Predicting Protein–Protein Interfaces that Bind Intrinsically Disordered Protein Regions

Eric T.C. Wong and Jörg Gsponer

Michael Smith Laboratories, *University of British Columbia, Vancouver, BC, Canada*
Department of Biochemistry and Molecular Biology, *University of British Columbia, Vancouver, BC, Canada*

Correspondence to Jörg Gsponer: *gsponer@msl.ubc.ca*
https://doi.org/10.1016/j.jmb.2019.06.010
Edited by Monika Fuxreiter

Abstract

A long-standing goal in biology is the complete annotation of function and structure on all protein–protein interactions, a large fraction of which is mediated by intrinsically disordered protein regions (IDRs). However, knowledge derived from experimental structures of such protein complexes is disproportionately small due, in part, to challenges in studying interactions of IDRs. Here, we introduce IDRBind, a computational method that by combining gradient boosted trees and conditional random field models predicts binding sites of IDRs with performance approaching state-of-the-art globular interface predictions, making it suitable for proteome-wide applications. Although designed and trained with a focus on molecular recognition features, which are long interaction-mediating-elements in IDRs, IDRBind also predicts the binding sites of short peptides more accurately than existing specialized predictors. Consistent with IDRBind's specificity, a comparison of protein interface categories uncovered uniform trends in multiple physicochemical properties, positioning molecular recognition feature interfaces between peptide and globular interfaces.

Introduction

Protein–protein interactions have a fundamental role in most biological processes and exhibit a remarkable diversity in structure. They enable the assembly of large cellular machines or transient complexes that mediate the transmission of cellular signals. Disruption of protein interactions by mutations can change cellular phenotypes and even lead to disease in multicellular organisms [1]. Therefore, significant effort has been made to map and characterize the molecular aspects of all protein interactions present in cellular systems [2].

Protein–protein interaction interfaces share common characteristics that make them differentiable from non-interface surfaces [3,4]. The most consistently discernable property is stronger evolutionary conservation of residues at the interface. Residue composition also differs between interface and non-interface surfaces, particularly when dividing the interface into core and rim regions [5]. Rim is the outer region that remains partially solvent accessible upon binding and typically consists of polar and charged residues. Core is the central region of the interface that is occluded from solvent upon binding and is generally hydrophobic, which is important since hydrophobic interactions are often the dominant contributors to binding affinity [6]. Concordantly, the core is enriched in hotspots, which are interface residues that contribute to a large fraction of the total interaction energy. Strong conservation of residues and enrichment in disease-causing mutations further highlight the importance of the core region in protein–protein interactions [7].

Properties that distinguish interface from non-interface residues are exploited in numerous computational methods that predict protein interaction sites on their given structures [8]. The property most commonly used by these protein interface predictors is evolutionary conservation, which is generally derived from multiple sequence alignments or, more rarely, the conservation of interface residues among structural neighbors [9]. Prediction methods can also exploit the physicochemical properties that distinguish interface from non-interface residues. These properties include hydrophobicity and solvent

accessibility [10–17]. Taking advantage of the plethora of available approaches, meta-predictors combine multiple predictors to increase prediction accuracy [18]. Protein–protein interface prediction is a challenging task, but the cumulation of advances in this field has led to predictors that are highly effective for protein characterization.

The majority of available protein interface predictors have been trained and tested on classical protein complexes in which all interaction partners are globular [10,12,16], generalizing the term globular to include all independently folding domains. However, this mode of interaction represents only a part of the currently recognized spectrum of interactions. Research over the last two decades has revealed that a large fraction of eukaryotic proteins contain intrinsically disordered protein regions (IDRs) and that these regions can also be involved in protein interactions. A steadily growing number of identified interactions are mediated by peptide motifs, which are also referred to as short linear motifs (SLiMs) [19]. Peptide motifs are typically segments up to 10 residues long that often bind to specific binding domains through conserved consensus sequence motifs [20,21]. A prolific example is certain proline-rich motifs that bind to SH3 domains [22]. Some peptide motifs occur in repeats or combinations, resulting in competitive or cooperative interactions [20].

In addition to the peptide motifs, much longer interaction-prone regions have been identified in IDRs [23]. Molecular recognition feature (MoRF) is a broad term encompassing interaction-mediating segments in IDRs at 10 to 70 residues in length that often fold upon binding [24–26]. They typically lack full-length consensus sequences, although they may contain shorter motifs. MoRFs form a diverse group with heterogeneous structures and interaction mechanisms. Unlike many peptide motifs, MoRFs can gain intricate folds upon binding, including substantial secondary structures and extended conformations that wrap around their binding partners [27]. Furthermore, MoRFs often contain regions that transiently sample secondary structures prior to binding, namely, preformed structural elements [28,29], and their interactions can involve anchoring, flanking, and linker subsegments [30]. For example, the 69-residue MoRF of p27, which contains a hydrophobic anchoring region as well as a partially preformed helix, folds up and binds to a heterodimer of globular proteins Cdk2 and cyclin A [31,32].

Interactions mediated by peptide motifs and MoRFs were initially assumed to result in complexes with well-defined quaternary structures, akin to quaternary structures of obligate complexes consisting of folded domains. However, this assumption was recently proven specious by the increasing evidence of IDRs exhibiting conformational heterogeneity in both their unbound and bound states, which was termed fuzziness in protein interactions [33,35]. Although MoRFs are classically characterized as regions that fold upon binding, they often exhibit multiple binding conformations [34]. A growing body of research recognizes fuzzy interaction modes and the contribution of both folded elements and transient contacts to IDR-mediated interactions [36]. The dynamic nature of MoRFs and broad energy landscapes of their bound and unbound states are proposed to favor their interactions with multiple partners [37,38]; a trait exploited by hub proteins that are central to signaling networks [39]. In essence, the emerging picture illustrates a continuum in the degree of conformational heterogeneity not only of individual protein structures, ranging from independently folding domain to IDRs [40], but also of protein interaction complexes, ranging from well-defined quaternary structures to highly heterogeneous complexes with no dominant binding mode [41,42].

While a comprehensive protein interaction interface predictor would be ideal, predictors have always targeted subsections of the interaction spectrum due to the diversity in attributes and the biases in our protein structure knowledgebase. The performance of predictors trained on globular protein complexes is suboptimal for IDR-mediated interactions because of the differences between the two classes of interactions. Chief among the features that distinguish interfaces between globular domains (globular interfaces for short) from interfaces between globular domains and peptides (peptide interfaces) is the deep pockets and grooves that are prominent in the latter [13,43]. Because peptides are intrinsically flexible, the conformational entropy cost in binding is a crucial component determining binding affinity. This energetic cost is counteracted by hotspots that mainly consist of hydrophobic and aromatic residues, often constituting the conserved motifs of the binding peptides. A key component of these hydrophobic interactions is the favorable solvent entropy change resulting from the desolvation of hydrophobic surfaces. Another strategy for mitigating the entropic cost of peptide binding was proposed based on the observation of bridging water molecules in unbound peptide interfaces [43]. Increasing awareness of differences between globular and peptide interfaces led to the development of specific predictors such as PeptiMap [13,43]. Similar to peptides, MoRFs tend to bind grooves on partner surfaces. However, protein interactions mediated by MoRFs are depleted in hydrogen bonds but further enriched in hydrophobic and electrostatics pairings when compared to interactions mediated by peptides [27,44]. Given the success of available interface predictors for binding sites of globular proteins and peptides, we aimed to develop a predictor for MoRF-binding sites (MoRF interface), which is a section of the interaction spectrum that has remained underserved.

To achieve our objective of developing a MoRF interface predictor, we combined two different machine learning approaches. The gradient boosted trees method was used first to train two prediction modules to identify core and rim residues. Prediction scores from these modules were then integrated through a conditional random field (CRF) to generate the final classification labels (Fig. S1). On a non-redundant test dataset of complexes between MoRF and globular proteins, the final predictor achieves a Matthews correlation coefficient (MCC) of 0.31 at a sensitivity of 51% for the separation of MoRF interface and non-interface residues. Surprisingly, our predictor also achieves an MCC of 0.36 at a sensitivity of 56% on peptide interfaces. In contrast, the predictor has an MCC of only 0.2 at a sensitivity of 33% on globular interfaces. Thus, our predictor, IDRBind, identifies protein interfaces that preferentially bind both peptide motifs and MoRFs, collectively the binding sites of IDRs. In agreement with IDRBind's performance on globular, peptide, and MoRF interfaces, a detailed analysis of multiple physicochemical measures places MoRF interfaces in the midst of a spectrum ranging between peptide to globular interfaces.

Results

MoRF complex datasets for predictor training and evaluation

Creating a computational tool that identifies MoRF interfaces on protein domains requires the assembly of training and testing datasets that consist exclusively of protein complexes that contain MoRFs and have minimal redundancy (see Methods). We first collected complexes for which at least one partner has experimental evidence for being intrinsically disordered prior to binding. Complexes were identified by using the IDEAL database in conjunction with a literature search [45]. Removal of redundant sequences resulted in 84 complexes. Although interface predictors can be trained on bound structures [16,17], meaning protein conformations found in the complexed form are used in training, performance evaluation requires a more stringent and realistic dataset that consists of protein structures in their unbound state. For 57 of the 84 complexes, we found unbound structures of the MoRF binding partners (MoRF partners). Thirty of these unbound structures were randomly allocated to the test set, which we named MoRF-test, and the rest were put in MoRF-train (i.e., 54 mixed bound and unbound structures of the MoRF partners). MoRF-test contains 4107 surface residues in total, including 186 core and 532 rim residues, as defined in the methods section. MoRF-train consists of 8863 residues, of which 427 and 1071 are core and rim, respectively. Although the resulting datasets are small, the stringent selection process allowed us to provide better estimates of the prediction performance and to form conclusions that are specific to MoRF interfaces.

While this work is focused mainly on the MoRF partners, the MoRFs were characterized to clarify the types of interaction elements contained in our datasets. The MoRFs range in lengths from 10 to 72 residues and have a mean of 25 residues. While they all have evidence of intrinsic disorder, some MoRFs in the two datasets transiently sample secondary structures in their unbound states (Table S1). Furthermore, MoRFs contain helices in 62% of the cases in their partner-bound form [46]. To assess the conformational diversity (i.e., fuzziness) of individual MoRFs, we calculated the mean RMSDs between NMR models of MoRFs after alignment of the MoRF partner structures. MoRFs in our datasets display a wide distribution of mean Cα RMSDs ranging between 0.6 and 17.0 Å (Fig. S2A). However, the primary interface residues, which we defined as MoRF residues that participate in the interaction in at least 50% of the NMR models, are not as structurally diverse as the remaining flanking residues. Notably, the segments containing primary interface residues of MoRFs in 18 out of 33 structures have mean Cα RMSDs below 3 Å. An example with relatively rigid primary interface residues is provided by the complex between XPC and p62, which is a subunit of transcription factor II H complex (TFIIH) that is recruited by XPC in the nucleotide excision repair pathway (PDB ID 2RVB [47]; Fig. S2B). In contrast, the structures of the complex between the activation domain of the herpes simplex virus protein VP16 and the human transcription initiation factor TFIIB [48] provide an example of a MoRF where models differ substantially in structure, thus resulting in high mean RMSD for the primary interface residues (Fig. S2B). Notably, while the MoRFs themselves display a wide conformational range, the residues on the partner surface to which they associate remain largely preserved across the models, providing a firm basis for the definition of MoRF interfaces that we aim to predict.

