TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Chemie der Biopolymere

Experimental identification and evolutionary analysis of homotypic transmembrane helix-helix interfaces

Yao Xiao

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

Die Dissertation wurde am 22.08.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 01.11.2018 angenommen.

ACKNOWLEDGEMENTS

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

Prof Dr. Dieter Langosch: for giving me the opportunity to carry out my doctoral research at his chair. This opportunity has made the biggest difference in my life. I thank him for the continuous support of my PhD study, for his patience, motivation, and immense knowledge.

Dr. Mark Teese: for supporting me in my project all the time. For always being so kind, helpful and motivating. His guidance helped my research and writing of the thesis. I could not have imagined having a better advisor for my PhD study.

Bo Zeng: for the wonderful collaborative experience over the last three years. For teaching me so much about bioinformatics.

Prof. Dr. Dmitrij Frishman: for his open-minded and generous support for the collaboration on this project, for his very helpful comments and suggestions, and additionally, for agreeing to review this thesis.

Doreen Tetzlaff: for her daily support in the lab, her generous preparation an uncountable number of LB plates media and for many nice conversations.

Dr. Markus Gütlich: for sharing his in-depth knowledge in almost everything and answering all kinds of questions about technical and scientific problems.

Walter Stelzer: for scientific and technical advice during the past years, for taking care of my computer and giving me access to the laboratory and research facilities.

Philipp: for the perfect wiki, lots of help, and many humorous and useful discussions.

My heartfelt thanks to all my present and past colleagues (Bo, Katja, Ayse, Philipp, Christoph, Julie, Martina, Elke, Ellen): for providing a great working atmosphere and for their help and conversations.

I gratefully acknowledge the funding of my PhD from the Chinese scholarship council.

I would like to thank my family for all their love and encouragement.

DECLARATION

This thesis is my own work. However, several sections of this thesis were the product of close collaboration with other researchers.

Chapter 4 contains results of scanning mutagenesis for a total of 10 novel TMDs. Approximately 12% (34/294) of the mutations were cloned by Nicola Berner, as part of a Bachelor thesis under my supervision.

Chapter 5 contains a comprehensive analysis of homotypic TM interfaces derived from my experimental data, data from other ETRA studies, NMR studies, and available crystal structures. The bioinformatic and statistical analysis of residue properties presented here is my own work, although some the underlying data used for analysis was derived from collaborations. The definition of interfacial residues from NMR and crystal data (Methods chapter 3.4) was conducted by Bo Zeng of the Dmitrij Frishman group within TU-München. Like myself, Bo Zeng measured residue conservation and polarity from multiple sequence alignments. To allow a co-author publication between Bo Zeng and myself, the separate bioinformatic methods were harmonised and finally incorporated into the thoipapy python package of Bo Zeng. Bo used these residue properties to create a machine-learning predictor, TM HOmodimer Interface Prediction Algorithm (THOIPA).

The construction and validation of THOIPA is not described here, but can be found in the thesis of Bo Zeng, and in the corresponding co-author publication.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The homotypic (self-self) interaction of transmembrane (TM) helices is a key type of protein-protein interaction in the membrane. It supports the dimerisation and oligomerisation of many bitopic membrane protein, such as receptor, and it is therefore vital for many cellular processes. However, the properties of the interface residues are poorly understood. Until now, there have been no quantitative studies on natural homotypic TM interfaces. The aim of this study was to quantify for the first time the importance of factors such as residue conservation, residue polarity, and the GxxxG motif. Experiments were performed using the ToxR assay, a powerful *Escherichia coli* reporter assay (ETRA), which allows the identification of homotypic TM interfaces in a natural membrane environment.

Interfaces of a total of 10 unique self-interacting TMDs from bitopic human proteins were determined using the ToxR assay. A total of 294 mutations at 224 positions are generated and their impact on the efficiency of self-interaction was tested. The novel interfaces were diverse and included GxxxG motifs, small residues, aromatic residues, aliphatic residues, and also one clear “leucine-zipper.” A dataset of 21 TMDs was created by combining these nine unique experimentally determined TMDs with another 12 ETRA studies from literature. This confirmed for the first time that mutation-sensitive positions in ETRA assays have α-helical periodicity, and that natural interfaces tend to be associated with GxxxG motifs, high conservation, and high polarity at interfacial positions. To obtain a broader perspective of TM homodimer interface properties, a comprehensive dataset of 54 self-interacting TMDs was created that combined experimental data from ETRA, NMR and crystallography studies. Extensive bioinformatic sequence analysis confirms the ETRA study in that homotypic TM interfacial residues tend to be conserved, coevolved, polar, and have a high depth in the membrane.

ZUSAMMENFASSUNG

Die homotypische Wechselwirkung von Transmembran-(TM) Helices ist eine wichtige Protein-Protein-Wechselwirkung in der Membran. Sie unterstützt die Dimerisierung und Oligomerisierung vieler bitopischer Membranproteine, wie etwa von Rezeptoren, und ist deshalb von entscheidender Bedeutung für viele zelluläre Prozesse. Jedoch sind die Eigenschaften der Aminosäuren an der Kontaktfläche zwischen den TM-Helices noch wenig verstanden. Bislang gab es nämlich keine quantitativen Studien zu homotypischen TM-Kontaktflächen natürlicher bitopischer Proteine. Ziel dieser Studie war es daher, erstmals die Bedeutung von Faktoren wie der evolutionären Konservierung und Polarität der Aminosäuren sowie des Auftretens des GxxxG Motifs zu quantifizieren. Die Experimente wurden mit dem ToxR-Assay, einem leistungsfähigen *Escherichia coli*-Reporter-Assay (ETRA), durchgeführt der die Profilierung von homotypischen TM-Kontaktflächen in einer natürlichen Membranumgebung ermöglicht.

Die Kontaktflächen von zehn selbst-interagierenden TMDs wurden unter Verwendung des ToxR-Assays bestimmt, indem die Auswirkung von 294 Punktmutationen an 224 Positionen auf die Effizienz der Wechselwirkung untersucht wurde. Die damit erhaltenen Kontaktflächen waren von diverser Aminosäurenzusammensetzung und enthielten GxxxG Motive, kleine, aromatische und aliphatische Reste sowie einen „Leucin-Zipper“ Motif. Ein Datensatz von 21 TMDs wurde erstellt, indem diese neun experimentell bestimmten TMDs mit weiteren 12 ETRA Studien aus der Literatur durch kombiniert wurden. Die Analyses dieses Datensatzes zeigt zum ersten Mal, dass die mutationsempfindlichen Positionen im ETRA-Assays eine Helix--Periodizität aufweisen, und dass natürliche Schnittstellen

mit dem Auftreten von GxxxG-Motiven sowie hoher Konservierung und Polarität der Aminosäuren in der Kontaktfläche assoziiert sind. Um eine breiteren Überblick zu den Eigenschaften der Kontaktflächen von TM Homodimeren zu erhalten, wurde ein umfassender Datensatz von 54 selbst-interagierenden TMDs erstellt der experimentelle Daten von ETRA, NMR und kristallographischen Studien vereint. Eine umfangreiche bioinformatische Sequenzanalyse bestätigt die ETRA Analyse insofern als dass homotypische TM-Kontakflächenreste tendenziell konserviert, koevolviert und polar, sowie tiefer in der Membran lokalisiert sind.

CHAPTER 1. INTRODUCTION

1.1 Membrane proteins

Integral membrane proteins constitute about a quarter of all proteins of currently sequenced genomes [45-47]. In humans, approximately 6,000 different membrane proteins are expressed [46, 48]. They take part in countless cellular processes, and comprise the majority of targets for pharmaceutical compounds [47].

Transmembrane (TM) proteins are typically classified according to their secondary structure. α-Helical TM proteins are the dominant type in eukaryotic membranes. β-barrel TM proteins 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. In general, TM domains are characterised by a high secondary structure, and a high degree of residue burial in the protein structure [49]. Presumably, this secondary structure helps shield the polar polypeptide backbone from the lipid environment. In contrast, a proportion of soluble proteins in aqueous biological solutions have no secondary structure, and are classified as disordered or unstructured [50, 51]. There are no known constructed membrane proteins. Contacts between transmembrane domains (TMDs) are considered to be very common [52].

α-Helical proteins are further classified by their topology in the membrane. α-Helical proteins that span the bilayer once are known as bitopic, or single-span membrane proteins. α-Helical proteins that span the bilayer more than once are known as polytopic, or multi-pass membrane proteins. Bitopic membrane proteins are highly abundant in eukaryotic genomes. In fact, they are more common than polytopic proteins with any other number of TM helices [47, 53].

1.2 Biological relevance of homotypic TMD interactions of bitopic proteins

TMD-TMD interaction can be divided into two categories, homotypic and heterotypic. Homotypic interactions are self-self interactions that are usually assumed to be TM dimers, although high oligomers such as trimer are known to exist [54]. Heterotypic interaction involves non-identical TM helices, such as the interaction between TM helices in a large, folded membrane protein.

Numerous studies have shown that the oligomerisation of the single TMD is required for in a wide variety of biological processes [47, 55-58]. Homotypic TMD interactions are common in receptor proteins, where they are vital for the transfer of a signal across the membrane [56]. For bitopic membrane proteins, the formation of specific TM dimers is particularly astounding, as it requires a highly specific molecular recognition based on a relatively small surface area. Signals can be transferred across the membrane by structural rearrangements of constitutive TM homodimers [59-61]. Increasing evidence suggests that this “ligand-induced-rotation” is a common feature within the receptor tyrosine kinase (RTK) family [62, 63]. TM homodimers are also found to have structural roles within large protein complexes such as photosystem II [64]. The monomer-dimer equilibrium of TMDs may also regulate intramembrane proteolysis, such as for the amyloid precursor protein (APP) implicated in Alzheimer’s disease [65].

1.3 Methods to determine homotypic TM interfaces

1.3.1 SDS-PAGE

Several homotypic TMD interactions were discovered due to the TMD-dependent dimerisation of proteins visualised by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS PAGE). Many membrane proteins retain their structures in the SDS micelle environment [66]. SDS PAGE has been used successfully to investigate some TMDs, such as GpA [67], BNIP3 [68] and ErbB [69]. However all of these TMDs were later investigated with ETRA methods (Section 1.3.2 below), which are not only to be faster, but offer a more natural membrane environment.

1.3.2 *E. coli* TM Reporter Assay (ETRA) techniques

ETRA techniques in combination with scanning mutagenesis have now been used for over 20 years to determine interfacial residues of TM homodimers. Most studies have used ToxR-based assays such as ToxR [70], TOXCAT [71], or the recently developed dsTβL [22]. Other ETRA techniques include GALLEX [72], BACTH [73, 74] and AraTM [75], which all utilise transcription activator domains, and BLaTM [76], which is based on a split enzyme.

In some cases, the interface seen in NMR has been confirmed in biological membranes using ETRA techniques. Early research on GpA and BNIP3 revealed interfacial residues that were generally consistent between SDS-assays [6, 77], ToxR [70, 78], and NMR analyses [1, 5, 79]. This suggested that the TM homodimers were relatively insensitive to the membrane environment. More recently, however, it has been shown that variations in sequence length and membrane environment can lead to drastic differences in TM homodimer structure and affinity [80]. For example, three different TM homodimer interfaces for ErbB2 have now been discovered: two with NMR [14, 81], and one using the dsTβL assay [22]. It is currently unknown if this is

an isolated case, or if the TM homodimer structures in the *Escherichia coli* (*E. coli*) membrane are consistently different from those in membrane mimetics.

1.3.2.1 The ToxR system

The ToxR assay is based on the ToxR transcriptional activator and is used to study TMD-TMD interactions in the inner membrane of *E. coli* [8, 70]. The ToxR transcription activator was originally from *Vibrio cholera, where it* that activates the expression of virulence factors. 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 [82].

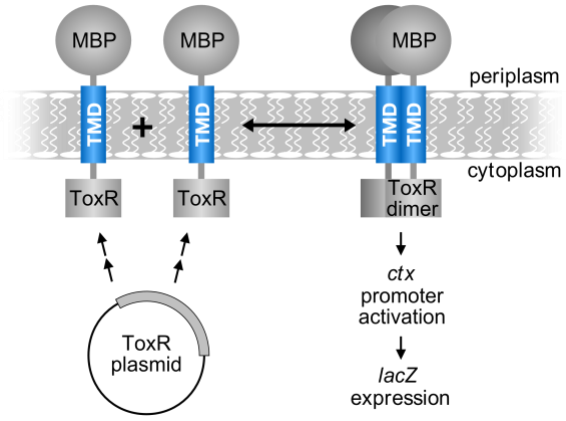

Figure 1-1: Overview of the ToxR system. The self-interaction of TMDs leads to the dimerisation of the cytoplasmic ToxR domains. The dimer then induces transcription activation of the ctx promoter and thus activate expression of the downstream lacZ gene. The periplasm MBP domain allows the detection of expression level the chimeric protein with antibodies and also for the analysis of correct membrane insertion. Adapted from [83].

In order to study the TMD interaction of bitopic membrane proteins, the membrane-spanning domain of the ToxR protein is replaced with the TMD of interest [70]. The

exchange of the TM segment does not affect the function of the ToxR cytoplasmic domain. The ToxR-activity assay utilises a chimeric protein consisting of the ToxR transcription activator at the cytoplasmic side of the membrane, the TMD of interest, and finally periplasmic maltose binding protein (MBP, encoding by malE) [22, 70]. 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 [8]. TMD-TMD interaction mediates the self-interaction of ToxR proteins in the cytoplasm leading to transcription activation of *ctx*-promoter. The reporter gene *LacZ* which encodes β-galactosidase is placed under the control of the *ctx*-promoter. β-galactosidase catalyses the hydrolysis of ortho-nitrophenyl-β-galactoside (ONPG) and produces the o-nitrophenol (ONP). The amount of ONP can be measured at 405 nm within the cell lysate. The rate of production of ONP, normalised to the cell concentration, thereby gives the relative homotypic interaction strength of a TMD.

1.3.2.1.1.1 PD28 membrane integration assay

The MBP domain at the C-terminus of the ToxR fusion protein allows a control assay to confirm the protein is correctly inserted into the membrane. The correctly inserted fusion proteins will have an MBP domain located at periplasm, and a ToxR domain located at the cytoplasm.

The MBP domain is part of a transporter system to take up maltose into the cell. MBP-deficient PD28 cells cannot grow in a minimal medium supplemented with maltose as the only carbon source. When grown in minimal media containing maltose as the only carbon resource, the correct integration of the ToxR-TMD-MBP fusion protein in the

inner membrane gives a measurable phenotype of faster growth. In contrast, the growth of cells containing the fusion construct lacking the transmembrane domain (ΔTM) should be strongly inhibited. ToxR-TMD-MBP constructs contain a TMD region that does not enter the membrane correctly will also show growth that is strongly inhibited.

1.3.3 Nuclear magnetic resonance (NMR) spectroscopy

NMR spectroscopy can be used for the structural characterisation of small proteins. NMR is therefore well suited to the structural analysis of isolated TM homodimers, which are typically analysed in detergent micelles or bicelles. The early characterisation of GpA, for example [1], showed a good correspondence with interface residues from earlier SDS-PAGE [77] and ETRA [70] experiments. To date, consensus structures have been generated based on NMR data for over 15 TM homodimer structures (Table 8-1) [1, 2, 5, 13, 14, 20, 24, 27, 30, 37, 38, 41-43]. These studies have been reviewed extensively [84, 85], and form the test dataset for de-novo structure determination [36, 86, 87]. 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 [13, 58], while in other cases it has been proposed that the TMD has multiple biologically relevant homodimer interfaces [58]. 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.

1.3.4 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 [88, 89]. As a consequence, no more than 2% of proteins in the Protein Data Bank (PDB) are transmembrane or globular proteins [90]. 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 membrane proteins [91].

Protein Data Bank of Transmembrane Proteins (PDBTM) database was created to collect transmembrane proteins from the PDB and defined their TMD by the TMDET algorithm [92]. The “crystal contacts” within the structures are often considered to be biologically relevant protein-protein interaction sites [93-96]. 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.

1.4 Properties of homotypic TMD interface residues predicted from case studies

There is little quantitative information on the residue properties of homotypic TMD interfaces. Most information is derived from case studies. Other findings have been discovered via a selection of artificial TMDs from “combinatorial libraries” that show strong homodimerisation [97]. A consistent theme implicated by numerous case studies is that homotypic interactions can be highly sequence specific. Single amino

acid mutations are well known to destabilise TM homodimers in natural membranes [6, 31], and strongly affect biological protein function [33]. In contrast to the studies of polytopic protein (detailed below), previous studies of bitopic TM interfaces have strongly focused on the role of simple sequence motifs. In fact, the discovery of a novel interface for a TMD of interest is sometimes described as a new sequence motif [98]. In this study, however, motifs were defined as a sequence that has been independently implicated in a number of different TMD interfaces.

1.4.1 GxxxG motif

Early case studies on TM homodimers focused heavily on the role of simple sequence motifs such as GxxxG and (small)xxx(small) [99]. The most common and best-characterized motif is the GxxxG motif, which was first detected by analysing the dimerisation of human glycophorin A (GpA) a major sialoglycoprotein of red blood cells (Figure 1-2) [70, 77]. The TM helix dimer of GpA adopted a negative crossing angle in, a tightly packed right-handed helix pair. The GxxxG motif consists of two Gly spaced four residues apart, placing both at the same helix side (assuming 3.6 residues per turn of the α-helix). Since the identification of the GxxxG motif within GpA [70], the motif has proved to be involved in the oligomerisation of the ErbB tyrosine kinase receptors [100], the yeast ATP synthase [101], the Helicobacter pylori vacuolating toxin [102], BNIP3 [6], HLA [103, 104] and other proteins as recently reviewed [97]. Because these data are derived from case studies, what is poorly understood is the relative predictive power of GxxxG motifs for identifying TMDs that form strong dimers [105]. Also uncertain is the exact prediction power of GxxxG motifs for the identification of interfacial residues within TMDs with strong self-interaction. In some cases, TMDs with GxxxG motifs do not appear to dimerise strongly, and others

dimerise via GxxxG-independent interfaces [21, 97, 106]. In some cases, the larger aliphatic residues are thought to cooperate with the small Gly residues (for GpA, GVxxGV) to maximise van der Waals contacts [19]. The effect of the neighbouring residues, known as “sequence context,” is also poorly understood. A sequence logo of the 26 GxxxG-dependent bitopic TMDs, aligned by their GxxxG motif, shows that there are no clear patterns visible in the sequence context (Figure 1-2). The one theory is that there is a complex sequence content consisting of residues that are favourable for the orientation of $C_\alpha H\cdots O{=}C$ main-chain/main-chain H-bond.

Senes lab developed the “Cα Transmembrane” (CATM) [107] method that can predict structures of GAS_{right} dimers. These dimers were assumed to be depended on $C_\alpha H\cdots O{=}C$ H-bond involved. They initially calculated the optimum angle of helices to maximise such bonds and provided a prediction-based model. The resulting models have a good correspondence to mutagenesis data for TOXCAT assays [11, 105, 107].

Figure 1-2: Sequence context of residues surrounding GxxxG motifs that facilitate homotypic TM interactions. The 26 TMDs with GxxxG-dependent self-interaction were collected from Teese and Langosch [97]. The sequences were aligned according to the GxxxG motif. A sequence logo was created using WebLogo. The height of the residue corresponds its frequency in that position of the alignment.

1.4.2 (small)xxx(small) motif

The (small)xxx(small) motif is a variant of the GxxxG motif where the Gly residues are occupied by any small residue. Small residues are typically Ser, Ala and Gly. In some cases, Cys is included. However, this is less crucial because of its low abundance. (small)xxx(small) motifs have been observed to drive homotypic helix-

helix association within natural membranes [18, 99, 108]. The importance of small residues at TM interfaces is well supported by recent NMR and crystal structure analyses [17, 109]. It has been suggested that small residues allow the close approach of the interacting helices, allowing favourable van der Waals interactions between interacting residues [1, 77]. Residues in TMDs are in general more buried than in soluble proteins, allowing increased van der Waals forces [110, 111]. Small residues that allow the tight packing of helices through “ridge-into-groove” or “knob-into-hole” packing. The lack of side-chain with association degrees of freedom means that dimerisation thus van der Waals forces occur without significant entropic loss [112, 113]. Alternatively, they allow the formation of non-canonical $C_\alpha H\cdots O=C$ H-bonding via the backbone C_α carbon [114, 115]. Small polar residues such as Ser contribute to the TMD-TMD interaction by forming a H-bond between side chains in the TMD or the C_α-H of one helix and a proton acceptor at the other helix [116]. Although strongly polar residues (e.g. Glu, His) create more specific, stronger H-bonds and salt bridges, they are rarely present within the TMD as their transfer into the membrane is thermodynamically unfavourable. Thus, polar residues are rare and usually buried in the interior of stable protein structures [117]. The substitution of polar residues to non-polar residues in the membrane protein is a common cause of genetic disease [118].

For predictive purposes, it has been argued that the more inclusive (small)xxx(small) motif is too common to be a strong indicator of self-interaction [97]. In a dataset of bitopic proteins over 60% of TMDs containing at least one (small)xxx(small) motif [97]. Even when the pattern is constrained to contain at least one Gly, the motif exists in 42% of all bitopic TM domains [19].

1.4.3 Leucine Zippers

This motif is loosely defined as a pattern of leucine, isoleucine or valine residues on one side of the helix face [119], using the heptad patterns based on coiled-coil nomenclature [120]. In coiled-coils, positions a and d in the heptad motif $[\underline{a}bc\underline{d}efg]_n$ are typically Ile or Leu residues, which are central to the interacting helices [121]. Helix pairs dimerising via a leucine zipper typically exhibit a positive crossing angle and pack as left-handed helix pairs [122]. In the absence of a defined motif, aliphatic side chains can contribute to the stability and specificity of TMD association through van der Waals interaction.

1.4.4 Aromatic residues and ionisable pairs

Aromatic interactions, either between two aromatic residues (π-π) or between a basic and an aromatic residue (cation-π) are another important feature of noncovalent interactions [123]. These interactions are known to enhance TMD dimerisation [117]. Due to the fact that aromatic residues have a strong propensity to face phospholipids, they are also thought to act as anchors for enhanced stability in the membrane. This could influence both helix tilting and hydrophobic mismatch [117]. In addition, the indole, phenol, and imidazole groups of aromatic residues can participate in H-bonding across the TM helix packing interface [117].

1.5 Properties of heterotypic TM interfaces that might also apply to homotypic interactions

Crystal databases contain an abundance of helix pairs that participate in heterotypic TM interactions [17, 109]. These are typically defined from folded polytopic

membrane proteins. These TMD-TMD interactions have been extensively and quantitatively analysed in several ways, giving an insight into the properties of heterotypic interface residues. They are typically referred to as “buried” residues within the polytopic protein structure, in comparison to the “lipid-accessible” outer residues [124]. Based on this knowledge, a large number of algorithms have been developed to aid the de-novo prediction of polytopic membrane protein structure [125-128].

This quantitive analysis of heterotypic TM interactions has yielded a more nuanced perspective on “dimerisation motifs”. TM helix pairs do not show strict adherence to any particular sequence motif [17, 109]. Indeed, interfacial residues are known to be quite diverse. Clustering of helix pairs based on their crossing angles and interhelical distance, however, has revealed clear patterns in amino acid propensity [17, 109]. However, it is difficult to be certain that the findings from crystal structure databases are relevant to TM homodimers. The helix pairs in crystal structure databases differ from TM homodimers in several ways. They typically belong to multi-pass rather than bitopic proteins. The interacting helices are not identical. Some multi-pass helices are quite short and/or polar. Finally, most helix pairs within crystal structures lie in an antiparallel configuration [129]. Nevertheless, several authors have assumed that most of the features of TM interfacial residues are shared between homotypic and heterotypic interaction [36, 87], including their high sequence conservation, polarity, and tendency to coevolve.

1.5.1 Evolutionary conservation

It has long been known that residues buried in membrane protein structures are well conserved [118, 124, 130, 131]. Residue conservation, in general, is obtained from

multiple sequence alignments (MSA) of homologues. Functionally important positions usually show the lowest tolerance of substitution and therefore the highest conservation [132]. Conservation has helped to predict important functional residues that are involved in protein folding, catalytic sites, ligand binding or protein-protein interactions [133]. The number of homologs is an important limiting factor especially for bitopic membrane proteins that are known to evolve rapidly [134].

1.5.2 Residue Polarity

Initially, it was thought that polytypic membrane proteins would be “inside-out” in comparison to soluble proteins, be internally polar and externally hydrophobic. This was later found not to be the case, however, residues buried in membrane protein structures are indeed more polar than those in contact with lipids [118, 124].

Several experimental and statistical methods have been used to assess amino acid hydrophobicity. For instance, Hessa et al. [135] designed several TMDs and quantify the membrane insertion efficiency of these segments. Four biological hydrophobicity scales, the Hessa scale [135], the KyteDoolittle scale [136], the Wimley scale [137] and the Engelman (GES) hydrophobicity scale [138] defines the contributions of individual AA in a position-specific manner.

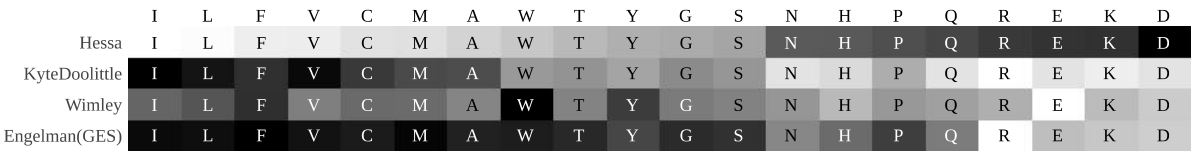

Figure 1-3: The KyteDoolittle, Hessa, Wimley and Engelman (GES) hydrophobicity scale The KyteDoolittle, Hessa, Wimley and Engelman (GES) for each amino acid when placed in a central position of a transmembrane helix.

The correlation between different scales is in general good but varies in some AA. For instance, Pro is hydrophilic in Hessa scale but hydrophobic in other three scales. Consider helix-breaking nature of Pro, the biological hydrophobicity scale used in this thesis is Engelman (GES) scale.

1.5.3 Residue coevolution

Residue coevolution is also known as covariance or evolutionary couplings. It is the tendency for correlated mutations to occur in residues that are in direct contact in a protein structure [139, 140]. In the evolutionary sequence analysis of homologues, a mutation to position A are therefore likely to be accompanied by a mutation to position B. This effect can be visualised in an example with the close proximity in a closely packed structure of a large residue A, to a small residue B. If B mutates to a larger residue, the fitness of the organism is decreased due to steric hindrance of the two large residues. The fitness cost can be mitigated by the reversion of B to a smaller residue, or alternatively, the mutation of residue A to a smaller residue.

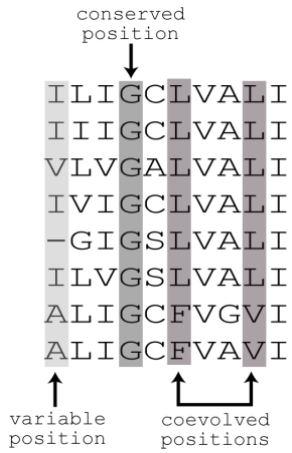

Figure 1-4. Illustration of residue coevolution, conservation and variability. A multiple sequence alignment against homologues is shown. Conserved, coevolved and variable positions are highlighted in grey.

A simple measure of coevolution is the mutual information (MI) between two positions. However, MI values have two problems. Firstly, they can be very high between two non-contacting residues, if both of these residues are in contact with a third residue (i.e. secondary correlations, [140]). Secondly, they cannot be used to detect residue pairs that are extremely well conserved [134]. In the last 10 years, there has been extensive progress in solving these two problems using methods known as direct-coupling analysis (DCA) [141, 142]. DCA methods often take a simple measure of coevolution (e.g. MI) as an input, and then apply a statistical model to distinguish direct contacts from indirect contacts. A popular DCA method yields the direct information (DI) coevolution score [143]. This is implemented in the EVfold software and online server (http://evfold.org), and other open-source implementations such as FreeContact [144].

The quality of measured coevolution values depends on the reliability of the MSA, numbers of homologues and the mean pairwise divergence levels in the MSAs [145]. The recent exponential increase in available sequences in public databases has

therefore increased the usefulness of coevolution scores in predicting contacting residues.

In a recent breakthrough, Wang and Barth [36] have shown that the coevolution scores can help identify interfacial residues of TM homodimers. They further used the coevolution scores to guide de-novo structure determination. They proposed that the sum of pairwise coevolution scores between all interfacial residues in available NMR structures are higher than those for non-interfacial residues. Because coevolution scores can only be calculated between non-identical residues, this suggests that even in symmetric dimers, interfacial residues often contact and coevolve with at least one other non-identical residue in the chain. This higher coevolution was only shown for retrospective analyses of TMDs with known interfaces, making it difficult to judge the importance of coevolution in comparison to other factors such as conservation. Thus far, this higher coevolution of interface residues has only shown for the small dataset of TMD homodimers investigated by NMR.

CHAPTER 2. MOTIVATION

As described in the previous chapter, single-span membrane proteins take part in countless biological processes. In many cases, their function has been shown to require dimerisation via their transmembrane domains (TMDs). Several features of TMD homodimer interfaces have been proposed, based on case studies. Mutational analyses have identified some features of TM homodimer interfaces, such as simple sequence motifs. However, we fundamentally lack a quantitative understanding of TM homodimer interface properties. Until now, the major bottleneck has been the small amount of natural TMDs with known interfaces, and a lack of quantitative studies.

Major unanswered questions include the following:

How prevalent are glycines at the interface in natural cellular membranes? Are sequence motifs such as GxxxG really indicative of natural TMD interfaces? Which amino acids are overabundant at interfacial positions? Is there a correlation between conservation of amino acids at certain positions in TMDs and their potential to drive self-interaction? Are interface residues more polar than non-interface residues?

The aims of this project were as follows:

- define novel interfaces for a total of 10 TM homodimers using the ToxR assay
- create the first comprehensive dataset of TM homodimer interfaces
    - create the first collection of ETRA data from literature
    - combine with interfaces identified via structural methods (NMR, X-ray crystallography)
- conduct the first quantitative analyses of interface residue properties

CHAPTER 3. MATERIALS AND METHODS

3.1 Experimental methods

3.1.1 Media, antibiotics, enzymes and antibodies

Two types of media were used in this work to grow *E. coli* cells. For the solid media, 1.5% (w/v) agar was added. All media were autoclaved before use.

LB medium (pH 7.0):	
1% (w/v)	Tryptone
0.5% (w/v)	Yeast extract
171 mM	NaCl
Adjust pH with NaOH	

SOB medium (pH 7.0):	
2% (w/v)	Tryptone
0.5% (w/v)	Yeast extract
8.6 mM	NaCl
2.5 mM	KCl
Adjust pH with NaOH	

The following antibiotics were used for the selection of bacterial strains:

Table 3-1: Antibiotics used for selection in liquid or solid media**.**

antibiotic	**final concentration**	**selection**
Ampicillin	100 µg/ml	FHK12 cells
Tetracycline	12.5 µg/ml	PD28 and XL1-Blue cells
Kanamycin	33 µg/ml	pToxRV plasmid

All restriction enzymes and DNA modifying enzymes (i.e. T4 DNA ligase and T4 polynucleotide kinase) were purchased from Fermentas and NEB. The digestion of plasmid DNA using restriction enzymes was performed in accordance with the manufacturer's protocol (NEB, Fermentas).

Antibodies used for Western blotting are listed in Table 3-2.

Table 3-2: Antibodies used for the detection of the maltose binding protein

antibody	dilution	source
Rabbit anti MBP antiserum	1:10,000	NEB
Anti-rabbit IgG Alkaline Phosphatase conjugate	1:7,500	Promega

3.1.2 Plasmids and bacterial strains

The following three *E. coli* strains and plasmids were used in this work:

label	Resist.	genotype	application	reference
FHK12	AmpR	F′lacIq lacZΔM15 proA+B+ ara Δ(lac-proAB) rpsL (ϕ80ΔlacZΔM15) attB::(ctx::lacZ) Ampr	ToxR assay	[8]
PD28	TetR	pop3325 ΔmalE444 Δ(srlR-recA)306:: Tn10	complementation assay	[12]
XL1-Blue	TetR	recA1, endA1, gyrA96, thi, hsdR17 (rK–, mK+), supE44, relA1, lac, [F′, proAB+, laclqZΔM15,: Tn10(Tetr)]	Cloning and plasmid preparation	Stratagene

3.1.3 Oligonucleotides

All primers were solved in ddH_2O and stored with a concentration of 100 pmol/µl at - 20 °C. The working concentration for primers was 10 pmol/µl. All primers used in this study were ordered from Invitrogen. The following table lists the used sequencing primers:

Table 3-3: Sequencing primers

primer	sequence	application
ToxRSeqDown	CCGTTATAGCCTTTATCGCCG	Binds the MBP region for sequencing of TMD region in anti-sense direction
ToxRSeqUp	CAATGTCGTGGCGAATAAATCGGCTC	Binds the ToxR domain for sequencing of TMD region in sense direction.

3.1.4 Preparation of chemically competent cells

Chemically competent cells of *E. coli* strains FHK12, PD28 and XL1-Blue were prepared as described below and transformed using standard heat-shock protocols.

3.1.4.1 Preparation of competent *E. coli* FHK12, *E. coli* PD28

The method of Inoue was used to prepare FHK12, PD28 chemically competent cells [146]. Initially, 250 ml SOB medium was inoculated with 10-12 colonies picked from an LB plate and subsequently incubated at 18 °C and 160 rpm. The optical density of the culture at 600 nm (OD_{600}) was measured at regular intervals. After reaching an OD_{600} between 0.4 and 0.6 the cells. Cells were first cooled on ice for 10 min. Then cells were sedimented by centrifugation at 2500 g for 10 min at 4 °C. The pellet was re-suspended in 80 ml ice-cold transformation buffer and incubated for a further 10 min on ice. Afterwards, the cells were centrifuged again at 2500 g at 4 °C for 10 min and re-suspended in 20 ml ice-cold transformation buffer. Cells were gently swirled, and DMSO added dropwise to a final concentration of 7%. Cells were further incubated for 10 min on ice, distributed into 100 µl aliquots into 1.5 ml tubes, and frozen in liquid N_2. All chemical competent cells were stored at -80 °C before use.

SOC medium (pH 7.0):	
10 mM	$MgCl_2$
10 mM	$MgSO_4$
20 mM	Glucose

Filter all solutions with 0.45 pore size. Store at 4 °C.

transformation buffer (pH 6.7):	
10 mM	PIPES
15 mM	$CaCl_2$
250 mM	KCl

Adjust the pH with KOH. Filter with 0.45 pore size. Store at 4 °C.

