to a good binary classification (ROCscore = 0.59), confirming that the FDS training set is really well-balanced with respect to this important criterion, which has been overlooked in the past.

Comparison of IChemPIC to Existing Methods. IChemPIC was compared to four state-of-the-art methods (NOXClass,[13] PISA,[9] DiMoVo,[12] and EPPIC[20]) with regard to their ability to predict the nature of PPIs originating from three different external test sets.

Regarding the diversity of interfaces in the herein presented FDS dataset, it is not surprising that the observed accuracy of existing methods is significantly lower than that reported in the seminal reports describing them.[9,13,12,20] NOXClass is remarkably sensitive (good true positive rate), but this comes at the cost of a much lower specificity (low true negative rate). In contrast, EPPIC and, to a lesser extent, DiMoVo are specific in detecting crystal assemblies (specificity = 0.949), but they perform less well in recognizing biological contacts (low sensitivity), notably when the interface area is small (<750 Å^2; Table S4). PISA, the method currently used in predicting macromolecular assemblies in the RCSB PDB, is the most stable with respect to all statistical parameters taken into account (Table 3).

Altogether, IChemPIC still appears to be the method of choice for a binary classification of protein–protein interfaces since it provides constant and robust performance at predicting both biological and crystal interfaces, whatever their size (see full predictions in Table S4). On one hand, it is inferior to NoxClass and PISA at detecting biological contacts, but it is much more accurate at predicting crystallographic interfaces. On the other hand, IChemPIC is less accurate than EPPIC at predicting crystallographic contacts, but it is significantly better at predicting biologically relevant interfaces (Table 3). Out of the five methods tested here, IChemPIC is notably the only method able to predict, with a similar good accuracy, either small biological interfaces or large crystal contacts.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005–2014

Table 3. Comparison of IChemPIC and State-of-the-Art Methods' Abilities To Predict the Status (Crystallographic, Biological) of the FDS External Test Set (n = 100)

	method				
statistics	IChemPIC[a]	NOXClass	DiMoVo[b]	PISA[c]	EPPIC[d]
sensitivity	0.740	0.878	0.480	0.771	0.667
precision	0.755	0.694	0.857	0.725	0.909
specificity	0.760	0.525	0.733	0.674	0.949
accuracy	0.750	0.719	0.538	0.725	0.826
F-measure	0.747	0.775	0.615	0.747	0.769

[a]Consensus predictions (biological or crystallographic) out of 10 independent RF models. In cases where there is an equal number of predictions for both properties, the interface is predicted to be crystallographic. Predictions were obtained using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC). [b]Thirty five entries common to the DiMoVo and FDS training sets were excluded. [c]Two entries (1i5h, 1y7q) could not be predicted by PISA (crystallographic data is either absent or incomplete); seven additional entries common to the PISA and FDS training sets were excluded. [b]Twenty nine entries common to the EPPIC and FDS training sets were excluded.

Since IChemPIC has been trained on the FDS dataset, it is fair to compare its performance on totally independent external test sets. We therefore chose three additional external datasets (IPAC,[7] Ponstigl,[10] and Bahadur[21]) containing PDB entries not present in the FDS training set. The first two sets have notably been used for benchmarking most tools similar to IChemPic. As stated before,[12,20] the Bahadur and Ponstingl datasets are not very informative because they have a strong bias toward small crystallographic and large biological contacts. As a consequence, all programs including IChemPIC achieve an excellent accuracy (0.85−0.95) at predicting the nature of these entries (Table 4). IChemPic notably exhibits the highest F-measure (0.932 and 0.870, respectively) on these two external sets, which indicates its robustness at predicting biological and crystallographic interfaces equally well (see full predictions in Tables S5 and S6).

The last external set (IPAC validation set 3)[7] is composed of 66 heterodimeric proteins of known crystal structure and experimentally determined binding constants. It notably permits the sensitivity of the method at predicting biological interfaces of quite different strengths to be evaluated. Out of the five methods, NOXClass presents the highest performance (only sensitivity is reported since crystallographic interfaces are lacking in this set) when applied to this dataset (Table 4). Surprisingly, this method never fails when it is applied to the lowest affinity complexes ($K_d < 10^{-5}$ M; Table S7). Given the propensity of NoxClass to overpredict biological interfaces in the previously examined external test sets (sensitivity ≫ precision; Tables 3 and 4), its excellent performance should be considered with extreme caution. Other methods are indeed sensitive to the strength of the corresponding complexes and logically failed more often for low-affinity than for high-affinity complexes (Table S7). Among these methods, IChemPic clearly exhibits the highest accuracy (Table 4).

