# TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung

und Umwelt

Fachgebiet für Bioinformatik

## Computational analysis and prediction of protein interaction sites in transmembrane proteins

BO ZENG

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Jan Baunbach

Prüfer der Dissertation:
1. Prof. Dr. Dmitrij Frishman
2. Prof. Dr. Dieter Langosch

Die Dissertation wurde am 17.12.2018 bei der Technischen Universität München eingereicht und durch die Fakultüt Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 18.03.2019 angenommen

# ABSTRACT

Alpha-helical membrane proteins constitute 25-30% of all proteins in all sequenced genomes and are vital for many biological processes, more than 50% of already reported drugs are transmembrane protein targeted. Due to the scarce 3D structure for this kind of protein, sequence-based interaction sites prediction tools are highly motivated for membrane protein structure prediction, mutagenesis, and better small molecular drug design.

In this thesis, firstly, for the transmembrane domains (TMDs) of single-pass membrane proteins, we have created the first machine learning algorithm for the prediction of TM homodimer interface residues. The Transmembrane HOmodimer Interface Prediction Algorithm (THOIPA) utilized evolutionary sequence information alone. We used 54 nonredundant self-interacting TMDs (20 experimental ETRA , 8 NMR and 25 crystal) as training and validation dataset, THOIPA obviously outperformed other currently available prediction methods according to the overall prediction performance of AUC or AUBOC10, it was particularly powerful for the prediction of the top residues involved in the interaction. Furthermore, we found that the interface residues involved in protein-protein interactions are significantly conserved, more co-evolved and more polar than non-interface residues, and the GxxxG motifs were overrepresented at TM interfaces, particularly when investigated in a natural membrane environment. The THOIPA code and standalone predictor is available at https://github.com/bojigu/thoipapy. The online webserver is available at http://www.thoipa.com/.

Secondly, we developed MBPred (Membrane Binding-site Prediction) for the interfacial residues prediction of alpha-helical membrane proteins, which is a suite of four individual RF models – MBPredTM, MBPredCyto, MBPredExtra, and MBPredAll – trained to predict protein interaction sites in transmembrane, cytoplasmic, and extracellular segments as well as in entire amino acid sequences, respectively. This study found that in comparison of non-interacting residues, interacting residues are more conserved, more co-evolved, and have interface position preferences along the protein segment and full sequence. The overall prediction performance (AUC) over 10-fold cross-validation for each of the four individual RF models are higher than 0.78. While for the 36 new independent test dataset, the prediction performance AUC can also reach on average of 0.75. The MBPred code and the standalone predictor is available at https://github.com/bojigu/MBPred.

# ZUSAMMENFAFSSUNG

Alpha-helikale Membranproteine machen 25-30% aller Proteine in allen sequenzierten Genomen aus und sind für viele biologische Prozesse unerlässlich, mehr als 50% der bereits berichteten Medikamente sind auf Transmembranproteine ausgerichtet. Aufgrund der knappen 3D-Struktur für diese Art von Protein sind sequenzbasierte Interaktionsstellen-Prädiktionswerkzeuge hoch motiviert für die Vorhersage von Membranproteinstrukturen, Mutagenese und besseres kleinmolekulares Wirkstoffdesign.

In dieser Arbeit, erstens, für die Transmembrandomänen (TMDs) von Single-Pass-Membranproteinen, haben wir den ersten maschinellen Lernalgorithmus für die Vorhersage von TM-Homodimer-Schnittstellenresten entwickelt. Der Transmembran HOmodimer Interface Prediction Algorithmus (THOIPA) verwendete evolutionäre Sequenzinformationen allein. Wir verwendeten 54 nicht-redundante selbst-interagierende TMDs (20 experimentelle ETRA, 8 NMR und 25 Kristall) als Trainings- und Validierungsdatensatz, THOIPA übertraf offensichtlich andere derzeit verfügbare Vorhersagemethoden entsprechend der gesamten Vorhersageleistung von AUC oder AUBOC10, es war besonders leistungsfähig für die Vorhersage der obersten Rückstände, die an der Interaktion beteiligt waren. Darüber hinaus fanden wir heraus, dass die an Protein-Protein-Interaktionen beteiligten Interface-Reste signifikant konserviert, ko-evolutionärer und polarer sind als nicht-interface-Reste, und die GxxxG-Motive waren an TM-Schnittstellen überrepräsentiert, insbesondere wenn sie in einer natürlichen Membranumgebung untersucht wurden. Der THOIPA-Code und der eigenständige

Prädiktor sind verfügbar unter https://github.com/bojigu/thoipapy. Der Online-Webserver könnte auch unter http://www.thoipa.com/. verfügbar sein.

Zweitens haben wir MBPred (Membrane Binding-site Prediction) für die Vorhersage von Grenzflächenrückständen von alpha-helikalen Membranproteinen entwickelt, eine Suite von vier einzelnen HF-Modellen - MBPredTM, MBPredCyto, MBPredExtra und MBPredAll -, die für die Vorhersage von Proteininteraktionsstellen in transmembranen, zytoplasmatischen und extrazellulären Segmenten sowie in ganzen Aminosäuresequenzen ausgebildet sind. Diese Studie ergab, dass im Vergleich zu nicht interagierenden Rückständen interagierende Rückstände konservierter, ko-evolutionärer und mit Präferenzen für die Schnittstellenposition entlang des Proteinsegments und der gesamten Sequenz sind. Die Gesamtvorhersageleistung (AUC) über die 10-fache Kreuzvalidierung für jedes der vier einzelnen HF-Modelle ist höher als 0.78. Während für den 36 neuen unabhängigen Testdatensatz die Vorhersageleistung AUC ebenfalls durchschnittlich 0.75 erreichen kann. Der MBPred-Code und der eigenständige Prädiktor sind verfügbar unter https://github.com/bojigu/MBPred.

# ACKNOWLEDGEMENTS

I would like to thank all people who helped and supported me during past years.

Prof. Dr. Dmitrij Frishman :: the best supervisor I can imagine, gives me a lot of freedom on scientific work, supports me from the beginning till the end by offering me assistance and inspiration. For highly encouraging trust on my work and providing me a lot of opportunities to exchange scientific knowledge. I also appreciate the help in improving my paper and thesis and I really learned a lot from it.

Prof. Dr. Dieter Langosch :: for initiating and supporting the fruitful collaboration project addressing TMD homodimer interface determination, for providing me the opportunity of catching a glimpse of membrane protein wet lab work and agreeing to review this thesis.

Dr. Mark Teese :: for guiding and supporting the TMD homodimer interface prediction study, thanks for organizing the weekly TMD homodimer protocol meeting and providing a lot of great ideas on the project, thanks for sharing your python programming experiences and a lot of work for editing the publication and this thesis.

Yao Xiao :: a great co-operator with a bunch of biology and wet lab experience, thanks for sharing your knowledge on biology and experiment and a lot of courage during my depression time on my Ph.D. study.

Peter Hönigschmid :: a great teammate on transmembrane group, thanks for all of your help during my transmembrane protein projects, I was greatly assisted on machine learning, transmembrane protein interaction, and paper writing.

Leonie Corry :: thanks for her excellent administrative help during my years in the lab.

Hongen Xu :: thanks a lot for sharing you knowledge about bioinformatic skills, and a lot of help for my research and life.

Jinglong Ru :: your help from both academic work and daily life are greatly appreciated.

I would like to express my gratitude to my other colleagues in Department of Genome-oriented Bioinformatics. Thanks Drazen Jalsovec for his work in maintaining IT infrastructure. Special thanks go to Yanping Zhang, Usman Saeed, Nermin Pinar Karabulut and Evans Kataka for insightful discussions and interesting conversations.

I would like to thank all of my friends who supported me a lot during my life in Munich, Yang Wang, Xiao Wu, Xiuli He, HaiLong He, Bing Zhou, Lianhua Li, Fangyin Cao, Hongchang Yang, Yuhong Mao, Yanli Zhang, Michael Schmidt, all of you are good friends I will cherish in my life.

I would like to acknowledge the financial support of the China Scholarship Council. I appreciate having such an opportunity to study in Technical University of Munich.

Finally, I wish to thank my parents Qingan Zeng and Fangmei Yuan, my sister Bo Zeng, my brother Yuan Zeng, my brother in law Mingjun Xia and my sister in law. All of your continued support is deeply appreciated.

# DECLARATION

This thesis is my own work. However, Chapter 2 of this thesis was the product of close collaboration with Professor Dieter Langosch Group (TUM) .

Chapter 2 is a cooperation work with Yao Xiao from the department of chemistry of Biopolymers (TUM). Yao Xiao mainly created the experimental mutagenesis data (ETRA), in correspondence with the NMR and crystal data created by myself. Yao also analyzed the residue properties such as conservation, polarity, co-evolution and residue depth, but I also helped Yao to preprocess and analyze these data.

In chapter 2, I only introduce my bioinformatic work but not the experimental work from Yao Xiao, for example, my work includes the creation of the NMR and crystal data, the complete dataset analysis, the creation of the THOIPA machine learning method, and the thoipa webserver. But most of the work I used the ETRA data which was created by Yao Xiao. The detail of the experimental method ToxR and others Yao used, the ETRA dataset, are described in the thesis of Yao Xiao.

Chapter 3 is an independent work of myself, but Peter Hönigschmid supported me lots of ideas, especially helped me a lot on the publication writing.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1. INTRODUCTION

It was well known that proteins are the driving horse of the cellular machinery, they are responsible for diverse functions ranging from molecular motors to signaling. Membrane proteins represent about 20–30% of the genome in a variety of different organisms [15-17]. Protein–protein interactions (PPI) within the membrane are involved in many vital cellular processes, especially understanding interface residue involved in PPI is critical to identity protein functions. Consequently, deficient oligomerization is associated with many diseases. The main target of this thesis is to analyze and predict the transmembrane PPI interfaces.

The following introduction aims at summarize present knowledge about transmembrane PPI. The first chapter briefly outlines the current view of biology membrane and membrane protein. In the following, the interaction of transmembrane protein, including the popular experimental PPI methods, the experimental techniques to study transmembrane domain (TMD) -TMD interactions, also the sequence motifs that mediating TMD-TMD interactions. Further, the current knowledge regarding the computational methods for the transmembrane protein interface residues prediction, briefly introduced homology-based, machine learning and co-evolution methods. Finally, a motivation and overview of this thesis.

## 1.1 Membranes and membrane proteins

### 1.1.1 Biological membranes

Biological membranes are lipid bilayers composed of various phospholipids with average thickness of 60 Å [18]. Biological membrane plays very important role as barrier between the intracellular content and the extracellular environment, they are also found within eukaryotic cells surrounding intracellular compartments such as the nucleus, mitochondria, chloroplasts, the endoplasmic reticulum (ER) and the Golgi apparatus. The Fluid-Mosaic-Model of Singer and Nicolson describes membranes as two-dimensional viscous fluids containing freely diffusing membrane proteins [19]. However, different lipid species are not alone distributed among the leaflets of a bilayer but also organized laterally in the plane.

### 1.1.2 Membrane proteins

Membranes are barriers which molecules generally cannot pass without assistance, and transmembrane proteins (TMP) embedded into the bilayer providing aid to transport signal or molecular. Around 30% of the genes in eukaryotic species are encoding integral membrane proteins [15-17]. In humans, approximately 6,000 different TMPs are expressed [20, 21]. They take part in countless cellular processes, and comprise the majority of targets for pharmaceutical compounds [22]. Based on the secondary structure difference, TMPs could be classified into alpha-helical TMP and beta-barrel TMP. Alpha-helical TMP constitutes between 20 and 30 percent of all ORFs in already sequenced genomes [23]. Beta-barrel TMPs are dominant in the outer membrane of gram-negative bacteria, and a small number are also found in the mitochondria and chloroplast organelles, which are of prokaryotic origin. The available information for TMP is scare in comparison with soluble proteins, there are less than 2% of all structure in the Protein

Data Bank (PDB) are corresponding to TMPs. Even though the number of TMP structures increases exponentially doubling approximately every third year.

## 1.2 Interaction of transmembrane proteins

Proteins are the major players in molecular recognition at the heart of all processes of life, they bind with other proteins to form supramolecular assemblies and elaborate molecular machines that perform all kinds of functions. PPI within membrane are vital for many cellular processes, many diseases were caused by the deficient PPIs [24]. TMPs transmit different signals between extracellular and intracellular environments, these signals play crucial roles such as homeostasis and signal transduction [25]. The signal transduction was initiated by the binding with ligands, which is believed to cause conformational change of the receptor and form oligomerization, this association will stimulate the function of many proteins, thus the deficient oligomerization will caused diseases such as cancer and amyloidal [26]. However, in contrast to PPI for soluble proteins, the knowledge on transmembrane protein interaction are quite limited because of the unique chemical and physical properties of membrane environment. Here I am going to introduce firstly the experimental methods used to detect protein-protein interactions. Secondly for transmembrane proteins, what are the art-of-state techniques to investigate the helix-helix interaction. Finally, an overview of the sequence motifs that mediating TMD-TMD interactions, which are the key factors for transmembrane PPIs.

### 1.2.1 Experimental methods for determining protein-protein interactions

During the last few decades, a variety of experimental methods for measuring PPIs have been developed [74]. Table 1-1 summarized most of the widely used experimental PPI detection methods. These methods were firstly categorized into high-throughput and low-throughput methods, the advantage of high-throughput method is the usage to screen a large quality of interaction, while the later doesn't have this capacity. Western blotting method [27] is the oldest method in the list, the second old method is the protein affinity chromatography [28], x-ray crystallography/NMR spectroscopy [29] are also the low-throughput methods. Among the other high-throughput methods which are capable of detecting various possible protein-protein interactions. Here we briefly describe two high throughput methods: Yeast two hybrid (Y2H) [30], and Tandem affinity purification (TAP) [31].

**Table 1-1: Experimental methods overview for the protein-protein interaction detection**

| Name | High-throughput | Type of interaction | Summary |
|---|---|---|---|
| Yeast two hybrid (Y2H) [30] | + | Direct physical | Yeast two-hybrid is typically carried out by screening a protein of interest against a random library of potential protein partners |
| Tandem affinity purification (TAP) [31] | + | Direct physical | TAP is based on the double tagging of the protein of interest on its chromosomal locus, followed by a two-step purification process and mass spectroscopic analysis |
| Protein microarrays | + | Direct physical | Microarray-based analysis allows the simultaneous analysis of thousands of parameters within a single experiment |
| X-ray crystallography [29] | - | Direct physical, and structure | X-ray crystallography enables visualization of protein structures at the atomic level and enhances the understanding of protein interaction and function |

| | | | |
|---|---|---|---|
| NMR spectroscopy [29] | - | Direct physical, and structure | NMR spectroscopy can even detect weak protein-protein interactions |
| Far western blotting [27] | - | Direct physical | far-western blotting uses a non-antibody protein which can bind the protein of interest to detect protein-protein interaction. |
| Protein affinity chromatography [28] | - | Direct physical | Affinity chromatography is highly responsive, can even detect weakest interactions in proteins, and also tests all the sample proteins equally for interaction |
| Synthetic Lethality [32] | + | Genetic interaction | Synthetic lethality is based on functional interactions rather than physical interaction |
| Co-expression [33] | + | Genetic interactions | proteins from the genes belonging to the common expression-profiling clusters are more likely to interact with each other than proteins from the genes belonging to different clusters |

#### 1.2.1.1 Yeast two hybrid (Y2H)

A yeast two-hybrid (Y2H) [30] experiment is a high-throughput screening method for protein interactions, and it greatly accelerated the speed for measuring protein interactions. It detects the physical interactions of proteins through the downstream activation of a reporter gene. How exactly this transcription is measured depends on the reporter gene. But most commonly it is done by auxotrophic selection, i.e. the ability of the yeast to grow on nutrient-restricted medium.

The advantages of a Y2H screen include: 1) that it is relatively fast and easy way to screen for PPIs; 2) it requires little hands-on time and technical skill and; 3) it is also able to be scaled up by screening yeast libraries of tagged “prey” proteins against a single “bait”, allowing thousands of potential interactions to be screened rapidly.

However, there are several disadvantages in the Y2H method. First, the interaction might not happen in yeast, since a queried protein may require a species specific folding protein, which may lack in yeast. Secondly the whole Y2H screening takes place in the yeast nucleus, thus if the proteins are not co-localised there, the interacting proteins are found to be noninteracting (false negative). Nonetheless, Y2H has been used to measure PPIs in worm [34], fly [35], and human [36].

#### 1.2.1.2 Tandem affinity purification

Tandem affinity purification (TAP) is a method for rapid protein complex purification, which allows rapid purification under native conditions of complexes, even when expressed at their natural level [31]. In first step of the technique, the protein of interest with the TAP tag first binds to beads coated with IgG, the TAP tag is then broken apart by an enzyme, and finally a different part of the TAP tag binds reversibly to beads of a different type. After the protein of interest has been washed through two affinity columns, it can be examined for binding partners [37].

There are some advantage of this method. Firstly, it doesn't require the prior knowledge of complex composition and can determine the protein partners quantitatively. Secondly, it often provides high yield [31]. Lastly, the TAP offers an effective and highly specific means to purify target protein.

However, there are also some disadvantages of this methods. Firstly, it is possible a tag added to a protein might obscure the new protein bind to its interacting partner. Also, the tag might influence the protein expression levels. Lastly a tag added to a protein might

not be sufficiently exposed to allow binding of the protein to the affinity beads or might affect protein function [31].

### 1.2.2 Techniques for studying TMD-TMD interactions within membrane

Transmembrane PPIs are partially or fully mediated by transmembrane domains (TMD), thus studying TMD-TMD interaction becomes very critical. Due to the hydrophobic property, TMPs are relatively insoluble in aqueous solution, which makes the structure study for TMD-TMD interaction experimentally difficult [38]. As helix-helix interactions in membrane are many times flexible by nature, in order to better understand the interaction flexibility within membrane, many techniques were developed in the past decades.

#### 1.2.2.1 The ToxR system

The ToxR system [39] is based on the ToxR transcriptional activator and can detect weak TMD–TMD interactions within the membrane environment of E. coli. In response to an external stimulus, the ToxR protein dimerises via its periplasmic domain. This leads to ToxR interactions at the cytoplasmic side, where the ToxR dimer binds to a tandemly repeated DNA segment within the ctx promoter dimerisation thereby activates transcription of linked virulence genes [40]. For the detection and characterization of high-affinity transmembrane domains with the ToxR system (Figure 1-1), the membrane-spanning domain of the ToxR protein is replaced with the TMD of interest [39], Furthermore, the maltose-binding protein (MBP, encoding by MalE) is attached as periplasmic domain which serves as control for correct membrane insertion as only constructs placing the MalE domain within the periplasm are able to complement the MalE

deficiency of E.coli PD28 cells. A plasmid coding for the chimeric protein (pToxRV) is introduced into the *E. coli* indicator-strain FHK12. In FHK12 cells, the reporter gene *lacZ* is under the control of the *ctx*-promoter [41]. TMD-TMD interaction mediates the self-interaction of ToxR proteins in the cytoplasm leading to transcription activation of *ctx*-promoter.

The ToxR system was designed to detect homo-oligomerization, but was modified further to detect hetero-oligomerization as well [42].

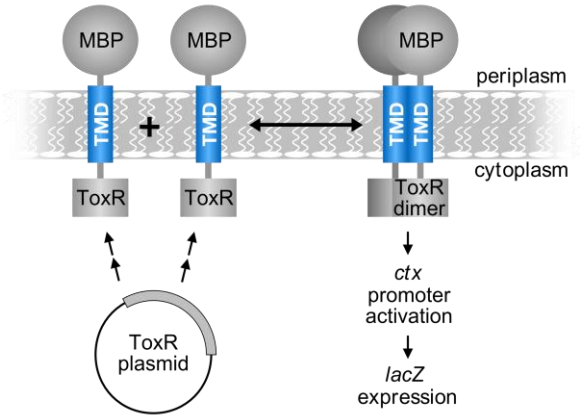



**Figure 1-1: The ToxR system overview.**After self-interaction of transmembrane domains, cytoplasmic ToxR dimers activate the transcription of reporter genes under the control of a ctx or ompU promoter. Periplasmic MalE domains allow for the analysis of correct membrane insertion. Adapted from [39].

#### 1.2.2.2 ETRA (*E. coli* TM Reporter Assay) methods

In combination with scanning mutagenesis, ETRA techniques have now been used for over 20 years to determine interfacial residues of TM homodimers. ToxR-based assays such as ToxR [39], TOXCAT [43], or the recently developed dsTβL [44] were mostly used in studies. Other ETRA techniques include GALLEX [45], BACTH [46, 47] and AraTM [48],

which all uses transcription activator domains, and BLaTM [49], which is based on a split enzyme.

In biological membranes, using ETRA techniques can confirm the interface seen in NMR. Early research on glycophorin A (GpA) and BNIP3 revealed interfacial residues that were generally consistent between SDS-assays [50, 51], ToxR [39, 52], and NMR analyses [1, 3, 53].

#### 1.2.2.3 NMR（Nuclear magnetic resonance) spectroscopy

NMR spectroscopy can be used to obtain information about the dynamic structure of small proteins. Thus, NMR is a good tool to analyse the isolated TMD homodimers, which are typically analysed in detergent micelles or bicelles. The early NMR structure determination of GpA, for example [1], showed a good correspondence with interface residues from earlier SDS-PAGE [50] and ETRA [39] experiments. To date, over 15 TM homodimer structures have been generated using NMR spectroscopy method (Table 5-1) [1-14]. These studies have been reviewed extensively [54, 55], and was used as the test dataset for de-novo structure determination [56-58]. A problem with the NMR dataset is the observation of multiple structures for each TMD, depending on the conditions of the experiment. In some cases, this has been attributed to differences in the lipid-like environment [4, 59], while in other cases it has been proposed that the TMD has multiple biologically relevant homodimer interfaces [59]. It should be noted that the protein concentrations used in a typical NMR experiment are far higher than that seen for individual proteins in biological membranes.

#### 1.2.2.4 X-ray crystallography

Membrane proteins are poorly amenable to crystallisation. The repertoire of TM helix-helix interactions is therefore poorly understood in comparison to soluble proteins. This is due to difficulties in expression, purification and crystallisation [60, 61]. As a consequence, no more than 2% of proteins in the PDB are TMPs [62]. Furthermore, many of these are close homologues, whose structures are not unique. As an example, stringent redundancy reduction of the entire PDB database resulted in the identification of less than 200 unique TMPs [63].

Protein Data Bank of Transmembrane Proteins (PDBTM) database was created to collect TMPs from the PDB and defined their TMD by the TMDET algorithm [64]. The “crystal contacts” within the structures are often considered to be biologically relevant PPI sites [65-68]. Some of these TMD interactions are “homodimer-like,” in that they involve a self-interaction of the same TM helix, between two identical proteins. However, until now, no-one has analysed the self-interacting helices explicitly, despite the fact that they might yield insights into the homotypic interactions of the TM helices of bitopic proteins.

### 1.2.3 Sequence motifs mediating TMD-TMD interactions

#### 1.2.3.1 Conserved motifs

A variety of conserved motifs in TMD were reported to mediate the TMD-TMD interactions: (i) GxxxG motif, which is the most common and best characterized motif for TMD-TIMD interaction, and it was detected to dimerize the of human glycophorin A (GpA) [69]; (ii)

Leucine zipper motif, which is loosely defined as a pattern of leucine, isoleucine or valine residues on one side of the helix face [70], and it controls the dimerization of the transmembrane domain of the platelet-derived growth factor β-receptor (PDGFβR) receptor [71]; (iii) PolarxxxPolar motif, in which the polar residue could be Ser, Thr, Glu, Gln, Asp and Asn, a specific case is the QxxS motif which forms the bacterial aspartate receptor (RAR-1) [72]; (iv) A Ser/Thr rich motif, This motif was found later in the transmembrane domain of Hepatitis C virus (HCV) non- structural protein 4B (NS4B) [73].

#### 1.2.3.2 Polar residues

Within the membrane, the interhelical association can be stabilized by formation of hydrogen bonds, such hydrogen boding is formed between a pair of TMDs through one or more polar residues. It has been studied that amino acids with two polar side-chain atoms have a greater tendency to drive TMD association than residues with only one side-chain polar atom [74] [75].The amino acids with two polar atoms can act simultaneously as a good hydrogen bond donor and acceptor and therefore form a more stable oligomer. It has been also found that non-polar-to-polar mutations in the TMDs of membrane proteins are associated with several diseases [26] [76]. For example, a specific Val→Glu mutation within the TMD of the ErbB2 oncogene product (Neu) [77] is known to induce ErbB2 dimerization and activation. Such activation of ErbB2 has been detected in a large fraction of breast and ovarian cancers [77].

#### 1.2.3.3 Charged residues

Positively charged amino acids, which are localized within the TMDs of membrane proteins, are known to have both functional and structural roles in the activity of these

proteins. Examples include their involvement in substrate recognition [78]. Charge–charge, or ionic, interactions between TMDs came from studies that probed the location of helices within the membrane. There, pairs of positively charged Lys and negatively charged Asp residues one helical turn apart placed a model helix deeper in the membrane than other spacings of the two residues [79]. Moreover, mutations that introduce positively charged residues into TMDs have been previously shown to be involved in human genetic diseases [76, 80]. TMDs are known to be involved in self- and hetero-assembly of membrane proteins. The charged amino acids may also affect the structure of the protein. Therefore, such mutations might interfere with the interactions and proper assembly of the TMDs[24].

#### 1.2.3.4 Aromatic residues

Aromatic residues serve as key structural elements that mediate the molecular recognition and the self-assembly of many membrane proteins including amyloid polypeptides, bacterial toxins and others [81, 82]. It has been reported that the indole, phenol, and imidazole groups of aromatic residues can participate in H-bonding across the TM helix packing interface, thus enhance the TMD dimerization [83]. Studies have also shown that a mutation of a single aromatic amino acid can abolish the ability of the corresponding amyloid peptide to form amyloid fibrils [84].

## 1.3 Computational prediction of protein-protein interaction interfaces

Proteins are the major catalytic agents, as structural elements and transporter, play important roles for signal transduction in cells, individual protein doesn’t function alone, it

binds to other proteins through interaction sites, residue mutations at protein-protein interaction sites usually cause disease. Hence, understanding the characteristics of these binding sites becomes more and more important for the protein design and even for the rational drug design, in addition it helps to better understand the mechanisms of macromolecular recognition. Many biochemical experimental approaches have been developed to identify the protein-protein interfaces, these techniques include X-ray crystallography [85] and NMR spectroscopy [86]. Alana scanning mutagenesis is another method for the interface determination at the residue level. The experimental strategies have some technical challenges, labor-intensive and cost of lots of money. According to the disadvantages of experimental methods, the computational methods for PPI interface prediction becomes extremely valuable. The computational methods could be briefly classified into three categories: (1) knowledge-based method, which is heavily dependent on the experimental structure of the homology data as template. (2) machine learning base method which utilize a dataset of experimentally defined interface residues to train a classification model, and use this model to predict interfacial residues of new proteins. (3) method dependent on the residue co-evolution, which assumed that interacting residues were evolved simultaneously . The multiple sequence alignment was used to identify such kinds of interfacial residues [87, 88]. In the past years, a broad range of computational methods for PPI interface prediction have been developed, some representative methods are summarized in Table 1-2, for each method, it gives the input protein type, sequence or structure, and the available web server, also a brief description of the method.

**Table 1-2: Three main types of protein-protein interaction interface prediction methods.**

| Type | Method | Input | Web server | Description |
|---|---|---|---|---|
| Homology-based | PS-HornPPI [89] | Sequence | http://ailab1.ist.psu.edu/PSHOMPPIv1.2/ | PS-HomPPI predict interfacial residues from the interfacial residues of homologous interacting proteins. PS-HomPPI classifies the templates into Safe, Twilight or Dark Zone, and use multiple templates from the best available zone to infer interfaces for query proteins. |
| | NPS-HornPPI [89] | Sequence | http://ailab1.ist.psu.edu/NPSHOMPPI/ | NPS-HomPPI is the non-partner-specific version of PS-HomPPIs. Without knowledge of the specific binding partner protein, it predicts residues that are likely to interact with other proteins. |
| | PredUs [90] | Structure | https://bhapp.C2b2.columbia.edu/PredUs/ | Inputs a query protein structure, PredUs uses a structural alignment method to identify structural neighbors, calculates the frequency of mapped contacts for each query protein residue and uses the logistic function to generate the residue-based interfacial score. |
| | IBIS [91] | Structure | https://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi | IBIS searches the experimentally determined interfaces for structural homologs of the query protein,, then clusters the interfaces of the homologs, and rank the clustered interfaces. |
| | PriSE [92] | Structure | http://ailab1.ist.psu.edu/prise/index.py | PriSE calculates a surface patch consisting of this target residues and its special neighbors, then searches surface patch database for similar surface patches with experimentally determined interface information, and weight them according to the similarity with the query surface patch. |
| Machine Learning | SPPIDER [93] | Structure | http://sppider.cchmc.org/ | SPPIDER uses the difference between predicted relative solvent accessibility and actual RSA of a residue as a feature to predict interface. |
| | PINUP [94] | Structure | http://sysbio.unl.edu/services/PINUP/ | PINUP uses a scoring function that is a linear combination of a side-chain energy, interface propensity, and residue conservation scores |
| | ProMate [95] | Structure | http://sysbio.unl.edu/services/PINUP/ | PINUP uses a scoring function that is a linear combination of a side-chain energy, interface propensity, and residue conservation scores |

| | | | | |
|---|---|---|---|---|
| | PIER [96] | Structure | http://abagyan.ucsd.edu/PIER/ | PIER predicts each surface patch as interfacial or not, using PLS (partial least squares) regression on the solvent accessibility values of 12 significantly over- and under-represented atomic groups at the interface |
| | Cons-PPISP [97] | Structure | http://pipe.scs.fsu.edu/ppisp.html | Cons-PPISP is a consensus neural network method for predicting protein–protein interaction sites. Features used include: position-specific scoring matrix, solvent accessibilities, and spatial neighbors of each residue |
| | Meta-PPISP [98] | Structure | http://pipe.scs.fsu.edu/meta-ppisp.html | Meta-PPISP is built on three individual web servers: cons-PPISP, PINUP, and ProMate. A linear regression method, using raw scores of the three severs as input, was trained on a set of 35 non-homologous proteins |
| | CPORT [99] | Structure | http://haddock.science.uu.nl/services/CPORT/ | CPORT is built on six individual web servers: WHISCY, PIER, ProMate, cons-PPISP, SPPIDER, and PINUP. The weights of a linear combination of the quantiles of the raw scores from the six servers were optimized on a set of complexes |
| | PAIRpred [100] | Sequence /Structure | http://combi.cs.colostate.edu/supplements/pairpred/ | PAIRpred uses multiple pairwise kernel SVMs to predict interacting residue pairs. Structural features used include: relative accessible surface area (rASA), residue depth, half sphere amino acid composition, protrusion index. Sequence features used include: PSSM and predicted rASA |
| | PpiPP [101] | Sequence | http://tardis.nibio.go.jp/netasa/ppipp/ | PpiPP trains 24 neural network predictors, and returns the average score of the 24 predictors as the final score. It uses a binary encoding of 20 types of amino acids plus PSSMs as features |
| | PSIVER [102] | Sequence | http://tardis.nibio.go.jp/PSIVER/ | PSIVER (Protein–protein interaction Sites prediction seVER) predicts protein–protein interaction sites using a PSSM and predicted accessibility as input for a Naive Bayes classifier |
| | WHISCY [103] | Structure | http://nmr.chem.uu.nl/Software/whiscy/ | WHISCY calculates a conservation score for each position of a MSA by summing up the scores in an adjusted Dayhoff matrix. It adjusts each conservation score using the interface propensity of the residue and |

| | | | | |
|---|---|---|---|---|
| | | | | smooth scores by considering surface neighbors to obtain the final prediction score |
| | Yan et al [104]. | Sequence | N/A | A two-stage classifier in which the first stage is a SVM interface predictor, and the second is a Naïve Bayes classifier trained on the predicted class labels from the SVM |
| | IntPred [105] | Structure | http://www.bioinf.org.uk /intpred/ | For a given PDB structure, IntPred uses sequence and structure information to create features that are the input to a random forest machine learning predictor, which will output a prediction label at either the surface patch- or residue-level. |
| Correlated Mutation [106] | i-Patch | MSA/ Structure | http://portal.stats. ox.ac.uk/userdata/ proteins/i-Patch/home.pl | The MSAs are concatenated based on knowledge about which pairs of proteins interact, and are used to calculate the correlated mutation scores for pairwise positions. A logistic model is trained on a combination of the propensities and the correlated mutation scores |

### 1.3.1 Homology-based methods

Homology-based method also called template-based method, based on the assumption that homologous proteins have significant similar sequence, structure and functional sites, homology-based approaches infer biological properties of a new protein from its homologs. This method has been broadly used in protein structure prediction [107], protein interaction partners prediction and protein function annotation [108, 109].

### 1.3.2 Machine learning-based methods

Even though homology-based methods are reliable, this kind of method also has some shortcomings because they require the experimentally determined interface residues of its homologs, this method is strictly limited when the homology template has no available

structure or unknown interfaces. In this case, machine learning method becomes an alternative approach for the interface prediction. The machine learning methods could be further categorized into structure-based or sequence-based methods, the former method required the input of 3D structure information, but the later method ask for only protein sequence. Structure method uses the protein structure to calculate the surface patch, based either on the residue-residue special distance, or based on a fixed number of the target residues, in which the surface patch consists of the target residue and its constant number of nearest surface residues. This type of methods have several obvious advantages over sequence-based methods. For example, it only needs to identify interfacial residues from the calculated surface patches instead of prediction for each protein residue. However, it also has some limitations, first of all, only few proportion of proteins exist structure, many functionally important proteins such as trans-membrane proteins are very hardly to be structurally conformed. Secondly, due to the conformational change existed before and after protein complex bound, the structural information extracted from unbound state may not exist in bound state anymore. Thirdly, structure-based method is very hardly to get the structure information of disordered proteins, especially for these disordered regions which only form structure after binding with their partners [18]. Thus, developing sequence-based methods is of great interest but predicting interfacial residues from sequence alone is very challenging and underdeveloped. The common input features for the sequence-based predictors includes physico-chemical properties such as hydrophobicity, residue charge and volume, or predicted structures such as solvent accessibility. It is also common to add the features of neighboring residues (sliding window) to represent the feature of the center residue.

Currently most of the structure-based predictors have higher prediction accuracy than sequence-based ones, this is because normally the interface residues are on the surface patches and structure methods can easily identify surface residues from interior residues. In addition, pair-wise interacting residues sometimes are spatially close in the 3D structure but far apart in the primary sequence of the protein. The spatial positions of residues are key for macromolecular recognition and the lack of the special information will weak the performance of sequence-based predictors.

