

# Interfaces Between Alpha-helical Integral Membrane Proteins: Characterization, Prediction, and Docking

Bian Li, Jeffrey Mendenhall, Jens Meiler\*

Department of Chemistry and Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA

## ARTICLE INFO

### Article history:

Received 14 February 2019  
Received in revised form 20 May 2019  
Accepted 21 May 2019  
Available online 25 May 2019

### Keywords:

Membrane protein interfaces  
Membrane protein docking  
Neural networks  
Weighted contact numbers

## ABSTRACT

Protein-protein interaction (PPI) is an essential mechanism by which proteins perform their biological functions. For globular proteins, the molecular characteristics of such interactions have been well analyzed, and many computational tools are available for predicting PPI sites and constructing structural models of the complex. In contrast, little is known about the molecular features of the interaction between integral membrane proteins (IMPs) and few methods exist for constructing structural models of their complexes. Here, we analyze the interfaces from a non-redundant set of complexes of  $\alpha$ -helical IMPs whose structures have been determined to a high resolution. We find that the interface is not significantly different from the rest of the surface in terms of average hydrophobicity. However, the interface is significantly better conserved and, on average, inter-subunit contacting residue pairs correlate more strongly than non-contacting pairs, especially in obligate complexes. We also develop a neural network-based method, with an area under the receiver operating characteristic curve of 0.75 and a Pearson correlation coefficient of 0.70, for predicting interface residues and their weighted contact numbers (WCNs). We further show that predicted interface residues and their WCNs can be used as restraints to reconstruct the structure  $\alpha$ -helical IMP dimers through docking for fourteen out of a benchmark set of sixteen complexes. The RMSD100 values of the best-docked ligand subunit to its native structure are  $<2.5$  Å for these fourteen cases. The structural analysis conducted in this work provides molecular details about the interface between  $\alpha$ -helical IMPs and the WCN restraints represent an efficient means to score  $\alpha$ -helical IMP docking candidates.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Alpha-helical integral membrane proteins (IMPs) play essential roles in biochemical and electrical signal transduction, molecular transport, and energy propagation among other critical cellular processes across the membrane. It was estimated that about one quarter of the human genome encodes  $\alpha$ -helical IMPs [1]. Frequently, these IMPs function as either homo-oligomers or hetero-oligomers with other IMPs [2–4].

Proteins can either form stable, obligate complexes via high-affinity protein-protein interactions (PPIs) or non-obligate, transient

complexes via low-affinity PPIs. Obligate PPIs are typically both structurally and functionally obligate: the participating IMPs are typically not found in isolation in vivo. One example is the KCNQ1 potassium channel, which forms homo-tetramers to create the channel pore in the center. It is believed to always bind one member of the KCNE protein family [5]. In contrast, IMPs that participate in transient PPIs can exist independently and complexes of this kind are usually involved in processes such as cellular signaling and receptor-ligand binding [6]. The characteristics of protein interfaces between globular proteins have been extensively studied in terms of size, amino acid composition, physicochemical texture, conservation, as well as coevolution of inter-subunit residue pairs. These properties usually differ for those PPIs that are transient versus those that are obligate [7]. In general, the interfaces in obligate complexes of soluble proteins are larger and more hydrophobic than those of non-obligate associations, for which, amino acid compositions are usually not drastically different from the rest of the protein surface [8–14]. It was also found that the interface is usually more conserved than the rest of the protein surface [15–17], although residues at the interfaces of obligate PPIs tend to be better conserved and exhibit much stronger coevolution with their interacting partners than those at the interfaces of transient PPIs [16].

*Abbreviations:* IMP, Integral membrane protein; WCN, Weighted contact number; PPI, Protein-protein interaction; RMSD, Root-mean-square distance; MSA, Multiple sequence alignment; OPM, Orientations of proteins in membranes; PDB, Protein data bank; PSSM, Position-specific scoring matrix; PCC, Pearson correlation coefficient; MAE, Mean absolute error; ROC, Receiver operating characteristic curve; AUC, Area under the ROC curve; TPR, True positive rate; TNR, True negative rate; PPV, Positive predictive value; PPM, Positioning of proteins in membrane.; RSA, Relative solvent accessibility.

\* Corresponding author at: 465 21st Ave South, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232-8725, USA.

E-mail address: [jens@meilerlab.org](mailto:jens@meilerlab.org) (J. Meiler).

<https://doi.org/10.1016/j.csbj.2019.05.005>

2001-0370/© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

These observations have been essential to our understanding of the biochemistry and biophysics of PPIs between globular, soluble proteins. In contrast, little is known about the characteristics of the interfaces between  $\alpha$ -helical IMPs. Here, we analyzed a non-redundant set of  $\alpha$ -helical IMP complexes whose structures have been experimentally determined to a high resolution to answer the following questions about IMP interfaces: 1) Is the IMP interface physico-chemically distinguishable from the rest of the protein surface? 2) Are residues in the interface more conserved than the rest of the protein surface? 3) Are there detectable coevolutionary signals across the interface? 4) How do the interfaces in obligate complexes differ from those in transient ones?

Our work consists of three parts that build on one another:

- 1) The analysis of IMP PPIs: We found that in  $\alpha$ -helical IMP complexes, while the aqueous region of the interface exhibits similar characteristics to interfaces between globular proteins, the interface in the membrane region is not significantly different from the rest of the surface in terms of average hydrophobicity. However, the interface is significantly better conserved than the rest of the surface both in the aqueous region and the membrane region, and residue pairs that are in physical contact at the interface of homo-oligomeric IMPs correlate more strongly than pairs not in contact. We restricted our correlation analysis on homo-oligomers because for hetero-oligomers, it requires construction of paired multiple sequence alignment (MSA) of interacting partners, and computationally, this problem becomes difficult by itself due to the existence of paralogs [18–20].
- 2) Sequence-based prediction of interface residues involved in IMP PPIs: Based on our findings, we also developed a neural network-based method that predicts weighted contact numbers (WCNs) of surface residues from evolutionary information. WCN is defined as the number of neighboring residues weighted by their proximity to the focal residue, it measures the local packing degree of residues within the protein tertiary structure [21,22]. We find that interface residues can be accurately identified based on the deviation of predicted WCN from the WCN computed from the protomer structure alone.
- 3) Structure-based docking of IMPs: Based on our previous findings in which we demonstrated that residue WCNs can be effectively used as restraints to improve *de novo* tertiary structure prediction for  $\alpha$ -helical IMPs [23], we implemented an algorithm which leverages the high discriminatory power of a WCN-based penalty score for accurate docking of  $\alpha$ -helical IMPs.

## 2. Methods

### 2.1. Data Set

A set of multi-pass  $\alpha$ -helical IMPs whose structures have been determined to a resolution of 2.5 Å or better and an R-free value of 0.3 or better were extracted from the Orientations of Proteins in Membranes (OPM) database [24] in March 2016. The data set was further refined by using the PISCES server [101] to reduce redundancy such that pairwise sequence identity between protein subunits is <25%. Proteins whose structures were not determined by X-ray crystallography or artificial chimeras were excluded from consideration. Classification of a complex as obligate or transient and assignment of biologically relevant oligomeric state (dimer, trimer, etc.) were carried out based on evidence found in the literature where the structure of the complex was reported. In summary, the data set consists of 36 obligate and nine transient complexes (Table 1). The bias toward more obligate complexes is not unexpected as the higher affinity and rigidity should facilitate crystallization and increase quality of the resulting structural model. The data set consists of 15 homodimers, twelve homotrimers, four homotetramers, two homopentamers, two homodecamers, one heterodimer, four heterotrimers, four heterotetramers, and one heteropentamer. It's worth mentioning that there is a spread in

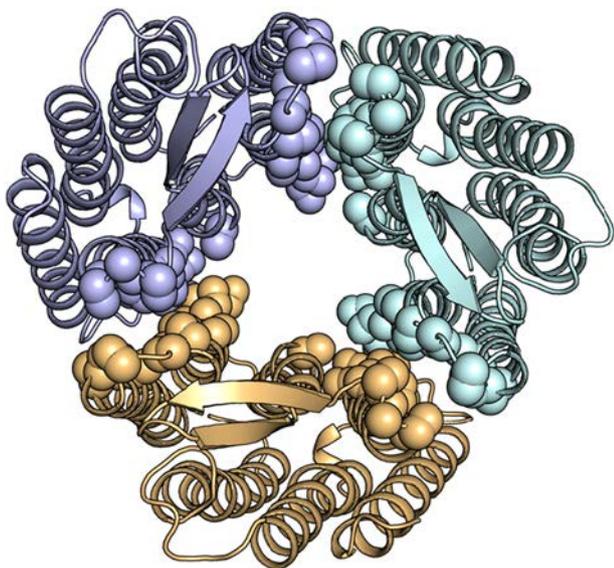
**Table 1**  
 $\alpha$ -helical IMP chains that form the oligomers in the data set.