Several features are associated with core and rim interface residues

Separating MoRF interface residues from non-interface residues is a binary classification task. Therefore, we decided to use receiver operating characteristic curves (ROC curves) to evaluate whether features that are classically used in the identification of interface residues are also useful in the prediction of MoRF interfaces. Moreover, we used ROC curves to gauge whether these features are better at predicting all interface residues or segregated core and rim residues. We utilized structure and sequence-based features that are

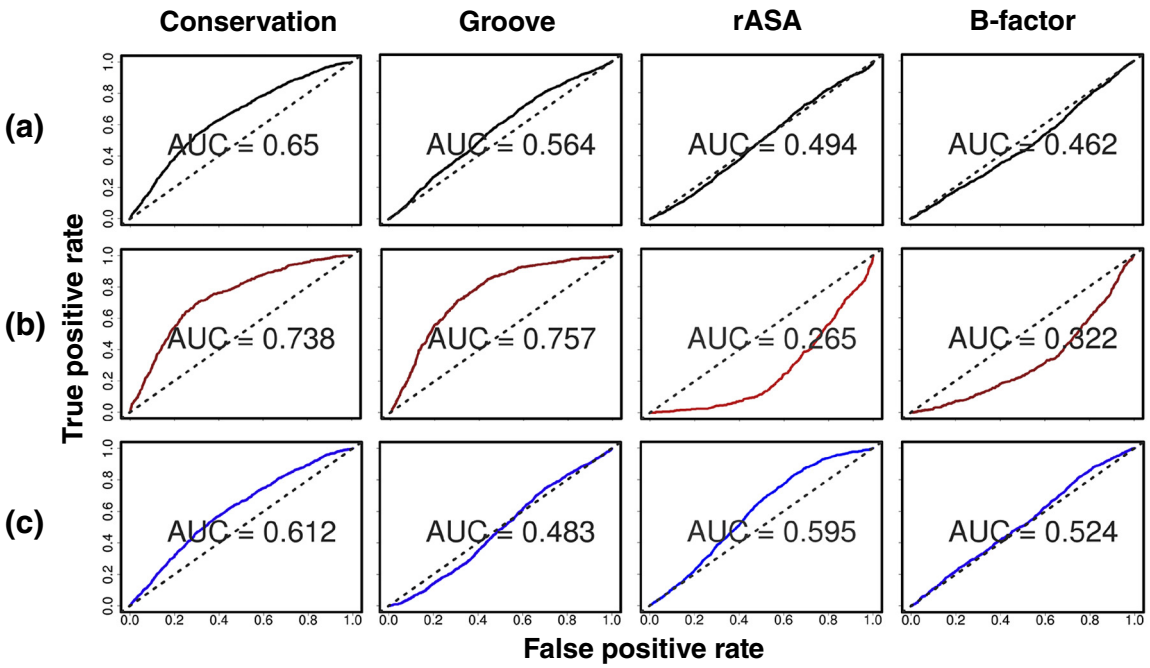

Fig. 1. ROC curves for interface *versus* non-interface classification using different individual feature scores. Non-interface residues are classified against all interface residues (A), core residues (B), and rim residues (C). Selected feature scores are shown from left to right: residue conservation score, groove score, rASA score, and *B*-factor estimate score. The dotted diagonal lines represent random classification, which has an AUC of 0.5. See Methods and Table S2 for detailed descriptions of the features.

either incorporated in existing interface prediction methods or deemed relevant to the prediction of MoRF interfaces. These features can be grouped broadly into six categories: residue composition, residue evolutionary conservation, residue relative accessible surface area (rASA), protein surface geometry, estimated *B*-factors and electrostatics of surface residues or their local environment. The scores of these features were calculated for the surface residues of each protein in the MoRF-train set, some of which using existing methods (see Methods and Table S2 for details).

The ROC curves and their corresponding area under curves (AUCs) of a selection of features tested are shown in Fig. 1A. We found that conservation feature scores decently separate interface and non-interface surface residues (AUC: 0.65). However, other feature scores lack the discriminative power to separate all interface from non-interface residues. Next, we segregated interface residues into core and rim (see Methods) and assessed the predictive power of the same features on discriminating non-interface residues from core and rim, respectively. Consistent with the literature, conservation scores are better at separating core residues and non-interface residues than all interface and non-interface residues (Fig. 1B). Most importantly, the other features selected are also able to separate core from non-interface residues. Two feature scores worth noting are a surface geometry feature designed to identify grooves and pockets (groove score) and rASA, a solvent-accessible surface area (SASA) score normalized per residue. Core residues tend to have higher groove scores than non-interface residues, suggesting that they are situated at the center of grooves in the protein surface. Correspondingly, core residues are likely to be less accessible than the bulk of surface residues. Some of the features have differently signed correlation with core and rim residues, which is reflected in AUCs above and below 0.5 for core and rim (Fig. 1B, C). For instance, rASA has a distinctive ability for differentiating core and rim residues because of the tendencies for core residues to be in grooves and rim residues to be more solvent exposed. Similarly, the *B*-factor estimates show that core residues are more rigid compared to both rim and non-interface residues. Overall, this analysis reveals that the predictive power of many features is amplified by separating the interface into core and rim regions.

Modules that predict core and rim residues

Based on these findings, we developed two distinct prediction modules that separate core and rim residues, respectively, from non-interface residues. To do so, we used gradient boosted trees models. The XGBoost software library [49] in R was

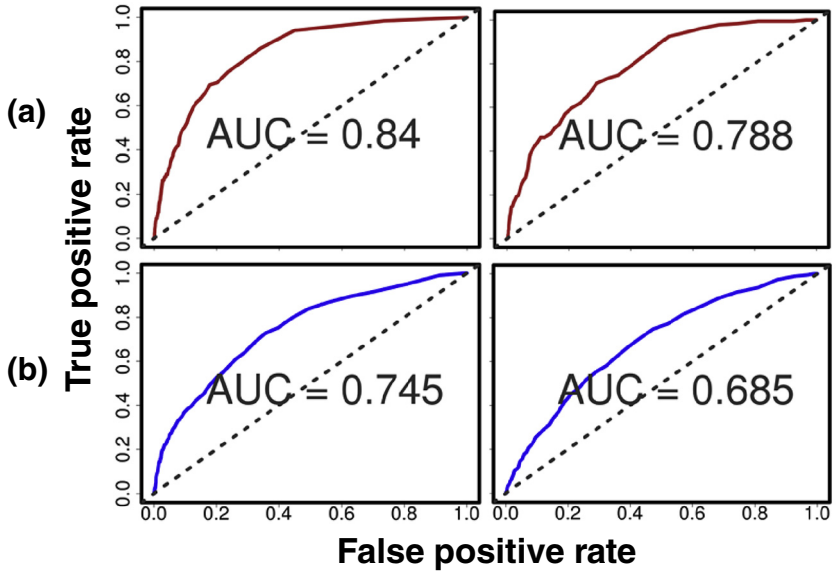

Fig. 2. ROC curves for core, rim and interface residue predictions. Left, ROC curves for the classification of core (A) and rim (B) residues *versus* non-interface residues made by CorePred and RimPred. Right, ROC curves for the classification of core (A) and rim (B) residues *versus* non-interface residues made by a single model trained to classify interface *versus* non-interface residues.

employed to train the two classification models. Thus, one gradient boosted trees model was trained to classify core and non-interface residues, while the other was trained to classify rim and non-interface residues. All features mentioned before as well as others (full list, see Table S2) were calculated for residues in the MoRF-train set and used in training. In addition, aggregated feature scores were created using surface residue patches. Surface residue patches are defined by combining each surface residue with its neighbors within a distance threshold (see Methods). For all residues in each surface patch, feature scores are aggregated (e.g., by taking averages or maximum scores) and added to the

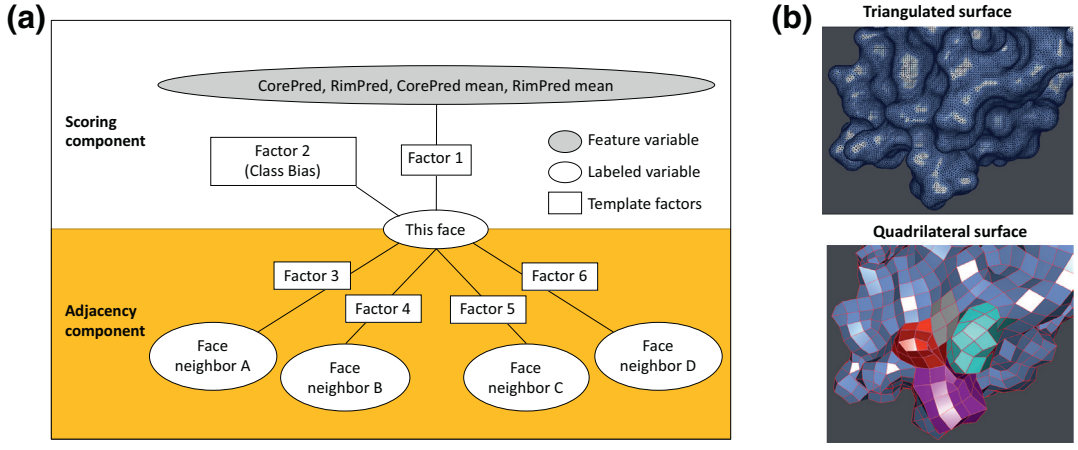

Fig. 3. Schematic of IDRBind CRF (A) and the quadrilateral mesh protein surface representation (B) used in it. (A) Variables used in the CRF are represented by ovals and factors by rectangles. There are feature variables (i.e., observed variables) in gray and labeled variables (i.e., unobserved variables) in white, which are the residue nodes for which we are predicting class labels. Each labeled variable is a face in the quadrilateral mesh surface. Factors connect the variables, describing their compatibility using weights and feature functions. The scoring component of the graph model is on top, while the adjacency component is highlighted with a yellow background below. (B) Quadrilateral mesh protein surface representation used to construct the graph model. Top, triangulated surface representation of a protein. Bottom, the same surface in quadrilateral representation. Quadrilateral faces identically colored in red, magenta, green and tan identify groups of faces that are mapped to the same protein residue.

Table 1. Performance evaluation of interface predictors on MoRF-test dataset

	TPR	FPR	Precision	Specificity	Accuracy	F1	MCC	meanMCC	SD error	NA count	*n*
IDRBind	0.51	0.17	0.39	0.83	0.77	0.44	0.31	0.31	0.04	0	30
ACCLUSTER3[a]	0.58	0.23	0.34	0.77	0.73	0.42	0.28	0.24	0.04	0	30
PeptiMap	0.32	0.13	0.34	0.87	0.78	0.33	0.20	0.18	0.05	0	30
ISMBLab	0.48	0.26	0.28	0.74	0.70	0.35	0.18	0.22	0.04	1	30
VORFFIP	0.78	0.51	0.24	0.49	0.54	0.36	0.20	0.15	0.03	1	30
CPORT	0.35	0.18	0.28	0.82	0.74	0.31	0.15	0.14	0.03	0	30
cons-PPISP	0.24	0.12	0.29	0.88	0.77	0.26	0.12	0.11	0.04	0	30
PredUs2-SVM	0.41	0.18	0.32	0.82	0.75	0.36	0.21	0.20	0.04	1	28
PrISE	0.45	0.31	0.23	0.69	0.65	0.30	0.11	0.07	0.03	0	30
ISPRED4	0.46	0.28	0.25	0.72	0.67	0.32	0.14	0.05	0.03	0	30
BindML	0.70	0.55	0.21	0.45	0.49	0.32	0.12	0.12	0.04	1	27
GHECOM	0.30	0.17	0.26	0.83	0.74	0.28	0.12	0.13	0.06	2	30
Fpocket	0.37	0.21	0.27	0.79	0.72	0.31	0.14	0.16	0.05	0	30

TPR, true-positive rate (i.e., sensitivity); FPR, false-positive rate; MCC, Matthews correlation coefficient; meanMCC, mean of MCC calculated per-structure; SD error, standard error of meanMCC; NA count, structures with absence of either any positive or negative predictions such that MCC cannot be calculated; *n*: number of structures processed.

[a] Number following ACCLUSTER denotes the number of top prediction clusters evaluated.

feature set of the central residue. Such aggregated features have been shown to be able to boost predictor performance [50]. From 2182 original features, 30 were chosen via feature selection. Feature selection and hyperparameters used by XGBoost were optimized for the two models separately through 20-fold cross-validation, measuring improvements in terms of AUC (see Methods). Using the optimized features and hyperparameters, the final models were retrained on the entire MoRF-train set. The resulting core and rim predictors, named CorePred and RimPred, achieved AUCs of 0.84 and 0.75 on the independent MoRF test set (Fig. 2, left). For comparison, we employed the same optimizations used for RimPred to train a gradient boosted trees model that separates all interface from non-interface residues. CorePred and RimPred show an advantage in performance compared to this model trained with all interface residues (Fig. 2, right). This result suggests the gains observed in the individual features upon separating the interface into two regions translate to the CorePred and RimPred models.

CorePred and RimPred scores are combined in the final predictions using CRFs

Next, we combined CorePred and RimPred outputs with a CRF model to create the final output of our new predictor that we call IDRBind. CRFs are discriminative undirected probabilistic graph models that take “neighborhood” context into account, in contrast to more common classifiers like CorePred and RimPred that evaluate each residue individually. Our CRF can be interpreted as having two components: a scoring and an adjacency component (Fig. 3A). The scoring component is trained to integrate residue-level CorePred and RimPred scores. In addition to these residue-level scores, the scoring component also uses the protein-average of CorePred and RimPred scores. The incorporation of these average scores enabled better handling of small proteins, for which predictions are known to be affected by a size bias [51] (discussed below and in Methods). The adjacency component of our CRF incorporates information from neighboring residues. It enforces penalties for spatially isolated core or rim residue predictions and enhances clusters of the same class, thereby smoothing the prediction results and removing outliers. While CRFs have been applied in this field, other predictors have used the protein sequence or distance thresholds to define adjacent residues [15,52,53]. Instead, we chose to use a network mesh generated from the molecular surface of proteins. Specifically, a quadrilateral surface mesh was generated (Fig. 3B and Methods). Each face in the quadrilateral surface was mapped to the closest residue. Because there are many more faces than residues, each residue of the protein can be mapped to multiple faces. Adjacent faces were defined as those sharing the same edge in the mesh, so each face has a maximum of four neighbors.

Faces of the protein mesh surface are represented by labeled variables (nodes) in the CRF that we generated through the software library FACTORIE [54]. Core, rim or non-interface labels of these variables are the prediction outputs of the CRF (Fig. 3A) and thus of IDRBind. Each labeled variable is connected to other variables through factors. Factors contain weights that describe the labels' compatibility with other variables. The scoring component of the CRF has the observed (input) feature variables CorePred, RimPred, CorePred average, and RimPred average. The adjacency component consists of pair-factors that connect adjacent labeled variables (i.e., faces sharing an edge). The prediction output of core,

rim, and non-interface labels is attained through approximate inference, for which we use a belief propagation algorithm for maximum *a posteriori* estimation (see Methods).