Practical Application of IChemPIC to PDB Biological Unit Files and Reasons for Its Failure. IChemPIC was next applied to classify 4950 nonredundant interfaces from the Dockground resource of protein−protein X-ray structures.[38] All of these structures are based on the proposed biological unit file (Biounit) inferred from PISA predictions and provided online by the RCSB PDB. About 30% (1493 in total) of these interfaces are, nevertheless, predicted as crystallographic by IChemPIC (Table S8). These discrepancies result from three major causes (Figure 6).

First, our method, like any other, is far from being perfect and fails in ca. 25% of test cases (recall Tables 3 and 4). In many of these cases, the error occurs because IChemPIC predicts interfaces and not quaternary structures. Hence, two chains may form a stable interface depending on the precise context of a much larger oligomeric state. For example, the isolated interface between CTLA-4 (chain B) and B7-2 (chain

Table 4. Performance of IChemPIC with Respect to State-of-the-Art Methods at Predicting the Status (Crystallographic, Biological) of Three Independent Benchmark Sets

	no. of interfaces			method				
test set	crystallographic	biological	statistics	IChemPIC[a]	NOXClass	DiMoVo	PISA	EPPIC
Bahadur[b]	20	122	sensitivity	0.902	0.938	n.a.[c]	0.918	0.885
			precision	0.965	0.892	n.a.	0.875	0.973
			specificity	0.800	0.450	n.a.	0.556	0.850
			accuracy	0.887	0.855	n.a.	0.835	0.880
			F-measure	0.932	0.915	n.a.	0.896	0.927
Ponstingl[d]	67	76	sensitivity	0.882	0.919	0.714	n.a.[e]	0.895
			precision	0.859	0.760	0.714	n.a.	0.840
			specificity	0.831	0.731	0.930	n.a.	0.806
			accuracy	0.858	0.822	0.887	n.a.	0.853
			F-measure	0.870	0.832	0.714	n.a.	0.866
IPACdb[f]	0	66	sensitivity	0.706	0.946	0.394	0.682	0.636

[a]Consensus predictions (biological or crystallographic) out of 10 independent RF models. In cases where there is an equal number of predictions for both properties, the interface is predicted to be crystallographic. Predictions were obtained using the IChemPIC server (http://bioinfo-pharma.u-strasbg.fr/IChemPIC). [b]One-hundred forty two PDB structures (122 biological, 20 crystallographic) not present in the IChemPIC training set. Entries present in NoxClass (n = 25), DiMoVo (n = 142), and PISA (n = 63) training sets were removed when the corresponding method was used for prediction. [c]Not applicable because DiMoVo was trained on the Bahadur set. [d]One-hundred forty three PDB structures (76 biological, 67 crystallographic) not present in the IChemPIC training set. Entries present in NoxClass (n = 14), DiMoVo (n = 72), and PISA (n = 109) training sets were removed when the corresponding method was used for prediction. [e]Not applicable because PISA was trained on the Ponstingl set. [f]Sixty six PDB heterodimeric structures (validation set 3) of known binding constants. Entries present in IChemPic (n = 15) and NoxClass (n = 10) training sets were removed when the corresponding method was used for prediction.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