Protein subunit	TMH count	Resolution (Å)	Oligomeric state	Obligate	Reference
1m0IA	7	1.5	Homotrimer	No	[25]
1m56A	12	2.3	Heterotetramer	Yes	[26]
1m56C	7	2.3	Heterotetramer	Yes	[26]
1q16C	5	1.9	Homodimer	No	[27]
1u7gA	11	1.4	Homotrimer	Yes	[28]
1yq3C	3	2.2	Heterotetramer	No	[29]
1yq3D	3	2.2	Heterotetramer	No	[29]
2a65A	12	1.7	Homodimer	Yes	[30]
2b12A	4	2.1	Homodecamer	Yes	[31]
2bs2C	5	1.8	Homodimer	No	[32]
2j8cM	5	1.9	Heterotrimer	Yes	[33]
2nq2A	10	2.4	Homodimer	Yes	[34]
2qtsA	2	1.9	Homotrimer	Yes	[35]
2uuuA	4	2.2	Homotrimer	Yes	[36]
2vpzC	8	2.4	Homodimer	No	[37]
2w2eA	6	1.2	Homotetramer	Yes	[38]
2wswA	12	2.3	Homotrimer	Yes	[39]
2yevC	7	2.4	Heterotrimer	Yes	[40]
2z73A	7	2.5	Homodimer	No	[41]
2zxeA	10	2.4	Heterotrimer	Yes	[42]
3b9yA	11	1.9	Homotrimer	Yes	[43]
3c02A	6	2.1	Homotetramer	Yes	[44]
3cx5C	8	1.9	Heteropentamer	Yes	[45]
3k3fA	10	2.3	Homotrimer	Yes	[46]
3klyA	6	2.1	Homopentamer	Yes	[47]
3m73A	10	1.2	Homotrimer	Yes	[48]
3oduA	7	2.5	Homodimer	No	[49,50]
3oufA	2	1.6	Homotetramer	Yes	[49]
3puwF	8	2.3	Heterodimer	Yes	[51]
3s8gA	13	1.8	Heterotrimer	Yes	[52]
3spcA	2	2.5	Homotetramer	Yes	[53]
3tijA	8	2.4	Homotrimer	Yes	[54]
4a01A	16	2.4	Homodimer	Yes	[55]
4bpmA	4	2.1	Homotrimer	Yes	[56]
4d2eA	3	2.3	Homotrimer	Yes	[57]
4dx5A	12	1.9	Homotrimer	Yes	[58]
4f4sA	2	1.9	Homodecamer	Yes	[59]
4jkvA	7	2.5	Homodimer	No	[60]
4mrsA	6	2.4	Homodimer	Yes	[61]
4o6mA	6	1.9	Homodimer	Yes	[62]
4o6yA	6	1.7	Homodimer	Yes	[63]
4qndA	3	1.7	Homodimer	Yes	[64]
4rngC	3	2.4	Homodimer	Yes	[65]
4u9nA	5	2.2	Homodimer	Yes	[66]
4wd8A	4	2.3	Homopentamer	Yes	[67]

transmembrane helix counts and each category of helix count is well represented except that no subunit with nine, fourteen, or fifteen helices satisfied our aforementioned data set curation criteria (Supporting Information Fig. S1).

### 2.2. Defining Interface Residues

Residue categorization was based on residues' weighted contact numbers (WCN). The WCN of a residue is defined as the number of its neighboring residues weighted by spatial proximity. The WCN of each residue in the data set was computed according to a procedure previously described [68]. A residue was then categorized as surface residue if its WCN < 8.38 and as core residue otherwise. This cutoff was determined based on a linear fit of WCN to relative solvent accessibility (RSA) ( $WCN = -9.78 \times RSA + 8.48$ ) and residues with an RSA > 0.01 are considered solvent exposed as in a previous study [16]. Interface residues were defined as those surface residues whose WCN increases by at least 1.0 upon the formation of a complex (see Fig. 1 for an example). Although some previous work has classified whether a residue is on the surface or in the core according to its solvent accessibility [16,69–72], we classified residues based on their WCN for two practical reasons. On the one hand, computing residue WCN is computationally less



**Fig. 1.** Defining interface residues according to residue weighted contact number (WCN). A surface residue is defined as an interface residue if its WCN increases by at least 1.0 ( $\Delta WCN \geq 1.0$ ) upon the formation of a protein complex. This figure shows interface residues, represented as spheres, identified according to the above criterion for the trimeric bacteriorhodopsin (PDB ID: 1m0l).

demanding than computing residue solvent accessibility. This makes it feasible to compute residue WCNs on the fly and use them directly for scoring in protein structure prediction and protein-protein docking predictions where an enormous amount of conformational space must be sampled [23]. On the other hand, we had previously developed a machine-learning framework that allowed accurate prediction of residue WCNs for  $\alpha$ -helical IMPs [68].

### 2.3. Site-specific Rate of Evolution

Site-specific rate of evolution was estimated using the Rate4Site program [73]. The input multiple sequence alignment (MSA) of homologs to each monomer was obtained by running HHblits on the Uniprot20 sequence database [74] with minimum coverage of query sequence (sequence of the monomer) being set to 75%, maximum sequence identity to query sequence being set to 90%, maximum pairwise sequence identity being set to 90%, and e-value cutoff for inclusion in result alignment being set to  $10^{-5}$ .

### 2.4. Mutual Information

We model a sequence site as a discrete random variable  $X$ , which takes on one of an alphabet of 20 possible letters  $A_X = (A, C, D, \dots, W, Y)$  with probabilities  $(p_A, p_C, p_D, \dots, p_W, p_Y)$ , with  $P(X = x) = p_x, p_x \geq 0$  and  $\sum_{x \in A_X} p_x = 1$ . The alphabet  $A_X$  contains one letter for each amino acid. Alignment gap was not considered because it introduces spuriously high conservation for alignment columns containing a high percentage of gaps.  $p_x$  is estimated by the relative frequency  $f_x$  of amino acid residue  $x$  at the column of an MSA:

$$f_x = \frac{1}{20} \lambda + M \left[ \lambda + \sum_{i=1}^M \delta(x, X_i) \right]$$

where  $M$  indicates alignment depth (the number of sequences aligned at position  $i$ ),  $\delta$  is the Kronecker delta function such that it evaluates to 1 if  $x = X_i$  and 0 otherwise.  $f_x$  is adjusted by a pseudocount parameter  $\lambda = 1$ .

Given two MSA columns  $X$  and  $Y$ , the mutual information  $I(X, Y)$  between them is defined as

$$I(X, Y) = \sum_{i=1}^{20} \sum_{j=1}^{20} p_{ij}(x_i, y_j) \log \frac{p_{ij}(x_i, y_j)}{p_i(x_i)p_j(y_j)}$$

$I(X, Y)$  is a general measure of association between two random variables  $X$  and  $Y$  (in this case, two alignment columns in an MSA), it equals zero if and only if  $X$  and  $Y$  are independent, and it is positive otherwise. Intuitively,  $I(X, Y)$  can be thought of as the average reduction in uncertainty about  $X$  that results from knowing the value of  $Y$ , and vice versa.

To correct for bias in  $I(X, Y)$  that may arise from phylogeny [75], entropy of the MSA columns [76], or background noise [77], and test for significance, we implemented a permutation test in which, for each pair of columns, one column was selected and permuted 200 times. This procedure was described in [78] in detail. Briefly, a mutual information was calculated for the pair after each permutation. If the number of permutations for which the mutual information is greater than that of the original column pair is  $\geq 2$  (1% of the total number of permutations), we reject the hypothesis that the column pair is correlated, and its mutual information is set to 0. Otherwise, the bias-corrected mutual information for the pair was computed by subtracting the average mutual information of the 200 permutations from that of the original column pair.

### 2.5. Training a Neural Network for Predicting WCN

We trained a neural network for predicting residue WCN from amino acid sequence. The target WCNs used in the training were computed from the structure of oligomers in the data set, using the procedure previously described [68]. Each residue was numerically described by rate of evolution of the sequence site, entries in a window of size 15 from the position-specific scoring matrix (PSSM) computed from an MSA of homologs of the protein family as previously described [79]. The output layer of the neural network consists of a single node for the residue-specific WCN. The hidden layer consists of 64 neurons. The neural network was regularized using the “dropout” technique [80] where 5% of units in the input layer and 50% of neurons in the hidden layer were randomly silenced during the presentation of each training case. Connection weights were iteratively tuned with back-propagation of errors [81] with the learning rate  $\eta$  being set to 0.1 and momentum factor  $\alpha$  being set to 0.1. The accuracy of WCN prediction was assessed by the Pearson correlation coefficient (PCC) and the mean absolute error (MAE) of predicted WCNs from target WCNs, computed as follows

$$PCC = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

where  $x_i$  and  $y_i$  denote the target and predicted WCNs of residue  $i$ , respectively, and  $n$  denotes the number of residues in the data set.

### 2.6. Predicting Interface Residues

Note that the WCN of a residue may be different depending on whether it is computed from the structure of the individual protomers or the complex. To make the distinction straightforward, we refer to WCNs computed from individual subunits as protomeric WCNs and those computed from complexes as oligomeric WCNs. For predicting interface residues, it is reasonable to assume that an experimental structure of each of the individual subunits is available, and as such, true protomeric WCNs can be computed from the structures of individual subunits. A surface residue is then predicted to be an interface residue

if its neural network-predicted oligomeric WCN is higher than its true protomeric WCN by at least 1.0.

The performance of the neural network on predicting interface residues was assessed by the area under the receiver operating characteristic curve (ROC) [82], or AUC. The AUC was estimated through cross-validation where the data set was partitioned into subsets such that proteins from the same SCOP superfamily [83] were placed in the same subset. Each subset was used exactly once as the test set for evaluating the performance of the neural network trained on the remaining of the data set. Effectively, a value of AUC was computed from each test subset. The final estimate of the AUC was computed as the mean of all the AUCs. The true positive rate (TPR), true negative rate (TNR), and positive predictive value (PPV) [84] were calculated using 1.0 as the threshold, where a true positive is defined as a correctly predicted interface residue and a true negative is a correctly predicted non-interface residue.

### 2.7. Integral Membrane Protein Docking

One of the advantages of modeling IMP complexes is that the membrane reduces the search space by imposing a restraint on the position of the subunits in the membrane. We designed a docking-based algorithm called BCL::MP-Dock for predicting structure of complexes where both binding partners are  $\alpha$ -helical IMPs. The algorithm assumes that both docking partners are IMPs and requires that their tertiary structures be available. The input structures to BCL::MP-Dock are a structure of the “receptor” and a structure of the “ligand”, both oriented in the membrane using the Positioning of Proteins in Membrane (PPM) server [24]. Generation of docking candidates begins with a random rotation of the ligand about the z-axis (e.g. the membrane normal) and translation of the ligand in the membrane to create a glancing contact with the receptor. The ligand is then randomly moved with respect to the receptor using a Monte Carlo search [85]. Translation along the z-axis is limited to  $<5.4 \text{ \AA}$  (one winding of an alpha-helix) and the step size of the tilt angle from the z-axis is limited to 0.05 rad (or  $\sim 2.8$  degree). The baseline scoring function of BCL::MP-Dock consists of a clashing term that represents van der Waals repulsion, a residue-pair contact potential term for interface interaction, and a radius of gyration term that favors compact packing between the two docking partners, which were described in detail previously [86].