### 1.3.3 Co-evolution methods

Co-evolution based statistical models, which operate under the assumption that interacting residues at the interface are likely to co-evolve and use a large multiple sequence alignment (MSA) to identify such residues [87, 88, 110]. Some correlated mutation prediction methods have been extended to predict intermolecular contacts between protein domains or between proteins, by concatenating two paired sequences [106, 111, 112]. However, the accuracy for prediction of intermolecular pairs is estimated to be 10 times lower than that for intramolecular pairs, suggesting that the signals of correlated mutations are weak, and it has been suggested that current methodologies of correlated mutation analysis are not suitable for intermolecular contact prediction [113].

### 1.3.4 Transmembrane protein interfaces prediction

Due to the experimental difficulty for transmembrane protein structure determination, many TMPs along with their homologous have no know 3D structure information. Hence,

it is not available to make the interface prediction using homology-based and the structure-based machine learning methods. Thus, a reliable sequence-based machine learning method becomes very necessary. In the past two decades, according to our knowledge, there was only one published paper in 2009 from Bordner [114] using random forest to predict the interface residues for alpha-helical and beta-barrel transmembrane proteins, this paper used only very basic features such as conservation, PSSM as input, and reached the prediction performance of AUC 0.75. Due to more TMP structures have been solved in recent years, more advanced multiple sequence alignment tools such as Hhblits [115] have been developed, more potentially useful prediction features could be calculated and used in the machine learning method. This method could be greatly improved and reach a high prediction accuracy.

### 1.4 Motivation and overview of this work

Helical integral membrane proteins are involved in diverse biological processes, with only a very limited number of available 3D protein structures and a high biological and medical importance, membrane proteins are an important research subject for structural bioinformaticians. Protein-protein interactions in membrane are involved in many cellular processes, the deficiency of oligomerization accounts for many diseases such as cancer and amyloidal diseases. Hence, here we will focus on PPIs in membrane milieu. Furthermore, reliable determining which specific amino acid residues form the interfaces between transmembrane protein-protein interactions is critical for understanding the structural and physicochemical determinants of protein recognition and binding affinity,

and has wide application in modeling and validating protein interactions predicted by high-throughput methods, in engineering proteins, and in prioritizing drug targets.

Within the following chapter 2 “Properties and prediction of homotypic transmembrane helix-helix interfaces”, results of experimental and computational study of self-interacting TMD interfaces will be presented. As mentioned in section 1.2.3, a variety of sequence motifs are known that promote TMD helix interaction within membrane proteins. Using the ToxR system, such sequence motifs for 10 new TMDs were identified by Yao Xiao from the group of Prof. Dieter Langosch (TUM). Subsequent properties of TMD interfacial residues were analyzed, we show that interfacial residues are statistically more conserved co-evolved, and polar than non-interfacial residues, and also more likely to be located in the hydrophobic core of the membrane. Using 60 features, I created The Transmembrane HOmodimer Interface Prediction Algorithm (THOIPA), which was particularly powerful for the prediction of the most important residues in the dimer.

Within chapter 3, entitled “Prediction of interfacial sites in alpha-helical membrane proteins”, it present MBPred (Membrane-protein Binding-sites Prediction), a sequence-based method for predicting interface residues in transmembrane proteins. MBPred utilizes a combination of four individual random forest models - MBPredTM, MBPredCyto, MBPredExtra, and MBPredAll – trained to predict residues involved in protein iterations in transmembrane, cytoplasmic, and extracellular segments as well as in the entire amino acid sequence, respectively.

Altogether, the main goal of the work presented in this dissertation is to better understand membrane protein-protein interactions, especially the interfacial residue which prompt the

interactions. We utilized the already known 3D structures of membrane proteins, get predictive features from the protein sequence and its homologs, finally build predictive machine learning models to correctly predict the interfacial residues of query membrane protein sequences.

# CHAPTER 2. PROPERTIES AND PREDICTION OF HOMOTYPIC TRANSMEMBRANE HELIX-HELIX INTERFACES

This chapter introduces the homotypic single-pass transmembrane protein helix-helix interfaces study, which is a cooperation work with Yao Xiao from the Biopolymer Chemistry department of TUM, I was responsible for the bioinformatic work such as data analysis and machine learning prediction software and web server development.

The transmembrane (TM) domains of single-pass membrane proteins often dimerize in the lipid environment. However only a few NMR dimer structures are available, and there is little quantitative information concerning the sequence properties of interfacial residues. An alternative method to profile TM homodimer interfaces is to use an E. coli TM Reporter Assay (ETRA) such as the well-known ToxR assay. In this study we also show that crystal structures contain self-interacting TM helices that have never been studied in detail. To quantitively analyze interface properties, we created a dataset of 54 self-interacting TM helices by combining the data from NMR, ETRA, and crystal studies. In a detailed analysis of the sequence properties, we show that interfacial residues are statistically more conserved and polar than non-interfacial residues, and also more likely to be located in the hydrophobic core of the membrane. GxxxG motifs were overrepresented at TM interfaces, particularly when investigated in a natural membrane environment. We found some evidence for the theory that interfacial residues are more coevolved. From this

information, we created the first machine learning algorithm for the prediction of TM homodimer interface residues. The Transmembrane HOmodimer Interface Prediction Algorithm (THOIPA) was particularly powerful for the prediction of the most important residues in the dimer. An objective listing of feature importance in THOIPA confirmed the predictive power of conservation, coevolution, polarity, and also the GxxxG motif. The THOIPA code and standalone predictor is available at https://github.com/bojigu/thoipapy/wiki. The THOIPA webserver is available at www.thoipa.com .

## 2.1 Introduction

Bitopic (single-pass) proteins make up to half of all integral membrane protein [116]. Their transmembrane (TM) helices are known to form strong, specific homodimers in cellular membranes [117, 118], with consequences for the functionalities of these proteins.

The ability to predict these PPIs would improve the knowledge of countless cellular processes, and pave the way for the design of therapeutic molecules. Due to technical difficulties associated with membrane proteins, the evolutionary profiles and structural forces of PPI interfaces in the membrane environment are poorly understood. Currently, the structures of only a handful of TM homodimers have been investigated by NMR spectroscopy and X-ray crystallography [119-121], and several of these belong to a single protein family, the receptor tyrosine kinases. Further interfaces have been identified in biological membranes using assays such as ToxR [122] and GALLEX [45], which we term E. coli TM reporter assay (ETRA) techniques. In combination with scanning mutagenesis,

these assays have exhaustively explored several additional TM helix-helix interfaces. Consequently, hundreds of potential TMD-TMD interfaces remain unexplored although there are many reports where limited mutagenesis provides sparse information on interface residues.

To close the gap between the numbers of well characterised TMD-TMD interfaces and the unknown ones, various methods have been devised previously to predict them from primary structure. These approaches rest on the known structural and evolutionary properties of TMD-TMD interfaces. These properties have been primarily derived from polytopic proteins the self-interaction of TMDs [123, 124].

There are currently four automated methods that identify TMD homodimer structures using energy functions: PREDDIMER [125, 126], EFDOCK-TM [56], TMDOCK [127] and CATM [128]. PREDDIMER and TMDOCK are easily accessible via online servers. EFDOCK-TM [56] relies on the output of the LIPS algorithm [129] and coevolution scores. LIPS was originally designed to predict lipid-facing residues in polytopic proteins and can identify a helix face with high conservation and polarity. EFDOCK-TM then identifies residue pairs via “evolutionary constraints”, as derived from sequence coevolution in the LIPS interface. Random combinations of evolutionary constraints are finally used to guide modelling via Rosetta. The PREDDIMER algorithm works by establishing the maximal complementarity of hydrophobic or hydrophilic surfaces of TMD homodimers. This is followed by geometry optimisation and structure refinement [125, 126]. The TMDOCK algorithm [127] threads a target amino acid sequence through several structural templates, followed by local energy minimisation. CATM is a specialised method that is only applicable to dimers driven by (small)xxx(small) motifs [128].

The general applicability of the current generation of predictors is limited by several key challenges. Firstly, there are only a few well-characterised homotypic TMD-TMD dimer structures by which the above algorithms have been validated. Secondly, and depending on the individual study, validation has been conducted using the $C_\alpha$ root mean square deviation (RMSD) for all [125-128] or subsets [56] of TMD residues, rather than reproducing residue-residue contacts. This limits their informative value. While the validation of soluble PPI site predictions has been standardised in the Critical Assessment of PRediction of Interactions (CAPRI) initiative [130, 131], there are no such guidelines for membrane proteins, nor have comparative assessments of predictive success been conducted. Thirdly, each of the above prediction algorithms generates an ensemble of possible dimer structures, which the user must interpret subjectively. Taken together, the automated prediction of a TMD-TMD interface structure remains a non-trivial task.

In this study, we create a comprehensive dataset of 54 self-interacting TM helices by combining data from ETRA, NMR and crystallography studies. We conducted a quantitative analysis comparing the sequence properties of interfacial and non-interfacial residues. We show that these PPI interfaces were associated with higher conservation, polarity, residue depth, and in some cases, higher coevolution. We also show the predictive power of motifs such as GxxxG. We then used such residue features to train the Transmembrane HOmodimer Interface Prediction Algorithm (THOIPA), a machine-learning-based method that compares favourably in its ability to predict TMD homodimer interfaces from primary structure.

## 2.2 Materials and Methods

### 2.2.1 NMR, crystal and complete datasets for THOIPA training and validation

An overview of the THOIPA datasets, machine learning features, and validation procedures is available in Figure 2-1. The “NMR” dataset was based on the 13 default dimer structures included in the validation by Wang et al. [132]. We updated the dataset to include the new NMR structures of the toll-like receptor 3 (PDB 2mk9, UniProt O15455, [133]), and high affinity nerve growth factor receptor (PDB 2n90, UniProt P04629, Nadezhdin et al. unpublished). Interface residues were defined based on the closest heavy-atom (non-hydrogen) distance between any atom in a residue pair. We first calculated the closest heavy-atom distance between the residue of interest and all other residues in any identical TMDs in the structure. Residue pairs with a heavy atom distance smaller than 3.5 Å were defined as interacting. The threshold of 3.5 Å was selected to ensure that the interface residues closely matched the interface provided by the authors of the published NMR studies [6, 9, 133-137]. All structures contained at least 4 interface residues. The NMR dataset contained 15 TMDs from 15 proteins, of which no two proteins shared more than 52% sequence identity, with a total of 115 interacting and 238 non-interacting residues.

The “crystal” dataset consisted of self-interacting TM helices extracted from crystal structures. The database “Non-redundant alpha” was downloaded from PDBTM [138], consisting of membrane proteins with annotated TM regions. Structures with a poor resolution (above 3.5 Å) were excluded. The dataset was made non-redundant by clustering full-length protein sequences with CD-HIT [139] using an amino acid sequence

identity cutoff of 40%. Interface residues for the crystal dataset were defined as described above for the NMR dataset. Self-interacting TMD helix pairs that had at least 4 unique interacting residues were retained. A second round of CD-HIT redundancy reduction was conducted based only on the TMD sequences. The final dataset was non-redundant at the 40% and 60% amino acid identity level for the full and TMD sequences, respectively. An exception in redundancy reduction was made for the dual topology fluoride ion channel (PDB 5nkq), where a single polytopic protein contributed two self-interacting TMDs with less than 20% identity to each other, TM1 and TM4.

Unlike the single-pass proteins within the NMR and ETRA datasets, most TM helices within the crystal dataset showed interaction with other helices within the multi-pass membrane protein. Such “folding contacts” are known to be conserved and polar [140, 141], however in this study they were classified as “non-interface” residues due to the lack of relevance to protein-protein interactions. To increase the accuracy of THOIPA towards TM homodimer interfaces within single-pass proteins, folding contacts were excluded from the training set. Folding contacts were still included in all validation procedures to allow a fair comparison between THOIPA and structure-based prediction algorithms. Folding contacts were defined based on heavy-atom distances to non-self TM helices using a 3.5 Å cut-off. The final crystal dataset contained 25 TMDs from 24 proteins, with a total of 167 interacting and 402 non-interacting residues, of which 55 were classified as folding contacts.

The “complete” dataset consisted of the combined ETRA, NMR and crystal structure datasets, with redundant proteins removed. It was used for training and validation of the THOIPA algorithm, and also to validate the accuracy of the LIPS, PREDDIMER and

TMDOCK algorithms. For redundancy reduction, we excluded TMDs based on an amino acid identity cut-off of 60% for TMDs and 40% for full protein sequences. For redundant TMDs with both ETRA and NMR data, the ETRA data was retained. This procedure resulted in removal of one protein (DDR2) from the ETRA dataset and seven proteins (EphA2, EGFR, ErbB2, ErbB3, and ErbB4, GpA and BNIP3) from the NMR dataset. The final “complete” dataset contains 54 proteins. All residue information is available in the Open Science Framework (OSF) data repository (osf.io/5cxpn/).

### 2.2.2 Key predictive features

Protein homologues were obtained via BLASTp against the NCBI non-redundant (nr) database. The BLAST query sequence consisted of the TMDs plus 20 adjacent residues on each side of the membrane. BLAST was conducted using the relatively permissive default settings to recruit the largest possible number of candidate homologues, including those distantly related to the query sequence. The TMD in the match sequence was identified based on the alignment to the query. The number of false positive hits was reduced by keeping only the alignments with fewer than 6 gaps and at least 20% sequence identity in the TMD region. Homologues with gaps in the query sequence were excluded. The remaining sequences were used to derive a multiple sequence alignment (MSA) from the original BLASTp pairwise alignments. For lipophilicity calculations that required information outside the immediate TMD region, the TMD plus five surrounding residues was extracted from the alignments, and the filtering procedure repeated as described above. Based on a careful manual verification of the resulting alignments we estimate that the number of false positives in our data does not exceed 2%. All alignments are available in the Open Science Framework (OSF) data repository (osf.io/5cxpn/).

A number of features were extracted from the MSA. These were used to examine interfacial properties, and also as input for the machine learning classifier THOIPA. Full details including formulae are available in 5.1.1 (Appendix A Supplementary Method). Briefly, conservation was calculated as $-S_{entropy}+3$, yielding positive values that increase with a decreasing rate of evolution. Polarity refers to the mean hydrophobicity of the residues at that position of the MSA, calculated using the Engelman (GES) hydrophobicity scale [142]. Positions with many polar/charged residues (Gln, Glu, etc) are therefore associated with high polarity scores. Residue depth refers to the relative position of the residue in the TMD, which ranged from 0 (first or last residue) to 1 (central residue). We calculated nine coevolution scores, using both direct interaction (DI) and mutual information (MI) scoring methods, yielding 18 coevolution features in total. Coevolution DI and MI scores were calculated from the MSA using FreeContact [143], an open source implementation of EVfold-mf DCA [144] and PSICOV [145].

### 2.2.3 Retrospective coevolution scores

Unlike the predictive coevolution scores listed above, retrospective coevolution scores required a previously defined homodimer interface, and could not be included as THOIPA features. We calculated retrospective coevolution scores for the interface residues of each TMD within the structural NMR and crystal datasets as previously described [132]. Briefly, for each TMD a list was created of all residue pairs that could be classified as “interacting”, based on closest heavy-atom distances below 3.5 Å. All other residue pairs were classified as “non-interacting.” In both cases, any residue pairs separated by more than 8 residues in the sequence were excluded (only <4% of the interacting pairs). For

each TMD, the mean DI coevolution score was then calculated for either the interacting, or the non-interacting residue pairs.

### 2.2.4 Machine learning and evaluation

Machine learning was conducted using extremely randomized trees [146], as implemented in python (sklearn ExtraTreesClassifier). ET is a tree ensemble method similar to the popular random forest algorithm. Like random forest, ET randomizes the selection of *K* features at each node, from which the best is chosen for the split. However, ET also randomizes the cut thresholds for each feature, leading to increased variance within each tree, but improved performance for the ensemble of trees [146]. For THOIPA training, the prediction was treated as a classification problem with two possible outcomes, “interface” or “non-interface”, as defined above for ETRA, NMR and crystal datasets. The number of trees was set to 100 and the number of features tested per node was set as 30% of the total number of input features (max_features = 0.3). The maximum tree depth was limited to 30 (max_depth = 30), and tree size was also limited by only splitting nodes with at least 3 samples (min_samples_leaf = 3). The quality of each split was judged using the entropy criterion. To improve the accuracy of measures of feature importance, each feature was only used once in the tree.

We evaluated THOIPA using the “leave-one-out” (LOO) cross-validation method, whereby the training dataset contained all TMDs except the one being tested. Validation is presented here using the classical receiver operating characteristic (ROC) curve. However during THOIPA development we focused primarily on performance in predicting the top ten residues important for the interaction. For this we developed a new evaluation

metric which we call the best-overlap curve (BO-curve). For the BO-curve it was necessary to rank the most important residues in the interaction from one to ten. This was done for both the experimental data, and the prediction results. For ETRA experimental data we used the disruption after mutation to rank residue importance. For interfaces derived from structures (either experimental, or predicted with PREDDIMER/TMDOCK) we used the closest heavy-atom distance. Although ranking based on heavy-atom distances is less precise than the standard rmsd approaches, it allowed us to directly compare evolutionary and structural prediction methods. These methods were also considered appropriate due to the moderate performance level of all predictors in this study. The full method and explanation is given in the 5.1.2. An area under the BO-curve for the top 10 residues in the interaction (AUBOC10) was calculated from the BO-curve as per the standard ROC AUC, and used to estimate overall performance against a dataset.

### 2.2.5 THOIPA implementation and prediction output availability

The web server implementing the machine learning for interface prediction will be publicly available at http://www.thoipa.com/. The full source code and a standalone version of THOIPA is available online (https://github.com/bojigu/thoipapy). As input THOIPA requires only a full-length protein sequence, as well as the sequence of the TM region of interest. The THOIPA output is a numerical value describing the potential of that residue to be located at a homodimer interface. The THOIPA interface predictions for the complete datasets are show in in heatmaps (Figure 5-1).

### 2.2.6 Comparison with other methods

We evaluated THOIPA, LIPS [141], PREDDIMER[125, 147] and TMDOCK [127] as predictors of TM homodimer interfaces. For LIPS we modified the output slightly. The original output identifies whether or not the residue of interest belongs to the helix face with the highest conservation and polarity. For each residue, this gives a value of 0 or 1. Such binary values are poorly amenable to validation. As described in the supplementary methods, we therefore modified these values slightly, so the predicted interface and non-interface residues were further ranked by their individual conservation and polarity. We submitted each TMD sequence to the online servers of PREDDIMER and TMDOCK. PREDDIMER required TMDs to be at least 20 residues. For TMDs shorter than 20 residues, we therefore extended the sequence by one residue at each end (starting at the C-term) until the length reached 20. The top structure in the output PDB file was used for validation. Similarly, the best structure for TMDOCK was obtained after submission of the TMD sequence to the online server. TMDOCK automatically truncated many of the TMDs. Therefore for each predictor, each TMD was validated separately after excluding any residues for which there was no interface data (PREDDIMER) or predicted structure (TMDOCK). The mean TMD length was 22.8 residues for THOIPA, LIPS and PREDDIMER, and 20.7 residues for TMDOCK.

### 2.2.7 Statistical significance.

Pairwise comparisons were conducted using an independent Student's *t*-test assuming equal variance, using bootstrapped data where indicated. *P*-values were represented as follows: *, $p<0.05$. **, $p<0.01$, ***, $p<0.001$.

## 2.3 Results

The aims of this study are laid out in Figure 2-1. First, we assembled a set of 54 well-characterised interfaces from a broad range of self-interacting TM helices (Table 2-1). The full homotypic TMD dataset comprises 21 TMDs investigated by ETRA techniques, 8 TMDs investigated by NMR and 25 TMDs investigated by X-ray crystallography. Second, a quantitative analysis of interface residue properties was conducted. Third, we developed THOIPA and compared its performance to TMDOCK and PREDDIMER.

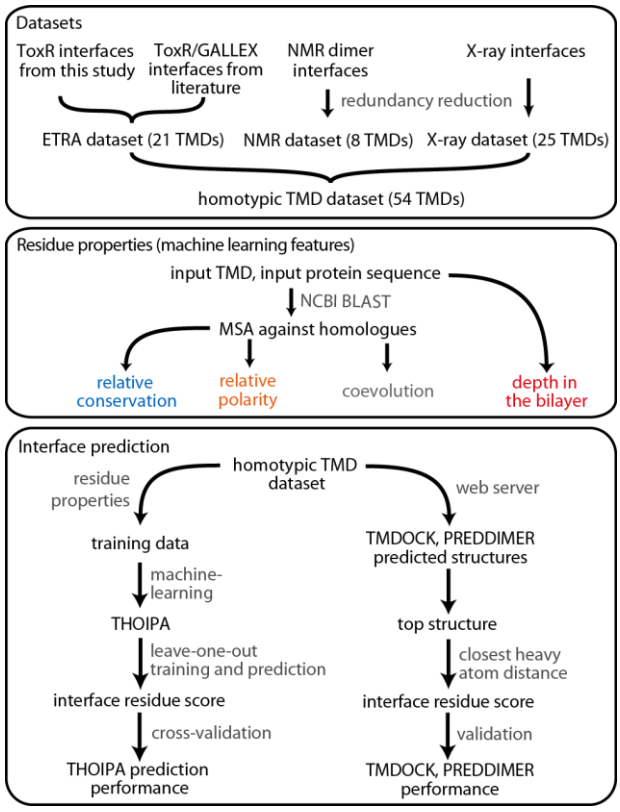



**Figure 2-1: Overview of datasets, residue properties, sequence analysis, machine learning and predictor validation conducted in this study.**

## Table 2-1: Interface residues of the homotypic TMD dataset.

| # | protein (acc[a]) (ref) | TMD sequence[b] |
|---|---|---|
| **ETRA TMDs** | | |
| **1** | **Ire1 (O75460)** | **ATIILSTFLLIGWVAFIITY** |
| **2** | **ATP1B1 (P05026) [148]** | **LLFYVIFYGCLAGIFIGTIQVMLLTI** |
| **3** | **PTPRG (P23470) [149]** | **IIPLIVVSALTFVCLILLIAVLV** |
| **4** | **Tie1 (P35590) [150]** | **LILAVVGSVSATCLTILAALLTLV** |
| **5** | **DDR1 (Q08345) [150]** | **ILIGCLVAIILLLLLIIALML** |
| **6** | **PTPRO (Q16827) [149]** | **VVVISVLAILSTLLIGLLLVTLIIL** |
| **7** | **Armcx6 (Q7L4S7) [151]** | **REVGWMAAGLMIGAGACYCV** |
| **8** | **PTPRU (Q92729) [149]** | **LILGICAGGLAVLILLLGAIIVII** |
| **9** | **Siglec7 (Q9Y286) [151]** | **VLLGAVGGAGATALVFLSFC** |
| 10 | GpA (P02724) [152] | LIIFGVMAGVIGTIL |
| 11 | ErbB2 (P04626) [152, 153] | LTSIISAVVGILLVVVLGVVFGIL |
| 12 | ITGB3 (P05106) [154] | VLLSVMGAILLIGLAALLI |
| 13 | ITGA2B (P08514) [155] | WVLVGVLGGLLLLTILVLAMW |
| 14 | FtsB (P0A6S5) [156] | TLLLLAILVWLQYSLWF |
| 15 | GP1BB (P13224) [157] | GALAAQLALLGLGLLHALLL |
| 16 | MPZ (P25189) [158] | YGVVLGAVIGGVLGVVLLLLLLFYVV |
| 17 | PTPRJ (Q12913) [149] | ICGAVFGCIFGALVIVTVGG |
| 18 | BNIP3 (Q12983) [118] | LLSHLLAIGLGIYIG |
| 19 | QSOX2 (Q6ZRP7) [159] | CVVLYVASSLFLMVMY |
| 20 | ADCK3 (Q8NI60) [160] | LANFGGLAVGLGFGALA |
| 21 | NS4A (Q99IB8) [161] | TWVLAGGVLAAVAAYCLAT |
| **NMR TMDs** | | |
| 22 | TLR3 (O15455, 2mk9) | FFMINTSILLIFIFIVLL |
| 23 | TYROBP (O43914, 2l34) | LAGIVMGDLVLTVLIALAVYFL |
| 24 | NTRK1 (P04629, 2n90) | LAVFACLFLSTLLLVL |
| 25 | APP (P05067, 2loh) | AIIGLMVGGVVIATVIVITLVML |
| 26 | PDGFRB (P09619, 2l6w) | VVVISAILALVVLTIISLIILIMLW |
| 27 | CD3ζζ (P20963, 2hac) | LCYLLDGILFIYGVILTALFL |
| 28 | EphA1 (P21709, 2k1k) | IVAVIFGLLLGAALLLGILVF |
| 29 | FGFR3 (P22607, 2lzl) | VYAGILSYGVGFFLFILVVAAVTLC |
| **X-ray TMDs** | | |
| 30 | KvAP (Q9YDF8, 1orqC4) | GKVIGIAVMLTGISALTLLIGTVSNMFQ |
| 31 | BR (Q8YSC4, 1xioA4) | GFLMSTQIVVITSGLIADL |
| 32 | PSII-M (Q8DHA7, 2axtM1)[d] | ATALFVLVPSVFLIILYV |
| 33 | Mgst1 (P08011, 2h8aA2) | HLNDLENIVPFLGIGLLYSL |
| 34 | Wza (Q9X4B7, 2j58A1)[d] | SQLVPTISGVHDMTETVRYI |
| 35 | OLI1 (P61829, 2wpdJ1) | AAKYIGAGISTIGLLGAGIGIA |
| 36 | MGST1 (O14684, 3dwwA2) | CLRAHRNDMETIYPFLFLGFVYS |
| 37 | p2X (Q6NYR1, 3h9vA2) | KFNIIPTLLNIGAGLALLGLVNVICDWIV |
| 38 | GluCl α (G5EBR3, 3rifA2) | IPARVTLGVTTLLTMTAQSAGIN |
| 39 | KCNJ12 (F1NHE9, 3spcA2) | PLAVFMVVVQSIVGCIIDSFMIGAIMAKM |
| 40 | fn ATPase F0 c-ring (Q8RGD7, 3zk1A1) | LGCSAVGAGLAMIAGLGPGIGEG |
| 41 | CorA (Q58439, 4ev6A1) | TMVTTIFAVPMWITGIYGMNF |
| 42 | CRACM1 (Q9U6B8, 4hksA1) | SWTSALLSGFAMVAMVE |
| 43 | CorA (Q9WZ31, 4i0uA1) | TIIATIFMPLTFIAGIYGMNF |
| 44 | NAD(P) transhydrogenase α2 (Q72GR9, 4o9pC1) | WSALYIFVLTAFLGYEL |
| 45 | AbgT (Q0VR69, 4r0cA7) | ITAMEVTMASMAGYLVLMFFAAQFVAWF |
| 46 | TspO (Q81BL7, 4ryiA2) | PGMTIGMIWAVLFGLIALSVA |
| 47 | mp ATPase F0 c-ring (A0A2S9G8T0, 4v1fA1) | GGLIMGGGAIGAGIGDGIAGNALI |
| 48 | TMEM16 (C7Z7K1, 4wisA10) | LKAWGLLLSILFAEHFYLVVQLAVR |
| 49 | Trpv1 (O35433, 5irzD6) | KAVFIILLLAYVILTYILLLNMLIALM |
| 50 | CRCB TM1 (Q7VYU0, 5nkqA1) | FIAIGIGATLGAWLRWVLG |
| 51 | CRCB TM3 (Q7VYU0, 5nkqA3) | AAVTGFLGGLTTFSTFSAETV |
| 52 | PC2 (Q13563, 5t4dA6) | RVLGPIYFTTFVFFMFFILLNMFLAIIN |
| 53 | BCNG-1 (O60741, 5u6oA6) | ITMLSMIVGATCYAMFVGHATALI |
| 54 | NadC (Q9KNE0, 5uldA9) | WKEIQKTADWGILLLFGGGLCL |

[a] Accession number (acc) from the UniProt database. The X-ray identification code (e.g. 1orqC4) consists of the PDB accession (e.g. 1orq), the protein chain (e.g. C), and the TMD number in the protein (e.g. 4).[b] Homotypic interface residues in the TMD sequences are underlined. [c] Bold text indicates new interfaces identified in the current study. [d] Two TMDs in the X-ray dataset derived from bitopic proteins.

### 2.3.1 Creation of a non-redundant dataset of TM homodimer interfaces

We combined the self-interacting helices of the ETRA dataset with those from NMR and crystal studies, resulting in the “complete” dataset of 54 self-interacting helices. The complete dataset was non-redundant at the 40% and 60% amino acid identity level for the full and TMD sequences, respectively. This allowed us to quantitatively and objectively analyze interface properties, which until now have been using case studies, artificial selection, or the small, highly-redundant NMR dataset. Our original NMR dataset consisted of 15 homodimer structures in total, comprising the 13 default structures used by Wang and Barth (2015) [162], and two more dimers structures that have been recently submitted (Table 5-1). Only eight of the NMR TMDs were added to the complete dataset. This was firstly due to high internal redundancy, as the NMR dataset contained six RTKs, and secondly to redundancy via existing TMDs in the ETRA dataset, specifically GpA, BNIP3, and ErbB2.

The novel crystal database consisted of 25 parallel, self-interacting TM helices within crystal structures of membrane proteins. The non-covalently associating, “homodimer-like” helices were identified from crystal structures annotated according to PDBTM [163]. As described in the methods, we only included TM helices with at least four interface residues, as defined by a 3.5 Å cut-off in the closest heavy-atom distance between residues. The vast majority of these (23/25) were helices from oligomeric multi-pass proteins. For example in the structure of TRPV1 (PDB 5irz), one of the six TM helices formed close “crystal contacts” to itself on the opposing protein. Of the two single-pass

TM proteins in the crystal dataset, psbM is an important component of the PSII dimer interface (PDB 2axt), and Wza forms a homomeric octamer (PDB 2j58). The helices of the crystal dataset had an average absolute (i.e. positive) crossing angle of 38.91°, slightly higher than the more parallel helices of the NMR dataset (28.65°).

For both the NMR and crystal datasets, interface residues were defined based on a 3.5 Å cut-off in closest heavy-atom distances, as defined in the methods. The proportion of interacting residues was similar between the ETRA (27%), NMR (33%) and crystal (28%) datasets. After redundancy reduction, the final complete dataset comprised 54 TMDs, the subsets of which contained 21 ETRA, 8 NMR, and 25 crystal TMDs. The complete dataset contained 352 interface residues and 780 non-interface residues (9.6 % folding contacts from crystal TMDs were removed). This corresponded to an average of 6.5 interface residues per TMD, comprising 31% of the TM residues. These objectively identified interface residues contained 19 of the 20 natural amino acid residue types, with the only exception being Lys. An undesirable feature of the “non-interface” residues within the crystal dataset was the presence of residues participating in heterotypic helix-helix interactions, usually the folding of multi-pass membrane protein. These “folding contacts” were identified based on a cutoff of 3.5 Å heavy atom distance, and comprised 55 (13.7%) of the non-interface residues within the crystal dataset. Folding contacts were removed from analyses of interface properties. All sequences, interface residues, and folding contacts are detailed in Figure 5-1 and in the Open Science Framework (OSF) data repository (osf.io/5cxpn/).

### 2.3.2 Determination of residue properties (THOIPA predictive features)

After creating the first large database of self-interacting TMDs, we began examining which residue properties could be used to distinguish interface from non-interface residues. We measured a large number of residue properties, which were also used as inputs (i.e. predictive features) for our machine learning predictor, THOIPA. Most of these properties were derived from multiple sequence alignments (MSA) against homologues, as outlined in Figure 2-1. A detailed description of all 60 features included in THOIPA is available in 5.1.1. Four of these features are examined here in detail (conservation, coevolution, relative polarity and relative depth), and are calculated as described in the methods. Briefly, conservation is a normalized form of entropy, with higher values indicating highly conserved residue positions. For the relative polarity, we first calculated the mean polarity of the residues at that position in the MSA. The relative polarity was the polarity divided by the mean polarity of the six surrounding residues. An Arg residue in the center of the TMD therefore scores much more highly than an Arg residue in the juxta membrane region. The relative depth is a simple measure of the position in TMD sequence, from the most central (value=1) to the most peripheral TMD residue (value=0).

Coevolution values are more complex. The output from EVfold includes pairwise scores between all possible residue pairs in the TMD[144]. Furthermore, the output includes a mutual information (MI) and a direct interaction (DI) value for each residue pair. The mutual information is a standard measure of coevolution between two residues, but is known to suffer a number of biases[145]. For example high scores can be seen for indirect contacts, such as when residues B and C are not in contact, but have a high MI score because they are both contacting and coevolving with residue A. A second bias in MI is the low score associated with high conservation. The EVfold algorithm calculates the DI

score by applying a statistical framework to the MI scores. The DI typically gives a better prediction of contacting residues. For prediction in THOIPA, and to understand interface properties, it was necessary to convert the pairwise coevolution scores to a single representative value at each residue position. For position i in the TMD, for example, we typically started with the MI and DI scores between i and all other residues in the TMD. For a 21 residue TMD, this comprises 20 residue pairs. One simple measure is simply the maximum DI value between all these residue pairs, which we refer to here as DImax. We also calculated this for the MI, yielding MImax. In total we calculated 16 different coevolution variants (5.1.1), in all cases for both MI and DI scores. These included mean values of selected residue pairs, maximum values of selected residue pairs, and finally whether or not the residue participated in a helical face with high overall coevolution. For all figures where a single representative coevolution metric was required, we used DImax.

### 2.3.3 Interface residues are conserved, coevolved, polar, and central in the TMD

The evolutionary conservation of residues was calculated from multiple sequence alignments (MSAs) against homologues. For interfacial residues, the average conservation is significantly ($p = 2.14*10^{-6}$) higher than that of their non-interface counterparts (Figure 2-2A). Thus, the interfaces are less likely to change during evolution than the remainder of a TMD. This contrasts with PPI interfaces in soluble proteins, where a higher conservation is only seen in selected conditions [164], and in some studies has been disputed entirely [165]. The interface residues are also distinguished by polarity. The difference between interface and non-interface residues is higher for “relative polarity”, i.e. polarity relative to the surrounding six residues (Figure 2-2B; p=0.0018), than for absolute polarity (p=0.029).

