Received: 17 January 2019 | Revised: 9 May 2019 | Accepted: 7 June 2019

DOI: 10.1002/prot.25757

RESEARCH ARTICLE

PROTEINS WILEY

Decrypting protein surfaces by combining evolution, geometry, and molecular docking

Chloé Dequeker[1] | Elodie Laine[1] | Alessandra Carbone[1,2]

[1]Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Paris, France

[2]Institut Universitaire de France (IUF), Paris, France

Correspondence
Elodie Laine, Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Paris, France.
Email: elodie.laine@upmc.fr
Alessandra Carbone, Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Paris, France and Institut Universitaire de France (IUF), Paris, France.
Email: alessandra.carbone@lip6.fr

Funding information
Institut Universitaire de France; Secrétariat Général pour l'Investissement, Grant/Award Number: Equip@Meso project - ANR-10-9 EQPX- 29-01; Secrétariat général pour l'investissement - Programme d'investissements d'avenir (PIA), Grant/Award Number: MAPPING - ANR-11-BINF-0003

Abstract

The growing body of experimental and computational data describing how proteins interact with each other has emphasized the multiplicity of protein interactions and the complexity underlying protein surface usage and deformability. In this work, we propose new concepts and methods toward deciphering such complexity. We introduce the notion of interacting region to account for the multiple usage of a protein's surface residues by several partners and for the variability of protein interfaces coming from molecular flexibility. We predict interacting patches by crossing evolutionary, physicochemical and geometrical properties of the protein surface with information coming from complete cross-docking (CC-D) simulations. We show that our predictions match well interacting regions and that the different sources of information are complementary. We further propose an indicator of whether a protein has a few or many partners. Our prediction strategies are implemented in the dynJET^2 algorithm and assessed on a new dataset of 262 protein on which we performed CC-D. The code and the data are available at: http://www.lcqb.upmc.fr/dynJET2/.

KEYWORDS
binding site, complete cross-docking, evolutionary conservation, interface prediction, protein-protein interaction

1 | INTRODUCTION

Proteins are main actors in biological processes and a detailed description of their interactions is expected to provide direct information on these processes and on the way to interfere with them.[1] Our knowledge of protein-protein interaction (PPI) networks[2] is largely incomplete, since the experimental assessment of all possible interactions of a protein is very challenging.[3,4] To overcome this limitation, recent efforts have been invested in the integration of direct and indirect experimental evidence and of computational predictions to better describe PPIs at the genome scale [5-11,40]. These efforts have revealed the complexity and multiplicity of PPIs. A protein may interact with several partners at the same time—each partner binding to a different site at its surface, or its surface may present a shared binding region that will be used by different partners at different moments of its lifetime. It is estimated that as much as 75% of the surface could potentially be used for PPIs.[12] In this context, there is a need for the development of tools able to decrypt protein surfaces at the residue level. A comprehensive description of protein surfaces and a precise identification of the residues involved in interactions are mandatory to identify cellular partners at large scale[11] and design drugs modulating PPIs.[13] Moreover, characterizing protein surfaces' properties may inform us on the number of partners a protein may have, and thus on the role of that protein in the cell.

Evolutionary, physicochemical, and geometrical properties have been shown to be relevant to PPIs,[14-25] and, based on them, in the past 20 years, a number of tools have been developed to predict interacting surfaces[24,26-33] (see[25,34] for surveys). Although some of these tools achieve very high accuracy against subsets of known

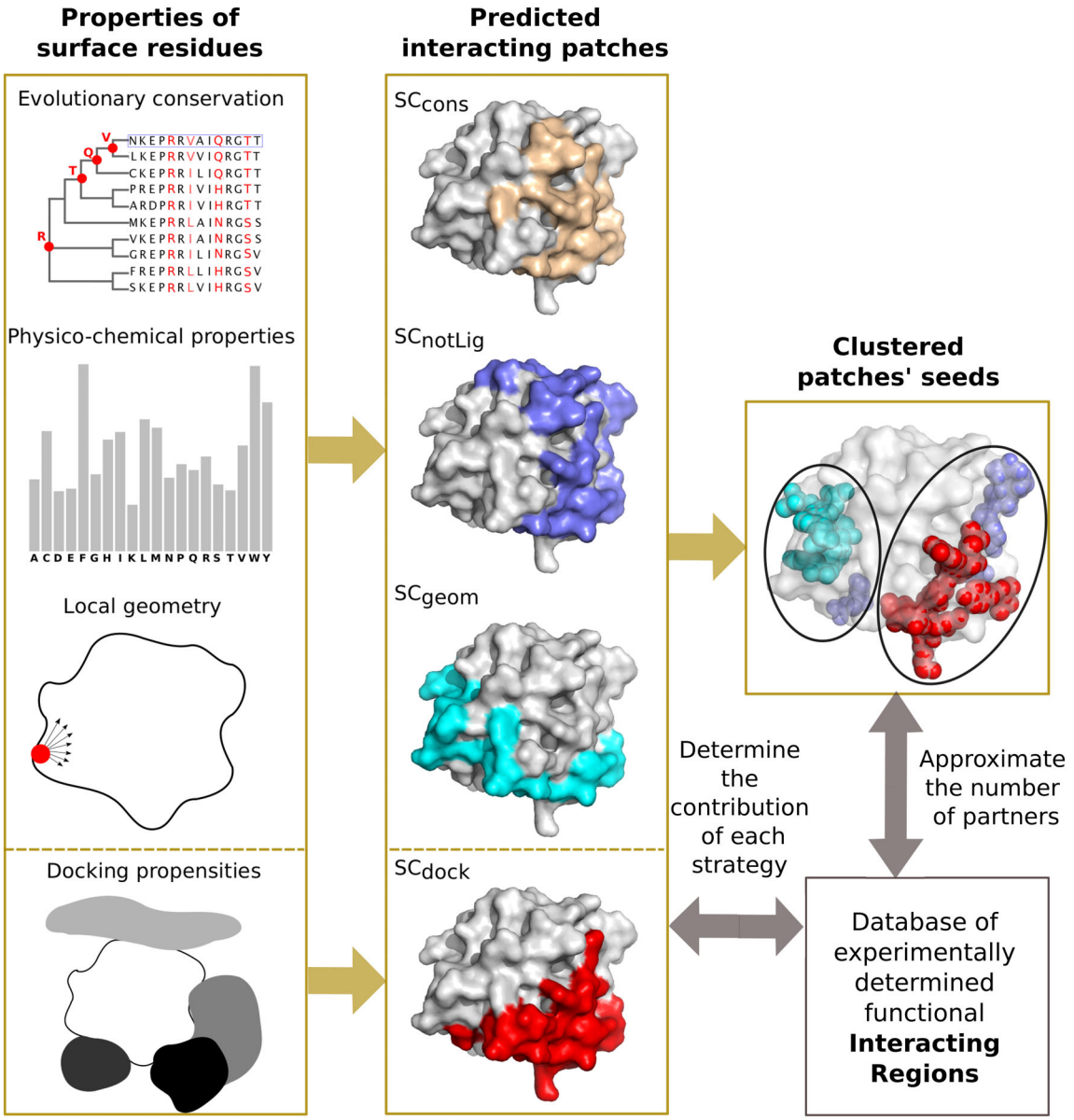

FIGURE 1 Schematic representation of our workflow. We consider four residue-based properties (left panel), namely evolutionary conservation, amino acid propensities to be found at an interface, local geometry, and propensities to be found in docked interfaces. We predict interacting patches at the surface of proteins by using four different strategies: SC_{cons}, SC_{notLig}, and SC_{geom} combines the first three properties, while SC_{dock} relies exclusively on the fourth property. We compare the predicted patches with a set of experimentally determined functional interacting regions. We analyze and cluster the predicted patches' seeds, from which they were grown, to precisely localize interacting regions and infer the number of partners used by each region [Color figure can be viewed at wileyonlinelibrary.com]

experimental binding sites, their predictions are generally much smaller than the expected interacting surface size.[12] Moreover, many tools do not propose sites but rather evaluate the probability of a residue to be involved in interactions. An orthogonal approach consists in exploiting molecular docking calculations. Docking methods were originally designed to predict the structure of a complex starting from the known structures of its components. Candidate conformations are evaluated based on properties reflecting the strength of the association, for example, shape complementarity, electrostatics, desolvation, and conformational entropy. By deriving statistics from the generated conformational ensemble, one can estimate the propensity of each protein surface residue to be found at a docked interface and use these propensities to identify binding sites.[35] This has been realized in single docking studies,[36-40] where two proteins known to interact are docked to each other, in arbitrary docking studies,[41] where proteins from a benchmark set are docked to arbitrarily chosen proteins, and in complete cross-docking (CC-D) studies,[11,42-45] where all vs all docking is realized on a given dataset.