While the analysis of the characteristics of interface residues compared to the rest of surface residues was conducted on a diverse set of non-redundant IMP oligomers, the BCL::MP-Dock algorithm is designed to work primarily with dimers. In the current study, we tested the idea of using predicted interface residues and their predicted WCNs as restraints for scoring docking solutions on a set of 16  $\alpha$ -helical IMP dimers

(Table 2). This benchmark set consists of all  $\alpha$ -helical IMP dimers satisfying our data set curation criteria (see Methods). The constituting subunits have as few as three and as many as sixteen transmembrane helices, and most of the subunits have moderate sizes with five, six, or seven transmembrane helices.

### 2.8. Computation of Enrichment

The enrichment was used to measure how capable a scoring function is to select the most accurate docking solutions from a pool of solutions. To calculate enrichment, a given set  $S$  of docking solutions are sorted by their RMSD100 values. The top 10% of the solutions with the lowest RMSD100 [102] values are put into the set  $T$  (true) and the rest of the solutions are put into the set  $F$  (false). The solutions in  $S$  are then sorted by their evaluated score. The top 10% of solutions with the lowest score are put into the set  $P$  (positive) and the rest are put into the set  $N$  (negative). The intersection of sets  $T$  and  $P$  are solutions that are correctly identified by the scoring function and referred to as  $TP$  (true positives). The intersection of sets  $F$  and  $P$  are solutions that are incorrectly identified by the scoring function and are referred to as  $FP$  (false positives). The enrichment value is then computed using the following formula:

$$\text{Enrichment} = \frac{TP}{TP + FP} / \frac{P}{P + N}$$

Intuitively,  $\frac{P}{P+N}$  represents that probability of obtaining a native-like model when choosing a model from  $S$  at random, whereas  $\frac{TP}{TP+FP}$  represents the probability of obtaining a native-like model when choosing from a set of models below an energy cutoff. By our experimental design,  $\frac{P}{P+N}$  has a constant value of 0.1, so the enrichment values cannot exceed 10.

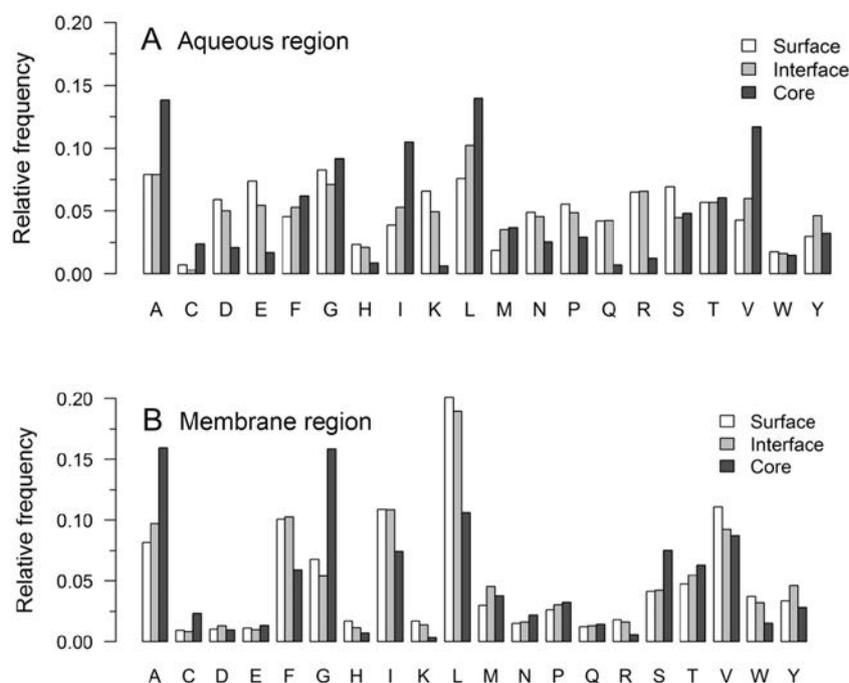
## 3. Results

Each protomer is divided into three disjoint regions: protein core, interface, and non-interface surface with no overlapping residues; and each region is further divided into two parts: the aqueous region and the membrane region based on the atomic coordinates and the calculated hydrophobic thickness of the membrane in which each complex resides, both were obtained from the OPM database [24]. The hydrophobic thicknesses provided by OPM were calculated by the PPM method, whose accuracy was thoroughly tested for a large set of IMPs whose orientations with respect to the lipid bilayer or membrane binding affinities have been experimentally studied [87].

**Table 2**  
Summary of the benchmark set of IMP complexes for testing the docking algorithm.

Protein ID	Resolution (Å)	Oligomeric state	Obligate	Protein name	TMH count in subunit
1q16_CF	1.9	Homodimer	No	Nitrate reductase A	5
2a65_AB	1.7	Homodimer	Yes	LEUTAA	12
2bs2_CF	1.8	Homodimer	No	Quinol fumarate reductase	5
2nq2_AB	2.4	Homodimer	Yes	Putative metal-chelate type ABC transporter	10
2vpz_CG	2.4	Homodimer	No	Polysulfide reductase from <i>Thermus thermophilus</i>	8
2z73_AB	2.5	Homodimer	No	Squid rhodopsin	7
3odu_AB	2.5	Homodimer	No	CXCR4 chemokine receptor	7
3puw_FG	2.3	Heterodimer	Yes	MBP-Maltose transporter transmembrane subunits	8/6
4a01_AB	2.4	Homodimer	Yes	H1-Translocating Pyrophosphatase	16
4jkv_AB	2.5	Homodimer	No	Human smoothed receptor	7
4mrs_AB	2.4	Homodimer	Yes	Bacterial Atm1-family ABC transporter	6
4o6m_AB	1.9	Homodimer	Yes	CDP-alcohol phosphotransferase	6
4o6y_AB	1.7	Homodimer	Yes	Cytochrome b561	6
4qnd_AB	1.7	Homodimer	Yes	Bacterial homologue of SWEET transporters	3
4rng_AC	2.4	Homodimer	Yes	Bacterial homologue of SWEET transporters (sequence identity with 4qnd_AB $<25\%$ )	3
4u9n_AB	2.2	Homodimer	Yes	Mg(2+) channel MgtE	5

Protein IDs are denoted as four-letter PDB ID followed by the chain IDs of the two subunits comprising the complex. Note that this benchmark set consists of all the dimers (fifteen homodimers + one heterodimer) from the data set used in our analysis of membrane protein interfaces.



**Fig. 2.** Amino acid composition of the core, interface, and the rest of the surface. (A) Amino acid composition in the aqueous region; just as amino acid composition in globular proteins and protein complexes, the relative frequencies of residues of hydrophobic nature increases in going from surface, to interface, and then to protein core, whereas those of residues of hydrophilic nature follows the opposite trend. (B) amino acid composition in the membrane region; due to its hydrophobic nature, the residue relative frequencies follow trends opposite to those in aqueous portion.

### 3.1. Amino Acid Composition and Interface Propensities

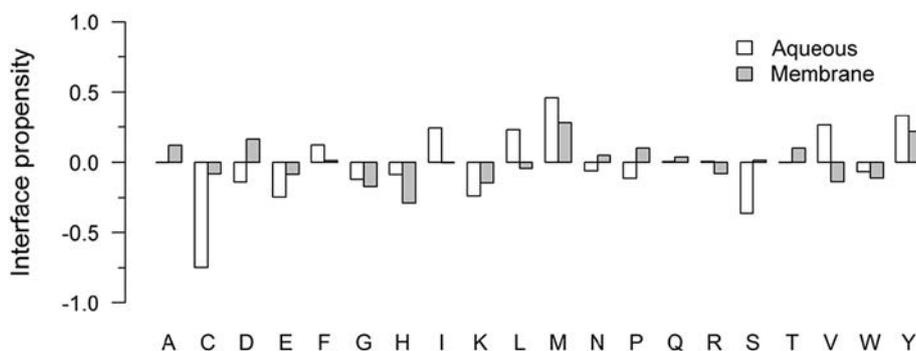
Fig. 2 compares the amino acid compositions of protein core, interface, and non-interface surface in the aqueous region and in the membrane region. Fig. 2A shows that in the aqueous region, protein cores have the highest frequencies of hydrophobic residues (e.g. Cys, Phe, Ile, Leu, Met, and Val), whereas surfaces have the highest frequencies of hydrophilic residues (e.g. Asp, Glu, His, Lys, Asn, Pro, Ser, and Trp), similar to the amino acid compositions of soluble protein complexes [17]. In the membrane, it is expected that hydrophobic residues are well represented on non-interface surfaces due to the hydrophobic nature of lipid tails. Interfaces and the IMP core should still be dominated by hydrophobic amino acids. However, some fraction of hydrophilic residues is expected for function, i.e. to build the polar pore in a channel or to lend selectivity to a PPI. Fig. 2B shows that non-interface surfaces in the membrane region have the highest frequencies of some hydrophobic residues (e.g. Ile, Leu, Val) as well as some hydrophilic residues (e.g. His, Lys, and Arg), whereas some hydrophilic residues have the highest frequencies in protein cores (e.g. Glu, Asn, Pro, Gln, Ser, and Thr). It is

also interesting that Gly and Ala, which have the smallest side-chain size, have the highest frequencies in protein cores both in the aqueous part and in the membrane. This is likely because Gly and Ala, which usually occur in GxxxG and AxxxA types of motifs, may facilitate more efficient helix packing [88,89].

The amino acid type distribution at the interface is significantly different from that on the rest of the surface for both the aqueous region ( $p = 1.7 \times 10^{-11}$ ,  $\chi^2$  test) and the membrane region ( $p = 0.011$ ,  $\chi^2$  test), albeit the difference is more pronounced in the aqueous region. Fig. 3 shows the interface propensity of each residue type. In the aqueous region, residue types that prefer interface over the rest of the surface are Phe, Ile, Leu, Met, Val, and Tyr. In the membrane, residue types that prefer interface over the rest of the surface are Ala, Asp, Met, Asn, Pro, Gln, Thr, and Tyr.