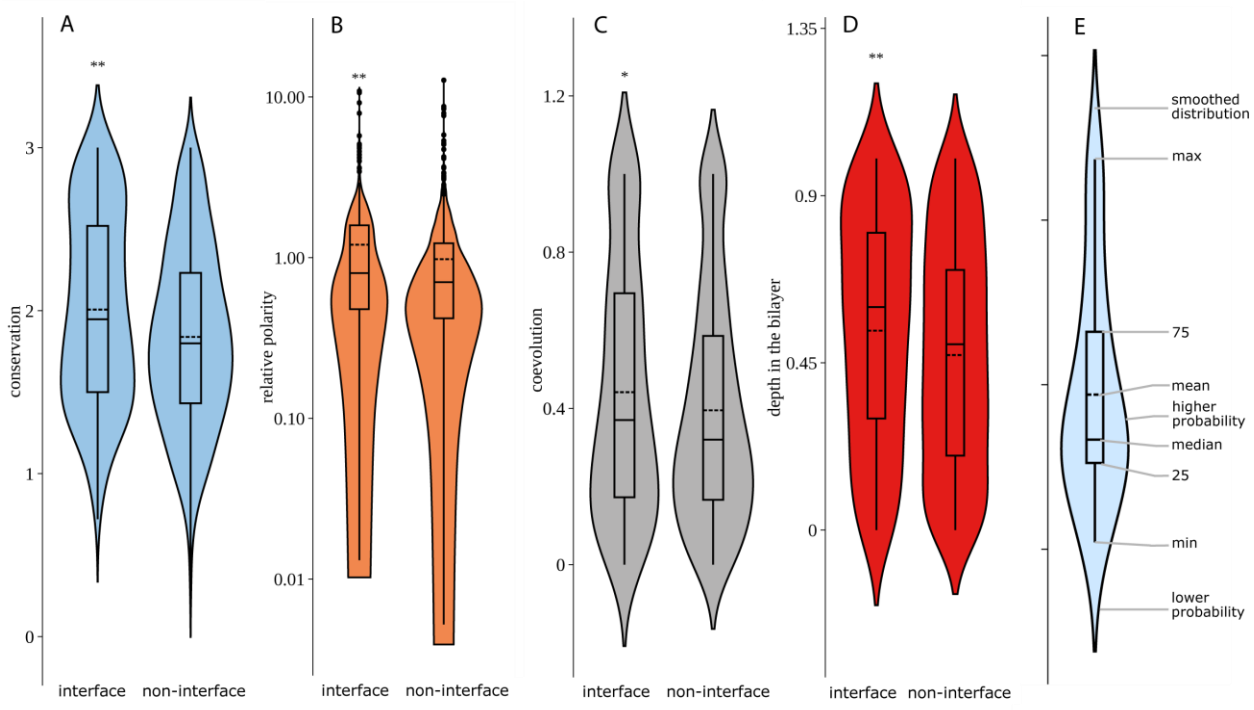



**Figure 2-2: Interface residues have a higher conservation, coevolution, relative polarity and relative depth in comparison to non-interface residues.** The violin plots display the means (dotted line) and medians (straight line) as well as a smoothed distribution of the data. (A) Conservation. (B) Relative polarity. (C) Coevolution (DImax; see 5.1.1). (D) Depth in the bilayer. (E) Components of the violin plot. Statistical significance was measured using a bootstrapped t-test (*, $p<0.05$. **, $p<0.01$).

Overall, the importance of conservation and polarity shown here is consistent with the known importance of these factors for polytopic membrane protein folding [129, 140, 166-168]. Interfacial residues also tend to be located deeply in the membrane (Figure 2-2D; $p = 7.94*10^{-4}$).

Sequence coevolution has been utilised previously in the prediction of polytopic membrane protein structures [169, 170] and of contacting residues in homotypic TMD interfaces [56]. Here, we tested 16 measures of coevolution (5.1.1) that do not require *a priori* knowledge of the interface and are thus termed “predictive” measures. Briefly,

pairwise mutual information (MI) and direct information (DI) scores were calculated from MSAs using EVfold [169, 171]. The 16 coevolution measures used here comprise the means or maxima of different pairwise coevolution values. Predictive coevolution measures are specific to individual positions in a sequence, and can therefore be used in machine learning approaches that take residue properties of single residues as input [172]. The simplest of these measures, DImax, is used as an example in the respective figures. The DImax is the maximum coevolution value between the residue of interest and all other residues in a TMD. DImax tends to be higher at interfaces (Figure 2-2D, $p = 0.015$).

From the 16 coevolution measures, ten differed significantly between interface and non-interface residues (DImax, DI4cum, DItop4mean, DItop8mean, MI3mean, MI4cum, MI4max, MI4mean, MItop4mean, MItop8mean, bootstrapped *t*-test, $p<0.05$). Of these ten, the DI values are higher at the interface, while most of the MI values are lower at the interface (Table 2-2). This reflects the fact that the MI values are artificially low at positions of high conservation (Figure 2-3). In contrast to DI values, MI values also decrease with the number of homologues (Figure 5-3). The predictive power of the different coevolution measures is only partially additive, as each of the above ten coevolution measures is correlated with at least one other coevolution measure ($R^2 > 0.5$).

**Table 2-2: Residue properties that differ between interface and non-interface residues**

| feature | higher for interface residues | p-value (t-test) | correlated features ($R^2 > 0.3$) |
|---|---|---|---|
| GxxxG | True | 6.38E-11 | SmxxxSm, G |
| conservation | True | 2.14E-06 | cons4mean |
| branched | False | 4.25E-05 | I, V, LIV |
| V | False | 7.99E-05 | LIV, branched |
| G | True | 0.000127 | GxxxG, SmxxxSm |

| | | | |
|---|---|---|---|
| cons4mean | True | 0.000271 | conservation |
| residue_depth | True | 0.000794 | |
| MI4max | False | 0.000823 | MImax, MItop4mean, MItop8mean, MI1mean, MI3mean, MI4mean |
| MI3mean | False | 0.001051 | MImax, MItop4mean, MItop8mean, MI4max, MI1mean, MI4mean |
| H | True | 0.001542 | |
| SmxxxSm | True | 0.001822 | GxxxG, A, G |
| relative_polarity | True | 0.001973 | polarity, polarity4mean, D, Q, DE, KR, QN |
| MItop8mean | False | 0.002308 | MImax, MItop4mean, MI4max, MI1mean, MI3mean, MI4mean |
| MItop4mean | False | 0.00375 | MImax, MItop8mean, MI4max, MI1mean, MI3mean, MI4mean |
| E | True | 0.005398 | polarity, polarity4mean, DE |
| MI4mean | False | 0.00542 | MImax, MItop4mean, MItop8mean, MI4max, MI1mean, MI3mean |
| LIV | False | 0.008051 | I, L, V, branched |
| DItop8mean | True | 0.010235 | DImax, DItop4mean, DI4max, DI4cum, MI4cum |
| DImax | True | 0.014904 | DItop4mean, DItop8mean, DI4max, DI4cum, MI4cum |
| DItop4mean | True | 0.020809 | DImax, DItop8mean, DI4max, DI4cum, MI4cum |
| polarity | True | 0.029957 | relative_polarity, polarity4mean, polarity1mean, D, E, K, N, Q, R, DE, KR, QN |
| MI4cum | True | 0.040042 | DImax, DItop4mean, DItop8mean, DI4max, DI4cum |
| DI4cum | True | 0.040042 | DImax, DItop4mean, DItop8mean, DI4max, MI4cum |
| QN | True | 0.040126 | polarity, relative_polarity, polarity4mean, polarity1mean, N, Q |
| I | False | 0.047567 | LIV, branched |

A previous study compared DI values of pairs of known interface residues and pairs of non-interface residues [56] (see: Figure 5-4). Since this approach requires *a priori* knowledge of the interface, we term it here a “retrospective” coevolution analysis which cannot be used for prediction. In a detailed analysis of retrospective coevolution (legend to Figure 5-4), we confirm that pairwise coevolution scores are higher between interface residues than between non-interface residues [56]. However, we also show that retrospective methods are biased by the non-random distribution of interface residues. Simply put, homotypic interfacial residues are often neighbours (Figure 5-4) and neighbouring residues have high coevolution scores [169, 171, 173]. In contrast, this bias is absent from predictive coevolution measures. Therefore, the higher predictive DI measures at interfaces (Figure 2-2C, Table 2-2) provide the first strong evidence of

enhanced coevolution between homotypic TMD interface residues. Our results suggest that the difference in coevolution between interfacial and non-interfacial residues exists, but is much more subtle than previously implied [56].

Separate analyses of the sub-datasets confirmed the general trends given in Figure 2-2 (Figure 5-2, Figure 5-5). The preferential coevolution of interface residues is strongest for TMDs of the X-ray dataset. This dataset has the highest number of homologues (Figure 5-5), which improves the accuracy of DI values [169, 171, 174].

A different way of presenting the data shown in Figure 2-2 is to calculate the percentages of TMDs where the mean value of a given property is higher for interface *vs*. non-interface residues. This method minimises the effects of the varying TMD lengths, overall conservation, and overall polarity. Accordingly, 62% to 70% of TMDs in the homotypic TMD dataset share higher interface conservation, coevolution, relative polarity and depth in the membrane (Figure 5-5A). The situation is similar when the sub-datasets were analysed separately (Figure 5-5B).

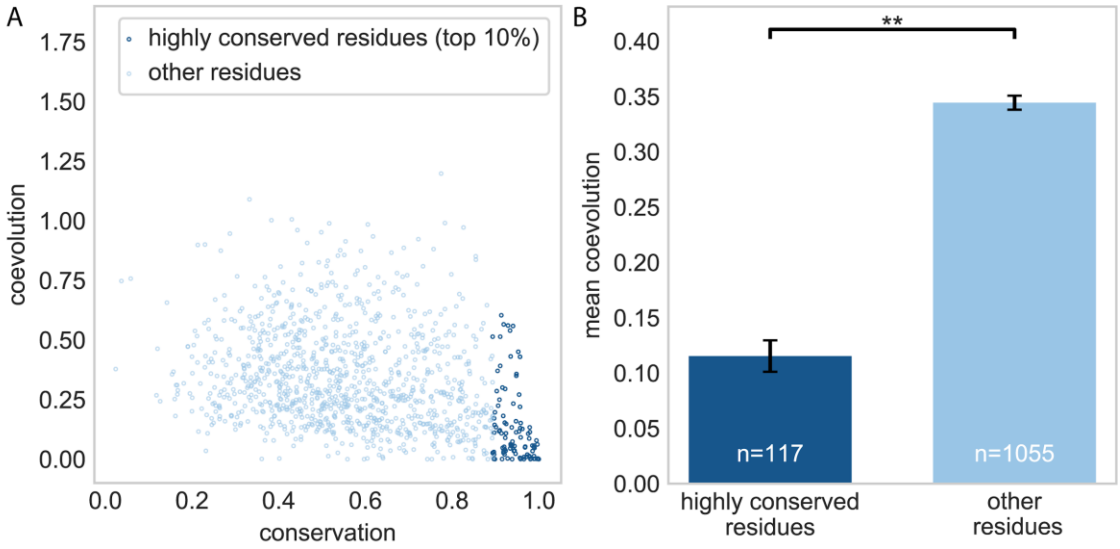



**Figure 2-3: Highly conserved residues have low mutual information (MI) coevolution scores.** For each residue in the complete dataset, conservation was plotted against coevolution. The coevolution score of coev_i4_MI is shown, which represents the mean coevolution of two residue pairs, i to i+4, and i to i-4. Folding contacts from crystal TMDs were removed from this analysis. (A) Scatterplot. Highly conserved residues (top 10%) are shown in a darker colour. (B) Barchart comparing the MI coevolution score of the most highly conserved residues (top 10%) with all other residues. In comparison to MI scores, the coevolution DI scores were less correlated to conservation, and tended to be high for highly conserved residues (data not shown). For data produced in this study, the mean ± SEM was shown.

### 2.3.4 Creation of an algorithm to predict the homodimer interfaces of TM helices

We developed THOIPA as the first machine-learning predictor of TM homodimer interfaces. Briefly, THOIPA uses extremely randomized trees [146], an ensemble technique similar to random forest [175]. It was trained as a classifier to predict the “interface” or “non-interface” definition of each residue in the complete dataset. A total of 60 input features were included for each residue, comprising various properties such as conservation, polarity, coevolution, and residue depth. As input, THOIPA requires only the sequences of the TMD and full protein. The output score for each residue represents

the probability that it lies at a homodimer interface. THOIPA is primarily designed as a tool to aid experimental or modelling approaches. Our aim was therefore to predict the most important residues for the dimer, rather than to accurately judge the contribution of all amino acids in the sequence. To this end, we optimized THOIPA using an in-house developed validation method, the BO-curve, which judges performance in predicting the top one to ten residues for the interaction (see 5.1.2). The importance of the residue in interaction was calculated using the disruption after mutation (ETRA), or the closest heavy-atom distances (NMR/crystal).

Here we show validation results with the BO-curve, but also with the more classical receiver operating characteristic (ROC)-curve methodology. We validated THOIPA against LIPS, PREDDIMER [125, 135, 147] and TMDOCK. Our validation could not include EFDOCK-TM [162], as the server was unavailable. We could not include another prediction algorithm, CATM [176], as this is only valid for interfaces based on (small)xxx(small) motifs. The energy minimization techniques PREDDIMER and TMDOCK each generate a number of feasible structures. In this case we simulated a de-novo prediction by assessing the accuracy of the top output structure according to the relevant prediction algorithm. As described above for experimental structures, the importance of each residue to the TMDOCK and PREDDIMER structure was calculated based on closest heavy-atom distances. Further details are available in the methods and THOIPA code. Heatmaps showing residue distances, prediction results, and important features including conservation are available for each TMD (Figure 5-1). A standalone version of THOIPA is available (https://github.com/bojigu/thoipapy/wiki).

THOIPA is designed to aid experimental and modelling approaches by identifying a small number of key interface residues in a TMD. For this purpose, THOIPA is vastly superior to PREDDIMER and TMDOCK (Figure 2-4). In the precision-recall plot, THOIPA showed excellent precision when considering the top residues (left-hand side of Figure 2-4*A*, and Figure 2-4*B*). This is a consistent feature of THOIPA, regardless of the experimental source of the TMD interface ( Figure 5-7). According to the best overlap (BO) validation developed in this study, the top one to five residues ranked by THOIPA are much more likely to comprise true positives than those of PREDDIMER and TMDOCK (Figure 2-4*C,D*). Performance peaked when the top two residues according to THOIPA were considered. Of the two most important residues in each of the 54 TMDs (108 residues in total), 34.2% were among the top two residues predicted by THOIPA. This compares favourably to a random prediction (9.4%). In the BO-validation curve, this yields a performance above random of 0.248 (i.e. 34.2% - 9.4%). Interestingly, THOIPA was also the best algorithm in the “fraction correctly predicted” analysis. This method was developed to assess the results of CAPRI competitions in terms of interface residue prediction [130], and represents a balanced assessment of overall predictive power towards all TMDs in a dataset. THOIPA performance was superior at nearly all precision-recall cutoff values (Figure 2-4*E,F*), followed closely by TMDOCK. In the CAPRI study, a precision-recall cutoff of 0.5 demarked a successfully predicted interface. At this cutoff, THOIPA correctly predicted a fraction of 0.43 of all interfaces in our dataset. By comparison, the best automated predictor of soluble interfaces, HADDOCK, had correctly previously predicted a fraction of only 0.38 of 20 CAPRI targets [130]. Thus, the performance of THOIPA is comparable to that of the best automated predictors of PPI in

soluble proteins [130, 131]. It has been previously reported that LIPS could successfully identify the NMR homodimer interfaces [56]. However, for the larger dataset presented here, the performance of LIPS is much lower than that of THOIPA (Figure 5-8, LIPS MCC=0.06, THOIPA maximum MCC=0.23). At the level of individual TMDs, THOIPA performance was best for ErbB2, BCNG-1 TM6, and Siglec7 interfaces (Figure 2-5), corresponding to a variety of interface types.

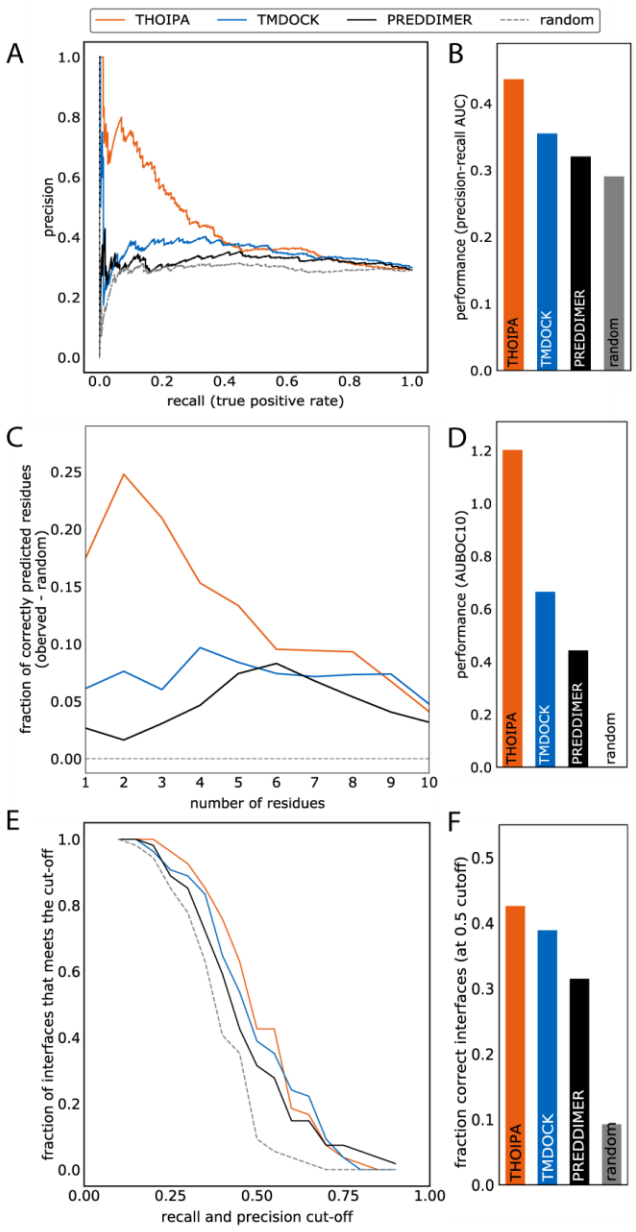


**Figure 2-4: THOIPA performance validation.** (A) Precision-recall curve. The closer the curve is towards the top right corner, the higher the overall performance. (B) Barchart of the area under the precision-recall curve. (C) Performance according to best overlap (BO) validation, a method developed here to report the number of residues at which peak performance is obtained (see *5.1.2*). (D) Area under the BO-curve for 1 to 10 examined residues (AUBOC10). (E) Fractions of correctly predicted interfaces at different accuracy levels. This is a validation method used previously in CAPRI [130]. (F) Fraction of TMDs with interfaces that were correctly predicted. An interface was defined as being correctly predicted if precision and recall both exceeded 0.5 [130].

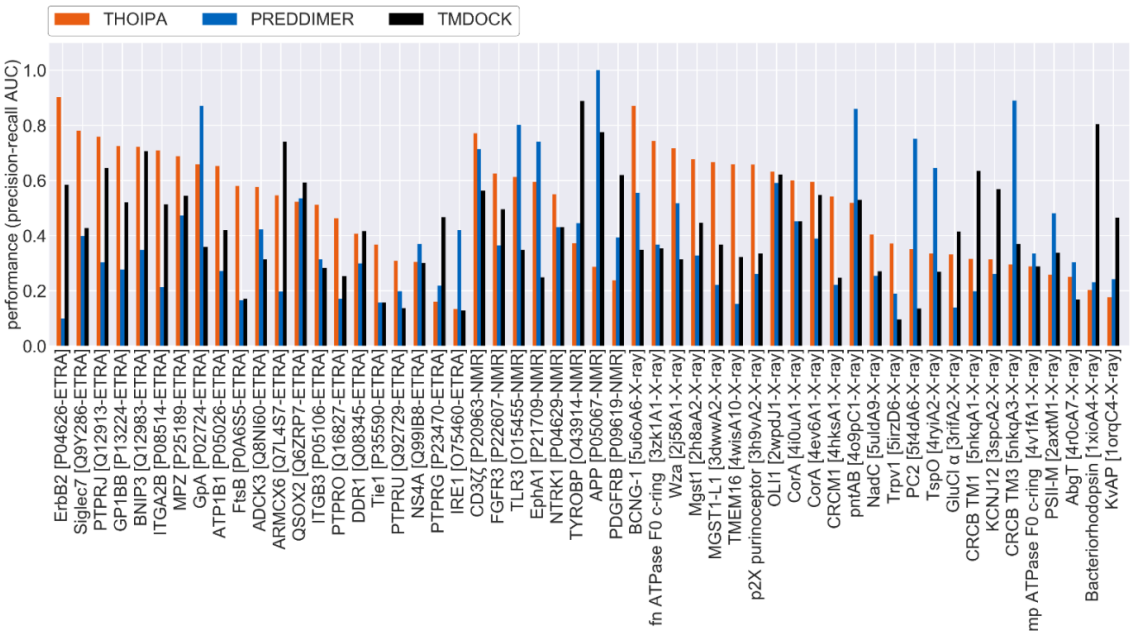



**Figure 2-5: Precision of predictors towards individual TMDs.** The precision-recall area under the curve (PR AUC) is shown. TMDs were ordered according to THOIPA performance within each dataset. For TMDs in the ETRA and NMR datasets, the respective UniProt accession

is shown. For TMDs in the X-ray dataset the reference number (e.g. 2h8aA2) is a concatemer of the PDB accession (e.g. 2h8a), the chain (e.g. A), and the TMD number (e.g. 2).

Ensemble machine-learning classifiers such as THOIPA can objectively rank input features according to their importance within the decision trees. The top input features according to THOIPA exhibit the following rank order: 1) participation of the residue in a GxxxG motif, 2) residue conservation, 3) residue depth in the bilayer, 4) the number of TMDs in the alignment, and 5-12) several measures of sequence coevolution (Figure 2-6). This corresponds well with the analysis of interface properties described above, and the current understanding of factors that are important to TMD-TMD interactions. Accordingly, most of the features important to THOIPA differed significantly between interface and non-interface residues in a Student's *t*-test (Table 2-2).

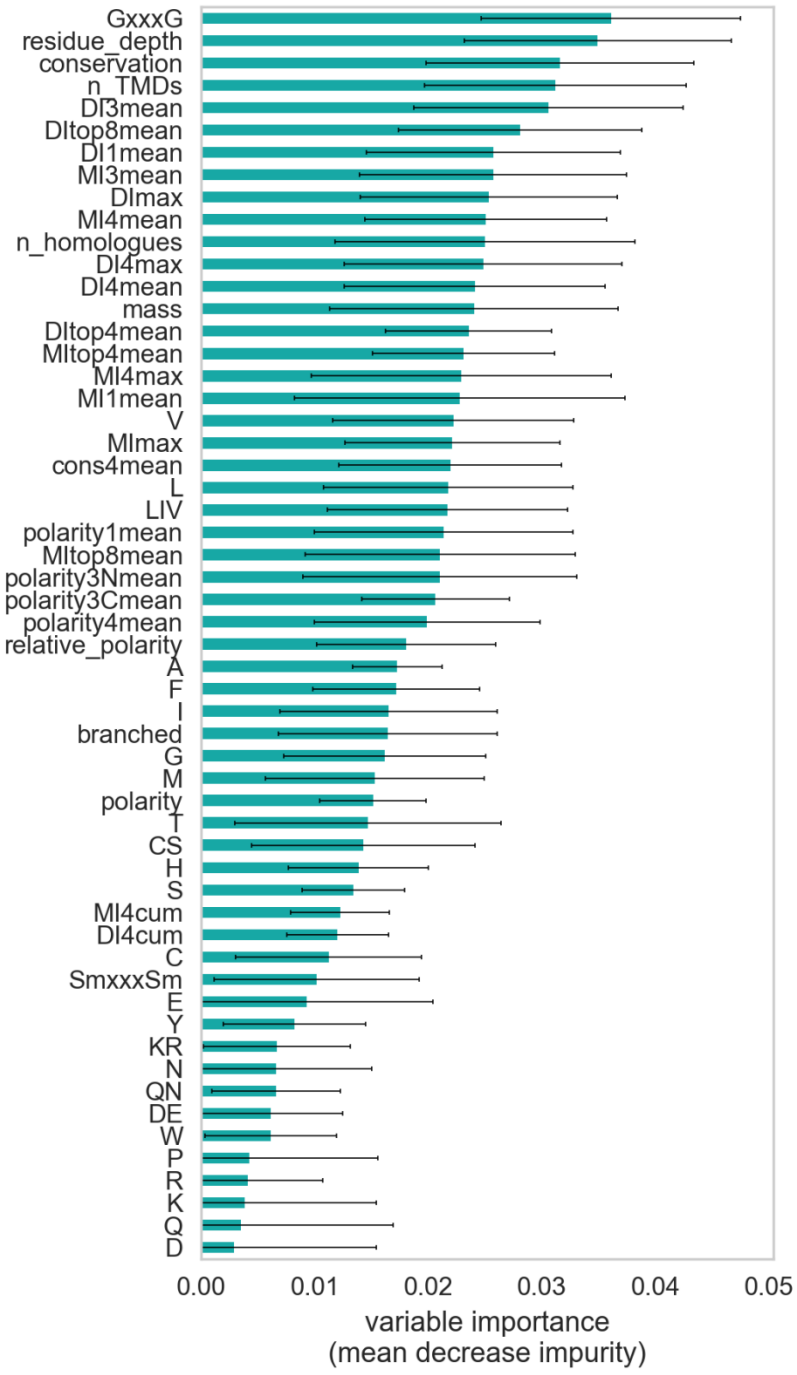



**Figure 2-6: Feature importances as ranked by THOIPA.** The relative importance of each feature was ranked during machine-learning training using standard methods, based on the mean decrease in impurity over respective nodes in the decision tree. In comparison to random forest, the extra-trees algorithm used in THOIPA is less susceptible to masking effects geurts [177]. Correlated features with similar predictive power therefore show a similar a level of importance.

## 2.4 Discussion

One of the major obstacles in understanding homotypic TMD-TMD interactions has been the small number of TMDs investigated via NMR. Our current knowledge of residue properties mostly stems from these case studies, artificial selection, or the analysis of TM helix pairs within polytopic membrane proteins. Until now, it was uncertain how many of these findings could be directly transferred to naturally evolved PPI interfaces. Here, we assembled the largest database of experimental, homotypic TMD-TMD interfaces. To do so, we extracted data from diverse sources, including our own experimental ToxR results, other ETRA and NMR studies from literature, and a novel database of self-interacting TM helices from crystal structures. This allowed us to compare the average sequence properties of interfacial and non-interfacial amino acids. We present several findings that are statistically robust, and directly relevant to the forces driving PPI in the membrane environment. However we show that there is a huge variability between TMDs in the properties of interfacial residues. We trained the first machine learning algorithm for the prediction of homotypic TM interfaces, which performed better than other automated methods. This also yielded an objective ranking of the most important features distinguishing interface and non-interfacial residues.

We present a novel dataset, consisting of self-interacting TM helices from crystal structures. In other fields, the identification of crystal contacts as biological interfaces is quite controversial. In our case, we only included TM helices that showed a close self-interaction. The helix pairs therefore resembled the TM homodimers seen in NMR studies. Their interfacial residues were highly conserved and coevolved, strongly suggesting that they are biologically important for the organism of interest. Most importantly, adding the

self-interacting helices from crystal studies greatly diversified the species of origin of the TMDs in our complete dataset. This makes our conclusions relevant to more than just human proteins, which comprised 93% of the proteins investigated by NMR and ETRA. Crystallography techniques are also completely unrelated to NMR and ETRA, reducing methodological bias in the analysis of the properties of interfacial residues.

So what are the underlying properties of the homotypic TM interfaces? Firstly, interface residues are significantly more conserved. This has been previously implied in numerous case studies [159, 178-182]. We find that interface residues were statistically more conserved than non-interface residues for all datasets, regardless of the experimental approach used. Also, residue conservation was objectively ranked by THOIPA as one of the most important predictive features. The TMD-TMD interfaces involved in PPI therefore resemble the permanently interacting interfaces found in multi-pass membrane proteins [23, 183]. Therefore for most of the TMDs examined, the homotypic interfaces are ancient and play an important role in the cell. However for predictive purposes the relative conservation of interfacial residues varied immensely between TMDs. In some cases this could be attributed to the presence of both homo- and hetero-dimer interfaces as proposed for ATP1B1 [180] and FtsB [184]. In other cases the interface may only be recently evolved, or perhaps tolerant of substitutions to other residues with similar properties.

The second feature associated with interface residues was their higher polarity. This suggests that PPI in the membrane are mediated by the same forces that drive the folding and stability within multi-pass membrane proteins [140, 141, 168]. The forces by which polar residues stabilise helix-helix interactions are described in detail elsewhere, but

typically involve H-bonds [185]. The strengths of these H-bonds between different functional groups and in different environments are still hotly disputed[186].

The third feature of interface residues was their high coevolution. The coevolution of contacting residues has allowed powerful predictions of protein folding and even protein-protein interacting partners [88, 187]. For the purposes of detecting TM homodimer interfaces, however, coevolution values had a modest contribution. Using a simple randomisation technique, we showed that retrospective methods may have overestimated the higher coevolution previously proposed for interface residues. Nevertheless this study provides support for hypothesis of Wang and Barth [162], who proposed that intra-helical coevolution values are stronger at homodimer interfaces. Our data supports this firstly by the observation that some coevolution values were higher for crystal interfaces (Figure 5-2), and secondly by the usefulness of coevolution values within THOIPA (FigBZ13). Biologically, this suggests that the TMD dimers, symmetric or not, depend on close contacts between non-identical interface residues. These contacts lead to coevolution, as a disruptive mutation in one residue is counterbalanced by a favourable mutation in the other. We attribute the lower predictive power of coevolution for the bitopic datasets (NMR/ETRA) to the lower number of available homologues, a problem that negatively impacts all coevolution methods [144]. The exponential increase in sequence data should greatly improve the usefulness of coevolution values in the future.

A simple but novel feature associated with interface residues is their depth in the membrane. Why did interface residues have a preference for the membrane hydrophobic core? This may suggest that helix-helix pairs are more stable when their interacting sites are deeper in the membrane, increasing the favourability of polar residue-residue

contacts in the absence of water. The fact that this effect was seen in TMD homodimers investigated by ETRA, NMR and crystal studies suggests that this is a genuine biological feature, rather than an artefact associated with a particular experimental technique. Further research is necessary to determine if this is common feature of protein-protein interactions mediated by membrane helices.

The excellent performance of THOIPA for the prediction of the most important interfacial residues suggests that it is a useful tool to guide experimental and structural modeling approaches. As yet it is not completely clear how evolutionary data and energy-based modeling can be most effectively combined to yield potential TM homodimer interfaces. Possibilities include 1) the addition of automatically generated structures as features within machine learning predictors, 2) the use of evolutionary data (e.g. THOIPA) to choose the most biologically relevant dimer structure, or 3) the use of evolutionary data to provide constraints during structural modeling. The latter approach was used in the EFDOCK-TM method of Wang and Barth, however the resultant structures have not yet been independently validated. In general, we support the premise that interface residues could be identified by combining LIPS and coevolution analyses. Our data suggest that EFDOCK-TM would be more successful for the ETRA than the NMR TMDs. For our small, non-redundant NMR subset, LIPS performed poorly, and coevolution was far inferior for interface prediction than simple measures such as residue depth.

Our stringent validation of TMDOCK and PREDDIMER was not only independent but in many cases blind, as these algorithms have never been tested against the new TMDs of the ETRA and crystal datasets. Although we use different validation measures, the performance of these algorithms was approximately in line with their published results.

We confirm that the newer TMDOCK algorithm performs slightly better, as previously claimed [127], either due to better structures or improved structure ranking. Automated structure ranking is clearly challenging, and complicated by the possibility of multiple biological interfaces. Currently, for a TMD of interest, we use PREDDIMER and TMDOCK by subjectively selecting the best automatically generated structure, based on available experimental and evolutionary data. The latter is available from the standalone THOIPA program, whose output includes LIPS and THOIPA predictions, and also the conservation, coevolution, and relative polarity for each position in the TMD.

In conclusion, we confirm that the interfaces of TM homodimers have a lot in common with the permanent interfaces within multi-pass membrane proteins. Key interface properties are conservation, coevolution, polarity, residue depth. However interfaces are diverse and difficult to predict. Nearly all residue types were found at interfaces. Many interfaces were poorly conserved, coevolved, or polar. Furthermore, our current knowledge suggests that many TMDs contain multiple homodimer interfaces [77], or additional heterodimer interfaces [180, 184]. In this challenging environment, we created THOIPA, the first machine-learning predictor of TM homodimer interfaces. The ranking of feature importances by THOIPA provided further support that these interface properties can help distinguish interface and non-interfacial residues within TM homodimers.

# CHAPTER 3. PREDICTION OF INTERACTION SITES IN α-HELICAL MEMBRANE PROTEINS

Many integral membrane proteins, just like their globular counterparts, form either transient or permanent multi-subunit complexes to fulfil specific cellular roles. Although numerous interactions between these proteins have been experientially determined, the structural coverage of the complexes is very low. Therefore, the computational identification of the amino acid residues involved in the interaction interfaces is a crucial step towards the functional annotation of all membrane proteins. Here, we present MBPred (Membrane-protein Binding-residues Prediction), a sequence-based method for predicting the interface residues in transmembrane proteins. A unique feature of our method is that it contains separate random forest models for two different use cases: a) when the location of transmembrane regions is precisely known from a crystal structure, and b) when it is predicted from sequence. In stark contrast to the aqueous-exposed protein segments, we found that the interaction sites located in the membrane are not enriched for evolutionary conservation, most likely due to their restricted amino acid composition. On the other hand, residue co-evolution proved to be a very informative feature, which has not so far been used for predicting interaction sites. MBPred reaches AUC, precision and recall values of 0.79/0.73, 0.69/0.51 and 0.55/0.48 on the cross-validation and independent test dataset, respectively, thus outperforming the previously published method of Bordner as well as all methods trained on globular proteins.

Moreover, we show that for the majority of complete interface patches, the method is capable of capturing more than 50% of the involved residues.

## 3.1 Introduction

A full understanding of a protein's function requires not only a possibly complete knowledge of its interaction partners, but also of the binding site location on its surface. Modern high-throughput assays, such as yeast two-hybrid or tandem affinity purification have generated data on close to 900,000 binary interactions [188]. By contrast, information about the specific interaction interfaces remains relatively scarce. Out of the 52,660 proteins from *Human* and the model organisms *E. coli* and *Bacillus subtilis* (bacteria), *S. cerevisiae* (fungi) and *Mus musculus* (Vertebrates) only 3,318 proteins (6.3%) have regions annotated as interaction sites in the Swiss-Prot [189] database, with only 12 of them annotated as 3D structure-derived. Furthermore, 1,722 and 39 proteins have annotations obtained by similarity-based transfer or motif and rule-based approaches, respectively. Less than a third of the proteins (1,296) were annotated based on publications involving experiments, such as alanine-scanning mutagenesis [190]. Manual inspection of these experimental annotations revealed that many of them are also based on 3d structures. In addition, 710 proteins have interface region annotations for which no evidence is provided. The percentage of proteins with available interface annotations ranges from less than 1% in bacteria to around 2% in *yeast* and 8-9% in vertebrates. For transmembrane proteins (TMPs), which constitute 20-30% of all proteins

the living cells [191], these numbers are even lower – only 0.5% in yeast and 5-6% in vertebrates.