3.1.4.2 Preparation of competent *E. coli* XL1-Blue

The preparation of XL1-Blue chemically competent cells was performed using the method of Chung [147]. Pre-warmed LB medium was inoculated with 1:100 v/v of an overnight culture. Cells were grown to an OD_{600} of 0.3. Cells were centrifuged at 1000 g, 10 min, and 4 °C. Cells were resuspended in 1:10 v/v ice-cold TSS buffer. Cells were distributed into 100 µl aliquots into 1.5 ml tubes. Freeze in liquid N_2, store at -80 °C.

TSS buffer:	
10% (w/v)	PEG 3000
50 mM	$MgCl_2$

Prepared in LB liquid medium. Filter with 0.45 pore size. Fresh prepared.

3.1.5 Transformation of competent cells

Chemical competent *E. coli* cells were transformed with plasmid DNA using standard heat shock methods. Initially, competent cells were thawed on ice for 30 s. Approximately 100 ng plasmid DNA (or 5 µl of ligation product) was added to 100 µl chemically competent cells. The mixture was incubated for 10 min on ice. Heat-shock was performed by submerging the cell suspension in a water bath at 42 °C for 60 s. This was followed by 5 min cooling on ice. Afterwards, 800 µl pre-warmed LB medium

was added to the cells, followed by incubation at 37 °C while shaking for 50 min. The cell suspension was then used for the inoculation of LB agar plates or liquid cultures with the appropriate antibiotics.

3.1.6 Extraction of plasmid DNA

The NucleoSpin Plasmid purification kit from Macherey-Nagel was used for the preparation of plasmid DNA. The manufacturer's methods for high copy plasmid purification were used. For extraction of plasmid or other DNA from an agarose gel, the NucleoSpin Extract II DNA extraction kit was utilised according to the manufacturer's protocol.

3.1.7 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate DNA fragments of varying sizes. It was performed in 1% (w/v) agarose gels with 0.5 µg/ml ethidium bromide (EtBr) in 1x TAE buffer using a voltage of 120 V for 30 min. DNA samples were mixed with 6x loading dye (Fermentas) before loading to the gel. A size standard (1 kb gene ladder) was used was used to estimate the size and concentration of DNA fragments. The DNA was viewed under UV light (312 nm).

1x TAE buffer (pH 8.6):	
40 mM	Tris free base
1 mM	Disodium EDTA
20 mM	Glacial Acetic Acid

3.1.8 Determination of DNA concentration

DNA concentration was measured using a UV spectrophotometer at 260 nm in a quartz cuvette using a dilution of 1:40 v/v DNA to ddH_20 or buffer. In addition, the absorption at 280 nm was recorded to detect possible contamination. The purity of the prepared DNA was judged by the ratio of OD_{260} to OD_{280}. An OD_{260}/OD_{280} ratio between 1.8 and 2 the DNA was considered as pure. For a pure DNA solution, an OD_{260} of 1 corresponds to a concentration of double-stranded DNA of 50 ng/µl.

3.1.9 DNA sequencing

DNA plasmids were sequenced in-house using a 3130 Genetic Analyzer (Thermo Fisher Scientific). For each sample, one master mix was prepared to contain the sequencing primer, polymerase, and stop nucleotides (BigDye Terminator v1.1 Cycle Sequencing Kit, Thermo Fisher Scientific). The PCR reaction was performed in a 96 well PCR plate (polypropylene), 86 mm, Sarstedt).

PCR	
Template DNA	125-250 ng
Primer (10 µM)	0.25 µl
BigDye	0.25 µl
5x BigDye buffer	1.00 µl
Adjust to a final volume of 5 µl with ddH2O.	

thermocycling condition	
96 °C	1 min
96 °C	10 s
50 °C	5 s
60 °C	4 min
	Hold at 4 °C

To obtain pure DNA for sequencing, ethanol/EDTA precipitation purification method was used.

1.25 µl of 125 mM EDTA was added to each well of the 96 PCR plate, followed by adding 18 µl of 95% EtOH (final concentration 67-71%). The sample was mixed by inverting. The mixture was rested at room temperature for 15 min away from light and

then centrifuged for 30 min at 3000 x g. The solvent was removed by tapping the inverted plate twice on tissue paper. Then the inverted plate was dried by centrifugation at 185 g for 1 min. 15 µl of 70% EtOH was added, followed by centrifugation at 1650 g for 15 min. The solvent was removed by tapping the inverted plate twice on a tissue. The inverted plate was centrifuged at 185 g for 1 min. The plate was rested for 15 min in the dark at room temperature to evaporate remaining EtOH. The sample was resuspended in 12 µl HIDI buffer. Before sequencing, the last step of denaturation of dsDNA was performed and 95 °C for 2 min then centrifuged shortly.

3.1.10 Cassette cloning

The cassette cloning protocol involves the hybridisation of short DNA fragments so that their overlapping ends mimic the products of a restriction digest [15]. Cassettes were designed to contain ~60 bp encoding the TMD of interest. The hybridised DNA was inserted into restriction digested plasmid vector using standard restriction-ligation techniques. To increase the likelihood that colonies contained a newly ligated plasmid rather than a re-ligated original plasmid, the original plasmid typically contained a TMD with an ApaI restriction site. ApaI restriction was performed before transformation to disrupt the original plasmids. The sense and antisense DNA oligomers were designed individually and ordered as standard DNA primers (Invitrogen). Silent mutations were introduced to avoid hairpin structures and self-annealing. In addition, all primers were optimised for *E. coli* codon usage (https://www.genscript.com/tools/codon-frequency-table). The pToxRV vector was digested with NheI and BamHI in two steps in Tango buffer, for 1 h respectively. The linear, digested DNA was purified by separation on an agarose gel. The DNA

oligomers were hybridised at a final concentration of 1 pmol/µl to form a short double-stranded DNA cassette. Ligation was performed with an excess of the cassette (100:1 molar ratio, 1 pmol cassette + 0.01 pmol vector). The following scheme shows the composition for ligation:

vector and oligonucleotide ligation	
33 ng	plasmid
2 µl	T4 DNA ligase buffer
1 U	T4 ligase
5 U	Polynucleotide kinase
1 pmol	hybridized cassette

Fill up to a volume of 20 µl with dH_2O.

thermocycling conditions	
30 min	37 °C
20 min	22 °C
20 min	16 °C
20 min	12 °C
Hold 4 °C	

An ApaI digestion was performed on the ligation product to remove original pToxRV plasmid. 1 µl of the ligation product was then used to transform XL1-Blue cells which were then spread on LB agar plate containing 33 µg/ml Kan. Clones were picked and plasmids extracted. The TMD and surrounding sequence were confirmed by sequencing with the ToxSeqDown primer by GATC Biotech (Konstanz, Germany). In all cases, the full chromatogram was aligned against the desired plasmid sequence, using the CLC Workbench software (CLC bio, USA).

3.1.11 Cassette cloning of multiple sequence frames each for novel TMDs

For each novel TMD that had never been previously tested in the ToxR system, we used cassette cloning to create four sequence frames (sequence variants) in the pToxRV vector as previously described [21]. This method is an efficient way to identify ToxR constructs that result in a signal when the TMDs interact. The method is based on the assumption that some strong dimers may not give a strong signal, as their

ToxR and TMD domains are not correctly oriented [21]. Four consecutive sequence frames were designed as follows, using IRE1 as an example (Figure 3-1).

original seq	**N-**……**LKDMATIILSTFLLIGWVAFIITYPLSMHQ**……**-C**
IRE1-0	nrasMATIILSTFLLIGWVAFIITgilinp
IRE1-1	nrasATIILSTFLLIGWVAFIITYgilinp
IRE1-2	nrasTIILSTFLLIGWVAFIITYPgilinp
IRE1-3	nrasIILSTFLLIGWVAFIITYPLgilinp

Figure 3-1: Cloning strategy to test four sequence frames. This strategy was used for all novel TMDs whose ToxR activity had never been tested. The Ire1 TMD (UniProt Accession No. O75460) is shown here as an example. In the original protein sequence, the predicted TMD according to UniProt is underlined. Four 20-amino acid sequence frames surrounding the TMD were selected and introduced into the ToxR-TMD-MBP fusion protein. The linker to the ToxR domain (nras) and the MBP domain (gilinp) is shown for each frame in grey. Each frame, labelled 0, 1, 2, or 3 is created via the stepwise insertion of a native amino acid at the C-terminus, along with the stepwise deletion of an amino acid at the N-terminus. Theoretically, these stepwise insertions rotate the potential TMD-TMD interfaces in relation to the ToxR domains, ensuring that at least one of the frames is correctly positioned to give a dimerisation signal.

The four 20-residue frames were designed manually to overlap the central hydrophobic section of the TMD. Four frames were tested in the ToxR system for each TMD listed in Table 4.1, with all TMD sequences tested summarised in Table 8-2. The frame with the highest ToxR signal assumed to have the optimal orientation, and was used for further experiments such as scanning mutagenesis.

3.1.12 ToxR assay

The ToxR assay was conducted essentially as described [70] using the pToxRV vector (Figure 3-2) system [148] that utilises the inducible arabinose promoter for protein expression [76]. Because the araBAD promoter is utilised in pToxRV, arabinose induces transcription of the mRNA corresponding to the fusion protein. In

contrast, whereas transcription is inhibited by isopropyl-β-D-thiogalactopyranoside (IPTG) [149]. To ensure a native level of protein expression, a very low level of transcription was used (0.0025% arabinose, 1 mM IPTG).

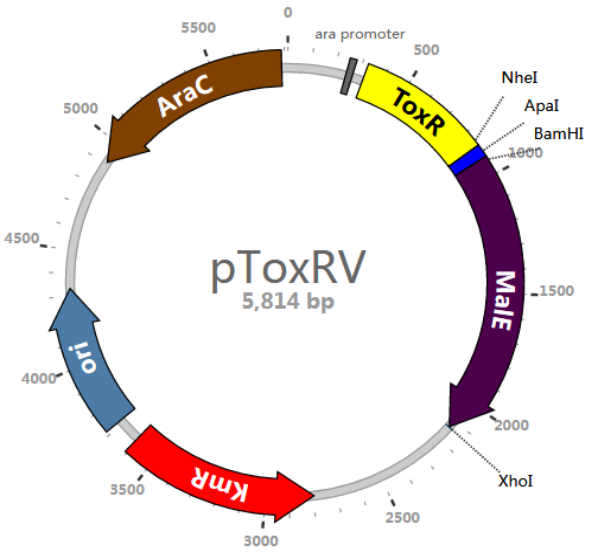

Figure 3-2: The pToxRV plasmid used in the ToxR assay. The ToxR-TMD-MBP fusion protein contains an N-terminal ToxR domain (yellow), a TMD (blue), and a maltose binding domain (MBP, purple). The fusion protein is under the control of the pBAD operator/promoter (black), whose expression is modulated by the AraC (brown). The plasmid contains the high copy origin ColE (light blue) and aminoglycoside O-phosphotransferase gene (KmR, kanamycin resistance) (red). The plasmid map image was created using Angular Plasmid.

Previous published ToxR assays were performed in glass tubes [70]. In this study, the ToxR assay was converted to use a 96 deep well plate (BRANDTECH) sealed with nonpermeable silicone lids (VWR, Germany). This new high-throughput test system can measure 32 samples with four replicates simultaneously in a small volume of only 300 µl medium each well. *E. coli* FHK12 cells containing a pToxRV plasmid were grown in the FHK 12 medium. Cell cultures were incubated for 16 h, 37 °C at 700 rpm in a shaker (orbital diameter 3 mm, Thermomixer HTM, HTA-BioTech, Germany).

FHK12 medium	
Arabinose	0.0025% (w/v)
IPTG	1 mM
Kanamycin	33 µg/mL

Freshly prepared in the LB medium.

Z-buffer (pH 7)	
$Na_2HPO_4.7H_2O$	60 mM
$NaH_2PO_4.H_2O$	40 mM
KCl	10 mM
$MgSO_4.7H_2O$	1 mM

Dissolved in ddH_2O, store at room temperature

Z-buffer with SDS	
Z-buffer	5 ml
SDS solution (10%)	1.6% (w/v)

Freshly prepared.

Z-buffer with ONPG	
Z-buffer	5 ml
ONPG	0.4% (w/v)

Freshly prepared.

Z-buffer with chloroform	
Z-buffer	10 ml
Chloroform	10% (v/v)
β-mercaptoethanol	1% (v/v)

Freshly prepared, vortex for 60 s, settling down at room temperature for > 10 min, use aqueous phase.

For measurement, 5 µl of each overnight culture were transferred to a 96 well microplate (FluoStar, BMG Labtech, Ortenberg, Germany) and mixed with 100 µl chloroform-saturated Z-buffer. The cell density (A_{600}) was determined by using a microplate reader (Molecular Devices, Sunnycal, CA, USA). Cells were then lysed by adding 50 µl of SDS Z-buffer and incubation at 28 °C for 5 min. 50 µl ONPG were added to each well and β-galactosidase activity was obtained by measuring the increase in absorbance (A_{405}) for 20 min at 28 °C using a microplate reader (Molecular Devices, Sunnycal, CA, USA). The initial velocity was calculated using the SoftMax Pro software (Molecular Devices). This was taken as the linear slope over

20 minutes. For samples with higher β-galactosidase activity, the initial velocity was measured using the first 10 data points.

The initial velocity is then divided by A_{600} to yield β-galactosidase activity per cell concentration as follows

$$m = \frac{v}{a} \qquad \textbf{Equation 3-1}$$

Where *m* is the β-galactosidase activity per cell in Miller Units, and *v* is the initial velocity, and *a* is the cell concentration measured using the absorbance at a path length of 0.4 cm.

3.1.13 Q5 site-directed mutagenesis

Mutagenesis of single amino acids in the TMD of pToxRV vectors was performed by Q5 site-directed mutagenesis, as developed by NEB. The method is based on the complete PCR amplification of the plasmid of interest by primers containing the desired mutation. Blunt-end cloning is used to religate the linear PCR product to a circular plasmid before transformation into *E. coli*. To ensure that most colonies contained plasmids formed by ligation (rather than the original template plasmid), the ligation mixture was digested with 10 U DpnI for at least 2 h at 37 °C before transformation. DpnI cleaves very often within methylated DNA, such as the original plasmid isolated from *E. coli*, but does not cut the PCR-amplified DNA containing the mutation of interest.

All the mutants were confirmed by DNA sequencing (Section 3.1.9). The full sequencing read was compared with the predicted plasmid using CLC Workbench.

volume	enzyme
0.2 µl	NEB Phusion polymerase
4 µl	HF buffer (supplied with the polymerase)
0.3 µl	Plasmid template (100 µg/µl)
1 µl	Forward primer (10 pmol/µl)
1 µl	Reverse primer (10 pmol/µl)
0.4 µl	dNTP
12 µl	ddH_2O

ligation	
5 µl	PCR product
2 µl	10x Tango buffer
2 µl	PEG 4000 (from T4 DNA ligase kit)
0.25 µl	T4 Polynucleotide kinase (1 U/µl)
1 µl	T4 DNA ligase (10 U/µl)

thermocycling condition	
98 °C	30 s
98 °C	10 s
60 °C	20 s
72 °C	30 s/kb
72 °C	5 min
4 °C	hold

During scanning mutagenesis of a TMD of interest, each TMD residue was first mutated to Ala, with the exception of Gly and Ala residues, which were mutated to Ile [83].

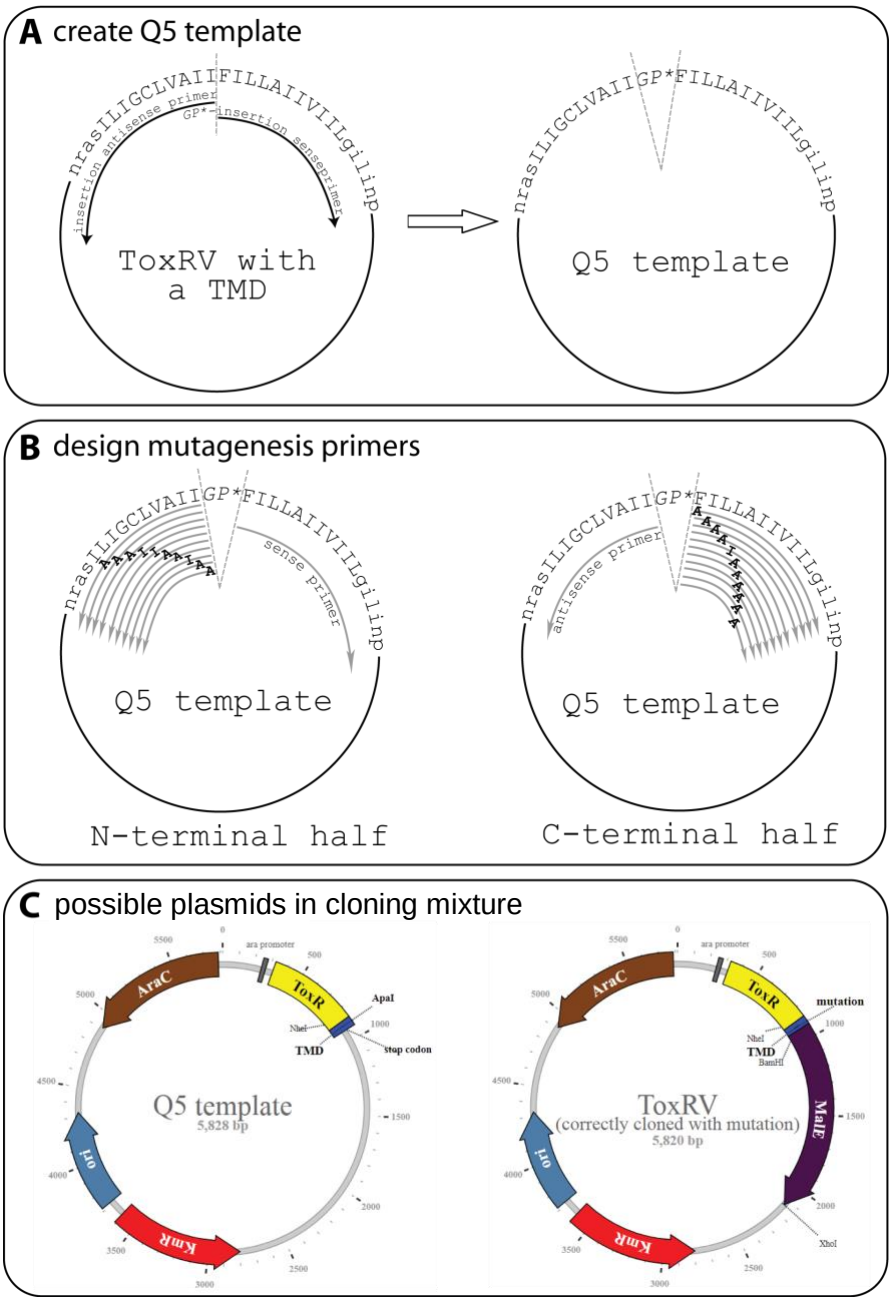

Figure 3-3: Protocol for the Q5 mutagenesis. Initially, the ToxRV plasmid was used as a template to insert the TMD of interest by cassette cloning. (A) Using Q5 mutagenesis, the sequence gggcccta was inserted into the middle of the TMD encoding region. This results in a frameshift, truncating the fusion protein and add a new restriction site. The inserted sequence (gggcccta) was added to the 5´ end of the forward primer. The created Q5 template was then used for mutagenesis. (B) Design of mutagenesis primers. To conserve resources multiple mutagenesis primers were paired with a single partner sense or antisense primer. (C) Plasmids in the cloning mixture after ligation. Even successful ligations contained a proportion of the original template plasmid, which could give rise to colonies on a plate. By using a dedicated Q5 template, correct clones can be further selected. The ToxRV plasmid correctly cloned with the mutation will skip the stop codon and express the

MBP protein. The plasmids map was created by the Angular Plasmid. The differences between the two plasmids for selection purpose is shown in Table 3-4.

Table 3-4: Selection methods that could be used to identify correct clones of Q5 mutagenesis.

	Q5 template plasmid	**ToxRV plasmid (correctly cloned)**	**selection method**
DpnI site	No	Yes	restriction before plating
ApaI site	Yes	No	restriction before sequencing
MBP	No	Yes	PD28 complementary assay
Full TMD	No	Yes	ToxR β-galactosidase assay
fusion protein size	66 kDa	23 kDa	SDS-PAGE

To confirm putative interface, additional mutations were attempted for positions that showed a 25% decrease or an increase of ToxR activity in comparison to the wildtype. In several cases, degenerate codons were used to yield mutations to multiple residues (Table 3-5). This increased the number of mutations achieved, without increasing the cost of mutation primers. An example was the insertion of the degenerate codon NRS, which encodes an equal distribution of Pro, Leu, Ala or Val (Table 3-5). Where degenerate primers were used, colonies were restreaked on LB-agar an extra time after transformation before isolation of plasmid, ensuring that clones contained only a single plasmid variant.

Table 3-5: Degenerate codons used for site-directed mutagenesis

codon	possible amino acids
GAD	I/L/F
GAV	V/F/L
NRS	P/L/A/V
VTC	V/I/L
DTC	F/I/V
NTC	L/F/V/I
SYN	P/A/L
GAN	V/I/L/F
GAH	I/V/F
BTC	V/L/F
HTC	I/L/F

3.1.14 Western blot and MBP complementation assays

The expression level of the ToxR proteins was confirmed by Western blotting with an antiserum recognising the MBP domain. 10 µl of whole cell lysates were loaded onto a 20% SDS-PAGE gel and then transferred to PVDF membranes for 1.5 h at 140 milliamps. Blots were blocked using 3% (w/v) non-fat dry milk powder in TBS-Tween buffer. They were then incubated with rabbit anti-MBP (NEB), followed by anti-rabbit-IgG (alkaline phosphatase conjugated, Promega).

The membrane integration of novel ToxR-TMD-MBP fusion proteins was confirmed using the PD28 complementation assay. The pToxRV plasmids were transformed into chemically competent *E. coli* PD28 cells. The overnight culture was centrifuged and the supernatant was discarded. The pellet was washed twice with PBS to remove the remaining LB medium. The cells after the last wash step were resuspended in the PBS. They were then transferred to M9 minimal medium containing 0.4% (w/v) maltose as the only carbon source. Growth occurred only if the C-terminal MBP-domain of ToxR chimeric proteins is successfully translocated to the periplasmic

space. Cell density (A_{650}) was measured every 24 h until 72 h and the growth kinetics compared to that of a construct lacking the TMD domain (ΔTM).

3.1.15 Orientation-dependence

For novel TMDs tested for the first time in the ToxR assay, the orientation dependence was calculated as previously described [21]. The orientation dependence is the difference between the highest and lowest ToxR signal for all four sequence frames. The orientation dependence is proposed to measure the optimal orientation of a TMD that has the highest relative affinity. The orientation dependence, *o*, was calculated as follows:

$$o = 1 - \frac{i}{a} \qquad \textbf{Equation 3-2}$$

where *i* and *a* are the lowest and highest mean β-galactosidase activity, respectively. The orientation dependence is therefore a value between 0 and 1. The *o* value close to 0 means the β-galactosidase activity is similar between four orientations, denoted as the low orientation-dependence. A value close to 1 represents high orientation-dependence. A TMD with more than 40% difference ($o > 0.4$) in reporter gene activity between different orientations is considered to be orientation-dependent. TMDs with $o < 0.4$ are denoted as weakly orientation-dependent.

3.2 Software and data repositories arising from this study

The vast majority of bioinformatics, statistical analysis and figure creation was conducted using the programming language Python, version 3.4. Coding tools

included the ipython/Jupyter notebook, and the integrated-development-environment, PyCharm (JetBrains, Czech Republic). All relevant code used in this study, including figure creation, has been released within one of the following open-source python packages.

datoxr

repository: https: //bitbucket.org/yaoxiaorepos/datoxr

PyPI repository: https://pypi.org/project/datoxr/

author: Yao Xiao

contributors: Mark Teese

content: a pipeline for analysis of ToxR data, automatic calculation of disruption, automatic identification of interface residues, and also the creation of figures analysing residue properties

installation: pip install datoxr

thoipapy

repository: https://github.com/bojigu/thoipapy

PyPI repository: https://pypi.org/project/thoipapy

author: Bo Zeng (Dmitrij Frishman group, Technische Universität München)

contributors: Mark Teese

content: a pipeline for download of homologous TMD sequences, calculation of residue properties including conservation, polarity and coevolution, and creation of a machine-learning classifier for the prediction of homotypic TM interface residues

pytoxr

repository: https://github.com/teese/pytoxr

PyPI repository: https://pypi.org/project/pytoxr/

main author: Mark Teese

content: some functions for the automated analysis of ToxR data

All data generated in this study has been released to the scientific community in the following repository:

repository: https://osf.io/txjev/.

The repository includes all experimental ToxR and scanning mutagenesis data, all data collected from literature, all defined interface residues, and all interface properties analysed in the course of this study.

3.3 Creation of the *E. coli* Transmembrane Reporter Assay (ETRA) dataset

3.3.1 Extraction of scanning mutagenesis data from the literature

All available scanning mutagenesis data in literature were collected. The selection was restricted to studies that aimed to identify natural TMD interfaces by disruptive mutations, using assays of helix homodimerisation in the *E. coli* inner membrane. TMDs were retained if the homodimerisation interface was identified with a high certainty, at least 75% of the TMD residues had been mutated, without major gaps in the data (at least 15 consecutive residues). This step yielded 12 proteins (Table 5-1). These TMDs were mainly analysed with ToxR-based experimental methods,

including ToxR, TOXCAT and dsTβL. Only one exception, NS4A of Hepatitis C, had been investigated in a study using the alternative GALLEX assay [33].

For TMDs extracted from literature, the scanning mutagenesis values were either taken from supplementary data or calculated from the bar height in figures. The dsTβL data for GpA and ErbB2 were kindly provided by Assaf Elazar. The data corresponded to Figure 5 of Elazar et al. [22]. The dsTβL (2 TMDs) and GALLEX (1 TMD) data were quite different from the ToxR-based assays that comprised all other investigated TMDs, and therefore required normalisation to yield comparable values. The dsTβL ΔΔapp association data was normalised by dividing by the mutation with the highest signal (L13T for GpA, I18G for ErbB2). These normalised values were considered equivalent to the fraction of wildtype activity in other studies. The original GALLEX data was inverted in comparison to ToxR signals (i.e. stronger dimers gave lower values), and the signal is typically represented on a log scale, in comparison to the linear signal derived from ToxR. To obtain comparable values to ToxR, for each GALLEX mutation the cube root of the fraction of wildtype activity was taken. Then the data was normalised from 0 to 1, so that the pBR322 control (no dimer) was equal to zero, and the wildtype was equal to 1. In this way, mutations with higher dimerisation than the wildtype gave scores larger than 1. The normalised values were further processed as for the ToxR data. One of the ToxR studies from literature also required normalisation. For ITGA2B the reported wildtype value was extremely low in comparison to values derived from mutations [40]. To obtain comparative values, the signal for each mutation was normalised to the median of all mutations, rather than to the wildtype.

3.3.2 Calculation of the disruption to the dimer signal after mutation

Disruption to the dimer signal was calculated for all mutations from all TMDs from experimental and literature data. In cases where a single position was mutated to multiple amino acids, the homodimer disruption was calculated by averaging the disruption for all mutations at that position. The calculation was according to:

$$d = 1 - \frac{w - m}{w} \qquad \textbf{Equation 3-3}$$

Where *d* is disruption, *w* and *m* are the self-affinity scores for the wildtype and the mutated construct. A positive *d* value indicates a decrease in the strength of self-interaction; a negative *d* value indicates that mutations at that position increased self-affinity.

3.3.3 Definition of interface residues from ETRA disruption data

Until now, no other study has attempted to objectively define interface residues based on the ETRA data for multiple TMDs. A subjective cut-off in disruption was therefore chosen (disruption value above 0.24) that gave at least three interface residues for all TMDs in the ETRA dataset.

To examine the α-helicity of the interfacial residues, the disruption index was fitted to an α-helical periodicity using the leastsq function of the scipy.optimize module in python and the formula y = a*sin(bx+c)+d. Fitting was conducted assuming perfect α-helicity (periodicity of 3.6) by keeping b constant at 2π/3.6.

3.4 Creation of NMR and crystal datasets of self-interacting TMDs

The interface residues of self-interacting TMDs from NMR and crystal studies were provided by Bo Zeng, as part of the co-authored publication of Yao Xiao and Bo Zeng "Properties and prediction of homotypic transmembrane helix-helix interfaces." Full methods are available in the associated publication. All code for the identification of interacting TMDs and interface residues from structural data is publicly available in the THOIPApy python package of Bo Zeng (https://github.com/bojigu/thoipapy/wiki).

The small number of TM homodimer structures obtained via NMR have been reviewed extensively, and used for validation of the prediction algorithms PREDDIMER [86], TMDOCK [87], and EFDOCK-TM [36]. The obtained NMR dataset was based on the 13 default dimer structures included in the validation of EFDOCK-TM by Wang et al. [36]. The dataset was updated by including the new NMR structure of the toll-like receptor 3 (PDB 2mk9, UniProt O15455) [43]. Three TMDs with NMR structures, GpA [1], BNIP3 [5] and ErbB2 [13], were excluded because they were already in the ETRA dataset. The dataset was made non-redundant to 60% identity of full protein and 40% of the TMD using CD-HIT [150]. The final NMR dataset contained 8 proteins (Table 8-1).

Unlike the well-studied NMR dataset, until now the TM "homodimer-like" helix pairs in crystal structures have never been identified and analysed. The "crystal" dataset was extracted from crystal structures using the membrane residue annotations obtained from the PDBTM [90]. 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 [150] using amino acid sequence identity cut off of 40% for the full protein. Interfacial residues for the crystal dataset were defined as described in Section 3.4. Self-interacting helices were retained that had at least four unique interacting residues. A second round of CD-HIT [150] redundancy reduction was

conducted based only on the TMD sequence. The final dataset was non-redundant to 40% amino acid identity of the full sequence, and 60% amino acid identity of the TMD. The final, non-redundant crystal database consisted of 25 parallel, self-interacting TM helices. The vast majority of these (23/25) were helices from oligomeric multipass proteins.

The interface of TMDs with known structures was defined by the heavy atom distance [132]. To be more specific, an interfacial residue is defined if any heavy atom (non-hydrogen atom) of the residue is within threshold diameter ($D_{thr)}$ in the interacting protein chain [151]. In this study, the closest heavy-atom distance between the residue of interest and all other residues in any identical TMDs in the structure was calculated. The threshold of 3.5 Å was selected to ensure that the interfacial residues were consistent between this and previously published studies (Table 8-1).

The final crystal dataset had a total of 167 interacting residues and 347 non-interacting residues. A unique feature of the crystal dataset was the presence of heterotypic contacts, comprising residues that had non-self contacts with residues other TM helices. The main focus of this study was to identify differences between TM homodimer interface residues and lipid-exposed non-interface residues in single-pass proteins. The heterotypic contacts were therefore considered undesirable, as they shared little in common with the lipid-exposed “non-interface” residues of the ETRA and NMR datasets. For most statistical analyses comparing interface and non-interface residues, these heterotypic contacts were removed from the dataset. They were defined exactly as for the homotypic contacts as described in Section 3.4, based on a closest heavy-atom distance cutoff of 3.5 A. A total of 55 heterotypic contacts were identified.

For comparisons of residue properties (e.g. conservation) between the interface and non-interface residues (Figure 5-6, Figure 5-8, Figure 5-10, Figure 5-11, Figure 5-13, Figure 5-14, Figure 5-15**,** Table 8-5), 55 (from 347) non-interacting residues were classified as heterotypic contacts and excluded from the analysis. For motif analyses, however, a continuous TMD sequence was required. Heterotypic contacts remained in the dataset, and all 347 residues were regarded as “non-interface” (Figure 5-16, Figure 5-17).

3.5 Creation of the homotypic TM dataset

Questions regarding the conservation, relative polarity, coevolution, and motifs of TM homodimer interfaces have until now been approached using case studies, artificial selection, or the small, highly redundant NMR dataset. To quantitatively and objectively analyse interface residue properties, all the parallel, self-interacting helices from ETRA, NMR and crystallographic studies were collected and combined.

The homotypic TM dataset was created by combining non-redundant ETRA, NMR and crystal datasets respectively. The homotypic TM dataset contained 54 TMDs in total, comprising 21 ETRA-derived interfaces, 8 NMR-derived interfaces, and 25 crystal-derived interfaces. Each subset, and also the entire database was non-redundant to 40% and 60% sequence identity of the full sequence and TMD, respectively. The proportion of interacting residues was similar among the ETRA (21%), NMR (39%) and crystal (32%) datasets. There are 347 interfacial residues and 770 non-interfacial residues included (55 heterotypic contacts from crystal TMDs were removed). This corresponded to an average of 6.4 interfacial residues per TMD, comprising 31% of the TM residues.

3.6 Determination of residue properties

3.6.1 Method harmonisation

To understand the properties of interface residues, a number of different residue features were calculated for each amino acid residue, including residue conservation, polarity, coevolution (a.k.a. covariance) to neighbouring residues, and the predicted depth in the membrane. Such features also formed the basis of a machine learning algorithm, which is described in the PhD thesis of Bo Zeng and is not discussed here. For publication in a peer-reviewed journal, the methods were harmonised, as described in detail in the acknowledgements. The initial python code used for the extraction of features such as conservation and polarity from multiple sequence alignments is publicly available in the datoxr package (https://yaoxiaorepos@bitbucket.org/yaoxiaorepos/datoxr.git), released by myself as part of this study. After harmonisation, methods for the calculation of residue properties were located in the THOIPApy package of Bo Zeng (https://github.com/bojigu/thoipapy). This strengthened the conclusions of the combined study by ensuring that the detailed analysis of sequence properties (shown here) utilised exactly the same underlying data as the machine-learning algorithm by Bo Zeng.

3.6.2 Multiple sequence alignments against homologues

Multiple sequence alignments (MSA) 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, position-specific scoring matrices (PSSM) were calculated to quantify the evolutionary profile of each amino acid in a TMD. The PSSM contained the frequencies of all 20 amino acids in each MSA column.

3.6.3 Sequence conservation

The conservation of a residue in this study refers to a normalised form of entropy, with higher values indicating highly conserved residue positions. The conservation was calculated from Shannon entropy ($S_{entropy}$) as follows:

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

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

Equation 3-4

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.

3.6.4 Polarity

Polarity was calculated for each position in the MSA, rather than the single residue in the sequence. 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 Engelman (GES)

hydrophobicity scale [138]. The final polarity score represented the mean 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).