As IDRBind predictions are categorical, we calculated well-established performance metrics instead of ROC curves to evaluate IDRBind's performance on MoRF-test (Table 1). In separating non-interface from interface residues, the union of rim and core classes, IDRBind achieves a MCC [55] of 0.31 at a sensitivity (true-positive rate, or TPR) of 51%. While MCC is a general indicator of predictive power, sensitivity measures the fraction of true interface the predictor identifies. To assess whether the CRF improved predictions, we compared IDRBind's predictions with the combined predictions of CorePred and RimPred. A naïve but effective way to combine CorePred and RimPred is to take the larger of the two scores for each residue. Table S3 shows the performance of this naïve model at varying thresholds. It reveals that for any threshold and associated TPR, the MCC of this naïve model is well below that of IDRBind on MoRF-test. Next, we analyzed the effect of protein size on predictions made by IDRBind and the naïve model. As mentioned before, size biases can negatively affect predictions. This issue is particularly important for small proteins that get over-scored by many interface predictors, meaning interfaces are predicted to be too large. This size bias can be eliminated by ranking scores for each protein individually [51], but IDRBind's categorical output cannot be ranked. Consequently, we calculated MCCs for each MoRF-test protein individually and then averaged over the full set. While the mean MCC of IDRBind of 0.31 is identical to the MCC calculated using all interface and non-interface residues in MoRF-test (Tables 1 and S4 for IDRBind performance on individual structures), the mean MCC of the naïve model drops for the majority of the thresholds tested (Table S3). In comparison, the scoring component of the CRF in isolation offers little improvement over the naïve model at comparable decision thresholds, that is, thresholds where the naïve model achieves sensitivity around 0.5 to 0.6, but the higher mean MCC provided by the scoring component relative to the naïve model is notable (Table S5). These observations suggest that the scoring component contributes to mitigating the size bias, while the adjacency component provides most of the performance gain over the CorePred and RimPred models.

IDRBind outcompetes existing predictors in the identification of MoRF-binding sites

We compared the performance of IDRBind on MoRF-test with that of various predictors of protein interfaces. Many more predictors have been developed than we can benchmark, so we selected representatives that allow for informative comparisons. The general interface predictors are trained predominately on complexes between globular proteins (globular interface). Of this group, ISPRED4 [15] consPPISP [16], BindML [17], and CPORT [18] represent a state-of-the-art as well as popular predictors that are often used for performance comparisons. PredUs [12] and PrISE [56] were chosen because they are a special class of predictors that uses structural homology. We also tested predictors of peptide-binding sites on globular proteins (peptide interface), and they are ACCLUSTER [11], PeptiMap [13], ISMBLab (peptide version) [10], and VORFFIP (i.e., peptide prediction from Multi-VORFFIP [14]). Lastly, because binding pockets are a distinctive feature of peptide binding sites, pocket-finding methods GHECOM [57] and Fpocket [58] were also included.

The performance measures used here are threshold dependent. Therefore, the threshold for each predictor was selected to maximize their MCC on MoRF-test. IDRBind's MCC of 0.31 is the highest of all predictors tested on the classification of interface *versus* non-interface residues in MoRF-test (Table 1). Predictors with the closest performance to IDRBind are peptide interface predictors. IDRBind also ranks the highest in terms of precision, followed by ACCLUSTER and PeptiMap. High precision is significant because it indicates the fraction of predicted interface residues that genuinely belongs to an interface, which can be interpreted as the usefulness of the predicted interface. Comparison of the mean MCCs gives IDRBind an even greater lead. Most importantly, the highest precision of the compared tools in conjunction with a sensitivity of 51% indicates that IDRBind produces conservative predictions with adequate coverage of the true interfaces, which makes it practical for MoRF interface predictions.

IDRBind accurately predicts the interface of well-known MoRF interaction partners

Several MoRF-test proteins are well-known proteins that play key roles in regulatory pathways of eukaryotic cells. We highlight some individual examples here to make the discussion on IDRBind's output more tangible. MRG15 is a transcription factor with various roles, including regulating the cell cycle. It associates with an intrinsically disordered region (IDR) of the protein Pf1, forming part of the Rpd3S/Sin3S corepressor complex [59,60]. The interaction involves two distinct hydrophobic sites on MRG15 that are both predicted accurately by IDRBind (Fig. 4A).

Importin-β is another protein found in MoRF-test. Critical to the transport of macromolecules between nucleus and cytoplasm, it interacts with cargo as well as adaptor proteins, including both MoRFs and globular proteins. In this case, the MoRF complex

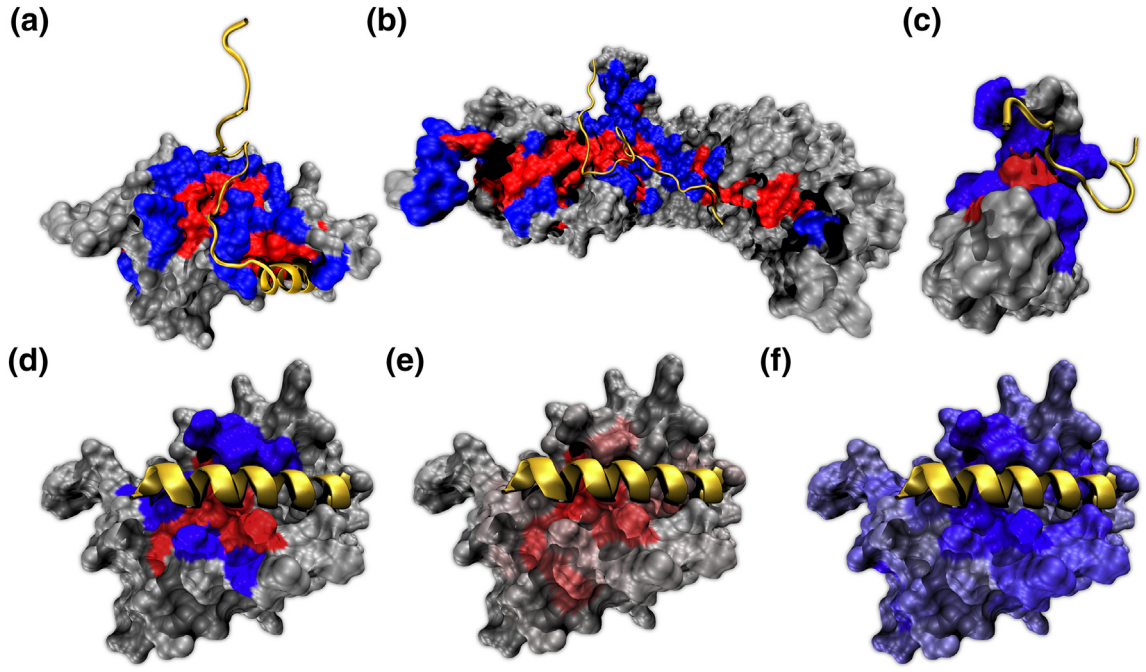

Fig. 4. Prediction results of select proteins mapped on their structures. Molecular surfaces of MRG15 (A), importin-beta (B), Pin1 (C), and Bcl-2 (D) are colored by IDRBind prediction labels for core, rim and non-interface in red, blue and gray, respectively. In panels E and F, Bcl-2's surface is colored by the core and rim scores provided by CorePred (E) and RimPred (F). The prediction labels (IDRBind), scores (CorePred or RimPred) and the protein surface are from the unbound structure. To visualize consistency between predictions and actual MoRF interfaces, MoRFs were placed on the unbound structures by aligning only the globular domains of the bound and unbound structures. The PDB IDs of the unbound and bound structures, respectively, are 2F5J and 2LKM (A), 1GCJ and 1M5N (B), 1I6C and 1I8H (C), and 1GJH and 5VAY (D–F).

is between the human importin-β and the parathyroid hormone-related protein (PTHrP) it transports [61,62]. The prediction was made using the unbound structure of mouse importin-β, which shares 98% sequence identity with the aligned region of the human protein. The concaved surface on importin-β N-terminal that associates with various binding partners was also correctly identified by IDRBind (Fig. 4B).

An example of a small MoRF partner protein that could be detrimentally affected by over-scoring due to size bias is the WW domain of Pin1, which interacts with tau as well as Cdc25 in phosphorylation-dependent interactions that act as regulatory switches [63]. The WW domain consists of only 39 residues and has a flat binding site, increasing the prediction difficulty. However, IDRBind can correctly identify some of the non-interface residues despite their scarcity (Fig. 4C).

Taking an example from outside our MoRF-test, we tested IDRBind on Bcl-2, a key regulator in apoptotic processes and the dysfunction of which is associated with multiple diseases, including cancer. IDRBind prediction correctly identifies the interface on of Bcl-2 that binds to the MoRF BH3 domain of Beclin-1 (PDB ID: 1GJH [64]; Fig. 4D). Figure 4E demonstrates CorePred's high specificity for grooves, practically excluding all protruding surfaces. RimPred prefers convex and solvent-exposed residues but produces more false positives than CorePred (Fig. 4F; for comparison of core and rim residue curvature scores, see also Fig. S3). However, the CRF integrates CorePred and RimPred scores and reduces false positives by suppressing residues that have relatively high RimPred scores but are isolated from the main interface patch (Fig. 4D).

Since MoRFs can exhibit significant structural diversity in their bound states, we specifically selected a MoRF that has diverse bound structures. The transcription activator–coactivator interaction between MED15–ABD1 and the MoRF of Gcn4–cAD constitutes a fuzzy complex [65]. Figure S4 shows the successful IDRBind prediction made on the *bound* structure of MED15–ABD1. While prediction on the bound structure may not reflect true performance, there are reasons to expect positive results. The interaction involves prominent electrostatics, a shallow groove, and hydrophobic patches, all of which are typical of MoRF interactions.

Due to the use of structural information, accurate predictions could be hindered by large conformational changes that occur in the partner upon binding, and calmodulin is a particularly challenging

Table 2. Performance measures on PEP

	TPR	FPR	Precision	Specificity	Accuracy	F1	MCC	meanMCC	SD error	NA count	*n*
IDRBind	0.56	0.13	0.38	0.87	0.83	0.45	0.36	0.38	0.05	0	25
ACCLUSTER1	0.42	0.10	0.38	0.90	0.84	0.40	0.31	0.31	0.09	0	25
ACCLUSTER3	0.63	0.26	0.26	0.74	0.73	0.37	0.26	0.26	0.05	0	25
ISMBLab	0.61	0.22	0.29	0.78	0.76	0.39	0.30	0.30	0.05	0	25
VORFFIP	0.57	0.09	0.48	0.91	0.87	0.52	0.44	0.45	0.07	1	25
VORFFIP-NR	0.32	0.06	0.39	0.94	0.87	0.35	0.28	0.23	0.08	0	11
CPORT	0.48	0.20	0.26	0.80	0.76	0.33	0.22	0.19	0.04	0	25

VORFFIP-NR reports the performance of VORFFIP on a subset of PEP that excludes peptide partners with greater than 95% sequence identity and 90% aligned length with their training data (see Methods).

example demonstrating this problem. Calmodulin is a calcium-binding protein that is involved in many regulatory processes. The interaction between the calcium-loaded calmodulin and GAD leads to substantial conformational changes in calmodulin. IDRBind identifies the part of the interface on the C-terminal domain of calmodulin but misses the region on the N-terminal domain (PDB ID 1DMO, 1NWD; Fig. S5) [66,67]. As illustrated by this example, large conformational changes can distort the interface region, leaving some interface residues spatially isolated or buried in the unbound structure, thus making predictions more difficult. Calmodulin is certainly an extreme case compared to the other MoRF-test and MoRF-train structures. While calmodulin interface residues have a Cα RMSD of 14 Å between the bound and unbound state, 67% of the interfaces on MoRF partners in our sets have an RMSD of less than 3 Å between their bound and unbound states (Fig. S2C). Certain conformational changes are acceptable for IDRBind prediction as evidenced by the calmodulin example and the evaluation of MoRF-test partners that show substantial conformational changes (Tables S1 and S4).

IDRBind identifies peptide but not globular interfaces, concordant with features differentiating IDR interfaces

Our finding that predictors developed for peptide interfaces come closest in their performance to IDRBind on MoRF-test motivated us to test IDRBind on peptide interfaces, specifically, the peptide interface dataset (PEP) used in the development of the peptide interface predictor PeptiMap [13]. IDRBind achieves an MCC and mean MCC of 0.36 and 0.38, respectively, on this peptide interface dataset (Table 2). These MCCs are the highest of the tested predictors, except for the peptide-binding model of VORFFIP, which was trained on some complexes included in the PEP. When VORFFIP was evaluated on the subset of PEP that it was not trained on, its performance was estimated to be substantially lower (VORFFIP-NR in Table 2).