### 3.2. Hydrophobicity

Since IMPs experience both lipid bilayer and aqueous environments at the same time, we analyzed the hydrophobicity of the interface, non-



**Fig. 3.** Interface versus surface propensity in the aqueous region and the membrane region for each amino acid type. The interface propensity of an amino acid type A,  $IP_A$ , was calculated according to  $IP_A = -\log^{f_{A,i}}/f_{A,s}$ , where  $f_{A,i}$  is its fraction at the interface and  $f_{A,s}$  is its fraction on the surface as a whole. A negative propensity value indicates that the amino acid is preferentially located at the interface.

interface surface region, and the protein core separately for these two environments. In the aqueous region, the interface is significantly more hydrophobic than non-interface surface region ( $p = 0.0020$ , paired *t*-test), and the protein core is significantly more hydrophobic than the interface ( $p = 1.2 \times 10^{-5}$ , paired *t*-test) (Fig. 4A), consistent with the observation made on soluble proteins [70]. In the membrane region, the protein core is significantly less hydrophobic than the noninterface surface region ( $p = 9.7 \times 10^{-6}$ , paired *t*-test) (Fig. 4B), due to the nonpolar nature of lipid tails. The protein core is also less hydrophobic than the interface ( $p = 0.00077$ , paired *t*-test). No significant difference in average hydrophobicity is found between the interface and the rest of the surface ( $p = 0.054$ , paired *t*-test). This may be partly because protomers of transient membrane protein complexes, when they are not part of a complex, need their interface hydrophobic enough to make favorable contact with lipid tails. In fact, in obligate complexes the interface is marginally less hydrophobic than the rest of the surface ( $p = 0.014$ , paired *t*-test) (Fig. 4C).

### 3.3. The Interface is Better Conserved than the Rest of the Surface in Obligate Oligomers

The selective pressure, and likewise conservation, is expected to be higher at the interface than the rest of the surface because molecular recognition is often achieved by making specific contacts at the interface. To test this hypothesis, we computed the average rate of evolution of the interface and that of the rest of the surface for each complex in the data set. A lower average rate of evolution signifies a stronger conservation.

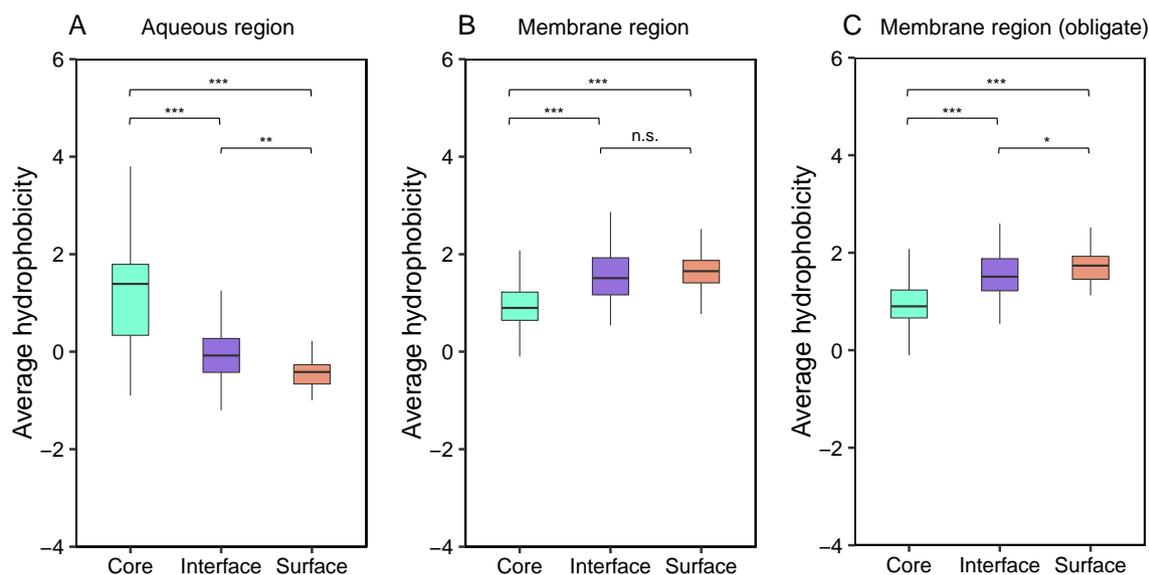
While the average rate of evolution of the interface is significantly lower than that of the rest of the surface in both the aqueous region ( $p = 0.0043$ , paired *t*-test) (Fig. 5A) and the membrane region ( $p = 0.00088$ , paired *t*-test) (Fig. 5D), we found that such difference stems from the selection pressure on obligate complexes. In obligate complexes, the interface is significantly better conserved than the rest of the surface in both the aqueous region ( $p = 0.00015$ , paired *t*-test) (Fig. 5B) and the membrane region ( $p = 3.3 \times 10^{-5}$ , paired *t*-test) (Fig. 5E). In contrast, in transient complexes, the interface is not more

conserved than the surface in either region ( $p = 0.87$  and  $0.81$ , respectively, paired *t*-test) (Fig. 5C and F). This confirms that, in general, the selection pressure on obligate complexes is stronger than on transient complexes [6,16].

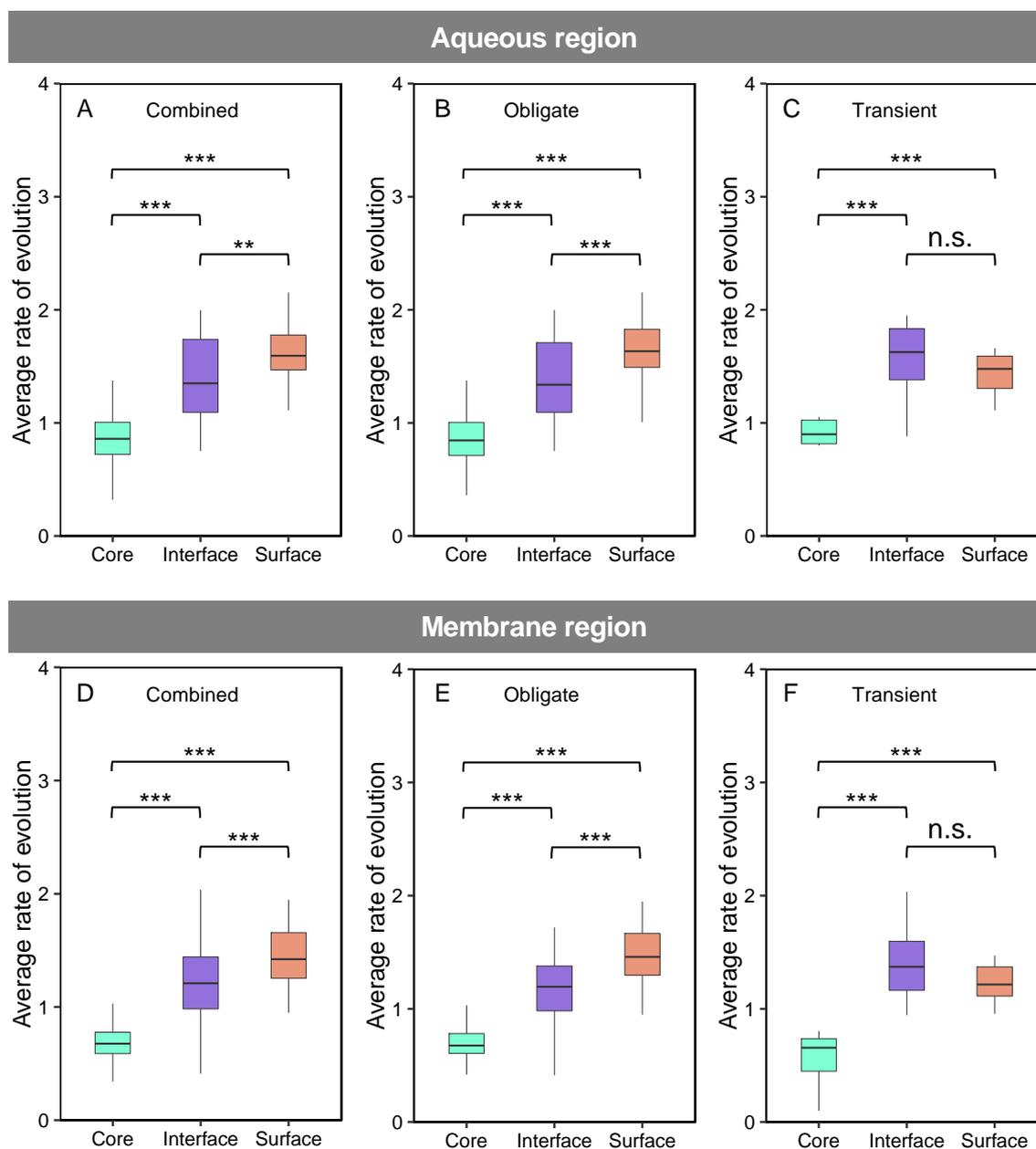
We also found that while the average rate of evolution of interface residues in the membrane region is significantly lower than that of interface residues in the aqueous region ( $p = 0.00016$ , paired *t*-test), it is worth noting that for many interfaces (12 out of 45 cases in our data set), the average rate of evolution is lower in the aqueous region than that in the hydrophobic region. This suggests that, for some membrane proteins, the aqueous region may be under stronger selection pressure than the membrane region for protein-protein interactions.