In TM helix-helix interactions, a high polarity is only associated with an important functional role if the residue is located in the hydrophobic core of the membrane [118, 124, 131]. In the crystal datasets, the TM residues determined by PDBTM often contained highly polar residues at the start or end of a TM sequence. This polarity usually indicated contact with the polar solvent, rather than an important role in the TM helix-helix interactions. A normalised version of polarity was therefore created, “relative polarity,” which tended to be higher for polar residues in the hydrophobic core, and lower for polar residues at the lipid-water interface. The relative polarity at position i is the polarity at the position i (calculated as described above) divided by the mean polarity of the 6 surrounding residues (i-3 to i+3, excluding i). For relative polarity, an Arg residue in the centre of the TMD therefore scores much more highly than an Arg residue in the juxtamembrane region.

3.6.5 Coevolution

In order to understand interface features, it was necessary to convert the pairwise coevolution scores to a single representative value at each residue position. MSA of homologues forms a protein family was used to search for correlated mutations that occur between contacted or spatially proximity residues.

Residue coevolution scores were derived based on DI. These were calculated as previously described [143], using the FreeContact implementation [144]. Briefly, for a pair of residues *i and j,* DI was calculated according to the following equation:

$$DI(i,j)=\sum_{A_i,A_j=1}^{q} P_{ij}^{Dir}(A_i,A_j)\ln\left(\frac{P_{ij}^{Dir}(A_i,A_j)}{f_i(A_i)f_j(A_j)}\right)$$ **Equation 3-5**

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 maximization approach, which calculates correlation scores for each pair of residues while taking into account all other pairs.

For position i in the TMD, for example, the direct information (DI) score was calculated 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 was referred here as DImax. For all figures in this study where a single representative coevolution metric was required, the DImax was used.

For the DI, the standard deviation of the values decreased with the number of homologues. To minimise these effects, the coevolution feature was normalised between 0 and 1 within each TMD.

3.6.6 Depth in the bilayer

The depth in the bilayer is a simple measure of the position in a TMD sequence, from the most central (value = 1) to the most peripheral (N-terminal or C-terminal) TMD residue (value = 0).

3.7 Analysis of interface residue properties

3.7.1 Calculation of amino acid frequency

The frequency (*f*) of each type of amino acid in the TMD sequence was calculated as the number of that particular amino acid divided by the total number of all amino acids.

$$f = x/t \qquad \textbf{Equation 3-6}$$

Where *x* is the number of a particular amino acid, and *t* is total numbers of all residue in TMDs.

3.7.2 Calculation of amino acid enrichment in the interface

The enrichment (*e*) of a particular amino acid is defined as the ratio of the frequency of occurrence of that amino acid at the interface compared to its frequency in the TMD [152, 153]. The relative enrichment of each amino acid types at the interface was calculated according to:

$$e = \frac{x/i}{t/a} \qquad \textbf{Equation 3-7}$$

where *x* is the number of the residue at the interface, *i* is the total number of interfacial residues, *t* is the total number of that residue in TMDs, *a* is the total number of all residues in TMDs.

3.7.3 Sequence conservation logo

WebLogo 3.0 (http://weblogo.threeplusone.com/create.cgi) was used with the standard parameters (no adjustment for composition, colour scheme ‘auto’, no error bars should be displayed) to visualise conserved residues within multiple alignments of TMD sequences [154].

3.7.4 Mapping of interface residues to a model helix

The FMAP [155, 156] is a web tool for the prediction of the structure of α-helices based on an input TMD sequence. The FMAP online server (http://www.membranome.org/server.php) can be used to predict α-helices in aqueous solution, protein molten globule state, micelles, and lipid bilayers.

3.7.5 Statistical significance

All pairwise comparisons were conducted with the *t*-test using bootstrapped data, as implemented in the python bootstrapped module. The bootstrap technique was introduced by Efron [157] as an alternative way to the usual non-parametric methods. It is a general-purpose technique for obtaining estimates of features of statistical estimators without making assumptions about the distribution of the data.

Similarly, confidence intervals in this study were made by using bootstrapped data, using 95% CI unless otherwise indicated. *P*-values were represented using standard symbols as follows: *, $p<0.05$. **, $p<0.01$, ***, $p<0.001$.

CHAPTER 4. RESULTS I: EXPERIMENTAL DETERMINATION OF INTERFACES IN NATURAL MEMBRANES

4.1 Identification of TMDs with strong self-affinity

There are several case studies where ETRA techniques such as ToxR have been used in combination with scanning mutagenesis to identify homotypic TM interfaces [3, 6, 9, 11, 21, 22, 25, 28, 31-33, 35, 39, 40, 44, 116]. This approach is most successful when the dimer strength is high, giving an easy discrimination between the wildtype signal, and the signal from mutations that decrease dimer strength. Therefore, the aim of the experiments described in this section was to identify TMDs with strong self-affinity in the ToxR assay (>80% GpA), which could be used later for the mutagenesis as described in Section 4.2. Two methods were used to identify TMDs with high self-affinity. Firstly, a bioinformatic identification of TMDs with a high sidedness of sequence conservation (Section 4.1.1), and secondly, based on the previous reports of their high dimerisation in literature (Section 4.1.2).

4.1.1 Self-affinity of a selection of human TMDs with a high conservation moment

The conservation moment is a measure given to a TM helix that describes the sidedness of conservation, assuming perfect α-helicity [83]. It had been proposed that there is a correlation between conservation moment and TMD affinity [15]. Therefore TMDs with a high conservation moment were selected from the human bitopic proteome and tested for self-interaction. Seven TMDs with a very high conservation moment (> 0.4) according to Ried et al. [158] were tested for self-affinity in the ToxR

system (Figure 4-1, supplementary Table 8-2). In each case, individual multiple sequence alignments were examined carefully to confirm that one side of the helix was indeed highly conserved.

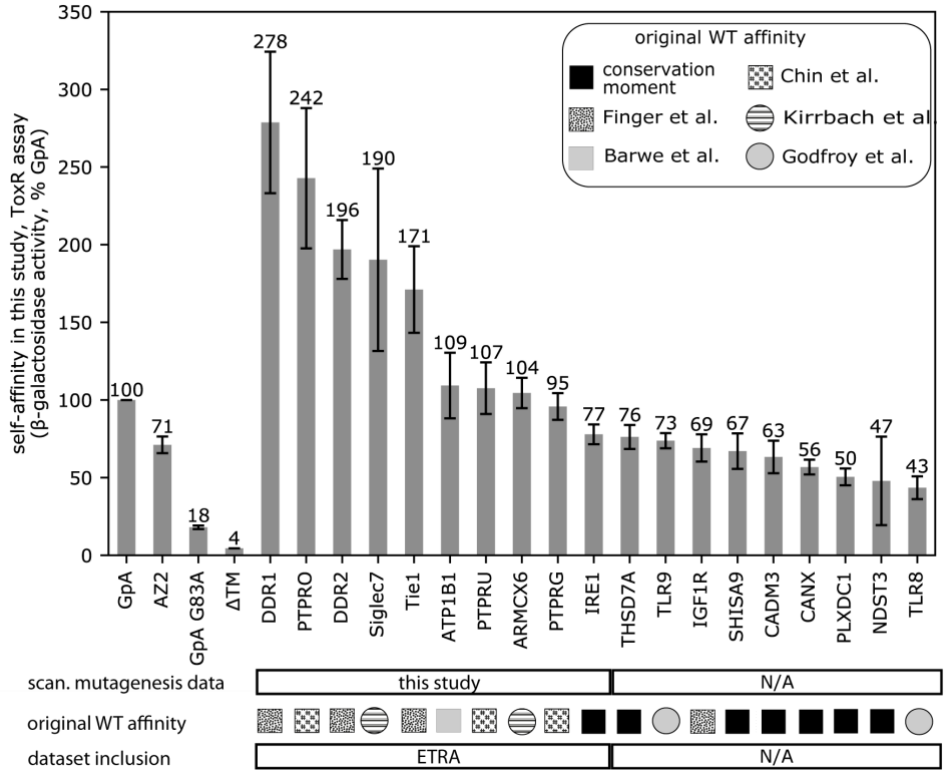

Figure 4-1: All TMDs tested for self-affinity by ToxR assay in this study. Shown are the data from thirteen TMDs tested from literature (Table 4-1), as well as seven TMDs tested due to their high conservation moment (Table 8-2). For TMDs tested due to their high conservation moment, the result from only the highest sequence frame is shown. All activities were normalised to the signal from GpA wildtype. Activities greater than AZ2 are considered as strong interaction. Activities lower than AZ2 and above G83A are considered as moderate interactions, and less than G83A were considered as weaker interactions. Data were obtained from at least 3 independent biological replicates (mean ± SEM). The availability of scanning mutagenesis data, and inclusion in the ETRA dataset are indicated in boxes below the x-axis. Literature references refer to the original description of the TMD as a strong homodimer with either the ToxR or TOXCAT assay.

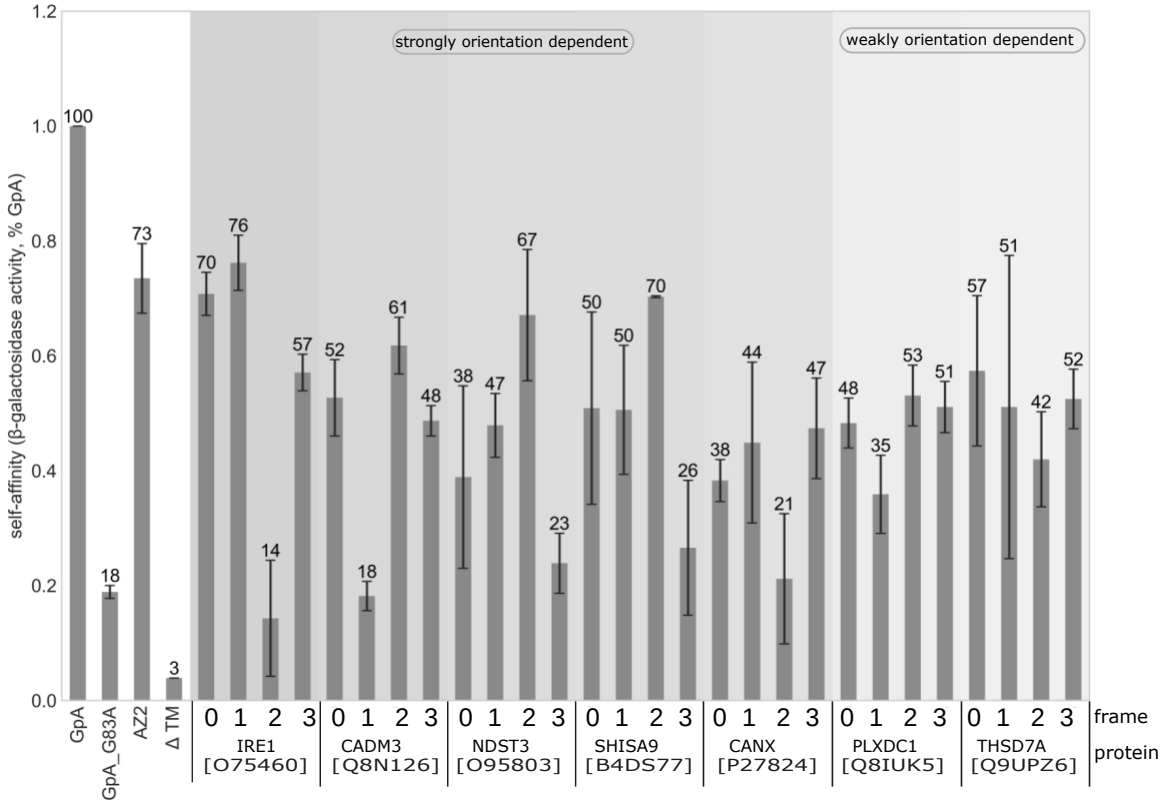

Figure 4-2*:* TMDs with a high conservation moment had a strong orientation dependence but moderate self-interaction. Data represent relative self-affinity activities (GpA= 100%) as measured with the ToxR reporter assay. For each TMD, the orientation-dependence was calculated as described in Section 3.1.15 by Equation 3-2. Results are sorted according to the ORD of interaction. Five TMDs show> 40% difference in relative self-affinity score which indicates strong orientation-dependence (dark shading of background). Two TMDs show weak orientation-dependence with 22% difference in relative self-affinity score (light shading of background) (n>3, mean ± SEM). See Table 8-2 for sequences.

For each TMD, four sequence “frames” were tested in the ToxR system as previously described [21]. This method ensures that the ToxR signal is not masked by orientation effects between the ToxR transcription activation domain and the TMD domain of the fusion protein. TMDs with large differences between the frames are described as having high “orientation dependence”, which is thought to indicate a high specificity of the TMD self-interaction [21]. Orientation dependence was calculated as described in Section 3.1.15. This revealed that five of seven TMDs tested showed a clear preference for one orientation (orientation dependence > 0.4, Figure 4-2). This was a

positive sign that may indicate a specific interface [21]. However, for these TMDs the ToxR signal of the highest frame was only modest. This would make interface detection scanning mutagenesis more difficult, as only a small decrease in signal could be detected. From the seven TMDs tested, only IRE1 was therefore chosen for scanning mutagenesis. To identify more TMDs with high self-affinity, others were from literature as described below.

4.1.2 Self-affinity of TMDs previously claimed to self-interact

Twelve TMDs suggested to have high self-association in an ETRA assay according to the literature were tested for ToxR activity in this study (Figure 4-1, Table 4-1). Most of the previous studies had used the TOXCAT system rather than the ToxR assay used here. Overall, 9/12 showed a strong self-affinity. In fact, a scatterplot revealed a strong correlation ($R^2 = 0.7$, Figure 4-3) between the previously published ToxR/TOXCAT signal, and the ToxR activities measured in this study. This showed that high self-affinity in the *E. coli* membrane is usually reproducible, even when a different ToxR system is used for analysis.

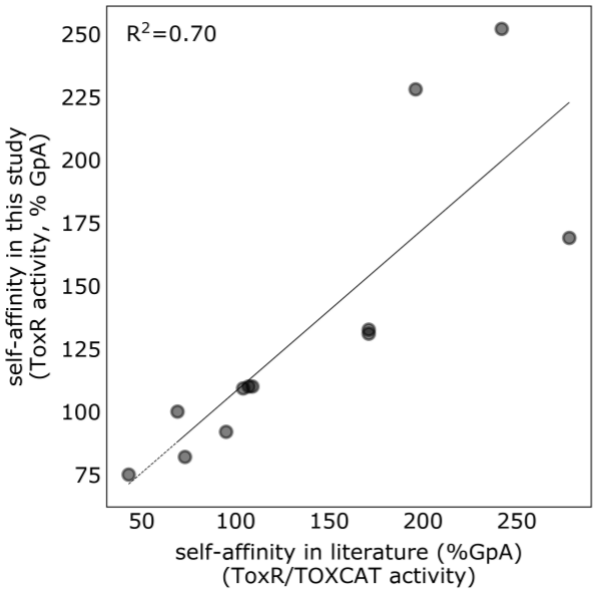

Figure 4-3: High self-affinity was confirmed for most of the twelve TMDs extracted from the literature. The TMDs listed in Table 4-1 were tested for self-affinity in the ToxR assay. Most of the self-affinity data from literature were obtained using the TOXCAT assay, rather than the ToxR assay used here.

The twelve TMDs included three of the receptor-like protein tyrosine phosphatases examined with TOXCAT by Chin et al. [9], PTPRU, PTPRG and PTPRO. All these three TMDs showed strong self-affinities in the ToxR system, and were selected for scanning mutagenesis. From the set of the toll-like receptors examined by Godfroy et al. [4], TLR8 and TLR9 were tested, however neither of them showed a high level of dimerisation. From the set of human receptor tyrosine kinases examined with TOXCAT by Finger et al. [3], DDR1, DDR2, LTK, IGF1R and TIE1 were tested. Of these, DDR1, DDR2 and TIE1 showed strong dimerisation in the ToxR system. Because no previous ETRA study has ever characterised the interface of two homologues, both DDR1 and DDR2 were both selected for scanning mutagenesis. ATP1B1 from Barwe et al. [25] was tested, which gave a high signal in the ToxR assay. From the ToxR experiments of Kirrbach et al. [21], Siglec7 and ARMCX6 were both confirmed to have a high level of dimerisation [21].

Table 4-1: TMDs tested in this study due to their apparent high self-affinity in previous publications

#	protein (acc[a])	TMD sequence	length	reference	self-affinity literature (GpA%)
1	TLR8 (Q9NR97)	VTAVILFFFTFFITTMVMLAALA	23	[4]	75
2	TLR9 (Q9NR96)	FALSLLAVALGLGVPMLHHL	20	[4]	82
3	DDR2 (Q16832)	ILIGCLVAIIFILLAIIVIIL	22	[3]	228
4	DDR1 (Q08345)	ILIGCLVAIILLLLLIIALML	22	[3]	169
5	TIE1 (P35590)	LILAVVGSVSATCLTILAALLTLV	24	[3]	131
6	AT1B1 (P05026)	LLFYVIFYGCLAGIFIGTIQVMLLTI	26	[25]	110
7	IGF1R (P08069)	LIIALPVAVLLIVGGLVIMLYVF	23	[3]	100
8	PTPRU (Q92729)	LILGICAGGLAVLILLLGAIIVII	24	[9]	110
9	PTPRG (P23470)	IIPLIVVSALTFVCLILLIAVLV	23	[9]	92
10	PTPRO (Q16827)	VVVISVLAILSTLLIGLLLVTLIIL	25	[9]	252
11	ARMCX6 (Q7L4S7)	REVGWMAAGLMIGAGACYCV	20	[21]	109.2
12	Siglec7 (Q9Y286)	VLLGAVGGAGATALVFLSFC	20	[21]	132.6

a Accession number (acc) is taken from the UniProt database.

4.1.3 Description of the nine unique TMDs chosen for scanning mutagenesis

The aim of this experimental section was to identify TMDs with high self-affinity (>80% GpA) for further scanning mutagenesis. This was successfully achieved, mostly by testing TMDs proposed in the literature have high self-affinity. The final set of proteins for scanning mutagenesis consisted of the following:

- eight TMDs previously shown to have high self-affinity in literature, confirmed here with the ToxR assay (DDR1, PTPRO, Siglec7, Tie1, ATP1B1, PTPRU, ARMCX6 and PTPRG).
- one TMD (IRE1) that had a high sidedness of conservation, and also a moderate-to-high level of self-affinity.

- one further TMD from literature, DDR2, that was homologous to DDR1. This was used to test the theory that homologous TMDs have similar homodimer interfaces.

The 10 TMDs all had strong self-affinity above 77% GpA. All these TMD sequences were checked for redundancy by the CD-HIT with the threshold of 70% identity of full protein. Only DDR1 and DDR2 were related, showing an amino acid identity of 71% for the TMD sequences and 53% similarity for full protein. The remaining proteins were non-redundant (amino acid identity < 20%).

4.2 Scanning mutagenesis of human TMDs with strong self-affinity reveals novel interfaces

Several studies have shown that TM helix-helix interaction is highly sequence-specific. It can be disrupted or enhanced by conservative amino acid changes at the most sensitive positions [67]. These mutation-sensitive positions are assumed to comprise the dimer interface. Residues that can be mutated without showing a major effect on dimerisation are suggested to be lipid-facing.

Scanning mutagenesis was conducted by systematically mutating each residue and monitoring their effects on dimerisation in the ToxR assay as described in Section 3.1.13. Overall, 294 mutations were conducted at 224 positions, of which most (179) were to Ala, and a large number (54) to Ile. The average number of mutants was 29.4 per TMD. Until now, the largest scanning mutagenesis study in a ToxR system has investigated only 2 TMDs [22]. This is therefore the largest experimental study of its kind. All previous studies conducted scanning mutagenesis for only one [3, 6, 9, 11, 21, 22, 25, 28, 31-33, 35, 39, 40, 44, 116], or two TMDs [24]. This high-throughput was achieved by modest improvements to existing methods, including reductions in

the number of necessary cloning primers (Section 3.1.13), the development of rapid methods to correctly identify correctly cloned plasmids (Table 3-4), and the conversion of ToxR growth assays to a high-throughput 96-well format (Section 3.1.12).

For positions that yielded altered dimerisation in comparison to the wildtype, Western blots (supplementary Figure 8-4) and a maltose complementation assay (PD28 assay, supplementary Figure 8-1) were conducted to confirm that the decrease was not due to insufficient protein expression or a failure of the constructs to properly insert into the membrane. The Western blot analysis resulted in the exclusion of some mutations for siglec7 (G4L, L17V, L17I, L17F) from further analysis, which showed both low expression and ToxR activity (data not shown). To assess a possible correlation between relative ToxR activity and relative cell growth, all measured ToxR activities (normalised to their corresponding wildtype values) of single mutants were plotted against their respective growth rate (Figure 8-2). The result showed that the efficiency of membrane integration does not correlate with the measured TMD affinity of TMD-TMD interaction ($R^2 = 0$). The change to ToxR β-galactosidase signal can therefore be attributed to change in TMD self-affinity. An analysis of the literature shows that most of the proteins have known vital roles in the cell, several of them had proposed roles for homodimers. In most cases further work (e.g. mutagenesis + functional assay) is required to understand the full implications of the TMD interface determined here.

4.2.1 Homotypic interaction of the ATP1B1 TMD is mediated by glycines

Within the cell, Na/K-ATPase generates the transmembrane potential via ion transport. This is an ATP-dependent process where the transport of three sodium ions out of the cell and two potassium ions into the cell occur in a single transport cycle. The Na/K-ATPase β-subunit (ATP1B1) TMD had been reported to form a strong homodimer and heterodimer with α subunit by Barwe et al. [25]. The ATP1B1 has a cell-cell adhesion function and its dimerisation is important for epithelial lumen formation [159]. Importantly for the primary function, the TMD contains a conserved motif YxxxYxxLxxxF that is involved in the hetero-oligomeric interaction between the β and α subunits [25]. The homodimerisation of the β subunit is less characterised. Barwe et al. [25] mutated only a single poorly conserved glycine G13 (G48 in the full sequence) that participated in a Gly zipper (GxxxGxxxG), and proposed that it was a key residue in the homodimer (β-β) interface.

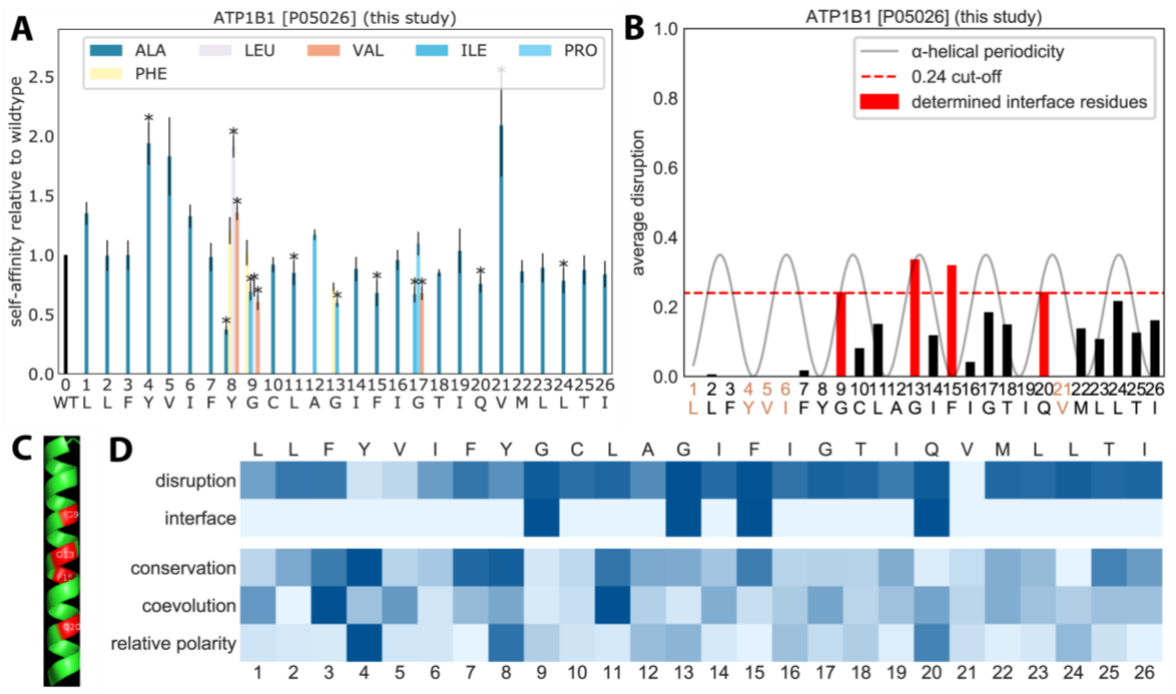

Figure 4-4: ATP1B1 TMD homodimer interface. [Full protein name: sodium/potassium-transporting ATPase subunit β-1, UniProt accession: P05026.] ToxR scanning mutagenesis data from this study, normalised to the wildtype TMD. Residue numbers are shown relative to the start of the TMD sequence used for scanning mutagenesis. Data presented are the mean ± SEM for at least 3 independent experiments. Statistical significance was calculated with a Student's *t*-test (*, p<0.05). B) Average disruption after mutation was calculated for

each residue position, based on the data in A. As described in Section 3.3.2, the disruption was calculated based on the average values of each individual mutation (e.g. for position 17, the average of G17I, G17P, and G17V mutations). A dotted red line indicates the 0.24 cut-off used to define interface residues. The bars for the identified interface residues are shown in red. For simplicity, only bars with positive disruption are shown. Positions with strong negative disruption (< -0.24) are indicated by orange text labels on the x-axis. The grey line shows the best fit of the data to α-helical periodicity. C) Location of interface residues on an ideal helix. Interfacial from data in B) are shown in red. The ideal helix model was created by the FMAP Server [132,133], based on the TMD sequence used for scanning mutagenesis. Peptide backbone is shown in cartoon format, with the N-terminus at the top. D) Heatmap showing average disruption after mutation, the defined interface, and residue properties. Conservation, relative polarity and coevolution were calculated from multiple sequence alignments against homologues using the thoipapy python package of Bo Zeng, as described in Section 3.2. Conservation was calculated from entropy, polarity using the GES scale [138], and coevolution using the DImax method (see associated publication, and Section 3.6). Darker shading indicates higher values. For the interface, dark shading indicates whether or not the position contains an interface residue according to the 0.24 cut-off shown in B).

The ToxR result from this study shows that the homotypic interface of ATP1B1 indeed includes the central G13 as proposed by Barwe et al. [25] (Figure 4-4). However the full interface was GxxxGxFxxxxQ (Figure 4-4 B), and was therefore a mixture of GxxxG, aromatic and strongly polar residues. ATP1B1 was unusual in the number of mutations that increased the dimer signal. Indeed, three mutations (Y4A, Y8L, and V21A) increased the dimerisation above 150% of the wildtype sequence (Figure 4-4 A). The defined interface was also unusual in that it had a very poor fit to α-helical periodicity (Figure 4-4 B). This may indicate that the helix contains a kink, or that one of the residues is indirectly necessary for helix-helix interactions.

4.2.2 Homotypic interaction of the human TIE1 TMD is mediated by hydrophobic residues

Receptor tyrosine kinases (RTK) are important proteins in signal transport across the membrane, and are implicated in many cellular functions and diseases such as cancer [160-162]. Initially, ligands bind to the soluble extracellular domain and trigger

receptor dimerisation [163], or the rearrangement of existing dimers [164]. Tie1 is a member of the endothelial-specific RTK family that is active in cell proliferation, migration and survival during angiogenesis [165]. Kontos et al. [166] showed that Tie1 is required for normal embryonic vascular development with genetic studies in mice. Targeted disruption of the Tie-1 gene results in a lethal phenotype with severe disruption to the normal integrity of the vasculature [167]. Currently, Tie1 has no defined ligand [168]. In a large TOXCAT study, Finger et al. [3] systematic compared the self-affinity of 58 TMDs from human RTKs. Tie1 gave the strongest signal of all RTKs in the study [3].

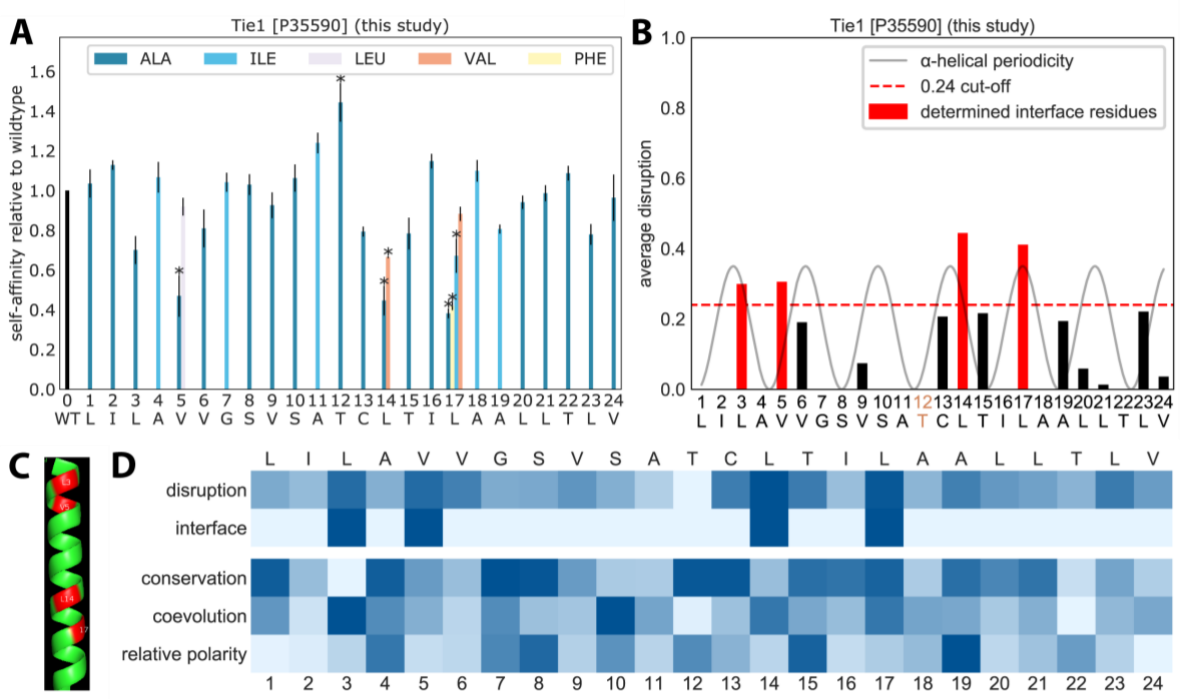

Figure 4-5: TIE1 TMD homodimer interface. Full protein name: tyrosine-protein kinase receptor Tie-1, UniProt accession: P35590. For the caption, see Figure 4-4.

Scanning mutagenesis revealed a long, hydrophobic interface of LxVxxxxxxxxLxxL was important for dimerisation of Tie1. Interestingly, mutations in a conserved AxxxS motif did not change the dimerisation propensity (Figure 4-5 D). The TMD itself was

very hydrophobic. 13/24 residues were Leu, Ile, or Val residues. However the interface did not fit a classical “leucine zipper” pattern (Figure 4-5 B).

4.2.3 Homotypic interaction of the Siglec7 TMD relies on GxxxG and (small)xxx(small) motifs

Sialic acid binding Ig-like lectins (siglecs) are a group of transmembrane receptors that are expressed primarily on the cells of the immune and hematopoietic systems [169]. With the ability to recognise sialic acids, they control cellular interactions and signalling events[170]. Siglec7 is a negative regulator present on NK cells which down-regulates innate immunity when activated [171]. The expression level of siglec7 strongly related to the liver injury, which makes it a potential biomarker for the prediction of mortality [172]. In addition, siglec7 display a high degree amino acid sequence identity with myeloid-specific CD33, a risk factor for Alzheimer’s disease (AD) [173]. Siglec7 was initially tested in the ToxR assay by Kirrbach et al. [21], who conducted a clustering study of human TMDs, and chose representatives for the analysis of self-affinity [21]. In the study, siglec7 showed the highest ToxR activity of all 33 TMDs tested. After testing the ToxR activity of four consecutive sequence frames for each TMD, siglec7 also showed the highest orientation-dependence [21]. Although some siglec proteins have been proposed to exist as disulfide-linked homodimers [174], little is known about the self-interaction of the TMD.

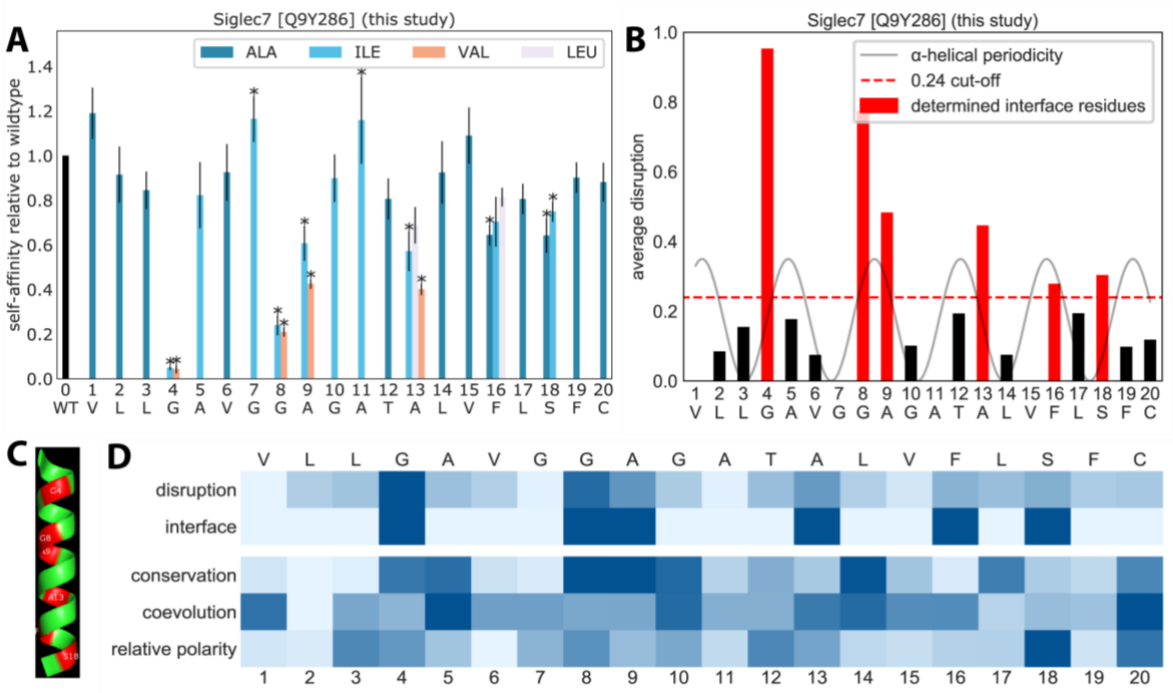

Figure 4-6: Siglec7 TMD homodimer interface. [Full protein name: sialic acid-binding Ig-like lectin 7, UniProt accession: Q9Y286.] For the caption, see Figure 4-4.