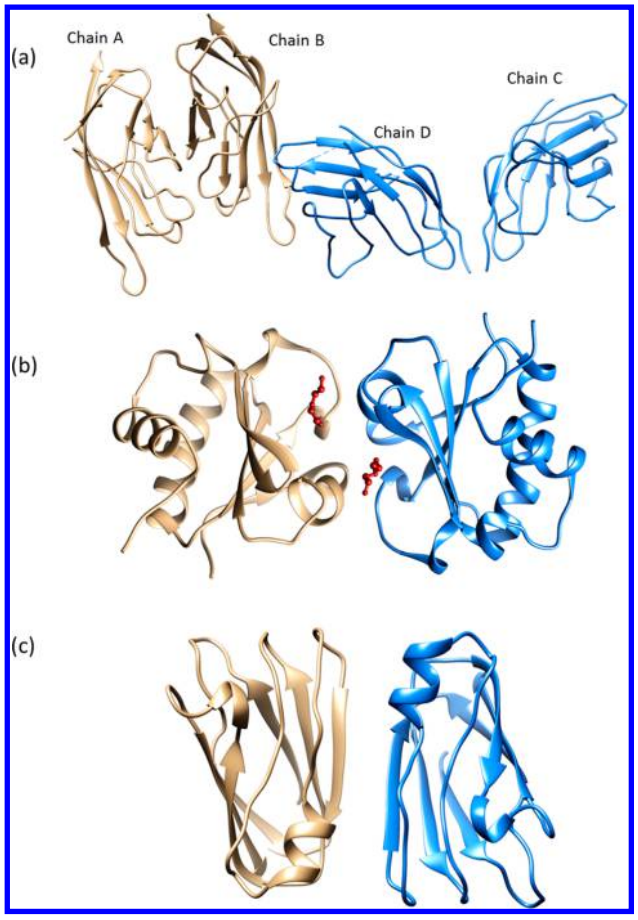


Figure 6. Typical examples of Dockground biological assemblies predicted crystallographic by IChemPIC. (A) CTLA-4 (chains A, B) /B7-2 (chains C, D) complex (PDB ID 1i85); (B) human phosducin-like protein 2 with bound PEG molecules (red ball and sticks) at the interface (PDB ID 3evi); (C) plastocyanin from the cyanobacterium Synechocystis sp. PCC 6803 (PDB ID 1pcs).

D) is predicted to be crystallographic (PDB ID 1I85; interface = 621 Å^2, nPTS = 42) since it exists only within a larger network (Figure 6A), explaining the periodic organization of these molecules within the immunological synapse at the cell surface.[39] Second, many of the small-sized interfaces (149 are smaller than 500 Å^2) are a clear consequence of the crystallization conditions, with either a salt or a precipitant-facilitating molecule at the interface. This case is nicely exemplified by the X-ray structure of human phosducin-like protein 2[40] (PDB ID 3EVI; interface = 422 Å^2, nPTS = 21), which presents two diethylene glycol molecules stabilizing an artifactual homodimeric interface (Figure 6B). Lastly, strong crystal packing may produce artificial interfaces, as shown here by the predicted biological assembly of a cyanobacterium plastocyanin (Figure 6C) with a perfect C_2 symmetry (PDB ID 1PCS; interface = 395 Å^2, nPTS = 6) but no biological relevance.[41]

From the present exercise, we estimate that ca. 15% of PDB biological units have a proposed oligomeric state that is likely to be biologically irrelevant. We therefore strongly suggest the usage of an accurate classifier like IChemPIC to reduce the number of such erroneous biological assemblies and enable the design of PPI inhibitors on biologically relevant interfaces.

CONCLUSIONS

We present a novel computational approach (IChemPIC) to distinguish between biologically relevant and crystallographic protein−protein interfaces. Since none of the existing benchmark datasets are satisfactory at this, notably for predicting small-sized ligandable biological interfaces, novel training and external test sets (FDS set) were defined and manually curated to afford (i) a comparable coverage of interface areas for existing crystallographic and biological interfaces and (ii) an application to small-sized protein−protein interfaces known to be modulated by low molecular weight drug-like compounds.

By describing the interface with a simple vector of 45 real numbers focusing on intermolecular interactions, machine learning methods can be used to classify interfaces as either crystallographic or biological. Due to its simplicity and low parametrization level, the Random Forest machine learning method was chosen to derive a model that distinguishes crystallographic from biological interfaces with a robust accuracy close to 80%. With respect to current state-of-the art methods, IChemPIC is the only approach able to predict with the same good accuracy the two categories of protein−protein interfaces, whatever the external test set. There are many advantages of using IChemPIC with respect to other methods: (i) the implicit inclusion of hydrogen atoms allows for using hydrogen bonds as descriptors for model development; (ii) the method can be applied to interfaces presenting post-translational modifications; (iii) the performance is independent of the size of the interface; and (iv) the applicability domain is large, ranging from small biological protein−protein interfaces (500 Å^2) to larger crystallographic contact (1500 Å^2).

We should acknowledge, however, that IChemPIC is currently parametrized to treat interfaces between two protein chains. For example, the three possible interfaces (AB, BC, AC) of an ABC heterotrimer will be predicted to be either crystallographic or biological, but no prediction will be made for interfaces between one chain and the two others. In other words, no prediction is made for the entire assembly, as in PISA, for example. This drawback explains some of the false negatives observed by IChemPIC and could be easily corrected by enabling the detection of all possible interactions between a single chain and its full protein environment. However, since our method is primarily aimed at further detecting all PDB biologically relevant interfaces amenable to small molecule disruption or stabilization, we prefer to restrict IChemPIC to treat only two-chain interfaces in order to exactly localize the interface to be targeted by a potential PPI modulator. IChemPIC can be used online (http://bioinfo-pharma.u-strasbg.fr/IChemPIC) starting from either a PDB identifier or a user-provided PDB input file.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00190.