PPI interfaces possess specific physico-chemical (amino acid composition, hydrophobicity, polarity), geometrical (accessible area, planarity), and evolutionary (conservation) properties [192], which makes them amenable to recognition by machine learning methods. A number of sequence- [193-197] and structure-based [95, 198] computational methods have been proposed to predict PPI interfaces in globular proteins. The latter group of methods tends to be more accurate, because they can leverage structure-level information such as solvent accessibility and the proximity of residues to each other, while the former one has the advantage of being applicable to the vast majority of proteins for which no experimental atomic structure is available. Most methods use machine learning techniques such as neural networks [199-202], support vector machines [203-206] and random forest [207-209] and almost all of them have been trained on globular proteins. The only method specifically geared towards predicting the interface residues in TMPs from sequence was proposed by Bordner in 2009 [114]. His random forest model, trained on evolutionary profiles extracted from 128 TMPs, achieved an average AUC of 0.75. Over the past 10 years, not only has the number of experimentally determined 3D structures of TMPs significantly increased but also database search tools such as HHblits have become much more sensitive [115]; additionally, vastly improved sequence co-evolution measures have become available [144, 210], providing powerful features for training machine learning algorithms.

Here we describe a novel computational method MBPred (Membrane-protein Binding-residues Prediction), which utilizes a combination of four individual random forest models

- MBPredTM, MBPredCyto, MBPredExtra, and MBPredAll – trained to predict residues involved in protein interactions in transmembrane, cytoplasmic, and extracellular segments as well as in the entire amino acid sequence, respectively. MBPredCombined merges the output of MBPredTM, MBPredCyto, MBPredExtra and is used when the location of the TM segments is known from structure or other experiments. Alternatively, MBPredAll is used when the location of the TM segments is unknown and therefore has to be predicted from sequence. The method was trained on 171 structures of $\alpha$-helical membrane proteins from 133 complexes and tested on an independent dataset of 36 structures. Since the Bordner's method does not appear to be available, no direct comparison with MBPred was possible. However, in our own implementation, a similar method only using evolutionary features achieved the AUC of 0.75 on the much larger dataset, while adding further features, such as TM helix orientation, residue co-evolution, and relative residue position with respect to the membrane, improved the AUC-based performance to 0.79. We also demonstrate that the surface patches consisting of amino acid residues classified by MBPred as interacting exhibit a significant overlap with the structure-derived interface regions. In 75% of the proteins, more than a half of the residues in the interface patches were correctly predicted.

## 3.2 Materials and Methods

### 3.2.1 Datasets

For training and benchmarking our method we created three datasets: i) comparison dataset (CompData), solely used for comparing the results with the previous work of

Bordner [114], ii) classification dataset (ClassData), for training and cross validating the classifier, and iii) an independent test dataset (TestData), for evaluating the performance of the final classifier.

#### 3.2.1.1 Comparison dataset

In the work of Bordner [114] a manually curated non-redundant dataset of TMP structures solved at better than 3.5 Å resolution was created. The original dataset included 64 α-helical multimeric complexes, 21 monomeric proteins, and 37 complexes of β-barrels. As our predictor is specialized on α-helical TMPs, we retained only the 135 α-helical subunits extracted from the α-helical multimeric complexes and the monomeric proteins. Although the Bordner dataset was claimed to be non-redundant, no specific information about the procedure used to reduce sequence redundancy is given in reference [114]. Clustering this dataset using CD-HIT [139] with a sequence identity cutoff of 30% resulted in a set of 101 unique protein sequences, further referred to as CompData (Table 5-2). The corresponding 3D coordinates were downloaded from the Protein Data Bank of Transmembrane Proteins (PDBTM) [138], which also provides information on the orientation of the TMP relative to the lipid bilayer calculated by TMDET [211]. This orientation information was used to extract the sequence positions of transmembrane and extramembranous regions, as explained below.

#### 3.2.1.2 Classification dataset

The classification dataset (ClassData) of transmembrane protein complexes was created to train our final prediction method and to assess its accuracy. The “Redundant Alpha”

dataset comprising 7374 TMP chains was obtained from PDBTM (version of June 2015). Biological complexes were first constructed by using the BIOMATRIX record given in the PDB file. Only the protein chains which have at least one biological contact with another protein were retained, and this dataset was made non-redundant at 30% sequence identity level. The resulting ClassData dataset included 171 unique proteins (Table 5-3), a 70% increase compared with the CompData dataset described above.

#### 3.2.1.3 Independent test dataset

In order to create an independent test dataset (TestData), all the new TMP chains added to the PDBTM database between June 2015 and June 2017 were downloaded and filtered using the same procedure as described for the classification data set (ClassData). Upon removing all proteins sharing more than 30% with any other sequence in ClassData or TestData itself, the latter dataset included 36 unique proteins (Table 5-4).

### 3.2.2 Definition of interacting residues

Defining interacting residues is a difficult task, as there are virtually as many definitions as there are publications on that topic. We selected three definitions, which use different distance cutoffs and different criteria for solvent accessible surface area (SASA) to determine whether or not two residues interact with each other, denoted as BordInter [114], FuchsInter [212] and RostInter [213]. The SASA is usually calculated using the 'rolling ball' algorithm [214], which considers a sphere of a particular radius to 'probe' the surface of the molecule. A typical value of the 'probe radius' is 1.4 Å, which approximates

the radius of a water molecule. The relative solvent accessibility (RSA) is the per-residue ratio between the calculated SASA and the maximum SASA for a particular residue. According to the BordInter definition, a residue residing in a TM segment is involved in an interaction if its RSA in the unbound state is larger than 0.2 and the distance between this residue and a residue of another protein chain in the same complex is below 4 Å. In this case, the inter-residue distance is defined as the nearest distance between any heavy atoms of the given residue pair. FuchsInter defines two residues to be in contact if they are situated on different TM segments and the minimal distance between any pair of heavy atoms is below 5.5 Å. Finally, according to the RostInter definition, residues Rx and Ry from two different chains X and Y are interacting if at least one pair of non-hydrogen atoms is closer than 6 Å, or Rx and Ry meet all of the following three conditions: (i) Rx and Ry change SASA after binding, (ii) Rx has no other interaction partners within 6 Å, (iii) from all residues that change SASA in chain Y, Ry is the closest residue to Rx. The impact of the contact definition on the total number of binding residues in ClassData and the overall prediction accuracy will be discussed in the *Results* section. Based on the above residue contact definitions, all residues were categorized into either the interacting or the non-interacting class.

### 3.2.3 Interface patches

Interface patches for each of the 171 ClassData proteins were identified based on the procedure proposed by Northey *et al*. [105]: i) for each TMP, the patch center residues with a relative solvent accessibility (RASA) of more than 25% are chosen, ii) for each patch center residue, the patch center atom with the highest absolute solvent accessible area is selected, iii) BiopTools [215] was applied to each patch center atom, resulting in

a total of 652 surface patches, iv) surface patches in which the RASA of the interface residues determined from structure makes up more than 50% of the total RASA were considered to be interface patches. The total of 249 interface patches were delineated.

### 3.2.4 TMP segments

TMPs contain three types of segments, which define their topology: extracellular (Extra), transmembrane (TM), and cytoplasmic (Cyto) segments. The complete sequence of the TMP is denoted as ‘All’ in this article. As the structure of the proteins in the training dataset is known, the TM regions were extracted according to the PDBTM definitions. Since PDBTM does not always contain information about the localization of extramembraneous regions (inside or outside), we used Phobius [216] predictions to verify sequence topology. A non-TM segment as defined by PDBTM was confirmed as cytoplasmic if the overlap between this segment and the cytoplasmic region predicted by Phobius was larger than the overlap between this segment and the predicted extracellular region. The same approach was used for extracellular regions.

Our classifier is solely trained on structure-derived topology, which is rarely available in real-world applications. We therefore benchmarked the method both on known and predicted topologies. If available, the experimentally determined topology can be submitted by the user for running predictions using the released model; otherwise, Phobius will be utilized to automatically predict the protein’s topology.

### 3.2.5 Multiple sequence alignments

Evolutionary information for each protein was gathered by searching the entire UniProt database *[217]* for related sequences using HHblits *[115]*, a fast HMM-HMM-based iterative sequence search tool. The benefit of HHblits is its ability to directly produce a multiple sequence alignment (MSA) out of the search results. In order to obtain maximum alignment size, we set the HHblits parameters to the following values: -Z 999999999 -B 999999999 -maxfilt 999999999 –id 99 –diff inf, which disable alignment filtering *[218]*. The arbitrary high numbers for –B and –Z maximize the number of sequences in the alignment and the summary hit list, while –maxfilt removes the limit on searched sequences. -id controls the maximum allowed pairwise sequence identity between the hits and was set to 99 in order to retrieve as many of non-identical sequences as possible. Finally, setting -diff to infinity removes the diversity filter, which would otherwise reduce the number of sequences in the MSA.

### 3.2.6 Random forest classification models

Random forest (RF) is an ensemble machine learning method, which can be used for classification and regression. The algorithm relies on a large number of decision trees and predicts the class for a particular input instance by the majority vote. During the model training, the input data for each tree is randomized in two ways. First, a subset of the total available features is selected randomly. By default, the size of this subset is equal to the square root of the total number of features. Second, the training data for each tree is sampled with replacement. The two parameters, which have the highest impact on prediction performance of the method, are the total number of trees created per each forest, and the number of features used for each tree. We set the number of trees to 2000,

and the number of features per tree was left at the default value of the square root of the total number of available features.

Since the data is highly unbalanced, *i.e.* the non-interacting class constitutes the majority of the data; the classifier will be biased towards the “major” class. To mitigate this problem, we tried two common approaches. The first one is based on cost sensitive learning and works by assigning a high cost to the misclassification of the minority class, therefore trying to minimize the overall cost. The second approach is to use a sampling technique, which can be either under-sampling the majority class, over-sampling the minority class, or a mixture of both. After comparing these methods, we chose the down-sampling method as it results in a lower prediction error (data not shown).

The final method MBPred (Membrane-protein Binding-residues Prediction) is a suite of four individual RF models – MBPredTM, MBPredCyto, MBPredExtra, and MBPredAll – trained to predict protein interaction sites in transmembrane, cytoplasmic, and extracellular segments as well as in entire amino acid sequences, respectively (Table 3-1). The combined output of the three segment-based models (MBPredTM, MBPredCyto, MBPredExtra) is referred to as MBPredCombined. The sole input for MBPredAll is a protein sequence, while MBPredCombined optionally takes as input the known membrane protein topology. If no user-supplied protein topology is available, MBPredCombined relies on Phobius predictions. Either way, MBPredCombined applies the three underlying models to predicted or known transmembrane, cytoplasmic, and extracellular segments and merges the predictions over the entire protein chain. Both MBPredCombined and MBPredAll thus classify all protein residues as interacting or non-

interacting, but using MBPredCombined also creates the opportunity to assess segment-based prediction quality.

**Table 3-1: The MBPred software suite consists of two main methods - MBPredCombined and MBPredAll.** The former combines three individual classifiers trained and tested on specific segment types separately, while the latter is trained on full protein sequences.

| Method name | Random forest classifier | Training data | Source of TM topology for training | Source of TM topology for prediction | Use case |
|---|---|---|---|---|---|
| MBPredCombined | MBPredTM | TM segments | Determined by structure and not used as feature | Determined by structure and not used as feature | TMP segments are known |
| | MBPredCyto | Cyto segments | | | |
| | MBPredExtra | Extra segments | | | |
| MBPredAll | MBPredAll | Entire TMP | Determined by structure and used as feature | Predicted from sequence and used as feature | TMP segments are not known |

### 3.2.7 Input features

The MBPred RF models were trained with three types of features derived from primary sequences (relative position, physical properties and segment) and five types of features calculated from MSAs (residue conservation, evolutionary profile (PSSM), cumulative and maximum co-evolution strength, and lipid accessibility) (Figure 3-1). The predictor output for a particular residue can be either 1 (interacting) or 0 (non-interacting).

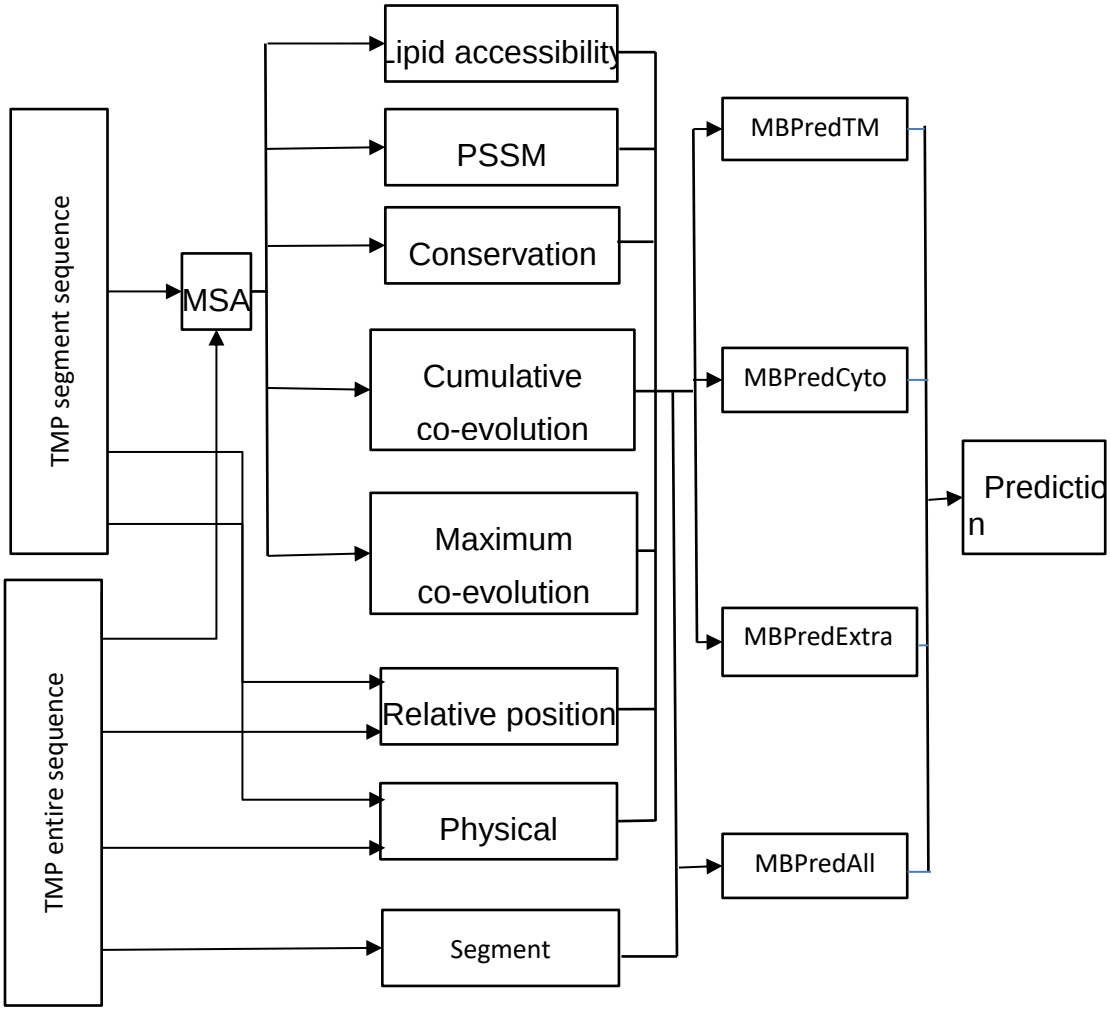



**Figure 3-1: Schematic overview of the RF-based classifiers of the MBPred suite for predicting interacting residues in TMPs.**

#### 3.2.7.1 Residue entropy and conservation

There is a strong selective pressure acting on functionally and structurally important residues [219], making them more evolutionary conserved. In particular, protein-binding interfaces are thought to be distinguishable from the rest of the protein surface by their higher degree of residue conservation [220]. The Shannon entropy for each residue at a certain position in the MSA was calculated according to the formula:

$$S_{entropy} = -\sum_{i=1}^{20} p_i log_2 p_i$$

where $p_i$ represents the observed frequency of the amino acid *i* in the alignment column. Lower values of Shannon entropy correspond to higher conservation. Entropy values were normalized to the range between 0 and 1 and transformed in such a way that higher values account for a stronger conservation:

$$\text{Conservation} = 1 - \frac{S_{entropy}}{\log(20)}$$

#### 3.2.7.2 PSSM

In this study, we used position specific scoring matrices (PSSM) to quantify the evolutionary profile of each amino acid in a protein sequence:

$$\text{PSSM(aa)} = \frac{rf_{aa}}{bf_{aa}}$$

where $rf_{aa}$ is the relative frequency of the amino acid *aa* in the MSA column at the position of interest and $bf_{aa}$ is the relative background frequency of the amino acid *aa*, which was calculated as the fraction of the given residue *aa* in the entire MSA. A 20-dimensional vector represents each PSSM position, with each element of the vector corresponding to one of the 20 amino acid types.

#### 3.2.7.3 Co-evolutionary strength

The term residue co-evolution refers to coordinated mutations of amino acids in two sequence positions to maintain energetically favorable interactions. Residue co-evolution is thus indicative of physical contacts between amino acids. We used two computational

measures to assess the strength of the co-evolutionary relationship between a pair of residues *i* and *j*. One of them is mutual information MI(i,j), which is calculated as

$$MI(i,j)=\sum_{A_i,A_j=1}^{q} f_{ij}(A_i,A_j)\,ln\,(\frac{f_{ij}(A_i,A_j)}{f_i(A_i)f_j(A_j)})$$

where $f_{ij}(A_i,A_j)$ is the observed frequency of amino acid pairs $A_i$, $A_j$ jointly occurring at positions i and j of a MSA, $f_i(A_i)$ and $f_j(A_j)$ are the overall probabilities of residue A at position i and residue A at position j, and q is the number of all possible residue pairs ($A_i$ , $A_j$).

MI calculation is solely based on the local residue pair probability in two MSA columns and does not consider transitivity effects [171]. This means that if residues in position *i* co-evolve with residues in position *k*, which in turn co-evolve with the residues in position *j*, the high value of MI(i,j) does not necessarily reflect a direct physical contact between these two protein sites. For this reason, we also employed an improved measure of residue co-evolution, called “direct information” [221], which is calculated as

$$DI(i,j)=\sum_{A_i,A_j=1}^{q} P_{ij}^{Dir}(A_i,A_j)\ln(\frac{P_{ij}^{Dir}(A_i,A_j)}{f_i(A_i)f_j(A_j)})$$

The local pair probability $f_{ij}(Ai,Aj)$ used in MI is replaced by the global pair probability $P_{ij}^{Dir}(A_i,A_j)$. The latter is calculated based on a global probability model using the entropy maximization approach, which calculates correlation scores for each pair of residues while taking into account all other pairs. For example, given a triple A, B and C where A-B and B-C are contacting pairs, direct interactions will avoid high correlation scores for the non-contacting pair A-C which would otherwise arise from transitive influence of the

contacting pairs [222]. Both MI and DI scores were calculated using the software tool FreeContact with all default parameters [223].

Residues important for protein function and structure tend to be involved in a large number of contacts and thus co-evolve with a large number of other residues [222]. For each residue we assessed its cumulative co-evolutionary strength as a measure of functional importance [110]. This was done by first ranking either MI or DI scores between the residue in question and any other residue paired with it in descending order. Co-evolutionary strength for this residue was then calculated as the sum of L best co-evolution scores [222]:

$$C_{DI}(i) = \sum_{(i,j \in l)} DI(ij)$$

$$C_{MI}(i) = \sum_{(i,j \in l)} MI(ij)$$

where *j* is a residue coevolving with *i*, $C_{DI}(i)$ and $C_{MI}(i)$ denote the cumulative DI and MI strength of the residue i, and $l$ is the list of top L highest ranking co-evolutionary pairs that residue *i* is involved in. L is either the sequence length of the whole protein (*all*) or the length of the segment (*seg*) where i and j are located in. Dependent on which definition of L was used, we distinguished between $C_{DI}^{seg}$, $C_{DI}^{all}$, $C_{MI}^{seg}$, $and\ C_{MI}^{all}$.

In addition to the cumulative co-evolutionary scores, another vector containing the maximum co-evolutionary scores for the current residue was established. According to the two possible definitions of L and the two co-evolutionary measures DI and MI, the vector contains the four maximum co-evolutionary scores: $M_{DI}^{seg}$, $M_{DI}^{all}$, $M_{MI}^{seg}$, $M_{MI}^{all}$.

#### 3.2.7.4 Relative Position

Two values were used to encode the relative positions of residues in the protein: i) the relative position of a residue in the topological segment it is located in (Rp1), calculated as the distance from the N-boundary of the segment to the residue position divided by the segment length, and ii) the relative position of a residue related to the entire protein sequence (Rp2), calculated as the position of the residue in the protein divided by the length of the protein.

#### 3.2.7.5 Lipid accessibility

LIPS (LIPid-facing Surface) [141] is a method for predicting TM helix-lipid interfaces from sequence information alone. The helix is split into seven helical faces, with each face being composed of a repeated pattern containing the anchor position *i* ranging from 1 to 7 (*i.e.* one anchor position for each face) and the residues at the positions *i+3*, *i+4* and *i+7*. Each helical face is assigned a LIPS score by multiplying the average alignment entropy with the lipophilicity of all residues contained in that face:

$$\text{LIPS} = \frac{\sum_{n=1}^{fn} e^{-\sum_{i=1}^{\text{rn}} \text{p(i)} * log_2 \text{p(i)}} * \sum_{n=1}^{fn} \sum_{i=1}^{\text{rn}} \text{p(i)} * \text{prop(i)}}{fn}$$

where *fn* is the number of residues in the face, *p(i)* and *rn* are the frequency of residue *i* and the number of residues in the alignment column, respectively and *prop(i)* is the lipophilicity propensity score of the residue *i*. The lipophilicity was calculated using the TMLIP2 scale [224], which provides two different lipophilicity values depending on whether the residue is located close (within 20% of the TM region length) to the membrane boundary or in the center of the TM. Amino acid residues Lys, Arg, Trp, Phe, and Leu get assigned a high lipophilicity value in the membrane boundary, while residues Ile, Leu, Phe and Val are given high lipophilicity values in the TM center. After ranking the

faces according to their LIPS score, each TM residue was represented as a vector consisting of three values: the rank of the predicted face (1-7) where the TM residue is situated (surfrank), the LIPS score of that face (surfscore), and the raw residue lipophilicity of the residue. Because surfrank, surfscore and the lipophilicity values are only applicable to the residues located in the TM segments, they were only used as input features for MBPredTM.

#### 3.2.7.6 Physico-chemical properties

Six physico-chemical properties of amino acids were obtained from the AAIndex [225] database; four of them use a continuous scale (hydrophobicity, polarity, charge and volume), while the other two provide a discrete value indicating if an amino acid is aliphatic or aromatic.

#### 3.2.7.7 Protein regions

The location of each residue was encoded by the letters ‘T’, ‘C’ or ‘E’ corresponding to the TM segments, the cytoplasmic or the extracellular parts of the protein (see section 3.2.4).

### 3.2.8 Feature importance

Mean decrease Gini (GiniDec) is a measure of feature importance included in the ‘randomForest’ R package. It is defined as the average gain of purity when splitting the data according to a given feature during the training process. Each time a variable is used for a split, the Gini coefficient is calculated before and after the split and the difference is averaged over all occurrences of that feature. Therefore, the higher the decrease in Gini coefficient is, the more important is the variable.

### 3.2.9 Measuring prediction performance

The overall prediction accuracy of MBPred was assessed based on ten-fold cross-validation. The dataset was randomly split into 10 equally sized bins. A random forest model was trained on data contained in 9 of these bins, and the remaining bin was used as a test data set to benchmark the model performance. This process was repeated 10 times, with each of the subsets serving as the test dataset once. The overall prediction performance was assessed by calculating the area under the ROC (Receiver Operating Characteristic) curve (AUC). The AUC can range from 0.0 to 1.0, with 0.5 indicating a totally random, while 1.0 a perfectly correct prediction.

To further quantify the overall prediction performance, we calculated precision, recall, Matthews Correlation Coefficient (MCC) and F1-score as follows:

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

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

$$\text{F1-score}=\frac{2*precision*recall}{precision+recall}$$

where TP (true positive) means interacting residues are correctly predicted to be interacting, FN (false negative) means interacting residues are incorrectly predicted not to be interacting, TN (true negative) means non-interacting residues are correctly predicted not to be interacting, and FP (false positive) means non-interacting residues are incorrectly predicted to be interacting.

## 3.3 Results and discussion

### 3.3.1 Feature analysis

#### 3.3.1.1 Binding residues are more conserved in the transmembrane portions of proteins

Residues mediating inter-molecular interactions tend to be evolutionarily conserved [226]. We compared sequence conservation calculated by the entropy-based score (section 2.6.1) between interacting and non-interacting residues in the three types of segments (TM, Cyto, and Extra) as well as in the full TMP sequences (All) (Figure 3-2). Alignment positions with more than 50% gaps were ignored. Interacting residues in the full TMP sequences are significantly more conserved than non-interacting residues ($p = 2.3e^{-10}$, t-test), but this difference is solely due to cytoplasmic ($p = 2.4e^{-5}$) and extracellular ($p = 2.4e^{-14}$) segments. There is no significant difference in the evolutionary conservation between interacting and non-interacting residues within the TM segments, presumably due to their restricted amino acid composition, *i.e*. the increased content of hydrophobic residues. Overall, TM residues exhibit stronger evolutionary conservation than Cyto and Extra residues.

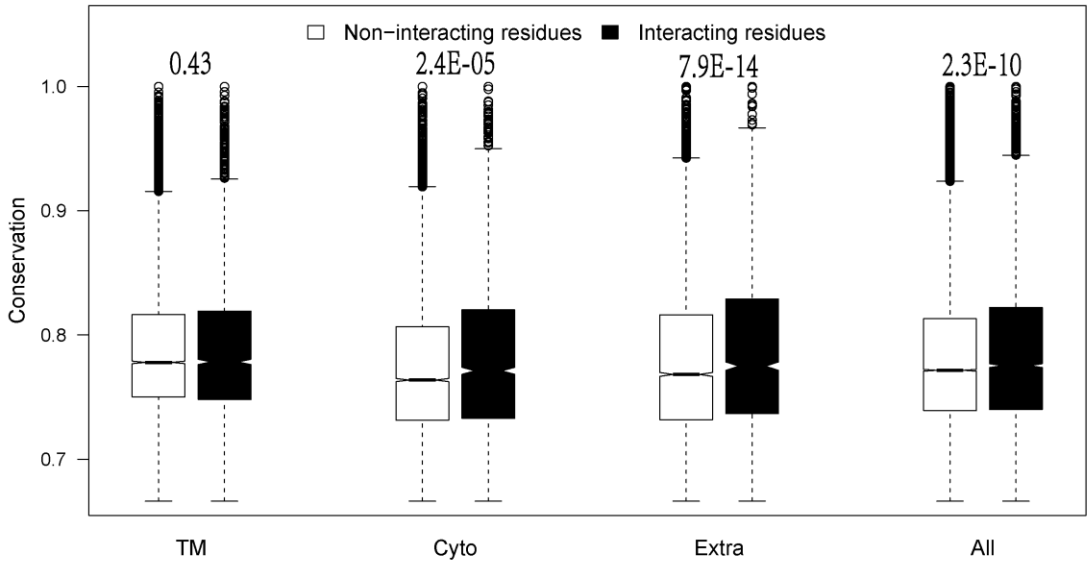



**Figure 3-2: Comparison of conservation scores between interacting and non-interacting residues in different TMP segments as well as in full sequences.** Interacting residues in the cytoplasmic and extracellular segments are much more conserved than non-interacting residues, while in the TM segments this difference is not significant.

#### 3.3.1.2 Interface residues tend to co-evolve with other interface residues

It has recently been shown that interacting residue pairs exhibit higher co-evolution scores than non-interacting residue pairs [132]. The cumulative and maximum co-evolutionary scores considered in this study would therefore be expected to contain a strong signal for potential residue contacts. Figure 3-3 shows the comparison of these two scores between interacting and non-interacting residues in individual TMP segments as well as the in the full protein sequences. The only distributions that exhibit no significant difference are those of $M_{DI}^{all}$ and $M_{MI}^{all}$ in TM segments (p=0.6558 and p=0.01372). We find that 77% of the amino acid pairs used for the $M_{DI}^{all}$ scores occurred in TM-TM residue

pairs, while only 11% and 12% of them were in TM-Cyto and TM-Extra residue pairs, respectively. These results imply that the residue pairs with the highest values of the co-evolutionary measures MI and DI in the TM segments are more likely to be involved in helix packing instead of intermolecular interactions. In all other cases, *i.e.* when considering either DI- or MI-based cumulative or maximum co-evolutionary scores in individual portions of proteins or in the entire sequences, the scores of interacting residues are significantly higher than the scores of non-interacting residues, thus constituting a promising signal for interface prediction.

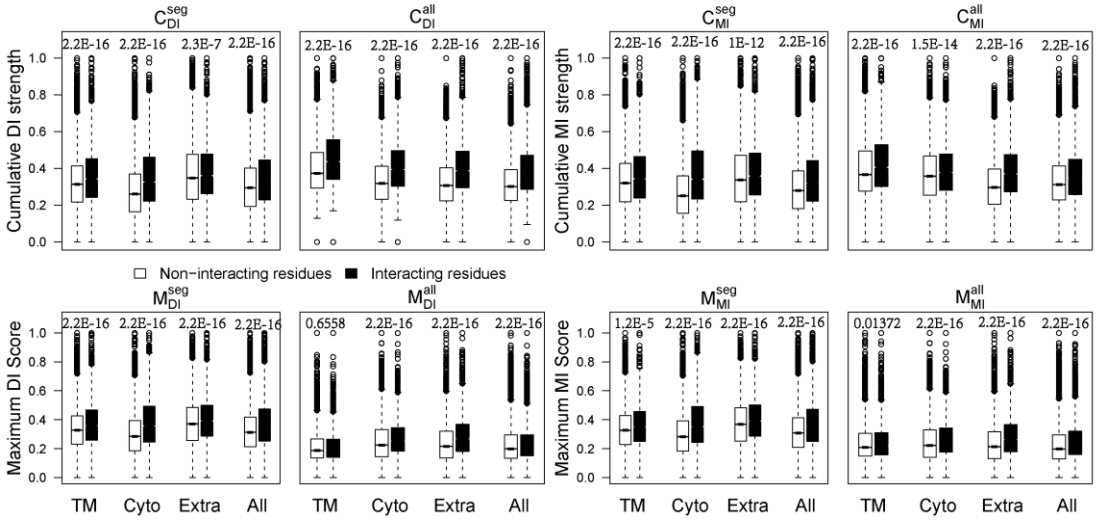



**Figure 3-3: Distribution of DI and MI scores in interacting and non-interacting residues in the three types of TMP segments (TM, Cyto, Extra) and the full sequences (All).** P-values indicate the significance of difference between co-evolutionary scores of the interacting and non-interacting residues.

### 3.3.2 Prediction performance of MBPred

Different portions of TMPs are known to have functionally and structurally distinct interaction landscapes. Interactions occurring in the transmembrane domains are critical for protein structure stabilization and biological activity [24], extracellular segments often

contain ligand binding sites, and cytoplasmic loops can be involved in intracellular signal transduction [227]. We therefore developed two alternative prediction techniques, one trained to recognize interaction sites in full-length protein sequences (MBPredAll), and another one (MBPredCombined) using an ensemble of three separate predictors (MBPredTM, MBPredCyto and MBPredExtra) trained on transmembrane, cytoplasmic, and extracellular portions of protein structures, respectively. These two techniques were compared based on four performance measures: ROC curve, precision-recall curve, MCC and F1-score, using ClassData for the 10-fold cross validation (Figure 3-4) and the independent TestData for benchmarking the final classifier (Figure 3-5). Because in real-world applications the correct transmembrane topology is often not known, the comparison was conducted using both experimentally determined and predicted protein segments. In the first step, MBPredCombined was compared with MBPredAll in terms of the overall and segment-wise performance. Subsequently, the predictions of MBPredAll were split into three parts according to the specific topological segments, *i.e.* TM, Cyto and Extra, and compared them with MBPredTM, MBPredCyto, and MBPredExtra predictions, respectively (Figure 3-4).

Using either ClassData or TestData, MBPredCombined performed slightly better than MBPredAll on segments derived from crystal structures (left part of Table 3-2). Consistently, MBPredTM, MBPredExtra and MBPredCyto achieved a higher accuracy on the segment types they were trained on, outperforming MBPredAll, which was trained without distinguishing between the topological segments. However, when using the segments predicted by Phobius (right part of Table 3-2), the opposite situation occurred, *i.e.* MBPredAll performed better than MBPredCombined when benchmarking the

segment-specific as well as the overall performance. MBPredCombined should therefore be preferred if the location of segments is exactly known, while MBPredAll works better for segments predicted from sequence. As expected, the performance of the three segment-based classifiers (MBPredTM, MBPredCyto, and MBPredExtra) drops significantly when applied to the segments they were not trained on (e.g. testing the performance of MBPredTM on extracellular segments) (Table 3-2).

The output of a classifier, which gives values between zero and one can be converted into the prediction of a class by applying a cutoff. A cutoff of 0.5, which is the default in most cases, is usually not optimal especially for imbalanced datasets. In order to find a proper classification cutoff for unknown data, we plotted the F1-score against all possible cutoffs using the cross-validation results on ClassData (Figure 3-4D). The F1-score is widely used to evaluate the success of an imbalanced binary classifier as it represents the harmonic mean of precision and recall. The optimal cutoff values were determined by taking the maximum F1-score for each of the four RF models (MBPredTM, MBPredCyto, MBPredExtra and MBPredAll), which reached their highest F1-scores of 0.493, 0.565, 0.479 and 0.509 at the cutoffs 0.31, 0.32, 0.27 and 0.27, correspondingly. Those optimal cutoffs, which maximized the F1-score are used in the final version of MBPred. Table 3-3 shows several performance metrics on ClassData and TestData after the application of these thresholds.