The TMD of siglec7 is rich in small residues, with 10/20 residues being G, A, S or C. Most of those small residues were highly conserved among homologues (Figure 4-6 D). The TMD contains one GxxxG motif and four (small)xxx(small) motifs. After mutagenesis of two residues, G10 and A13, Kirrbach et al. [21] suggested they might be important for the self-interaction.

Full scanning mutagenesis of the TMD revealed a clear interface of GxxxGAxxxAxxFxS. It therefore is dominated by small residues, and includes both GxxxG and a (small)xxx(small) motifs. The interface residues were obtained with high certainty, with mutations to a second residue always confirming the initial result (Figure 4-6 A). The unusual “back-to-back” arrangement of (small)xxx(small) motifs showed weak fit to α-helical periodicity (Figure 4-6 B).

4.2.4 Homotypic interaction of the ARMCX6 TMD depends on a GxxxG motif, aliphatic and ionisable residues

The function of human Armcx6 is still unknown. The self-interaction of the TMD was initially tested as part of the clustering study by Kirrbach et al. [21]. It showed the second highest ToxR activity of 33 human TMDs tested.

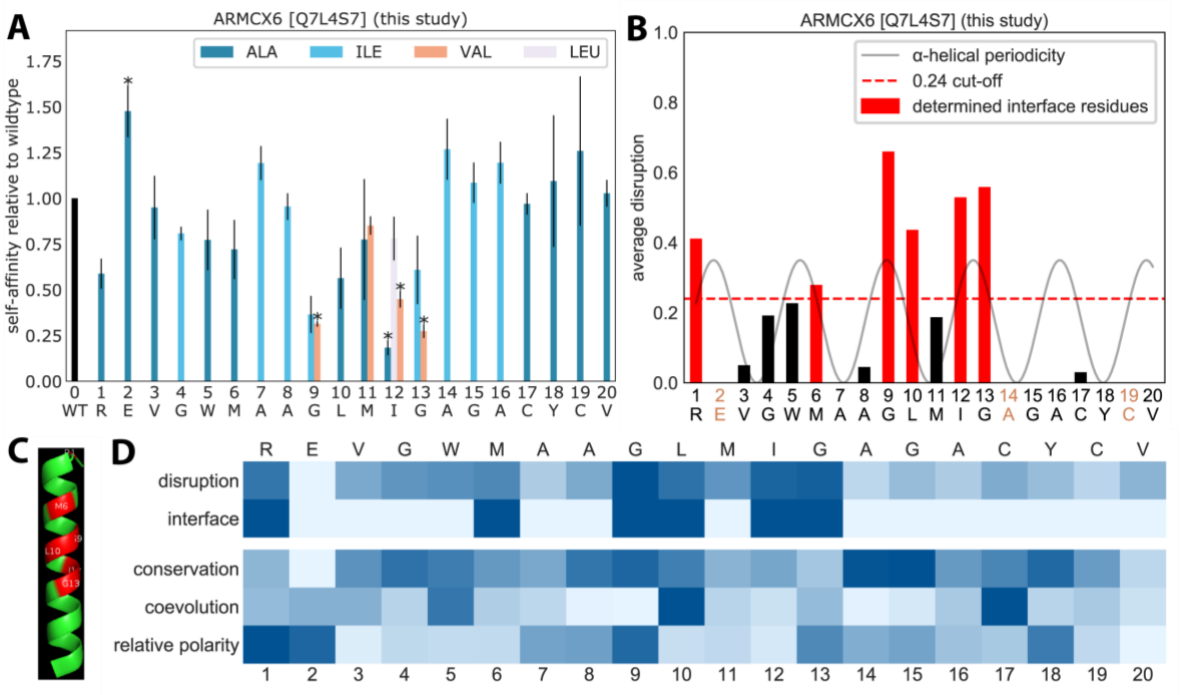

Figure 4-7: ARMCX6 TMD homodimer interface. [Full protein name: protein ARMCX6, UniProt accession: Q7L4S7.] For the caption, see Figure 4-4.

The TMD interface of Armcx6 was identified as RxxxxMxxGLxIG, and showed an excellent fit to α-helical periodicity (Figure 4-7 B). Small residues are common in the TMD (Figure 4-7 D). The TMD contains four (small)xxx(small) motifs and one GxxxG motif, all of which are highly conserved among homologues. Of these motifs, however, only the GxxxG motif is involved in dimerisation (Figure 4-7 B). This data agrees with the previous finding based on limited mutagenesis that G9 and G13 were mutation sensitive positions [21]. Interestingly, the aliphatic residues L10 and I12 neighbouring the Gly residues support the dimerisation (Figure 4-7 B). This mirrors the proposed

role of the Val residues in GpA [70] which also neighbour Gly residues. Mutation of an arginine residue in the juxtamembrane region drastically decreased the self-affinity (Figure 4-7 B). Recent research has shown that positively charged residues in the juxtamembrane region are vital for membrane insertion of some TMDs [22, 175]. This raises the possibility that the arginine residue increases membrane insertion and therefore indirectly the ToxR signal, rather than being located directly at the interface. However, the good alignment of the arginine residue with the GxxxG motif on one side of the helix suggests that it is indeed at the interface. The ionisable residue is therefore proposed to cooperate with the GxxxG motif as seen for the histidine of BNIP3 [6].

4.2.5 Homotypic interaction of PTPRU depends on small and aliphatic residue

This study includes the scanning mutagenesis of TMDs from three receptor-like protein tyrosine phosphatases (RPTPs), being PTPRO, PTPRU, and PTPRG. RPTPs regulate phosphotyrosine levels of the cell [176], and are important for the nervous system [177]. Like receptor tyrosine kinases (RTKs), RPTPs are associated with cancer development [178]. There are 22 known RPTPs in humans, divided into eight structural classes, based primarily on motifs in their extracellular domains (ECDs) [179]. The dimerisation of RPTPs has physiological significance, as the dimerisation state can be regulated by extracellular ligands and changes in oxidation state [180]. Dimerisation through TMD interactions has been suggested as a general mechanism for signal transfer by RPTPs [9].

PTPRU is essential for the maintenance of epithelial integrity and in the regulation of the Wnt/β-catenin-signalling pathway [106]. The role of TMD self-interaction in signal transfer by PTPRU is not known.

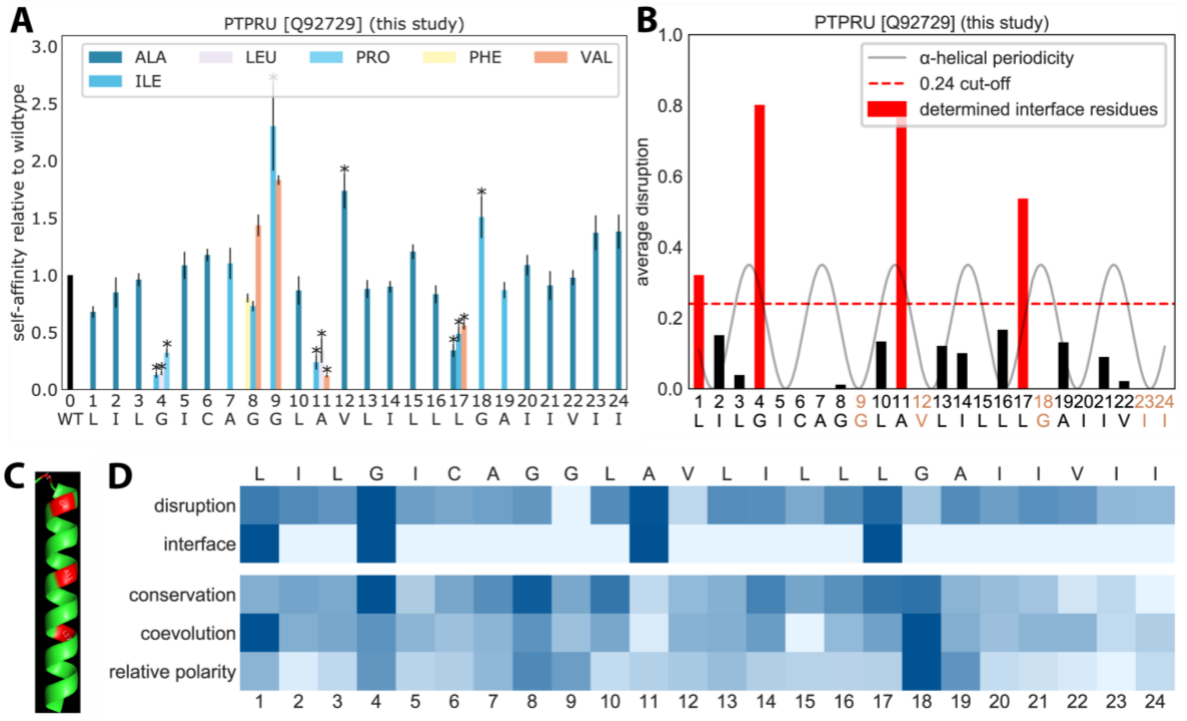

Figure 4-8: PTPRU TMD homodimer interface. [Full protein name: receptor-type tyrosine-protein phosphatase U, UniProt accession: Q92729.] For the caption, see Figure 4-4.

The results show a broad homotypic interface of LxxGxxxxxAxxxxxL, which showed a moderate fit to α-helical periodicity (Figure 4-8 B). Despite the presence of an N-terminal GxxxG motif and two (small)xxx(small) motifs, the experiments show that only one of the Gly residues (G4) supports the dimerisation (Figure 4-8 B). This is despite the fact that both G4 and G8 were highly conserved among homologues (Figure 4-8 D). The sequence was rich in aliphatic residues (V/I/L) that formed two leucine zipper motifs, however neither of these was important for self-interaction.

Several mutants enhanced dimerisation. Most of them involved mutations from a polar residue to a hydrophobic residue (G to I, G to V) (Figure 4-8 A). These mutations increased the average hydrophobicity of the TMD that may have improved membrane

insertion. In fact, the G9I mutation yielded a 2-fold increase in self-association in comparison to the wildtype. To exclude the chance that this was due to unexpected mutations in the plasmid backbone during cloning, a second mutation was applied to this plasmid that reverted the sequence back to the wildtype. The reverted wildtype plasmid (G9I_I9G) had the same ToxR activity as the original wildtype (supplementary Figure 8-5 A). The result confirmed that the massive increase in self-affinity was a sequence-specific effect caused by the G9I mutation, and was not due to any rare mutations in the backbone plasmid that could occur during cloning.

4.2.6 Homotypic interaction of the PTPRG TMD depends on highly conserved residues

PTPRG is a potential tumour suppressor in nasopharyngeal carcinoma (NPC) [178]. Some cancers have lower expression levels of PTPRG [181]. The exact role of homotypic TM interactions in signal transfer is not known.

The objectively defined interface TxxxLxxxxxL excluded a preceding Ser by only a small margin, which would have resulted in a central SxxTxxxL with a good fit to α-helical periodicity (Figure 4-9 B). The interface therefore combined small/polar and large aliphatic residues. Interestingly, the TMD contained many hydrophobic residues (L/V/I) that were highly conserved (Figure 4-9 D), forming three leucine zipper motifs in total. However, these highly conserved Leu residues were not found at the interface.

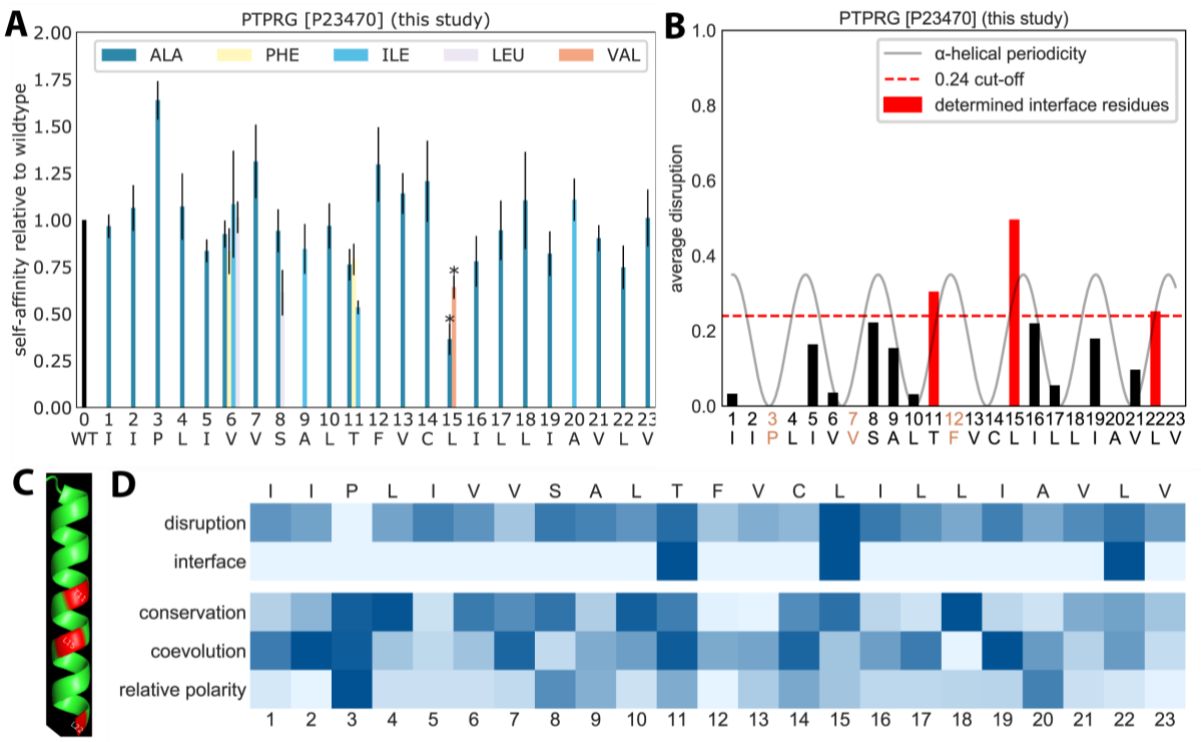

Figure 4-9: PTPRG TMD homodimer interface. [Full protein name: receptor-type tyrosine-protein phosphatase gamma, UniProt accession: P23470.] For the caption, see Figure 4-4.

4.2.7 Homotypic interaction of the PTPRO TMD depends on multiple hydrophobic residues

PTPRO is mainly expressed in the developing nervous system and involved in axon guidance [182]. PTPRO can exist in a disulfide-linked dimerised state in living cells, as determined by a Western blot analysis [183]. One of its substrates is the neurotrophin-3 (NT-3) receptor tropomyosin-related kinase C (TrkC) [183]. Dimerisation of PTPRO is proposed to decrease phosphatase activity. The TM dimer interface has never been shown. The wildtype showed an extreme high homotypic interaction as the ToxR signal is 242% of GpA. It was the TMD with the second highest ToxR activity of all TMDs tested in this study.

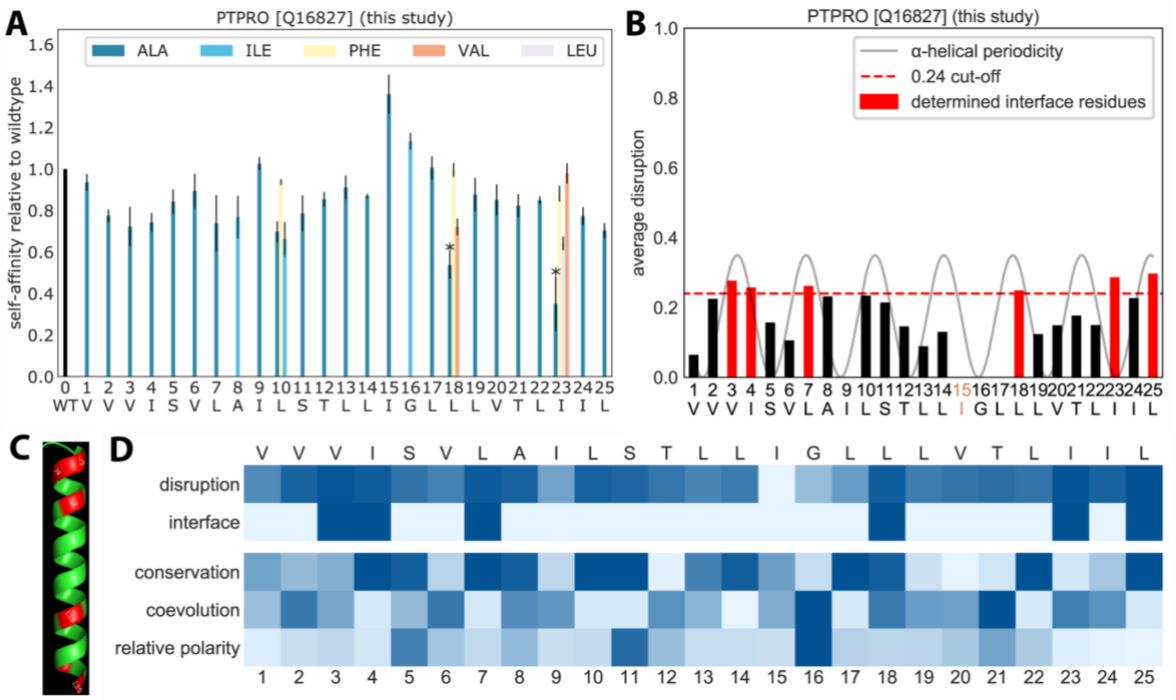

Figure 4-10: PTPRO TMD homodimer interface. [Full protein name: receptor-type tyrosine-protein phosphatase O, UniProt accession: Q16827.] For the caption, see Figure 4-4.

The sequence of the PTPRO TMD contained three conserved leucine zippers (Figure 4-10 D). The results show that the interface VIxxLxxxxxxxxxLxxxxIxL is dependent on hydrophobic residues (Figure 4-10 B). However, the interface does not form a typical leucine zipper. The defined interface also did not fit the α helical periodicity (Figure 4-10 B). Since the interface was extremely broad this might indicate that presence of a kink the likelihood of kinks between the residues. The overall disruption of each residue was modest, which indicated that multiple residues might contribute weakly to the interaction (Figure 4-10 A). Dimerisation of her PTPRO TMD could not be fully disrupted by single amino acid substitutions.

4.2.8 Homotypic interaction of the IRE1 TMD relies on a highly conserved tryptophan residue

IRE1 is located in the endoplasmic reticulum and acts as an essential proximal sensor of the unfolded protein response pathway in mammalian cells [184]. Uniquely, it contains an N-terminal luminal domain that detects unfolded proteins, a cytoplasmic kinase domain, and also a cytoplasmic mRNA endonuclease domain [185]. Unfolded proteins in the lumen are thought to increase dimerisation or oligomerisation of IRE1 [186]. This leads to autophosphorylation of the kinase domain in the cytoplasm, and subsequent activation of the endonuclease domain, which then splices the mRNA of XBP1 [187]. This signalling cascade is central to the unfolded protein response, an important process in age-related diseases. XBP1 over-expression is known to reduce the expression levels of APP [188], the precursor to the toxic aβ42 fragment implicated in Alzheimer disease. Deletion of the IRE1 endonuclease domain dramatically reduces Alzheimer-related symptoms in mice [189].

The TMD was chosen in this study because of the high sidedness of conservation of the TMD region (Table 8-2). Despite the importance of the self-interaction, the TM homodimer interface is unknown, and this is the first ETRA study to investigate it fully.

The results show that the interface of IRE1, WVxxxxT, is strongly dependent on a highly-conserved Trp residue located toward the centre of the TMD (Figure 4-11 B). Trp residues are common at the lipid interface (juxtamembrane region), however their potential role in TMD interactions has been shown [34]. Although the ToxR activity of the wildtype was not particularly high (77% GpA), the interface was nevertheless highly sensitive to mutations. This was already suggested by the high orientation dependence of the TMD (Figure 4-2). It supports the possibility that the other TMDs with high sidedness of conservation but modest ToxR signal (Figure 4-2) may still be an excellent candidate for scanning mutagenesis.

Like the G9 mutation of PTPRU, a mutation was observed in the IRE1 TMD that greatly enhanced the dimer signal. L9A approximately doubled the ToxR signal in comparison to the wildtype. As before, mutating the L9A plasmid back to the wildtype sequence completely restored the expected wildtype signal (supplementary Figure 8-5 B). The result confirmed that the massive increase in self-affinity was a sequence-specific effect caused by the L9A mutation, and was not due to any rare mutations in the backbone plasmid that could occur during cloning.

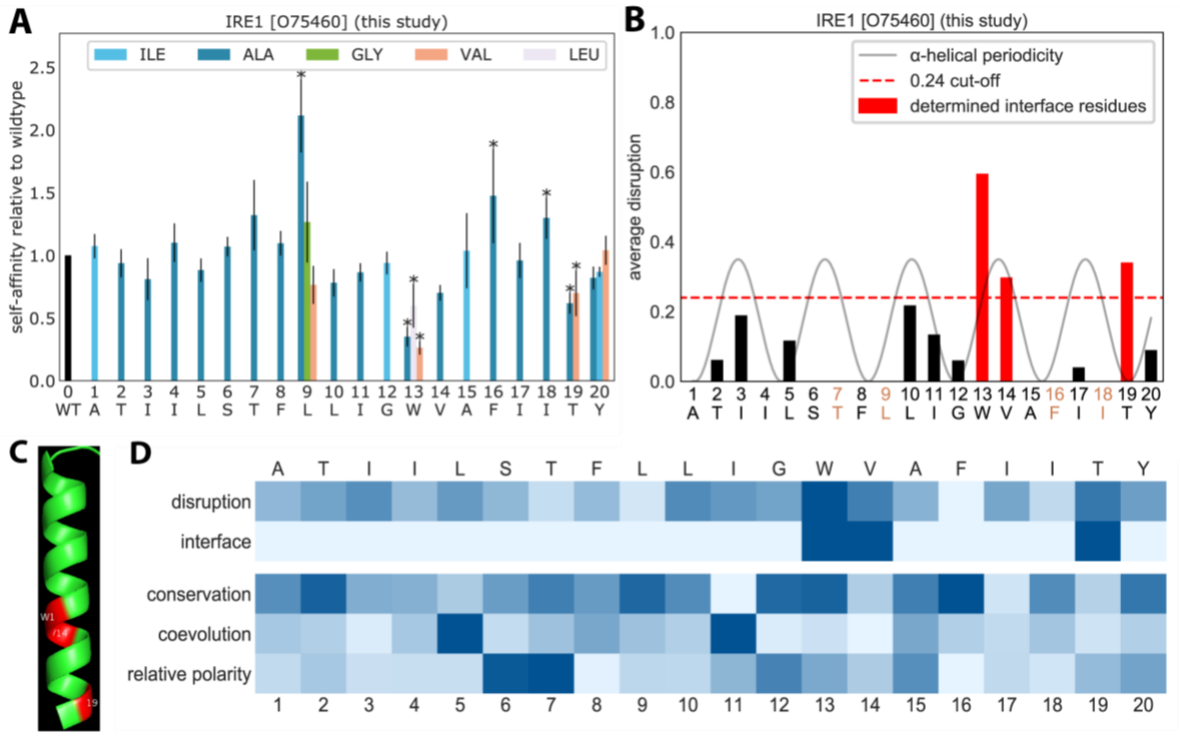

Figure 4-11: IRE1 TMD homodimer interface. [Full protein name: serine/threonine-protein kinase/endoribonuclease IRE1, UniProt accession: O75460.] For the caption, see Figure 4-4.

4.2.9 Homotypic interface of DDR1 and DDR2 TMDs is conserved between homologues, and depends on a leucine zipper

Discoidin domain receptors DDR1 and DDR2 belong to the receptor tyrosine kinase (RTK) family, which have extremely well-characterised, biologically relevant TM

homodimers [190]. They are characterised by the extracellular discoidin (DS) homology domain and are widely expressed in normal and malignant tissues [191]. They are activated by different types of collagen [190] and controlled developmental processes and also associated with several human diseases, including tissue fibrosis disease and several types of cancer [192, 193]. DDRs form constitutive dimers in the cell membrane [194]. The DDR1 protein is thought to be important for cell attachment, migration, survival, and proliferation [195]. DDR2 plays an important role in osteoblast and chondrocyte differentiation where it regulates cell proliferation and controls remodelling of the extracellular matrix [196]. Previous TOXCAT experiments with DDR1 and/or DDR2 include a proteome-wide test of all human RTKs [3], and also some limited mutagenesis to investigate dimerisation and collagen-induced signalling [191]. These showed that highly disruptive GP double mutations (LL->GP) in the Leu-rich region of the TMD disrupted both collagen-induced transmembrane signalling and TOXCAT self-association [191]. In contrast, a single A8V mutation to the GxxxA motif had no effect. However until now, the full TM dimer interface was unknown. DDR1 and DDR2 were the only homologues in this study. They have been reported to dimerise independently of an N-terminal AxxxG motif [3, 106].

Scanning mutagenesis revealed a long interface for DDR1 of IxxxxLxxIIxxLxxIIxxML (Figure 4-12 B). For all these residues to exist at the dimer interface would imply a parallel helix dimer with low crossing angles (Figure 4-12 C). The defined interface had a very strong α-helical periodicity (Figure 4-12 B). The LxxIIxxL pattern corresponds to [abcdefga] of a classical heptad motif, and therefore classifies as a “leucine zipper” interface [119].

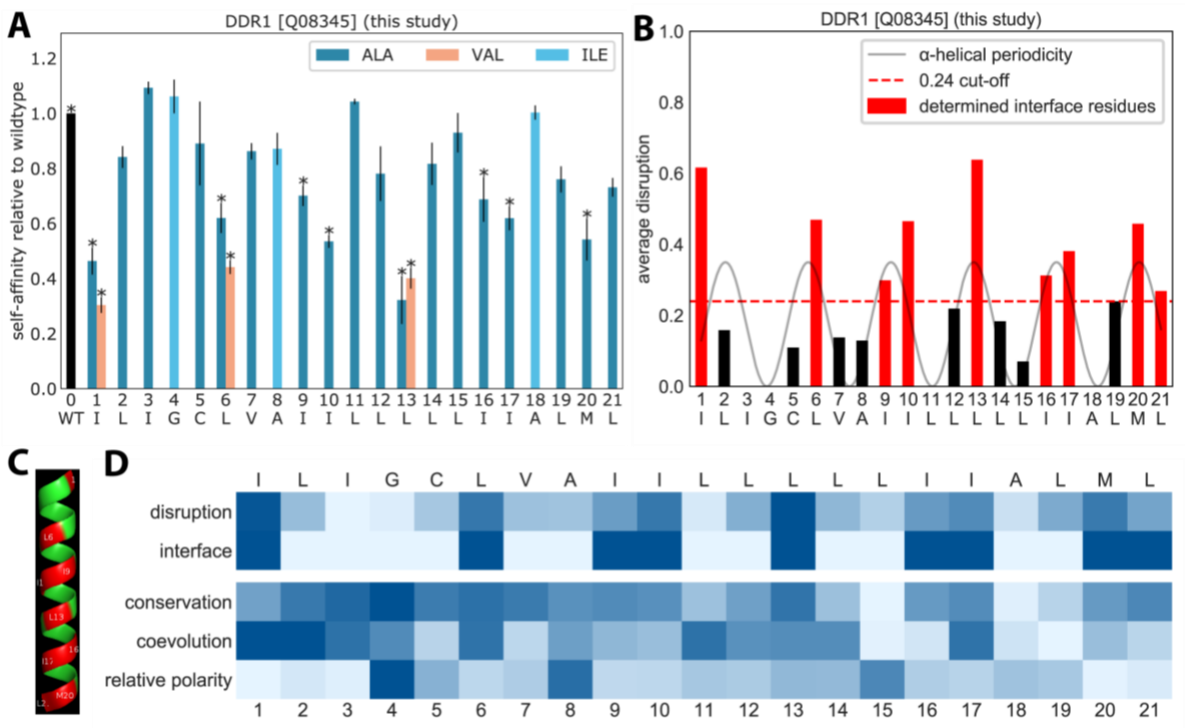

Figure 4-12: DDR1 TMD homodimer interface. [Full protein name: epithelial discoidin domain-containing receptor 1, UniProt accession: Q08345.] For the caption, see Figure 4-4.

The DDR2 interface was judged to be LIxxLxxxIxxLxxIIxxIL (Figure 4-13 B) which suggested a long, Ile/Leu dominated interface as described for DDR1 (Figure 4-12). The mutation-sensitive residues of DDR2 showed a remarkable helical periodicity (Figure 4-13 C). A comparison of the DDR1 and DDR2 interfaces revealed a high similarity (Figure 4-14). This confirms one previous study, which also found that homologous TMDs showed similar ToxR activity [21, 83]. This is the first time that homologous TMDs have been shown in ETRA experiments to have similar interface residues.

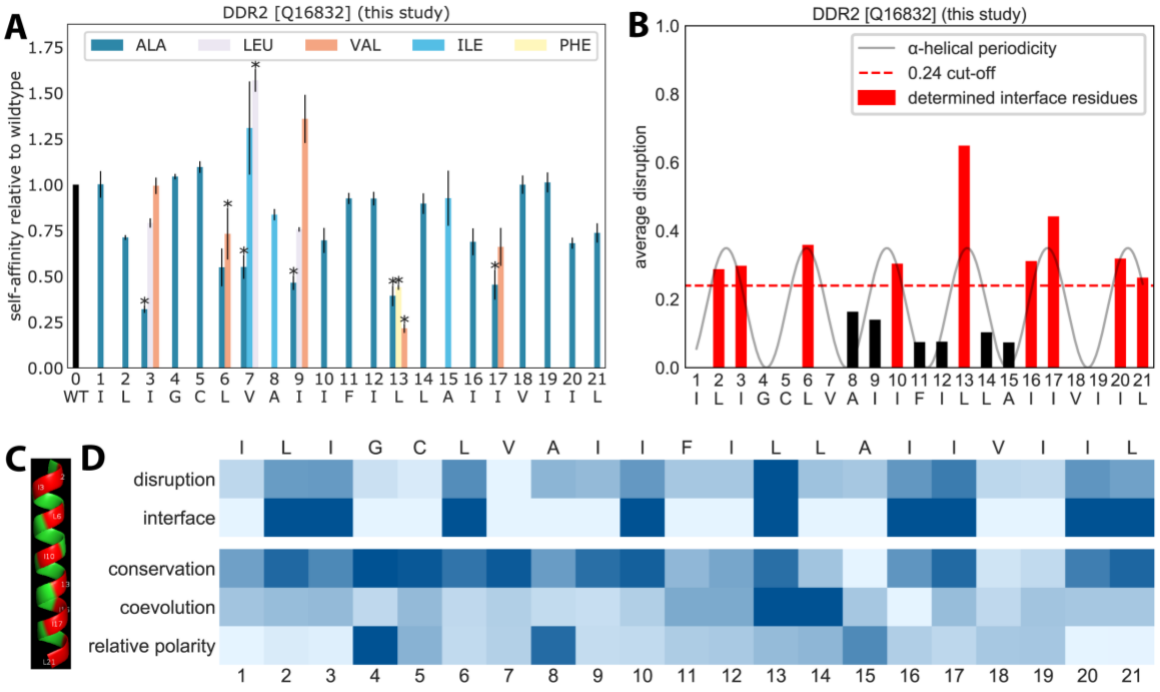

Figure 4-13: DDR2 TMD homodimer interface. [Full protein name: discoidin domain-containing receptor 2, UniProt accession: Q16832.] For the caption, see Figure 4-4.

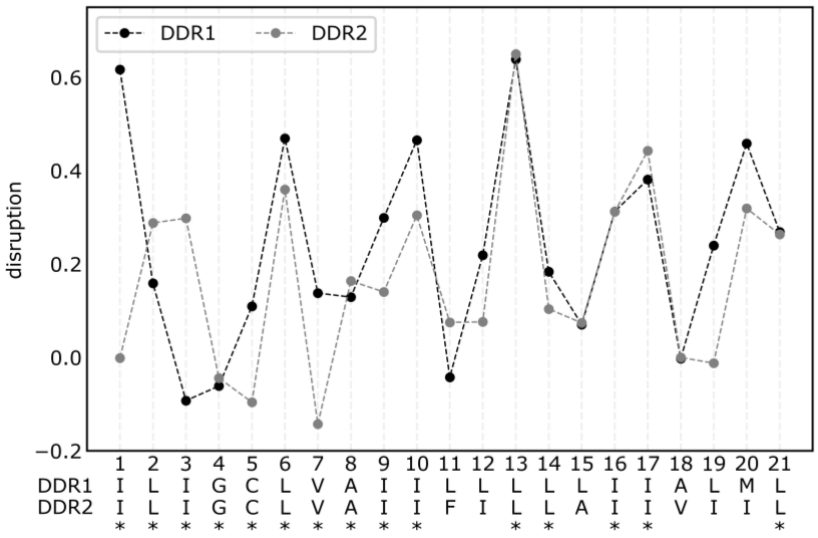

Figure 4-14: The defined interface for homologues DDR1 and DDR2 were similar. Disruption of DDR1 (Figure 4-12 B) and DDR2 (Figure 4-13 B) were plotted along the TMD sequence. TMD sequence are shown on the x-axis. Stars indicate identical residues of DDR1 and DDR2.

CHAPTER 5. RESULTS II: FEATURES OF INTERFACIAL RESIDUES

The nine unique novel homotypic TM interfaces described in CHAPTER 4 represent over 40% of all known interfaces investigated by ETRA techniques. The aim of this section is to combine these new interfaces with all those previously described in literature. The combined ETRA data were then evaluated to improve our quantitative understanding of interface residue properties. This section includes the manual collection of data from all other ETRA studies. In addition, via a collaboration with Bo Zeng of the Dmitrij Frishman group of the Technical University of Munich, it was possible to analyse the sequence properties of many other TM homodimers investigated by NMR or X-ray crystallography techniques. The full “homotypic TM database” contained 54 TMDs with known interfaces characterised via ETRA, NMR or crystallography techniques. An overview of created datasets and calculated residue properties is available in Figure 5-1. The machine learning predictor of interface residues (THOIPA) constructed by Bo Zeng utilised exactly the same homotypic TM database and residue properties. This required extensive collaboration and harmonisation (Section 3.4) and of the used approaches. Due to this harmonisation, it was possible to confirm that all of the major residue properties described here were important for machine-learning prediction of homotypic TM interface residues. The details are available in the thesis of Bo Zeng and the associated co-first-author manuscript (Section CHAPTER 10).