In the present study, we combine these different types of information to decipher the complexity of protein surfaces and give clues about the many interactions a protein may have (Figure 1). Given a protein, we predict patches at its surface based on some intrinsic properties of that surface and on properties inferred from the behavior of the protein with respect to others in docking calculations. We

assess our predictions against a new set of experimentally known functional interfaces, detected at the surface of 262 proteins and of their close homologs. We demonstrate that considering only one single complex for a given protein leads to underestimate the proportion of its surface involved in functional interactions and to the incorrect assessment of protein interface prediction algorithms. To cope with this issue, we introduce the new concept of interacting region (IR) as a protein surface region used by one or several partners. IRs are defined by merging overlapping interacting sites (IS) extracted from different protein complex structures. We show that our predictions better match IRs compared to ISs and capture the interface variability induced by molecular flexibility. Our approach includes sequence-based analysis, which allows the detection of signals even when the interface is "hidden." Interestingly, we highlight a few cases where docking enables unveiling interfaces that could not be detected otherwise. We further exploit the way in which our predicted patches are grown, starting from a seed that is progressively extended. Specifically, we demonstrate that predicted patches' seeds can be used to localize IRs with high precision and to determine whether a protein has a few or many partners.

We provide sets of experimentally known interaction sites and regions and CC-D results for our dataset of 262 proteins, along with a computational tool, called dynJET[2], for predicting interacting patches based either on protein sequence and structure analysis or on any pre-computed residue based property. All data and implemented code are available at: http://www.lcqb.upmc.fr/dynJET2/.

2 | MATERIALS AND METHODS

2.1 | Datasets

2.1.1 | Proteins: P-262

We defined a dataset of 262 protein chains and associated Protein Data Bank (PDB) structures featuring both single and multiple partners interactions. This dataset was extracted from a larger set of 2246 protein chains defined in the scope of the HCMD2 project (see http://www.ihes.fr/ carbone/HCMDproject.htm), for which we performed CC-D. We considered the subset of PDB structures comprising a protein complex previously reported in Reference 45, from which we excluded: (a) only C-α structures, (b) chains for which docking results were missing, (c) chains forming coiled-coils complexes, (d) deprecated PDB codes, (e) chains for which no biologically relevant interface (see subsection *Surfaces and interfaces* for definition) could be found in the whole PDB (considering 90% sequence identity, see below). The remaining 262 protein chains comprise on average 200.5 ± 131.2 residues (Table S1). This indicates a large variation of protein size inside the dataset (21 residues for the smallest protein vs 789 residues for the largest one). Based on the information recovered from the PDB, the proteins were manually classified, following and extending the classification proposed in Reference 46. We defined 11 functional classes: 16 bound antibodies (AB), 25 complex subunits (C), 60 enzymes (E), 10 enzyme regulators (ER), 9 G proteins (G), 6 antigens from the immune system (I), 23 receptors (R), 24 structural proteins (S), 16 substrates/inhibitors (SI), 7 transcription factors (TF) and 66 proteins with other function (O).

2.1.2 | Protein interfaces: PPI-262 and $\text{PPI-262}_{\text{ext}}$

We defined two datasets of experimental protein-protein interfaces, namely PPI-262 and $\text{PPI-262}_{\text{ext}}$ (Figure S1). Both datasets comprise only interfaces buried within "biological units" or "biological assemblies," as annotated by the authors of the PDB structure or by PISA software.[47] This ensures that the interfaces we consider carry a biological meaning. PPI-262 comprises 329 ISs (see definition below, in subsection *Surfaces and interfaces*) extracted from the PDB files associated to P-262 and $\text{PPI-262}_{\text{ext}}$ comprises 370 IRs (see definition below, in subsection *Surfaces and interfaces*) defined from PDB files of close homologs of the proteins from P-262.

To construct $\text{PPI-262}_{\text{ext}}$ (see Figure S1), we first searched for homologs of the 262 proteins from P-262 in the PDB. We downloaded the pre-computed set of PDB structures clustered at 90% sequence identity from ftp://resources.rcsb.org/sequence/clusters/. This set was determined using BLASTClust with the arguments -p T -b T -S 90. We then filtered out structures with a resolution poorer than 5 Å resolution. 23 642 functional ISs were detected on these structures and were then mapped onto the query proteins from P-262 by performing global pairwise sequence alignment (using the blosum62 matrix, with the Biopython package[48]). ISs were then merged into IRs.

2.2 | Complete cross-docking

Given an ensemble of proteins, CC-D consists in docking each protein against all others in the dataset, including itself. CC-D was performed on P-262 using the MAXDo (Molecular Association via Cross Docking) rigid-body coarse-grained docking program.[42] Statistics were computed from the generated conformations (docking poses) to determine the propensity of each residue from each protein to be found in a docked interface. We define the interface propensity (IP) of residue i, belonging to protein P, as[11,42]:

$$IP_P(i) = \frac{N_{int,P}(i)}{N_{pos,P}} \quad (1)$$

where $N_{pos,P}$ is the total number of docking poses considered for protein P and $N_{int,P}(i)$ is the number of docking poses where residue i lies in the interface. In order to limit the number of docked interfaces to reconstruct, which is the most time-consuming part of the analysis, we considered only the lowest energy docking poses (less than 2.7 kcal from the best-scored pose, as described in Reference 11). This led to ~50 000 docking poses for each protein pair. Thus, for a given protein P, we considered about 50 000 × 262 = 13 100 000 docking poses.

Below, we shall use $IP_P(i)$ values to formally define a normalized interaction propensity score, called NIP, that dynJET[2] uses in order to predict interface sites.

2.3 | Residue-based properties

Four measures, T_{JET}, PC, CV, and NIP, are used to evaluate single residues in a protein and to define scores for the prediction of protein interfaces.

T_{JET} reflects the evolutionary conservation level of a residue, and is computed from phylogenetic trees constructed by using sequences, homologous to a query sequence and sampled by a Gibbs-like approach.[23] The Gibbs-like approach extracts N representative subsets of N sequences[23] in a way that, within each subset, the proportions of sequences sharing [20-39]%, [40-59]%, [60-79]%, and [80-98]% sequence identity with the query sequence are similar (ideally, about one quarter for each group of identity). Sequences in a subset are then aligned using CLUSTALW2[49] and a distance tree is constructed from the alignment based on the Neighbor Joining algorithm.[50] From each tree T, a *tree trace level* is computed for each position in the query sequence: it corresponds to the level n in the tree T where the amino acid at this position appeared and remained conserved thereafter (see Reference 23 for a more precise definition). Let us recall that this definition of evolutionary trace is notably different from the measure defined in References 14,51 to rank protein residues.

Then, tree trace levels are averaged over the N trees to get statistically significant values, which we denote *relative trace significances*, or T_{JET}, and which are calculated as follows[23]:

$$T_{JET}(j) = \frac{1}{M_j}\sum_{t=1}^{M_j}\frac{L_t - l_j^t}{L_t} \quad (2)$$

where l_j^t is the tree trace level of residue r_j in tree t, L_t is the maximum level of t and M_j is the number of trees where a nonzero tree trace level was computed for r_j. T_{JET} values vary in the interval [0,1] and represent averages, over all trees of residues, of evolutionary conservation levels.

PC indicates the physicochemical propensity specific to amino acids located at a protein interface. The original values, taken from,[52] range from 0 to 2.21 and are scaled here between 0 and 1 for the calculation of residue scores.

CV is the circular variance, a measure of the density of protein around a residue. Formally, the circular variance of a fixed point in 3D space is computed from the vectorial distribution of a set of neighboring points around it.[53] Specifically, the CV value of an atom i is expressed as:

$$CV(i) = 1 - \frac{1}{n_i}\left|\sum_{j\neq i, r_i \leq r_c}\frac{\vec{r_{ij}}}{||\vec{r_{ij}}||}\right| \quad (3)$$

where n_i is the number of atoms distant by less than r_c Å from atom i and $\vec{r_{ij}}$ is the coordinate vector from atom i to its neighbor j. If atom i is buried within the protein, then the resultant of the vectors toward its neighbors will be small and its CV value will be close to one. On the contrary, if the atom is located in a protruding region, the vectors toward its neighbors in the protein will share the same direction, their resultant will be high, and hence the CV value will be close to zero. We compute the CV value of a residue a_k as the average of the atomic CVs, over all atoms of a_k: $CV(a_k) = \frac{1}{N_{a_k}}\sum_i CV(i)$, where N_{ak} is the number of atoms in a_k. Compared to solvent accessibility, CV changes more smoothly from the surface to the interior of the protein,[54] and is thus less sensitive to small conformational changes. CV values are scaled between 0 (most protruding residues) and 1 (least protruding residues) for the calculation of residue scores.

NIP is the normalized form of the Interface Propensity score *IP*, defined in Equation (1), that reflects the propensity of a residue to be found at the interface. In order to compare *IP* scores among proteins, we normalize it, as done in Reference 11: a positive NIP value indicates that the residue i is favored to occur at potential binding sites, and a negative NIP value indicates that it is disfavoured. NIP is defined as:

$$NIP_P(i) = \frac{IP_P(i) - < IP_P(j) >_{j\in P}}{\max(IP_P(j))_{j\in P} - < IP_P(j) >_{j\in P}} \quad (4)$$

where $<IP_P(j)>_{j\in P}$ and $\max(IP_P[j])_{j\in P}$ are the average *IP* and the maximum *IP*, respectively, computed over all the residues j in P. The NIP value represents how often a residue is docked on the retained conformations (ie, those conformations that have less than 2.7 *kcal/mol* energy difference from the best one, as explained above).