### 3.4. Contacting Interface Residue Pairs Show Stronger Correlation than Non-Contacting Pairs

Previously, studies have shown that residue pairs in inter-domain interfaces in globular proteins tend to be correlated [91] and that inter-subunit contacting residues in single-pass homo-oligomers co-evolve strongly [92]. However, it remains unclear whether these observations hold true for multi-pass membrane protein complexes. We reasoned that inter-subunit contacting residue pairs in multi-pass membrane protein complexes are also correlated more strongly than non-contacting pairs. To test this hypothesis, we analyzed the strength of correlation, quantified by mutual information, of contacting and non-contacting residue pairs for all homo-oligomers in our data set. A residue *i* in one subunit forming the interface is considered in physical contact with another residue *j* in the other subunit if the separation between any heavy atom of *i* and *j* is  $\leq 4.0$  Å. Fig. 6 compares the distribution of mutual information between pairs of residues in contact with that of the mutual information between pairs of residues not in contact. We observe a shift toward higher mutual information ( $p = 2.0 \times 10^{-10}$ , Mann-Whitney *U* test) for residue pairs that are in contact. This shift suggests that on average, at the interface of multi-pass membrane protein complexes, residue pairs that are in physical contact tend to correlate more strongly than those that are not in contact.



**Fig. 4.** Distribution of average hydrophobicity of core, interface, and noninterface surface of  $\alpha$ -helical IMPs. The average hydrophobicities were computed according to the Kyte & Doolittle hydrophobicity scale [90]. In this scale, arginine is assigned a hydrophobicity of  $-4.50$  and is the least hydrophobic; isoleucine is assigned  $4.50$  and is the most hydrophobic. (A) Distribution of average hydrophobicity in the aqueous region; the protein core is significantly more hydrophobic than the interface, which is in turn, significantly more hydrophobic than the surface. (B) Distribution of average hydrophobicity in the membrane region; the protein core is significantly less hydrophobic than the noninterface surface region. The protein core is also less hydrophobic than the interface, however, no significant difference in average hydrophobicity is found between the interface and the rest of the surface. (C) Distribution of average hydrophobicity for obligate oligomers; the interface is less hydrophobic than the rest of the surface, although it is only marginally significant. \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ , n.s.: not significant.

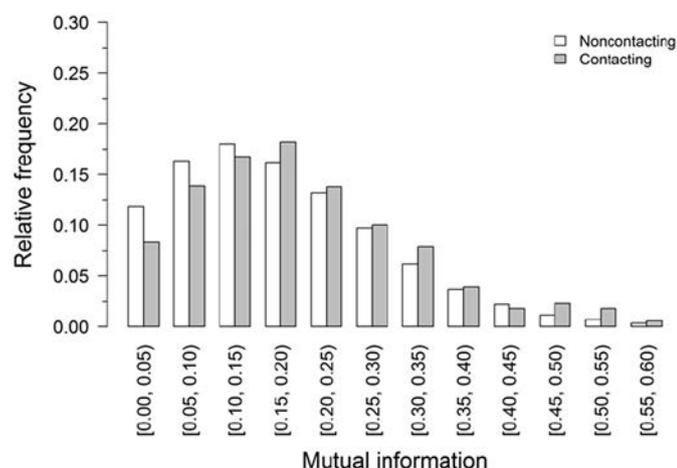


**Fig. 5.** Distribution of average rate of evolution of core, interface, and noninterface surface of  $\alpha$ -helical IMPs. Distribution of average rate of evolution for the aqueous region taking (A) obligate and transient oligomers together, (B) obligate oligomers only, (C) transient oligomers only; and for the membrane region taking (D) obligate and transient oligomers together, (E) obligate oligomers only, (F) transient oligomers only. \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ , n.s.: not significant.

A recent study suggests that there is a subtle trade-off between evolutionary and energetic constraints in protein-protein interactions [93]. The authors showed that for some classes of protein complexes, evolutionary constraints play the key role in defining the interaction surface, whereas in others, energetic constraints emerge as more important. While their analysis was conducted exclusively on soluble protein complexes, their conclusion might as well apply to membrane protein oligomers. In fact, from the 32 homo-oligomers for which we were able to compute inter-subunit residue-pair mutual information, a randomly selected contacting residue pair will have a higher mutual information than that of a randomly selected non-contacting residue pair in twenty cases (using Mann-Whitney  $U$  test and a cutoff  $p$ -value of 0.05, Supporting Information Table S1). Whereas in the remaining twelve homo-oligomers, other factors such as energetic constraint may play a more important role than evolutionary constraint.

### 3.5. Predicting Interface Residues in the Membrane

We first evaluate the accuracy of the trained neural networks in predicting WCNs. As shown in the scatter plot in Fig. 7, the Pearson correlation coefficient between true and predicted WCNs is 0.70 and the mean absolute error of predicted WCNs from true WCNs is 1.58. We then evaluate the performance of the trained neural networks in distinguishing residues at the interface from those on the rest of the surface given experimental structures for the protomers. A score for a residue to be at the interface is computed by subtracting its protomeric WCN from its predicted WCN, this score is termed  $\Delta_{\text{WCN}}$ . Surface residues whose  $\Delta_{\text{WCN}} \geq 1.0$ , a threshold chosen for consistency with the threshold used in the definition of interface residues, are predicted to be interface residues, resulting in a TPR of 0.88, a TNR of 0.54, and a PPV of 0.39.



**Fig. 6.** Comparison on the distribution of the mutual information between pairs of residues in contact with that of the mutual information between pairs of residues not in contact. There is a shift of the distribution of mutual information between contacting residues toward higher values.

**Fig. 7** shows the ROC curve plotted using cross-validated predictions from neural networks trained to distinguish interface residues from residues on the rest of the surface. While the AUC of our neural network is 0.75, comparable to a previous random forest-based protein-protein binding site predictor for membrane proteins [69], we'd like to highlight that our work addressed several limitations of the method introduced in [69] (Supplementary discussion).

### 3.6. Docking Membrane Proteins Using Predicted WCNs as Restraints

The BCL::MP-Dock algorithm was evaluated on a set consisting of 1 heterodimer and 15 homodimer structures. The results for global searches from fully randomized starting positions and orientations are detailed in Table 3. Of the 15 homodimer cases, BCL::MP-Dock is able to reconstruct the structure of the complex for 14 cases where the best RMSD100 of the docked ligand subunit to its native structure is  $<2.5$  Å (2.5 Å is the threshold of resolution used in creating the data set). The two cases where the best RMSD100 of the docked ligand is  $>2.5$  Å are the bacterial Atm1-family ABC transporter (PDB ID: 4mrs\_AB) and the heterodimeric transmembrane subunits of the

maltose ABC transporter (PDB ID: 3puw\_FG). To confirm the effectiveness of the sampling scheme of BCL::MP-Dock, we also computed the means of RMSD100 of the docked ligand of the top 1% models ranked by RMSD100. Again, except for 3puw\_FG and 4mrs\_AB, the mean RMSD100 is either less than or only slightly  $>3.0$  Å. 3puw\_FG is a heterodimer where helices from the two subunits are domain-swapped and 4mrs\_AB is a homodimer with bulky cytosolic domains. Both pose a significant challenge to the sampling scheme of BCL::MP-Dock.

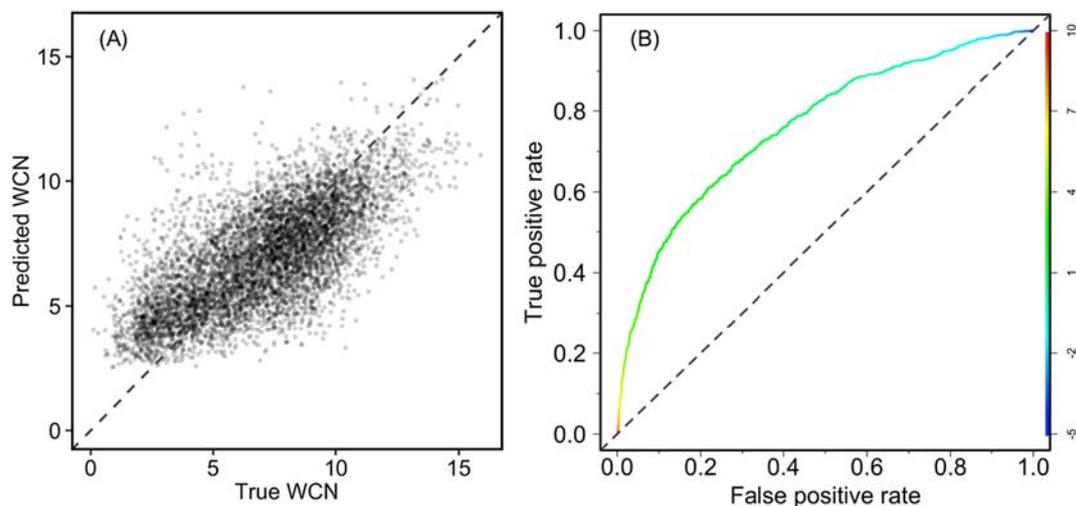
To assess the baseline scoring function, we computed the means of RMSD100 of the docked ligand of the 1% models ranked by score v.s. ranked by RMSD100. Compared to the means of RMSD100 of the top 1% models ranked by RMSD100, the baseline scoring function is remotely effective in only three cases (2a65\_AB, 2vpz\_CG, and 4a01\_AB). While the enrichment of the baseline scoring function is  $>1.0$  for eleven cases, it is  $>2.0$  for only five cases (Table 3). These results indicate that the baseline scoring function has difficulty in identifying docked ligands that are top-ranked by RMSD100 in most cases.

Previously we demonstrated that residue WCNs [23] and other experimental or simulated restraints [94] can be effectively incorporated into scoring functions to improve de novo tertiary structure prediction for  $\alpha$ -helical membrane proteins. Likewise, we hypothesized that a scoring term that assigns a penalty score to docked models according to the magnitude of the deviation the WCNs of predicted interface residues from their predicted WCNs will also be highly effective:

$$\text{Penalty score} = \sqrt{\frac{1}{n} \sum_{i \in \text{predicted interface residues}}^n (\text{WCN}_i - \text{WCN}_i^p)^2}$$

where  $\text{WCN}_i$  is the WCN of interface residue  $i$  computed based on the docked model,  $\text{WCN}_i^p$  is the WCN of interface residue  $i$  predicted by the neural network. The predicted WCNs of the resulting interface residues of each docking partners were used as restraints. Note that, to avoid potential overestimation of the effectiveness of this approach, the WCNs used as restraints in docking the partners of each complex are predicted by a neural network trained with a data set containing no members from the SCOP superfamily [83] of the complex being tested for docking.