To complete the comparison, we also evaluated IDRBind on globular interfaces. Docking Benchmark 5 [68] is a popular benchmark for docking and is dominated by complexes between globular proteins, so we used it as our globular interface dataset, named DB5 (we customized DB5 for our needs; see Methods). IDRBind achieves an MCC of 0.20 on DB5 (Table 3). Cons-PPISP and PrISE, globular interface predictors that were chosen for their facilities for processing large datasets, achieve higher MCCs than IDRBind on DB5. Notably, IDRBind's sensitivity of 0.33 is especially low when compared to its performance on MoRF-test and PEP. However, this sensitivity does mean that a small portion of globular interface residues is positively predicted by IDRBind, which is well illustrated by the interface prediction made for the cytokine interleukin-2. Interleukin-2's interfaces with its globular receptor proteins coincide with IDRBind's two relatively small predicted interface patches, one of which contains a hotspot that is known to be susceptible to small-molecule binding (Fig. S6; PDB

Table 3. Performance measures on DB5

	TPR	FPR	Precision	Specificity	Accuracy	F1	MCC	meanMCC	SD error	NA count	*n*
IDRBind	0.33	0.13	0.31	0.87	0.79	0.32	0.20	0.19	0.013	3	339
cons-PPISP	0.28	0.07	0.39	0.93	0.83	0.33	0.24	0.20	0.016	5	337
PrISE	0.48	0.20	0.28	0.80	0.76	0.36	0.23	0.26	0.012	0	331

See Table 1 for more details.

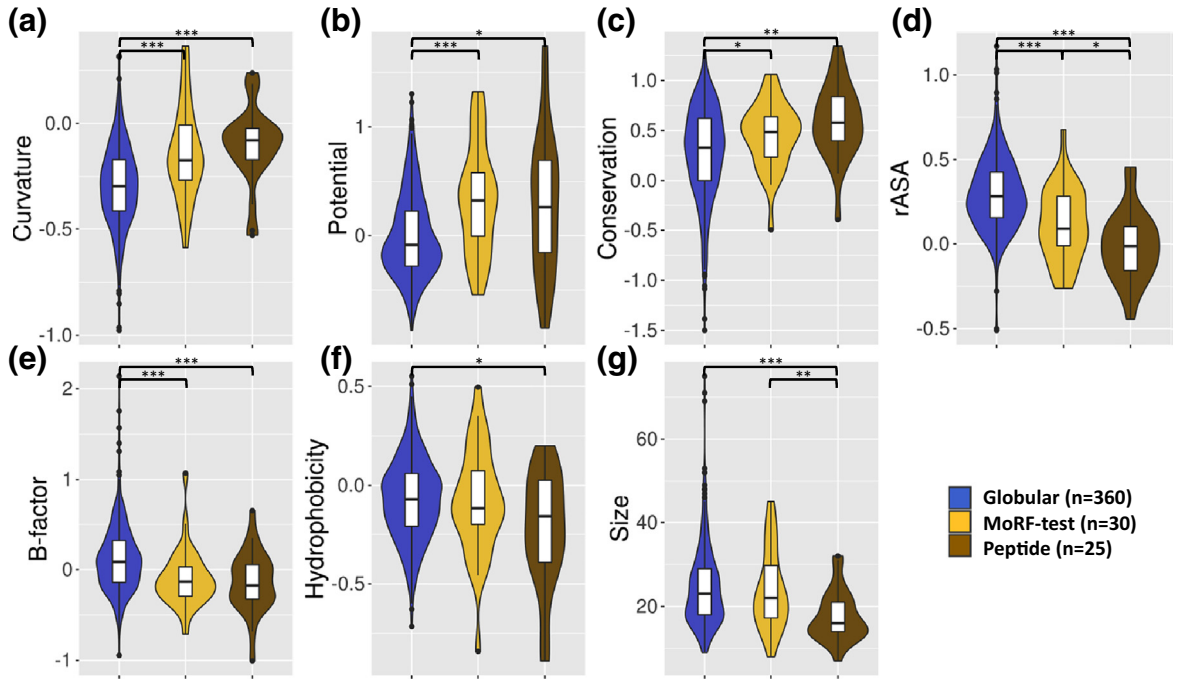

Fig. 5. Feature comparison for globular (blue), MoRF (yellow), and peptide (brown) interfaces. Compared are interface curvature (A), potential (B), conservation (C), rASA (D), *B*-factor estimate (E), negative hydrophobicity score (F), and size (G). Wilcoxon rank sum tests are used to calculate *p*-values, and significance is denoted by asterisks: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

ID 1M4C, 2B5I, 1PY2 [69–71]). Overall, this analysis demonstrates that IDRBind performs well not only on MoRF but also peptide interfaces while having a much lower MCC on globular interfaces.

Given the performance on these three interface classes, we analyzed the interface properties of MoRF, DB5, and PEP complexes to reveal commonalities and differences between them. Figure 5 shows the averaged *Z*-scores (see Methods) for several distinguishing features of residues found in the interfaces of the three types of complexes. As a reference, we also provide the interface size for each (Fig. 5G). *Z*-scores are chosen because they are more informative for features such as conservation and *B*-factor estimates that vary between individual proteins and thus needed to be normalized. Both MoRF and peptide interfaces have significantly higher averaged feature scores than globular interfaces for the curvature, electrostatic surface potential, and conservation. The feature that best distinguishes MoRF and peptide from globular interfaces is curvature, suggesting that the former have more concaved surfaces than the latter. Consistent with this interpretation, MoRF and even more so peptide interfaces have significantly lower rASA and estimated *B*-factor values compared to globular interfaces, indicating that residues in MoRF and peptide interfaces tend to be more removed from the solvent in those concave surface regions and are generally less dynamic. While these feature score distributions appear to place MoRF and peptide interfaces closer together, the overlap in all three classes is substantial. Furthermore, the un-normalized values of the same features (Figs. S7 and S8) place MoRF in the middle of the three interface types or even closer to globular interfaces. More importantly, MoRFs are consistently placed between peptide and globular interfaces in the spectrum of features characterizing protein interfaces, mirroring IDRBind's performance on the three major classes.

Discussion

The accurate prediction of protein interface residues is a key challenge in computational biology that pushes the boundaries of our understanding of protein–protein interactions and machine learning. Numerous methods have been developed that predict interaction interfaces, and the majority are intended for general interface prediction, which is dominated by globular interfaces [3,16–18,52,56,72–74]. More recently, the need for more specialized prediction was recognized, leading to peptide interface predictors [10,11,13,14]. Here, we introduced IDRBind, a structure-based prediction method designed for identifying protein surface residues with the potential to interact with IDRs, with an emphasis on MoRFs, filling the gap between peptide and globular interface prediction. A differentiating factor of IDRBind is its architecture. In the first step, the two prediction

modules CorePred and RimPred predict core and rim interface residues based on their distinct characteristics. In the second step, the CorePred and RimPred scores are combined by a unique grid-structured CRF that considers the compatibility of neighboring residue predictions and favors the output of interface patches. This architecture solves several challenges we identified for predicting interface residues.

One of the challenges is in fully utilizing the distinctiveness of core and rim residues despite the overlap between the two regions. The core region is enriched in conserved residues critical for the interaction and is distinct in multiple features, most notable are surface geometries such as grooves and curvatures that are specific to binding sites of IDRs, including both MoRFs and peptides. Furthermore, core residues are fewer in number and clustered in a narrow region. Accordingly, CorePred predicts fewer residues but identifies core residues with higher accuracy measured at performance-optimizing thresholds. Analogous to a gasket that keeps water out, rim residues surround the core and are more often polar or charged. Rim residues are spatially close to non-interface, and many of its residues are likely not functionally critical, which is reflected in their weak correlation with conservation scores. The weakly discernable characteristics of individual rim residues lead RimPred to rely more on the surface patch feature scores and produce noisier predictions compared to CorePred. Despite the differences, the two regions overlap since they are defined based on artificial rASA thresholds that do not directly correspond to the other physicochemical differences between the residue classes (see Methods). The goal is to use the characteristics of core and rim to predict interface and not to classify core *versus* rim. To that end, the usage of two models allows explicit optimization on core and rim independently while forgoing penalties for misclassification between them during training.

In the subsequent step, the CRF handles the challenge of integrating CorePred and RimPred while adjusting for large scale dependencies between residues. First, because there is a disproportionately large ratio of interface over non-interface residues in small proteins, protein interface predictors tend to over-score smaller proteins, leading to the size bias. This problem could be exacerbated by the graph model during inference (prediction) as the long-distance coupling between labeled variables in these over-scored proteins drives all labels to the interface classes [75]. By applying a penalty to proteins with high average CorePred and RimPred scores, we increased the mean MCC value on MoRF-test, which suggests improved handling of the size bias. Second, gradient boosted trees, support-vector machines and random forest are more commonly used and well-proven methods, but they weakly account for spatial relationships through features that include local neighborhood properties and do not generate the interface clusters that are desired. CRFs can better handle the spatial dependencies of protein residue classification, and there is precedence for this use case for CRFs [15,52,53]. ISPRED4 uses CRF for this purpose, and the authors noted an increase in precision, which is consistent with our observations for IDRBind. This increase in precision likely comes from removing outlying residues that score relatively high on CorePred or RimPred but do not have adjacent interface residues. The use of CRF leads to a noticeable improvement in performance. Notably, the improvement is achieved despite the fact that some neighborhood properties are already integrated into CorePred and RimPred, demonstrating the high effectiveness of the CRF.

Thanks to this architecture, IDRBind achieves a high MCC of 0.31 on the MoRF-test set, which is comparable to the MCCs of general predictors on globular interfaces. Specifically, compared with two routinely benchmarked predictor, IDRBind's MCC surpasses or is comparable to the MCC cons-PPISP and PredUs achieve on globular datasets [15,76]. For more perspective, PredUs's performance always closely follows the top predictor in benchmarks showcasing state-of-the-art general predictors [10,56]. This implies that the usability of MoRF interface predictions approaches that of predictions made for globular interfaces. This is also true regarding peptide interfaces. However, benchmarking results should always be interpreted with care. Comparing between predictors is inherently hard due to differences in evaluation methodology as well as unknown overlaps between training and testing datasets [9]. A striking example of the effect of overlapping datasets is presented in Table 2 where the exclusion of PEP peptide interfaces that VORFFIP was exposed to during training reduces its estimated performance by a large margin. Further complicating the matter is the incomplete annotation of interfaces due to either the criteria used in defining interfaces or incomplete structural data. Inadequacies aside, IDRBind outcompetes existing predictors on the MoRF-test. Specifically, IDRBind achieves the highest MCC on this set.

Comparison of MCCs reveals that peptide interface predictors such as ACCLUSTER and PeptiMap are the most comparable to IDRBind. Just like how IDRBind is created for MoRFs yet performs well on peptide interfaces, these peptide interface predictors show the same is true in reverse. IDRBind performs not only well but even better on peptide interfaces than MoRF interfaces that it was designed for because key feature scores that CorePred and RimPred rely on favor the prediction of peptide interfaces even more than MoRF interfaces. Prominent among key features that show this trend are surface geometry scores and rASA. Conversely,

these scores are not as effective for the prediction of globular interfaces, resulting in a lower performance of IDRBind on the DB5 set. The distinct feature distribution present in peptide, MoRF, and globular interfaces may be, in part, explained by the distinct thermodynamic and kinetic aspects of binding of these protein regions. Being intrinsically disordered, both MoRFs and peptides incur a larger average cost in entropy for binding than globular domains. Likely compensating for the entropy cost, MoRF and peptide interfaces show tendencies to consist of residues that are more hydrophobic, buried, and rigid, which are properties that thermodynamically favor binding. MoRFs are also noted for their central role in many signaling pathways. Notably, strong electrostatic interactions are associated with regulatory proteins [77]. The electrostatic potential is marginally stronger at MoRF and peptide interfaces of the unbound partner structures as measured by the electrostatic potential feature score (Fig. 5B). Strikingly, the electrostatic contributions to binding depend on the complementarity between the binding proteins, yet the electrostatic potential on the partner surface alone was strong enough to play an important role in our prediction model (Fig. S9). Regarding IDRs, this long-range force is proposed to promote fast on and off kinetics beneficial to signaling and regulatory switches [78].

While the characteristics above are shared between all IDR interfaces, there is one distinguishing aspect between MoRFs and peptides. MoRFs are longer than peptides, allowing MoRFs to adopt more complex folded conformations and resulting in more variation in interface size and morphology. MoRF interfaces span a broad range in sizes, and the larger interaction surfaces bring the characteristics of MoRF interfaces closer to globular interfaces. An attribute that correlates with interface size is the magnitude of binding-induced conformational change in globular complexes [26,79]. Concordantly, the relatively small peptide interfaces have been demonstrated to be essentially unchanged upon binding [43], in contrast with the range in RMSDs between the bound and unbound structures of MoRF partner that we found. Conformational changes most directly affect geometry-based features such as the surface groove score. These factors dilute the feature scores that are key indicators of IDR interfaces, increasing the difficulty of predicting MoRF interfaces compared to peptides and requiring IDRBind to have a higher sensitivity to these features. The result is a predictor that can handle a broader range of IDR lengths compared to peptide interface predictors.