These four residue-based properties were previously shown to be useful for the prediction of protein interfaces.[11,23,24,33,41,43,44] T_{JET}, PC, CV are computed using dynJET^2, a modified version of JET^2[24] that handles NIP values, as described below. For each measure, values are scaled between 0 and 1.

2.4 | Surfaces and interfaces

A residue is considered to be at the surface of the protein if it displays at least 5% of relative accessible surface area (rasa), as computed by Naccess.[55] Experimental and predicted interfaces are exclusively comprised of surface residues.

Experimental interfaces are detected on known PDB complex structures, considering only biological assemblies. We define two types of interfaces, namely *interacting sites* (ISs) and *interacting regions* (IRs). ISs are detected in single PDB structures by using the INTBuilder software[56] (www.lcqb.upmc.fr/INTBuilder/) with a distance threshold of 5 Å. They may represent single- or multiple-partner interactions (Figure 2A). For instance, let us consider a ternary complex comprised of proteins P_1, P_2, and P_3. If both P_2 and P_3 bind to P_1 but are not in contact with each other, then we will define two single-partner ISs at the surface of P_1 (Figure 2A, middle panel). By contrast, if P_2 and P_3 are in contact with each other (less than 5 Å away), then we will define one multi-partner IS at the surface of A (Figure 2A, right panel). IRs are defined by merging several ISs. Two ISs, namely IS_1 and IS_2, of reasonable sizes (more than five residues), will be merged into an IR if their maximum overlap with respect to their respective sizes (max [overlap(IS_1, IS_2), overlap(IS_2, IS_1)]) is greater than an arbitrarily chosen threshold of 60%. In practice, this threshold gave us the most realistic IRs given the experimental information. This merging criterion is

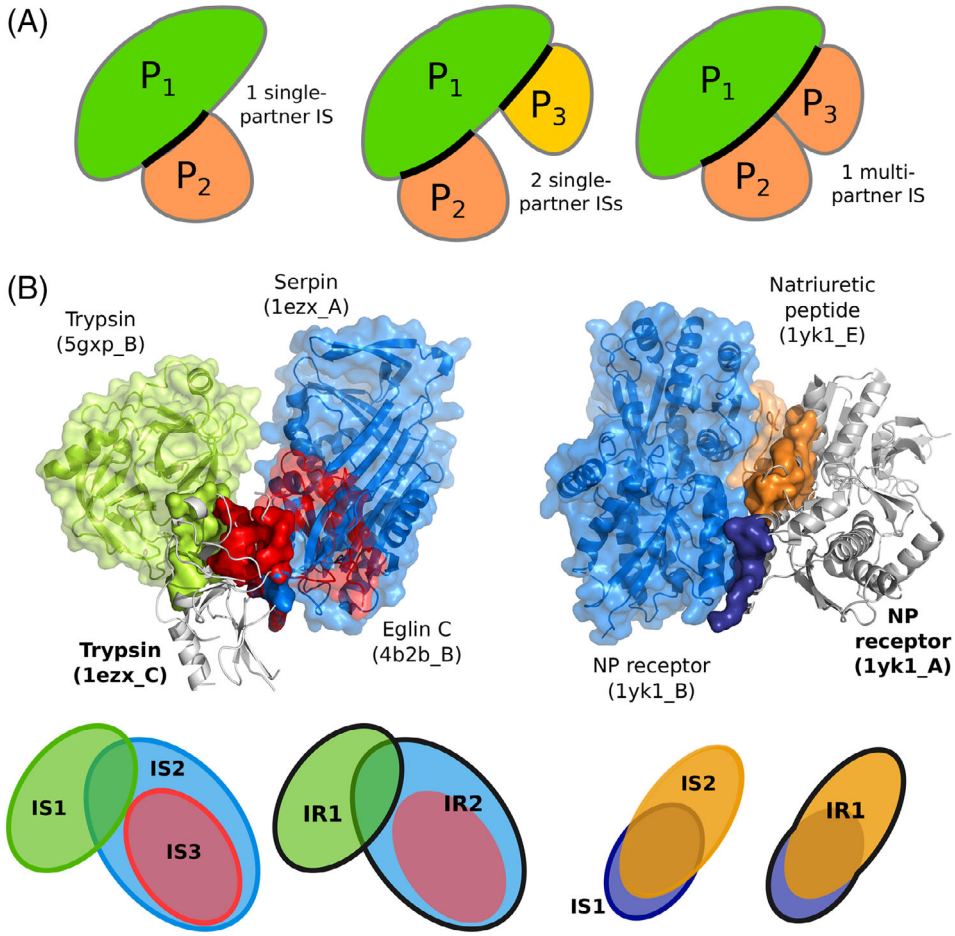

FIGURE 2 Examples and schema illustrating the notions of interacting site and interacting region. A, Schematic representation of single- and multiple-partner interacting sites. Three proteins are considered, namely P_1, P_2, and P_3. The ISs defined on P_1 are highlighted by thick black lines. See the materials and methods section for a precise definition of the sites issued from multiple partners. B, Two examples of the usage of the protein surface by different partners. The query proteins are displayed as gray cartoons, their interacting sites as opaque colored surfaces, and their partners as colored cartoons and transparent surfaces. Left: trypsin (1ezx_C, in gray) interacts with itself (5gxp_B, in green), serpin (1ezx_A, in blue) and eglin C (4b2b_B, in red). The three corresponding ISs lead to the definition of 2 IRs, as depicted on the schema at the bottom, where each IR is contoured by a thick line. Right: the natriuretic peptide receptor forms a homodimer (1yk1_A, in gray, and 1yk1_B, in blue) to bind its substrate (1yk1_E, in orange). The 2 ISs detected at the surface of one receptor monomer (1yk1_A, in gray) are merged into an IR

relaxed when dealing with very small ISs. Namely, we request that an IS comprising at most five residues shares at least one residue in common with another IS to be merged with it. The merging procedure is iterated over all ISs for a given protein. To construct **PPI-262** and **PPI-262**$_{ext}$, we retained only ISs and IRs of reasonable sizes, that is, comprising more than five residues.

Predicted interfaces are identified by the dynJET^2 software (www.lcqb.upmc.fr/dynJET2/). Given the sequence and the structure of a query protein, dynJET^2 predicts the location of potential protein binding sites on the protein surface.[24] dynJET^2 implements a clustering algorithm and scoring strategies specifically aimed at detecting the different layers of a protein interface, namely the *support*, the *core* and the *rim*.[57] These three layers are defined for known experimental interfaces by comparing their solvent accessibilities in the presence and absence of the partner.[24] Support residues are buried with and without the partner, core residues become buried upon binding to the partner and rim residues are exposed in the presence and absence of the partner. A threshold of 25% relative solvent accessibility is used to determine whether a residue is buried or not. To approximate these layers, our algorithm[24] first identifies a small cluster of highly scored residues, called the seed. Seeds closer than 5 Å are merged. Then, the detected seeds are progressively extended, and the resulting residue clusters are merged if they are in contact (< 5 Å away). Importantly, the way residues are picked up based on their scores (defined below) differs between seed and extension, such that the detected signal is very strong in the seed and progressively fades away as the extension is grown. Finally, an outer layer is added to form what we call a *predicted patch*. We used the iterative mode of dynJET^2 (i-dynJET^2) and considered a residue to be predicted as interacting if it was detected at least twice over 10 iterations (as done in Reference 24).

Four scoring schemes or strategies are implemented in dynJET^2 (compared to three in JET^{224}):

SC_{cons} targets very conserved residues (identified by the T_{JET} score) to form a seed which is then extended using both T_{JET} and PC scores. An outer layer is added considering both PC and CV scores. SC_{cons} is intended to detect diverse protein binding sites.

SC_{notLig} detects both seed and extension layers using a combination of T_{JET} and CV scores. It aims at detecting highly conserved residues that are not buried too deeply beneath the surface of the protein. The outer layer is defined based on PC and CV scores, as in SC_{cons}. SC_{notLig} specifically distinguishes protein interfaces from small ligand binding sites.

SC_{geom} disregards evolutionary information and solely employs PC and CV scores for detecting all three layers of the interface. SC_{geom} yields consistent predictions for interfaces displaying very low conservation signal, eg, antigen binding sites.

SC_{dock} applies the NIP score of the residues to all three layers (core, extension, and outer layer). The usage of NIP is motivated by the observation that proteins tend to dock to their cognate partners and also to noninteractors via the same region at their surface.[11,41-43] This scoring scheme is new in dynJET^2.

To evaluate the performance of our predictions, we mainly relied on the F1-score, which is the harmonic mean of precision and recall. Predicted patches were compared to ISs from **PPI-262** and IRs from $\textbf{PPI-262}_{\text{ext}}$. The union of all predicted interface residues was also compared to the union of all experimental interface residues, for each protein.