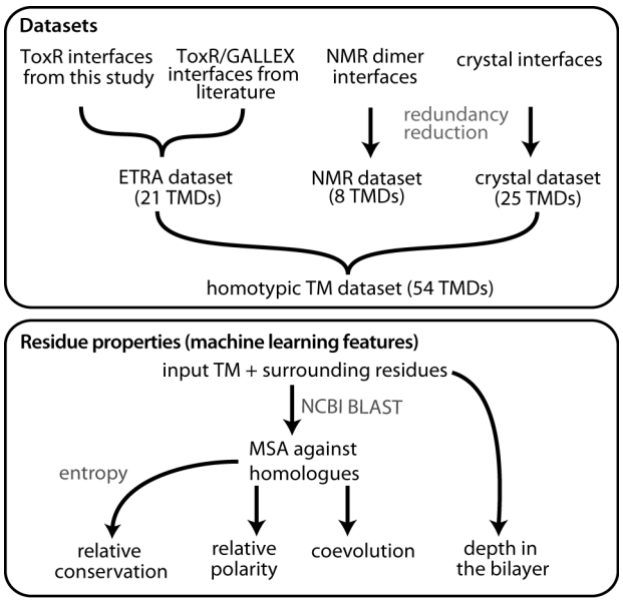

Figure 5-1: Overview of datasets and residue properties analysed in this study.

The conclusions described in this chapter are mostly derived from the full homotypic TM database of 54 TMDs. Analyses of the individual datasets were also conducted to confirm that trends were universal or identify residue properties that more strongly associated with interface residues in a particular dataset. After redundancy reduction, the NMR dataset contained only eight TMDs. In any analyses of the individual datasets, emphasis should therefore be placed on the larger ETRA and crystal datasets, although the NMR dataset was always included for completeness.

5.1 Creation of the first large dataset of homotypic TM interface

5.1.1 Creation of the *E. coli* TMD Reporter Assay (ETRA) dataset

The *E. coli* Transmembrane Reporter Assay (ETRA) dataset comprises TMDs whose homodimerisation interface has been determined in the *E. coli* inner membrane. The ETRA dataset includes the scanning mutagenesis data of nine unique novel

interfaces derived from the experimental section of this study. Of the homologues DDR1 and DDR2 examined in Section CHAPTER 4, only DDR1 was included.

The full ETRA dataset included 21 TMDs from non-redundant human bitopic proteins. Scanning mutagenesis data for twelve TMDs were taken from the literature. These had all been investigated using ToxR-based methods such as ToxR/TOXCAT/dsTβL, except for the NS4A TMD, which had been investigated using the GALLEX assay (Table 5-1). The scanning mutagenesis data was standardised between different studies (Section 3.3.2). For each mutation, the disruption to self-affinity was calculated by comparison to the signal of the wild-type sequence (Section 3.3.2). Important residues for the interaction were characterised by low signal after mutation, and therefore a high disruption. Interfacial residues were defined objectively as any position with a mean disruption value above 0.24.

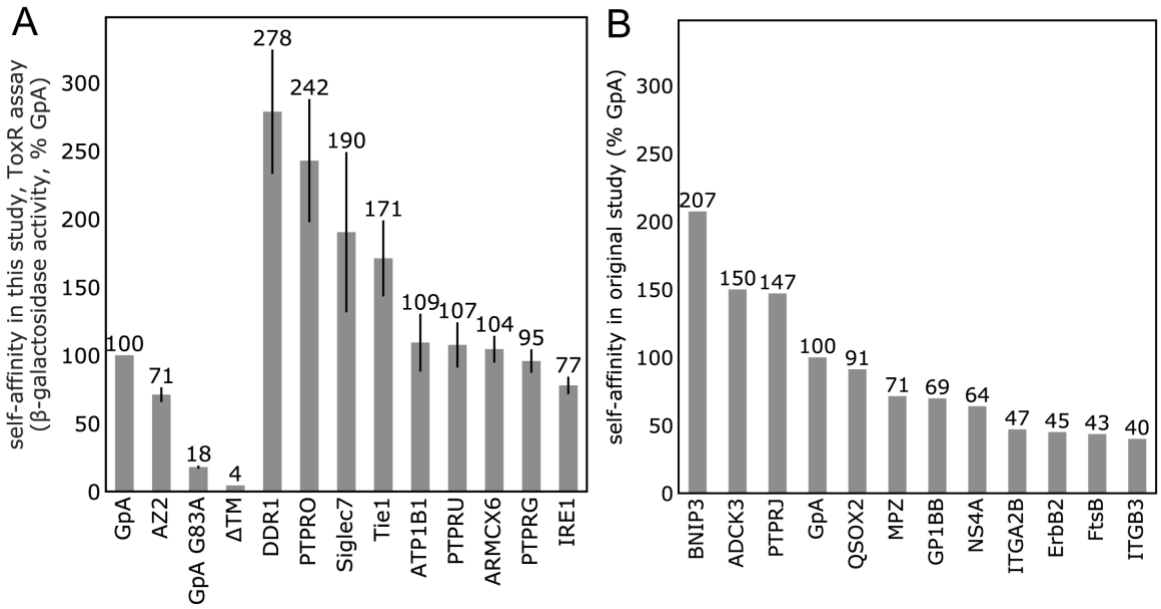

Figure 5-2: The ETRA dataset is primarily comprised of TMDs with strong self-affinity. (A) The self-affinity in a ToxR-based system is shown for all TMDs include in the ETRA dataset. In all cases, the affinity shown is relative to the glycophorin A (GpA). Controls AZ2, GpA G83A and ΔTM represent moderate, weak and no homodimers respectively. For data produced in this study, the mean ± SEM of biological replicates (separate transformation, n > 3) is shown. (*B*) ETRA TMDs with scanning mutagenesis data from the literature. Note that because the dsTβL mutagenesis assay only measures relative affinity, different sources were used for the ErbB2 scanning mutagenesis data [22], and reported TOXCAT signal

shown here [44]. NS4A was the only member of the ETRA dataset that was derived from GALLEX rather than a ToxR-based assay. An equivalent %GpA for NS4A was calculated as described in Section 3.3.1.

Table 5-1: TMDs for which scanning mutagenesis data were extracted from the literature

protein (acc[a])	TMD sequence[b]	reference
BNIP3 (Q12983)	LLSHLLAIGLGIYIG	[6]
ADCK3 (Q8NI60)	LANFGGLAVGLGFGALA	[11]
PTPRJ (Q12913)	ICGAVFGCIFGALVIVTVGG	[9]
GpA (P02724)	LIIFGVMAGVIGTIL	[22]
QSOX2 (Q6ZRP7)	CVVLYVASSLFLMVMY	[28]
MPZ (P25189)	YGVVLGAVIGGVLGVVLLLLLLFYVV	[31]
GP1BB (P13224)	GALAAQLALLGLGLLHALLL	[32]
NS4A (Q99IB8)	TWVLAGGVLAAVAAYCLAT	[33, 34]
ITGA2B (P08514)	WVLVGVLGGLLLLTILVLAMW	[35]
ErbB2 (P04626)	LTSIISAVVGILLVVVLGVVFGIL	[22]
FtsB (P0A6S5)	TLLLLAILVWLQYSLWF	[39]
ITGB3 (P05106)	VLLSVMGAILLIGLAALLI	[40]

[a] Accession number (acc) is taken from the UniProt database.
[b] Interfacial residues are underlined. These were determined objectively using the methods described in Section 3.3.3.

The length of TMDs in the ETRA dataset ranged between 20 and 26 amino acids. The average length was 20.5 residues. The ETRA dataset contained 862 mutations at 432 positions. Among them, the BNIP3 had the largest number of mutations (101) [197] while QSOX2 has the least, with mutations at only 16 positions [158]. The "disruption index" was used to calculate the overall impact of one or more mutations at a position on dimer affinity [39] (Equation 3-4). The average number of interfacial residues was 5.2 AA per TMD. In general, 21% of residues were interface positions.

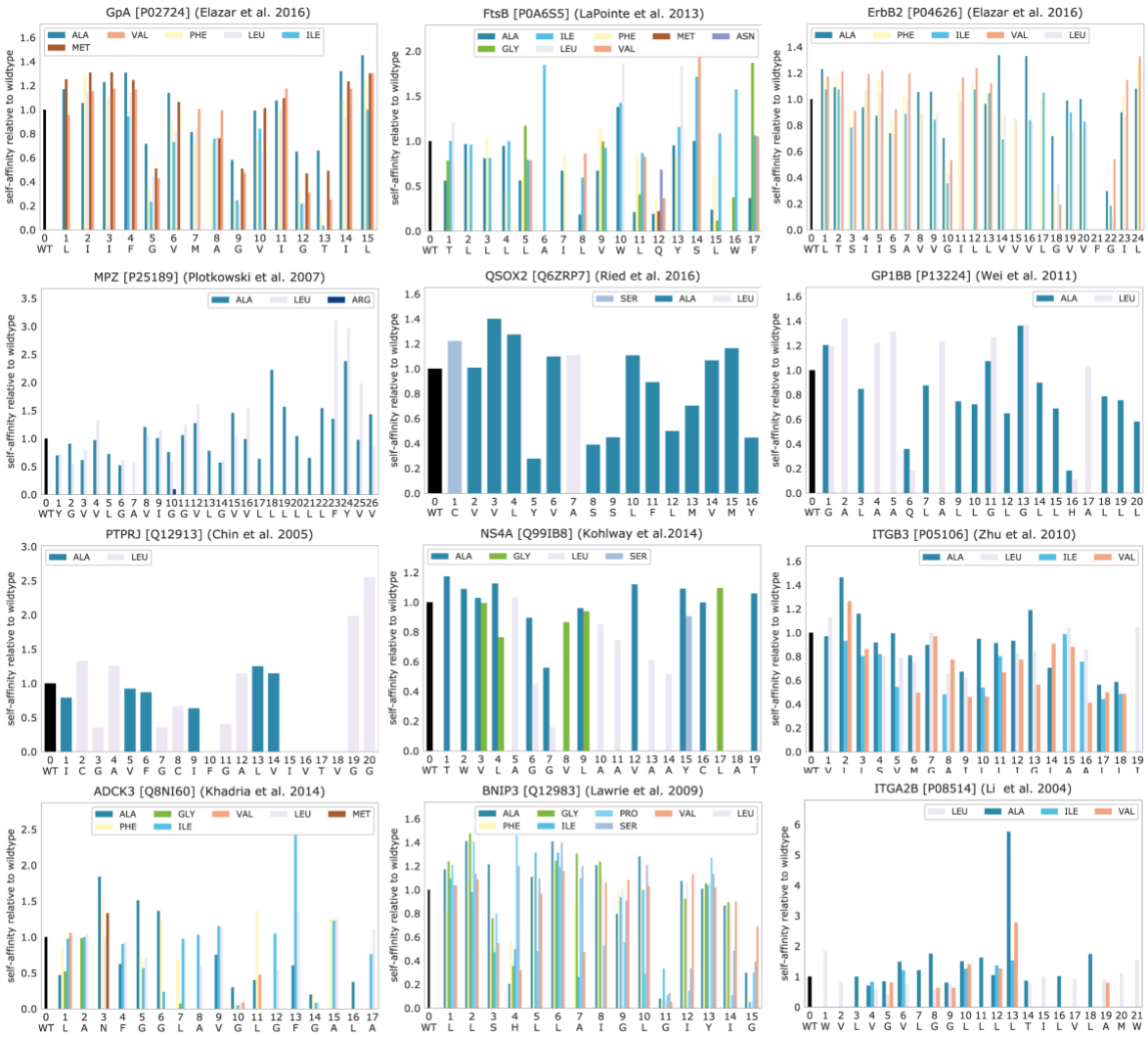

Figure 5-3 Scanning mutagenesis data for TMDs whose data were derived from literature. In all cases, the data is normalised to the wildtype sequence of that TMD. Normalisation applied to GpA, ErbB2, ITGB3, and NS4A is described in Section 3.3.13.3.1

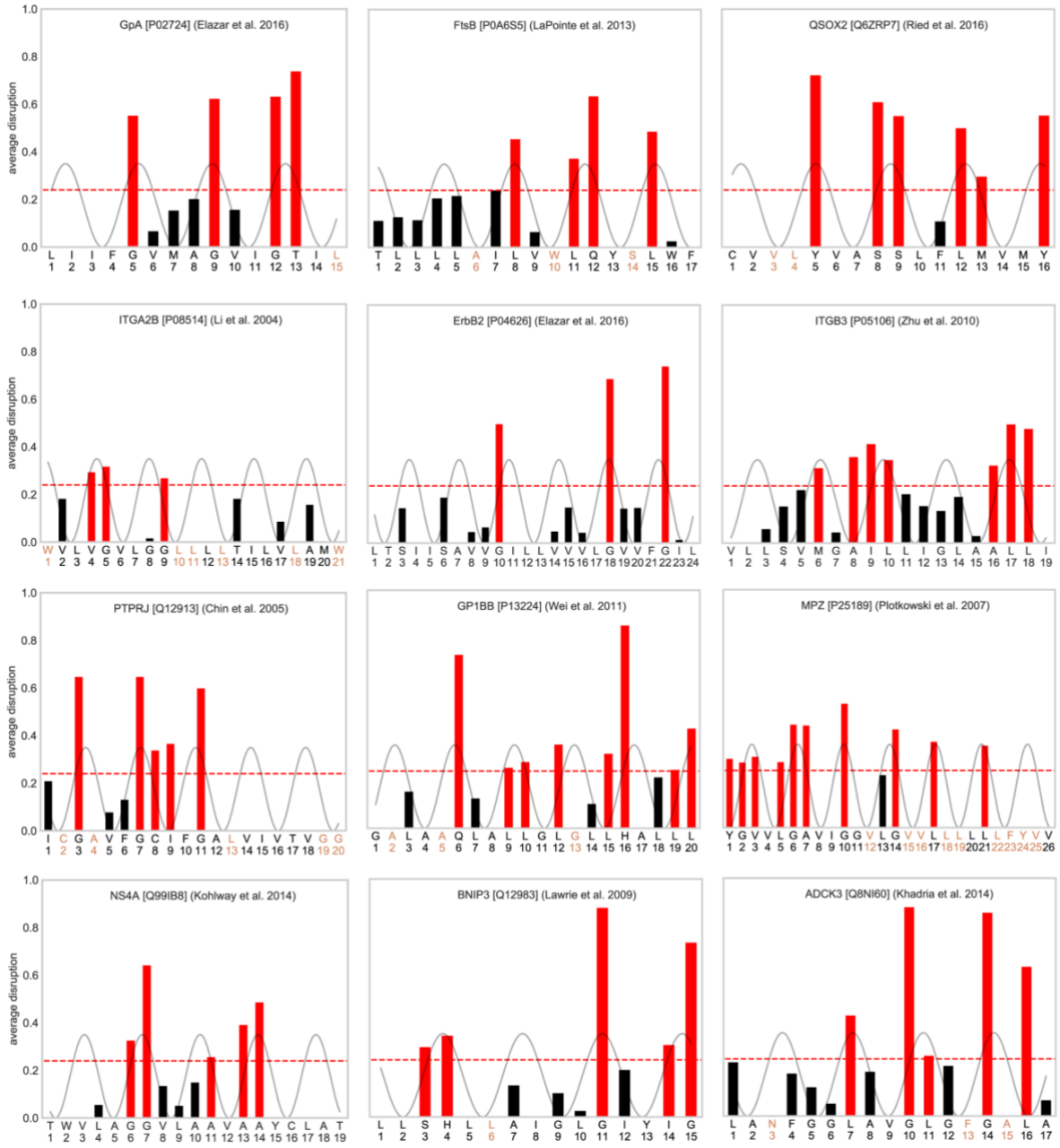

Figure 5-4: Homotypic interface residues extracted from literature. The mutation data presented in Figure 5-3 was converted to a disruption index for each residue. A low ToxR signal for mutations (indicative of an interface) gives a high disruption. Interfacial residues are shown in red, as determined objectively using the methods described in Section 3.3.3. For each TMD, residues were classified as "interface" that gave a disruption above 0.24 (red dotted line). Thus each TMD has a minimum of three interfacial residues, as described in Section 3.3.2. Only positive disruption values are shown. Positions where the disruption index was < -0.24 (i.e., where mutations greatly increased ToxR signal) are shown with orange text labels on the x-axis. The disruption index was fitted to a sine curve with α-helical periodicity (grey).

5.1.2 Addition of interfaces investigated by NMR and X-ray crystallography studies

The ETRA dataset described above was considered still too small to draw strong conclusions about homotypic TM interface properties. Via a collaboration with Bo Zeng, further self-interacting TMDs with characterised interfaces were then identified from studies that used other techniques, specifically NMR and X-ray crystallography.

NMR structures of over 15 TM homodimers have been reviewed extensively [36], and used to validate existing molecular modelling approaches [36, 86, 87]. Despite this, they had never been reduced to a single non-redundant dataset for quantitative analyses. As described in Section 3.4, redundancy reduction of the available NMR structures resulted in only eight TMDs that could be added to the homotypic TM dataset.

The crystal structures of membrane proteins have been extremely important for quantitative analyses of TM helix-helix interactions [17, 109, 124, 131]. The contacts between membrane proteins in crystal structures have been characterised and used to create machine-learning prediction algorithms. However, these methods are not truly de-novo, as they assume that a high-resolution accurate protein structure is available. Until now, the self-interacting helix pairs observable in the crystal structure data have never been extracted and studied in isolation.

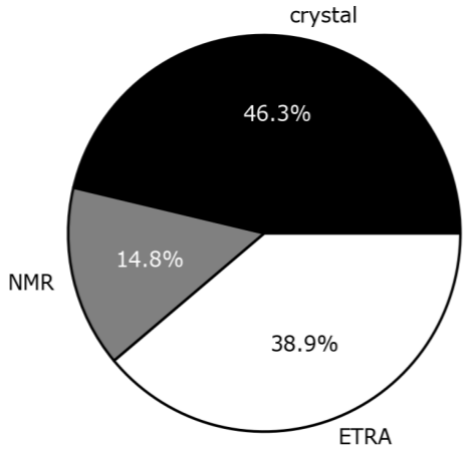

Figure 5-5: Most TMDs in the homotypic TM dataset were derived from experimental data using ETRA or X-ray crystallography structure techniques.

A total of 25 non-redundant, self-interacting TM helices were identified in crystal structures as described in Section 3.4. These primarily belonged to polytopic membrane proteins, and therefore the structure of the full proteins was fundamentally different from the bitopic TMDs examined in the ETRA and NMR dataset. However, in the context of TM helix-helix interactions, the crystal dataset was distinguished from the other datasets only by the TMD hydrophobicity (more polar residues), and the presence of concurrent heterotypic TM helix-helix interactions between non-identical TMDs of the polytopic proteins. Despite the fact that the self-interacting helixes from crystal structures had never been analysed, they were actually much more numerous than the TMDs in the well-studied NMR dataset (Figure 5-5).

5.2 Interfacial residues tend to be conserved, coevolved, polar and centrally located

Previous studies have shown that residues with many sequence contacts (i.e. high degree of burial in the protein) are known to have high sequence conservation amongst homologues [118]. This is also true for membrane proteins [124, 131]. In

comparison to residue contacts within folded proteins, the residue contacts involved in protein-protein interactions are often transient and/or difficult to confirm experimentally. In fact, some studies of soluble proteins have found no significant difference between the conservation of protein-protein interfaces and other solvent-exposed residues [133].

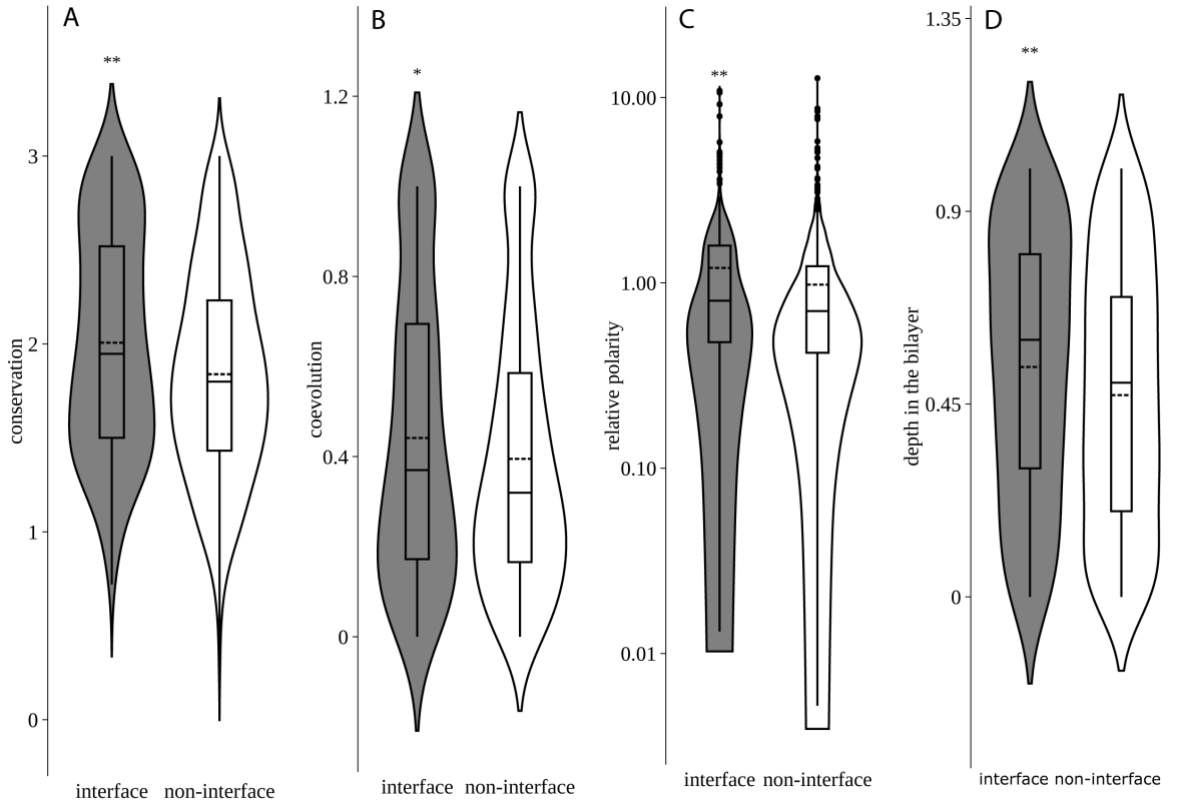

Figure 5-6: Interfacial residues are more conserved, coevolved, polar and centrally located in comparison to non-interfacial residues. An analysis of residue features of the non-redundant homotypic TMD dataset is shown. Interfacial residues were defined as described in Section 3.3.3. For the crystal dataset, heterotypic contacts were excluded from the analysis. Individual analyses of ETRA, NMR and crystal datasets are available in Figure 5-8. Statistical significance was measured using a bootstrapped t-test comparing interface and non-interfacial residues (*, $p<0.05$. **, $p<0.01$). In the violin plot, whiskers show min and max, the box represents the interquartile range, a solid line indicates the median, and a dotted line indicates the mean. (A) Conservation. (B) Coevolution (DImax). (C) Relative polarity, representing polarity divided by the mean polarity of the surrounding six residues (y-axis is presented on a $\log_{10}$ scale to show a wide range of values). (D) Residue depth in the bilayer. This was based on position in the TMD sequence, from the most central (value=1) to the most peripheral TMD residue (value=0). The interfaces of NMR and crystal TMDs were provided by Bo Zeng, and residue properties were calculated using the THOIPApy software package.

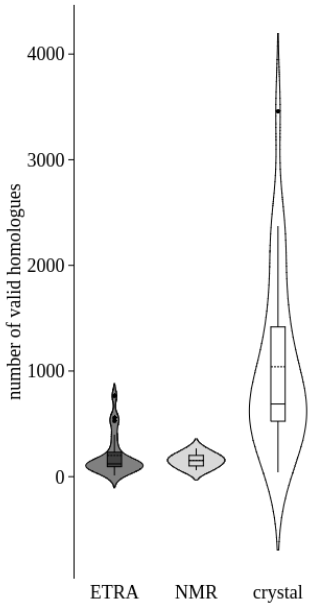

Figure 5-7: Number of valid homologues for TMDs of each dataset. The mean number of homologues was 201, 154, and 1040 for the ETRA, NMR and crystal datasets respectively. Filtering and redundancy reduction of homologues is described in Section 3.4. Violin plots were constructed from the data as described in Figure 5-6.

Here, it is shown for the first time that the interfacial residues mediating homotypic TM interactions are significantly more conserved than non-interfacial residues. This was true not only for the homotypic TMD dataset (Figure 5-6), but also for the crystal, NMR and ETRA datasets alone (Figure 5-8). A statistical analysis using 95% confidence intervals (CI) confirmed the importance of conservation for the large ETRA and crystal subsets **(**Table 8**-**5). The CI analysis also revealed that overall conservation scores (both interface and non-interface) were lower for the crystal dataset. This can be attributed to the larger number of homologues in the crystal dataset (Figure 5-7). Many conservation scores are weakly dependent on the number of sequences in the alignment [198].

A measure of sequence coevolution was also significantly higher for interfacial residues (Figure 5-6). Coevolution values are particularly powerful for the identification of interacting helical pairs in multi-pass membrane proteins [199, 200],

but have only been applied to homotypic TM interactions in one other study [36]. Coevolution scores such as direct information (DI) are pairwise values between non-identical residues that are calculated from multiple sequence alignments.

The coevolution score used in this chapter is the DImax, which represents the maximum DI score of this residue with any other residue in the TMD (see Section 3.6.5 for methods). Unlike the strong result obtained for conservation, the extent of the difference in coevolution between interface and non-interface residues was moderate (**Table** 8**-**5). Nevertheless, this suggests that interfacial residues were often distinguished by having a high coevolution score with some other residue or residues within the TMD. This study therefore provides some support to the report by Wang and Barth [36], who suggest that coevolution scores can help predict homotypic TM interfaces.

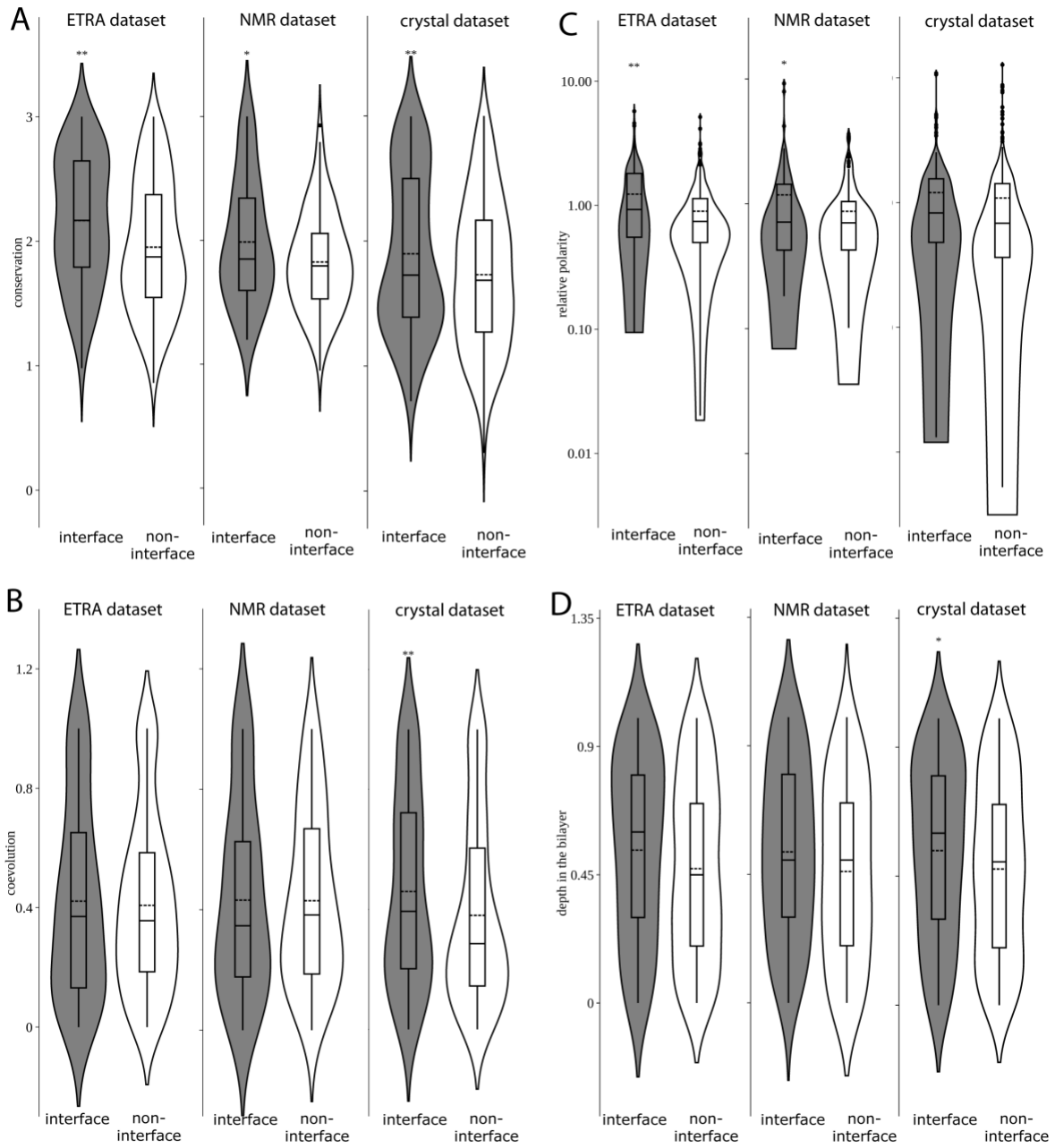

Figure 5-8: Interfacial residue properties for each dataset. ETRA, NMR and crystal datasets were examined. For crystal dataset, folding-contact residues were excluded from this analysis. For the caption, see Figure 5-6.

The importance of polar residues in membrane proteins is known to depend on their depth in the bilayer [118]. Polar residues at juxtamembrane regions are common and unlikely to indicate an important role in protein-protein interactions [201]. Here it is shown that relative polarity (relative to the six immediate surrounding residues) was significantly higher than non-interfacial residues (Figure 5-6, Figure 5-8). This

enrichment of polar residues at protein-protein interaction (PPI) interfaces of membrane proteins was consistent with previous analyses of the more permanent TMD helix interfaces found in crystal structures [124, 131]. The extent of the difference in polarity was not as great as the difference seen in conservation (**Table 8-5**). In membrane proteins, it is known that there is a complex interplay between residue depth, polarity, and conservation, whereby highly polar residues located centrally in the TMD were well conserved [118]. In this study, however, there was no linear correlation between conservation and polarity ($R^2 < 0.05$) or relative polarity ($R^2 < 0.05$). As expected, the two polarity measures (polarity and relative polarity) were correlated with each other ($R^2 = 0.56$).

Depth in the membrane is a novel residue property investigated in this study. As the structures of the ETRA dataset are unknown, the true relative depth in the membrane could not be measured exactly, and was therefore estimated from the position of the residue in the TMD sequence. Here it is shown for the first time that homotypic TM interface residues have a higher depth in the bilayer than non-interface residues (Figure 5-6 D). The importance of depth in the membrane is consistent with reports that H-bonding between helices is more favourable in the apolar hydrophobic core [202], and the higher polarity of interface residues in the homotypic TM dataset (Figure 5-6 C).

5.2.1 The successful exclusion of possible spurious correlations

It has been shown that residues with a high depth in the membrane are strongly conserved [203]. It is shown above that interfacial residues had a higher depth in the membrane (Figure 5-6 D). Could the higher conservation of interface residues simply

be an artefact (spurious correlation) caused by the distribution of interfacial residues in the sequence? To test this hypothesis, the analysis performed in Figure 5-6 was repeated using a randomisation approach.

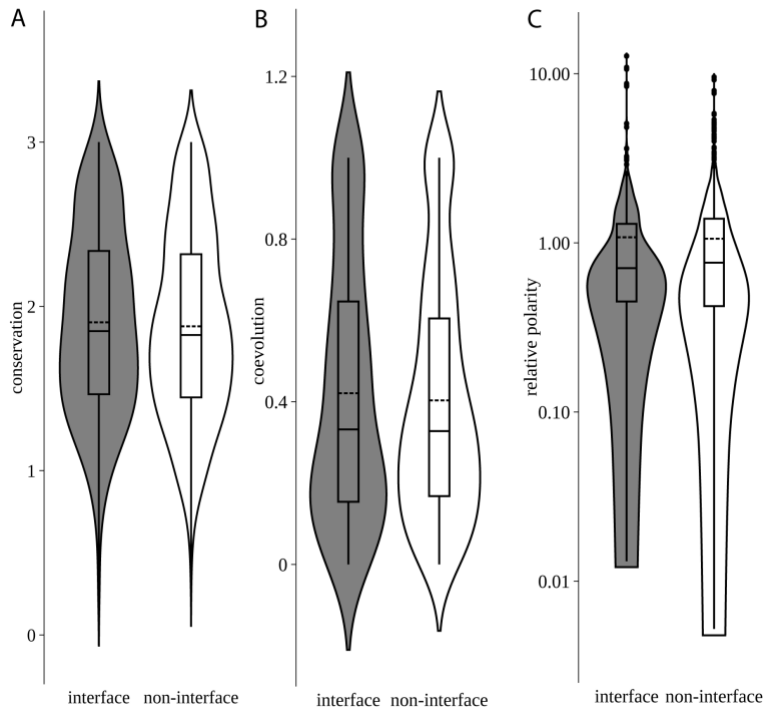

Figure 5-9: Randomisation of interfacial residues rejects the hypothesis of a spurious correlation. The analysis in Figure 5-6 was repeated by using interfacial residues that were chosen at random from another TMD. For the caption, see Figure 5-6. This figure confirms that the correlations were seen in A, B, and C of Figure 5-6 are not spurious correlations due to the "confounding factor" of residue depth indicated by D of Figure 5-6.

For each TMD, the position of interfacial residues was taken at random from another TMD in the dataset. This preserved the distribution of residue positions (depth in the membrane) in the entire dataset, but ensured that most of the positions labelled as "interface" residues were mostly non-interface positions. If conservation, coevolution or relative polarity were spurious correlations due to the presence of a third factor (depth in the membrane), it would be expected that they would still show higher values for interface residues, even after randomisation.

Instead, the results show that the higher conservation, coevolution and relative polarity of interfacial residues were completely abolished after randomisation (p = 0.507, 0.386 and 0.777 respectively, bootstrapped t-test Figure 5-9). This confirmed that none of these factors was a spurious relationship due to their non-random distribution in the TMD sequence.