We defined a quantity called enrichment to measure how capable a scoring function is to select the most accurate docked models from a



**Fig. 7.** Performance of the neural network-based interface residue prediction model trained using interface and surface residues in the membrane region only. (A) Scatter plot of true WCN versus WCN predicted by the neural network. The Pearson correlation coefficient between true WCN and predicted WCN is 0.70 and the mean absolute error of predicted WCN from true WCN is 1.58. (B) The ROC curve for interface residue classification with an AUC of 0.75. The ROC curve is colored according to classification thresholds and the scale on the right vertical axis indicates the mapping between colors and classification thresholds  $\Delta\text{WCN}$ .

**Table 3**

Summary of the global docking of IMPs using predicted interface residue WCN as restraints.

Protein ID	# Interface residues	TPR	PPV	MAE	Best RMSD100 (Å)		Mean RMSD100 of top 1% models ranked by RMSD100 (Å)		Mean RMSD100 of top 1% models ranked by score (Å)		Enrichment	
					$\emptyset_{\text{WCN}}$	$P_{\text{WCN}}$	$\emptyset_{\text{WCN}}$	$P_{\text{WCN}}$	$\emptyset_{\text{WCN}}$	$P_{\text{WCN}}$	$\emptyset_{\text{WCN}}$	$P_{\text{WCN}}$
					1q16_CF	46	0.67	0.17	2.09	1.1	1.1	2.2
2a65_AB	65	0.64	0.14	1.74	0.7	0.8	1.4	1.4	5.3	5.9	3.0	3.1
2bs2_CF	54	0.67	0.22	1.81	1.2	1.3	2.9	2.9	28.0	33.3	0.6	0.6
2nq2_AB	81	0.92	0.28	1.89	0.7	1.1	2.9	2.2	40.0	5.1	0.4	6.6
2vpz_CG	52	0.61	0.21	2.18	0.7	0.9	2.0	1.9	10.5	24.2	2.8	0.8
2z73_AB	66	0.69	0.14	2.16	1.9	2.6	4.2	4.2	26.1	28.7	0.7	0.8
3odu_AB	81	0.86	0.07	2.50	0.9	1.2	3.5	4.1	35.2	30.0	1.1	1.3
3puw_FG	123	0.93	0.45	1.99	3.2	1.0	6.9	7.4	25.7	13.4	1.8	3.5
4a01_AB	50	0.72	0.52	1.46	0.6	0.7	2.4	2.6	5.4	4.3	4.1	8.0
4jkv_AB	58	0.70	0.28	2.14	0.9	0.5	2.4	2.1	19.8	8.0	2.2	4.4
4mrs_AB	55	0.75	0.38	1.77	2.9	9.5	16.4	15.8	29.1	26.5	0.5	1.0
4o6m_AB	24	0.55	0.50	1.46	0.6	0.9	2.3	2.2	12.4	3.4	2.1	4.8
4o6y_AB	56	0.95	0.36	1.77	1.1	0.8	2.0	1.8	21.4	5.6	1.6	5.6
4qnd_AB	30	0.81	0.70	1.70	1.7	0.9	3.0	2.9	12.7	5.3	1.8	4.6
4rng_AC	53	0.87	0.49	1.49	1.0	1.2	3.0	2.9	21.8	4.1	1.5	7.6
4u9n_AB	72	0.90	0.63	1.86	0.7	1.2	3.1	2.8	17.3	4.9	1.4	6.2

TPR: true positive rate; PPV: positive predictive value; MAE: mean absolute error;  $\emptyset_{\text{WCN}}$  indicates docking using the baseline scoring function, which lacks use of predicted  $\Delta_{\text{WCN}}$ ;  $P_{\text{WCN}}$  indicates adding predicted  $\Delta_{\text{WCN}}$  to the baseline scoring function. Cases for which the enrichment of the penalty score is  $<2.0$  are colored in red and those for which the enrichment is higher than 2.0 are colored in green.

pool of candidates. A scoring function with an enrichment of  $r$  means that models top-ranked by the scoring function are  $r$  times more likely than a randomly selected set of models to contain native-like models. As shown in Table 3, the enrichment of the WCN-based penalty score is improved in 13 cases, compared to the enrichment of the baseline scoring function. The number of cases where the enrichment is  $>2.0$  is increased from five to ten cases. The cases where the enrichment of the penalty score is worse than that of the baseline scoring function or  $<2.0$  is likely due to poor classification of interface residues or poor accuracy of WCN prediction. For example, cases for which the enrichment of the penalty score is  $<2.0$  (colored in red in Table 3) have an average TPR, PPV, and MAE of 0.71, 0.20, and 2.09, whereas the average TPR, PPV, and MAE for cases for which the enrichment is  $>2.0$  (colored in green in Table 3) are 0.80, 0.44, and 1.75, respectively. To demonstrate the effectiveness of the penalty score, we plot the “score v.s. rmsd” relationship and compare the best docked model in the 1% models ranked by this score to the native complex structure for cases where the enrichment is greater 2.0 (Fig. 8).

#### 4. Discussion

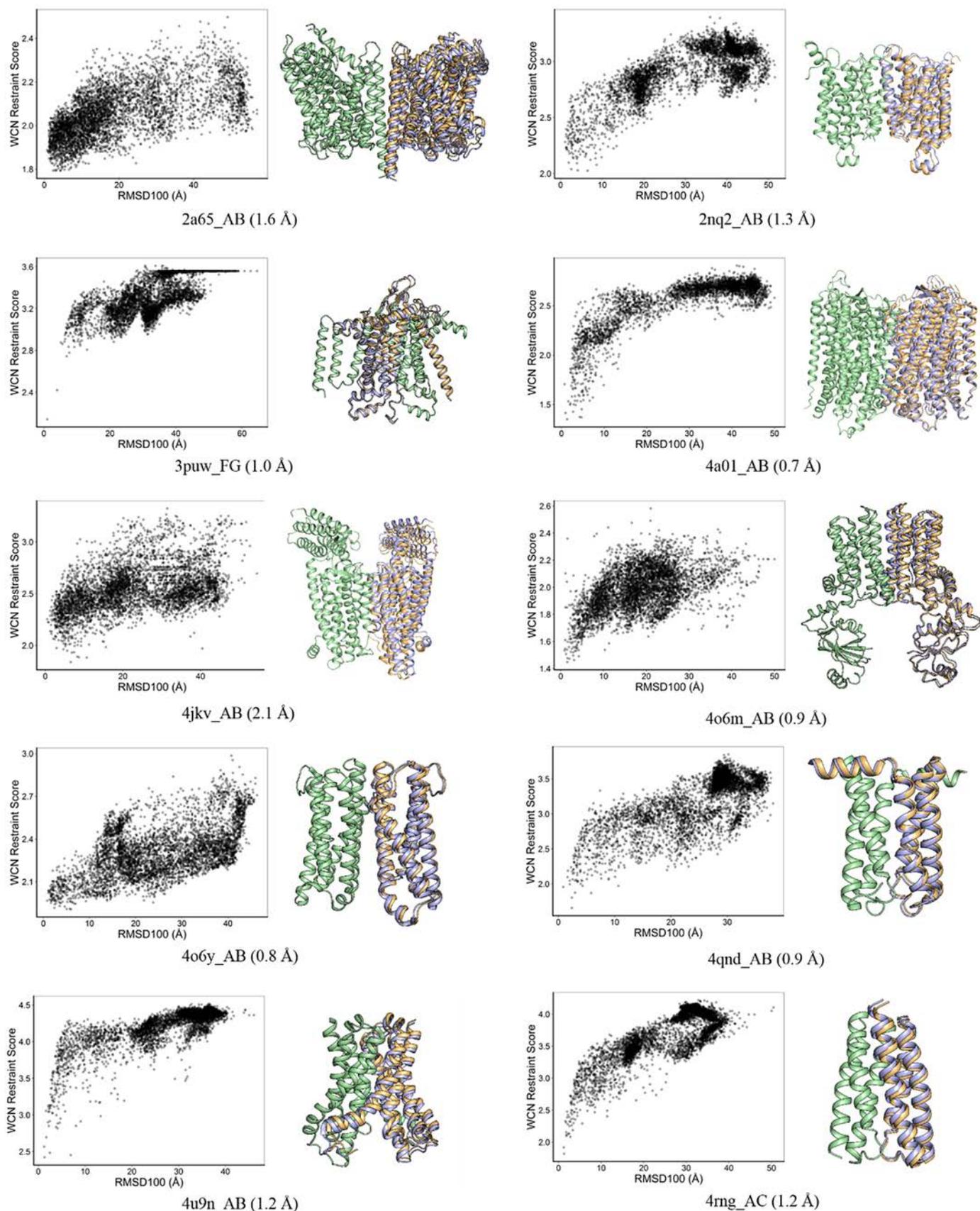
Herein, we compiled a non-redundant set of oligomers of  $\alpha$ -helical IMPs whose structure have been determined to high resolution and studied the properties of the interfaces in terms of hydrophobicity, amino acid composition, interface propensity of amino acids, evolutionary conservation, and correlation of amino acid pairs at the interface. We found that the aqueous portion of the interface in oligomers of IMPs have similar characteristics to the interfaces between oligomers of globular proteins [11,14,17,70,95]. Within the membrane, while the average hydrophobicity of the interface is similar to the rest of the lipid-contacting surface, however, the interface is significantly more conserved. We also observed that contacting residue pairs across the interface tend to correlate more strongly than non-contacting residue pairs.

While our results suggest that interfaces within the membrane are not significantly different from the rest of the surface for the global

physicochemical compositions, two important questions remain: are specific motifs of the GxxxG type enriched in the interface of multi-pass IMP oligomers? And if so, how do such motifs contribute to multi-pass IMP oligomerization? It is known that the GxxxG motif plays an important role in stabilizing single-pass homodimers [88,89]. Within multi-pass IMPs, previous studies have indicated that GxxxG motifs can be important for both folding and oligomerization and it has been suggested that the presence of a GxxxG type of motif alone is only a weak predictor of protein dimerization in the membrane [89]. However, we are not aware of any study that has systematically examined whether the GxxxG motifs have a higher propensity for inter-subunit than intra-subunit interactions. Given the added complexities in the process of multi-pass IMP folding and oligomerization, separate in-depth studies are needed to answer these questions.