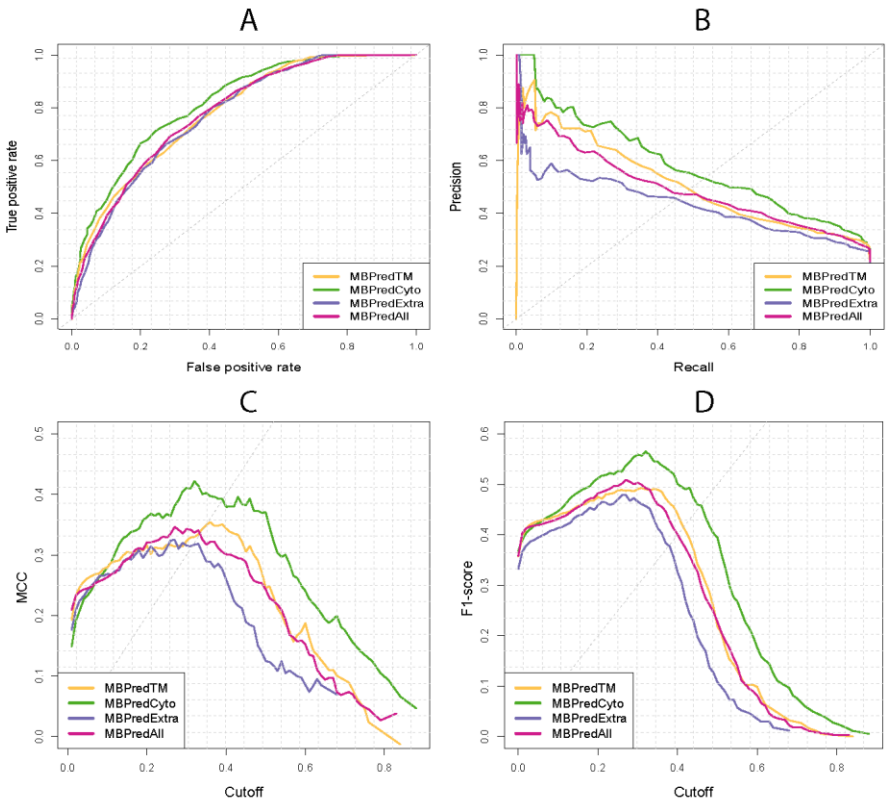



**Figure 3-4: Different performance measures of the classifiers during the 10-fold cross-validation using the ClassData dataset**: ROC curve, precision-recall curve, MCC and F1-score in MBPredTM, MBPredCyto, MBPredExtra and MBPredAll respectively.

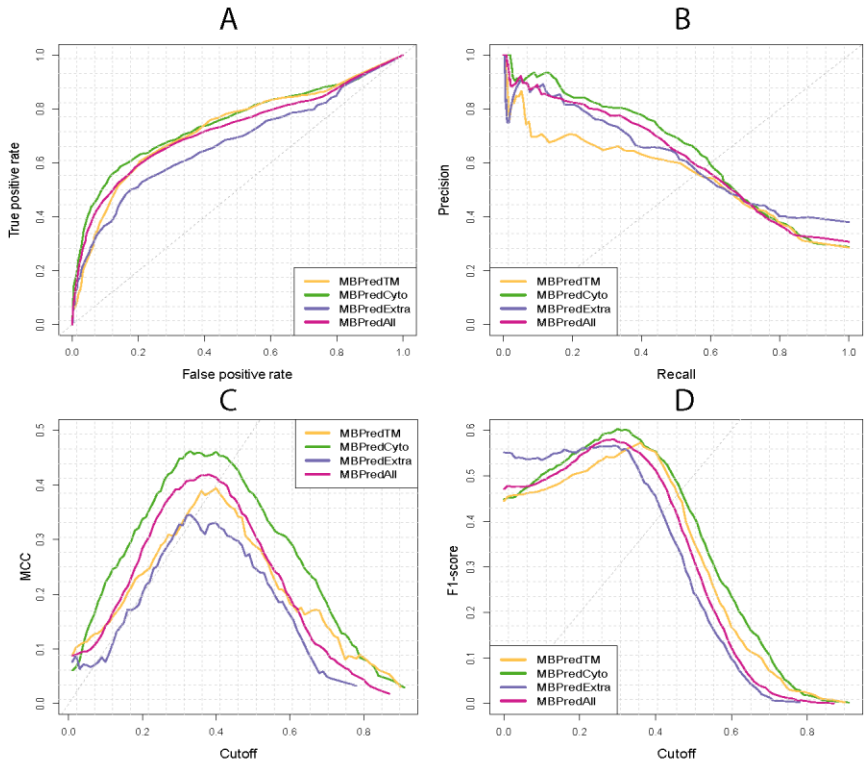



**Figure 3-5: Different performance measures of the classifiers on the new independent TestData dataset:** ROC curve, precision-recall curve, MCC and F1-score in MBPredTM, MBPredCyto, MBPredExtra and MBPredAll respectively.

**Table 3-2: AUC performance of predictors on ClassData or TestData.**

| | Structure | | | | Phobius | | | |
|---|---|---|---|---|---|---|---|---|
| | Cyto | Extra | TM | All | Cyto | Extra | TM | All |
| | | | | ClassData | | | | |
| MBPredAll | 0.801 | 0.769 | 0.776 | 0.782 | 0.746 | 0.711 | 0.758 | **0.742** |
| MBPredCyto | 0.818 | 0.695 | 0.696 | | 0.722 | 0.690 | 0.680 | |
| MBPredExtra | 0.722 | 0.772 | 0.688 | | 0.694 | 0.703 | 0.681 | |
| MBPredTM | 0.718 | 0.656 | 0.782 | | 0.653 | 0.630 | 0.738 | |
| MBPredCombined | | | | **0.791** | | | | 0.721 |
| | | | | TestData | | | | |
| MBPredAll | 0.745 | 0.672 | 0.724 | 0.721 | 0.709 | 0.674 | 0.720 | **0.704** |
| MBPredCyto | 0.760 | 0.626 | 0.605 | | 0.688 | 0.629 | 0.625 | |
| MBPredExtra | 0.581 | 0.685 | 0.581 | | 0.628 | 0.644 | 0.614 | |
| MBPredTM | 0.612 | 0.594 | 0.753 | | 0.594 | 0.597 | 0.714 | |
| MBPredCombined | | | | **0.732** | | | | 0.682 |

*Note:* Left part of the table: segments derived from crystal structures. Right part of the table: segments predicted by Phobius. Upper part of the table: average AUC values over 10-fold cross-validation on ClassData. Lower part of the table: AUC values on TestData. Gray shading: segment classifiers applied to segments they were not trained on. Bold: higher scores when comparing MBPredAll and MBPredCombined.

**Table 3-3: Performance metrics using structure derived TM segments for ClassData and TestData after application of the adjusted threshold.**

| | **ClassData** | | | | **TestData** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **MCC** | **F1-score** | **Precision** | **Recall** | **MCC** | **F1-score** |
| **MBPredAll** | 0.842 | 0.432 | 0.351 | 0.571 | 0.692 | 0.402 | 0.346 | 0.509 |
| **MBPredTM** | 0.747 | 0.507 | 0.389 | 0.604 | 0.591 | 0.422 | 0.328 | 0.493 |
| **MBPredCyto** | 0.783 | 0.541 | 0.444 | 0.640 | 0.664 | 0.491 | 0.422 | 0.565 |
| **MBPredExtra** | 0.865 | 0.418 | 0.386 | 0.563 | 0.633 | 0.385 | 0.324 | 0.479 |
| **MBPredCombined** | 0.689 | 0.547 | 0.423 | 0.610 | 0.507 | 0.481 | 0.351 | 0.493 |

### 3.3.3 Comparison of MBPred with Bordner's method

We investigated how well MBPred predicts interaction sites in transmembrane regions compared with the method of Bordner using the CompData dataset for training and BordCont as a contact definition (see *Materials and Methods*). Bordner's method, which only takes into account evolutionary conservation, PSSMs, and physical properties, achieved an AUC of 0.75 (Figure 3-6). Adding the LIPS related features, co-evolution features, and the relative position of residues in the protein, *i.e.* all the features used in MBPred, resulted in the AUC increase to 0.77, 0.78, and 0.79, respectively.

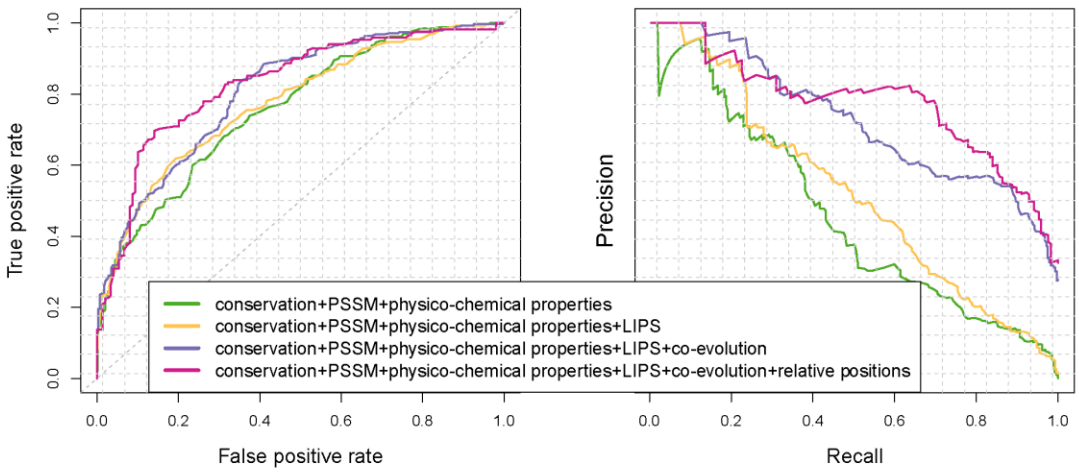



**Figure 3-6: ROC curves.** (A) and precision-recall curves (B) for predicting interacting residues using different feature combinations. Green color: evolutionary features and physico-chemical properties (Bordner's method) (AUC 0.75); orange color: with added LIPS-related features (AUC 0.77); blue color: with added co-evolutionary signal (AUC 0.78); purple color: with added relative positions of residues (AUC 0.79).

### 3.3.4 Variable importance

As described in section 3.2.8, GiniDec measures how much a feature contributes to the separation of the classes and therefore how important it is for the classification success. **Figure** 3-7 shows the variable importance GiniDec for our four individual RF models. First of all, conservation, co-evolution, and relative positions are among the highly important features. In each of the four individual RF models, conservation was in the top five features with an average GiniDec of 240.61. $C_{DI}^{all}$ was among the top six features with an average GiniDec of 248.66, and Rp2 was placed in the top four features with an average GiniDec of 261.53. Among the three LIPS related features in MBPredTM, the surfscore (GiniDec of 153.53) was more important than lipophilicity (GiniDec of 143.72) and surfrank (GiniDec of 71.05). Secondly, Rp2 is always among the top four features and

performs better than Rp1. The average GiniDec of Rp1 (201.04) in the four RF models was lower than Rp2 (261.53). Thirdly, all the six physicochemical properties were always the least important features in each of the individual RF models. Finally, among the 20 amino acids, we found that cysteine ranked first, first, fifth and first in MBPredTM, MBPredCyto, MBPredExtra, and MBPredAll, respectively, with an average GiniDec of 283.02, probably due to its lower abundance compared to the other 19 residues (Figure 3-8). Looking at the importance of amino acids in the individual segment types and the full protein sequence, it is noteworthy that the most important residues in MBPredTM were cysteine, methionine, histidine and aspartic acid, as opposed to alanine, leucine, glycine and valine reported by Bordner [26]. This could be due to the different contact definitions, larger training dataset, or the prevalence of our new features making the residue probabilities less important. In our dataset, the most prevalent residues at the interaction sites were all hydrophobic - leucine, isoleucine, valine, glycine and alanine - while the charged amino acids aspartic acid, glutamic acid, histidine, arginine and lysine were the least frequent. Overall, without distinguishing between interacting and non-interacting sites, polar and charged residues occur more frequently in cytoplasmic and extracellular regions than transmembrane regions, as one would expect.

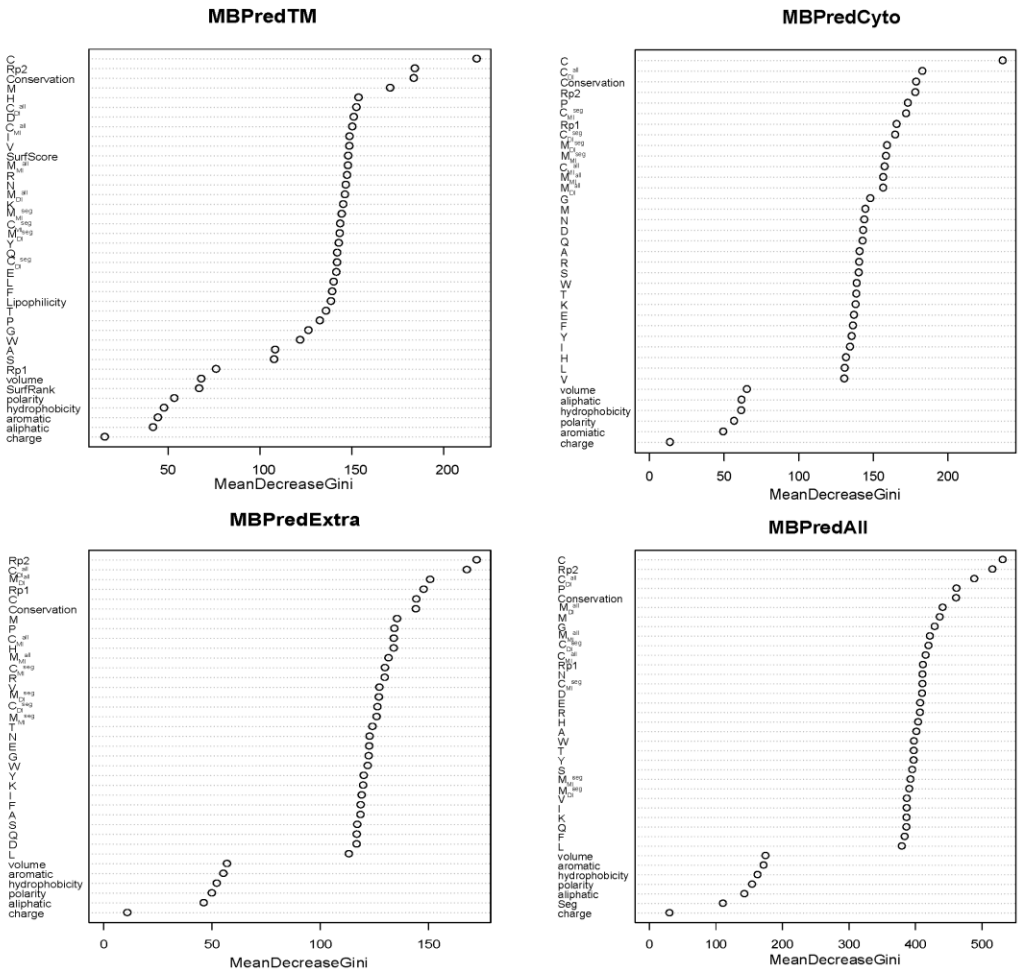



**Figure 3-7: Variable importance for four individual RF models** - MBPredTM, MBPredCyto, MBPredExtra and MBPredAll - measured with GiniDec. The importance of three LIPS related features was only measured in MBPredTM and one segment indicator feature (Seg) was

presented in MBPredAll. Features related to evolutionary conservation, co-evolution, and relative residue position exhibit the highest importance.

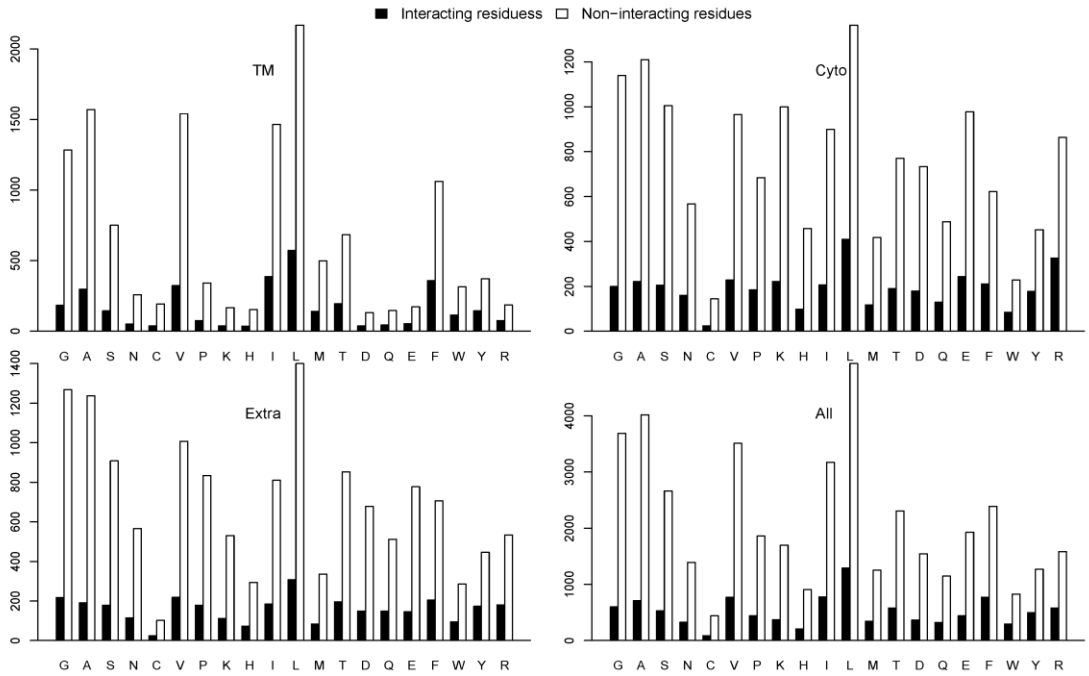



**Figure 3-8: Occurrences of amino acids in protein interaction sites (black) and non-interacting sites (white) in the four segment types.**

### 3.3.5 Impact of residue interact definition on classifier performance

To understand how residue contact definitions described in section 3.2.2 affect the classifier performance, we compared the average AUC of the four individual predictors using different contact definitions. In each segment type and therefore also in the entire sequence using RostInter results in the highest number of interacting residues, followed by FuchsInter and BordInter (Table 3-4). The same ranking becomes apparent when comparing the prediction performance; for example, for MBPredTM the AUC values were 0.792 (RostInter), 0.784 (FuchsInter), and 0.776 (BordInter). The prediction performance is thus clearly correlated with the choice of the contact definition and the resulting number

of interacting residues. The RostInter, FuchsInter and BordInter definitions are based on the heavy atom distance thresholds of 6.0, 5.5 and 4.0 Å, respectively. Expectedly, the definition with the least strict distance threshold of 6.0 Å (RostInter) yields a higher number of interacting residues than a definition using 5.5 Å (FuchInter) or 4.0 Å (BordInter). The correlation between the number of interacting residues and the prediction performance can be attributed to the reduced class imbalance caused by a higher percentage of residues labeled as interacting. Based on this assessment, we utilized the RostInter definition of interacting residues in our final classifiers.

**Table 3-4**: **Predicted number of interacting residues and prediction performance** (AUC) of the four classifiers (MBPredTM, MBPredCyto, MBPredExtra and MBPredAll) when using three different residue contact definitions.

| Contact definition | MBPredTM | | | MBPredCyto | | | MBPredExtra | | | MBPredAll | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ni | Nni | AUC | Ni | Nni | AUC | Ni | Nni | AUC | Ni | Nni | AUC |
| BordInter | 3308 | 13457 | 0.776 | 3815 | 14985 | 0.797 | 3171 | 14083 | 0.777 | 10294 | 42525 | 0.773 |
| FuchInter | 4552 | 12213 | 0.784 | 5085 | 13715 | 0.812 | 4335 | 12919 | 0.786 | 13792 | 38847 | 0.787 |
| RostInter | 5047 | 11718 | 0.792 | 5581 | 13219 | 0.824 | 4768 | 12486 | 0.802 | 15396 | 37423 | 0.804 |

Note: Ni=the number of interacting residues, Nni=the number of non-interacting residues.

### 3.3.6 Case study: predicting the interaction interface for the photosystem II D2 protein

To illustrate the performance of MBpred, we present the prediction of interface residues for the photosystem II (PSII) D2 protein from *Thermosynechococcus elongates* (PDB

entry 4PJ0, chain D [228]). Photosystem II (PSII), a large protein complex consisting of 20 subunits, is a light-driven water plastoquinone oxidoreductase that uses light energy to abstract electrons from H2O, generating O2 and a proton gradient subsequently used for ATP formation [228]. PSII D2 is required for the assembly of a stable PSII complex. After excluding this protein from the ClassData, using the structure-derived topology of the protein, our classifier MBPredCombined achieved an AUC of 0.8549, with 193 out of 205 binding residues predicted correctly (Figure 3-9) . Chain D is situated in the center of the PSII complex and has an interface to 15 out of 20 subunits, with the exception of chains I, R, Y and Z. MBPredCombined was able to infer all of the actual interfaces, from the smallest one, which consists of just one interacting residue with chain K, to the largest one with chain A. The latter consists of 126 residues of which 97 have been predicted correctly. Nine of the 15 interfaces were predicted with a perfect coverage (K, U, J, V, T, M, F, O, E). The only interface poorly predicted by MBPredCombined was to chain C, where only 2 out of 6 interface residues were predicted correctly.

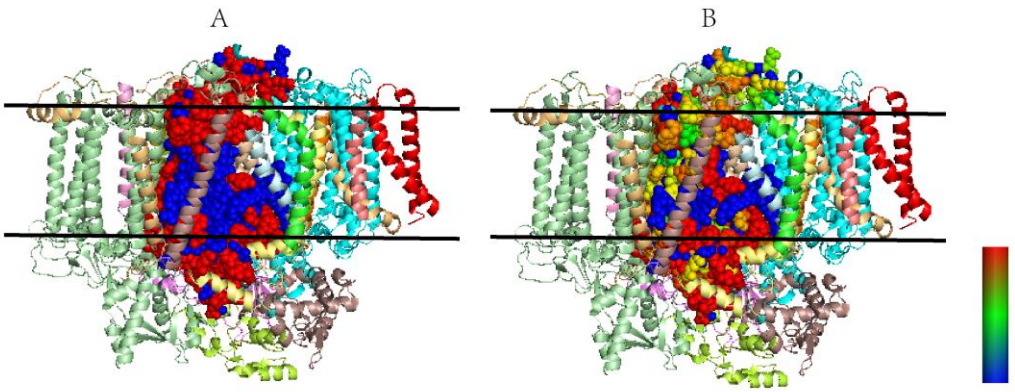



**Figure 3-9: MBpred prediction for the photosystem II D2 protein (PDB entry 4PJ0, chain D).** Amino acids of the photosystem II D2 protein are shown as colored spheres while the other

subunits are displayed using the cartoon representation. (A): Actual interacting (red spheres) and non-interacting (blue spheres) residues. (B): Interacting residues predicted by MBPredCombined. The spheres are colored depending on the random forest output, and therefore the interaction likelihood from red (interacting) to blue (not interacting).

### 3.3.7 Prediction of interaction interfaces

Functionally important residues often cluster together and form patches on the protein surface. Especially in TMPs, interactions tend to be quite large to exclude lipids between the interaction partners [229, 230]. Yet, most of the binding energy comes from small regions within interface patches, the so-called hot spots [231]. In order to determine how well MBPred predicts entire interface patches rather than individual residues, we calculated patches for the 171 ClassData proteins as described in *Materials and Methods* to measure the overlap between the predicted interface residues and the interface patches determined from the structure. For each interface patch the percentage of predicted interacting residues was determined. For 75% (186/249) of the interface patches the overlap was over 50% and 23 patches were exactly predicted (Figure 3-10).

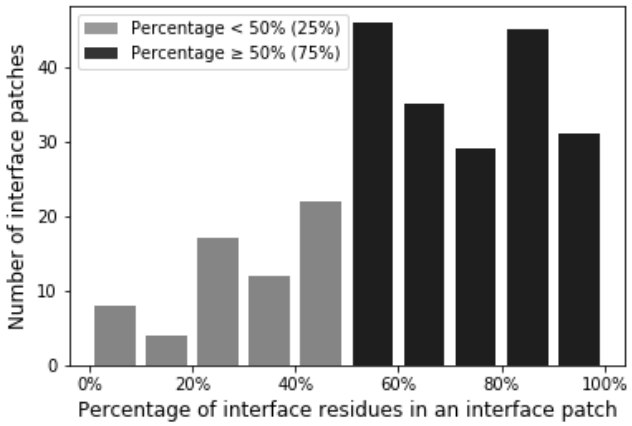


**Figure 3-10: Percentage of predicted interface residues in the interface patches.** Black and gray bars represent the numbers of interface patches which overlap with the interface residues predicted by MBPred by more and less than 50%, respectively.

### 3.3.8 Comparison to PSIVER - a method for globular proteins

To assess the benefits of a method tailored specifically to TMPs, we compared MBPred with PSIVER [195, 232], a sequence-based method for predicting interacting residues in globular proteins. PSIVER relies on the definition of interface residues based on the decrease of solvent accessibility during complex formation. More specifically, surface residues are defined as those having RASA greater than 5%, and if their absolute solvent accessibility decreases by more than 1 $Å^2$ upon complex formation, such surface residues are considered to be part of an interface. We ran PSIVER predictions for the 36 TestData proteins and compared the ROC and precision-recall curves of MBPred and PSIVER using only those residues, for which both interface definitions were consistent (Figure 3-11). MBPred achieved an AUC of 0.78 and average precision of 0.58, and thus outperformed PSIVER (AUC of 0.59 and an average precision of 0.22) by a wide margin. As PSIVER itself was reported to surpass other interface residue prediction methods in terms of accuracy, we conclude that TMPs with their distinct biophysical characteristics require a specialized method to achieve state of the art results.

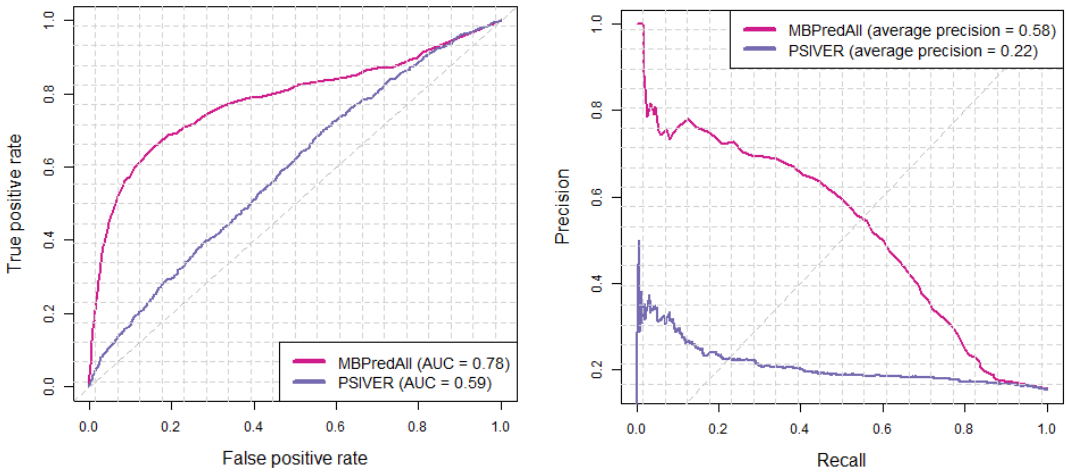



**Figure 3-11: Comparison of PSIVER and MBPred using ROC (left) and precision-recall (right) curves.**

### 3.3.9 Availability

The full source code and a standalone version of MBPred are available from https://github.com/bojigu/MBPred.git.

## 3.4 Conclusions

α-helical transmembrane proteins form complexes and bind to ligands in order to perform their biological functions. Elucidating the precise location of the binding sites is an indispensable part of protein functional annotation. Here we present a machine learning approach called MBPred for predicting interacting residues in α-helical membrane proteins from sequence alone. MBPred was developed having two application scenarios in mind. In the first situation, the user wishes to identify potential interaction sites for a

protein whose 3D structure and therefore, the topology, is known. In this case, MBPred uses an ensemble of three classification models, MBPredTM, MBPredCyto, MBPredExtra, which were separately trained on transmembrane, cytoplasmic, and extracellular regions, respectively. In the second scenario, the 3D structure is not known and the protein topology is predicted from sequence by appropriate methods, in this case Phobius. In this context, MBPred uses a single model trained on the entire protein sequence without distinguishing between different protein regions. We demonstrate that employing the method specifically tailored to one of these two distinct use cases allows achieving a higher prediction accuracy. Compared to non-interacting residues, interacting residues are significantly more conserved in the cytoplasmic and extracellular segments while no significant difference in evolutionary conservation could be established in the transmembrane regions. We speculate that the interacting residues are more conserved than non-interacting residues independently of the location but the restricted amino acid composition does not allow discerning this trend in the transmembrane regions. Interacting residues also exhibit significant higher co-evolutionary scores, indicating that interfaces involving multiple residues evolve in a coordinated fashion. The only type of residue pairs exhibiting even higher co-evolutionary scores were those involved in transmembrane helix packing. It is hoped that MBPred will become a useful tool for guiding experimental and theoretical investigations of membrane protein interactions.

# CHAPTER 4. SUMMARY

Overall, this thesis aims at a better understanding of protein-protein interactions occurring in alpha-helical membrane proteins. It intends to provide new algorithms specially developed for membrane proteins that can be used to predict interfacial residues on either single-pass homodimers (THOIPA) or the full sequence of alpha-helical membrane protein (MBPred). THOIPA and MBPred methods only require a membrane protein sequence input, and the presented method can predict and give a prioritized list of which residues participate in transmembrane protein-protein interactions. These two methods have potential application in guiding the experimental verification of membrane protein interactions, structure-based drug discovery, and also in predicting the membrane protein complex structures. Based on the obtained results presented in chapter 2 and 3, two interfacial residues prediction software were developed for membrane proteins.

### 4.1.1 Software development for homotypic helix-helix interfaces prediction

One of the major obstacles in understanding homotypic TMD-TMD interactions has been the small number of TMDs investigated via NMR. In order to compare the sequence properties of interfacial and non-interfacial residues, we firstly collected 54 experimental, NMR and crystal TMDs which form homotypic interfaces. We then analysed the underlying properties of the homotypic TM interfaces. We show that firstly, interface residues were statistically more conserved than non-interface residues for all datasets, regardless of the experimental approach used. Secondly, interface residues have their higher polarity, the forces by which polar residues stabilise helix-helix interactions are

described in detail elsewhere, but typically involve H-bonds. Thirdly, interfacial residues have higher co-evolution scores. Biologically, this suggests that the TMD dimers, symmetric or not, depend on close contacts between non-identical interface residues. These contacts lead to coevolution, as a disruptive mutation in one residue is counterbalanced by a favourable mutation in the other. Thirdly, interface residues have a preference in the membrane hydrophobic core. This may suggest that helix-helix pairs are more stable when their interacting sites are deeper in the membrane, increasing the favourability of polar residue-residue contacts in the absence of water.

We trained the first machine learning algorithm for the prediction of homotypic TM interfaces (THOIPA), which performed better than other automated methods. The ranking of feature importance by THOIPA provided further support that these interface properties can help distinguish interface and non-interfacial residues within TM homodimers. The excellent performance of THOIPA for the prediction of the most important interfacial residues suggests that it is a useful tool to guide experimental and structural modelling approaches.

### 4.1.2 Software development for alpha-helical membrane protein interface prediction

Alpha-helical membrane proteins interact with other proteins to fulfil specific cellular processes, identifying the amino acid residues involved in the interaction is crucial to annotate the function of the protein. However, rare of these interactions have been experimental conformed. Therefore, we developed the machine learning method (MBPred) to predict the interface residues using the sequence information alone. MBPred

could be divided into MBPredCombined (when the location of transmembrane region is known from structure) and MBPredAll (the transmembrane region is unknown and need prediction). We show that interfacial residues are significantly more conserved than non-interfacial residues in cytoplasmic and extracellular regions, but not significant difference in transmembrane regions, this may indicate that interface residues are indeed more conserved but do not allow discerning the trend perceived in soluble segments. We also found that interface residues have significant higher co-evolutionary scores. MBPred has an overall high prediction performance, reaches AUC, precision and recall values of 0.79/0.73, 0.69/0.51 and 0.55/0.48 on the cross-validation and independent test dataset, respectively, thus outperforming the previously published method of Bordner as well as all methods trained on globular proteins. We believe that MBPred will become a useful tool for guiding experimental and theoretical investigations of membrane protein interactions.

# CHAPTER 5. APPENDIX

## 5.1 AppendixA: Supplementary Methods

### 5.1.1 Calculation of residue properties

In this section we describe the calculation of physico-chemical, structural and evolutionary properties of amino acid residues at each sequence position (designated *i*) of transmembrane domains (TMD). In total, 56 features were considered. Some of the properties were derived from multiple sequence alignments (MSA). These were gathered by searching the NCBI non-redundant database for related sequences using BLASTp. Homologues were filtered by keeping only the alignments with fewer than 6 gaps and at least 20% sequence identity in the TMD region. Only homologues with unique TM sequences were retained (non-redundant to 100% sequence identity). In addition, we calculated position specific scoring matrices (PSSM) to quantify the evolutionary profile of each amino acid in a TMD. A PSSM contains the frequencies of all 20 amino acids in each MSA column.

**Residue coevolution (16 features)**

We employed the FreeContact implementation [233] of EVfold [234] to calculate coevolution scores between all possible residue pairs in the TMDs. For each residue pair, the EVfold output includes the values of mutual information (MI) and direct interaction (DI). Mutual information is a standard measure of coevolution between two residues but is known to be prone to several biases [234, 235]. For example, high scores can be seen

for indirect contacts, e.g. when residues B and C are not in contact, but both make contacts and coevolve with residue A. A second bias in MI is the low score associated with high conservation (Figure 5-9). Direct coupling analysis (DCA) is a global statistical inference method that aims to disentangle direct and indirect contacts and counter the effects of high conservation. DCA yields an adjusted DI score for each residue pair. For prediction in THOIPA, and to understand interface properties, it was necessary to convert the pairwise coevolution scores to a single representative value at each residue position. We included 16 such coevolution measures, comprising nine MI and nine DI values, respectively. In all cases, the predictive coevolution value was the maximum or mean from a selected number of residue pairs that included the residue of interest.

Briefly, for a pair of residues *i and j,* MI *was* calculated as:

$$MI(i,j)=\sum_{A_i,A_j=1}^{q} f_{ij}(A_i,A_j)\,ln\left(\frac{f_{ij}(A_i,A_j)}{f_i(A_i)f_j(A_j)}\right)$$

where $f_{ij}(A_i,A_j)$ is the observed frequency of amino acid pairs $A_i$, $A_j$ jointly occurring at positions i and j of an MSA, $f_i(A_i)$ and $f_j(A_j)$ are the overall probabilities of residue A at position i and residue A at position j, and q is the number of all possible residue pairs ($A_i$ , $A_j$).

DI was calculated according to the following equation:

$$DI(i,j)=\sum_{A_i,A_j=1}^{q} P_{ij}^{Dir}(A_i,A_j)\ln\left(\frac{P_{ij}^{Dir}(A_i,A_j)}{f_i(A_i)f_j(A_j)}\right)$$

Here, the local pair probability $f_{ij}(Ai,Aj)$ used in MI is replaced by the global pair probability $P_{ij}^{Dir}(A_i, A_j)$. The latter is calculated based on a global probability model using the entropy maximisation approach, which calculates correlation scores for each pair of residues while considering all other pairs [169, 171, 234].