5.3 The interface shows a strong helical pattern, but residue properties show only weak helicity

TMDs were aligned according to the residue judged to be the most important for the self-interaction. To identify this residue, an “interface score” was created for all residues in the homotypic TMD dataset, representing the importance of the residue for the self-interaction. For the ETRA subset, this was based on the average disruption after mutation. For the NMR and crystal subsets this was based on the closest heavy-atom distances between residue side chains. Once the TMDs were aligned based on their most important residue, at each position (relative to the central position) the average, conservation, coevolution and relative polarity were calculated.

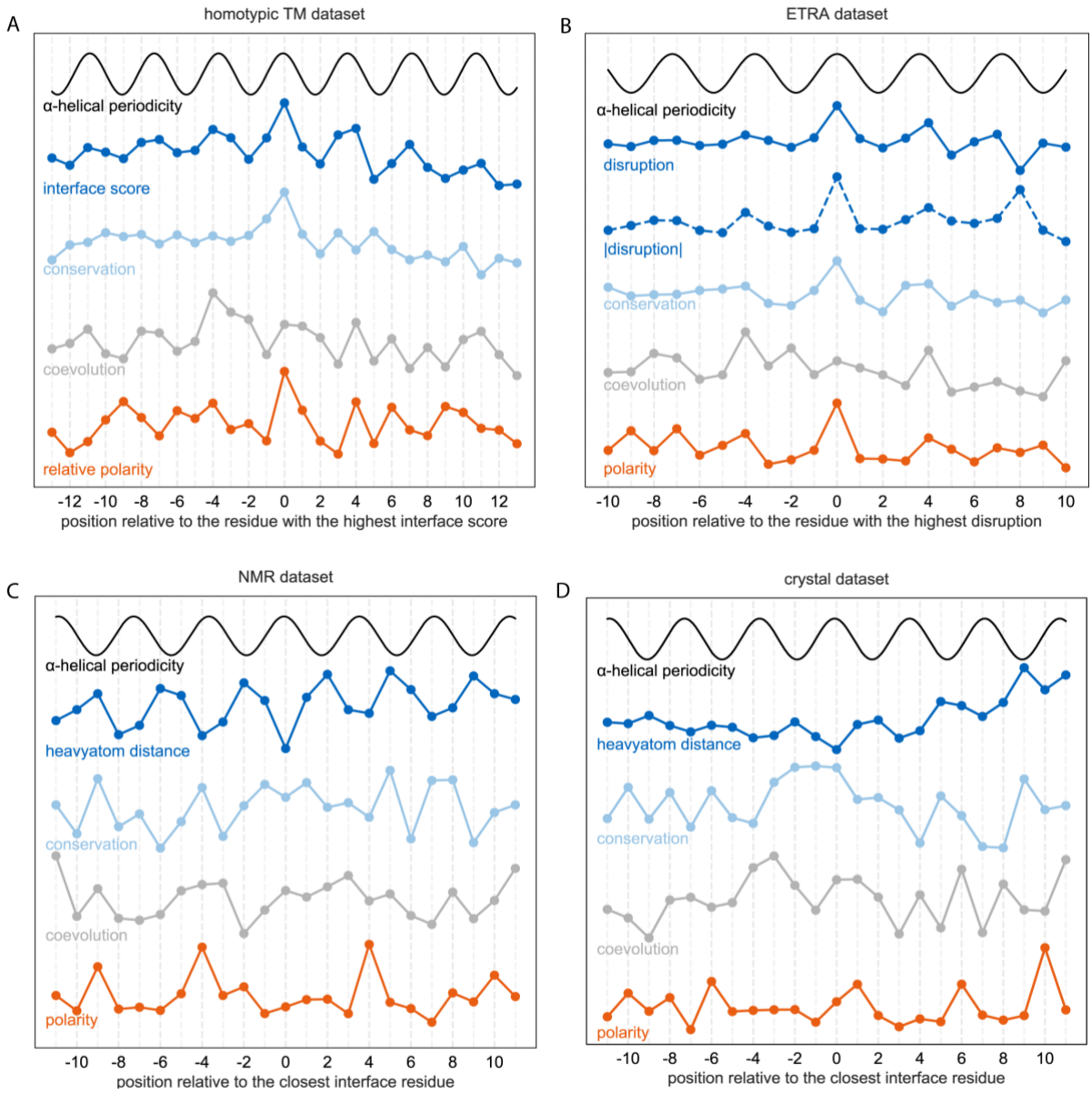

Figure 5-10: Interfacial residues were strongly α-helical, but only weak patterns were seen for conservation, coevolution and polarity. The sequences in three datasets were aligned centrally according to the most important interfacial residue. The mean values for the interface score, conservation, coevolution (DImax) and relative polarity were then calculated for each position in the alignment. Heterotypic contacts from the crystal dataset were excluded. (A) The homotypic TMD dataset. The most important interfacial residue in the TMD was defined as the residue with the closest heavy-atom distance (NMR and crystal TMDs) or the highest disruption after mutation (ETRA TMDs). To create an interface score that was compatible between ETRA and NMR/crystal TMDs, the ETRA disruption were normalised between -0.4 and +0.4, where any scores at or above +0.4 were equal to 1. Similarly, the closest heavy-atom distances of NMR/crystal were normalised between 2 and 10, whereby any heavy-atom distance of 10 or above was equal to 0. The mean of the normalised interface score at each position was then calculated exactly as for the other features. (B) The ETRA dataset. The most important interfacial residue in the TMD was defined as the residue with the highest disruption after mutation. (C) The NMR dataset. The most important interfacial residue in the TMD was defined as the residue with the closest heavy-atom distance NMR. (D) The crystal dataset. The most important interfacial residue in the TMD was defined as the residue with the closest heavy-atom distance. The interfaces

of NMR and crystal TMDs were provided by Bo Zeng, and residue properties were calculated using the THOIPApy software package.

The combined interface score of the homotypic TM dataset was exceptionally α-helical (Figure 5-10 A). This helical pattern even extended up to twelve residues upstream and downstream. It has been proposed that buried or interacting TM residues are on average more polar and conserved than other helix faces [36, 124]. However, the α-helicity of conservation, coevolution and relative polarity was weak in the homotypic TMD dataset, despite isolated peaks at positions i, i-4 and i+4. (Figure 5-10 A). Overall, the data was consistent with the results above (Figure 5-6), that there are only modest differences in conservation, coevolution and relative polarity values between interface and non-interface residues.

Importantly, the ETRA dataset showed a strong α-helical periodicity in the role of the residues (Figure 5-10 B). This helicity of disruption after scanning mutagenesis has previously been shown for individual ToxR case studies [39, 197], but never examined for all available TMDs with ETRA data. This strongly suggests that the majority of mutation-sensitive residues in ETRA data indeed located on a helical face, and are therefore directly involved in helix-helix contacts. The position with the highest importance for the TMD interaction, denoted i, showed a dominant peak of conservation and relative polarity. Peaks on the same side of the helix face, such as i-4 and i+4, were notably found. Surprisingly, the α-helicity of coevolution (DImax) values was stronger at position i-4 and i+4 than at the central position, i. As the central position, i, was highly conserved, this might simply reflect the difficulty in obtaining accurate coevolution values for highly conserved residues [204].

5.4 The evolutionary footprint associated with interfaces is unique for each individual TMD

In a total dataset of 1172 residues, interface residues were associated with a slight increase of conservation, coevolution and polarity. To understand this variability further, the number of TMDs were counted whose interface residues were, on average, more conserved, coevolved or polar than then non-interface residues. This revealed a high variability within individual TMDs. As an example, for 37% of the TMDs in the homotypic TMD dataset, interfacial residues were not more conserved than non-interfacial residues (Figure 5-11 A). Also, 33% of the TMDs had interface residues that were less polar than non-interface residues (Figure 5-11 A).

The individual analysis of the ETRA, NMR, and crystal datasets confirmed the overall trends seen in Figure 5-6 and described above. However, the residue property with the most prominence varied depending on the experimental technique. For ETRA this was relative polarity. For NMR this was the depth in the membrane. And for crystal structures this was the coevolution score (i.e. DImax). As the datasets are still quite small, further research is necessary to understand whether these differences are indeed a feature of each experimental approach.

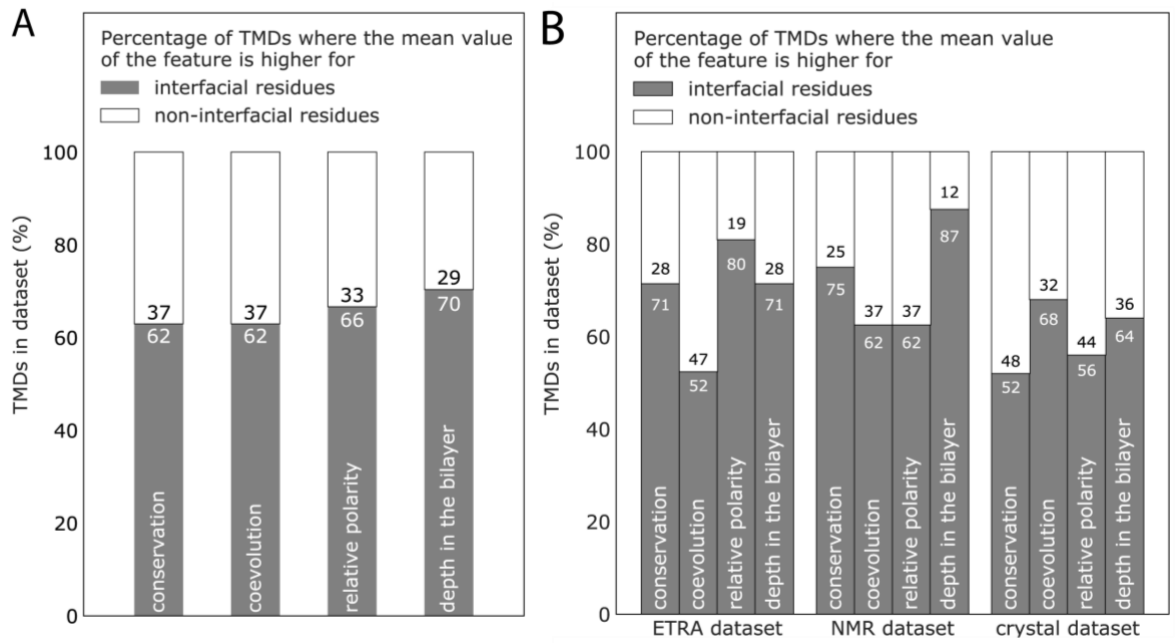

Figure 5-11: Individual TMDs have unique structural requirements, leading to high variability in residue features of interfaces. The percentage of TMDs is shown where the residue features (e.g. conservation) were on average higher for interfacial residues than non-interfacial residues. For the remaining TMDs (white), the residue features were higher for non-interfacial residues. This high variability shows that each TMD interface is under unique evolutionary selection pressure and that the maintenance of function may not necessarily require high conservation, coevolution or relative polarity. As a reference, within the bar, the percentage of residues involved were also shown. (A) Analysis of the homotypic TMD dataset. (B) Analysis of three datasets. For the crystal dataset, heterotypic contacts were excluded, as detailed in the methods. The interfaces of NMR and crystal TMDs were provided by Bo Zeng, and residue properties were calculated using the THOIPApy software package.

These results show clearly that the prediction of interface residues in TMDs based on any one single factor (e.g. conservation) is likely to have a high failure rate. Prediction of interface residues should therefore take into account many factors simultaneously, and rely heavily on any available experimental data.

Four members of the ETRA dataset (BNIP3, ADCK3, FtsB and siglec7) showed a good correlation between the disruption after mutation and the conservation value (R> 0.5) (Figure 5-12). However, overall only 15/21 TMDs showed a positive correlation (R> 0, Figure 5-12). The high conservation of the homodimer interface for FtsB (R= 0.524) was somewhat surprising, as it is also known to have an alternative interface

for hetero-oligomerisation [205]. In contrast, ATP1B1 clearly had a hetero-dimer interface [25] that was more conserved than the homo-dimer interface examined here, resulting in a slightly negative correlation. In addition, for GpA and a number of TMDs without known heterodimer partners, the interfacial residues that were less conserved than non-interfacial residues. Indeed, for the model homodimer GpA, there was a weakly negative correlation between the role at the interface and residue conservation (Figure 5-12, R= -0.253).

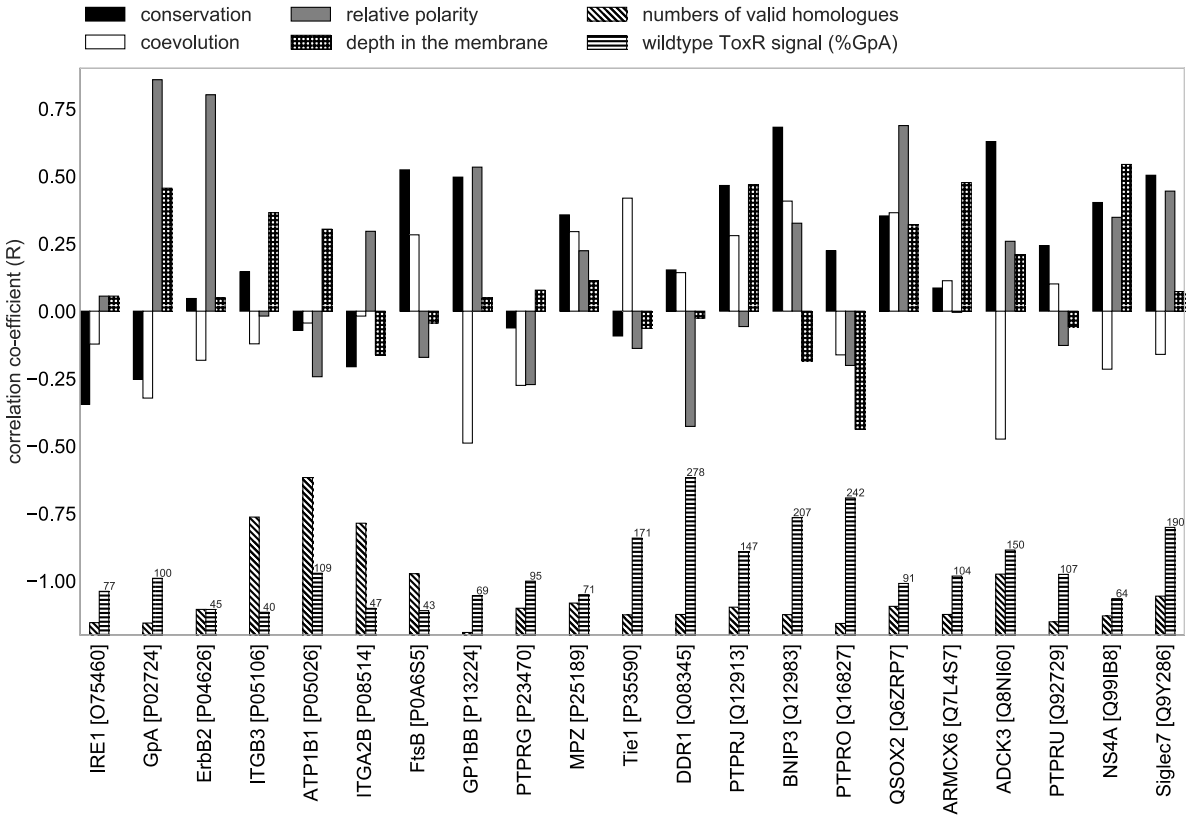

Figure 5-12: Correlations between ETRA disruption and residue properties reveals a high variability in the evolutionary footprint of TM homodimer interfaces. For each TMD in the ETRA dataset, the disruption after mutation was plotted against residue conservation, coevolution (DImax), relative polarity and depth in the bilayer. This was conducted only for the ETRA dataset, where the disruption index directly measured the impact of the residue on dimerisation. The R value was calculated to determine which factors gave a linear correlation to the importance of the homodimer. A positive R-value indicates a good correlation, suggesting that interfacial residues are conserved, coevolved, polar or centrally located. However, in many cases, a negative R value was observed, suggesting that the importance of the interaction was inversely correlated with the residue property. Note that this analysis was strongly affected by positions with negative disruption, representing mutations that increased self-affinity. The results, therefore, differ slightly to the binary comparison of interface and non-interfacial residues used in other analyses (e.g. Figure 5-6, Figure 5-8 and Figure 5-11). The number of available homologues and the

strength of the original homodimer is shown as a reference in the lower section of the graph. The residue properties were calculated using the THOIPApy software package of Bo Zeng.

In total, 20/21 TMDs showed an interface with a negative correlation for at least one of the important residue properties (Figure 5-12). The work in this chapter therefore shows that interface residues are on average more conserved, coevolved, polar and buried in the membrane. However very few interfaces have all of the above features.

5.5 TMD-TMD dimerisation is mediated by glycine and strongly polar residues

As expected, hydrophobic residues (Leu, Ala, Val and Ile) constituted the majority of residues in the TMD (Figure 5-13), However the percentage of these residues (LIVI) was low for the crystal dataset (65%, 67% and 47% for ETRA, NMR and crystal dataset respectively) (Figure 5-14 A).

Strongly polar residues (D, E, K, R, H, N, P and Q) are highly unfavourable within TMD regions [126]. A previous analyse of TMD sequences revealed that only about 25% of TM helices contain one or more strongly polar residues [206]. In general, they constitute only 4–6% of the total amino acid composition in TMDs [19]. Although polar residues create a thermodynamically unfavourable situation, the possible formation of salt bridges and H-bonds can reduce the energetic cost. This is easier in larger membrane proteins. Bitopic TMDs, in comparison are more exposed to the surrounding lipid media. Correspondingly, strongly polar residues were relatively rare in the homotypic TMD dataset as a whole (Figure 5-13), and especially in the bitopic ETRA (1.85%) and NMR (1.75%) datasets. Strongly polar residues were much more common in the (mostly polytopic) crystal dataset (10.31%).

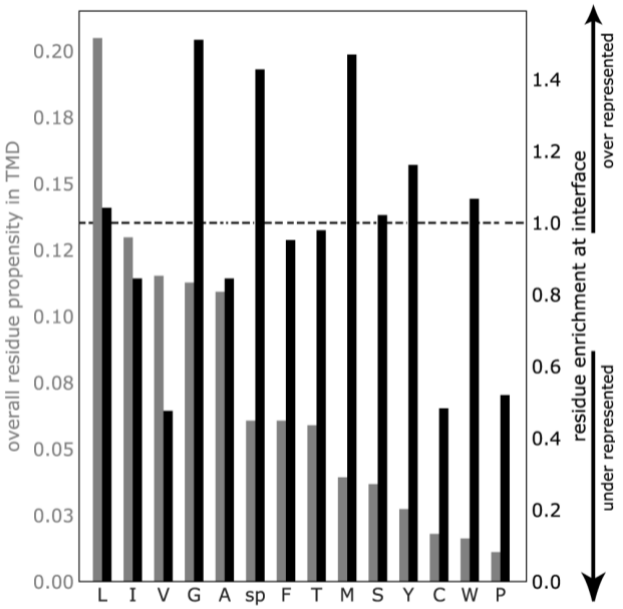

Figure 5-13: Gly, Met, and strongly polar residues are enriched at homotypic TM interfaces. The strongly polar residue types (sp = Asp, Glu, Lys, Arg, Asn, Gln, His) were combined, due to a lack of data when analysed individually. The residue enrichment at the interface is the proportion of the residue type at interfacial positions, divided by the proportion of the residue type within all TMD sequences, as described in Section 3.7.2. Values above 1 indicate over-representation at the interface. The interfaces of NMR and crystal TMDs were provided by Bo Zeng.

Strongly polar residues were highly enriched at homotypic TM interfaces (Figure 5-13). The importance of strongly polar residues was consistent with the higher relative polarity of interfaces shown above (Figure 5-6). The high propensity of strongly polar residues to be at a TMD-TMD interface (heterotypic or homotypic) is consistent with their high conservation in TM regions [118, 201]. Strongly polar residues can contribute to TMD interactions via side-chain/side-chain and side-chain/main-chain H-bonding [116]. Ionisable residues are proposed to form salt-bridges [207], which may depend on the depth of the side-chain in the membrane and the protonation state of the side-chain. Polar residues can contribute to the dimerisation only when they are appropriately placed, allowing the formation of interhelical H-bonds [206]. However, strongly polar residues are highly unfavourable for membrane insertion [208]. Thus, the number of bitopic TMDs whose interaction

depended on strongly polar residues was limited. Note that the apparent over-representation of strong polar residues at the interface for the ETRA dataset (Figure 5-9 B) may not be representative. The entire ETRA dataset of TMDs contained only eight strong polar residues, of which six were found at the interface.

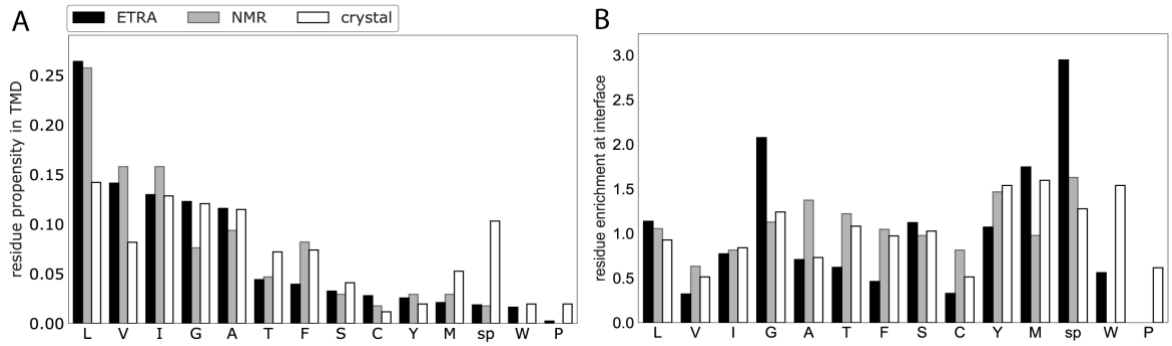

Figure 5-14: Overall frequency of amino acids within TMD sequence and their homotypic interfaces. Heterotypic contacts from crystal TMDs were removed from this analysis. The strongly polar (sp) residue types (Asp, Glu, Lys, Arg, Asn, Gln, His) were combined, due to a lack of data when analysed individually. A) Residue propensity within the TMD of the ETRA, NMR and crystal dataset. The propensity was calculated by numbers of that particular amino acid divide by total numbers of residues in that dataset as described in Section 3.7.1. The calculated composition was plotted for each residue type. B) Enrichment of amino acids at the interface of the ETRA, NMR and crystal dataset. Interfacial residues were defined as described in Section 3.3.2 and 3.4. The enrichment is the proportion of the residue type at interfacial positions, divided by the proportion of the residue type at the TMD sequence, as described in Section 3.7.2. Values above 1 indicate over-representation at the interface. Residues are ordered according to the frequency of occurrence within the ETRA dataset. The interfaces of NMR and crystal TMDs were provided by Bo Zeng.

5.6 Gly plays a key role in TMD dimerisation

Gly is a unique amino acid that lacks a side chain and known as an α-helix destabilising factor in soluble proteins [209]. In TMDs, Gly is thought to facilitate helix packing by allow close packing to facilitate van der Waals forces and/or H-bonding [1, 210].

In the homotypic TMD dataset, Gly residues were strongly enriched at the interface (Figure 5-13). This was especially strong for the ETRA dataset (Figure 5-9 B). The involvement of Gly residues is very robust due to the high number of residues involved.

From 128 Gly residues in total, 59 were found at the interface. Although Gly residues are usually mentioned in connection to the GxxxG motif, the interfaces containing glycine were often quite diverse. In fact, a correlation analysis of residues found together at interfaces showed Gly-Gly pairs at an interface were not more likely than Gly-Leu or Gly-Ala (data not shown). None of the interfaces consisted solely of Gly residues.

Positions with Gly residues were distinguished by high conservation, coevolution, relative polarity, and depth in the bilayer (Figure 5-15). This analysis was conducted for all Gly residues, regardless of their role at the TMD interface. For the ETRA dataset, the mutation of positions with Gly residues led not only to the biggest decreases in dimer signal, but also to the biggest increases. This is seen by the high |disruption| (absolute disruption) values at positions with Gly. Positions with Gly were clearly the most sensitive to mutagenesis.

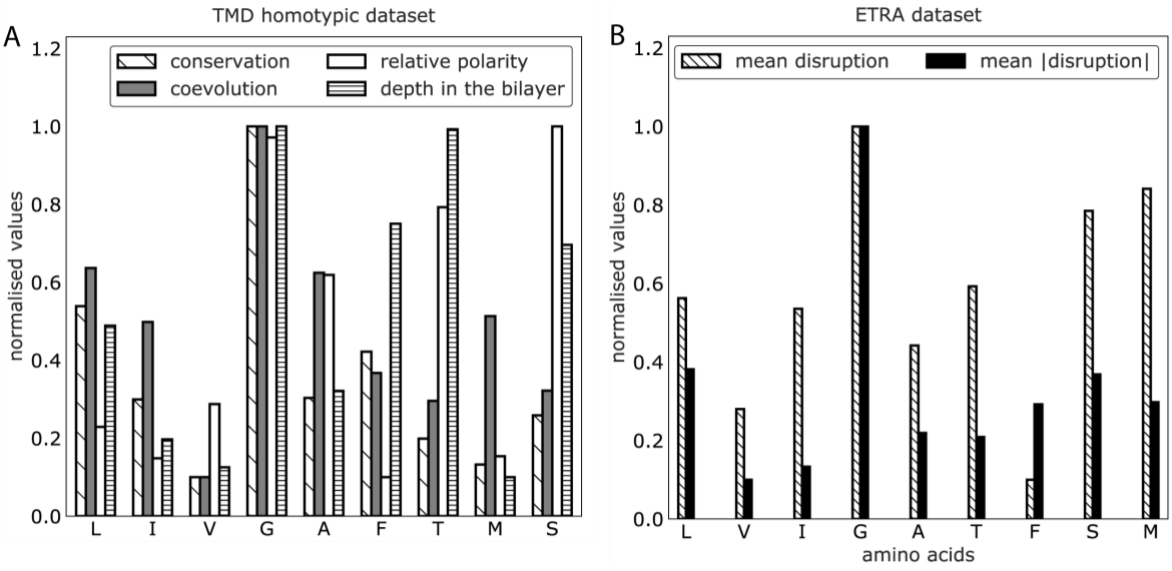

Figure 5-15: Positions with Gly residues show high conservation, coevolution, polarity, and depth in the bilayer. (A)Mean values are shown for all residues (interface and non-interface). Data for the seven residues present more than 40 times in the dataset are shown. Each data type (conservation, coevolution, relative polarity and depth in the bilayer) was normalised between 0.1 and 1 for comparison. (B) Within the ETRA database, positions with Gly residues showed the highest disruption and the absolute disruption (|disruption|). All residues (interface and non-interface) in the ETRA dataset were examined.

Data for the most abundant seven residue types are shown. Note that for the standard disruption score, mutations that give positive disruption (decreased dimerisation relative to wildtype) and negative disruption (increased dimerisation relative to wildtype) can even out to give a disruption value of zero. In contrast, the (absolute) |disruption| score is increased whenever a mutation leads to any change in the dimer strength. Each data type (disruption, |disruption|,) was normalised between 0.1 and 1 for comparison. The interfaces of NMR and crystal TMDs were provided by Bo Zeng.

Relative polarity was calculated from residues in the multiple sequence alignment rather than for the amino acid used in the experiments. However the relative polarity for Gly and Ser position was consistent with the polarity of Gly and Ser according to the Engelman (GES) hydrophobicity scale used in this study [138]. Positions with Gly were associated with the second-highest relative polarity after Ser, when only the most common residues types in TMDs are taken into account. Overall, the high conservation, coevolution, relative polarity and relative depth of Gly residues shows that the importance of Gly is not an isolated feature. Instead, it was supported by the entire evolutionary footprint associated with TM homodimer interfaces.

5.7 A quantitative analysis of GxxxG motifs confirms their over-abundance at natural TM interfaces

To find out sequence-specific patterns associated with homotypic TMD helix-helix interactions, the frequency of simple sequence motifs was analysed. The observed abundance of motifs in TMDs or interfaces was compared to the expected randomise value. This was derived from random sequences with the same amino acid propensity and length as the original sequence.

Numerous case studies and artificial selection experiments have shown that GxxxG motifs can drive homotypic TM interactions (reviewed in [97]). The GxxxG and related (small)xxx(small) motif is unusually abundant in TM regions considering the overall

proportion of Gly residues. This is no evidence, however, that all these motifs are involved in TMD interactions. Until now, there have been no quantitative studies that have proven the over-abundance of GxxxG motifs at natural homotypic TM interfaces.

The analysis in this study shows that 57% of GxxxG motifs were found at a TM homodimer interface (Figure 5-16). The abundance is therefore far higher than that expected by random chance. The over-representation GxxxG motifs at the interface were highest for ETRA TMDs (Figure 5-17). Nevertheless, almost half of the motifs in the homotypic TMD dataset were not found at the interface. The GxxxG motif alone therefore is not sufficient for interface prediction. Instead, the GxxxG motif, like sequence conservation, should be considered a strong predictive factor for interface residues.

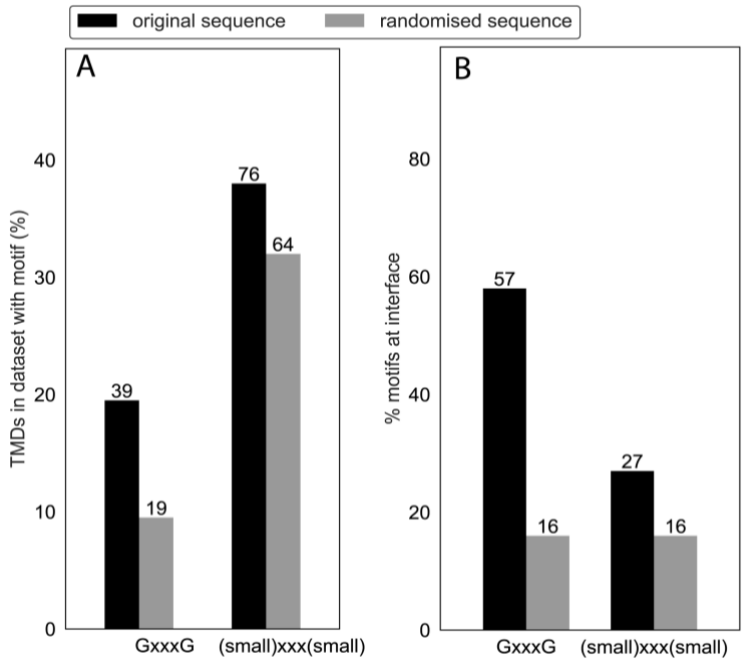

Figure 5-16: GxxxG motifs are strongly associated with interfaces. Small residues within the (small)xxx(small) motif were defined as Gly, Ala, Ser or Cys. To understand the expected abundance, random sequences were created with the same amino acid propensity and length as each original sequence. The mean result for 100 randomised sequences is shown. Note that the heterotypic contacts of the crystal dataset (see Section 3.4) were not removed from this analysis, to allow unbiased motif identification and randomisation. A) Analysis of motif abundance in the full TMD sequence. Values higher than random suggest that the motif is overrepresented in the dataset. B) Analysis of the propensity of motifs to

contain interfacial residues. A motif was counted only if both residues were located at the predicted interface. The interfaces of NMR and crystal TMDs were provided by Bo Zeng.

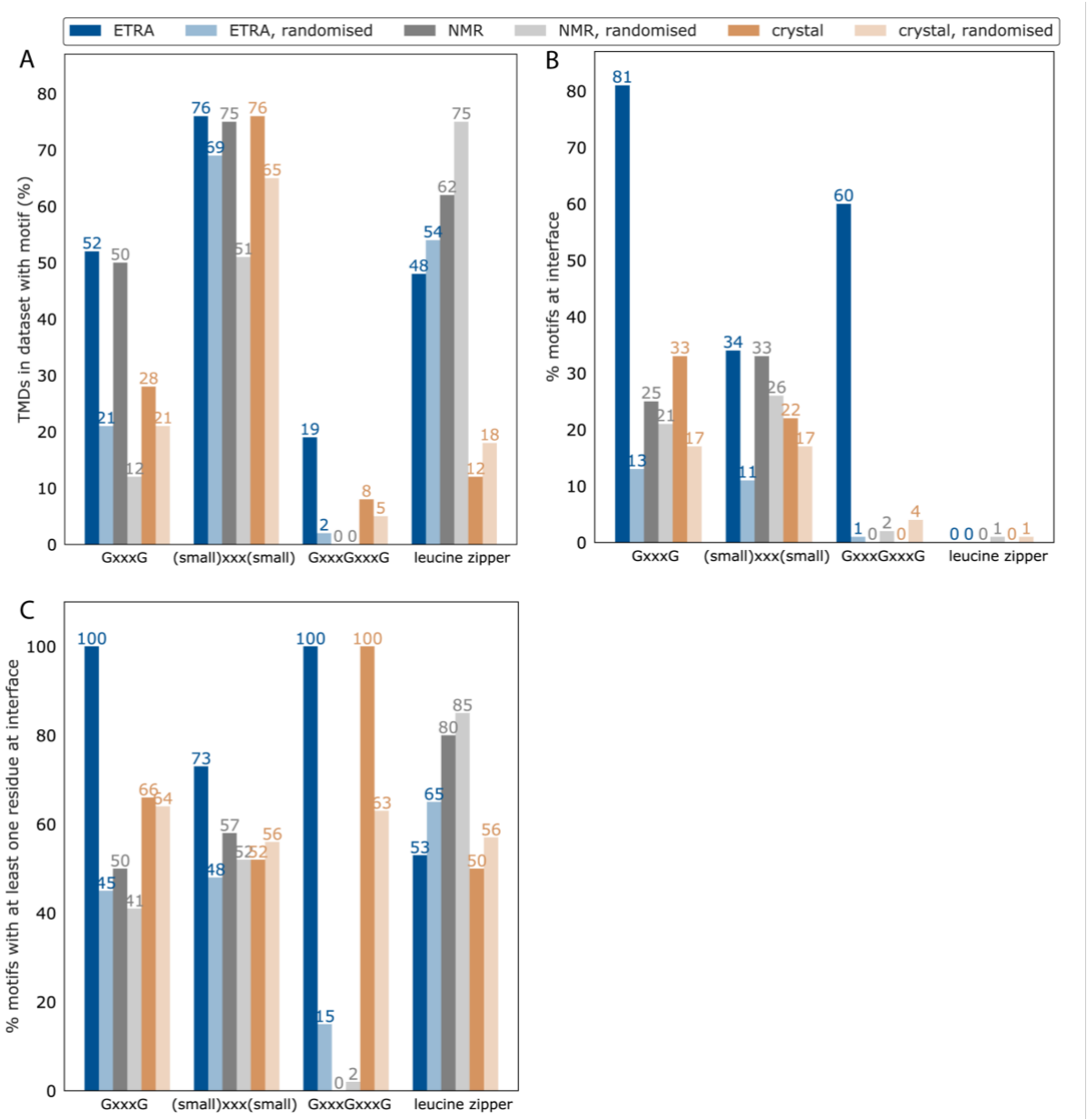

Figure 5-17: A detailed analysis reveals the importance of GxxxG motifs for the ETRA dataset. Interfacial residues were defined based on heavy-atom distances (crystal, NMR) or disruption after mutation (ETRA) as defined in the methods. Small residues within the (small)xxx(small) motif were defined as Gly, Ala, Ser or Cys. The leucine zipper was defined as LxxLLxL, where L was either Leu, Ile or Val. Randomised values were derived from random sequences with the same amino acid propensity and length as the original sequence. The mean value for 100 randomised sequences is shown. The interfaces of NMR and crystal TMDs were provided by Bo Zeng. A) Analysis of motif abundance in the full TMD sequence (interface and non-interface). Values higher than random suggest that the motif is overrepresented. B) Analysis of the propensity of motifs to contain interfacial residues. Values higher than random suggest that the motif is indicative of an interfacial residue. The higher-than-random value for the ETRA dataset and the (small)xxx(small) motif is mostly

due to the GxxxG motifs. After excluding glycines, the (small)xxx(small) motif is no longer associated with interfacial residues any more than a random selection (8% motifs with all residues at the interface, in comparison to 12% random). Note that in this analysis, a single residue could participate in more than one motif (i.e. GxxxGxxxG contains two GxxxG motifs). C) Percentage of motifs where at least one residue was at the interface.