2.5 | Seeds clustering

Seeds generated by dynJET^2's different scoring schemes were collected. The SC_{cons} seeds were discarded because almost half of their residues were shared with the SC_{notLig} seeds (Figure S3a), they were bigger than the other seeds (Figure S3b) and we observed that they often extended over several other seeds. We considered that these characteristics would make SC_{cons} seeds perform badly in locating different IRs. The atoms belonging to the seeds collected from SC_{notLig}, SC_{geom}, and SC_{dock} were then classified by applying hierarchical clustering using the average linkage method. A threshold distance of 23 Å was used to define the clusters. This value yielded the best match between the number of clustered seeds and the number of IRs.

2.6 | Conformational variability of IRs

For each IR, the RMSD of its backbone atoms (or, if not possible, its C-α atoms) was computed between the query structure from **P-262** and each of the homologous structures on which the IR was detected. For each homologous structure, only the subset of residues detected on this structure were considered to compute the RMSD. RMSD values were then averaged over the homologous structures (including the query structure if the IR was also detected on it). This gives us a single RMSD value for each IR.

2.7 | Number of partners

To count how many different partners a protein has, we considered all known homologs of the protein in the PDB and their partners. We clustered the partners depending on their sequence homology: two partners were classified in different clusters if they shared less than 90% sequence identity. This threshold in agreement with the criteria we applied to protein chains and their homologs. The number of clusters provides an estimation of the number of partners for the protein.

2.8 | Comparison with other methods

SPPIDER[58] (accessed at http://sppider.cchmc.org) was applied on the 262 protein chains from **P-262**. Multi-VORFFIP[59] (accessed at www.bioinsilico.org/cgi-bin/SUPER_VORFFI/htmlVORFFI/home) was applied on a subset of 252 protein chains. The 10 proteins that were discarded for the analysis either belonged to the training set of Multi-VORFFIP or produced an error when running the tool. For both tools, we considered residues displaying a probability above 0.5 as predicted to interact. Residues separated by less than 5 Å were subsequently clustered to form predicted patches.

3 | RESULTS

3.1 | From interacting sites to interacting regions

Our analyses were performed on a set of protein complex crystallographic structures, which we call **P-262**, involving 262 protein chains (see Materials and Methods and Table S1). From these experimental structures, we defined two sets of functional interfaces, **PPI-262** and $\textbf{PPI-262}_{\text{ext}}$ (see Materials and Methods and Figure S1). **PPI-262** comprises 329 ISs, where each IS corresponds to one functional interaction described by one structure. This classical definition of a protein interacting site is very restrictive and does not account for the interface variability that may come from structure ensembles. Indeed, the definition of the interface between two given proteins may vary from one structure to another, depending on the crystallization conditions, on the quality of the data/model and/or on the inherent flexibility of the assembly. What is more, the notion of IS masks the complexity of protein surface usage by multiple partners. This motivated us to define the new concept of IR, obtained by merging overlapping ISs (≥ 60% overlap). Based on the observation that functional interfaces are conserved across closely related homologs,[60] we collected all functional ISs involving the query proteins from **P-262** or their close homologs (≥ 90% sequence identity) from the PDB.[61] This amounted to 23 642 ISs, which were merged into 370 IRs to define our second "extended" dataset, $\textbf{PPI-262}_{\text{ext}}$. The two examples in Figure 2B illustrate the complexity of the experimental interaction surfaces in our

datasets. Binding sites may be disjoint, overlapping or included in others (Figure 2B, on the left), and they may be defined by the interaction with another copy of the same protein, other proteins or peptides (Figure 2B, on the right). The two examples show five ISs (three on the left and two on the right), which were merged into three distinguished IRs (two on the left and one on the right, contoured by thick forest green lines). In all those cases, the IRs result from the merging of ISs that represent binary interactions with different partners. In addition, IRs may also be defined from several ISs representing a single interaction, but whose binding mode slightly differs from one PDB structure to another (see below).

3.2 | Prediction of interacting patches

We predicted interacting patches at the surface of the proteins from **P-262** by relying on four residue properties (see Materials and Methods for precise definitions): evolutionary sequence conservation inferred from the analysis of homologous sequences, physicochemical properties expected at the interface based on experimentally known complex structures, local geometry computed on the protein 3D structure, and propensities to be found at docked interfaces inferred from CC-D calculations (Figure 1). The first three properties are used to derive three different scoring strategies (SC_{cons}, SC_{notLig}, and SC_{geom}) aimed at identifying different types of protein-protein interfaces (see Materials and Methods and Reference 24). Each SC explicitly describes the role of each one of the properties with respect to the expected support-core-rim structure of interacting sites.[57] Evolutionary sequence conservation is used in SC_{cons} and SC_{notLig} to target a very conserved residues on the protein surface. In SC_{notLig}, it is combined with local geometry to avoid small-ligand binding pockets, which are usually more deeply buried than protein-protein interfaces. As both SC_{cons} and SC_{notLig} are designed to target conserved sites, their predictions often overlap substantially.[24] By contrast, SC_{geom} disregards conservation and uses only physicochemical properties and local geometry to capture highly protruding interfaces not necessarily conserved through evolution. The fourth property, inferred from docking, is used exclusively in a fourth strategy, SC_{dock} (see Materials and Methods). It reflects the propensity of each protein residue to bind partners and non-partners in docking calculations. To evaluate docking conformations, we used a coarse-grained empirical energy function comprising a Lennard-Jones potential for van der Waals interactions and a Coulomb potential for electrostatics.[62] The four SCs are implemented in dynJET^2, an upgraded version of the JET^2 method.[24]

3.3 | Estimation of the protein surface involved in functional interactions

We used both experimental interfaces and predicted patches to estimate the proportion of protein surface involved in functional interactions. On average, experimental ISs from **PPI-262** cover ~30% of the protein surface (Figure 3A). Hence, by looking at this dataset, one may infer that the residues involved in functional interactions generally represent less than a third of the protein surface. However, when looking at **PPI-262**$_{ext}$ (Figure 3B), which comprises experimental IRs defined from close homologs, the coverage increases up to ~50%. Moreover, a significant number of proteins (32) have their surface completely or almost completely covered by functional interactions (coverage ≥80%). This suggests that most of the proteins from **P-262** engage in multiple interactions with different partners. One can notice some difference between the functional classes (Figure S2a). At one end of the spectrum, the G proteins and the receptors use only one third of their surface for interactions, on average. At the other end, the antigens and the structural proteins use about two-thirds of their surface.

The estimation provided by the union of predicted patches is slightly higher, ~60% on average (Figure 3C). The associated distribution resembles that of the union of experimental IRs from **PPI-262**$_{ext}$ (compare Figure 3C with 3B), except for two notable differences at the extremities. This observation is statistically supported by a Mann-Whitney *U*-test: while the full distributions are different (*P*-value = .002), the truncated distributions (without values below 20% and above 95%) are indistinguishable (*P*-value = .37). The differences at the extremities are the following: the minimum coverage is higher for predictions than for experimental interfaces (18% vs 6.2%), and there are more proteins completely or almost completely covered (≥ 80%) by predictions than by experimental interfaces. The first difference can be explained by the specifics of dynJET^2 clustering algorithm, which discards very small predictions (see Materials and Methods and Reference 23). The second difference suggests that all functional interfaces have not been yet experimentally characterized.

When looking at individual patches instead of their union, we found that patches predicted from docking (SC_{dock}) display sizes similar to those of experimental ISs (compare Figure 3D,F). Both types of interfaces represent about one quarter of the protein surface, on average. By contrast, conserved (SC_{cons}, SC_{notLig}) predicted patches are bigger, covering about one third of the protein surface, on average (Figure 3G,H). Their size distributions are similar to that of experimental IRs (compare with Figure 3E). These three types of interfaces are highly variable, with standard deviations in the [24-28]% range. Finally, patches predicted based on local geometry (SC_{geom}) are the smallest (Figure 3I), representing 16% of the protein surface.