***MItop4mean, DItop4mean, MItop8mean, DItop8mean.*** The coevolution scores (MI or DI) between the residue of interest and all other residues in the TMD were calculated and ranked from highest to lowest. The mean was then calculated for the top-scoring 4 and 8 residue pairs.

***MI1mean, DI1mean, MI3mean, DI3mean, MI4mean, DI4mean.*** Mean coevolution (MI or DI) between two residue pairs, i and i+x, as well as i and i-x, where x represents a distance of 1, 3 or 4 residues.

***MImax, DImax.*** The maximum coevolution value (MI or DI) between the residue of interest and all residues in the TMD.

***MI4max, DI4max.*** The maximum coevolution value (MI or DI) between the residue of interest and the eight neighbouring residue positions (i-4 to i+4).

***MI4cum, DI4cum.*** The coevolution values (MI or DI) between all possible residue pairs in the TMD were measured and sorted from highest to lowest. All unique residues in the top 4 residue pairs were identified. A boolean value was then created, describing whether the residue of interest was among these residues.

**Normalisation of coevolution-based features**

For our data, the mean MI coevolution values were found to decrease with an increasing number of homologues (Figure 5-3). For both MI and DI, the standard deviation of the values decreased with the number of homologues (Figure 5-3). To minimise these effects, we normalised the coevolution features described above between 0 and 1 within each TMD before the application of statistical analyses or machine learning.

**Homologues and evolutionary sequence conservation (3 features)**

***n_homologues.*** The number of homologues in the MSA.

***conservation.*** Conservation was assessed based on the Shannon entropy ($S_{entropy}$) as follows:

$$S_{entropy} = -\sum_{i=1}^{20} p_i \log p_i$$

$$\text{conservation} = -S_{entropy} + 3$$

where $p_i$ represents the observed frequency of amino acid *i* in the given MSA column. Conservation thus takes positive values that increase with a decreasing rate of evolution.

***cons4mean.*** Mean conservation of the three residue positions i, i-4 and i+4.

**Polarity (6 features)**

***polarity.*** Polarity was calculated for each position in the MSA. The PSSM of amino acid frequencies was first adjusted to exclude gaps, ensuring that the sum of the amino acid frequencies was 1. The proportion of each residue type was multiplied by the respective value in the GES (Engelman) hydrophobicity scale [142]. The GES scale was chosen as it offered consistently high performance during THOIPA development and validation. The final polarity score represented the sum of these products for all 20 residues. According to the GES scale, higher values correspond to higher polarity (e.g. positions rich in Lys or Glu).

***relative polarity.*** Polarity at position i divided by the mean polarity of the 6 surrounding residues (i-3 to i+3, excluding i).

***polarity1mean, polarity4mean.*** Mean polarity of three positions i, i-x, and i+x, where x is equal to one, or four.

***polarity3Nmean, polarity3Cmean.*** Mean polarity of the 3 N-terminal and 3 C-terminal residues relative to the residue of interest, respectively.

**Presence or absence of helix interaction motifs (2 features)**

***GxxxG.*** A boolean variable describing whether a given residue participates in a GxxxG motif.

***SmxxxSm.*** A boolean variable describing whether a given residue participates in a (small)xxx(small) motif, with small residues defined as Gly, Ala, Ser or Cys.

**Amino acid and di-peptide composition (27 features)**

***A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, CS, QN, KR, DE, LIV.*** The evolutionary profile (position-specific scoring matrix, PSSM) comprised the fraction of each residue at that position in the multiple sequence alignment (MSA). In addition, the propensities for residue groups with highly similar properties were combined, such as positively charged (KR), negatively charged (DE), strongly polar uncharged (QN), and large aliphatic (LIV) residues. For example, the feature “LIV” is the combined fraction of Leu, Ile, and Val residues at that position in the multiple sequence alignment.

***mass.*** Mass of the amino acid in the TMD of interest, taken from AAindex [236].

***branched.*** A boolean variable indicating whether or not a residue is classified as a β-branched amino acid, according to AAindex [236]. β-branched residues comprised Ile, Val, and Thr.

**Structural properties (2 features)**

***residue depth.*** Relative position of the residue in the TMD, where 1 represents a central residue, and 0 represents either the most N-terminal or C-terminal residue. This was rounded to one significant figure (0, 0.1, 0.2 etc) to prevent our machine learning method from remembering exact residue positions, rather than learning general interface properties.

***n_TMDs.*** The number of TM helices in the full protein, as predicted by Phobius [237]. This variable can take the following values: 0, 1, 2, 3, and 4. For the training dataset of well-studied membrane proteins, the value 0 indicated an erroneous prediction by

Phobius that the sequence encoded a soluble protein. The values 1, 2 and 3 represent the predicted number of TM helices, and 4 represents the prediction of 4 or more TM helices in the protein.

### 5.1.2 Best overlap (BO) validation

During THOIPA development, we aligned our validation method to the goal of predicting the small number of key residues involved in homotypic TMD interactions. This is especially appropriate for ToxR data, where a few key residues were usually mutation-sensitive, and the remaining residues comprise a noisy background. The validation method required the following properties:

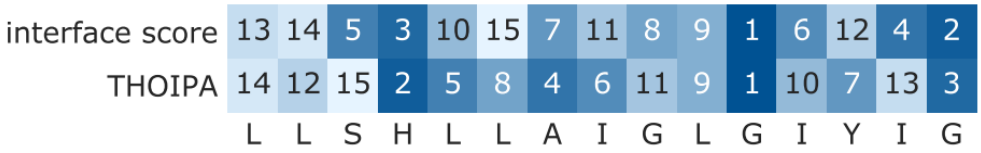
1) to measure performance in identification of top interface residues within a TMD (rather than accuracy for all residues)
2) to indicate the number of top residues at which performance is best
3) to give a measure of individual performance for each TMD
4) to give a measure of overall performance for a dataset that applies equal weights to each TMD, rather than to each residue

Such validation is complicated by the wide range in TMD lengths, which varied from 15 to 29 residues. Standard statistical analyses based on p-values were also unsuitable due to the small sample size in a single TMD (~20 residues). We therefore developed our own validation method that fulfilled all the above requirements. It is based on the overlap of two groups of selected residues from the TMD:

1) the sample of residues corresponding to the top (top 1, top 2, top 3, etc) residues according to experimental data
2) the sample of residues corresponding to the top (top 1, top 2, top 3, etc) residues according to a predictor

The method we refer to here as best overlap (BO) validation has been rigorously tested using randomly generated predictions. It is illustrated by the following example where we validate THOIPA prediction against experimental ToxR data represented by disruption after mutation (interface score). Dark shading indicates high disruption, or high THOIPA score.


We first ranked the experimental and prediction scores, where 1 represented the most important residue for the TMD interaction.




We assessed the overlap in residues between the experimental and prediction data for a particular sample size. The sample size represented the number of “top” residues examined according to the experiment or predictor. For example, at sample size 5, we determined how many of the top 5 residues according to experimental data were among the top 5 predicted by THOIPA. We calculated the observed overlap, and the expected overlap by random chance as follows

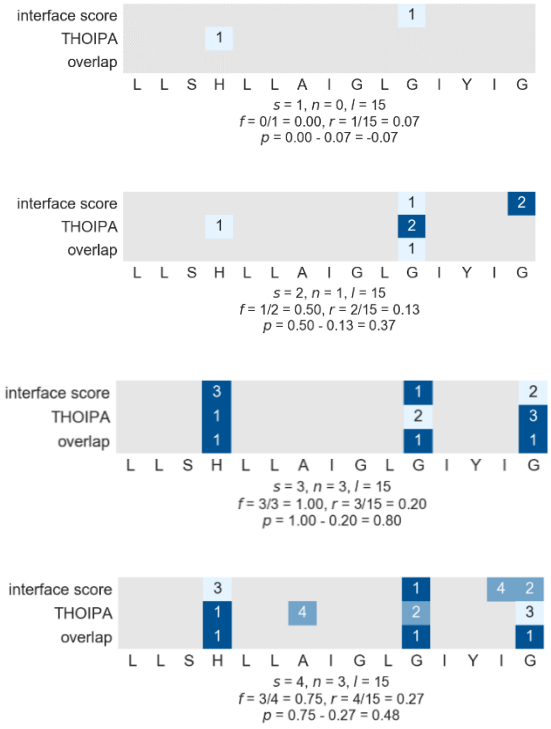
$$f = \frac{n}{s},\ r = \frac{s}{l}$$

where *f* is the observed fraction of overlapping residues, *r* is the expected random fraction of overlapping residues, *n* is the observed number of overlapping residues at that sample

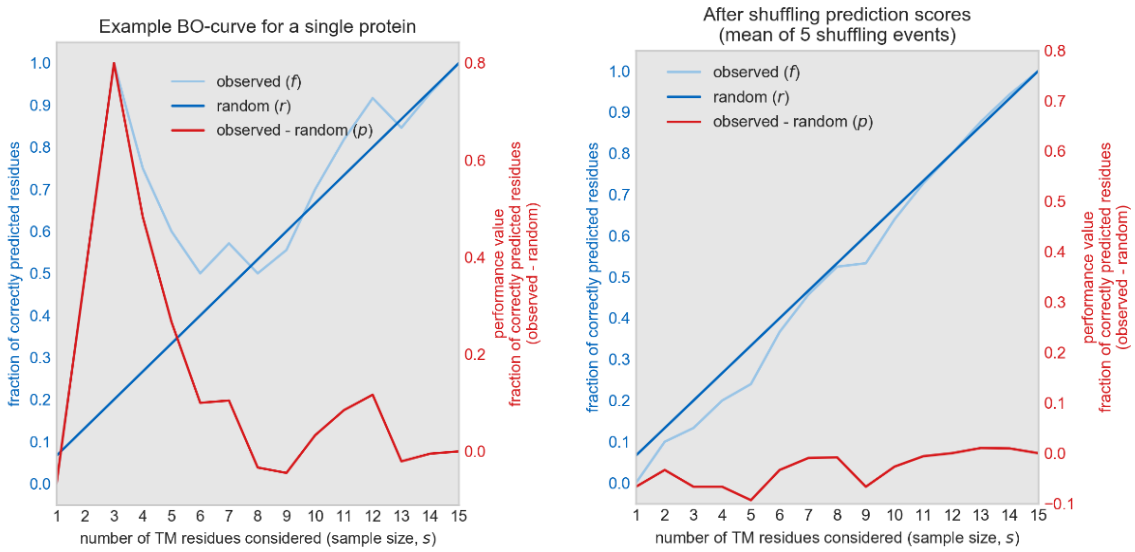
size, *s* is the sample size, and *l* is the length of the TMD. The performance is calculated as follows:

$$p = f - r$$

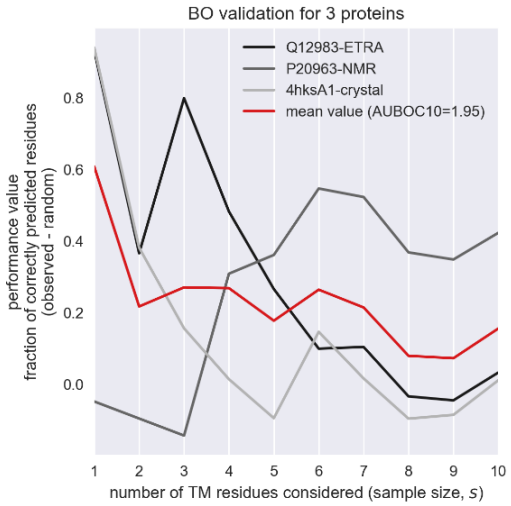
where *p* is simply the fraction of correctly predicted (overlapping) residues, minus the fraction expected by random chance. This is visually demonstrated below for sample sizes 1 to 4, where the “overlap” row represents whether the experiment and predictor have an overlap at that position:




A plot of *f*, *r* and *p* over the entire TMD length of a single protein shows that the random values rise quickly, limiting the possible performance above random.



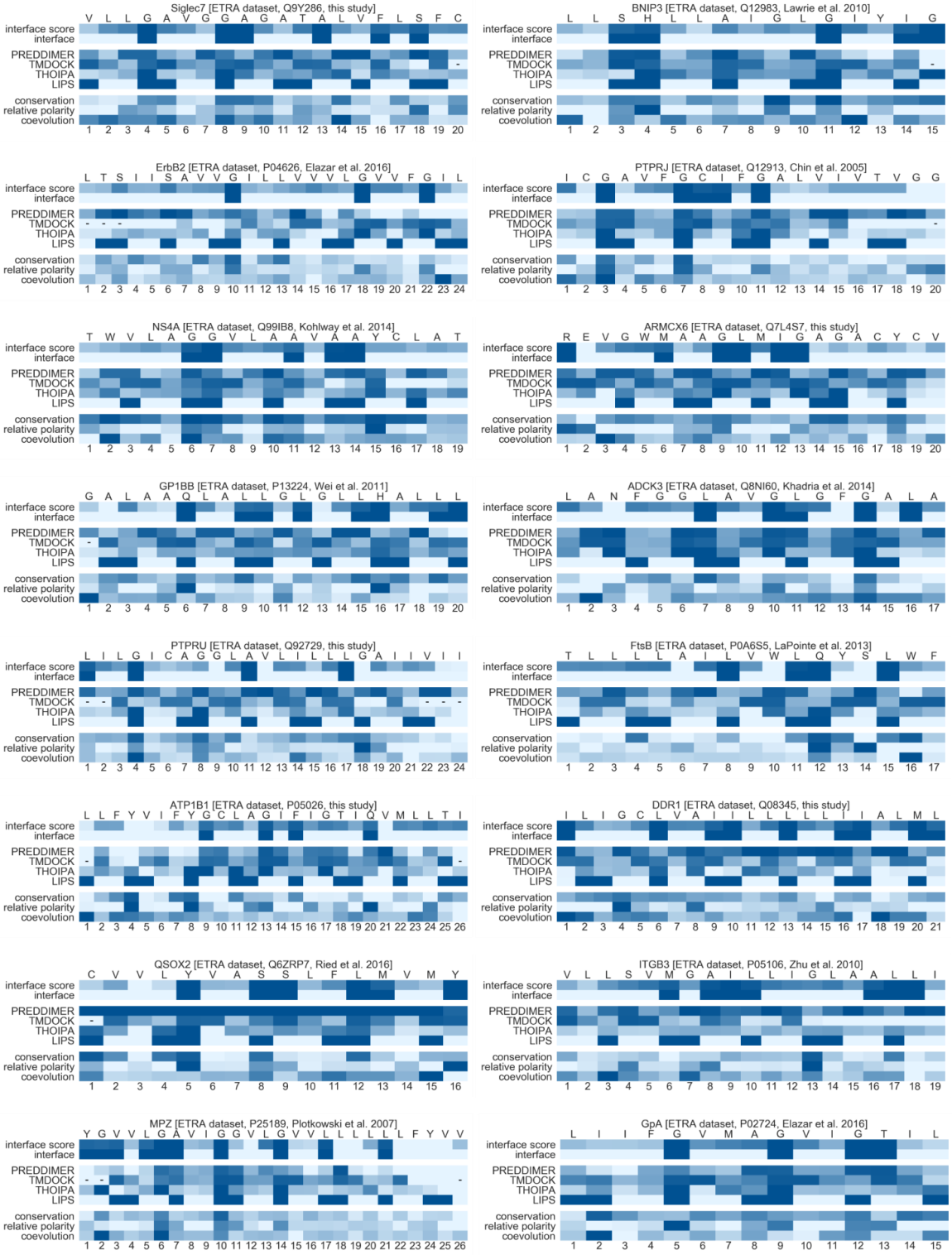

The accuracy of the calculated random scores is confirmed by processing shuffled prediction values (above right). The performance of shuffled predictions clearly conforms to calculated random values. For datasets of multiple proteins, we typically plotted the performance above random, *p* (observed – random), for sample sizes of 1 to 10.

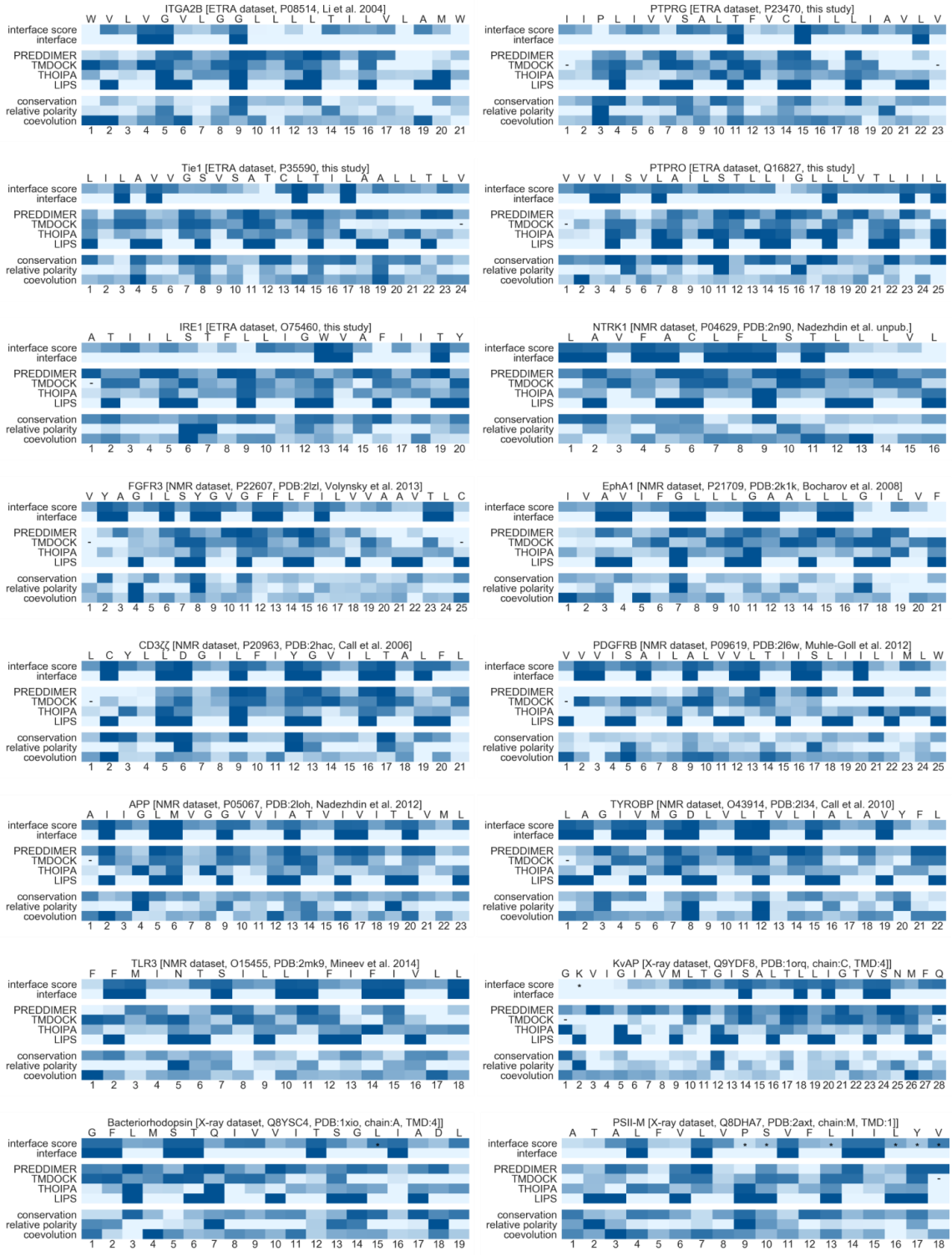


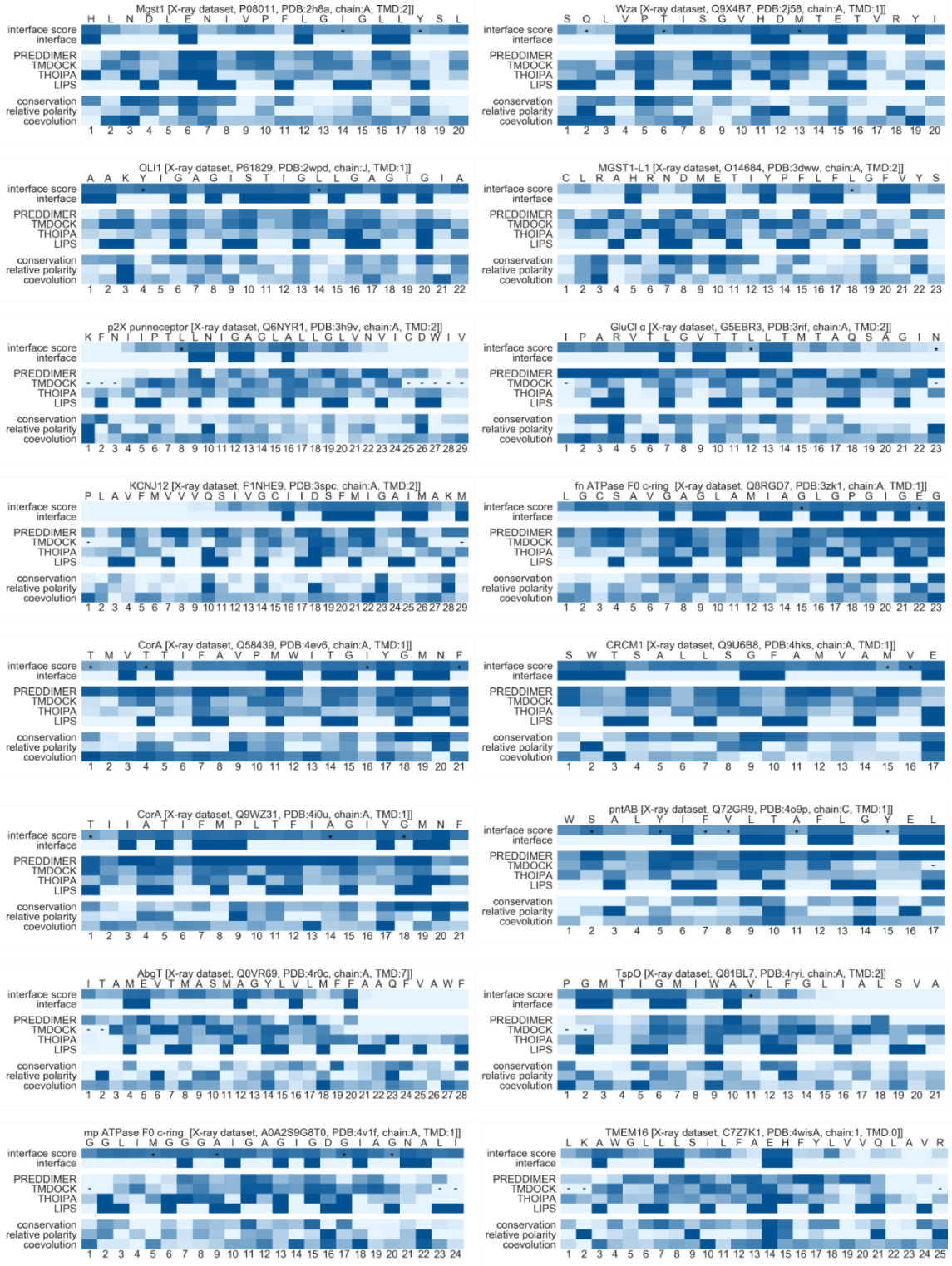

To visualise the overall performance for a dataset of multiple TMDs, we typically calculated the mean performance value for all proteins in the dataset at each sample size. As a performance value for individual TMDs or a dataset, we took the area under the

curve of p for the desired sample size (e.g. area under the BO curve from 1 to 10, AUBOC10). Considering that there were approximately 6.5 interface residues per TMD in our datasets, a smaller sample size (e.g. AUBOC6) is most appropriate. However, we show AUBOC10 values here to avoid bias against the structural predictors TMDOCK and PREDDIMER, whose performance in BO-curves tended to peak at a higher sample size than THOIPA.

## 5.2 Appendix B: Supplementary Figures




ITGA2B [ETRA dataset, P08514, Li et al. 2004]
PTPRG [ETRA dataset, P23470, this study]
Tie1 [ETRA dataset, P35590, this study]
PTPRO [ETRA dataset, Q16827, this study]
IRE1 [ETRA dataset, O75460, this study]
NTRK1 [NMR dataset, P04629, PDB:2n90, Nadezhdin et al. unpub.]
FGFR3 [NMR dataset, P22607, PDB:2lzl, Volynsky et al. 2013]
EphA1 [NMR dataset, P21709, PDB:2k1k, Bocharov et al. 2008]
CD3ζζ [NMR dataset, P20963, PDB:2hac, Call et al. 2006]
PDGFRB [NMR dataset, P09619, PDB:2l6w, Muhle-Goll et al. 2012]
APP [NMR dataset, P05067, PDB:2loh, Nadezhdin et al. 2012]
TYROBP [NMR dataset, O43914, PDB:2l34, Call et al. 2010]
TLR3 [NMR dataset, O15455, PDB:2mk9, Mineev et al. 2014]
KvAP [X-ray dataset, Q9YDF8, PDB:1orq, chain:C, TMD:4]]
Bacteriorhodopsin [X-ray dataset, Q8YSC4, PDB:1xio, chain:A, TMD:4]]
PSII-M [X-ray dataset, Q8DHA7, PDB:2axt, chain:M, TMD:1]]
interface score
interface
PREDDIMER
TMDOCK
THOIPA
LIPS
conservation
relative polarity
coevolution


Mgst1 [X-ray dataset, P08011, PDB:2h8a, chain:A, TMD:2]]
Wza [X-ray dataset, Q9X4B7, PDB:2j58, chain:A, TMD:1]]
OLI1 [X-ray dataset, P61829, PDB:2wpd, chain:J, TMD:1]]
MGST1-L1 [X-ray dataset, O14684, PDB:3dww, chain:A, TMD:2]]
p2X purinoceptor [X-ray dataset, Q6NYR1, PDB:3h9v, chain:A, TMD:2]]
GluCl α [X-ray dataset, G5EBR3, PDB:3rif, chain:A, TMD:2]]
KCNJ12 [X-ray dataset, F1NHE9, PDB:3spc, chain:A, TMD:2]]
fn ATPase F0 c-ring [X-ray dataset, Q8RGD7, PDB:3zk1, chain:A, TMD:1]]
CorA [X-ray dataset, Q58439, PDB:4ev6, chain:A, TMD:1]]
CRCM1 [X-ray dataset, Q9U6B8, PDB:4hks, chain:A, TMD:1]]
CorA [X-ray dataset, Q9WZ31, PDB:4i0u, chain:A, TMD:1]]
pntAB [X-ray dataset, Q72GR9, PDB:4o9p, chain:C, TMD:1]]
AbgT [X-ray dataset, Q0VR69, PDB:4r0c, chain:A, TMD:7]]
TspO [X-ray dataset, Q81BL7, PDB:4ryi, chain:A, TMD:2]]
mp ATPase F0 c-ring [X-ray dataset, A0A2S9G8T0, PDB:4v1f, chain:A, TMD:1]]
TMEM16 [X-ray dataset, C7Z7K1, PDB:4wisA, chain:1, TMD:0]]
interface score
interface
PREDDIMER
TMDOCK
THOIPA
LIPS
conservation
relative polarity
coevolution

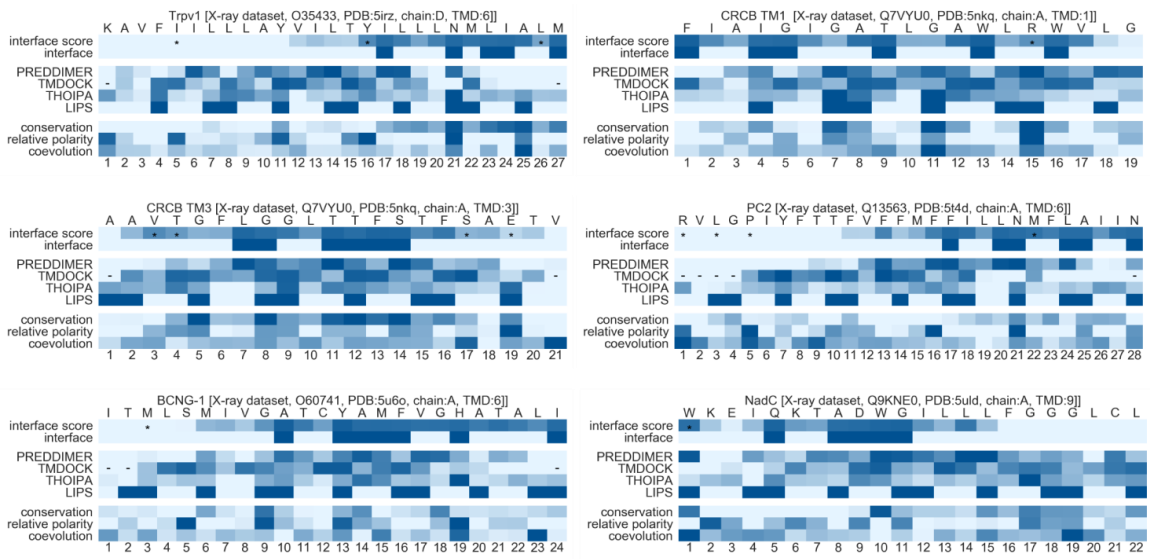



**Figure 5-1**: **A comparison of interface scores (see *Appendix A Supplementary Methods*), designated interface residues, prediction scores and their evolutionary and physical properties.** Darker shading indicates higher values for the interface score (see *Appendix A Supplementary Methods*), interface (boolean value based on >0.24 disruption or <3.5 Å heavy atom distance), PREDDIMER and TMDOCK scores (as calculated from closest heavy-atom distances in the top structures and normalised from 8 Å to 2.5 Å; values of 8 Å and above correspond to 0, and values 2.5 Å and below correspond to 1), THOIPA score (derived from leave-one-out validation and normalised from 0.15 to 0.5), LIPS score (boolean value describing participation in the helix face with the highest conservation and polarity), conservation (normalised from 1.5 to 3), relative polarity (normalised from 0.5 to 2.5) and coevolution (DImax, normalised from 0 to 1). A hyphen (-) indicates TMD positions truncated by TMDOCK, for which there was no structural prediction. For TMDs of the X-ray dataset, a star (*) in the interface score indicates residues that were involved in heterotypic TMD interactions (heterotypic contacts).

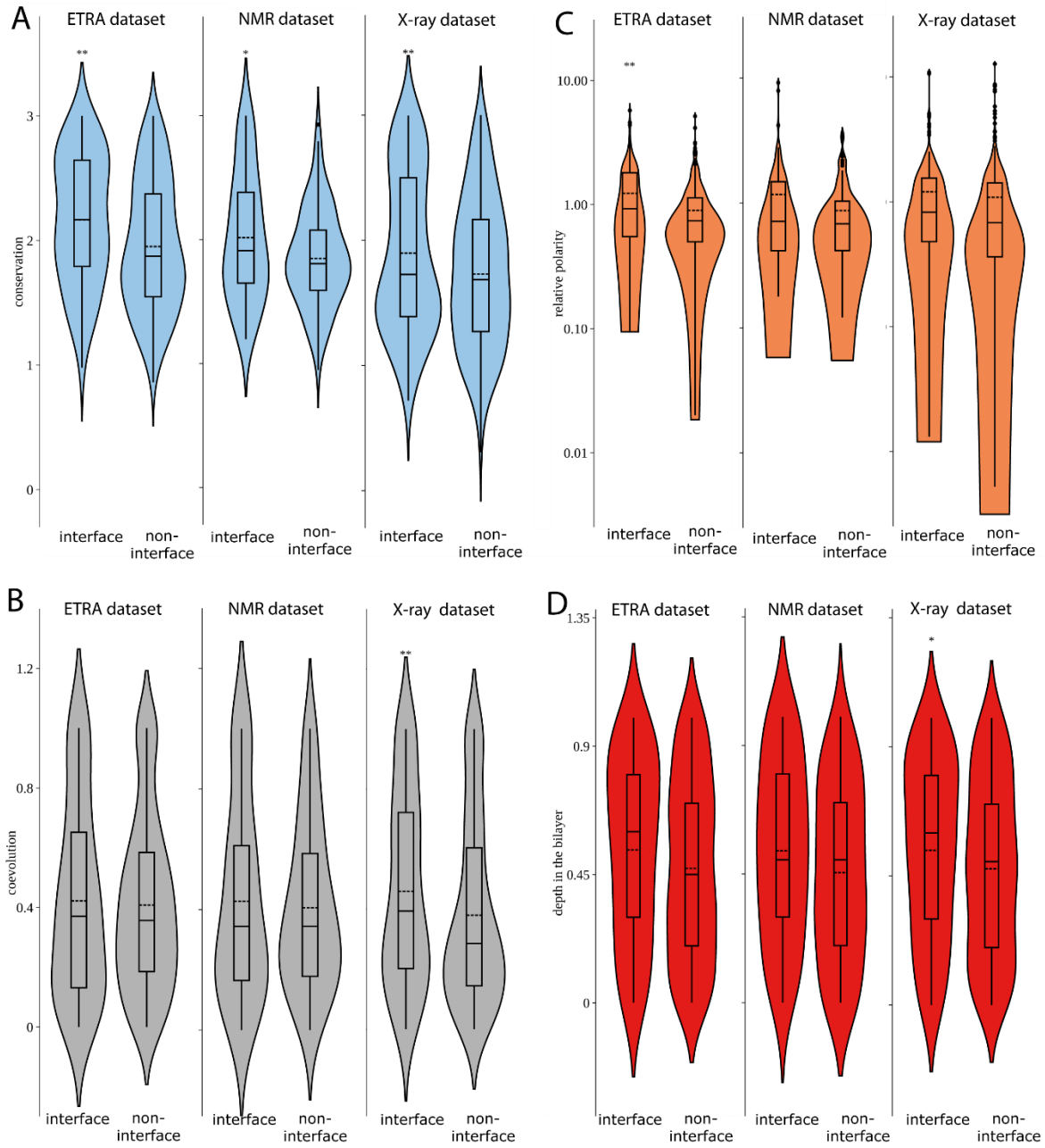



**Figure 5-2**: **Analysis of residue features associated with interface residues, within each dataset separately.**

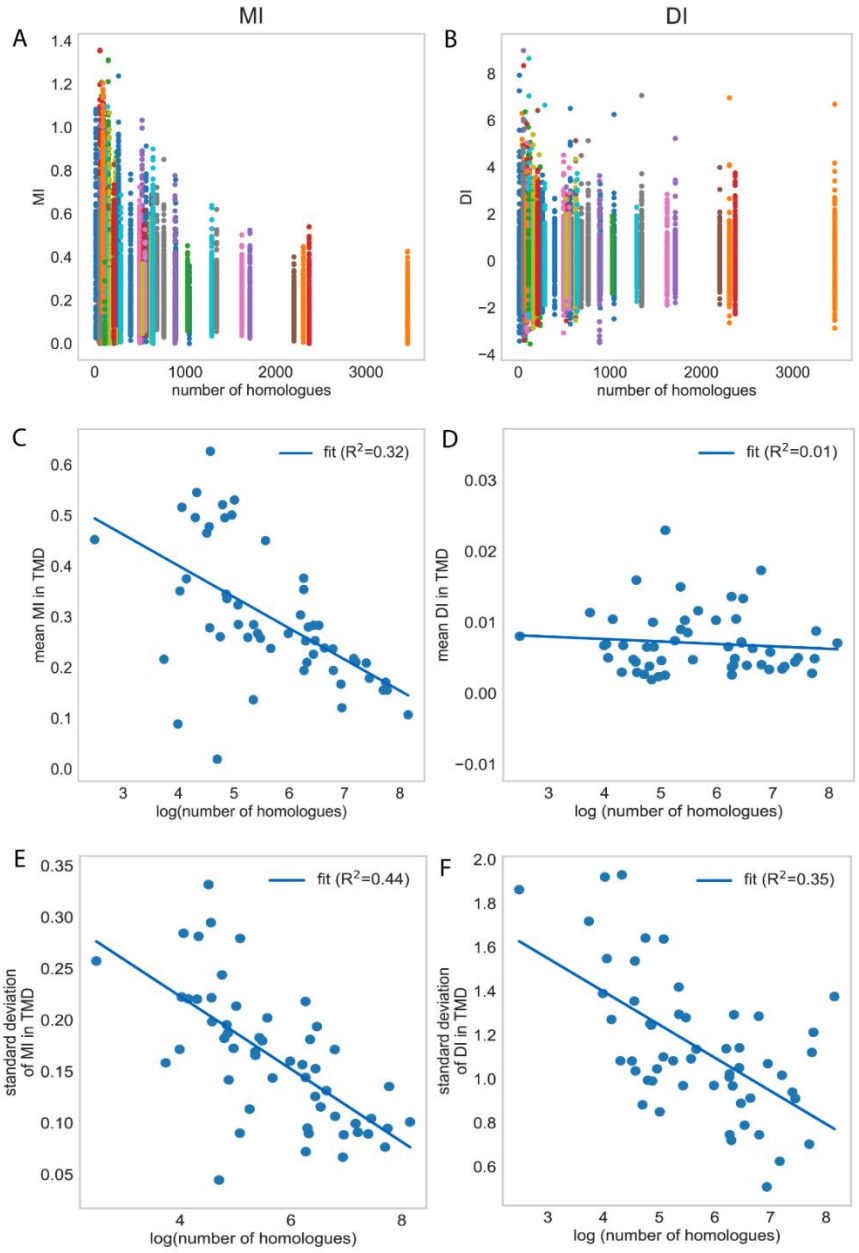



**Figure 5-3: Relationship between coevolution values and the number of homologues.**

This figure justifies the use of normalised coevolution values in the statistical analyses and machine learning. Values were normalised was conducted within each TMD described in the SI Methods. (A, B) Raw data for MI and DI values at each residue position, plotted against the number of homologues. The number of homologues is a discrete value shared by all residues in a TMD. (C, D) The mean MI values within each TMD are negatively correlated to the number of homologues. (E, F) The standard deviation of MI and DI values within each TMD is negatively correlated with the number of homologues. Due to these factors, the normalisation of coevolution values within each TMD was necessary before analysis.