The more inclusive (small)xxx(small) motif was highly abundant in TMDs (76% of TMDs), but only slightly over-represented at interfaces (Figure 5-16, Figure 5-17). This was consistent with a number of ToxR-based studies [18, 105, 108, 211], which tended to show a dominant role of Gly residues at interfaces, and a secondary role for other small residues. Leucine zippers were not represented at interfaces more than expected by random chance (Figure 5-17).

As previously described [19, 97], the GxxxG and (small)xxx(small) motifs were more abundant in the TMD sequences than expected by random chance (Figure 5-16, Figure 5-17). Note that these data did not necessarily imply a causal relationship to the overabundance at interfaces. Theoretically, the overabundance of the GxxxG motif within TMDs could be due to other factors such as favourable membrane insertion, or TM-lipid interactions.

Two GxxxG motifs form a Gly zipper motif (GxxxGxxxG), which has been shown to be overrepresented in transmembrane proteins [29]. In the ETRA dataset, the Gly zipper was found five times in four TMDs (18% of TMDs, Figure 5-17), It is therefore relatively rare. In three of these cases, the Gly zipper motif was completely involved in the interaction. Although the numbers are small, this preliminary data suggests that the GxxxGxxxG motif is a powerful indicator of a homotypic TMD interfacial residues.

Overall this data shows quantitatively for the first time that (small)xxx(small), GxxxG and Gly zipper motifs indeed have prediction power for the identification of interface residues. Mutation of residues in these motifs, as done in many case studies, is

indeed strong for the identification of true interface residues. However it is difficult to determine true interfaces without experimental data, as many of these motifs were not located at the helix-helix interface. This is consistent with previous studies showing that the role of the motifs is strongly determined by sequence context [206].

CHAPTER 6. DISCUSSION

6.1 Interfaces are diverse

Until now, researchers in the field of homotypic TM interactions have shown an intense interest in the analysis of simple sequence motifs such as GxxxG, and have rarely emphasised the diversity of homotypic TM interfaces. In fact, new interfaces in the literature with up to six residues are sometimes described as “motifs” despite the fact that such a long sequence is almost certainly unique. There is currently no evidence of the convergent evolution of homotypic TM interfaces [21]. By creating non-redundant ETRA, NMR, and crystal datasets of self-interacting TM helices, this study has illustrated the incredible diversity of residue combinations that give rise to self-interaction. In the homotypic TM dataset, 19 from the 20 natural amino acids were found at interfaces. All TM interfaces were unique. However this study shows that on average, the interface residues share some common features.

6.2 Interface residues are often conserved

In this study, it is shown for the first time that TM homodimer interfacial residues are significantly more conserved than non-interfacial residues (Figure 5-6 A). This has been previously implied in numerous case studies [25, 39, 103, 158, 212, 213]. This finding is in line with the well-established finding that residues buried in the protein structure are more conserved in both soluble and membrane proteins [136, 214]. Despite the identification in recent years of many highly-conserved protein-lipid interactions [215], it remains a fact that, on average, lipid-facing residues have low conservation. Several exceptions are noted. The most prominent of these was the well-researched TMD of GpA. In the alignment against homologues, glycines of the

central GxxxG motif in GpA were often exchanged with small or polar residues, and are therefore poorly conserved. A broader analysis revealed that in the full homotypic TM dataset, over 37% (Figure 5-11 A) of the TMDs had interface residues that were, on average, less conserved than non-interface residues.

High residue conservation invariably indicates that the retained residue is important for the survival of the organism. The higher conservation of interfacial residues of the crystal, NMR and ETRA datasets is evidence that the majority of these homodimer contacts are biologically relevant. Many of the interfacial residues were polar, and it is known that polar residues in TM regions are highly conserved [118].

In some cases, the homodimer interface may involve the same helix face as a more biologically relevant heterodimer. There are several examples of TM helices with reported homo and heterodimers, including integrin αIIbβ3 [40] and syndecans [216]. In the ETRA dataset, ATP1B1 TMD was an example that showed a poor correlation between conservation and disruption after mutation (Figure 5-12 E). It contained a highly conserved FYxxFYxxLxxxF motif [25], of which only the final phenylalanine formed part of the ToxR homodimer interface (Figure 4-4 B). Instead of facilitating homotypic interactions, these conserved residues may instead be responsible for hetero-contacts with α-subunit [25]. The presence of a shared heterodimer interface does not exclude a biological function for the homodimer.

If the strong homodimers prevent the formation of a biologically relevant heterodimer interface, how would the cell prevent homodimer formation? Could the homodimer/heterodimer equilibrium itself be important? To answer these questions, a more detailed understanding of TMD interactions in the cell is required. Cellular reporter assays such as ToxR are particularly powerful, as they provide evidence of strong, specific dimerisation of the TMD in a natural membrane environment. The

derived ToxR interfaces are an excellent guide for mutation studies aiming to determine the role of the homodimers in the cells of the original organism. Recent advances in heterotypic *E. coli* reporter assays may allow the determination of both homo- and hetero-typic interfaces [76].

6.3 Interfacial residues are sometimes coevolved

Residue coevolution, or covariance, is based on the observation that a natural mutation in a buried residue is often accompanied by a mutation in a contacting partner residue [200, 217]. The patterns discerned from multiple-sequence alignments against homologues allow the identification of contacting residue pairs [200]. The usefulness of coevolution data has dramatically improved in recent years due to the exponential increase in sequence data, and also the development of sophisticated coevolution algorithms [139]. Coevolution is now a key factor in the identification of interacting helix pairs within multipass membrane proteins, the first step in de-novo structure determination [36, 200].

However the results in this study it is confirmed that coevolution data is somewhat less useful for identifying interface residues in symmetric TM homodimers. In this case, the most important contact residue is often the identical position in the opposing chain, and therefore has no coevolution score. In their pioneering work, the Barth lab argues that the pairwise coevolution scores between interfacial residues are higher than the coevolution scores between non-interfacial residues [36]. These metrics, however, are interface-dependent, in that they can only be calculated where the interface is already known. Here it is shown that the linear correlation between the interface-independent metrics and a role at the interface is relatively weak (Figure

5-12). For the homotypic TM and crystal datasets, significantly high coevolution values for interfacial residues than non-interfacial residues were observed (Figure 5-6 B, Figure 5-8 B). However, this was not true for the small ETRA and the NMR dataset (Figure 5-8 B). This indicates that current measures of coevolution are poorly adapted to the investigation of TM homodimers. There is a strong need for improved algorithms that are specifically designed to investigate contacts between residues that are located close to one another in the polypeptide chain. There is also a strong need for more homologous sequences. The number of homologues currently available in public databases is a known limitation of all coevolution methods [140]. The exponential increase in available sequence data should greatly improve the usefulness of coevolution values for TM homodimer interface prediction.

6.4 Interfacial residues are often polar

Homodimer interfacial residues were significantly more polar than non-interfacial residues for the homotypic TM, NMR and ETRA datasets. This trend is in line with many studies that show the higher polarity of buried, interfacial residues [118]. The importance of polar residues in TMD interactions is also supported by the observation that polar residue substitutions are the most common disease-causing mutations in membrane proteins [118]. GpA is an excellent example of the importance of polarity, as the central GxxxGxxG motif was strongly polar in multiple sequence alignments, but poorly conserved. A contrary example is ErbB3, where polarity is of little consequence, and the NMR interface effectively ignores a series of polar positions [27]. Modelling studies have had trouble reproducing the ErbB3 dimer structure for this reason [36]. Surprisingly the importance of polarity was quite low for the crystal dataset (Figure 5-8 C). The initial hypothesis to explain this result was hypothesised

that this was due to the unique nature of the homodimers within larger crystal structures, which also have hetero-contacts between non-identical TM helices. However, the importance of polarity was minimal in the crystal dataset, regardless whether hetero contact residues were removed (data not shown).

A problem with a single measure for polarity is the large discrepancy between different available hydrophobicity scales [135-137, 205]. In particular, biological hydrophobicity scales often disagree on the insertion costs of positively charged residues. Due to the positive inside rule, these residues can strongly promote membrane insertion, and in some scales, are scored as favourable [126]. However, this favourability depends on the position relative to the membrane. An arginine at the cytoplasmic side of the membrane is likely to be important for insertion, whereas an arginine located in the centre of the membrane is extremely likely to be a buried contact residue. The LIPS algorithm solves this problem by using separate hydrophobicity scales for central and peripheral TM residues [124]. However, the LIPS approach is very arbitrary, as it applies a different hydrophobicity scale to the first and last 5 residues of the TMD. The LIPS method is therefore dependent on the method used to define TM regions. In this study, a more consistent and logical method is used, whereby the “relative polarity” is the polarity divided by the polarity of the surrounding six residues. This successfully lowered the polarity score for charged residues at the interface regions. As a measure of the success of this approach, both polarity and relative polarity were statistically higher for interface residues. However the relationship with relative polarity was much stronger (Table 8-4). Further studies should therefore test various measures of relative polarity to identify those that have the strongest predictive power for TMD interfaces.

6.5 The depth in the membrane is a novel indicator of interface residues

A simple but novel feature associated with interfacial residues is their depth in the bilayer. This may suggest that helix-helix pairs are more stable when their interacting sites are deeper in the bilayer, increasing the favourability of polar residue-residue contacts in the absence of water [10, 218, 219]. This effect was seen in TMD homodimers investigated by ETRA, NMR and crystal studies. This 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 a common feature of protein-protein interactions mediated by membrane helices.

6.6 TMD interface properties do not form strong helical patterns

After aligning the sequences to the position with the highest disruption, some helical patterns were seen in the data for conservation, polarity and coevolution (Figure 5-10). The weak nature of the patterns seen in this study is in line with previous studies of TMD-TMD interactions in polytopic membrane proteins [17, 109].

Previous studies proposed that helix faces follow either a tetrad $[\underline{a}bc\underline{d}]_n$ [220] or heptad $[\underline{a}bc\underline{de}f\underline{g}]_n$ [113] motifs. The tetrad motif is associated with right-handed helix pairs and the heptad motif with left-handed helix pairs. In this study, we utilised ETRA-derived interfaces, whose helix-helix orientation could not be assigned. This justifies our approach to use a fit to sine with a periodicity of 3.6, rather than a heptad or tetrad motif. It has been argued that the heptad motif is superior for predicting interfaces for both left and right-handed helix pairs [124]. This may simply be because the 7-residue (heptad) repeating element has a much closer fit to a periodicity of 3.6 than a 4-residue (tetrad) repeating element. In this study, it is argued that a fit to sine at a

periodicity of 3.6 is superior to both of these rigid repeating elements for cases where the helix-helix orientation is mixed in the database, or unknown. In any case, the data presented in this study suggests that averaging features over a helical periodicity will only show a weak benefit for interface prediction, because the helical patterns are not strong (Figure 5-10). A larger dataset of characterised TMD interfaces is needed to determine if this weak link to helicity is indeed a feature of homotypic TM interactions.

6.7 Gly residues dominate many homotypic TM interfaces

The abundance of small residues at interfaces is thought to allow closer helix-helix contacts. This may increase van der Waals forces, or allow the formation of Cα-H-bonds. [1, 67, 105, 115, 221]. However, all of these features could be supported to be a greater or lesser extent by other small residues such as Ser, Ala, and Cys. It is currently not clear why Gly residues dominate TMD-TMD interactions.

Gly residues proved to be most important for the ETRA dataset (Figure 5-14 B). The ETRA dataset contained only human proteins, and human TMDs are known to be relatively Gly-rich [97]. However, this does not explain the dominance of Gly in the ETRA dataset, as the NMR and crystal datasets also had a similar proportion of Gly residues in their TMD sequences (Figure 5-14 A). The vast majority of interfacial Gly in the ETRA dataset were derived from the previous literature (i.e. GpA [22], BNIP3 [6], ADCK3 [11], ITGA2B [35], and ErbB2 [22]), rather than the newly released interfaces (ARMCX6, PTPRU and siglec7). Specifically, the authors of previous papers have noted their tendency to investigate TMDs that were rich in Gly residues that were proposed to mediate interfaces [9, 11, 103]. However, this bias can only explain a small proportion of the important Gly residues in the dataset.

Another possible reason for the dominance of Gly is that by measuring mutation sensitivity, ToxR disruption more faithfully reflects the importance of each residue. In contrast, the role of each residue in the crystal and NMR datasets was estimated based on distances between the heavy atoms of contacting residues. Are all contacting residues vitally important for the dimer? A 3.5 Å cut-off for NMR structures was chosen in this study, so that the interface roughly matched the interface proposed by authors in each NMR study (Table 8-1). Attempts to define interfaces by C-α distances, C-β distances, or relative surface accessibility led to interface residues that did not match those subjectively chosen by the authors (data not shown). One possibility is that the authors have over-estimated the accuracy of the contacts seen in heavy-atom distances, and that C-α or C-β distances give a better understanding of biologically relevant, contacting residues. For example, the highly-regarded CASP (The Critical Assessment of Techniques for Protein Structure Prediction) system for ranking de-novo structural predictions in soluble proteins is heavily reliant on C-β distances [222]. Further studies should therefore investigate the optimum definition of interface residues from structural data.

Could Gly-dependent interfaces give usually high dimer signals in ETRA assays? Other studies appear to support this theory. In two cases (ErbB2 and APP) [22, 223], the ToxR dimer interfaces were based on a GxxxG motif, whereas available NMR interfaces were not [13, 223]. A previous selection of strongly dimerising TMDs from random libraries yielded almost exclusively Gly-rich interfaces [18]. More recent artificial selection studies have resulted in a broader range of amino acids within strong dimers, but Gly residues are invariably overrepresented [74]. Also, in a TOXCAT study of TMDs with (small)xxx(small) motifs, those with Gly residues tended to give a higher dimerisation signal [105]. However, some date obtained in this study

argue against the ability of Gly and GxxxG to generally give high ETRA signals. From the ToxR experiments, there was not a single Gly residue among the interfaces of the three TMDs with the highest ToxR dimerisation score (DDR1, PTPRO and DDR2).

6.8 GxxxG motif has predictive power for interface identification

This is the first study to show quantitatively the overabundance of GxxxG motifs at natural heterotypic TM interfaces. There has been considerable debate regarding the predictive power of GxxxG motifs, as they are found not to mediate TMD interactions in several different examples [97]. Nevertheless, the data shown here is in broad agreement with most studies in the field, where GxxxG motifs are considered key indicator of homotypic TM interfaces [6, 11, 21, 22, 31, 70, 224, 225]. The GxxxG motif should therefore be considered a predictive feature that can be used to identify interface residues, along with other features such as residue conservation, polarity, coevolution, and residue depth.

It is important to note that GxxxG motifs do not always convey dimerisation. This conclusion is corroborated by the solved structures of ErbB3 [27] and DAP12 [20]. The interface of these dimers does not dependent on the GxxxG motif; despite the fact that the motif is present in the sequence.

In this study, the (small)xxx(small) motif exhibited a much more moderate overabundance at interfaces than the GxxxG motif (Figure 5-17). This result is consistent with a number of ToxR-based studies, which tend to show a dominant role of Gly residues at interfaces, and a secondary role for other small residues [18, 105, 108, 211]. Overall, it supports the hypothesis that Gly is a “special case,” whose

importance in promoting homotypic TM interactions is greater than other small residues with similar chemical properties.

CHAPTER 7. CONCLUSION AND OUTLOOK

In this thesis I conducted the following:

- Defined interfaces of a total of 10 TMDs by using the ToxR-assay combined with scanning mutagenesis.
- Created the first *E. coli* TM reporter assay (ETRA) dataset of 21 TMDs, whose interfaces were defined in this study or taken from literature.
- Analysed the sequence properties of the first homotypic TM dataset, comprising 54 TMDs from ETRA, NMR and crystallography studies.
- Proved that properties of TMD homodimer interfaces from the ETRA and NMR datasets have a lot in common with the permanent interfaces within multi-pass membrane proteins, such as higher conservation, coevolution and relative polarity.
- Proved that homotypic TM interfaces have a high depth in the bilayer and have statistically overrepresented abundance of the GxxxG motif.

Table 7-1: Summary of findings.

	origin of theory[a]	key ref	applicable to homotypic interfaces?	statistically robust for natural TMDs?	supported by this study[b]
conservation	heterotypic interface	[7]	unknown	●	●
polarity	heterotypic interface	[10]	unknown	●	●
small residues	heterotypic interface	[17]	unknown	●	●
GxxxG	homotypic interface	[18, 19]	●	unknown	●
(small)xxx(small)	homotypic interface	[23]	●	unknown	●
leucine zipper	homotypic interface	[26]	●	unknown	○
glycine zipper	homotypic interface	[29]	●	unknown	●
coevolution	homotypic interface	[36]	●	unknown	●
membrane depth	this study	N/A	N/A	N/A	●

[a] heterotypic interface refers to studies of TMD interactions within polytopic membrane proteins. Homotypic interface refers to ETRA or NMR case studies using bitopic membrane proteins.
[b] ● strongly supported. ● weakly supported. ○ not support or not enough data.

The major findings in this study concern the residue properties of TM homodimer interfaces, as summarised in (Table 7-1). Some of these residue properties (conservation, coevolution, polarity) had been previously shown for heterotypic interfaces. However it had not been known if they were applicable to homotypic interfaces. Many others had been derived from case studies of TM homodimers or artificial selection studies. Until now, it had not been known whether these findings were statistically robust for natural TMDs. And finally, this study also presents a completely new feature of TM homodimer interface residues, their depth in the membrane.

This study identified several trends associated with interface properties. For predictive purposes, however, these trends are quite weak. One thing that is not well understood is why such a large number of interfaces had low conservation,

coevolution, or polarity. Since all these factors were calculated based on multiple sequence alignments, this might reflect the fact that some interfaces are rapidly acquired or lost, and do not leave a strong evolutionary footprint. More studies should be made to understand how long interfaces are retained, and whether recently-evolved interfaces can be identified. Here we show that the DDR1 and DDR2 homologues proteins have the same TM interface.

This study shows clearly that the GxxxG motif is overrepresented in natural TM interfaces to a far greater extent than the broader (small)xxx(small) motif. What is not proven, however, is a clear role of other motifs such as the glycine zipper (e.g. GxxxGxxxG) and the leucine-zipper. In the case of the former, this is difficult because of the rare nature of the motif, making the statistical analysis difficult. In the case of the leucine zipper, the problem is the unclear definition of the motif in the literature. The preliminary analysis presented here suggests that the leucine zipper plays only a minor role in TMD interactions, however individual aliphatic residues (e.g. Leu, Ile, Val) are very common at interfaces. Further studies are necessary to define the motifs more specifically, and to determine their exact role at homotypic TM interfaces.

This study has focused in the properties of TM homodimers. However it has been proposed that TMDs of bitopic proteins can have multiple homodimer interfaces [60], or additional heterodimer interfaces [25, 205]. Newly developed methods, such as BlaTM [76] provide the possibility to analysis heterodimeric TMDs. Further studies should therefore attempt the high-throughput analysis of TM heterodimer interface residues. Scanning mutagenesis of both TMDs that participate in a heterodimer would give equivalent data to the ETRA dataset as shown here. The properties of the TM heterodimer interface residues could then be analysed in detail, as concluded in this thesis.

CHAPTER 8. APPENDIX

Table 8-1: 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]	[5]
2jwa	ErbB2 [P04626]	[13]
2k1k	EphA1 [P21709]	[14]
2l34	TYROBP [O43914]	[20]
2k9y	EphA2 [P29317]	[24]
2l9u	ErbB3 [P21860]	[27]
2loh	APP [P05067]	[30]
2l6w	PDGFRB [P09619]	[37]
2lcx	ErbB4 [Q15303]	[38]
2m0b	EGFR [P00533]	[41]
2lzl	FGFR3 [P22607]	[42]
2mk9	TLR3 [O15455]	[43]
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 8-2: Sequence frames of TMDs tested with a high conservation moment..

protein (acca)	conservation moment	frames	TMD sequences
PLXDC1 (Q8IUK5)	0.49	0	GTIVGIVLAVLLVAAIILAG
		1	TIVGIVLAVLLVAAIILAGI
		2	IVGIVLAVLLVAAIILAGIY
		3	VGIVLAVLLVAAIILAGIYI
NDST3 (O95803)	0.49	0	TVILLATFCMVSIIISAYYL
		1	VILLATFCMVSIIISAYYLY
		2	ILLATFCMVSIIISAYYLYS
		3	LLATFCMVSIIISAYYLYSG
SHISA9 (B4DS77)	0.45	0	LIVYIICGVVAVMVLVGIFT
		1	IVYIICGVVAVMVLVGIFTK
		2	VYIICGVVAVMVLVGIFTKL
		3	YIICGVVAVMVLVGIFTKLG
CANX (P27824)	0.43	0	WLWVVYILTVALPVFLVILF
		1	LWVVYILTVALPVFLVILFC
		2	WVVYILTVALPVFLVILFCC
		3	VVYILTVALPVFLVILFCCS
IRE1 (O75460)	0.42	0	MATIILSTFLLIGWVAFIIT
		1	ATIILSTFLLIGWVAFIITY
		2	TIILSTFLLIGWVAFIITYP
		3	IILSTFLLIGWVAFIITYPL
CADM3 (Q8N126)	0.42	0	HAIIGGIVAFIVFLLLIMLI
		1	AIIGGIVAFIVFLLLIMLIF
		2	IIGGIVAFIVFLLLIMLIFL
		3	IGGIVAFIVFLLLIMLIFLG
THSD7A (Q9UPZ6)	0.40	0	TWVYGVAAGAFVLLIFIVSM
		1	WVYGVAAGAFVLLIFIVSMI
		2	VYGVAAGAFVLLIFIVSMIY
		3	YGVAAGAFVLLIFIVSMIYL

a Accession number (acc) is taken from the UniProt database.
b The calculated conservation moment according to Ried et al. [15, 16].

Table 8-3: Sequence and interface residues of TMDs in the ETRA dataset

#	protein (acc[a])	TMD sequence	self-affinity[b]	reference
1	DDR1 (Q08345)[c]	ILIGCLVAIILLLLLIIALML	278	[3]
2	PTPRO (Q16827)[c]	VVVISVLAILSTLLIGLLLVTLIIL	242	[9]
3	Siglec7 (Q9Y286)[c]	VLLGAVGGAGATALVFLSFC	190	[21]
4	Tie1 (P35590)[c]	LILAVVGSVSATCLTILAALLTLV	171	[3]
5	ATP1B1 (P05026)[c]	LLFYVIFYGCLAGIFIGTIQVMLLTI	109	[25]
6	PTPRU (Q92729)[c]	LILGICAGGLAVLILLLGAIIVII	107	[9]
7	ARMCX6 (Q7L4S7)[c]	REVGWMAAGLMIGAGACYCV	104	[21]
8	PTPRG (P23470)[c]	IIPLIVVSALTFVCLILLIAVLV	95	[9]
9	IRE1 (O75460)[cd]	ATIILSTFLLIGWVAFIITY	77	N/A
10	BNIP3 (Q12983)	LLSHLLAIGLGIYIG	207	[6]
11	ADCK3 (Q8NI60)	LANFGGLAVGLGFGALA	150	[11]
12	PTPRJ (Q12913)	ICGAVFGCIFGALVIVTVGG	147	[9]
13	GpA (P02724)	LIIFGVMAGVIGTIL	100	[22]
14	QSOX2 (Q6ZRP7)	CVVLYVASSLFLMVMY	91	[28]
15	MPZ (P25189)	YGVVLGAVIGGVLGVVLLLLLLFYVV	71	[31]
16	GP1BB (P13224)	GALAAQLALLGLGLLHALLL	69	[32]
17	NS4A (Q99IB8)	TWVLAGGVLAAVAAYCLAT	64	[33]
18	ITGA2B (P08514)	WVLVGVLGGLLLLTILVLAMW	47	[35]
19	ErbB2 (P04626)	LTSIISAVVGILLVVVLGVVFGIL	45	[22, 44]
20	FtsB (P0A6S5)	TLLLLAILVWLQYSLWF	43	[39]
21	ITGB3 (P05106)	VLLSVMGAILLIGLAALLI	40	[40]

[a] Accession number (acc) is taken from the UniProt database.
[b] Self-affinity (% GpA) in ToxR, TOXCAT, or GALLEX assay
[c] Scanning mutagenesis from this study. Reference declares self-affinity of wildtype TMD.
[d] Sequence for the frame with the highest self-affinity is shown

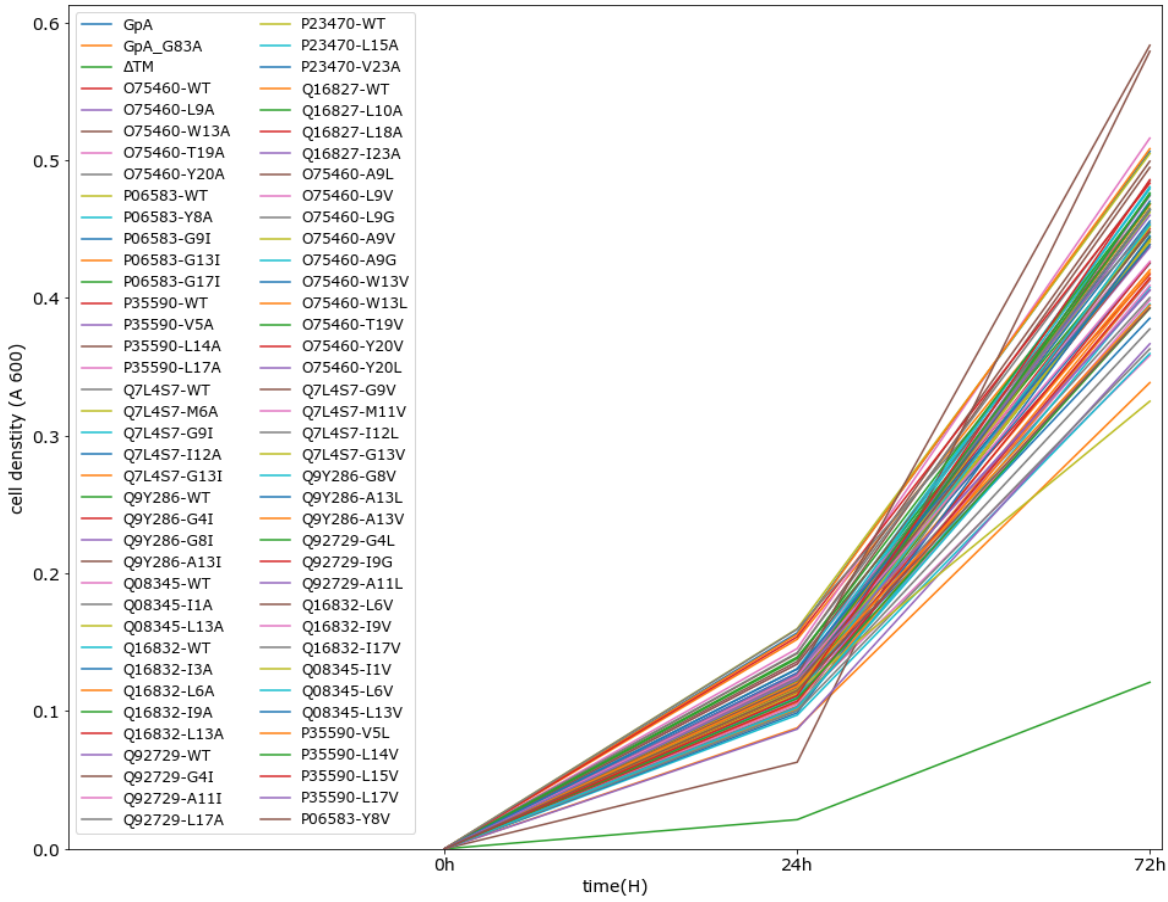

Figure 8-1: Membrane integration of the ToxR-TMD-MBP fusion proteins determined by PD28 MBP complementation assay. Plasmids with various TMDs were transformed to PD28 strains. After incubation for at least 16 h in M9 minimal medium containing 0.4% maltose as a sole carbon source, the cell density was obtained by measuring the A600 at 0h, 24h and 72h. All clones were considered to express membrane-integrated ToxR proteins correctly since the slope of their growth curves is at least 50% of GpA. The ΔTM construct should demonstrate no incorporation into the membrane.

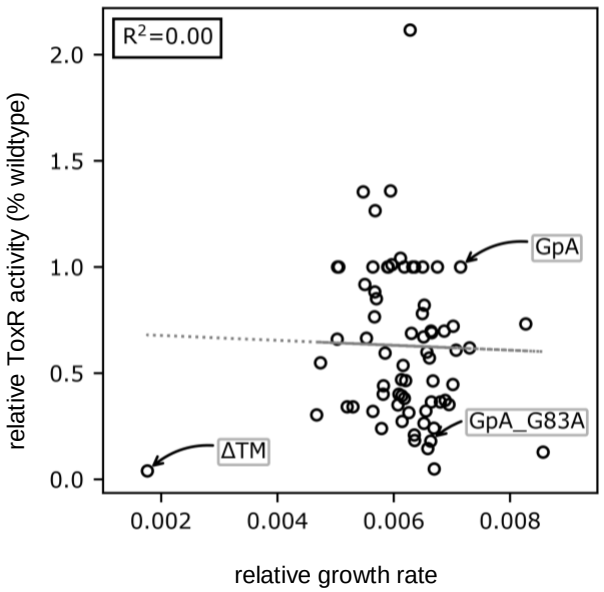

Figure 8-2: Self-affinity was confirmed not to correlate with membrane insertion. Cell density (A_{600}) for each mutant was monitored at 0 h, 24 h and 72 h. The growth rate for each sample was calculated. The measured self-affinity score for each sample was normalised to its wildtype and plotted against its PD28 grow rate, which reflects membrane insertion. The A_{600} for all mutants are present in supplementary Figure 8-1. Note that TMD affinity is not dependent on the different levels of member insertion.

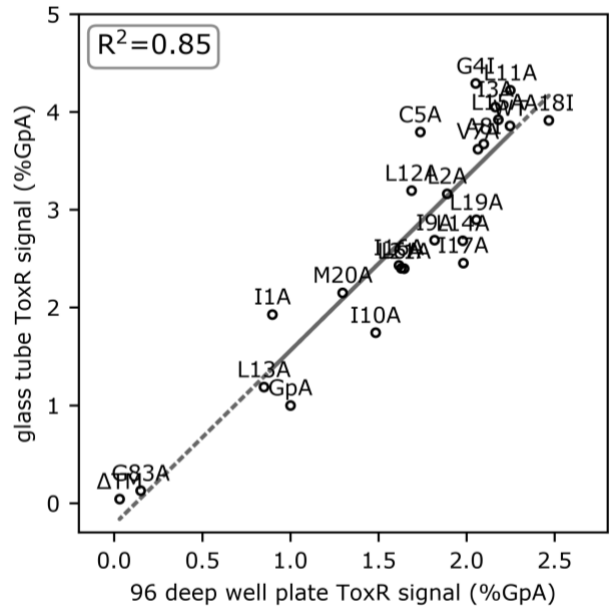

Figure 8-3: Validation of the new 96-well ToxR . 20 mutants from DDR1 TMD were tested ToxR activities by using the glass tube and the 96 deep well plates. For each system, at least three biological replicates were performed. A single assay contained three technical replicates for the glass tube assay or four technical replicates for the 96 deep well plates.

All ToxR activities were normalised to its respective GpA wildtype. GpA, G83A and ΔTM were included as controls.