3.4 | Assessment of the predictions and contribution of each SC

The identification of a protein's set of interacting residues is important to understand the determinants of molecular association. For each protein, we compared the union of all predicted patches with the union of all ISs (respectively IRs) from **PPI-262** (resp. **PPI-262**$_{ext}$). To do so, we relied on the F1-score, which reflects the balance between precision (or positive predictive value) and recall (or sensitivity). The average F1-score on **PPI-262** is 0.41 ± 0.24 and it increases up to 0.57 ± 0.19 on **PPI-262**$_{ext}$ (Figure 4A). This increase reflects a global shift of the F1-score distribution toward higher values (**P-value = 10^{-4}** with the Mann-Whitney *U* test). The proportion of proteins with very good predictions (F1-score > 0.6) increases from 18% to 46%, while that of proteins with very poor predictions (F1-score < 0.2) drastically reduces from about one quarter to

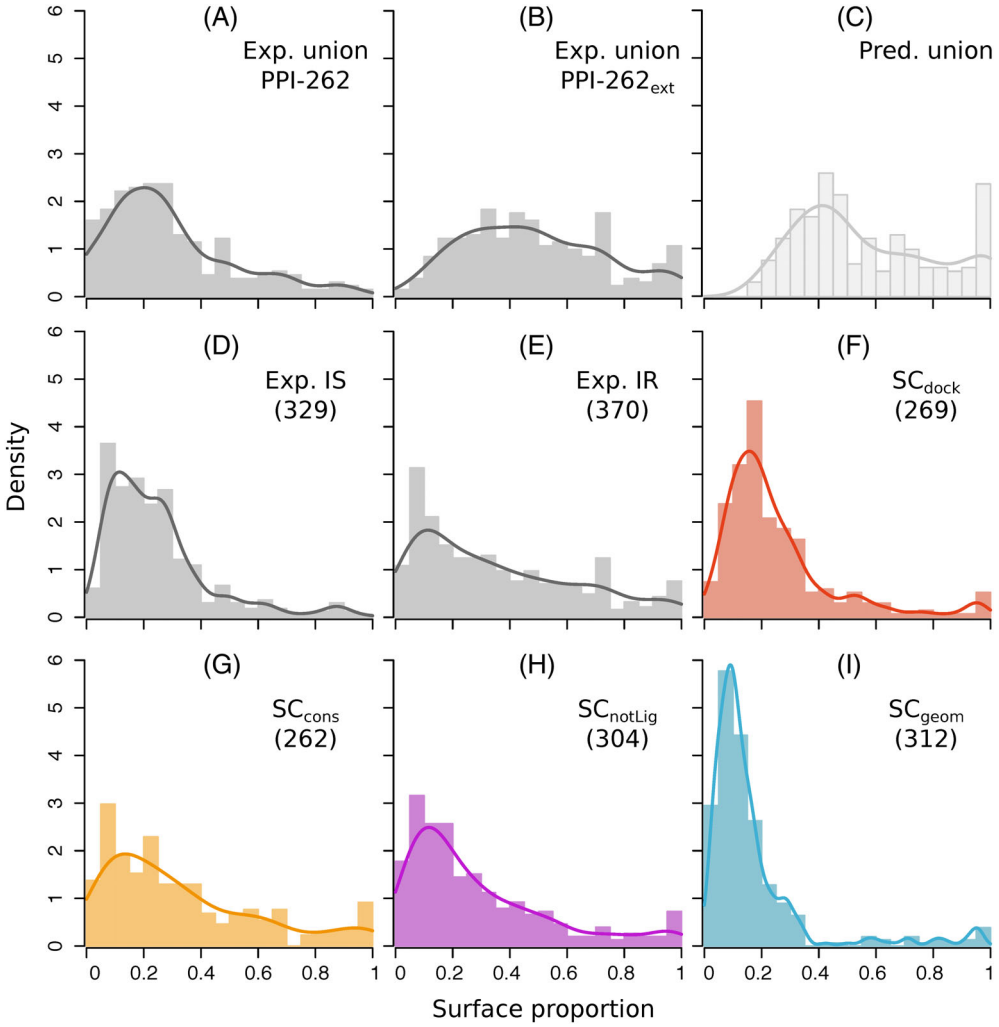

FIGURE 3 Proportion of protein surface covered by experimental interfaces and predicted patches. Distribution are reported for: (A) the union of ISs from **PPI-262**, (B) the union of IRs from **PPI-262**$_{ext}$, (C) the union of patches predicted by dynJET^2, (D) individual ISs from **PPI-262**, (E) individual IRs from **PPI-262**$_{ext}$,(F-I) individual patches predicted by each SC. The union of ISs, IRs or predicted patches is realized for each protein. Notice that the sizes of the predicted patches do not add up when considering their union, since several of them overlap [Color figure can be viewed at wileyonlinelibrary.com]

4%. These results highlight the importance of considering all available experimental information to properly evaluate protein interface predictions. Predicted residues that would be considered as false positives when looking at the restricted dataset, **PPI-262**, are actually involved in interactions with other partners, as revealed by the extended dataset, **PPI-262**$_{ext}$.

We further investigated to what extent the partitioning of protein surfaces into patches predicted by the different SCs matches experimental IRs (Figure 4B,C,D). None of the SC is sufficient on its own to detect all IRs (Figure 4C, in orange, purple, cyan, and red). This observation is also illustrated by the two examples of Figure 5A,C, where several SCs are necessary to capture the entirety of the experimental signal. Combining SC_{cons}, SC_{notLig}, and SC_{geom} enables increasing the average F1-score by about 0.1 compared to individual SCs, and drastically reducing the number of completely missed IRs to only 28 over 370 (7.6%, Figure 4B, in marine). This is indicative of the complementarity of the three SCs in their coverage of the protein surface, as already observed in.[24] Accounting for SC_{dock} patches further enhances the quality of the predictions up to an average F1-score of 0.54 (compare boxplots in marine and darkblue). In particular, the detection of the binding sites at the surface of antibodies and G proteins is very sharp, with average F1-scores of 0.67 and 0.64 (Figure S2b). By contrast, the interfaces of the receptors and the enzymes regulators are the most difficult to detect, with both average F1-scores equal to 0.46.

To better characterize the contribution of docking-based information, we compared the predictive performance of SC_{cons}, SC_{notLig}, SC_{geom}, either considered individually or altogether, with that of SC_{dock} (Figure 5B). We observed that the vast majority of IRs is better detected by the former than the latter (points below the diagonal, 68% on top and 72% at the bottom). Hence, evolutionary conservation, physicochemical properties and local geometry are generally able to better capture protein interface signals than the coarse-grained empirical energy function used in the docking experiment. Nevertheless, there are a number of cases where docking-based data provide valuable information to improve predictions by unveiling interfaces that could not be detected otherwise. An example is given by the anticoagulation Factor X (Figure 5C), where one of its three IRs (in white) is very well detected by SC_{dock} (in red, F1-score = 0.74) but completely missed by the other SCs.

3.5 | Predictions capture interface variability coming from molecular flexibility

Accurately accounting for molecular flexibility remains a challenge for protein interface and interaction prediction. We looked at how our

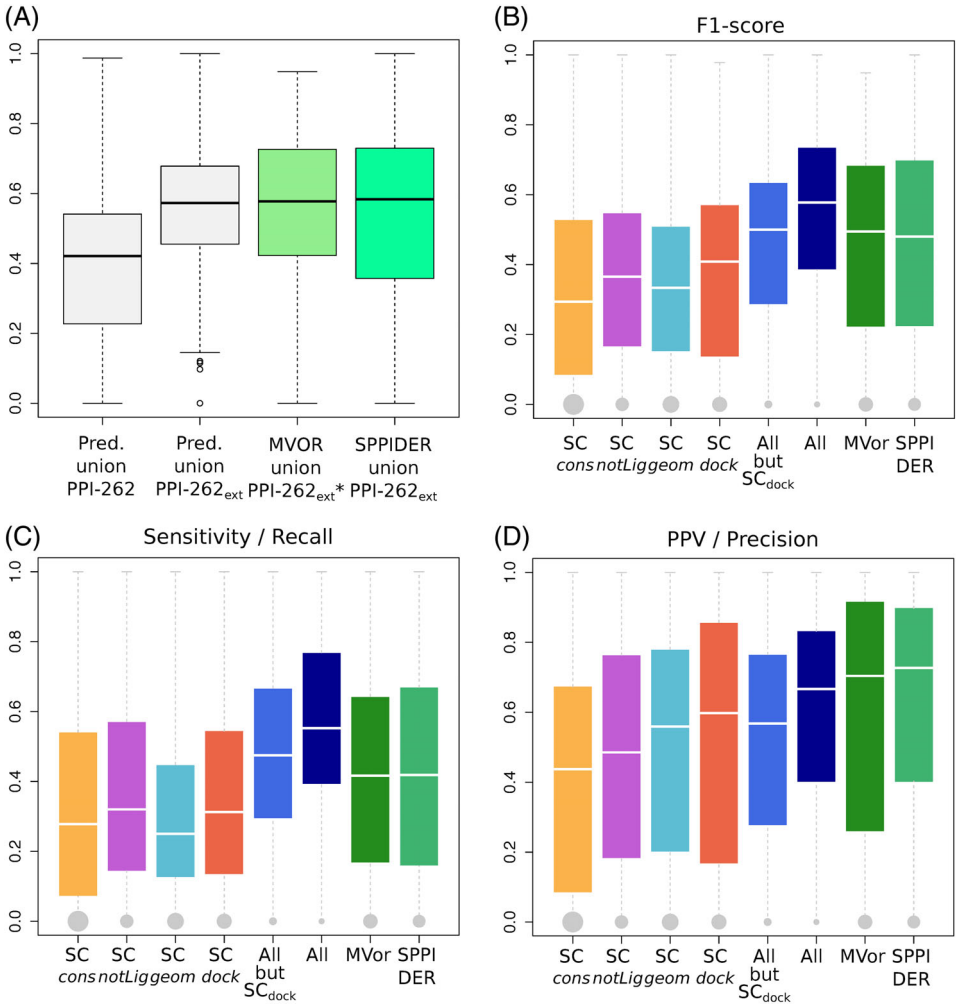

FIGURE 4 Agreement between experimental interfaces and predicted patches. A, Distributions of F1-scores computed for the union of dynJET^2 predictions (boxes in light gray), Multi-VORFFIP predictions (box in light green) and SPPIDER predictions (box in spring green). dynJET^2 predictions were assessed against the union of ISs from **PPI-262** and of IRs from **PPI-262**$_{ext}$. Multi-VORFFIP predictions were assessed against the union of IRs from **PPI-262**$_{ext}$*, a subset from **PPI-262**$_{ext}$ involving 252 protein chains. SPPIDER predictions were assessed against the union of IRs from **PPI-262**$_{ext}$. B-D, Agreement between predicted patches and experimental IRs from **PPI-262**$_{ext}$. For each IR, the best-matching patch or combination of patches predicted by the strategies/methods indicated in x-axis is retained. The performance measures are the following: (B) F1-score, (C) sensitivity (recall), (D) positive predicted value (precision). The sizes of the gray dots are proportional to the number of IRs that could not be detected at all [Color figure can be viewed at wileyonlinelibrary.com]