Given the prevalence of IDR-mediated interactions in complex biological pathways, IDRBind could be useful for high throughput analysis. With the accumulating protein sequence and structural data, computational analysis is fast becoming a key method for understanding biological pathways. For example, Interactome INSIDER is a tool that integrates protein–protein interactions, genetic variation, and structural information from both experiments and prediction [50]. However, given an estimated 122,000 MoRFs in the human proteome [20], a specialized method like IDRBind could be useful for filling in the gaps from experiments and other prediction methods. There is eagerness in the community for mapping protein–protein interaction interfaces as well as modeling their structures, and for IDR-mediated interactions even more so due to their potential as interesting drug targets. Protein–protein interfaces are considered more difficult to target than traditional small-molecule binding sites because protein interfaces are relatively flat and much larger [80]. It is likely easier to design drugs to disrupt MoRF and peptide interfaces, which have grooves and pockets. In fact, the MoRF-binding site of Bcl-2 is the target of drugs under development (Fig. S10) [81].

In summary, IDRBind is a structure-based prediction method designed for identifying protein surface residues with the potential to interact with MoRFs. MoRFs-mediated interactions are ultimately a subclass of protein–protein interactions that cannot be precisely discriminated based on binding interface properties alone. Because MoRF and peptide interfaces share characteristic features that our predictor exploits, IDRBind is a leading predictor of IDR interfaces spanning both MoRFs and peptides. To facilitate the usage of the new method, IDRBind predictions are available through a web server at https://idrbind.msl.ubc.ca/.

Methods

Construction of protein complex datasets

MoRF complex datasets were built using the IDEAL database and a literature search [45]. MoRF complexes comprise a MoRF interacting with one or more folded domains (i.e., globular proteins), except for interactions between two MoRFs. A folded protein segment mediating an interaction can be classified as a MoRF if there is experimental evidence that it is intrinsically disordered, meaning it does not have one well-defined tertiary structure in its unbound state. The 2016 July release of IDEAL database was used, which provided 183 MoRFs complex structures. When combined with our literature search, this resulted in 229 MoRF complexes. Complexes with redundant MoRF partners were removed using a length-dependent sequence identity threshold defined by Rost with parameter n of zero [82]. We identified unbound structures of MoRF partners through a BLAST search of the PDB

database, excluding structures with DNA and RNA molecules. We used NCBI's blast2 program to search for proteins with 95% or higher identity and 90% or higher aligned length [83]. The quality of the structures was checked manually, excluding antibodies and some unbound structures with small molecules blocking the MoRF interface, resulting in 84 non-redundant MoRF partners, 57 of which have unbound structures. Thirty unbound structures were selected randomly for MoRF-test, and the remainder were placed into MoRF-train.

Globular (DB5) and peptide (PEP) protein datasets are based on established datasets of bound and unbound structures that have already been used to benchmark the performance of multiple existing predictors [43,68]. Docking Benchmark 5 was modified by removing antibody complexes, complexes where the bound and unbound protein sequences were not well-aligned, and complexes where the interface in the unbound structure was obstructed. The resulting DB5 has 360 interacting domains, with some domains appearing more than once but with unique interactions. These repeated domains are included in the feature score comparisons between interface classes, but they are excluded in performance measures for which we selected only one interface per domain (Table 3). Importantly, because our pipeline for benchmarking MoRF and peptide interface focuses on one specific interface for each complex, as opposed to accounting for all general interfaces given a multimeric complex, benchmarks on DB5 will underestimate the performance with regard to general interface predictions. Extensive assessments of general interface predictors on globular interfaces were done in other studies and were not repeated here [56,76]. For instance, ISPRED4's performance evaluation on Docking Benchmark 5 represents a more accurate assessment for predictors on general interfaces [15]. The PEP dataset was derived from the 30-peptide-complex test dataset of ACCLUSTER, which was sourced from the database peptiDB [11,43]. Because peptide and MoRF complexes are similar, peptide binding partners (i.e., peptide partners) in PEP that are also in the MoRF datasets had to be removed. CD-HIT-2D from the CD-HIT package was employed, using the parameters "word_length" of five, sequence identity of 0.6, and length difference of 0.5 to remove redundant peptide partners [84]. Five out of the original 30 PEP structures were removed. All the peptide interface predictors evaluated on MoRF-test were also benchmarked with PEP, excluding PeptiMap because it was optimized on PEP.

Surface, core, and rim residue definition

Surface, core, and rim residue classes were defined using SASA calculations. Areaimol from the CCP4 suite was used to calculate SASA [85]. The segregation of core and rim residues has been shown to be effective in analyzing protein interfaces [86]. A modified set of criteria for core, rim, and surface residues was devised to increase the number of surface residues and maximize the effectiveness of core and rim predictions. For each complex structure (i.e., MoRF bound to a MoRF partner), Areaimol was used to calculate the SASA of both the protein complex and the MoRF partner protein in isolation, where the isolated MoRF partner is simply the complex structure with the MoRF atoms removed. The SASA of each residue was then normalized by its SASA in a Gly-X-Gly peptide in an extended conformation to generate rASAs [86]. Surface residues were then defined as those with an rASA $>5\%$ in the isolated MoRF partner. Of these surface residues, those with a change in rASA when going from the isolated MoRF partner to the complex structure were identified as interface residues. The core contains interface residues with less than 7% rASA in the complex, while the remaining interface is part of the rim. Lastly, the residue class labels were mapped to an unbound structure if available. The same definitions were used for globular-globular and peptide-globular complexes.

Prediction evaluation comparisons with existing predictors

There are differences between predictors, so our evaluation method aims to convert the different prediction outputs to a consistent format while making the assessments as fair as possible. For instance, IDRBind predicts core and rim class labels, which must be merged into a single interface class for evaluation (see the Quantification and statistical analysis section for details on scoring metrics). Some predictors, including IDRBind and ISPRED4 [15], only make predictions on surface residues, so only surface residues were evaluated across all predictors. Individual residues without prediction scores were ignored and not held against the predictor. This could occur because of differences in definitions of surface residues or issues with interpreting the input structure.

Some predictors output multiple interface clusters, as opposed to scores for individual residues. These clusters must be converted to interface and non-interface residue labels to conform to our evaluation scheme. This is more common among predictors of ligand binding pockets, that is, GHECOM [57] and Fpocket [58], but ACCLUSTER [11] and PeptiMap [13] also do this. Their prediction outputs are in the form of multiple binding sites. While each site could denote a separate interaction, we chose to combine them because the IDR interfaces are large in comparison, especially those of MoRFs. Each site

is composed of either a cluster of partner residues or a cluster of probes mimicking the predicted ligand positions. In the latter case, we used a distance threshold of 4.5 Å to label the probes' adjacent protein residues as interface for evaluation. Furthermore, the individual clusters are often ranked based on prediction confidence. Thus, we tried to select and merge the optimum number of the top-ranking clusters that provided the best performance for the predictor. The number of clusters selected for ACCLUSTER is indicated by the number following its name in Tables 1 and 2. Using three top clusters from PeptiMap prediction resulted in its highest performance in MoRF-test. Whereas the top two pockets from the Fpocket predictions were utilized. Lastly, only the top pocket was evaluated from the GHECOM predictions.

Similarly, other predictors were also evaluated such that their reported performance is reasonably optimized for the respective benchmark datasets. Predictors that output individual residue scores were evaluated at multiple thresholds that were selected to give reasonable estimates of maximum MCC on each testing dataset. The performance at the threshold resulting in the highest MCC is reported for each predictor in Tables 1 to 3. In addition, some servers provide multiple models. ISMBLab [10] consists of a family of predictors, including one for peptide interfaces, which we evaluated in Tables 1 and 2. Multi-VORFFIP also provides multiple models. We evaluated the peptide interface model (i.e., PDB file output labeled as EBS) from Multi-VORFFIP [14]. Analogously, the PredUs2 model combines two separate components. The SVM component of PredUs2 provided better results in MoRF-test, so we chose to report that instead of the performance of the full predictor [12].

There are two exceptional cases where we altered the datasets to evaluate the predictors better. One such case is in the evaluation of Multi-VORFFIP on the PEP dataset where there is a substantial overlap between their training set and PEP. The dataset used by Multi-VORFFIP was sourced from Petsalaki *et al*. [87]. We ran CD-HIT-2D with the word length of 4 and the thresholds 95% identity and 80% alignment length. CD-HIT-2D clusters the sequences and returns a subset of the PEP with the redundant sequences removed. The result is reported in Table 2. Another case is in the evaluation of ACCLUSTER, which allows the option to input peptide sequences up to 30 residues in length. Thus, we reported the evaluation on a subset of MoRF-test that have MoRF sequences satisfying this length restriction in Table S6 for completeness.

Having outlined the general criteria used in the evaluation process, we will list some of the parameters and options we selected for individual predictors for the interested readers. Most often, we used default parameters because we assumed them to be the suggested settings. However, we did try some parameters that seemed better suited to our use case. ISMBLab outputs numeric scores as well as clustered binary predictions, and we chose the latter due to the higher performance. On the other hand, we chose the numeric predictions over the binary predictions for cons-PPISP. Similarly, the probability output of PrISE at the threshold chosen based on MCC provided higher performance than its binary output for both MoRF-test and DB5 [56]. PrISE server predictions were calculated with the inclusion of highly homologous proteins to increase its performance. CPORT predictions were carried out with the threshold set to sensitive to provide sensitivity comparable to the rest of the predictors [18]. BindML predictions were obtained through their web server instead of their downloadable software. The parameters used for the two pocket detection methods were set in an effort to obtain larger interface pockets compared to the default options. GHECOM predictions were calculated using multi-scale pocket detection with parameters "-rs" of 2.5, "-rli" of 4, "-rlx" of 13, and "-br" of 1. Fpocket predictions were calculated using parameters "-M" of 13, "-l" of 45, "-r" of 5, and "-D" of 2.

As mentioned above, there are differences between the subset of residues that are processed by each predictor, so we present a corrected MoRF-test benchmark for completeness. IDRBind does not return residues that are largely buried because they lack some feature scores, such as the groove scores that are only calculated for residues close to the surface. Tables S7 and S8 provide the performance measures of predictors assessed on only the subset of residues with IDRBind predictions.

Calculating feature scores

Conservation

Conservation scores were calculated using methods and software from Capra and Singh [88]. The calculation of these conservation scores required multiple sequence alignments (MSA), which we generated using a procedure we optimized for protein interfaces. Specifically, sequence homologs from the NCBI nr table (ftp:/ftp.ncbi.nlm.nih.gov/mmdb/nrtable/) [89] were identified using BLAST (blastp version 2.2.28+) with the *E*-value threshold 0.000001. Homologous sequences were only used when the sequence identity was greater than 60% by default, but a lower threshold of 40% was used if fewer than 50 sequences were initially returned. MSAs were built from the homologous sequences using MAFFT (version 7.407) with the fftns option [90,91]. Sequences that were aligned to less than 80% of the reference sequence length were removed. Subsequently, alignment gaps were removed from the reference sequence, and a new

MSA was built with MAFFT. Conservation was calculated using this MSA. Two *Z*-scores were calculated for the conservation scores over individual proteins, with one based on all residues and the other only on surface residues (see Table S2).

Electrostatics and groove

Electrostatics-derived features and groove scores were both measured from a 3D grid built around the protein. The grid points were spaced 1.2 Å apart and extended five grid points beyond the X, Y, and Z coordinate extremities of the protein. To determine whether each grid point is solvent exposed, we used EDTSurf to calculate the triangulated solvent accessible surface and used Inpolyhedron, a Matlab module by contributor Sven in File Exchange (https://www.mathworks.com/matlabcentral/fileexchange), to identify the grid points outside the triangulated surface (i.e., solvent-exposed) [92]. To calculate groove scores, we then used an approach very similar to the one employed by pocket-finding method LIGSITE [93]. In short, binding pockets are large indentations on the protein surface that can be identified by searching for void spaces that are surrounded by the protein on multiple sides. In detail, at each solvent-exposed grid point, we scanned along several evenly spaced lines (e.g., the axes) that intersect the target point, searching for the protein surface. Lines that intersect the protein on both ends were counted for the calculation of groove scores. Thus, if we scan along six lines, the score will range from 0 to 6, where a score of 6 suggests that the grid point is surrounded by protein in 12 directions. We can determine whether each line intersects the protein because the lines were placed along the 3D grid where each point was labeled as inside or outside of the protein. Each line was extended from the target point by 12 grid points in both directions. Two groove score variants were calculated with 6 and 13 lines per grid, respectively, and the feature scores were mapped onto protein residues using a 4.5-Å threshold to define adjacent atoms. The score was then averaged over each residue.

The electrostatic potential was calculated using DelPhi, and the related feature scores were derived from the potential and electrostatic field measured at the solvent-exposed grid points [94]. To prepare the protein structure for DelPhi, Profix from the JACKAL package was used to fill in missing atoms and Reduce to replace or add hydrogens [95–97]. The processed PDB file was then formatted with CHARMM param22 atom names using the MMTSB Tool Set [98,99]. The parameters used for DelPhi are similar to those used previously (see Supplementary Methods for details) [27,100]. Electrostatic potential and electrostatic field gradients were mapped to the residue atoms within a 4.5-Å threshold and then averaged over each residue, resulting in feature scores for potential and field. The electrostatic potential scale has both positive and negative values, so we also calculated the absolute value of the potential, which could better reflect the strength of the potential. For additional variants of field scores, see Supplementary Methods.