Based on the observation that the interface is significantly more conserved than the rest of the surface in the membrane, we adapted our previously-developed neural network-based method [68] to distinguish interface residues from residues on the rest of the surface. This classification was based on the WCNs of surface residues predicted by the neural network. While the performance of this method is comparable to the random forest-based binary classifier developed by Bordner [69], its strength lies in that it predicts residues' real-valued WCN. Residues' WCN is an effective restraint for improving the fraction of native contacts in predicted structural models for *de novo* prediction of tertiary structures of  $\alpha$ -helical IMPs [23], and we have herein shown that it can be further used to derive an effective score for selecting native-like docking candidates of IMP complexes.

The sampling problem in docking IMPs is inherently more tractable than that in docking globular proteins, due to the reduced translational and rotational degree of freedom with respect to the membrane normal. Despite this simplification in sampling, docking IMPs is still a challenging problem in terms of scoring native-like poses. Intuitively, we expected that matching relatively polar patches on the IMPs transmembrane surface, so as to shield polar regions from the membrane, would be beneficial, however, we unexpectedly found that intra-membrane interface regions do not differ from non-interface



**Fig. 8.** Examples of oligomers for which WCNs helped identify the correct docking solutions. Plots on WCN restraint score v.s. RMSD100 relationship, and comparison of the best docked model in the 1% models ranked by WCN restraint score to the native oligomer structure (pale green: native receptor subunit, pale blue: native ligand subunit, light orange: docked ligand subunit) for cases where the enrichment of the WCN restraint score is higher than 2 using  $\Delta WCN = 1$  as the threshold for classifying interface residue. The values in parentheses are the RMSD100 of the models shown.

surface regions in terms of hydrophobicity. Shape complementarity, or packing volume is similarly of limited use due to IMPs having relatively cylindrical shapes. Using restraints derived from evolutionary analysis may represent an effective approach to narrowing down the set of viable docking candidates for IMPs. In fact, the effectiveness of using predicted coevolving residue pairs to identify native-like docking solutions has been demonstrated previously for globular proteins [91,96–100] and some isolated cases of IMPs [97,98].

While coevolutionary analysis has the benefit of pinpointing specific interacting residue pairs, it requires construction of paired MSA of interacting partners, and computationally, this problem becomes difficult by itself due to the existence of paralogs [18–20]. We have taken a novel approach by first predicting WCNs of surface residues from sequence for monomeric vs. multimeric proteins and using the difference between these predictions to identify likely interface residues. Next, we used the predicted interface residues as “sticky” points and their predicted WCNs as restraints for ranking docking solutions. This approach is computationally efficient and easy to implement as it eliminates the necessity of creating paired MSAs. We have also demonstrated the effectiveness of this approach using a benchmark set of 16  $\alpha$ -helical IMP oligomers.

The effectiveness of our approach depends on the accuracy of the prediction of WCNs and the identification of interface residues. Structural models ranked top 1% by the WCN restraint score are closer to the native structure (lower average RMSD100) than models ranked 1% by the baseline scoring function are when either the PPV for classifying interface residues is above 0.28 or the MAE for WCN prediction is below 2.0 (Table 3). When the PPV drops below 0.28 and the MAE goes above 2.0, the WCN score starts to act against identifying correct docking solutions. We realize that such criteria are arbitrary due to the limited size of the benchmark set and that in a real-life scenario a user of our method wouldn't know the PPV or MAE for their protein of interest. Thus, we'd like to point out that non-obligate oligomers where structural or functional constraint on the interface may be too weak to leave detectable evolutionary information on interface residues or incidental oligomers where the interface is neither structurally nor functionally relevant [6] are particularly challenging to our method. In fact, five out of the six non-obligate dimers in the benchmark set have a PPV <0.28 or a MAE >2.0. This also highlights the need for both computational and experimental studies to better understand non-obligate interactions between membrane proteins and the development of more accurate methods for predicting interaction sites.

Nevertheless, the current study provides important insights into the interface of  $\alpha$ -helical IMP oligomers and presents an effective and robust approach for docking  $\alpha$ -helical IMPs. The statistics obtained may give insights into the development of methods that are more accurate in predicting interface residues, while the docking approach is valuable for constructing structural models for  $\alpha$ -helical IMP oligomers.

## Software and Data Availability

BCL::MP-Dock has been integrated into the Biochemical Library (BCL) software suite that is being actively developed. It is available at <http://www.meilerlab.org/bclcommons> under academic and business site licenses. The BCL source code is published under the BCL license and is available at <http://www.meilerlab.org/bclcommons>. Weighted contact numbers of amino acid residues in helical membrane proteins can be predicted using BCL::TMH-Expo via its webserver: [http://www.meilerlab.org/servers/tmh\\_expo](http://www.meilerlab.org/servers/tmh_expo). The raw data, processed data, and protocol capture for this work are available at <http://dx.doi.org/10.17632/cbk98yrszd.1>.

## Declarations of Interest

None.

## Sources of Funding

This project was supported by National Institutes of Health (NIH) Grant R01 HL122010 and R01 GM080403. B. Li was also supported by American Heart Association Pre-Doctoral Fellowship Award 16PRE27260211.

## Appendix A. Supplementary Data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2019.05.005>.

## References

- [1] Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L. Prediction of the human membrane proteome. *Proteomics* 2010;10:1141–9.
- [2] Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean BD, et al. Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* 2012;489:585–9.
- [3] Daley DO. The assembly of membrane proteins into complexes. *Curr Opin Struct Biol* 2008;18:420–4.
- [4] Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, et al. Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci U S A* 2005;102:12123–8.
- [5] Jespersen T, Grunnet M, Olesen SP. The KCNQ1 potassium channel: from gene to physiological function. *Physiology* 2005;20:408–16.
- [6] Nooren IMA, Thornton JM. Diversity of protein-protein interactions. *EMBO J* 2003;22:3486–92.
- [7] Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. *Structure* 2010;18:1233–43.
- [8] Ansari S, Helms V. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins* 2005;61:344–55.
- [9] Dey S, Pal A, Chakrabarti P, Janin J. The subunit interfaces of weakly associated homodimeric proteins. *J Mol Biol* 2010;398:146–60.
- [10] Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 1996;93:13–20.
- [11] Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* 1997;272:121–32.
- [12] Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* 1997;272:133–43.
- [13] Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol* 1999;285:2177–98.
- [14] Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* 2003;325:991–1018.
- [15] Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 2004;13:190–202.
- [16] Mintseris J, Weng ZP. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 2005;102:10930–5.
- [17] Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. *Protein J* 2008;27:59–70.
- [18] Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. *Proc Natl Acad Sci U S A* 2016;113:12180–5.
- [19] Burger L, van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* 2008;4:165.
- [20] Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci U S A* 2016;113:12186–91.
- [21] Shih CH, Chang CM, Lin YS, Lo WC, Hwang JK. Evolutionary information hidden in a single protein structure. *Proteins* 2012;80:1647–57.
- [22] Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol Biol Evol* 2014;31:135–9.
- [23] Li B, Mendenhall J, Nguyen ED, Weiner BE, Fischer AW, Meiler J. Improving prediction of helix-helix packing in membrane proteins using predicted contact numbers as restraints. *Proteins* 2017;85:1212–21.
- [24] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* 2012;40:D370–6.
- [25] Schobert B, Cupp-Vickery J, Hornak V, Smith SO, Lanyi JK. Crystallographic structure of the K intermediate of bacteriorhodopsin: conservation of free energy after photoisomerization of the retinal. *J Mol Biol* 2002;321:715–26.
- [26] Svensson-Ek M, Abramson J, Larsson G, Tornroth S, Brzezinski P, Iwata S. The X-ray crystal structures of wild-type and EQ(L-286) mutant cytochrome c oxidases from *Rhodospirillum rubrum*. *J Mol Biol* 2002;321:329–39.
- [27] Bertero MG, Rothery RA, Palak M, Hou C, Lim D, Blasco F, et al. Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. *Nat Struct Biol* 2003;10:681–7.
- [28] Khademi S, O'Connell J, Remis J, Robles-Colmenares Y, Miericke LJW, Stroud RM. Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å resolution. *Science* 2004;305:1587–94.