predicted patches matched experimental interfaces undergoing variations from one structure to another. We focused on the 78 IRs from **PPI-262**$_{ext}$ which are only slightly (<1.5 times) bigger than the corresponding IS(s) from **PPI-262**. Four examples are illustrated on Figure 6. The difference between the IR and the original IS(s) typically reflects the interface variability between different crystallographic structures of the same complex. For example, about 20 structures of the same hetero-4-mer involving Caspase-1 are available in the PDB and contribute to the definition of one IR (Figure 6, top right). For the vast majority of these IRs (>85%), the precision reached by dynJET^2 predictions is equal to or greater than that computed on the corresponding ISs (compare black/white and colored surfaces on Figure 6). These results reveal that there exists a non-negligible variability inherent to protein interfaces and that dynJET^2 predictions is generally able to capture it.

We also assessed the robustness of our predictions with respect to conformational changes. For each IR from **PPI-262**$_{ext}$, we calculated the conformational deviation of its backbone atoms between the query structure from **P-262**, on which our predictions were computed, and the structures of its homologs (see Materials and Methods). Almost all (95%) IRs display average conformational deviations lower than a 4 Å (Figure S4). The extent of the deviation is not correlated to the quality of the predictions (Pearson correlation coefficient of 0.05 between RMSD and F1-score, Figure S4). This indicates that dynJET^2 predictions are robust to small to medium conformational changes.

3.6 | Predicted patches' seeds describe the multiplicity of interactions

Almost all (94%) IRs from **PPI-262**$_{ext}$ were detected, at least partially, by considering predictions issued by all four SCs (Figure 4G). Some predicted patches display a good or very good match with a single IR. For example, the interface between profilin and human VASP (Figure 5A, in black) is very well detected by SC$_{cons}$ (Figure 5A, in beige, *Sens* = 0.63, *PPV* = 0.61). Another example is given by the interface between the heavy and light chains of the anticoagulation factor X which is well detected by SC$_{dock}$ (Figure 5C), in red, *Sens* = 0.82, *PPV* = 0.68). Some other patches cover several IRs, as exemplified by SC$_{geom}$ in Profilin (Figure 5A, in cyan) and SC$_{cons}$ in the factor X's heavy chain (Figure 5C, in beige). These cases are ambiguous if one considers a single SC. However, by crossing the information coming from different SCs, one may infer the existence of several IRs and thus resolve the ambiguity. For instance, the presence of a SC$_{cons}$

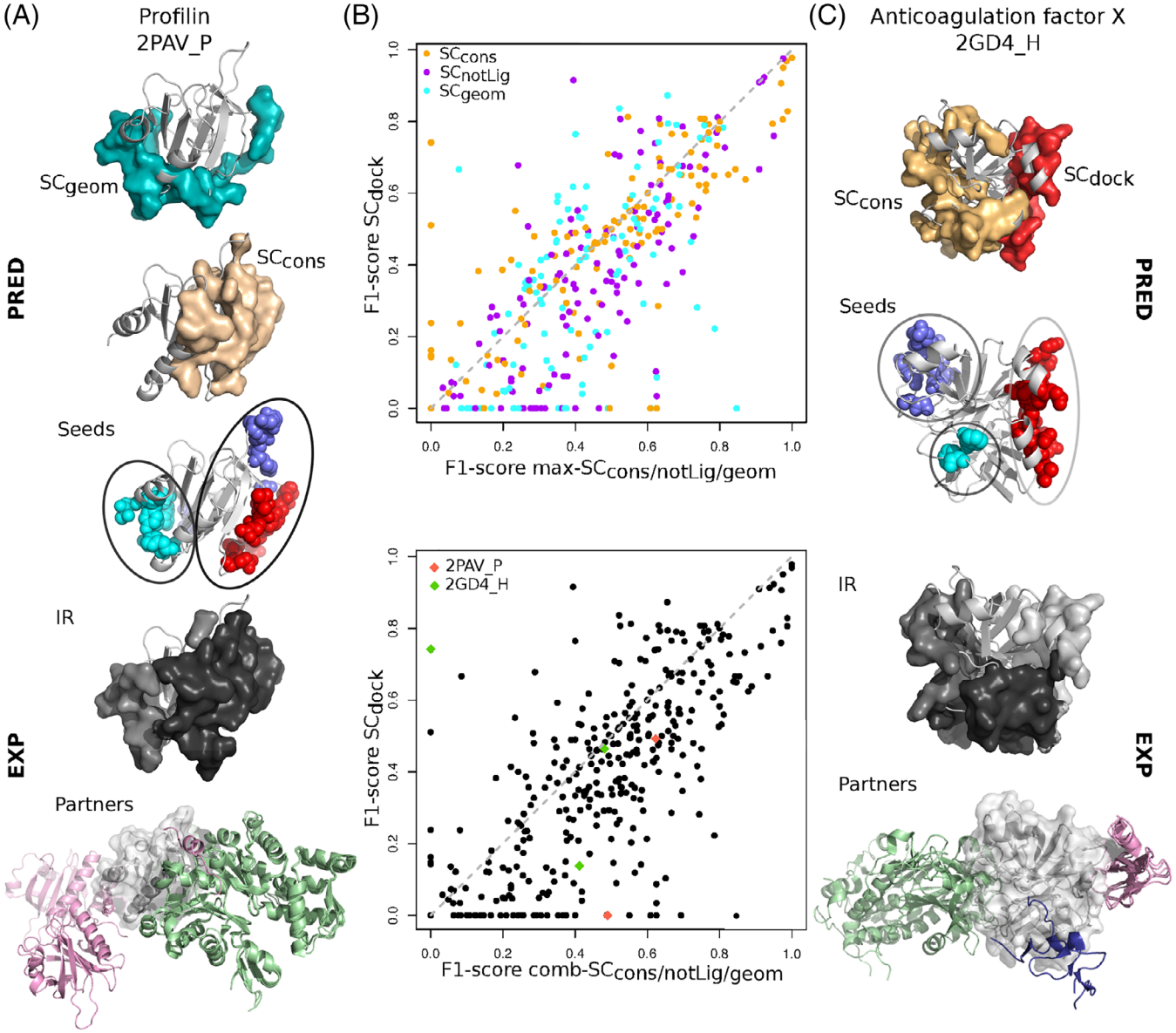

FIGURE 5 Examples and comparison of predictions. A, Profilin (light gray cartoon) displayed with the patches predicted by SC_{cons} (in beige) and SC_{geom} (in cyan), the patches' clustered seeds, two experimental IRs from **PPI-262**$_{ext}$ (in gray tones) and the corresponding partners (colored cartoons); (B) Scatterplot of F1-scores computed for the best-matching patch or combination of patches, among SC_{cons}, SC_{notLig}, SC_{geom} (x-axis), and from SC_{dock} (y-axis) against experimental IRs from **PPI-262**$_{ext}$. In cases where a combination of several patches is retained, the patches either come from a single SC (on top) or from several SC (at the bottom, x-axis). (C) Heavy chain of the anticoagulation factor X (light gray cartoon) displayed with the patches predicted by SC_{cons} (beige) and SC_{dock} (red), the patches' clustered seeds, the three experimental IRs from **PPI-262**$_{ext}$ (in gray tones) and the corresponding partners (colored cartoons) [Color figure can be viewed at wileyonlinelibrary.com]

patch partially overlapping with the SC_{geom} patch at the surface of Profilin (Figure 5A) could be used as an indicator of the existence of two IRs and of the fact that the SC_{geom} patch extends over these two IRs.

To test whether this type of reasoning could be generalized, we systematically investigated how predicted patches were distributed over experimental IRs. For this, we explicitly considered the patches' seeds, which are the first groups of residues being detected by dynJET[2] clustering algorithm. We collected all seeds generated by SC_{notLig}, SC_{geom}, and SC_{dock} and clustered them based on 3D proximity (note that SC_{cons} seeds were not considered for this analysis, see Materials and Methods). The total number of resulting clustered seeds is 562, which corresponds to 2.14 seeds per protein on average. By comparison, the average number of IRs is 1.4. About one quarter of the seeds are completely inside an IR (100% precision) and almost than half of the seeds detect an IR with very high (≥80%) precision (Figure 7A). In the examples of Profilin and factor X, the number of seeds is equal to the number of IRs and each seed points to a different IR (Figure 5A,C).