Residue composition

Residue composition scores were calculated using principal components of amino acid indices generated from the protr package in R [101]. This package provided 531 complete indices from the AAindex database [102] for the 20 canonical amino acids, and principal component analysis (PCA) was used to obtain orthogonal indices. The first five principal components were used, which accounts for 95% of the variance of the original indices. Each residue was assigned its indexed scores for principal components 1 to 5 to obtain residue composition scores 1 to 5, and we also calculated surface patch averaged scores for each residue together with its neighboring residues within a 13-Å radius (see list in Table S2). Notably, residue composition score 1 has a strong inverse correlation with hydrophobicity, with lower values for hydrophobic residues [Pearson correlation with hydrophobicity index 0.93 (AAindex ID: FASG890101); $p = 1.6 * 10^{-9}$].

B-factor

B-factor estimate scores were based on the thermal fluctuation estimates from Gaussian network modeling (GNM) provided by the ProDy Python package [103]. GNM calculations for proteins were carried out with a 7.3-Å cutoff for pairwise interactions between α-carbons.

Curvature and roughness

Besides groove scores, we also used the surface geometry features curvature and roughness. The curvature scores were calculated with the help of Surface Racer [104]. Surface Racer is set to calculate molecular surface using a probe radius of approximately 1.4 Å, adjusting the radius at 0.01-Å increments if the program fails on a protein. The roughness score of protein surfaces was calculated using rufness from the HotPatch package [105,106].

Feature patterns 1 to 5

To capture the unique elongated profile of MoRF interfaces, we devised the additional feature scores pattern 1 to 5. MoRF interfaces have core residues in

the center with rim residues surrounding them, creating elongated donut shapes. Core residues have higher groove and conservation scores that contrast with those of the rim. Therefore, we combined the groove and conservation scores and searched the protein surface for such patterns. More details on the feature pattern 1 to 5 can be found in the Supplementary Methods.

Aggregated feature scores

Additional feature scores were derived by aggregating neighboring residue scores. Surface patches were defined using 6-, 9-, and 12-Å thresholds between α-carbons. The individual feature scores of the residues in each surface patch were then aggregated. The aggregation functions we used are taking average, average of absolute values, maximum value, minimum value, average of top 50% of values, and average of bottom 50% of values. In addition, *Z*-scores were calculated based on the original feature scores and the aggregated feature scores. The *Z*-scores of each residue were calculated with respect to the surface residue scores of the protein to which it belongs.

Constructing CorePred, RimPred, and IDRBind CRF

CorePred and RimPred were both trained on the MoRF-train set using the XGBoost library in R [49]. They are both gradient boosted trees models built using the binary logistic objective. These two models were optimized through feature selection and hyperparameter optimization, and 20-fold cross-validation was used for evaluation in both processes. Residues from the same protein were grouped together during cross-validation since adjacent residues may have very similar feature scores. For feature selection, features were grouped into conservation, groove, curvature, roughness, field, potential, *B*-factor estimate, rASA, residue composition 1 (i.e., hydrophobicity), residue composition 2, residue composition 3, residue composition 4, residue composition 5, Pattern1/2/3, and Pattern4/5. Grouping of features into categories was inspired by Meyer *et al*. [50]. This reduces the search space of feature selection in an expert-guided manner. The initial set of 30 features was selected based on the highest Pearson correlation to the core or rim residues within each feature group, selecting two from each group. The R package rBayesianOptimization was used to optimize XGBoost hyperparameters based on AUC performance. Sampling points of the hyperparameter space were determined by rBayesianOptimization's "ucb" function at kappa of 1 and epsilon of 0. Unfortunately, optimization of more than five parameters at once is time intensive with this method. Therefore, we chose to optimize two sets of XGBoost hyperparameters separately. Parameters1 consists of nrounds, max_delta_step, and gamma, while Parameters2 consists of max_depth, eta, colsample_bytree, min_child_weight, and subsample. Starting with a rough set of manually selected parameters, we first optimized Paramter1 with 80 random initial points and 20 iterations of Bayesian optimization. This was followed by optimization of Paramters2 with 100 random initial points and 30 iterations of Bayesian optimization. The resulting set of hyperparameters was used during feature selection, which involved iterations through the 30 features and replacing each with features of the same category and selecting the feature resulting in the highest AUC in cross-validation. For CorePred, the aggregated feature scores were removed, except for residue composition features averaged over 13-Å surface patches. Following feature selection, fine adjustments were made to Parameters2 with 100 initial points and 30 iterations of Bayesian optimization.

To generate the CRFs, a network representation of the protein surface was created for each protein using the EDTSurf and Instant Meshes programs. EDTSurf was used to calculate a triangulated molecular surface of the protein using a 1.42-Å probe [92]. Instant Meshes takes the triangulated surface and converts it to a quadrilateral mesh (Fig. 3B) [107], which we used to define the graph nodes and edges. The desired scale of the edges was set to 3.5 Å along with the following settings: two smoothing steps, deterministic algorithms, intrinsic mode, and align to boundaries. The scaling of the mesh was selected with the goal of representing each low rASA residue through at least one quadrilateral face while limiting the number of faces associated with each highly solvent-exposed residue. The averaged coordinate of the four vertices of each face was mapped to the closest residue with an upper limit distance of 6 Å. Faces that share an edge were considered neighbors, defining the edges of the adjacency component of our CRF.

The CRF model combines the output scores from CorePred and RimPred while accounting for the compatibility between neighboring residue labels (Fig. 3A). FACTORIE, a software library for creating factor graph models, was used to generate CRFs [54]. Factor graphs are bipartite graphs with factors and variables. Factors contain the functions describing the compatibility between variables with weights that are optimized during training. The factors we used are template factors, meaning that factors of the same type share the same weights across the whole graph and all graphs. The variables include the feature variables (observed) and the labeled variables (unobserved), where the labeled variables are defined by the faces on the mesh surface described above and each one has a discrete label core, rim, or non-interface. The feature variables

were generated as a vector of four continuous values: CorePred score, RimPred score, and the protein averages of the two scores. The protein-averaged scores were introduced to counteract a protein size bias in prediction scores [51]. Thus, our model applies a penalty to core and rim labels for proteins with high averaged CorePred and RimPred scores.

Our CRF consists of the scoring component and adjacency component. For each labeled variable, the scoring component is composed of a feature variable vector, a factor describing the feature variables' compatibility with the labeled variable, as well as a factor associated with class bias (Fig. 3A). The class bias factor provides one weight value per label class, allowing adjustments to the weight of each class during training. The adjacency component consists of the paired-factors defined using the surface mesh, connecting neighboring labeled variables. Using a quadrilateral mesh to define connections of the adjacency component of our graph model ensures that each labeled variable is limited to a maximum of four neighbors. Thus, the paired factors making the connections will exert constraints evenly across all labeled variables. In summary, the scoring component combines CorePred and RimPred by applying weights to both and adjusting for size bias, while the adjacency component accounts for the compatibility of neighboring labeled variables.

The two components of our graph model were optimized sequentially. First, the scoring component was trained on the CorePred and RimPred scores from MoRF-train that were calculated using the leave-one-out method, calculating prediction scores on one protein by using XGBoost models built on the rest. For training, we adjusted the representation of each class label by oversampling. We did this for each protein individually by starting with the initial set of labeled variables and added to it by sampling with replacement the under-represented label classes until we had 33% core, 33% rim, and 34% non-interface. Thus, the three classes of labeled variables were roughly equally represented during training. In FACTORIE, a likelihood Example is a data structure encapsulating a training example and calculates gradients using maximum likelihood during training. Each labeled variable provided two training examples: one that treats the interface as a single class, and one that also evaluates core *versus* rim classification. This encouraged the model to learn interface detection instead of putting excessive weights on core and rim segregation. The weight parameters were optimized through batch training 60 iterations on the Examples with the AdaGrad optimizer and parameter averaging [54,108].

Subsequently, the adjacency component was optimized on top of the fully trained scoring component based on the same MoRF-train set. The adjacency component consists of just one type of factor, which is a paired factor touching two labeled variables. The weight parameters for these paired factors were decided upon with the intent to favor clustering of nodes with the same class label, counting core and rim as one class. The idealized model of the interface regions also places the rim around the core, so the case of core being adjacent to non-interface was made to be the most unfavorable. More specifically, the weight tensor of the paired factors between labeled variables was set manually to 1 for matching classes, 1 between core and rim, −1 between core and non-interface, and −0.5 for the rim and non-interface. This weight tensor was modified by multiplying a coupling strength parameter *c*. For the optimization of this parameter *c*, the MCC was evaluated for predictions made on MoRF-train while varying *c*. The resulting MCC was plotted with *c*, and a polynomial of degree two was fitted in order to select the *c* that maximizes the MCC. For the final prediction, maximum a posteriori (MAP) estimation on the combined two-component model is accomplished through loopy belief propagation [109]. Ten steps of loopy belief propagation are enough for convergence.

Finally, the predicted graph labels are mapped back to the protein residues. This is done hierarchically with core taking precedence over rim and rim taking precedence over non-interface. For example, if at least one labeled variable associated with a residue is predicted to have a core label, the whole residue takes on the core prediction.

Figure generation

Figures of protein structures were generated using VMD with the molecular surfaces calculated from MSMS [110,111]. Figure 3B is the exception where the molecular surface was calculated with EDTSurf and the figures are screenshots from Instant Meshes [92,107]. The placement of MoRFs onto the unbound structures of MoRF partners was done through aligning the bound and unbound MoRF partner structures only using the MultiSeq alignment tool in VMD [112]. Figures of protein structures were edited using GIMP (https://www.gimp.org/). The R package ggplot2 was used for the violin plots [113]. The *p*-value labels were calculated using the Wilcoxon rank sum test in R.

Quantification and statistical analysis

To evaluate the full range of values for the feature scores as well as prediction modules such as CorePred and RimPred, the ROC curves and AUCs were generated using the ROCR package [114] in R (https://www.r-project.org/). For the evaluation of predictions with binary values, or with thresholds applied to continuous values, we use the

following measures. True positive (TP) and false positive (FP) are the numbers of residues correctly and incorrectly classified as interface, respectively. True negative (TN) and false negative (FN) are the numbers of residues correctly and incorrectly classified as non-interface, respectively. TPR, also referred to as sensitivity, is defined as TP/(TP + FN). False-positive rate (FPR) is defined as FP/(FP + TN). Precision is TP/(TP + FP). Specificity is 1-FPR. Accuracy is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

F1 score is defined as:

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

MCC is:

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

Structure RMSD calculations

Structure alignment and RMSD calculations presented in Fig. S2 and Table S1 were performed using the McLachlan algorithm implemented in the program ProFit (http://www.bioinf.org.uk/software/profit/) [115]. All alignments and calculations were carried out on Cα atoms. For each NMR MoRF complex structure, the MoRF's primary interface residues were defined as residues classified as part of the interface in at least 50% of the models using the rASA definition of interface residues described above. The flanking residues consist of the residues that are not part of the primary interface. The RMSDs of the MoRFs were calculated by aligning the structure of the MoRF partner before calculating the RMSD of the specified MoRF residues (i.e., all residues, primary interface, flanking). The mean RMSDs were calculated by taking the quadratic mean of pairwise RMSDs between all NMR models [116]. Five out of 38 MoRFs were omitted from the analysis because they were missing flanking regions. The interface RMSD between the bound and unbound states of MoRF partners was calculated by aligning the non-interface residues and subsequently calculating RMSD of the interface residues. The RMSD of the whole MoRF partner structure was also calculated for comparison. Whenever multiple NMR models exist for the MoRF partner structure, the first model was used.

Prediction server

A web server was developed to allow users to utilize our prediction method. Using protein structure and sequence submitted in the PDB and FASTA format, respectively, the IDRBind server returns results in a modified PDB file with the class label in the *B*-factor column for viewing in programs such as VMD [110]. In addition, the user can download a text file containing the IDRBind class labels, CorePred scores, and RimPred scores. By default, prediction jobs are placed in a queue displayed in the queue page where links to the result page of each job are shown. Users can also opt to have their jobs remain private, which hides them from the queue page.

Data and software availability

The IDRBind method is available on the web server at https://idrbind.msl.ubc.ca/. The MoRF-train and MoRF-test datasets used in training and benchmarking are available in Table S1.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.06.010.

Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada, the University of British Columbia, and the Michael Smith Foundation for Health Research for research support. We also thank Matthew Jacobson for the development of the website, and Nawar Malhis for useful discussion on the website development and data analysis.

Received 8 March 2019;
Received in revised form 1 June 2019;
Available online 15 June 2019

Keywords:
protein–protein interactions;
protein interface prediction;
intrinsically disordered proteins;
molecular recognition features;
protein interface prediction benchmarking

Abbreviations used:
IDR, intrinsically disordered region; MoRF, molecular recognition feature; CRF, conditional random field; ROC, receiver operating characteristic; AUC, area under curve; TPR, true positive rate; FPR, false-positive rate; MCC, Matthews correlation coefficient; SASA, solvent-accessible surface area; rASA, relative accessible surface area.

References

[1] E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, A. Godzik, A Pan-Cancer Catalogue of CANCER DRIVER PROTEIN INTERACTION INTERFACES, PLoS Comput.