Set of 200 crystallographic interfaces (FDS dataset); set of 200 biological interfaces (FDS dataset); descriptors of protein−protein interfaces; prediction of 100 protein−protein interfaces (FDS external set) by various methods; prediction of 142 protein−protein interfaces (Bahadur external set) by various methods; prediction of 143 protein−protein interfaces (Ponstingl external set) by

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

various methods; prediction of 66 protein−protein interfaces (IPAC validation set 3) by various methods; list of Dockground interfaces predicted by PPI-Ichem (PDF).

AUTHOR INFORMATION

Corresponding Author

*Phone: +33 3 68 85 42 35. Fax: +33 3 68 85 43 10. E-mail: rognan@unistra.fr.

Present Address

†(J.D.) Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the National Center for Scientific Research (CNRS, Institut de Chimie) and the Alsace Region for the doctoral fellowship to F.D.S. The High-Performance Computing Center (University of Strasbourg, France) and the Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) are acknowledged for allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg, Germany) for providing an executable version of Protoss. O. Sperandio, X. Morelli, P. Roche, and E. Kellenberger are acknowledged for critical reading of the manuscript and helpful discussions.

REFERENCES

(1) Ivanov, A. A.; Khuri, F. R.; Fu, H. Targeting Protein-Protein Interactions as an Anticancer Strategy. *Trends Pharmacol. Sci.* **2013**, *34*, 393−400.

(2) Villoutreix, B. O.; Kuenemann, M. A.; Poyet, J.-L.; Bruzzoni-Giovanelli, H.; Labbé, C.; Lagorce, D.; Sperandio, O.; Miteva, M. A. Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology. *Mol. Inf.* **2014**, *33*, 414−437.

(3) Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **2007**, *450*, 1001−1009.

(4) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2012**, *33*, 224−232.

(5) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic acids research* **2000**, *28*, 235−242.

(6) Janin, J. Specific Versus Non-Specific Contacts in Protein Crystals. *Nat. Struct. Biol.* **1997**, *4*, 973−974.

(7) Mitra, P.; Pal, D. Combining Bayes Classification and Point Group Symmetry under Boolean Framework for Enhanced Protein Quaternary Structure Inference. *Structure* **2011**, *19*, 304−312.

(8) Henrick, K.; Thornton, J. M. PQS: A Protein Quaternary Structure File Server. *Trends Biochem. Sci.* **1998**, *23*, 358−361.

(9) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.

(10) Ponstingl, H.; Henrick, K.; Thornton, J. M. Discriminating between Homodimeric and Monomeric Proteins in the Crystalline State. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 47−57.

(11) Ponstingl, H.; Kabir, T.; Thornton, J. M. Discriminating Between Homodimeric and Monomeric Proteins in the Cristalline State. *J. Appl. Crystallogr.* **2003**, *36*, 1116−1122.

(12) Bernauer, J.; Bahadur, R. P.; Rodier, F.; Janin, J.; Poupon, A. DiMoVo: A Voronoi Tessellation-Based Method For Discriminating Crystallographic and Biological Protein-Protein Interactions. *Bioinformatics* **2008**, *24*, 652−658.

(13) Zhu, H.; Domingues, F. S.; Sommer, I.; Lengauer, T. Noxclass: Prediction of Protein-Protein Interaction Types. *BMC Bioinf.* **2006**, *7*, 27.

(14) Liu, Q.; Kwoh, C. K.; Li, J. Binding Affinity Prediction for Protein-Ligand Complexes Based on Beta Contacts and B Factor. *J. Chem. Inf. Model.* **2013**, *53*, 3076−3085.

(15) Tsuchiya, Y.; Kinoshita, K.; Naiamura, H. Analyses of Homo-Oligomer Interfaces of Proteins from the Complementarity of Molecular Surface, Electrostatic Potential and Hydrophobicity. *Protein Eng., Des. Sel.* **2006**, *19*, 421−429.

(16) Block, P.; Paern, J.; Hullermeier, E.; Sanschagrin, P.; Sotriffer, C. A.; Klebe, G. Physicochemical Descriptors to Discriminate Protein-Protein Interactions in Permanent and Transient Complexes Selected by Means of Machine Learning Algorithms. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 607−622.