We also investigated whether seeds could be used to infer the number of partners a protein has (Figure 7B). For this, we looked at the properties of the seeds lying completely or almost completely (PPV

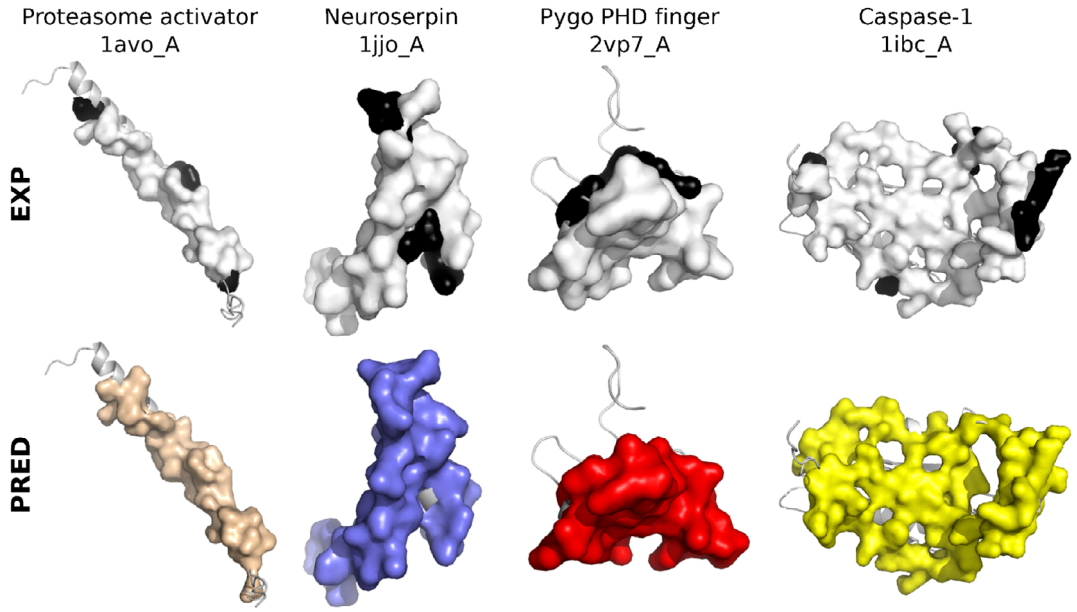

FIGURE 6 Examples of predictions whose precision is higher on the IR compared to the IS. The query protein structure from `P-262` is displayed as a gray cartoon. The experimental and predicted interfaces are displayed as opaque surfaces: on top, the IS is colored in white and the additional residues belonging to the IR are in black; at the bottom, the SC_{cons}, SC_{notLig}, and SC_{dock} patches predicted for 1avo_A, 1jjo_A, 2vp7_A are in wheat, purple and red, respectively, and the best combination of patches predicted for 1ibc_A is in yellow. The precision increases from 79% to 91% for 1avo_A, from 76% to 92% for 1jjo_A, from 70% to 83% for 2vp7_A and from 75% to 84% for 1ibc_A [Color figure can be viewed at wileyonlinelibrary.com]

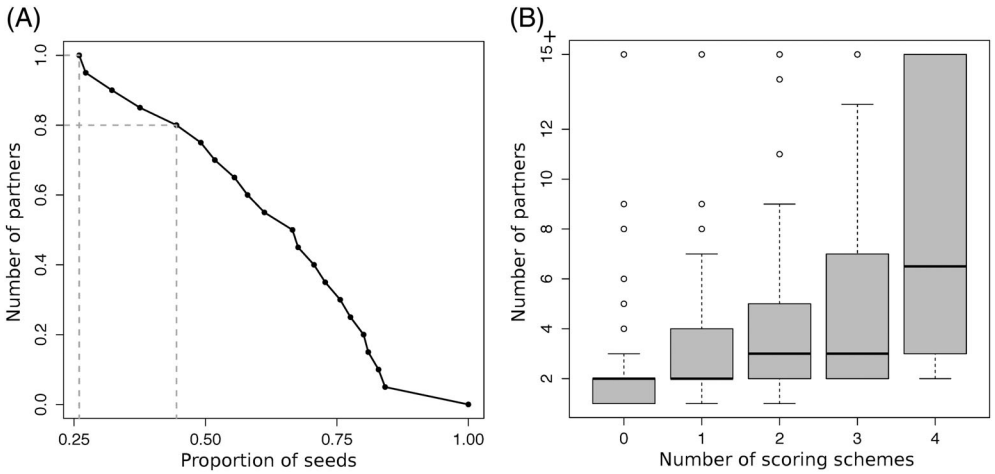

FIGURE 7 Ability of the patches' seeds to detect IRs and estimate the number of partners. (A) Cumulative distribution of patches' seeds precision in detecting IRs. Each x-value corresponds to the proportion of seeds with precision higher than the y-value. Dotted segments emphasize the points with $y = 1$ and $y = 0.8$. (B) Number of partners for each IR vs number of scoring schemes predicting a seed in the IR

≥80%) within an IR. We observed that the number of partners binding to an IR increases with the number of scoring schemes predicting one or more seeds within the IR (Figure 7B, Pearson correlation of 0.52). This means that IRs displaying a multiplicity of signals relevant to protein interactions tend to attract more partners. Hence, the accumulation of seeds with different properties in a protein region can be used as an indicator that this region will likely interact with many partners.

3.7 | Comparison with other state-of-the art interface predictors

We compared dynJET[2] predictions to those of Multi-VORFFIP[59] and SPPIDER,[58] two state-of-the-art machine learning methods. Multi-VORFFIP integrates a broad set of residue descriptors including solvent accessibility, energy terms, sequence conservation, crystallographic B-factors and Voronoi Diagrams-derived contact density, in a two-steps random forest ensemble classifier. It was applied to a subset of 252 protein chains from `P-262` (see Materials and Methods). SPPIDER implements a consensus based classifier that combines 10 different neural networks. It integrates various sequence- and structure-based features, those contributing the most being based on solvent accessibility prediction. It was applied on all proteins from `P-262`. The distributions of F1-scores obtained for the union of the residues predicted by Multi-VORFFIP and SPPIDER, respectively, are similar to that obtained for the union of dynJET[2] predictions (Figure 4A, compare the second, third and fourth boxes). However, the performance of Multi-VORFFIP and SPPIDER in detecting individual IRs is lower than that of dynJET[2] (Figure 4B,C,D, compare darkblue and green boxplots). The average

F1-score is lower, the number of very good predictions is significantly lower and the number of missed IRs is much larger (52 for MultiVORFFIP, 46 for SPPIDER, vs 21 for dynJET^2). In particular, dynJET^2 is more sensitive than the two other tools (Figure 4D).

4 | DISCUSSION

Protein surfaces are used in multiple ways in cellular partners association. A comprehensive and accurate description of protein surfaces should account for multiple partners, molecular flexibility (from slight rearrangements to conformational changes), disorder and post-translational modifications. In this work, we have analyzed a pool of proteins with different functions to address the two first aspects.

In line with a previous study,[12] we found that the protein surface involved in functional interactions is probably much bigger than anticipated. An accurate estimation of this surface is mandatory for the correct assessment of protein interface prediction methods. However, such an estimation is still beyond reach as we do not know the exact number of cellular partners a protein has and how these partners use its surface in solution. To move forward, we have introduced the notion of interacting region, which results from combining several overlapping interacting sites detected in experimental complex structures. By taking into account all known homologs of our query proteins and their crystallographic complexes, we could synthesize over 23 000 ISs into a relatively small number of IRs (1.4 per protein chain). This procedure permitted shedding light on the variability of binding modes and on the multiplicity of protein surface usage. We observed that, within an IR, some residues are specific to the interaction with one partner while others are shared between different partners, and possibly between another copy of the same protein and other proteins or peptides. The proportion of shared residues can be high, indicating that IRs can serve as "binding platforms" for very different partners. In the evaluation of dynJET^2 predictions, we could appreciate that a large amount of predicted patches better matched IRs, compared to ISs. This result is expected from a good protein interface prediction algorithm, as the notion of IR seems more biologically pertinent than that of IS in many cases, especially when the IR synthesizes the variability inherent to structure ensembles of the same complex.

Our predictions were generated based on three sequence- and structure-based properties of monomeric protein surfaces and also on residue propensities inferred from docking calculations. The latter reflect how energetically favorable the interatomic interactions established between a protein and many other proteins are. The energy was evaluated by a combination of Lennard-Jones and Coulomb potentials. We have systematically assessed the contribution of this physical description of protein interactions to the prediction of interfaces. We have shown that in most cases, its predictive power is lower than that of the sequence and structure-based descriptors. Nevertheless, it brings complementary information, which helps to improve the accuracy of the predictions and, in some cases, it even unveils binding sites that could not be detected otherwise.