Biol. 11 (2015)https://doi.org/10.1371/journal.pcbi.1004518.
[2] O. Keskin, N. Tuncbag, A. Gursoy, Predicting protein–protein interactions from the molecular to the proteome level, Chem. Rev. 116 (2016) 4884–4909, https://doi.org/10.1021/acs.chemrev.5b00683.
[3] S. Jones, J.M. Thornton, Prediction of protein–protein interaction sites using patch analysis, J. Mol. Biol. 272 (1997) 133–143, https://doi.org/10.1006/jmbi.1997.1233.
[4] N.J. Burgoyne, R.M. Jackson, Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces, Bioinformatics. 22 (2006) 1335–1342, https://doi.org/10.1093/bioinformatics/btl079.
[5] P. Chakrabarti, J. Janin, Dissecting protein–protein recognition sites, Proteins. 47 (2002) 334–343, https://doi.org/10.1002/prot.10085.
[6] B. Vallone, A.E. Miele, P. Vecchini, E. Chiancone, M. Brunori, Free energy of burying hydrophobic residues in the interface between protein subunits, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 6103–6107, https://doi.org/10.1073/pnas.95.11.6103.
[7] A. David, M.J.E. Sternberg, The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease, J. Mol. Biol. 427 (2015) 2886–2898, https://doi.org/10.1016/j.jmb.2015.07.004.
[8] L.C. Xue, D. Dobbs, A.M.J.J. Bonvin, V. Honavar, Computational prediction of protein interfaces: a review of data driven methods, FEBS Lett. 589 (2015) 3516–3526, https://doi.org/10.1016/j.febslet.2015.10.003.
[9] R. Esmaielbeiki, K. Krawczyk, B. Knapp, J.-C. Nebel, C.M. Deane, Progress and challenges in predicting protein interfaces, Brief. Bioinform. 17 (2016) 1–15, https://doi.org/10.1093/bib/bbv027.
[10] C.-T. Chen, H.-P. Peng, J.-W. Jian, K.-C. Tsai, J.-Y. Chang, E.-W. Yang, J.-B. Chen, S.-Y. Ho, W.-L. Hsu, A.-S. Yang, Protein–protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces, PLoS One 7 (2012), e37706. https://doi.org/10.1371/journal.pone.0037706.
[11] C. Yan, X. Zou, Predicting peptide binding sites on protein surfaces by clustering chemical interactions, J. Comput. Chem. 36 (2015) 49–61, https://doi.org/10.1002/jcc.23771.
[12] H. Hwang, D. Petrey, B. Honig, A hybrid method for protein–protein interface prediction, Protein Sci. 25 (2016) 159–165, https://doi.org/10.1002/pro.2744.
[13] A. Lavi, C.H. Ngan, D. Movshovitz-Attias, T. Bohnuud, C. Yueh, D. Beglov, O. Schueler-Furman, D. Kozakov, Detection of peptide-binding sites on protein surfaces: the first step towards the modeling and targeting of peptide-mediated interactions, Proteins. 81 (2013) 2096–2105, https://doi.org/10.1002/prot.24422.
[14] J. Segura, P.F. Jones, N. Fernandez-Fuentes, A holistic in silico approach to predict functional sites in protein structures, Bioinformatics. 28 (2012) 1845–1850, https://doi.org/10.1093/bioinformatics/bts269.
[15] C. Savojardo, P. Fariselli, P.L. Martelli, R. Casadio, ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model, Bioinformatics. 33 (2017) 1656–1663, https://doi.org/10.1093/bioinformatics/btx044.
[16] H. Chen, H.-X. Zhou, Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data, Proteins. 61 (2005) 21–35, https://doi.org/10.1002/prot.20514.
[17] D. La, D. Kihara, A novel method for protein–protein interaction site prediction using phylogenetic substitution models, Proteins. 80 (2013) 126–141, https://doi.org/10.1002/prot.23169.A.
[18] S.J. de Vries, A.M.J.J. Bonvin, CPORT: a consensus Interface predictor and its performance in prediction-driven docking with HADDOCK, PLoS One 6 (2011), e17695. https://doi.org/10.1371/journal.pone.0017695.
[19] M.-H. Seo, P.M. Kim, The present and the future of motif-mediated protein–protein interactions, Curr. Opin. Struct. Biol. 50 (2018) 162–170, https://doi.org/10.1016/j.sbi.2018.04.005.
[20] P. Tompa, N.E. Davey, T.J. Gibson, M.M. Babu, A million peptide motifs for the molecular biologist, Mol. Cell 55 (2014) 161–169, https://doi.org/10.1016/j.molcel.2014.05.032.
[21] C. Mooney, G. Pollastri, D.C. Shields, N.J. Haslam, Prediction of short linear protein binding regions, J. Mol. Biol. 415 (2012) 193–204, https://doi.org/10.1016/j.jmb.2011.10.025.
[22] N. Kurochkina, U. Guha, SH3 domains: modules of protein–protein interactions, Biophys. Rev. 5 (2013) 29–39, https://doi.org/10.1007/s12551-012-0081-z.
[23] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm, J. Mol. Biol. 293 (1999) 321–331, https://doi.org/10.1006/jmbi.1999.3110.
[24] R. van der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631, https://doi.org/10.1021/cr400525m.
[25] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, V.N. Uversky, Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362 (2006) 1043–1059, https://doi.org/10.1016/j.jmb.2006.07.087.
[26] V. Vacic, C.J. Oldfield, A. Mohan, P. Radivojac, M.S. Cortese, V.N. Uversky, A.K. Dunker, Characterization of molecular recognition features, MoRFs, and their binding partners, J. Proteome Res. 6 (2007) 2351–2366, https://doi.org/10.1021/pr0701411.
[27] E.T.C. Wong, D. Na, J. Gsponer, On the importance of polar interactions for complexes containing intrinsically disordered proteins, PLoS Comput. Biol. 9 (2013), e1003192. https://doi.org/10.1371/journal.pcbi.1003192.
[28] B. Ma, M. Shatsky, H.J. Wolfson, R. Nussinov, Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations, Protein Sci. 11 (2002) 184–197, https://doi.org/10.1110/ps.21302.
[29] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, J. Mol. Biol. 338 (2004) 1015–1026, https://doi.org/10.1016/j.jmb.2004.03.017.
[30] R. Pancsa, M. Fuxreiter, Interactions via intrinsically disordered regions: what kind of motifs? IUBMB Life 64 (2012) 513–520, https://doi.org/10.1002/iub.1034.
[31] E.R. Lacy, Y. Wang, J. Post, A. Nourse, W. Webb, M. Mapelli, A. Musacchio, G. Siuzdak, R.W. Kriwacki, Molecular basis for the specificity of p27 toward cyclin-dependent kinases that regulate cell division, J. Mol. Biol. 349 (2005) 764–773, https://doi.org/10.1016/j.jmb.2005.04.019.

[32] A.A. Russo, P.D. Jeffrey, A.K. Patten, J. Massagué, N.P. Pavletich, J. Massague, N.P. Pavletich, Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A–Cdk2 complex, Nature. 382 (1996) 325–331, https://doi.org/10.1038/382325a0.
[33] M. Fuxreiter, Fuzziness in protein interactions—a historical perspective, J. Mol. Biol. 430 (2018) 2278–2287, https://doi.org/10.1016/j.jmb.2018.02.015.
[34] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions, Trends Biochem. Sci. 33 (2008) 2–8, https://doi.org/10.1016/j.tibs.2007.10.003.
[35] M. Miskei, C. Antal, M. Fuxreiter, FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies, Nucleic Acids Res. 45 (2017) D228–D235, https://doi.org/10.1093/nar/gkw1019.
[36] M. Fuxreiter, Fold or not to fold upon binding - does it really matter? Curr. Opin. Struct. Biol. 54 (2018) 19–25, https://doi.org/10.1016/j.sbi.2018.09.008.
[37] V.N. Uversky, Unusual biophysics of intrinsically disordered proteins, Biochim. Biophys. Acta - Rev. Cancer. 1834 (2013) 932–951, https://doi.org/10.1016/j.bbapap.2012.12.008.
[38] P. Jemth, X. Mu, Å. Engström, J. Dogan, A frustrated binding interface for intrinsically disordered proteins, J. Biol. Chem. 289 (2014) 5528–5533, https://doi.org/10.1074/jbc.M113.537068.
[39] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29, https://doi.org/10.1038/nrm3920.
[40] M.M. Babu, R.W. Kriwacki, R.V. Pappu, Versatility from protein disorder, Science. 337 (2012) 1460–1461, https://doi.org/10.1126/science.1228775.
[41] A. Borgia, M.B. Borgia, K. Bugge, V.M. Kissling, P.O. Heidarsson, C.B. Fernandes, A. Sottini, A. Soranno, K.J. Buholzer, D. Nettels, B.B. Kragelund, R.B. Best, B. Schuler, Extreme disorder in an ultrahigh-affinity protein complex, Nature. 555 (2018) 61–66, https://doi.org/10.1038/nature25762.
[42] D.A. Merle, A. Witternigg, C. Tam-Amersdorfer, C. Hartlmüller, E. Spreitzer, E. Schrank, S. Wagner-Lichtenegger, O. Werzer, K. Zangger, A.J. Kungl, T. Madl, N.H. Meyer, S.F. Falsone, Increased aggregation tendency of alpha-synuclein in a fully disordered protein complex, J. Mol. Biol. (2019)https://doi.org/10.1016/j.jmb.2019.04.031.
[43] N. London, D. Movshovitz-Attias, O. Schueler-Furman, The structural basis of peptide–protein binding strategies, Structure. 18 (2010) 188–199, https://doi.org/10.1016/j.str.2009.11.012.
[44] B. Mészáros, P. Tompa, I. Simon, Z. Dosztányi, Molecular principles of the interactions of disordered proteins, J. Mol. Biol. 372 (2007) 549–561, https://doi.org/10.1016/j.jmb.2007.07.004.
[45] S. Fukuchi, S. Sakamoto, Y. Nobe, S.D. Murakami, T. Amemiya, K. Hosoda, R. Koike, H. Hiroaki, M. Ota, IDEAL: intrinsically disordered proteins with extensive annotations and literature, Nucleic Acids Res. 40 (2012) D507–D511, https://doi.org/10.1093/nar/gkr884.
[46] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers. 22 (1983) 2577–2637, https://doi.org/10.1002/bip.360221211.
[47] M. Okuda, M. Kinoshita, E. Kakumu, K. Sugasawa, Y. Nishimura, Structural insight into the mechanism of TFIIH recognition by the acidic string of the nucleotide excision repair factor XPC, Structure. 23 (2015) 1827–1837, https://doi.org/10.1016/j.str.2015.07.009.
[48] H.R.A. Jonker, R.W. Wechselberger, R. Boelens, G.E. Folkers, R. Kaptein, Structural properties of the promiscuous VP16 activation domain, Biochemistry. 44 (2005) 827–839, https://doi.org/10.1021/bi0482912.
[49] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16, ACM Press, New York, New York, USA, 2016: pp. 785–794. doi:https://doi.org/10.1145/2939672.2939785.
[50] M.J. Meyer, J.F. Beltrán, S. Liang, R. Fragoza, A. Rumack, J. Liang, X. Wei, H. Yu, Interactome INSIDER: a structural interactome browser for genomic studies, Nat. Methods 15 (2018) 107–114, https://doi.org/10.1038/nmeth.4540.
[51] J. Martin, Benchmarking protein–protein interface predictions: why you should care about protein size, Proteins. 82 (2014) 1444–1452, https://doi.org/10.1002/prot.24512.
[52] Z. Dong, K. Wang, T.K.L. Dang, M. Gültas, M. Welter, T. Wierschin, M. Stanke, S. Waack, CRF-based models of protein surfaces improve protein–protein interaction site predictions, BMC Bioinformatics. 15 (2014) 277, https://doi.org/10.1186/1471-2105-15-277.
[53] T. Wierschin, K. Wang, M. Welter, S. Waack, M. Stanke, Combining features in a graphical model to predict protein binding sites, Proteins. 83 (2015) 844–852, https://doi.org/10.1002/prot.24775.
[54] A. McCallum, K. Schultz, S. Singh, FACTORIE: Probabilistic programming via imperatively defined factor graphs, Proc. 22nd Int. Conf. Neural Inf. Process. Syst. (2009) 1249–1257http://papers.nips.cc/paper/3654-factorie-probabilistic-programming-via-imperatively-defined-factor-graphs.pdf.
[55] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta - Protein Struct. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.
[56] R.A. Jordan, Y. EL-Manzalawy, D. Dobbs, V. Honavar, Predicting protein–protein interface residues using local surface structural similarity, BMC Bioinformatics. 13 (2012), 41. https://doi.org/10.1186/1471-2105-13-41.
[57] T. Kawabata, Detection of multiscale pockets on protein surfaces using mathematical morphology, Proteins. 78 (2010) 1195–1211, https://doi.org/10.1002/prot.22639.
[58] V. Le Guilloux, P. Schmidtke, P. Tuffery, Fpocket: an open source platform for ligand pocket detection, BMC Bioinformatics. 10 (2009) 168, https://doi.org/10.1186/1471-2105-10-168.
[59] T. Xie, R. Graveline, G.S. Kumar, Y. Zhang, A. Krishnan, G. David, I. Radhakrishnan, Structural basis for molecular interactions involving MRG domains: implications in chromatin biology, Structure. 20 (2012) 151–160, https://doi.org/10.1016/j.str.2011.10.019.
[60] P. Zhang, J. Zhao, B. Wang, J. Du, Y. Lu, J. Chen, J. Ding, The MRG domain of human MRG15 uses a shallow hydrophobic pocket to interact with the N-terminal region of PAM14, Protein Sci. 15 (2006) 2423–2434, https://doi.org/10.1110/ps.062397806.
[61] G. Cingolani, J. Bednenko, M.T. Gillespie, L. Gerace, Molecular basis for the recognition of a nonclassical nuclear localization signal by importin beta, Mol. Cell 10 (2002) 1345–1353, https://doi.org/10.1016/S1097-2765(02)00727-X.