- [29] Huang LS, Sun G, Cobessi D, Wang AC, Shen JT, Tung EY, et al. 3-Nitropropionic acid is a suicide inhibitor of mitochondrial respiration that, upon oxidation by complex II, forms a covalent adduct with a catalytic base arginine in the active site of the enzyme. *J Biol Chem* 2006;281:5965–72.
- [30] Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E. Crystal structure of a bacterial homologue of Na<sup>+</sup>/Cl<sup>-</sup>-dependent neurotransmitter transporters. *Nature* 2005;437:215–23.
- [31] Murata T, Yamato I, Kakinuma Y, Leslie AGW, Walker JE. Structure of the rotor of the V-type Na<sup>+</sup>-ATPase from *Enterococcus hirae*. *Science* 2005;308:654–9.
- [32] Madej MG, Nasiri HR, Hilgendorff NS, Schwalbe H, Lancaster CRD. Evidence for transmembrane proton transfer in a dihaem-containing membrane protein complex. *EMBO J* 2006;25:4963–70.
- [33] Koepke J, Krammer EM, Klungen AR, Sebban P, Ullmann GM, Fritzsche G. pH modulates the quinone position in the photosynthetic reaction center from *Rhodospirillum rubrum* in the neutral and charge separated states. *J Mol Biol* 2007;371:396–409.
- [34] Pinkett HW, Lee AT, Lum P, Locher KP, Rees DC. An inward-facing conformation of a putative metal-chelate-type ABC transporter. *Science* 2007;315:373–7.
- [35] Jasti J, Furukawa H, Gonzales EB, Gouaux E. Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. *Nature* 2007;449:316–23.
- [36] Molina DM, Wetterholm A, Kohl A, McCarthy AA, Niegowski D, Ohlson E, et al. Structural basis for synthesis of inflammatory mediators by human leukotriene C-4 synthase. *Nature* 2007;448:613–6.
- [37] Jormakka M, Yokoyama K, Yano T, Tamakoshi M, Akimoto S, Shimamura T, et al. Molecular mechanism of energy conservation in polysulfide respiration. *Molecular Biology; Nat Struct Mol Biol* 2008;15:730–7.
- [38] Fischer G, Kosinska-Eriksson U, Aponte-Santamaria C, Palmgren M, Geijer C, Hedfalk K, et al. Crystal structure of a yeast aquaporin at 1.15 Å resolution reveals a novel gating mechanism. *PLoS Biol* 2009;7:e1000130.
- [39] Schulze S, Koster S, Geldmacher U, van Scheltinga ACT, Kuhlbrandt W. Structural basis of Na<sup>+</sup>-independent and cooperative substrate/product antiport in CaT. *Nature* 2010;467:233–6.
- [40] Lyons JA, Aragao D, Slattery O, Pisljakov AV, Soulimane T, Caffrey M. Structural insights into electron transfer in caa(3)-type cytochrome oxidase. *Nature* 2012;487:514–8.
- [41] Murakami M, Kouyama T. Crystal structure of squid rhodopsin. *Nature* 2008;453:363–7.
- [42] Shinoda T, Ogawa H, Cornelius F, Toyoshima C. Crystal structure of the sodium-potassium pump at 2.4 Å resolution. *Nature* 2009;459:446–50.
- [43] Li X, Jayachandran S, Nguyen HHT, Chan MK. Structure of the nitrosomonas europaea Rh protein. *Proc Natl Acad Sci U S A* 2007;104:19279–84.
- [44] Newby ZER, O'Connell J, Robles-Colmenares Y, Khademi S, Miercke IJ, Stroud RM. Crystal structure of the aquaglyceroporin PfAQP from the malarial parasite *Plasmodium falciparum*. *Molecular Biology; Nat Struct Mol Biol* 2008;15:619–25.
- [45] Solmaz SRN, Hunte C. Structure of complex III with bound cytochrome c in reduced state and definition of a minimal core interface for electron transfer. *J Biol Chem* 2008;283:17542–9.
- [46] Levin EJ, Quick M, Zhou M. Crystal structure of a bacterial homologue of the kidney urea transporter. *Nature* 2009;462:757–61.
- [47] Waight AB, Love J, Wang DN. Structure and mechanism of a pentameric formate channel. *Molecular Biology; Nat Struct Mol Biol* 2010;17:31–7.
- [48] Chen YH, Hu L, Punta M, Bruni R, Hillerich B, Kloss B, et al. Homologue structure of the SLAC1 anion channel for closing stomata in leaves. *Nature* 2010;467:1074–80.
- [49] Derebe MG, Sauer DB, Zeng WZ, Alam A, Shi N, Jiang YX. Tuning the ion selectivity of tetrameric cation channels by changing the number of ion binding sites. *Proc Natl Acad Sci U S A* 2011;108:598–602.
- [50] Wu BL, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, et al. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* 2010;330:1066–71.
- [51] Oldham ML, Chen J. Snapshots of the maltose transporter during ATP hydrolysis. *Proc Natl Acad Sci U S A* 2011;108:15152–6.
- [52] Tiefenbrunn T, Liu W, Chen Y, Katritch V, Stout CD, Fee JA, et al. High resolution structure of the ba3 cytochrome c oxidase from *Thermus thermophilus* in a lipidic environment. *PLoS One* 2011;6:e22348.
- [53] Hansen SB, Tao X, MacKinnon R. Structural basis of PIP2 activation of the classical inward rectifier K<sup>+</sup> channel Kir2.2. *Nature* 2011;477:495–8.
- [54] Johnson ZL, Cheong CG, Lee SY. Crystal structure of a concentrative nucleoside transporter from *Vibrio cholerae* at 2.4 Å resolution. *Nature* 2012;483:489–93.
- [55] Lin SM, Tsai JY, Hsiao CD, Huang YT, Chiu CL, Liu MH, et al. Crystal structure of a membrane-embedded H<sup>+</sup>-translocating pyrophosphatase. *Nature* 2012;484:399–403.
- [56] Li DF, Howe N, Dukkipati A, Shah STA, Bax BD, Edge C, et al. Crystallizing membrane proteins in the lipidic mesophase. Experience with human prostaglandin E2 synthase 1 and an evolving strategy. *Crystr Growth Des* 2014;14:2034–47.
- [57] Boland C, Li DF, Shah STA, Haberstock S, Dotsch V, Bernhard F, et al. Cell-free expression and in meso crystallisation of an integral membrane kinase for structure determination. *Cell Mol Life Sci* 2014;71:4895–910.
- [58] Eicher T, Cha HJ, Seeger MA, Brandstatter L, El-Delik J, Bohnert JA, et al. Transport of drugs by the multidrug transporter AcrB involves an access and a deep binding pocket that are separated by a switch-loop. *Proc Natl Acad Sci U S A* 2012;109:5687–92.
- [59] Symersky J, Osowski D, Walters DE, Mueller DM. Oligomycin frames a common drug-binding site in the ATP synthase. *Proc Natl Acad Sci U S A* 2012;109:13961–5.
- [60] Wang C, Wu HX, Katritch V, Han GW, Huang XP, Liu W, et al. Structure of the human smoothened receptor bound to an antitumour agent. *Nature* 2013;497:338–43.
- [61] Lee JY, Yang JG, Zhitnitsky D, Lewinson O, Rees DC. Structural basis for heavy metal detoxification by an Atm1-type ABC exporter. *Science* 2014;343:1133–6.
- [62] Sclari G, Clarke OB, Tomasek D, Kloss B, Tabuso S, Byfield R, et al. Structural basis for catalysis in a CDP-alcohol phosphotransferase. *Nat Commun* 2014;5.
- [63] Lu PL, Ma D, Yan CY, Gong XQ, Du MJ, Shi YG. Structure and mechanism of a eukaryotic transmembrane ascorbate-dependent oxidoreductase. *Proc Natl Acad Sci U S A* 2014;111:1813–8.
- [64] Xu Y, Tao YY, Cheung LS, Fan C, Chen LQ, Xu S, et al. Structures of bacterial homologues of SWEET transporters in two distinct conformations. *Nature* 2014;515:448–52.
- [65] Wang J, Yan CY, Li YN, Hirata K, Yamamoto M, Yan N, et al. Crystal structure of a bacterial homologue of SWEET transporters. *Cell Res* 2014;24:1486–9.
- [66] Takeda H, Hattori M, Nishizawa T, Yamashita K, Shah STA, Caffrey M, et al. Structural basis for ion selectivity revealed by high-resolution crystal structure of Mg<sup>2+</sup> channel MgtE. *Nat Commun* 2014;5.
- [67] Yang TT, Liu Q, Kloss B, Bruni R, Kalathur RC, Guo YZ, et al. Structure and selectivity in bestrophin ion channels. *Science* 2014;346:355–9.
- [68] Li B, Mendenhall J, Nguyen ED, Weiner BE, Fischer AW, Meiler J. Accurate prediction of contact numbers for multi-spanning helical membrane proteins. *J Chem Inf Model* 2016;56:423–34.
- [69] Bordner AJ. Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics* 2009;10:312.
- [70] Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. *Proteins* 2005;60:353–66.
- [71] Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. *J Mol Biol* 1987;196:641–56.
- [72] Miller S, Lesk AM, Janin J, Chothia C. The accessible surface area and stability of oligomeric proteins. *Nature* 1987;328:834–6.
- [73] Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 2002;18(Suppl. 1):S71–7.
- [74] Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;9:173–5.
- [75] Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci U S A* 2000;97:3288–91.
- [76] Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004;56:211–21.
- [77] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008;24:333–40.
- [78] Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers Jr RG, Haussler D. Information-theoretic dissection of pairwise contact potentials. *Proteins* 2002;49:7–14.
- [79] Li B, Mendenhall JL, Kroncke BM, Taylor KC, Huang H, Smith DK, et al. Predicting the functional impact of KCNQ1 variants of unknown significance. *Circ Cardiovasc Genet* 2017;10.
- [80] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;15:1929–58.
- [81] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533–6.
- [82] Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (Roc) curve. *Radiology* 1982;143:29–36.
- [83] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–40.
- [84] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;16:412–24.
- [85] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys* 1953;21:1087–92.
- [86] Woetzel N, Karakas M, Staritzbichler R, Muller R, Weiner BE, Meiler J. BCL:Score—knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS One* 2012;7:e49242.
- [87] Lomize AL, Pogozheva ID, Mosberg HL. Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *J Chem Inf Model* 2011;51:930–46.
- [88] Senes A, Engel DE, DeGrado WF. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol* 2004;14:465–79.
- [89] Teese MG, Langosch D. Role of GxxxG motifs in Transmembrane domain interactions. *Biochemistry* 2015;54:5125–35.
- [90] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105–32.
- [91] Pazos F, Helmer Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 1997;271:511–23.
- [92] Wang Y, Barth P. Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy. *Nat Commun* 2015;6:7196.
- [93] Marchetti F, Capelli R, Rizzato F, Laio A, Colombo G. The subtle trade-off between evolutionary and energetic constraints in protein-protein interactions. *J Phys Chem Lett* 2019;10:1489–97.
- [94] Li B, Fooks M, Heinze S, Meiler J. Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Crit Rev Biochem Mol Biol* 2018;53:1–28.

- [95] Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient protein-protein interactions. *Protein Eng Des Sel* 2011;24:635–48.
- [96] Madaoui H, Guerois R. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc Natl Acad Sci U S A* 2008;105:7708–13.
- [97] Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 2014;3.
- [98] Ovchinnikov S, Kinch L, Park H, Liao YX, Pei JM, Kim DE, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* 2015;4.
- [99] Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci U S A* 2017;114:E2662–71.
- [100] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;106:67–72.
- [101] Wang G, Dunbrack Jr RL. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–91.
- [102] Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci* 2001;10:1470–3.