We have highlighted several cases where almost the entire protein surface is involved in functional interactions. This finding challenges the role of specificity in the evaluation of protein interface prediction methods and rather put the emphasis on precision. We have demonstrated the usefulness of the prediction patches' three-layer structure by showing that the patches' seeds enabled precisely locating and discriminating IRs at the protein surface. We have further shown that the seeds could help determine whether a protein has a few or many partners. Future work will aim at getting a more accurate estimation of the number of partners. Moreover, with the help of future PPI data, it seems achievable to associate functions to the partners binding on different surface areas, described by different seeds on a region.

Although dynJET^2 predictions match reasonably well experimentally identified interacting regions, the match is not perfect. Given the degree of complexity we have highlighted in the usage of a protein surface, it seems legitimate to ask whether perfect match with experimental interfaces is an attainable goal for protein interface predictions algorithms that, like dynJET^2, do not use any knowledge about the query protein's partners. An accurate estimation of the maximum level of agreement one could expect would be most valuable. Besides, even without perfect match, dynJET^2 predictions can be fully exploited to guide experiments. For example, the above-mentioned ability of patches' seeds to precisely locate IRs has implications for the control and modulation of existing protein-protein interactions. Mutating seed residues should impact the binding of the associated partners. Another way to go would be to design interactors that bind to the predicted patches. Indeed, dynJET^2 algorithm provides a mean to delineate regions at the protein surface with sizes similar to those of experimental interacting sites or regions and complying with a few properties known to be relevant to protein-protein association. Thus, in addition to detecting regions being actually used to interact, it also reveals the potentiality of other regions to interact.

ACKNOWLEDGMENTS

The MAPPING project (ANR-11-BINF-0003, Excellence Programme "Investissement d'Avenir"); funds from the Institut Universitaire de France; the access to the HPC resources of the Institute for Scientific Computing and Simulation (Equip@Meso project - ANR-10-EQPX-29-01, Excellence Program "Investissement d'Avenir"); the World Community Grid (WCG, www.worldcommunitygrid.org) and WCG volunteers that allowed us to perform cross-docking experiments with MAXDo on the HCMD2 dataset; Alexey Porollo for handling his script to submit large scale projects to SPPIDER's webserver.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

ORCID

Alessandra Carbone https://orcid.org/0000-0003-2098-5743

REFERENCES

1. Baker M. Proteomics: the interaction map. *Nature*. 2012;484(7393):271-275.
2. Bonetta L. Protein-protein interactions: Interactome under construction. *Nature*. 2010;468(7325):851-854.
3. Huttlin EL, Ting L, Bruckner RJ, et al. The BioPlex network: a systematic exploration of the human Interactome. *Cell*. 2015;162(2):425-440.
4. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;159(5):1212-1226.
5. Meyer MJ, Beltran JF, Liang S, et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat Methods*. 2018;15(2):107-114.
6. Hayashi T, Matsuzaki Y, Yanagisawa K, Ohue M, Akiyama Y. MEGADOCK-web: an integrated database of high-throughput structure-based protein-protein interaction predictions. *BMC Bioinformatics*. 2018;19(Suppl 4):62.
7. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;45(D1):D362-D368.
8. Garzon JI, Deng L, Murray D, Shapira S, Petrey D, Honig B. A computational interactome and functional annotation for the human proteome. *Elife*. 2016;10:5.
9. Tsuji T, Yoda T, Shirai T. Deciphering supramolecular structures with protein-protein interaction network modeling. *Sci Rep*. 2015;5:16341.
10. Mosca R, Ceol A, Aloy P. Interactome3D: adding structural details to protein networks. *Nat Methods*. 2013;10(1):47-53.
11. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. *PLoS Comput Biol*. 2013;9(12):e1003369.
12. Tonddast-Navaei S, Skolnick J. Are protein-protein interfaces special regions on a protein's surface? *J Chem Phys*. 2015;143(24):243149.
13. Szilagyi A, Nussinov R, Csermely P. Allo-network drugs: extension of the allosteric drug concept to protein- protein interaction and signaling networks. *Curr Top Med Chem*. 2013;13(1):64-77.
14. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*. 1996;257(2):342-358.
15. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*. 1997;6(1):53-64.
16. Larsen TA, Olson AJ, Goodsell DS. Morphology of protein-protein interfaces. *Structure*. 1998;6(4):421-427.
17. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol*. 1998;280(1):1-9.
18. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng*. 2000;13(2):77-82.
19. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*. 2001;43(2):89-102.
20. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins*. 2002;47(3):334-343.
21. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*. 2004;338(1):181-199.
22. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics*. 2007;23(2):e13-e16.
23. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput Biol*. 2009;5(1):e1000267.
24. Laine E, Carbone A. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLOS Comput Biol*. 2015;11(12):1-32.
25. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel JC, Deane CM. Progress and challenges in predicting protein interfaces. *Brief Bioinform*. 2016;17(1):117-131.
26. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*. 2002;18(Suppl 1):S71-S77.
27. Glaser F, Pupko T, Paz I, et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*. 2003;19(1):163-164.
28. Cheng G, Qian B, Samudrala R, Baker D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res*. 2005;33(18):5861-5867.
29. Innis CA. siteFiNDER—3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res*. 2007;35(Web Server issue):W489-W494.
30. Tyagi M, Thangudu RR, Zhang D, Bryant SH, Madej T, Panchenko AR. Homology inference of protein-protein interactions via conserved binding sites. *PLoS One*. 2012;7(1):e28896.
31. Xue LC, Dobbs D, Honavar V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinf*. 2011;12(1):244.
32. Esmaielbeiki R, Nebel JC. Scoring docking conformations using predicted protein interfaces. *BMC Bioinf*. 2014;15(1):171.
33. Ripoche H, Laine E, Ceres N, Carbone A. JET2 viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. *Nucleic Acids Res*. 2017;45(7):4278.
34. Aumentado-Armstrong TT, Istrate B, Murgita RA. Algorithmic approaches to protein-protein interaction site prediction. *Algorithms Mol Biol*. 2015;10:7.
35. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol*. 2004;335(3):843-865.
36. Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinform*. 2008;9(1):447.
37. Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins*. 2010;78(15):3085-3095.
38. Hwang H, Vreven T, Pierce BG, Hung JH, Weng Z. Performance of ZDOCK and ZRANK in CAPRI rounds 13–19. *Proteins*. 2010;78(15):3104-3110.
39. de Vries SJ, Bonvin AM. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PloS One*. 2011;6(3):e17695.
40. Hwang H, Vreven T, Weng Z. Binding interface prediction by combining protein–protein docking results. *Proteins*. 2014;82(1):57-66.
41. Martin J, Lavery R. Arbitrary protein- protein docking targets biologically relevant interfaces. *BMC Biophysics*. 2012;5(1):7.
42. Sacquin-Mora S, Carbone A, Lavery R. Identification of protein interaction partners and protein-protein interaction sites. *J Mol Biol*. 2008;382(5):1276-1289.
43. Vamparys L, Laurent B, Carbone A, Sacquin-Mora S. Great interactions: how binding incorrect partners can teach us about protein recognition and function. *Proteins*. 2016;84(10):1408-1421.
44. Laine E, Carbone A. Protein social behavior makes a stronger signal for partner identification than surface geometry. *Proteins*. 2017;85(1):137-154.
45. Lagarde N, Carbone A, Sacquin-Mora S. Hidden partners: using cross-docking calculations to predict binding sites for proteins with multiple interactions. *Proteins*. 2018;86(7):723-737.
46. Mintseris J, Wiehe K, Pierce B, et al. Protein-Protein Docking Benchmark 2.0: an Update. *Proteins*. 2005;60(2):214-216.
47. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mole Biol*. 2007;372(3):774-797.

48. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423.
49. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947-2948.
50. Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*. 1988;5(6):729-731.
51. Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*. 2004;336(5):1265-1282.
52. Negi SS, Braun W. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J Mol Model*. 2007;13(11):1157-1167.
53. Mezei M. A new method for mapping macromolecular topography. *J Mol Graph Model*. 2003;21:463-472.
54. Ceres N, Pasi M, Lavery R. A protein solvation model based on residue burial. *J Chem Theo Comput*. 2012;8(6):2141-2144.
55. Hubbard SJ, Thornton JM. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London 1993;2.
56. Dequeker C, Laine E, Carbone A. INTerface builder: a fast protein-protein Interface reconstruction tool. *J Chem Inf Model*. 2017;57(11):2613-2617.
57. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol*. 2010;403:660-670.
58. Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions. *Proteins*. 2007;66(3):630-645.
59. Segura J, Jones PF, Fernandez-Fuentes N. A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics*. 2012;28(14):1845-1850.
60. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *PNAS*. 2003;100(10):5772-5777.
61. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
62. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*. 2003;12(6):1271-1282.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Dequeker C, Laine E, Carbone A. Decrypting protein surfaces by combining evolution, geometry, and molecular docking. *Proteins*. 2019;87:952–965. https://doi.org/10.1002/prot.25757