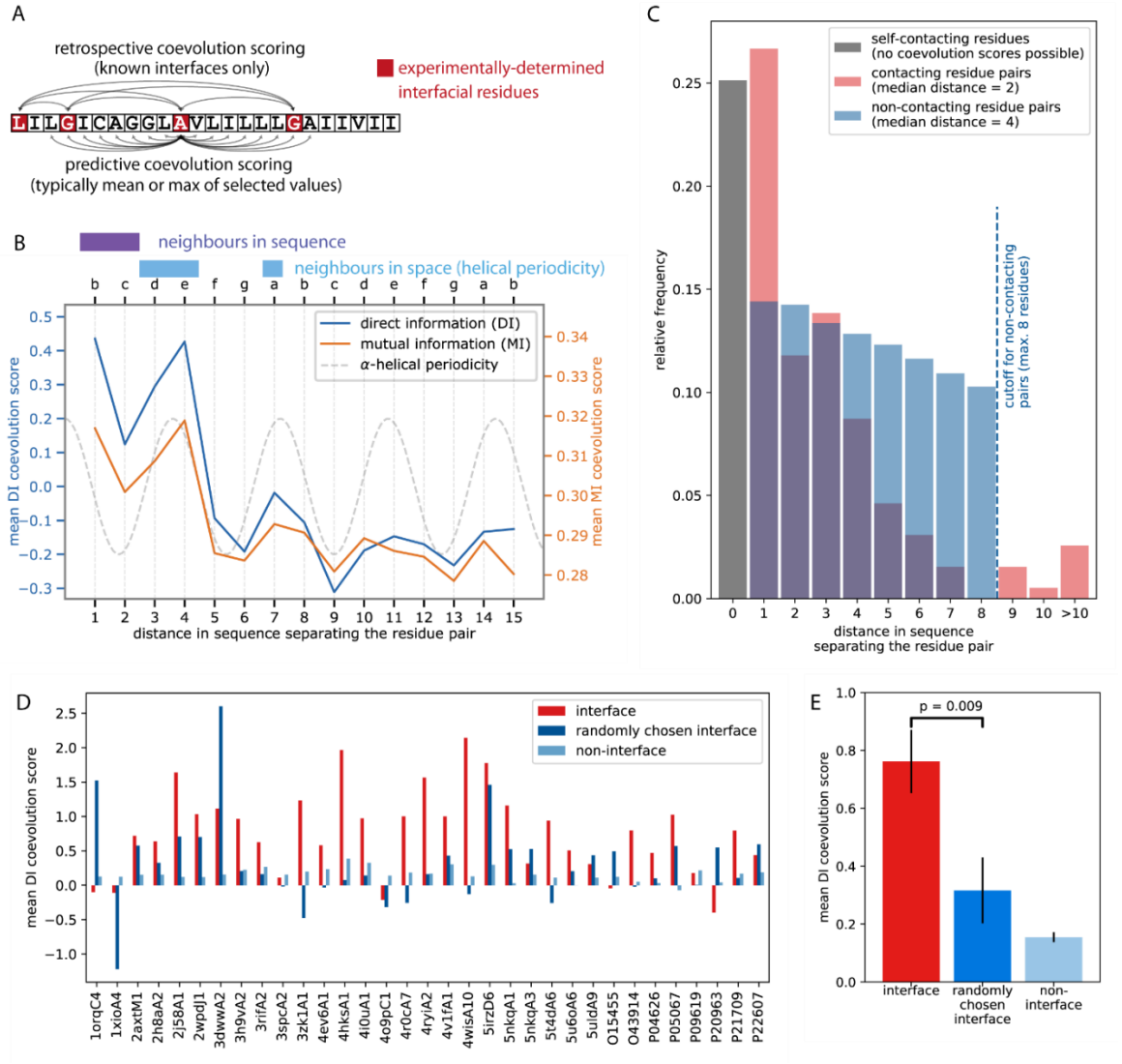



**Figure 5-4: Coevolution of contacting (interface) residues in the NMR and X-ray datasets is biased by the "neighbour effect"**. (A) Schematic illustration comparing the calculation of retrospective and predictive coevolution values. Retrospective methods require a known interface and thus can compare overall coevolution values between interface and non-interface residues. Predictive methods calculate the mean or maximum of pairwise values, involving any residue of interest[172]. A previous retrospective study [56] had proposed that in TM homodimers, coevolution is higher for interface than non-interface residues. This conclusion was drawn from an analysis comparing mean pairwise DI of contacting residue pairs to the mean pairwise DI of non-contacting residue pairs. In the previous study, non-contacting residues separated by up to 8 residues in sequence were considered. When we applied this method to our own NMR and X-ray dimers, we noticed a few effects and biases that compromised the retrospective evolution approach. (B) The neighbour effect. Coevolution scores are intrinsically higher for residues being close

in the protein sequence (i.e., at distances of 1 or 2), or close in space at α-helical periodicity. The relationship between the separation distance and coevolution scores was calculated using the MI and DI scores from the homotypic TMD dataset. The helical pattern followed by both scores shows good correspondence to that previously described for soluble α-helices [173]. Residues are aligned with the classical heptad motif (abcdefg), where a is the reference residue. Neighbours in sequence of the residue of interest, a, occupy positions b (direct neighbourhood) and c (separated by one residue). Direct neighbours of a in space comprise positions d and e, assuming perfect α-helicity. (C) There is a clear difference in the distribution of the distances separating contacting (interface) or non-contacting (non-interface) residue pairs. We noticed that the number of contacting residues was much smaller than the number of non-contacting residues. Thus, the distance in the sequence between contacting residues (median=2) was half the distance between non-contacting residues (median=4). Since the coevolution values are dependent on residue distance (part B) shorter distances between interface residues artificially raise coevolution values of interface residues. Interface residues generally appear more co-evolved, even for dimer configurations that are not found in the organism of interest. D) Mean DI coevolution values for all TMDs in the NMR and X-ray datasets. The mean DI was calculated between all contacting (=interface) and non-contacting (=non-interface) residue pairs as previously described [56]. For each TMD, we also calculated the score for a randomly chosen interface after substituting its contacting amino acids by contacting residue positions of an unrelated TMD. E) Mean values of the data shown in part D. Comparing original and randomly chosen interfaces shows that a proportion of the higher coevolution scores in the original interfaces is due to the fact that average distances between contacting residue pairs are smaller than those of non-contacting ones, i.e., the neighbour effect. Retrospective analyses comparing coevolution of interface and non-interface residues (compare red and light blue bars) are therefore biased. The comparison of real and randomly chosen interfaces leads to a reduction of this bias. However, with only 33 NMR and X-ray TMDs available, it is not certain if such randomisation can completely remove the bias inherent in retrospective analyses. In comparison, analyses based on predictive coevolution measures (**Figure 2-2**, **Figure 5-2**, **Table 2-2**) offer a far more rigorous comparison of residue properties between interface and non-interface residues.

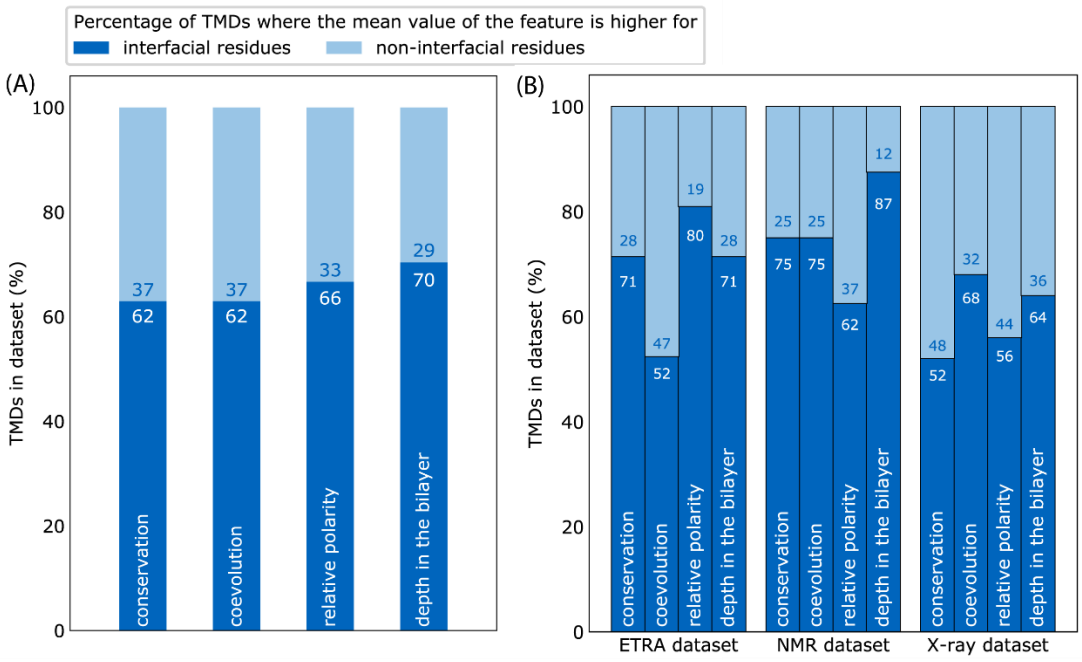



**Figure 5-5: Individual TMDs have unique structural requirements, leading to high variability in residue properties of interfaces.** (A) Percentage of TMDs of the homotypic TMD dataset (ETRA, NMR and X-ray) where the mean value of a given residue property is higher for interface or non-interface residues, respectively. (B) Values as in part A, but calculated for each dataset separately.

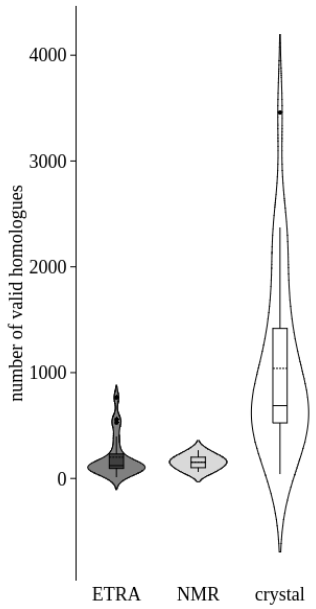



**Figure 5-6: Number of valid homologues for TMDs of each dataset.** The mean number of homologues was 201, 154, and 1040 for the ETRA, NMR and crystal datasets respectively. Filtering and redundancy reduction of homologues. Violin plots were constructed from the data as described in Figure 2-2.

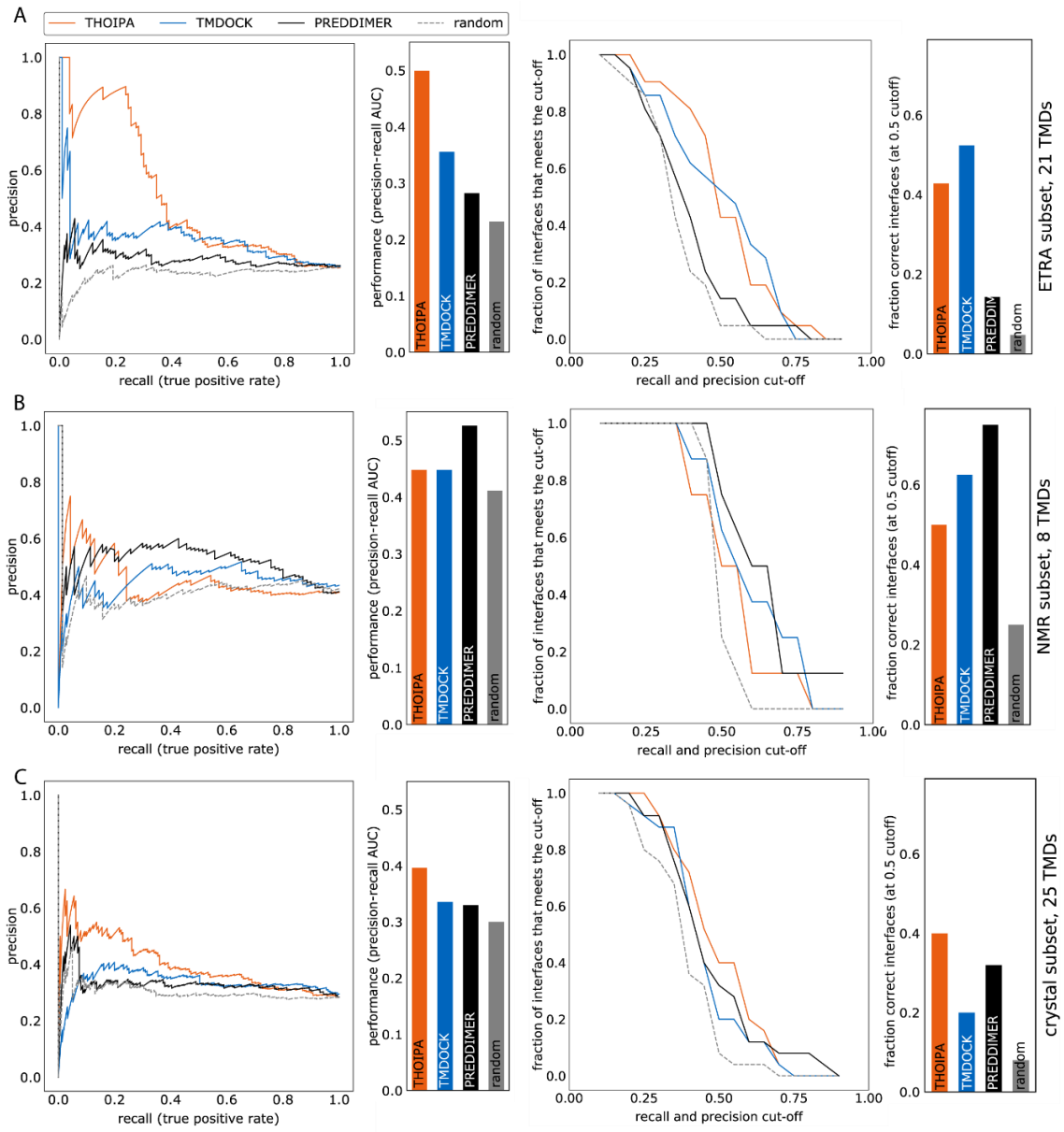



**Figure 5-7: Validation of THOIPA performance towards the ETRA, NMR and X-ray datasets.** The underlying residue predictions are identical to those used in **Fig. 6**, however datasets were validated separately. For each dataset, the precision-recall curve (and AUC barchart) is shown on the left, and the fraction of correctly predicted interfaces (and associated barchart at cut-off = 0.5) is shown on the right. (A) ETRA TMDs. (B) NMR TMDs. (C) X-ray TMDs.

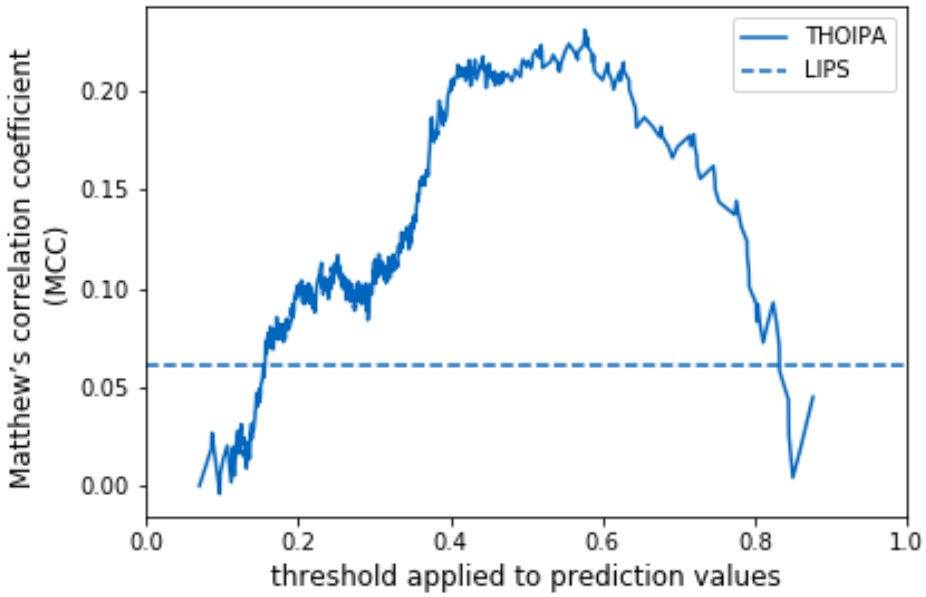



**Figure 5-8: Comparison of THOIPA and LIPS performance.** Validation is shown for the full TMD homodimer dataset. LIPS gives binary prediction results (interface or non-interface) that could not be analysed with precision-recall curves. Instead, LIPS predictions were validated against the interface residues from experimental data using the Matthews correlation coefficient (MCC). Higher values indicate stronger prediction. The MCC for LIPS was 0.06 (dotted line). For THOIPA, the MCC depended on the chosen threshold. For most THOIPA thresholds (0.2 to 0.8), the THOIPA performance is superior to LIPS. The maximum MCC achieved by THOIPA is 0.23, at a threshold of 0.58. Note that this threshold is far higher than the average THOIPA prediction (0.31, due to machine learning with 31% of the residues classified as interface residues). The

MCC analysis therefore provides further evidence that THOIPA performance is best at high thresholds, corresponding to a small number of key interface residues.

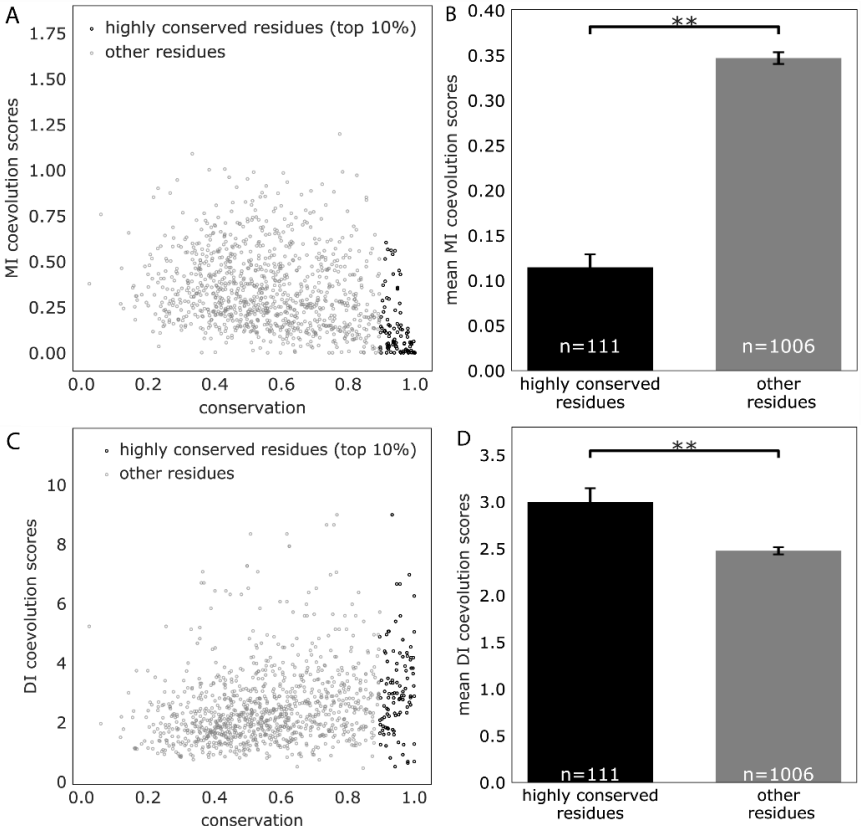



**Figure 5-9: Highly conserved residues are associated with low MI and high DI coevolution scores.** For each residue in the homotypic TMD dataset, conservation was plotted against coevolution, represented by MI4mean and DI4mean (see *SI Methods*), before normalisation of values within each TMD. (A) Scatterplot of MI values against conservation. Highly conserved residues (top 10%) are shown in a darker colour. (B) Barchart comparing the most highly conserved residues with all other residues, in respect to MI4mean values. (C) Scatterplot of DI values against conservation. (D) Barchart comparing the most highly conserved residues with all other residues, in respect to DI4mean values. Means ± SEM. Statistical significance was tested using a bootstrapped t-test.

## 5.3 Appendix C: Supplementary Tables

**Table 5-1: Accession and reference for TMDs with known NMR structures**

| PDB[a] | protein (acc[b]) | reference |
|---|---|---|
| 1afo | GpA [P02724] | [1] |
| 2hac | CD3ζζ [P20963] | [2] |
| 2j5d | BNIP3 [Q12983] | [3] |
| 2jwa | ErbB2 [P04626] | [4] |
| 2k1k | EphA1 [P21709] | [5] |
| 2l34 | TYROBP [O43914] | [6] |
| 2k9y | EphA2 [P29317] | [7] |
| 2l9u | ErbB3 [P21860] | [8] |
| 2loh | APP [P05067] | [9] |
| 2l6w | PDGFRB [P09619] | [10] |
| 2lcx | ErbB4 [Q15303] | [11] |
| 2m0b | EGFR [P00533] | [12] |
| 2lzl | FGFR3 [P22607] | [13] |
| 2mk9 | TLR3 [O15455] | [14] |
| 2n90 | NTRK1 [P04629] | Nadezhdin et al. unpublished |

[a] Accession number (PDB) is taken from the PDB database.
[b] Accession number (acc) is taken from the UniProt database.

**Table 5-2: Composition of the CompData dataset.** For 101 α-helical transmembrane proteins PDB IDs and chain names are given.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1FFT:A | 1FFT:B | 1FFT:C | 1H2S:A | 1JB0:A | 1JB0:F | 1JB0:I | 1JB0:K | 1JB0:M | 1JB0:X |
| 1KF6:C | 1KF6:D | 1KQF:B | 1KQF:C | 1LGH:A | 1LGH:B | 1LNQ:A | 1M56:B | 1M56:C | 1M56:D |
| 1NEK:C | 1NEK:D | 1NKZ:A | 1OTS:A | 1Q16:C | 1Q90:A | 1Q90:B | 1Q90:G | 1Q90:N | 1Q90:R |
| 1RH5:A | 1RH5:B | 1RZH:L | 1RZH:M | 1S5L:B | 1S5L:C | 1S5L:D | 1S5L:E | 1S5L:F | 1S5L:H |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1S5L:I | 1S5L:J | 1S5L:K | 1S5L:L | 1S5L:M | 1S5L:T | 1S5L:X | 1S5L:Z | 1V54:D | 1V54:G |
| 1V54:I | 1V54:J | 1V54:K | 1V54:L | 1V54:M | 1VF5:B | 1VF5:D | 1VF5:F | 1VF5:G | 1VF5:H |
| 1XL4:A | 1XME:A | 1XME:B | 1YEW:A | 1YEW:B | 1YEW:C | 1ZCD:A | 2BHW:A | 2BL2:A | 2BS2:C |
| 2FYU:E | 2FYU:G | 2FYU:K | 2H88:C | 2H88:D | 2HYD:A | 2IH3:C | 2IUB:A | 2J58:A | 2J8S:A |
| 2NQ2:A | 2NWL:A | 2O01:G | 2O01:H | 2O01:I | 2O01:J | 2O01:L | 2OAR:A | 2Q67:A | 2QTS:A |
| 2R6G:F | 2RDD:B | 2VL0:A | 2VV5:A | 2YVX:A | 3CX5:C | 3CX5:D | 3CX5:H | 3CX5:I | 3D31:C |
| 3EAM:A | | | | | | | | | |

**Table 5-3: Composition of the ClassData dataset.** For 171 $\alpha$-helical transmembrane proteins PDB IDs and chain names are given.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1E7P:F | 1FFT:B | 1FFT:C | 1FX8:D | 1H2S:A | 1JB0:A | 1JB0:F | 1JB0:L | 1KQG:E | 1KQG:F |
| 1L0V:C | 1L0V:D | 1NTK:C | 1NTK:D | 1NTK:E | 1NTK:G | 1NTK:J | 1NTK:K | 1P84:T | 1P8I:A |
| 1RWT:F | 1SIW:F | 1T9W:A | 1ZCD:A | 2ACZ:G | 2B6P:D | 2FBW:C | 2FBW:D | 2J7A:F | 2NUU:A |
| 2NWL:A | 2OAR:D | 2OCC:A | 2OCC:B | 2OCC:C | 2OCC:D | 2OCC:G | 2OCC:I | 2ONK:D | 2R9R:D |
| 2VPY:G | 2VV5:D | 2WX5:H | 2WX5:L | 2YEV:B | 2ZW3:D | 3B9W:A | 3HD7:A | 3HD7:B | 3I5D:A |
| 3KCU:D | 3KDP:D | 3M6E:A | 3MP7:B | 3NCY:B | 3NE5:A | 3ODU:A | 3OR6:D | 3QNQ:A | 3QS4:A |
| 3RKO:J | 3RKO:L | 3RKO:M | 3SYP:D | 3T9N:D | 3TDR:D | 3TIJ:A | 3UX4:A | 3WME:A | 3WU2:C |
| 4A01:A | 4BEM:J | 4BPM:A | 4BRB:A | 4BW5:A | 4C9Q:A | 4CZ9:A | 4DXW:D | 4EV6:D | 4FTP:A |
| 4FZ0:A | 4GD3:A | 4GD3:T | 4GPO:A | 4GX0:E | 4H1D:A | 4H1W:A | 4HE8:D | 4HEA:A | 4HG6:B |
| 4HKS:E | 4HUQ:T | 4I0U:D | 4IFF:A | 4J72:A | 4KHZ:F | 4KHZ:G | 4KJS:D | 4LLH:A | 4LMK:D |
| 4LP8:D | 4LTO:D | 4M8J:A | 4MND:A | 4N4Y:B | 4N7W:A | 4O93:A | 4O93:D | 4OR2:A | 4P6V:B |
| 4P6V:C | 4P6V:E | 4P6V:F | 4PHZ:E | 4PHZ:F | 4PHZ:G | 4PIR:A | 4PJ0:B | 4PJ0:D | 4PL0:A |
| 4PV1:A | 4PV1:C | 4PV1:D | 4PXF:A | 4Q4A:A | 4Q7C:A | 4QNC:A | 4QTN:A | 4R0C:A | 4R6Z:D |
| 4R9U:A | 4RDQ:D | 4RFS:S | 4RI2:A | 4TL3:A | 4TNW:D | 4TPJ:A | 4TQU:M | 4TQU:N | 4TSY:A |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4U4W:A | 4U5B:D | 4U9N:A | 4WD7:D | 4WGW:A | 4WIS:A | 4XK8:H | 4XU5:A | 4Y7K:D | 4YCR:A |
| 4YMK:D | 4YMS:D | 4YTM:C | 4YZF:A | 4YZI:A | 4Z7F:A | 4Z90:D | 5A43:A | 5A63:A | 5A63:B |
| 5A63:C | 5A63:D | 5BW8:C | 5C3L:A | 5C3L:B | 5C3L:C | 5C65:B | 5C78:D | 5DJQ:B | 5DJQ:C |
| 5EKE:D | | | | | | | | | |

**Table 5-4: Composition of the independent TestData dataset.** For 36 $\alpha$-helical transmembrane proteins PDB IDs and chain names are given.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3JCU:H | 3JCU:I | 3JCU:T | 3JCU:W | 3JCU:X | 4Y28:K | 5AZD:A | 5B0W:A | 5B1A:J | 5B1A:K |
| 5B1A:L | 5B1A:M | 5B57:A | 5B5E:A | 5B5E:M | 5B5E:T | 5B5E:Z | 5BN2:A | 5BQG:A | 5C2T:D |
| 5DJQ:N | 5EG1:A | 5EIY:A | 5FL7:K | 5HV9:A | 5I32:A | 5JJE:B | 5JNQ:A | 5KAF:Y | 5L22:B |
| 5MKK:A | 5MRW:C | 5MRW:D | 5UL7:A | 5X3X:Q | 5X5Y:G | | | | |

# CHAPTER 6. LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| TM | transmembrane |
| TMPs | transmembrane proteins |
| PPI | protein-protein interaction |
| AUC | area under curve |
| SASA | solvent accessible surface area |
| RSA | relative solvent accessibility |
| Mem | transmembrane |
| Cyto | cytoplasmic segment |
| Extra | extra-cellular segment |
| MSA | multiple sequence alignment |
| RF | Random Forest |
| PSSM | position-specific scoring matrices |
| MI | mutual information |
| DI | direct interacting |
| Rp | relative position |
| Y2H | Yeast two hybrid |
| TAP | Tandem affinity purification |
| ETRA | *E. coli* TM reporter assay |
| GPCR | G-protein coupled receptor |
| GpA | glycophorin A |
| IPTG | isopropyl β-D-1-thiogalactopyranoside |
| lacZ | gene coding for β-galactosidase |
| LB | lysogeny broth |
| malE | maltose binding protein E (gene encoding MBP) |
| MBP | maltose binding protein |
| MAM | meprin, A-5 protein, and receptor protein-tyrosine phosphatase mu |
| NMR | nuclear magnetic resonance |
| PDB | protein data bank |
| PDBTM | Protein Data Bank of Transmembrane Proteins |
| TrkC | receptor tropomyosin-related kinase C |

| | |
|---|---|
| RPTPs | receptor-like protein tyrosine phosphatases |
| SDS | sodium dodecyl sulfate |
| TMD | transmembrane domain |
| TM | transmembrane |

# CHAPTER 7. PUBLICATIONS ARISING FROM THIS THESIS

Yao Xiao[‡], Bo Zeng[‡], Nicola Berner, Dmitrij Frishman, Dieter Langosch, and Mark George Teese

(submitted) Properties and prediction of homotypic transmembrane helix-helix interfaces.

[‡]*co-first-authorship. The authors contributed equally to this work.*

Bo Zeng, Peter Hönigschmid, and Dmitrij Frishman

(submitted) Prediction of interaction sites in alpha-helical membrane proteins.

# REFERENCES


1. MacKenzie, K.R., J.H. Prestegard, and D.M. Engelman, *A transmembrane helix dimer: structure and implications.* Science, 1997. **276**(5309): p. 131-133.
2. Call, M.E., et al., *The structure of the ζζ transmembrane dimer reveals features essential for its assembly with the T cell receptor.* Cell, 2006. **127**(2): p. 355-68.
3. Bocharov, E.V., et al., *Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger.* J Biol Chem, 2007. **282**(22): p. 16256-66.
4. Bocharov, E.V., et al., *Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state.* Journal of Biological Chemistry, 2008. **283**(11): p. 6950-6956.
5. Bocharov, E.V., et al., *Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1.* J Biol Chem, 2008. **283**(43): p. 29385-95.
6. Call, M.E., K.W. Wucherpfennig, and J.J. Chou, *The structural basis for intramembrane assembly of an activating immunoreceptor complex.* Nature immunology, 2010. **11**(11): p. 1023.
7. Bocharov, E.V., et al., *Left-handed dimer of EphA2 transmembrane domain: helix packing diversity among receptor tyrosine kinases.* Biophysical journal, 2010. **98**(5): p. 881-889.
8. Mineev, K., et al., *Spatial structure and dimer–monomer equilibrium of the ErbB3 transmembrane domain in DPC micelles.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2011. **1808**(8): p. 2081-2088.
9. Nadezhdin, K.D., et al., *Dimeric structure of transmembrane domain of amyloid precursor protein in micellar environment.* FEBS letters, 2012. **586**(12): p. 1687-1692.
10. Muhle-Goll, C., et al., *Hydrophobic matching controls the tilt and stability of the dimeric platelet-derived growth factor receptor (PDGFR) β transmembrane segment.* Journal of Biological Chemistry, 2012. **287**(31): p. 26178-26186.
11. Bocharov, E.V., et al., *Structural and thermodynamic insight into the process of "weak" dimerization of the ErbB4 transmembrane domain by solution NMR.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2012. **1818**(9): p. 2158-2170.
12. Endres, N.F., et al., *Conformational coupling across the plasma membrane in activation of the EGF receptor.* Cell, 2013. **152**(3): p. 543-556.
13. Bocharov, E.V., et al., *Structure of FGFR3 transmembrane domain dimer: implications for signaling and human pathologies.* Structure, 2013. **21**(11): p. 2087-2093.