(17) Mintseris, J.; Weng, Z. Atomic Contact Vectors in Protein-Protein Recognition. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 629−639.

(18) Valdar, W. S.; Thornton, J. M. Protein-Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 108−124.

(19) Elcock, A. H.; McCammon, J. A. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 2990−2994.

(20) Duarte, J. M.; Srebniak, A.; Scharer, M. A.; Capitani, G. Protein Interface Classification by Evolutionary Analysis. *BMC Bioinf.* **2012**, *13*, 334.

(21) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A Dissection of Specific and Non-Specific Protein-Protein Interfaces. *J. Mol. Biol.* **2004**, *336*, 943−955.

(22) Chakrabarti, P.; Janin, J. Dissecting Protein-Protein Recognition Sites. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 334−343.

(23) Bourgeas, R.; Basse, M. J.; Morelli, X.; Roche, P. Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2p2i Database. *PLoS One* **2010**, *5*, e9598.

(24) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. A., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, W.; Wolf, R. M.; Wu, X.; Kollman, P. A. *Amber*, version 14; University of California: San Francisco, CA. http://ambermd.org/.

(25) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305−320.

(26) Redundancy in the Protein Data Bank. http://www.rcsb.org/pdb/statistics/clusterStatistics.do.

(27) Rognan, D. Rational Design of Protein-Protein Interaction Inhibitors. *MedChemComm* **2015**, *6*, 51−60.

(28) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. A Structure-Based Benchmark for Protein-Protein Binding Affinity. *Protein Sci.* **2011**, *20*, 482−491.

(29) Gavenonis, J.; Sheneman, B. A.; Siegert, T. R.; Eshelman, M. R.; Kritzer, J. A. Comprehensive Analysis of Loops at Protein-Protein Interfaces for Macrocycle Design. *Nat. Chem. Biol.* **2014**, *10*, 716−722.

(30) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.

(31) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623−637.

(32) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287−2299.

(33) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R news* **2002**, *2*, 18−22.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014

(34) *R: A language and environment for statistical computing*, version 3.2.0; R Foundation for Statistical Computing: Vienna, Austria. http://www.r-project.org/.
(35) Chang, Y. S.; Graves, B.; Guerlavais, V.; Tovar, C.; Packman, K.; To, K. H.; Olson, K. A.; Kesavan, K.; Gangurde, P.; Mukherjee, A.; Baker, T.; Darlak, K.; Elkin, C.; Filipovic, Z.; Qureshi, F. Z.; Cai, H.; Berry, P.; Feyfant, E.; Shi, X. E.; Horstick, J.; Annis, D. A.; Manning, A. M.; Fotouhi, N.; Nash, H.; Vassilev, L. T.; Sawyer, T. K. Stapled Alpha-Helical Peptide Drug Development: A Potent Dual Inhibitor of Mdm2 and Mdmx for P53-Dependent Cancer Therapy. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E3445−3454.
(36) Breiman, L. Random Forests. *Mach Learn* **2001**, *45*, 5−32.
(37) Touw, W. G.; Bayjanov, J. R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; van Hijum, S. A. F. T. Data Mining In the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle? *Briefings Bioinf.* **2013**, *14*, 315−326.
(38) Douguet, D.; Chen, H. C.; Tovchigrechko, A.; Vakser, I. A. DOCKGROUND Resource for Studying Protein-Protein Interfaces. *Bioinformatics* **2006**, *22*, 2612−2618.
(39) Schwartz, J. C.; Zhang, X.; Fedorov, A. A.; Nathenson, S. G.; Almo, S. C. Structural Basis for Co-Stimulation by The Human Ctla-4/B7−2 Complex. *Nature* **2001**, *410*, 604−608.
(40) Lou, X.; Bao, R.; Zhou, C. Z.; Chen, Y. Structure of the Thioredoxin-Fold Domain of Human Phosducin-Like Protein 2. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2009**, *65*, 67−70.
(41) Romero, A.; De la Cerda, B.; Varela, P. F.; Navarro, J. A.; Hervas, M.; De la Rosa, M. A. The 2.15 Å Crystal Structure of a Triple Mutant Plastocyanin from the Cyanobacterium Synechocystis sp. pcc 6803. *J. Mol. Biol.* **1998**, *275*, 327−336.

DOI: 10.1021/acs.jcim.5b00190
J. Chem. Inf. Model. 2015, 55, 2005−2014