[62] S.J. Lee, N. Imamoto, H. Sakai, A. Nakagawa, S. Kose, M. Koike, M. Yamamoto, T. Kumasaka, Y. Yoneda, T. Tsukihara, The adoption of a twisted structure of importin-β is essential for the protein–protein interaction required for nuclear transport, J. Mol. Biol. 302 (2000) 251–264, https://doi.org/10.1006/jmbi.2000.4055.

[63] R. Wintjens, J.M. Wieruszeski, H. Drobecq, P. Rousselot-Pailley, L. Buée, G. Lippens, I. Landrieu, 1H NMR study on the binding of Pin1 Trp–Trp domain with phosphothreonine peptides, J. Biol. Chem. 276 (2001) 25150–25156, https://doi.org/10.1074/jbc.M010327200.

[64] A.M. Petros, A. Medek, D.G. Nettesheim, D.H. Kim, H.S. Yoon, K. Swift, E.D. Matayoshi, T. Oltersdorf, S.W. Fesik, Solution structure of the antiapoptotic protein bcl-2, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 3012–3017, https://doi.org/10.1073/pnas.041619798.

[65] L.M. Tuttle, D. Pacheco, L. Warfield, J. Luo, J. Ranish, S. Hahn, R.E. Klevit, Gcn4-mediator specificity is mediated by a large and dynamic fuzzy protein–protein complex, Cell Rep. 22 (2018) 3251–3264, https://doi.org/10.1016/j.celrep.2018.02.097.

[66] K.L. Yap, T. Yuan, T.K. Mal, H.J. Vogel, M. Ikura, Structural basis for simultaneous binding of two carboxy-terminal peptides of plant glutamate decarboxylase to calmodulin, J. Mol. Biol. 328 (2003) 193–204 http://www.ncbi.nlm.nih.gov/pubmed/12684008 (accessed May 21, 2019).

[67] M. Zhang, T. Tanaka, M. Ikura, Calcium-induced conformational transition revealed by the solution structure of apo calmodulin, Nat. Struct. Biol. 2 (1995) 758–767 http://www.ncbi.nlm.nih.gov/pubmed/7552747 (accessed May 21, 2019).

[68] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, Z. Weng, Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2, J. Mol. Biol. 427 (2015) 3031–3041, https://doi.org/10.1016/j.jmb.2015.07.016.

[69] M.R. Arkin, M. Randal, W.L. DeLano, J. Hyde, T.N. Luong, J.D. Oslob, D.R. Raphael, L. Taylor, J. Wang, R.S. McDowell, J.A. Wells, A.C. Braisted, Binding of small molecules to an adaptive protein–protein interface, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 1603–1608, https://doi.org/10.1073/pnas.252756299.

[70] C.D. Thanos, M. Randal, J.A. Wells, Potent small-molecule binding to a dynamic hot spot on IL-2, J. Am. Chem. Soc. 125 (2003) 15280–15281, https://doi.org/10.1021/ja0382617.

[71] X. Wang, M. Rickert, K.C. Garcia, Structure of the quaternary complex of interleukin-2 with its alpha, beta, and gammac receptors, Science. 310 (2005) 1159–1163, https://doi.org/10.1126/science.1117893.

[72] S.J. de Vries, A.D.J. van Dijk, A.M.J.J. Bonvin, WHISCY: what information does surface conservation yield? Application to data-driven docking, Proteins. 63 (2006) 479–489, https://doi.org/10.1002/prot.20842.

[73] H. Zellner, M. Staudigel, T. Trenner, M. Bittkowski, V. Wolowski, C. Icking, R. Merkl, PresCont: predicting protein–protein interfaces utilizing four residue properties, Proteins. 80 (2012) 154–168, https://doi.org/10.1002/prot.23172.

[74] H. Neuvirth, R. Raz, G. Schreiber, ProMate: a structure based prediction program to identify the location of protein–protein binding sites, J. Mol. Biol. 338 (2004) 181–199, https://doi.org/10.1016/j.jmb.2004.02.040.

[75] J. Besag, Statistical analysis of dirty pictures, J. Appl. Stat. 20 (1993) 63–87, https://doi.org/10.1080/02664769300000059.

[76] S. Maheshwari, M. Brylinski, Predicting protein interface residues using easily accessible on-line resources, Brief. Bioinform. 16 (2015) 1025–1034, https://doi.org/10.1093/bib/bbv009.

[77] C. Nilofer, A. Sukhwal, A. Mohanapriya, P. Kangueane, Protein–protein interfaces are vdW dominant with selective H-bonds and (or) electrostatics towards broad functional specificity, Bioinformation. 13 (2017) 164–173, https://doi.org/10.6026/97320630013164.

[78] D. Ganguly, S. Otieno, B. Waddell, L. Iconaru, R.W. Kriwacki, J. Chen, Electrostatically accelerated coupled binding and folding of intrinsically disordered proteins, J. Mol. Biol. 422 (2012) 674–684, https://doi.org/10.1016/j.jmb.2012.06.019.

[79] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein–protein recognition sites, J. Mol. Biol. 285 (1999) 2177–2198, https://doi.org/10.1006/JMBI.1998.2439.

[80] J.A. Wells, C.L. McClendon, Reaching for high-hanging fruit in drug discovery at protein–protein interfaces, Nature. 450 (2007) 1001–1009, https://doi.org/10.1038/nature06526.

[81] A.J. Souers, J.D. Leverson, E.R. Boghaert, S.L. Ackler, N. D. Catron, J. Chen, B.D. Dayton, H. Ding, S.H. Enschede, W.J. Fairbrother, D.C.S. Huang, S.G. Hymowitz, S. Jin, S.L. Khaw, P.J. Kovar, L.T. Lam, J. Lee, H.L. Maecker, K.C. Marsh, K.D. Mason, M.J. Mitten, P.M. Nimmer, A. Oleksijew, C.H. Park, C.M. Park, D.C. Phillips, A.W. Roberts, D. Sampath, J.F. Seymour, M.L. Smith, G.M. Sullivan, S.K. Tahir, C. Tse, M.D. Wendt, Y. Xiao, J.C. Xue, H. Zhang, R.A. Humerickhouse, S.H. Rosenberg, S.W. Elmore, ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets, Nat. Med. 19 (2013) 202–208, https://doi.org/10.1038/nm.3048.

[82] B. Rost, Twilight zone of protein sequence alignments, Protein Eng. 12 (1999) 85–94http://www.ncbi.nlm.nih.gov/pubmed/10195279.

[83] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402, https://doi.org/10.1016/j.automatica.2016.09.031.

[84] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT suite: a web server for clustering and comparing biological sequences, Bioinformatics. 26 (2010) 680–682, https://doi.org/10.1093/bioinformatics/btq003.

[85] F.M. Richards, The interpretation of protein structures: total volume, group volume distributions and packing density, J. Mol. Biol. 82 (1974) 1–14, https://doi.org/10.1016/0022-2836(74)90570-1.

[86] E.D. Levy, A simple definition of structural regions in proteins and its use in analyzing Interface evolution, J. Mol. Biol. 403 (2010) 660–670, https://doi.org/10.1016/j.jmb.2010.09.028.

[87] E. Petsalaki, A. Stark, E. García-Urdiales, R.B. Russell, Accurate prediction of peptide binding sites on protein surfaces, PLoS Comput. Biol. 5 (2009), e1000335. https://doi.org/10.1371/journal.pcbi.1000335.

[88] J. a Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics. 23 (2007) 1875–1882, https://doi.org/10.1093/bioinformatics/btm270.

[89] J.F. Gibrat, T. Madej, S.H. Bryant, Surprising similarities in structure comparison, Curr. Opin. Struct. Biol. 6 (1996)

377–385 http://www.ncbi.nlm.nih.gov/pubmed/8804824 (accessed October 3, 2018).
[90] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780, https://doi.org/10.1093/molbev/mst010.
[91] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066, https://doi.org/10.1093/nar/gkf436.
[92] D. Xu, Y. Zhang, Generating triangulated macromolecular surfaces by euclidean distance transform, PLoS One 4 (2009), e8140. https://doi.org/10.1371/journal.pone.0008140.
[93] M. Hendlich, F. Rippmann, G. Barnickel, LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, J. Mol. Graph. Model. 15 (1997) 359–363, https://doi.org/10.1016/S1093-3263(98)00002-3.
[94] L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, E. Alexov, DelPhi: a comprehensive suite for DelPhi software and associated resources, BMC Biophys. 5 (2012) 9, https://doi.org/10.1186/2046-1682-5-9.
[95] D. Petrey, Z. Xiang, C.L. Tang, L. Xie, M. Gimpelev, T. Mitros, C.S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I.Y.Y.Y. Koh, E. Alexov, B. Honig, Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling, Proteins. 53 (Suppl. 6) (2003) 430–435, https://doi.org/10.1002/prot.10550.
[96] J.M. Word, S.C. Lovell, J.S. Richardson, D.C. Richardson, Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation, J. Mol. Biol. 285 (1999) 1735–1747, https://doi.org/10.1006/jmbi.1998.2401.
[97] S. Sarkar, S. Witham, J. Zhang, M. Zhenirovskyy, W. Rocchia, E. Alexov, DelPhi Web Server: a comprehensive online suite for electrostatic calculations of biological macromolecules and their complexes, Commun. Comput. Phys. 13 (2013) 269–284 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3966485/ (accessed October 4, 2018).
[98] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B 102 (1998) 3586–3616, https://doi.org/10.1021/jp973084f.
[99] M. Feig, J. Karanicolas, C.L. Brooks 3rd, MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology, J Mol Graph Model. 22 (2004) 377–395, https://doi.org/10.1016/j.jmgm.2003.12.005.
[100] F.B. Sheinerman, B. Honig, On the role of electrostatic interactions in the design of protein–protein interfaces, J. Mol. Biol. 318 (2002) 161–177, https://doi.org/10.1016/S0022-2836(02)00030-X.
[101] N. Xiao, D.-S. Cao, M.-F. Zhu, Q.-S. Xu, Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, Bioinformatics. 31 (2015) 1857–1859, https://doi.org/10.1093/bioinformatics/btv042.
[102] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 27 (1999) 368–369http://www.ncbi.nlm.nih.gov/pubmed/9847231.
[103] A. Bakan, L.M. Meireles, I. Bahar, ProDy: protein dynamics inferred from theory and experiments, Bioinformatics. 27 (2011) 1575–1577, https://doi.org/10.1093/bioinformatics/btr168.
[104] O.V. Tsodikov, M. Thomas Record, Y.V. Sergeev, Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature, J. Comput. Chem. 23 (2002) 600–609, https://doi.org/10.1002/jcc.10061.
[105] F.K. Pettit, E. Bare, A. Tsai, J.U. Bowie, HotPatch: a statistical approach to finding biologically relevant features on protein surfaces, J. Mol. Biol. 369 (2007) 863–879, https://doi.org/10.1016/j.jmb.2007.03.036.
[106] F.K. Pettit, J.U. Bowie, Protein surface roughness and small molecular binding sites, J. Mol. Biol. 285 (1999) 1377–1382, https://doi.org/10.1006/jmbi.1998.2411.
[107] W. Jakob, M. Tarini, D. Panozzo, O. Sorkine-Hornung, Instant field-aligned meshes, ACM Trans. Graph. 34 (2015) 1–15, https://doi.org/10.1145/2816795.2818078.
[108] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159, https://doi.org/10.1109/CDC.2012.6426698.
[109] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, CA, USA, 1988https://dl.acm.org/citation.cfm?id=52121, Accessed date: 16 February 2019.
[110] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38, https://doi.org/10.1016/0263-7855(96)00018-5.
[111] M.F. Sanner, A.J. Olson, J.C. Spehner, Reduced surface: an efficient way to compute molecular surfaces., Biopolymers. 38 (1996) 305–20. doi:10.1002/(SICI)1097-0282(199603)38:3≪305::AID-BIP4≫3.0.CO;2-Y.
[112] E. Roberts, J. Eargle, D. Wright, Z. Luthey-Schulten, MultiSeq: unifying sequence and structure data for evolutionary analysis, BMC Bioinformatics. 7 (2006) 382, https://doi.org/10.1186/1471-2105-7-382.
[113] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016http://ggplot2.org.
[114] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics. 21 (2005) 3940–3941, https://doi.org/10.1093/bioinformatics/bti623.
[115] A.D. McLachlan, Rapid comparison of protein structures, Acta Crystallogr. Sect. A. 38 (1982) 871–873, https://doi.org/10.1107/S0567739482001806.
[116] A. Kuzmanic, B. Zagrovic, Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors, Biophys. J. 98 (2010) 861–871, https://doi.org/10.1016/j.bpj.2009.11.011.