14. Mineev, K.S., S.A. Goncharuk, and A.S. Arseniev, *Toll‐like receptor 3 transmembrane domain is able to perform various homotypic interactions: An NMR structural study.* FEBS letters, 2014. **588**(21): p. 3802-3807.

15. Jayasinghe, S., K. Hristova, and S.H. White, *MPtopo: A database of membrane protein topology.* Protein Sci, 2001. **10**(2): p. 455-8.

16. White, S.H., *The progress of membrane protein structure determination.* Protein Sci, 2004. **13**(7): p. 1948-9.

17. White, S.H. and G. von Heijne, *The machinery of membrane protein assembly.* Curr Opin Struct Biol, 2004. **14**(4): p. 397-404.

18. von Heijne, G. and D. Rees, *Membranes: reading between the lines.* Curr Opin Struct Biol, 2008. **18**(4): p. 403-5.

19. Singer, S.J. and G.L. Nicolson, *The fluid mosaic model of the structure of cell membranes.* Science, 1972. **175**(4023): p. 720-31.

20. Liu, J. and B. Rost, *Comparing function and structure between entire proteomes.* Protein Science, 2001. **10**(10): p. 1970-1979.

21. Käll, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* Journal of molecular biology, 2004. **338**(5): p. 1027-1036.

22. Hubert, P., et al., *Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye?* Cell adhesion & migration, 2010. **4**(2): p. 313-324.

23. Wallin, E. and G. von Heijne, *Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.* Protein Sci, 1998. **7**(4): p. 1029-38.

24. Fink, A., et al., *Transmembrane domains interactions within the membrane milieu: principles, advances and challenges.* Biochim Biophys Acta, 2012. **1818**(4): p. 974-83.

25. Hantgan, R.R., et al., *Ligand binding promotes the entropy-driven oligomerization of integrin alpha IIb beta 3.* J Biol Chem, 2003. **278**(5): p. 3417-26.

26. Therien, A.G. and C.M. Deber, *Oligomerization of a peptide derived from the transmembrane region of the sodium pump gamma subunit: effect of the pathological mutation G41R.* J Mol Biol, 2002. **322**(3): p. 583-50.

27. Wu, Y., Q. Li, and X.Z. Chen, *Detecting protein-protein interactions by Far western blotting.* Nat Protoc, 2007. **2**(12): p. 3278-84.

28. Cuatrecasas, P., *Protein purification by affinity chromatography. Derivatizations of agarose and polyacrylamide beads.* J Biol Chem, 1970. **245**(12): p. 3059-65.

29. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 5): p. 905-21.

30. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.

31. Puig, O., et al., *The tandem affinity purification (TAP) method: a general procedure of protein complex purification.* Methods, 2001. **24**(3): p. 218-29.

32. Kaelin, W.G., Jr., *The concept of synthetic lethality in the context of anticancer therapy.* Nat Rev Cancer, 2005. **5**(9): p. 689-98.

33. Jansen, R., D. Greenbaum, and M. Gerstein, *Relating whole-genome expression data with protein-protein interactions.* Genome Res, 2002. **12**(1): p. 37-46.

34. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.

35. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.

36. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

37. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nat Biotechnol, 1999. **17**(10): p. 1030-2.

38. Rath, A., R.M. Johnson, and C.M. Deber, *Peptides as transmembrane segments: decrypting the determinants for helix-helix interactions in membrane proteins.* Biopolymers, 2007. **88**(2): p. 217-32.

39. Langosch, D., et al., *Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator.* J Mol Biol, 1996. **263**(4): p. 525-30.

40. Miller, V.L., R.K. Taylor, and J.J. Mekalanos, *Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein.* cell, 1987. **48**(2): p. 271-279.

41. Kolmar, H., et al., *Membrane insertion of the bacterial signal transduction protein ToxR and requirements of transcription activation studied by modular replacement of different protein substructures.* EMBO J, 1995. **14**(16): p. 3895-904.

42. Gerber, D. and Y. Shai, *In vivo detection of hetero-association of glycophorin-A and its mutants within the membrane.* J Biol Chem, 2001. **276**(33): p. 31229-32.

43. Russ, W.P. and D.M. Engelman, *TOXCAT: A measure of transmembrane helix association in a biological membrane.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(3): p. 863.

44. Elazar, A., et al., *Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane.* Elife, 2016. **5**.

45. Schneider, D. and D.M. Engelman, *GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane.* Journal of Biological Chemistry, 2003. **278**(5): p. 3105-3111.

46. Sawma, P., et al., *Evidence for new homotypic and heterotypic interactions between transmembrane helices of proteins involved in receptor tyrosine kinase and neuropilin signaling.* J Mol Biol, 2014. **426**(24): p. 4099-111.

47. Steindorf, D. and D. Schneider, *In vivo selection of heterotypically interacting transmembrane helices: Complementary helix surfaces, rather than conserved interaction motifs, drive formation of transmembrane hetero-dimers.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2017. **1859**(2): p. 245-256.

48. Su, P.-C. and B.W. Berger, *Identifying key juxtamembrane interactions in cell membranes using AraC-based transcriptional reporter assay (AraTM).* Journal of Biological Chemistry, 2012. **287**(37): p. 31515-31526.

49. Schanzenbach, C., et al., *Identifying ionic interactions within a membrane using BLaTM, a genetic tool to measure homo-and heterotypic transmembrane helix-helix interactions.* Scientific reports, 2017. **7**: p. 43476.

50. Lemmon, M.A., et al., *Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices.* Journal of Biological Chemistry, 1992. **267**(11): p. 7683-7689.

51. Sulistijo, E.S. and K.R. MacKenzie, *Sequence dependence of BNIP3 transmembrane domain dimerization implicates side-chain hydrogen bonding and a tandem GxxxG motif in specific helix–helix interactions.* Journal of molecular biology, 2006. **364**(5): p. 974-990.

52. Lawrie, C.M., E.S. Sulistijo, and K.R. MacKenzie, *Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes.* J Mol Biol, 2010. **396**(4): p. 924-36.

53. Sulistijo, E.S. and K.R. MacKenzie, *Structural basis for dimerization of the BNIP3 transmembrane domain.* Biochemistry, 2009. **48**(23): p. 5106-5120.

54. Bugge, K., K. Lindorff‐Larsen, and B.B. Kragelund, *Understanding single‐pass transmembrane receptor signaling from a structural viewpoint—what are we missing?* The FEBS journal, 2016. **283**(24): p. 4424-4451.

55. Valley, C.C., A.K. Lewis, and J.N. Sachs, *Piecing it together: Unraveling the elusive structure-function relationship in single-pass membrane receptors.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2017. **1859**(9): p. 1398-1416.

56. Wang, Y. and P. Barth, *Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy.* Nature Communications, 2015. **6**: p. 7196.

57. Polyansky, A.A., et al., *PREDDIMER: a web server for prediction of transmembrane helical dimers.* Bioinformatics, 2013. **30**(6): p. 889-890.

58. Lomize, A.L. and I.D. Pogozheva, *TMDOCK: an energy-based method for modeling α-helical dimers in membranes.* J Mol Biol, 2017. **429**(3): p. 390-398.

59. Bocharov, E.V., et al., *Structure elucidation of dimeric transmembrane domains of bitopic proteins.* Cell adhesion & migration, 2010. **4**(2): p. 284-298.

60. Tusnády, G.E. and I. Simon, *Topology of membrane proteins.* Journal of chemical information and computer sciences, 2001. **41**(2): p. 364-368.

61. Mus-Veteau, I., *Membrane proteins production for structural analysis*. 2014: Springer.

62. Kozma, D., I. Simon, and G.E. Tusnady, *PDBTM: Protein Data Bank of transmembrane proteins after 8 years.* Nucleic acids research, 2012. **41**(D1): p. D524-D529.

63. Bernhofer, M., et al., *TMSEG: Novel prediction of transmembrane helices.* Proteins: Structure, Function, and Bioinformatics, 2016. **84**(11): p. 1706-1716.

64. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank.* Nucleic acids research, 2005. **33**(suppl_1): p. D275-D278.

65. Kobe, B., et al., *Crystallography and protein–protein interactions: biological interfaces and crystal contacts*. 2008, Portland Press Limited.

66. Janin, J. and F. Rodier, *Protein–protein interaction at crystal contacts.* Proteins: Structure, Function, and Bioinformatics, 1995. **23**(4): p. 580-587.

67. Ponstingl, H., K. Henrick, and J.M. Thornton, *Discriminating between homodimeric and monomeric proteins in the crystalline state.* Proteins: Structure, Function, and Bioinformatics, 2000. **41**(1): p. 47-57.

68. Bahadur, R.P., et al., *A dissection of specific and non-specific protein–protein interfaces.* Journal of molecular biology, 2004. **336**(4): p. 943-955.

69. Lemmon, M.A., et al., *Glycophorin-a Dimerization Is Driven by Specific Interactions between Transmembrane Alpha-Helices.* Journal of Biological Chemistry, 1992. **267**(11): p. 7683-7689.

70. Gurezka, R., et al., *A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments.* Journal of Biological Chemistry, 1999. **274**(14): p. 9265-9270.

71. Oates, J., G. King, and A.M. Dixon, *Strong oligomerization behavior of PDGFbeta receptor transmembrane domain and its regulation by the juxtamembrane regions.* Biochim Biophys Acta, 2010. **1798**(3): p. 605-15.

72. Sal-Man, N., D. Gerber, and Y. Shai, *The identification of a minimal dimerization motif QXXS that enables homo- and hetero-association of transmembrane helices in vivo.* J Biol Chem, 2005. **280**(29): p. 27449-57.

73. Dawson, J.P., J.S. Weinger, and D.M. Engelman, *Motifs of serine and threonine can drive association of transmembrane helices.* J Mol Biol, 2002. **316**(3): p. 799-805.

74. Choma, C., et al., *Asparagine-mediated self-association of a model transmembrane helix.* Nature Structural Biology, 2000. **7**(2): p. 161-166.

75. Zhou, F.X., et al., *Polar residues drive association of polyleucine transmembrane helices.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2250-5.

76. Partridge, A.W., R.A. Melnyk, and C.M. Deber, *Polar residues in membrane domains of proteins: molecular basis for helix-helix association in a mutant CFTR transmembrane segment.* Biochemistry, 2002. **41**(11): p. 3647-53.

77. Fleishman, S.J., J. Schlessinger, and N. Ben-Tal, *A putative molecular-activation switch in the transmembrane domain of erbB2.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 15937-40.

78. Ding, P.Z. and T.H. Wilson, *The effect of modifications of the charged residues in the transmembrane helices on the transport activity of the melibiose carrier of Escherichia coli.* Biochem Biophys Res Commun, 2001. **285**(2): p. 348-54.

79. Chin, C.N. and G. von Heijne, *Charge pair interactions in a model transmembrane helix in the ER membrane.* J Mol Biol, 2000. **303**(1): p. 1-5.

80. Partridge, A.W., A.G. Therien, and C.M. Deber, *Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease.* Proteins, 2004. **54**(4): p. 648-56.

81. Belbeoc'h, S., et al., *Elements of the C-terminal t peptide of acetylcholinesterase that determine amphiphilicity, homomeric and heteromeric associations, secretion and degradation.* Eur J Biochem, 2004. **271**(8): p. 1476-87.

82. Ramachandran, R., R.K. Tweten, and A.E. Johnson, *Membrane-dependent conformational changes initiate cholesterol-dependent cytolysin oligomerization and intersubunit beta-strand alignment.* Nat Struct Mol Biol, 2004. **11**(8): p. 697-705.

83. Bocharov, E.V., et al., *Structure-functional insight into transmembrane helix dimerization*, in *Protein Engineering*. 2012, InTech.

84. Azriel, R. and E. Gazit, *Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation.* J Biol Chem, 2001. **276**(36): p. 34156-61.

85. Shi, Y., *A glimpse of structural biology through X-ray crystallography.* Cell, 2014. **159**(5): p. 995-1014.

86. Gobl, C., et al., *NMR approaches for structural analysis of multidomain proteins and complexes in solution.* Prog Nucl Magn Reson Spectrosc, 2014. **80**: p. 26-63.

87. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.

88. hopf, *Sequence co-evolution gives 3D contacts and structures of protein complexes.* 2014.

89. Xue, L.C., D. Dobbs, and V. Honavar, *HomPPI: a class of sequence homology based protein-protein interface prediction methods.* BMC Bioinformatics, 2011. **12**: p. 244.

90. Zhang, Q.C., et al., *PredUs: a web server for predicting protein interfaces using structural neighbors.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W283-7.

91. Shoemaker, B.A., et al., *Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites.* Nucleic Acids Res, 2010. **38**(Database issue): p. D518-24.

92. Jordan, R.A., et al., *Predicting protein-protein interface residues using local surface structural similarity.* BMC Bioinformatics, 2012. **13**: p. 41.

93. Porollo, A. and J. Meller, *Prediction-based fingerprints of protein-protein interactions.* Proteins, 2007. **66**(3): p. 630-45.

94. Liang, S., et al., *Protein binding site prediction using an empirical scoring function.* Nucleic Acids Res, 2006. **34**(13): p. 3698-707.

95. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 2004. **338**(1): p. 181-99.

96. Kufareva, I., et al., *PIER: protein interface recognition for structural proteomics.* Proteins, 2007. **67**(2): p. 400-17.

97. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data.* Proteins, 2005. **61**(1): p. 21-35.

98. Qin, S. and H.X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction.* Bioinformatics, 2007. **23**(24): p. 3386-7.

99. de Vries, S.J. and A.M. Bonvin, *CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK.* PLoS One, 2011. **6**(3): p. e17695.

100. Minhas, F., B.J. Geiss, and A. Ben-Hur, *PAIRpred: partner-specific prediction of interacting residues from sequence and structure.* Proteins, 2014. **82**(7): p. 1142-55.

101. Ahmad, S. and K. Mizuguchi, *Partner-aware prediction of interacting residues in protein-protein complexes from sequence data.* PLoS One, 2011. **6**(12): p. e29104.

102. Murakami, Y. and K. Mizuguchi, *Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites.* Bioinformatics, 2010. **26**(15): p. 1841-8.

103. de Vries, S.J., A.D. van Dijk, and A.M. Bonvin, *WHISCY: what information does surface conservation yield? Application to data-driven docking.* Proteins, 2006. **63**(3): p. 479-89.

104. Yan, C., D. Dobbs, and V. Honavar, *A two-stage classifier for identification of protein-protein interface residues.* Bioinformatics, 2004. **20 Suppl 1**: p. i371-8.

105. Northey, T., A. Baresic, and A.C.R. Martin, *IntPred: a structure-based predictor of protein-protein interaction sites.* Bioinformatics, 2017.

106. Hamer, R., et al., *i-Patch: interprotein contact prediction using local network information.* Proteins, 2010. **78**(13): p. 2781-97.

107. Yachdav, G., et al., *PredictProtein--an open resource for online prediction of protein structural and functional features.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W337-43.

108. Yu, H., et al., *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-18.

109. Loewenstein, Y., et al., *Protein function annotation by homology-based inference.* Genome Biol, 2009. **10**(2): p. 207.

110. Marks, D.S., T.A. Hopf, and C. Sander, *Protein structure prediction from sequence variation.* Nat Biotechnol, 2012. **30**(11): p. 1072-80.

111. Pazos, F., et al., *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 1997. **271**(4): p. 511-23.

112. Skerker, J.M., et al., *Rewiring the specificity of two-component signal transduction systems.* Cell, 2008. **133**(6): p. 1043-54.

113. Halperin, I., H. Wolfson, and R. Nussinov, *Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.* Proteins, 2006. **63**(4): p. 832-45.

114. Bordner, A.J., *Predicting protein-protein binding sites in membrane proteins.* BMC Bioinformatics, 2009. **10**: p. 312.

115. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. **9**(2): p. 173-5.

116. Fagerberg, L., et al., *Prediction of the human membrane proteome.* PROTEOMICS, 2010. **10**(6): p. 1141-1149.

117. Brosig, B. and D. Langosch, *The dimerization motif of the glycophorin A transmembrane segment in membranes: Importance of glycine residues.* Protein Science, 1998. **7**(4): p. 1052-1056.

118. Lawrie, C.M., E.S. Sulistijo, and K.R. MacKenzie, *Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: Roles for sequence context in helix-helix association in membranes.* Journal of Molecular Biology, 2010. **396**(4): p. 924-936.

119. Bugge, K., K. Lindorff-Larsen, and B.B. Kragelund, *Understanding single-pass transmembrane receptor signaling from a structural viewpoint—what are we missing?* The FEBS Journal, 2016. **283**(24): p. 4424-4451.

120. Bocharov, E.V., et al., *Helix-helix interactions in membrane domains of bitopic proteins: Specificity and role of lipid environment.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2017. **1859**(4): p. 561-576.

121. Valley, C.C., A.K. Lewis, and J.N. Sachs, *Piecing it together: Unraveling the elusive structure-function relationship in single-pass membrane receptors.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2017. **1859**(9): p. 1398-1416.

122. Langosch, D., et al., *Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator.* J Mol Biol, 1996. **263**(4): p. 525-30.

123. Ridder, A., et al., *Tryptophan supports interaction of transmembrane helices.* Journal of Molecular Biology, 2005. **354**(4): p. 894-902.

124. Johnson, R.M., K. Hecht, and C.M. Deber, *Aromatic and cation-π interactions enhance helix-helix association in a membrane environment.* Biochemistry, 2007. **46**(32): p. 9208-9214.

125. Polyansky, A.A., P.E. Volynsky, and R.G. Efremov, *Multistate organization of transmembrane helical protein dimers governed by the host membrane.* Journal of the American Chemical Society, 2012. **134**(35): p. 14390-14400.

126. Polyansky, A.A., et al., *PREDDIMER: A web server for prediction of transmembrane helical dimers.* Bioinformatics, 2014. **30**(6): p. 889-890.

127. Lomize, A.L. and I.D. Pogozheva, *TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes.* Journal of Molecular Biology, 2017. **429**(3): p. 390-398.

128. Mueller, B.K., S. Subramaniam, and A. Senes, *A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(10): p. E888-E895.

129. Adamian, L. and J. Liang, *Prediction of transmembrane helix orientation in polytopic membrane proteins.* BMC Structural Biology, 2006. **6**: p. 13.

130. Lensink, M.F. and S.J. Wodak, *Blind predictions of protein interfaces by docking calculations in CAPRI.* Proteins: Structure, Function and Bioinformatics, 2010. **78**(15): p. 3085-3095.

131. Lensink, M.F., et al., *The challenge of modeling protein assemblies: the CASP12-CAPRI experiment.* Proteins: Structure, Function and Bioinformatics, 2018. **86**: p. 257-273.

132. Wang, Y. and P. Barth, *Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy.* Nat Commun, 2015. **6**: p. 7196.

133. Mineev, K.S., S.A. Goncharuk, and A.S. Arseniev, *Toll-like receptor 3 transmembrane domain is able to perform various homotypic interactions: An NMR structural study.* FEBS Letters, 2014. **588**(21): p. 3802-3807.

134. Kohlway, *Hepatitis C virus RNA replication and virus particle assembly require specific dimerization of the NS4A protein.* 2014.

135. Muhle-Goll, C., et al., *Hydrophobic matching controls the tilt and stability of the dimeric platelet-derived growth factor receptor (PDGFR) α transmembrane segment.* Journal of Biological Chemistry, 2012. **287**(31): p. 26178-26186.

136. Call, M.E., et al., *The structure of the ζζ transmembrane dimer reveals features essential for its assembly with the T Cell Receptor.* Cell, 2006. **127**(2): p. 355-368.

137. Bocharov, E.V., et al., *Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1.* Journal of Biological Chemistry, 2008. **283**(43): p. 29385-29395.

138. Kozma, D., I. Simon, and G.E. Tusnady, *PDBTM: Protein Data Bank of transmembrane proteins after 8 years.* Nucleic Acids Res, 2013. **41**(Database issue): p. D524-9.

139. Li, W. and A. Godzik, *CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.

140. Beuming, T. and H. Weinstein, *A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins.* Bioinformatics, 2004. **20**(12): p. 1822-1835.

141. Adamian, L. and J. Liang, *Prediction of transmembrane helix orientation in polytopic membrane proteins.* Bmc Structural Biology, 2006. **6**.

142. Engelman, D.M., T.A. Steitz, and A. Goldman, *Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins.* Annual review of biophysics and biophysical chemistry, 1986. **15**: p. 321-353.

143. Kaján, L., et al., *FreeContact: Fast and free software for protein contact prediction from residue co-evolution.* BMC Bioinformatics, 2014. **15**(1).

144. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.

145. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-90.

146. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine Learning, 2006. **63**(1): p. 3-42.

147. Polyansky, A.A., et al., *PREDDIMER: a web server for prediction of transmembrane helical dimers.* Bioinformatics, 2014. **30**(6): p. 889-90.

148. Barwe, S.P., et al., *Janus model of the Na,K-ATPase β-subunit transmembrane domain: distinct faces mediate α/β assembly and β-β homo-oligomerization.* Journal of Molecular Biology, 2007. **365**(3): p. 706-714.

149. Chin, C.N., J.N. Sachs, and D.M. Engelman, *Transmembrane homodimerization of receptor-like protein tyrosine phosphatases.* FEBS Letters, 2005. **579**(17): p. 3855-3858.

150. Finger, C., C. Escher, and D. Schneider, *The single transmembrane domains of human receptor tyrosine kinases encode self-interactions.* Science Signaling, 2009. **2**: p. 89.

151. Kirrbach, J., et al., *Self-interaction of transmembrane helices representing pre-clusters from the human single-span membrane proteins.* Bioinformatics, 2013. **29**(13): p. 1623-1630.

152. Elazar, A., et al., *Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane.* eLife, 2016. **5**: p. e12125.

153. Gerber, D., N. Sal-Man, and Y. Shai, *Two motifs within a transmembrane domain, one for homodimerization and the other for heterodimerization.* Journal of Biological Chemistry, 2004. **279**(20): p. 21177-21182.

154. Zhu, H., et al., *Specificity for homooligomer versus heterooligomer formation in integrin transmembrane helices.* Journal of Molecular Biology, 2010. **401**(5): p. 882-891.

155. Li, R., et al., *Dimerization of the transmembrane domain of integrin αIIb subunit in cell membranes.* Journal of Biological Chemistry, 2004. **279**(25): p. 26666-26673.

156. LaPointe, L.M., et al., *Structural organization of FtsB, a transmembrane protein of the bacterial divisome.* Biochemistry, 2013. **52**(15): p. 2574-2585.

157. Wei, P., et al., *The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif.* Protein Science, 2011. **20**(11): p. 1814-1823.

158. Plotkowski, M.L., et al., *Transmembrane domain of myelin protein zero can form dimers: Possible implications for myelin construction.* Biochemistry, 2007. **46**(43): p. 12164-12173.

159. Ried, C.L., C. Scharnagl, and D. Langosch, *Entrapment of water at the transmembrane helix-helix Interface of Quiescin Sulfhydryl Oxidase 2.* Biochemistry, 2016. **55**(9): p. 1287-1290.

160. Khadria, A.S., et al., *A gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3.* Journal of the American Chemical Society, 2014. **136**(40): p. 14068-14077.

161. Kohlway, A., et al., *Hepatitis C virus RNA replication and virus particle assembly require specific dimerization of the NS4A protein transmembrane domain.* Journal of Virology, 2014. **88**(1): p. 628-642.

162. Wang, Y. and P. Barth, *Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy.* Nature Communications, 2015. **6**.

163. Kozma, D., I. Simon, and G.E. Tusnády, *PDBTM: Protein data bank of transmembrane proteins after 8 years.* Nucleic Acids Research, 2013. **41**(D1).

164. Xue, L.C., D. Dobbs, and V. Honavar, *HomPPI: a class of sequence homology based protein-protein interface prediction methods.* BMC Bioinformatics, 2011. **12**(1): p. 244.

165. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Science, 2004. **13**(1): p. 190-202.

166. Stevens, T.J. and I.T. Arkin, *Substitution rates in α-helical transmembrane proteins.* Protein Science, 2001. **10**(12): p. 2507-2517.

167. Walters, R.F.S. and W.F. DeGrado, *Helix-packing motifs in membrane proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(37): p. 13658-13663.

168. Zhang, S.Q., et al., *The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions.* Structure, 2015. **23**(3): p. 527-541.

169. Hopf, T.A., et al., *Three-dimensional structures of membrane proteins from genomic sequencing.* Cell, 2012. **149**(7): p. 1607-1621.

170. Fuchs, A., et al., *Co-evolving residues in membrane proteins.* Bioinformatics, 2007. **23**(24): p. 3312-3319.

171. Marks, D.S., et al., *Protein 3D structure computed from evolutionary sequence variation.* PLoS One, 2011. **6**(12): p. e28766.

172. Teixeira, P.L., et al., *Membrane protein contact and structure prediction using co-evolution in conjunction with machine learning.* PLOS ONE, 2017. **12**(5): p. e0177866.

173. Caporaso, J.G., et al., *Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics.* BMC Evolutionary Biology, 2008. **8**(1): p. 327.

174. Avila-Herrera, A. and K.S. Pollard, *Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species.* BMC Bioinformatics, 2015. **16**: p. 268.

175. Svetnik, V., et al., *Boosting: an ensemble learning tool for compound classification and QSAR modeling.* J Chem Inf Model, 2005. **45**(3): p. 786-99.

176. Mueller, B.K., S. Subramaniam, and A. Senes, *A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical C alpha-H hydrogen bonds.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(10): p. E888-E895.

177. Louppe, G., et al., *Understanding variable importances in forests of randomized trees.* Adv. Neur. Inform. Proc. Syst., 2013. **1**: p. 431-439.

178. Cosson, P. and J.S. Bonifacino, *Role of transmembrane domain interactions in the assembly of class II MHC molecules.* Science, 1992. **258**(5082): p. 659-662.

179. Lin, X., et al., *Two types of transmembrane homomeric interactions in the integrin receptor family are evolutionarily conserved.* Proteins: Structure, Function, and Bioinformatics, 2006. **63**(1): p. 16-23.

180. Barwe, S.P., et al., *Janus model of the Na,K-ATPase β-subunit transmembrane domain: distinct faces mediate α/β assembly and β-β homo-oligomerization.* J Mol Biol, 2007. **365**(3): p. 706-14.

181. Paulhe, F., et al., *Dimerization of Kit-ligand and efficient cell-surface presentation requires a conserved Ser-Gly-Gly-Tyr motif in its transmembrane domain.* FASEB Journal, 2009. **23**(9): p. 3037-3048.

182. LaPointe, L.M., et al., *Structural organization of FtsB, a transmembrane protein of the bacterial divisome.* Biochemistry, 2013. **52**(15): p. 2574-85.

183. Illergard, K., A. Kauko, and A. Elofsson, *Why are polar residues within the membrane core evolutionary conserved?* Proteins, 2011. **79**(1): p. 79-91.

184. Khadria, A.S. and A. Senes, *The transmembrane domains of the bacterial cell division proteins FtsB and ftsL form a stable high-order oligomer.* Biochemistry, 2013. **52**(43): p. 7542-7550.

185. Hong, H., *Toward understanding driving forces in membrane protein folding.* Arch Biochem Biophys, 2014. **564**: p. 297-313.

186. Dawson, J.P., et al., *Sequence context strongly modulates association of polar residues in transmembrane helices.* J Mol Biol, 2003. **331**(1): p. 255-62.

187. de Juan, D., F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nature Reviews Genetics, 2013. **14**(4): p. 249-261.

188. Kerrien, S., et al., *IntAct--open source resource for molecular interaction data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D561-5.

189. The UniProt, C., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.

190. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spots--a review of the protein-protein interface determinant amino-acid residues.* Proteins, 2007. **68**(4): p. 803-12.

191. Frishman, D. and H.W. Mewes, *Protein structural classes in five complete genomes.* Nat Struct Biol, 1997. **4**(8): p. 626-8.

192. Nooren, I.M. and J.M. Thornton, *Diversity of protein-protein interactions.* EMBO J, 2003. **22**(14): p. 3486-92.

193. Ofran, Y. and B. Rost, *Predicted protein-protein interaction sites from local sequence information.* FEBS Lett, 2003. **544**(1-3): p. 236-9.

194. Res, I., I. Mihalek, and O. Lichtarge, *An evolution based classifier for prediction of protein interfaces without using protein structures.* Bioinformatics, 2005. **21**(10): p. 2496-2501.

195. Murakami, Y. and K. Mizuguchi, *Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites.* Bioinformatics, 2010. **26**(15): p. 1841-1848.

196. Meyer, M.J., et al., *Interactome INSIDER: a structural interactome browser for genomic studies.* Nat Methods, 2018. **15**(2): p. 107-114.

197. Kamisetty, H., S. Ovchinnikov, and D. Baker, *Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.* Proc Natl Acad Sci U S A, 2013. **110**(39): p. 15674-9.

198. Fernandez-Recio, J., M. Totrov, and R. Abagyan, *Identification of protein-protein interaction sites from docking energy landscapes.* J Mol Biol, 2004. **335**(3): p. 843-65.

199. Zhou, H.X. and Y. Shan, *Prediction of protein interaction sites from sequence profile and residue neighbor list.* Proteins, 2001. **44**(3): p. 336-43.

200. Fariselli, P., et al., *Prediction of protein--protein interaction sites in heterocomplexes with neural networks.* Eur J Biochem, 2002. **269**(5): p. 1356-61.

201. Wang, B., et al., *Predicting protein interaction sites from residue spatial sequence profile and evolution rate.* FEBS Lett, 2006. **580**(2): p. 380-4.

202. Chen, C.T., et al., *Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces.* PLoS One, 2012. **7**(6): p. e37706.

203. Bordner, A.J. and R. Abagyan, *Statistical analysis and prediction of protein-protein interfaces.* Proteins-Structure Function and Bioinformatics, 2005. **60**(3): p. 353-366.

204. Koike, A. and T. Takagi, *Prediction of protein-protein interaction sites using support vector machines.* Protein Engineering Design & Selection, 2004. **17**(2): p. 165-173.

205. Bradford, J.R. and D.R. Westhead, *Improved prediction of protein-protein binding sites using a support vector machines approach.* Bioinformatics, 2005. **21**(8): p. 1487-1494.

206. Zellner, H., et al., *Prescont: Predicting protein-protein interfaces utilizing four residue properties.* Proteins-Structure Function and Bioinformatics, 2012. **80**(1): p. 154-168.

207. Li, B.Q., et al., *Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS.* PLoS One, 2012. **7**(8): p. e43927.

208. Sikic, M., S. Tomic, and K. Vlahovicek, *Prediction of Protein-Protein Interaction Sites in Sequences and 3D Structures by Random Forests.* Plos Computational Biology, 2009. **5**(1).

209. Segura, J., P.F. Jones, and N. Fernandez-Fuentes, *Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams.* Bmc Bioinformatics, 2011. **12**.

210. Marks, D.S., *Protein 3D Structure Computed from Evolutionary Sequence Variation.* 2011.

211. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates.* Bioinformatics, 2005. **21**(7): p. 1276-7.

212. Fuchs, A., A. Kirschner, and D. Frishman, *Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks.* Proteins, 2009. **74**(4): p. 857-71.

213. Hamp, T. and B. Rost, *Alternative protein-protein interfaces are frequent exceptions.* PLoS Comput Biol, 2012. **8**(8): p. e1002623.

214. Richmond, T.J., *Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect.* J Mol Biol, 1984. **178**(1): p. 63-89.

215. Porter, C.T. and A.C. Martin, *BiopLib and BiopTools--a C programming library and toolset for manipulating protein structure.* Bioinformatics, 2015. **31**(24): p. 4017-9.

216. Kall, L., A. Krogh, and E.L. Sonnhammer, *Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W429-32.

217. Magrane, M. and C. UniProt, *UniProt Knowledgebase: a hub of integrated protein data.* Database (Oxford), 2011. **2011**: p. bar009.

218. Honigschmid, P. and D. Frishman, *Accurate prediction of helix interactions and residue contacts in membrane proteins.* J Struct Biol, 2016. **194**(1): p. 112-23.

219. Wang, X.F., et al., *Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach.* PLoS One, 2011. **6**(10): p. e26767.

220. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.

221. Marks, D.S., et al., *Protein 3D structure computed from evolutionary sequence variation.* PLoS ONE, 2011. **6**(12).

222. Hopf, T.A., et al., *Three-dimensional structures of membrane proteins from genomic sequencing.* Cell, 2012. **149**(7): p. 1607-21.

223. Kajan, L., et al., *FreeContact: fast and free software for protein contact prediction from residue co-evolution.* BMC Bioinformatics, 2014. **15**: p. 85.

224. Adamian, L., et al., *Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins.* Proteins, 2005. **59**(3): p. 496-509.

225. Kawashima, S., H. Ogata, and M. Kanehisa, *AAindex: Amino Acid Index Database.* Nucleic Acids Research, 1999. **27**(1): p. 368-369.

226. Guharoy, M. and P. Chakrabarti, *Conserved residue clusters at protein-protein interfaces and their use in binding site identification.* BMC Bioinformatics, 2010. **11**: p. 286.

227. Wegener, K.L. and I.D. Campbell, *Transmembrane and cytoplasmic domains in integrin activation and protein-protein interactions (review).* Mol Membr Biol, 2008. **25**(5): p. 376-87.

228. Hellmich, J., et al., *Native-like photosystem II superstructure at 2.44 A resolution through detergent extraction from the protein crystal.* Structure, 2014. **22**(11): p. 1607-15.

229. Yin, H. and A.D. Flynn, *Drugging Membrane Protein Interactions.* Annu Rev Biomed Eng, 2016. **18**: p. 51-76.

230. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.

231. Moreira, I.S., et al., *SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots.* Sci Rep, 2017. **7**(1): p. 8007.

232. Maheshwari, S. and M. Brylinski, *Predicting protein interface residues using easily accessible on-line resources.* Brief Bioinform, 2015. **16**(6): p. 1025-34.

233. Kaján, L., et al., *FreeContact: Fast and free software for protein contact prediction from residue co-evolution.* BMC Bioinformatics, 2014. **15**(1): p. 85.

234. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(49): p. E1293-E1301.

235. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-190.

236. Kawashima, S. and M. Kanehisa, *AAindex: Amino acid index database.* Nucleic Acids Research, 2000. **28**(1): p. 374.

237. Käll, L., A. Krogh, and E.L.L. Sonnhammer, *A Combined Transmembrane Topology and Signal Peptide Prediction Method.* Journal of Molecular Biology, 2004. **338**(5): p. 1027-1036.