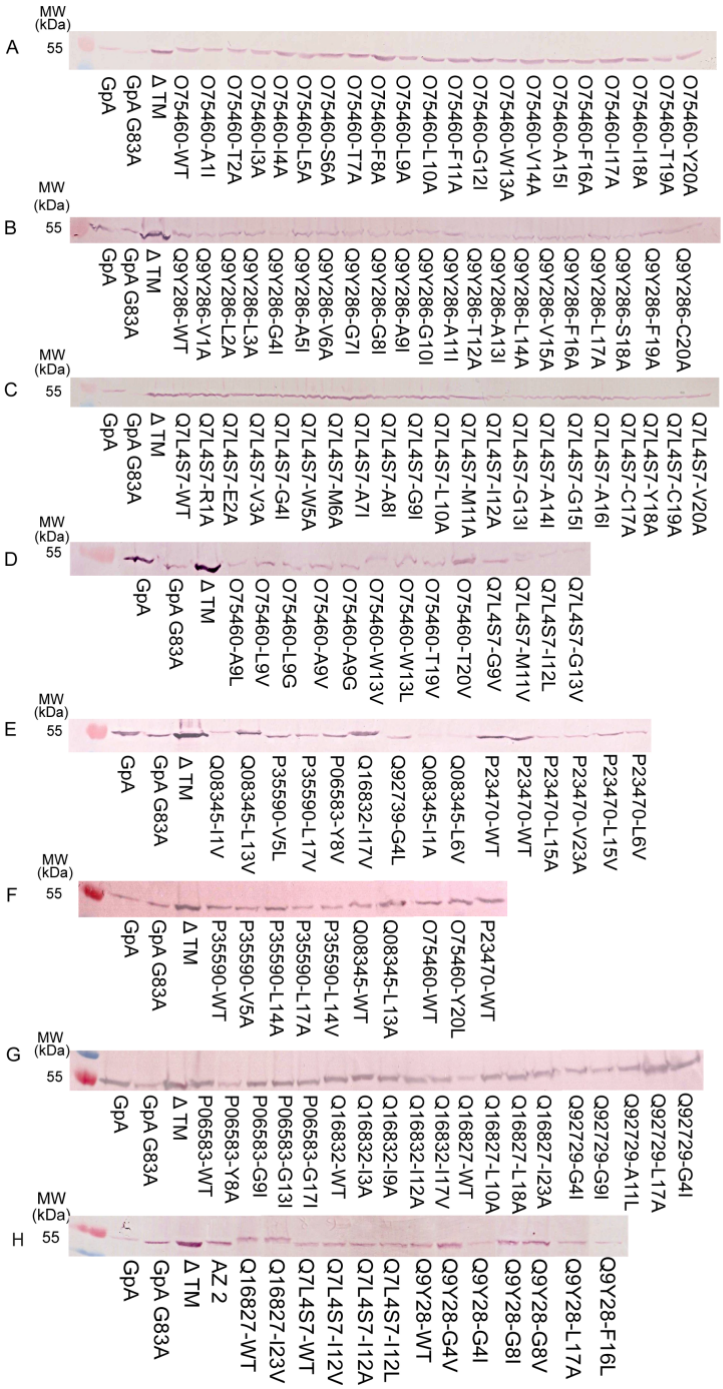

Figure 8-4: Expression level of ToxR-TMD-MBP fusion protein . In order to determine protein expression level of mutants, equal amounts of bacterial cells were separated by

SDS-PAGE and subjected to a Western blot with the primary rabbit anti-MBP antibody followed by anti-rabbit-IgG.

Table 8-4: Bootstrapped T-test for data in Figure 5-6 B, compare polarity with relative polarity.

dataset	relative polarity	polarity
TM homotypic	0.0014	0.0306
ETRA	0.0002	0.0014
NMR	0.0452	0.3810
crystal	0.3554	0.4572

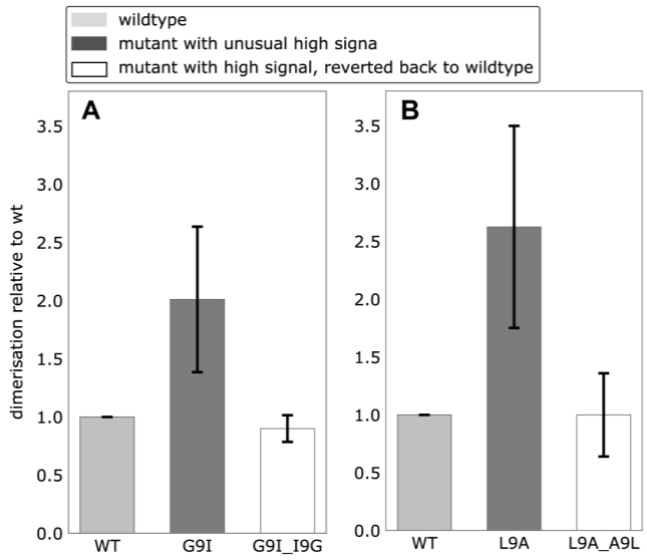

Figure 8-5: Reversion back to wildtype self-affinity confirms that mutations that dramatically increase the ToxR signal are sequence-specific. Two mutants from PTPRU and IRE1 doubled the ToxR signal in comparison to their wildtype. To confirm the increased self-affinity was not due to the unexpected mutation of plasmids, these two mutants were mutated by reverting the sequence back to the wildtype sequence. (A) G9 of PTPRU was mutated to I (G9I). Then G9I was reverted back to the wildtype sequence (G9I_I9G). The self-affinity was checked by the ToxR assay. (B) L9 from IRE1 was mutated to A (L9A). Then L9A was mutated back to the wildtype sequence (L9A_A9L). Then the ToxR assay was conducted to these mutations to test self-association. Data presented are the mean ± SEM. n > 3 independent experiments.

Table 8-5: 95% bootstrapped confidence intervals for residue features in Figure 5-6 and Figure 5-8.

dataset	int/non-int	CI upper	CI lower	*[a]
conservation				
homotypic TM	interface	1.84	1.96	
	non-interface	1.84	1.92	
ETRA	interface	2.07	2.26	*
	non-interface	1.89	2.01	
NMR	interface	1.88	2.11	
	non-interface	1.75	1.9	
crystal	interface	1.8	1.99	*
	non-interface	1.67	1.79	
coevolution				
homotypic TM	interface	0.39	0.46	
	non-interface	0.38	0.42	
ETRA	interface	0.36	0.48	
	non-interface	0.38	0.44	
NMR	interface	0.36	0.51	
	non-interface	0.38	0.49	
crystal	interface	0.41	0.51	
	non-interface	0.35	0.41	
relative polarity				
homotypic TM	interface	0.95	1.26	
	non-interface	0.99	1.14	
ETRA	interface	1.07	1.43	*
	non-interface	0.83	0.97	
NMR	interface	0.92	1.68	
	non-interface	0.76	1.02	
crystal	interface	1.04	1.49	
	non-interface	0.96	1.26	
depth in the bilayer				
homotypic TM	interface	0.51	0.57	*
	non-interface	0.44	0.48	
ETRA	interface	0.48	0.59	
	non-interface	0.44	0.51	
NMR	interface	0.46	0.6	
	non-interface	0.4	0.52	
crystal	interface	0.49	0.58	
	non-interface	0.44	0.51	

a * indicates CI upper and CI lower are non-overlapped.
Bo Zeng and Dr. Mark Teese defined NMR and crystal TMDs datasets and interfaces. Analysis and CI calculation were done by myself.

CHAPTER 9. LIST OF SYMBOLS AND ABBREVIATIONS

% (v/v)	volume percent
% (w/v)	weight percent
AA	amino acid
APS	ammonium persulfate
APP	amyloid precursor protein
ctx	cholera toxin promotor
CI	confidence interval
ddH_2O	double-distilled water
dNTP	desoxynucleotide triphosphate
dsDNA	double stranded DNA
dsTβL	deep-sequencing TOXCAT-β-lactamase
DCA	direct-coupling analysis
EDTA	ethylenediaminetetraacetic acid
E. coli	*Escherichia coli*
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 mu	meprin, A-5 protein, and receptor protein-tyrosine phosphatase
NMR	nuclear magnetic resonance
NT-3	neurotrophin-3
OD	optical density
ONP	ortho-nitrophenol
ONPG	ortho-nitrophenyl-β-D-galactopyranoside
PAGE	polyacrylamide-gel electrophoresis
PBS	phosphate buffered saline

PCR	polymerase chain reaction
PDB	protein data bank
PNK	polynucleotide kinase
PPI	protein-protein interactions
rpm	rounds per minute
RTK	receptor tyrosine kinase
TrkC	receptor tropomyosin-related kinase C
RPTPs	receptor-like protein tyrosine phosphatases
SDS	sodium dodecyl sulfate
sp	strongly polar
TBS	tris buffered saline
TBS-T	tris buffered saline with tween20
TEMED	tetramethylethylenediamine
TMD	transmembrane domain
TM	transmembrane
Tris	tris (hydroxymethyl)-aminomethane

CHAPTER 10. PUBLICATIONS ARISING FROM THIS THESIS

Yao Xiao‡, Bo Zeng‡, 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.*

REFERENCES

[1] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, A transmembrane helix dimer: structure and implications, Science 276(5309) (1997) 131-133.

[2] M.E. Call, J.R. Schnell, C. Xu, R.A. Lutz, J.J. Chou, K.W. Wucherpfennig, The structure of the ζζ transmembrane dimer reveals features essential for its assembly with the T cell receptor, Cell 127(2) (2006) 355-68.

[3] C. Finger, C. Escher, D. Schneider, The single transmembrane domains of human receptor tyrosine kinases encode self-interactions, Science Signaling 2(89) (2009) ra56-ra56.

[4] J.I. Godfroy III, M. Roostan, Y.S. Moroz, I.V. Korendovych, H. Yin, Isolated Toll-like receptor transmembrane domains are capable of oligomerization, PloS one 7(11) (2012) e48875.

[5] E.V. Bocharov, Y.E. Pustovalova, K.V. Pavlov, P.E. Volynsky, M.V. Goncharuk, Y.S. Ermolyuk, D.V. Karpunin, A.A. Schulga, M.P. Kirpichnikov, R.G. Efremov, I.V. Maslennikov, A.S. Arseniev, Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger, J. Biol. Chem. 282(22) (2007) 16256-66.

[6] E.S. Sulistijo, 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, J. Mol. Biol. 364(5) (2006) 974-990.

[7] T.L. Blundell, D. Donnelly, J.P. Overington, S.V. Ruffle, J.H. Nugent, Modeling α-helical transmembrane domains: The calculation and use of substitution tables for lipid-facing residues, Protein Sci. 2(1) (1993) 55-70.

[8] H. Kolmar, F. Hennecke, K. Götze, B. Janzer, B. Vogt, F. Mayer, H.-J. Fritz, Membrane insertion of the bacterial signal transduction protein ToxR and requirements of transcription activation studied by modular replacement of different protein substructures, The EMBO journal 14(16) (1995) 3895-3904.

[9] C.-N. Chin, J.N. Sachs, D.M. Engelman, Transmembrane homodimerization of receptor-like protein tyrosine phosphatases, FEBS Lett. 579(17) (2005) 3855-3858.

[10] H. Gratkowski, J.D. Lear, W.F. DeGrado, Polar side chains drive the association of model transmembrane peptides, Proc. Natl. Acad. Sci. 98(3) (2001) 880-885.

[11] A.S. Khadria, B.K. Mueller, J.A. Stefely, C.H. Tan, D.J. Pagliarini, A. Senes, A Gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3, J. Am. Chem. Soc. 136(40) (2014) 14068-14077.

[12] P. Duplay, S. Szmelcman, H. Bedouelle, M. Hofnung, Silent and functional changes in the periplasmic maltose-binding protein of Escherichia coli K12: I. transport of maltose, J. Mol. Biol. 194(4) (1987) 663-673.

[13] E.V. Bocharov, K.S. Mineev, P.E. Volynsky, Y.S. Ermolyuk, E.N. Tkach, A.G. Sobol, V.V. Chupin, M.P. Kirpichnikov, R.G. Efremov, A.S. Arseniev, Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state, J. Biol. Chem. 283(11) (2008) 6950-6956.

[14] E.V. Bocharov, M.L. Mayzel, P.E. Volynsky, M.V. Goncharuk, Y.S. Ermolyuk, A.A. Schulga, E.O. Artemenko, R.G. Efremov, A.S. Arseniev, Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1, J. Biol. Chem. 283(43) (2008) 29385-95.

[15] C.L. Ried, S. Kube, J. Kirrbach, D. Langosch, Homotypic interaction and amino acid distribution of unilaterally conserved transmembrane Helices, J. Mol. Biol. 3(420) (2012) 251-257.

[16] T.J. Stevens, I.T. Arkin, Substitution rates in α-helical transmembrane proteins, Protein Sci. 10(12) (2001) 2507-2517.

[17] S.-Q. Zhang, D.W. Kulp, C.A. Schramm, M. Mravic, I. Samish, W.F. DeGrado, The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions, Structure 23(3) (2015) 527-541.

[18] W.P. Russ, D.M. Engelman, The GxxxG motif: A framework for transmembrane helix-helix association, J. Mol. Biol. 296(3) (2000) 911-919.

[19] A. Senes, M. Gerstein, D.M. Engelman, Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions, J. Mol. Biol. 296(3) (2000) 921-36.

[20] M.E. Call, K.W. Wucherpfennig, J.J. Chou, The structural basis for intramembrane assembly of an activating immunoreceptor complex, Nat. Immunol. 11(11) (2010) 1023.

[21] J. Kirrbach, M. Krugliak, C.L. Ried, P. Pagel, I.T. Arkin, D. Langosch, Self-interaction of transmembrane helices representing pre-clusters from the human single-span membrane proteins, Bioinformatics 29(13) (2013) 1623-30.

[22] A. Elazar, J. Weinstein, I. Biran, Y. Fridman, E. Bibi, S.J. Fleishman, Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane, Elife 5 (2016).

[23] D. Schneider, D.M. Engelman, Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions, J. Mol. Biol. 343(4) (2004) 799-804.

[24] E.V. Bocharov, M.L. Mayzel, P.E. Volynsky, K.S. Mineev, E.N. Tkach, Y.S. Ermolyuk, A.A. Schulga, R.G. Efremov, A.S. Arseniev, Left-handed dimer of EphA2 transmembrane domain: helix packing diversity among receptor tyrosine kinases, Biophys. J. 98(5) (2010) 881-889.

[25] S.P. Barwe, S. Kim, S.A. Rajasekaran, J.U. Bowie, A.K. Rajasekaran, 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. 365(3) (2007) 706-14.

[26] J.M. Mason, K.M. Arndt, Coiled coil domains: stability, specificity, and biological implications, Chembiochem 5(2) (2004) 170-176.

[27] K. Mineev, N. Khabibullina, E. Lyukmanova, D. Dolgikh, M. Kirpichnikov, A. Arseniev, Spatial structure and dimer–monomer equilibrium of the ErbB3 transmembrane domain in DPC micelles, Biochim. Biophys. Acta Biomembr. 1808(8) (2011) 2081-2088.

[28] C.L. Ried, C. Scharnagl, D. Langosch, Entrapment of water at the transmembrane helix-helix interface of quiescin sulfhydryl oxidase 2, Biochemistry 55(9) (2016) 1287-90.

[29] S. Kim, T.-J. Jeon, A. Oberai, D. Yang, J.J. Schmidt, J.U. Bowie, Transmembrane glycine zippers: physiological and pathological roles in membrane proteins, Proc. Natl. Acad. Sci. USA 102(40) (2005) 14278-14283.

[30] K.D. Nadezhdin, O.V. Bocharova, E.V. Bocharov, A.S. Arseniev, Dimeric structure of transmembrane domain of amyloid precursor protein in micellar environment, FEBS Lett. 586(12) (2012) 1687-1692.

[31] M.L. Plotkowski, S. Kim, M.L. Phillips, A.W. Partridge, C.M. Deber, J.U. Bowie, Transmembrane domain of myelin protein zero can form dimers: possible implications for myelin construction, Biochemistry 46(43) (2007) 12164-12173.

[32] P. Wei, X. Liu, M.H. Hu, L.M. Zuo, M. Kai, R. Wang, S.Z. Luo, The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif, Protein Sci. 20(11) (2011) 1814-23.

[33] A. Kohlway, N. Pirakitikulr, F.N. Barrera, O. Potapova, D.M. Engelman, A.M. Pyle, B.D. Lindenbach, Hepatitis C virus RNA replication and virus particle assembly require specific dimerization of the NS4A protein transmembrane domain, J. Virol. 88(1) (2014) 628-42.

[34] N. Sal-Man, D. Gerber, I. Bloch, Y. Shai, Specificity in transmembrane helix-helix interactions mediated by aromatic residues, J. Biol. Chem. 282(27) (2007) 19753-19761.

[35] R. Li, R. Gorelik, V. Nanda, P.B. Law, J.D. Lear, W.F. DeGrado, J.S. Bennett, Dimerization of the transmembrane domain of Integrin alphaIIb subunit in cell membranes, J. Biol. Chem. 279(25) (2004) 26666-73.

[36] Y. Wang, P. Barth, Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy, Nat. Comms. 6 (2015) 7196.

[37] C. Muhle-Goll, S. Hoffmann, S. Afonin, S.L. Grage, A.A. Polyansky, D. Windisch, M. Zeitler, J. Bürck, A.S. Ulrich, Hydrophobic matching controls the tilt and stability of the dimeric platelet-derived growth factor receptor (PDGFR) β transmembrane segment, J. Biol. Chem. 287(31) (2012) 26178-26186.

[38] E.V. Bocharov, K.S. Mineev, M.V. Goncharuk, A.S. Arseniev, Structural and thermodynamic insight into the process of "weak" dimerization of the ErbB4 transmembrane domain by solution NMR, Biochim. Biophys. Acta Biomembr. 1818(9) (2012) 2158-2170.

[39] L.M. LaPointe, K.C. Taylor, S. Subramaniam, A. Khadria, I. Rayment, A. Senes, Structural organization of FtsB, a transmembrane protein of the bacterial divisome, Biochemistry 52(15) (2013) 2574-2585.

[40] H. Zhu, D.G. Metcalf, C.N. Streu, P.C. Billings, W.F. Degrado, J.S. Bennett, Specificity for homooligomer versus heterooligomer formation in integrin transmembrane helices, J. Mol. Biol. 401(5) (2010) 882-91.

[41] N.F. Endres, R. Das, A.W. Smith, A. Arkhipov, E. Kovacs, Y. Huang, J.G. Pelton, Y. Shan, D.E. Shaw, D.E. Wemmer, Conformational coupling across the plasma membrane in activation of the EGF receptor, Cell 152(3) (2013) 543-556.

[42] E.V. Bocharov, D.M. Lesovoy, S.A. Goncharuk, M.V. Goncharuk, K. Hristova, A.S. Arseniev, Structure of FGFR3 transmembrane domain dimer: implications for signaling and human pathologies, Structure 21(11) (2013) 2087-2093.

[43] K.S. Mineev, S.A. Goncharuk, A.S. Arseniev, Toll-like receptor 3 transmembrane domain is able to perform various homotypic interactions: An NMR structural study, FEBS Lett. 588(21) (2014) 3802-3807.

[44] D. Gerber, N. Sal-Man, Y. Shai, Two motifs within a transmembrane domain, one for homodimerization and the other for heterodimerization, J. Biol. Chem. 279(20) (2004) 21177-21182.

[45] A. Krogh, B. Larsson, G. Von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1, J. Mol. Biol. 305(3) (2001) 567-580.

[46] J. Liu, B. Rost, Comparing function and structure between entire proteomes, Protein Sci. 10(10) (2001) 1970-1979.

[47] P. Hubert, P. Sawma, J.-P. Duneau, J. Khao, J. Hénin, D. Bagnard, J. Sturgis, Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye?, Cell Adhes. Migr. 4(2) (2010) 313-324.

[48] L. Käll, A. Krogh, E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, J. Mol. Biol. 338(5) (2004) 1027-1036.

[49] A. Oberai, N.H. Joh, F.K. Pettit, J.U. Bowie, Structural imperatives impose diverse evolutionary constraints on helical membrane proteins, Proc. Natl. Acad. Sci. 106(42) (2009) 17747-17750.

[50] T.S. Ulmer, B. Yaspan, M.H. Ginsberg, I.D. Campbell, NMR analysis of structure and dynamics of the cytosolic tails of integrin αIIbβ3 in aqueous solution, Biochemistry 40(25) (2001) 7498-7508.

[51] R. Li, C.R. Babu, K. Valentine, J.D. Lear, A.J. Wand, J.S. Bennett, W.F. DeGrado, Characterization of the monomeric form of the transmembrane and cytoplasmic domains of the integrin β3 subunit by NMR spectroscopy, Biochemistry 41(52) (2002) 15618-15624.

[52] A. Fink, N. Sal-Man, D. Gerber, Y. Shai, Transmembrane domains interactions within the membrane milieu: principles, advances and challenges, Biochim. Biophys. Acta 1818(4) (2012) 974-83.

[53] L. Fagerberg, K. Jonasson, G. von Heijne, M. Uhlén, L. Berglund, Prediction of the human membrane proteome, Proteomics 10(6) (2010) 1141-1149.

[54] P.N. Reardon, H. Sage, S.M. Dennison, J.W. Martin, B.R. Donald, S.M. Alam, B.F. Haynes, L.D. Spicer, Structure of an HIV-1–neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer, Proc. Natl. Acad. Sci. 111(4) (2014) 1391-1396.

[55] D.T. Moore, B.W. Berger, W.F. DeGrado, Protein-protein interactions in the membrane: sequence, structural, and biological motifs, Structure 16(7) (2008) 991-1001.

[56] E. Li, K. Hristova, Receptor tyrosine kinase transmembrane domains: Function, dimer structure and dimerization energetics, Cell Adhes. Migr. 4(2) (2010) 249-254.

[57] C. Scott, S. Griffin, Viroporins: structure, function and potential as antiviral targets, J. Gen. Virol. 96(8) (2015) 2000-2027.

[58] E.V. Bocharov, P.E. Volynsky, K.V. Pavlov, R.G. Efremov, A.S. Arseniev, Structure elucidation of dimeric transmembrane domains of bitopic proteins, Cell Adhes. Migr. 4(2) (2010) 284-298.

[59] T. Moriki, H. Maruyama, I.N. Maruyama, Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain1, J. Mol. Biol. 311(5) (2001) 1011-1026.

[60] S.J. Fleishman, J. Schlessinger, N. Ben-Tal, A putative molecular-activation switch in the transmembrane domain of erbB2, Proc. Natl. Acad. Sci. 99(25) (2002) 15937-15940.

[61] M. Vilar, I. Charalampopoulos, R.S. Kenchappa, A. Simi, E. Karaca, A. Reversi, S. Choi, M. Bothwell, I. Mingarro, W.J. Friedman, Activation of the p75 neurotrophin receptor through conformational rearrangement of disulphide-linked receptor dimers, Neuron 62(1) (2009) 72-83.

[62] M.I. Tao RH, All EGF(ErbB) receptors have preformed homo- and heterodimeric structures in living cells., J. Cell Sci. 121(19) (2008) 3207-3217.

[63] G.V. Sharonov, E.V. Bocharov, P.M. Kolosov, M.V. Astapova, A.S. Arseniev, A.V. Feofanov, Point mutations in dimerization motifs of the transmembrane domain stabilize active or inactive state of the EphA2 receptor tyrosine kinase, J. Biol. Chem. 289(21) (2014) 14955-14964.

[64] L.-X. Shi, W.P. Schröder, The low molecular mass subunits of the photosynthetic supracomplex, photosystem II, Biochim. Biophys. Acta. 1608(2-3) (2004) 75-96.

[65] E. Winkler, A. Julius, H. Steiner, D. Langosch, Homodimerization protects the amyloid precursor protein C99 fragment from cleavage by γ-secretase, Biochemistry 54(40) (2015) 6149-6152.

[66] D.V. Tulumello, C.M. Deber, SDS micelles as a membrane-mimetic environment for transmembrane segments, Biochemistry 48(51) (2009) 12096-12103.

[67] M.A. Lemmon, J.M. Flanagan, H.R. Treutlein, J. Zhang, D.M. Engelman, Sequence specificity in the dimerization of transmembrane. alpha.-helixes, Biochemistry 31(51) (1992) 12719-12725.

[68] E.S. Sulistijo, T.M. Jaszewski, K.R. MacKenzie, Sequence-specific dimerization of the transmembrane domain of the “BH3-only” protein BNIP3 in membranes and detergent, J. Biol. Chem. 278(51) (2003) 51950-51956.

[69] A.M. Stanley, K.G. Fleming, The transmembrane domains of ErbB receptors do not dimerize strongly in micelles, J. Mol. Biol. 347(4) (2005) 759-772.

[70] D. Langosch, B. Brosig, H. Kolmar, H.J. Fritz, Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator, J. Mol. Biol. 263(4) (1996) 525-30.

[71] W.P. Russ, D.M. Engelman, TOXCAT: A measure of transmembrane helix association in a biological membrane, Proc. Natl. Acad. Sci. USA 96(3) (1999) 863.

[72] D. Schneider, D.M. Engelman, GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane, J. Biol. Chem. 278(5) (2003) 3105-3111.

[73] P. Sawma, L. Roth, C. Blanchard, D. Bagnard, G. Cremel, E. Bouveret, J.P. Duneau, J.N. Sturgis, P. Hubert, Evidence for new homotypic and heterotypic interactions between transmembrane helices of proteins involved in receptor tyrosine kinase and neuropilin signaling, J. Mol. Biol. 426(24) (2014) 4099-111.

[74] D. Steindorf, D. Schneider, In vivo selection of heterotypically interacting transmembrane helices: Complementary helix surfaces, rather than conserved interaction motifs, drive formation of transmembrane hetero-dimers, Biochim. Biophys. Acta Biomembr. 1859(2) (2017) 245-256.

[75] P.-C. Su, B.W. Berger, Identifying key juxtamembrane interactions in cell membranes using AraC-based transcriptional reporter assay (AraTM), J. Biol. Chem. 287(37) (2012) 31515-31526.

[76] C. Schanzenbach, F.C. Schmidt, P. Breckner, M.G. Teese, D. Langosch, Identifying ionic interactions within a membrane using BLaTM, a genetic tool to measure homo-and heterotypic transmembrane helix-helix interactions, Scientific reports 7 (2017) 43476.

[77] M.A. Lemmon, J.M. Flanagan, J.F. Hunt, B.D. Adair, B.-J. Bormann, C.E. Dempsey, D.M. Engelman, Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices, J. Biol. Chem. 267(11) (1992) 7683-7689.

[78] C.M. Lawrie, E.S. Sulistijo, 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. 396(4) (2010) 924-36.

[79] E.S. Sulistijo, K.R. MacKenzie, Structural basis for dimerization of the BNIP3 transmembrane domain, Biochemistry 48(23) (2009) 5106-5120.

[80] H. Hong, J.U. Bowie, Dramatic destabilization of transmembrane helix interactions by features of natural membrane environments, J. Am. Chem. Soc. 133(29) (2011) 11389-11398.

[81] K.S. Mineev, E.V. Bocharov, Y.E. Pustovalova, O.V. Bocharova, V.V. Chupin, A.S. Arseniev, Spatial structure of the transmembrane domain heterodimer of ErbB1 and ErbB2 receptor tyrosine kinases, J. Mol. Biol. 400(2) (2010) 231-243.

[82] V.L. Miller, R.K. Taylor, J.J. Mekalanos, Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein, Cell 48(2) (1987) 271-279.

[83] J. Kirrbach, Mapping the uuman single-span membrane proteome for self-interacting transmembrane domains, Technische Universität München, Thesis (04.10.2012).

[84] K. Bugge, K. Lindorff-Larsen, B.B. Kragelund, Understanding single-pass transmembrane receptor signaling from a structural viewpoint—what are we missing?, The FEBS journal 283(24) (2016) 4424-4451.

[85] C.C. Valley, A.K. Lewis, J.N. Sachs, Piecing it together: Unraveling the elusive structure-function relationship in single-pass membrane receptors, Biochim. Biophys. Acta Biomembr. 1859(9) (2017) 1398-1416.

[86] A.A. Polyansky, A.O. Chugunov, P.E. Volynsky, N.A. Krylov, D.E. Nolde, R.G. Efremov, PREDDIMER: a web server for prediction of transmembrane helical dimers, Bioinformatics 30(6) (2013) 889-890.

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

[88] G.E. Tusnády, I. Simon, Topology of membrane proteins, J. Chem. Inform. Comput. Sci. 41(2) (2001) 364-368.

[89] I. Mus-Veteau, Membrane proteins production for structural analysis, Springer2014.

[90] D. Kozma, I. Simon, G.E. Tusnady, PDBTM: Protein Data Bank of transmembrane proteins after 8 years, Nucleic Acids Res. 41(D1) (2012) D524-D529.

[91] M. Bernhofer, E. Kloppmann, J. Reeb, B. Rost, TMSEG: Novel prediction of transmembrane helices, Proteins 84(11) (2016) 1706-1716.

[92] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, Nucleic Acids Res. 33(suppl_1) (2005) D275-D278.

[93] B. Kobe, G. Guncar, R. Buchholz, T. Huber, B. Maco, N. Cowieson, J.L. Martin, M. Marfori, J.K. Forwood, Crystallography and protein–protein interactions: biological interfaces and crystal contacts, Portland Press Limited, 2008.

[94] J. Janin, F. Rodier, Protein–protein interaction at crystal contacts, Proteins 23(4) (1995) 580-587.

[95] H. Ponstingl, K. Henrick, J.M. Thornton, Discriminating between homodimeric and monomeric proteins in the crystalline state, Proteins 41(1) (2000) 47-57.

[96] R.P. Bahadur, P. Chakrabarti, F. Rodier, J. Janin, A dissection of specific and non-specific protein–protein interfaces, J. Mol. Biol. 336(4) (2004) 943-955.

[97] M.G. Teese, D. Langosch, Role of GxxxG motifs in transmembrane domain interactions, Biochemistry 54(33) (2015) 5125-5135.

[98] M. Weber, D. Schneider, Six amino acids define a minimal dimerization sequence and stabilize a transmembrane helix dimer by close packing and hydrogen bonding, FEBS Lett. 587(11) (2013) 1592-6.

[99] D. Schneider, D.M. Engelman, Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions, J Mol Biol 343(4) (2004) 799-804.

[100] J.M. Mendrola, M.B. Berger, M.C. King, M.A. Lemmon, The single transmembrane domains of ErbB receptors self-associate in cell membranes, J. Biol. Chem. 277(7) (2002) 4704-4712.

[101] G. Arselin, M.F. Giraud, A. Dautant, J. Vaillier, D. Brethes, B. Coulary-Salin, J. Schaeffer, J. Velours, The GxxxG motif of the transmembrane domain of subunit e is involved in the dimerization/oligomerization of the yeast ATP synthase complex in the mitochondrial membrane, The FEBS Journal 270(8) (2003) 1875-1884.

[102] M.S. McClain, H. Iwamoto, P. Cao, A.D. Vinion-Dubiel, Y. Li, G. Szabo, Z. Shao, T.L. Cover, Essential role of a GXXXG motif for membrane channel formation by Helicobacter pylori vacuolating toxin, J. Biol. Chem. 278(14) (2003) 12101-12108.

[103] P. Cosson, J.S. Bonifacino, Role of transmembrane domain interactions in the assembly of class II MHC molecules, Science 258(5082) (1992) 659-662.

[104] G. King, A.M. Dixon, Evidence for role of transmembrane helix–helix interactions in the assembly of the Class II major histocompatibility complex, Mol. BioSys 6(9) (2010) 1650-1661.

[105] S.M. Anderson, B.K. Mueller, E.J. Lange, A. Senes, Combination of Cα–H hydrogen bonds and van der Waals packing modulates the stability of GxxxG-mediated dimers in membranes, J. Am. Chem. Soc. 139(44) (2017) 15774-15783.

[106] S. He, Y. Liang, F. Shao, X. Wang, Toll-like receptors activate programmed necrosis in macrophages through a receptor-interacting kinase-3–mediated pathway, Proc. Natl. Acad. Sci. 108(50) (2011) 20054-20059.

[107] B.K. Mueller, S. Subramaniam, A. Senes, A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα–H hydrogen bonds, Proc. Natl. Acad. Sci. 111(10) (2014) E888-E895.

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

[109] J. Wang, J.X. Qiu, C. Soto, W.F. DeGrado, Structural and dynamic mechanisms for the function and inhibition of the M2 proton channel from influenza A virus, Curr. Opin. Struct. Biol. 21(1) (2011) 68-80.

[110] M.T. De Marothy, A. Elofsson, Marginally hydrophobic transmembrane α-helices shaping membrane protein folding, Protein Sci. 24(7) (2015) 1057-1074.

[111] N.H. Joh, A. Oberai, D. Yang, J.P. Whitelegge, J.U. Bowie, Similar energetic contributions of packing in the core of membrane and water-soluble proteins, J. Am. Chem. Soc. 131(31) (2009) 10846-10847.

[112] D. Walther, F. Eisenhaber, P. Argos, Principles of helix-helix packing in proteins: the helical lattice superposition model, J. Mol. Biol. 255(3) (1996) 536-553.

[113] D. Langosch, J. Heringa, Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils, Proteins 31(2) (1998) 150-159.

[114] S.O. Smith, D. Song, S. Shekar, M. Groesbeek, M. Ziliox, S. Aimoto, Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers, Biochemistry 40(22) (2001) 6553-6558.

[115] A. Senes, I. Ubarretxena-Belandia, D.M. Engelman, The Cα—H··· O hydrogen bond: A determinant of stability and specificity in transmembrane helix interactions, Proc. Natl. Acad. Sci. 98(16) (2001) 9056-9061.

[116] F.X. Zhou, H.J. Merianos, A.T. Brunger, D.M. Engelman, Polar residues drive association of polyleucine transmembrane helices, Proc. Natl. Acad. Sci. 98(5) (2001) 2250-2255.

[117] E.V. Bocharov, K.V. Pavlov, P.E. Volynsky, R.G. Efremov, A.S. Arseniev, Structure-functional insight into transmembrane helix dimerization, Protein Eng., InTech2012.

[118] K. Illergård, A. Kauko, A. Elofsson, Why are polar residues within the membrane core evolutionary conserved?, Proteins 79(1) (2011) 79-91.

[119] R. Gurezka, R. Laage, B. Brosig, D. Langosch, A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments, J. Biol. Chem. 274(14) (1999) 9265-9270.

[120] E.K. O'Shea, J.D. Klemm, P.S. Kim, T. Alber, X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil, Science 254(5031) (1991) 539-544.

[121] E. Li, W.C. Wimley, K. Hristova, Transmembrane helix dimerization: beyond the search for sequence motifs, Biochim. Biophys. Acta Biomembr. 1818(2) (2012) 183-193.

[122] D. Langosch, I.T. Arkin, Interaction and conformational dynamics of membrane-spanning protein helices, Protein Sci. 18(7) (2009) 1343-1358.

[123] R.M. Johnson, K. Hecht, C.M. Deber, Aromatic and cation− π interactions enhance helix− helix association in a membrane environment, Biochemistry 46(32) (2007) 9208-9214.

[124] L. Adamian, J. Liang, Prediction of transmembrane helix orientation in polytopic membrane proteins, BMC Struct. Biol. 6 (2006) 13.

[125] P. Hönigschmid, D. Frishman, Accurate prediction of helix interactions and residue contacts in membrane proteins, J. Struct. Biol. 194(1) (2016) 112-123.

[126] G. Von Heijne, Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225(2) (1992) 487-494.

[127] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305(3) (2001) 567-80.

[128] P. Barth, J. Schonbrun, D. Baker, Toward high-resolution prediction and design of transmembrane helical protein structures, Proc. Natl. Acad. Sci. 104(40) (2007) 15682-15687.

[129] S.Q. Zhang, D.W. Kulp, C.A. Schramm, M. Mravic, I. Samish, W.F. DeGrado, The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions, Structure 23(3) (2015) 527-41.

[130] Y. Park, S. Hayat, V. Helms, Prediction of the burial status of transmembrane residues of helical membrane proteins, BMC Bioinformatics 8(1) (2007) 302.

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

[132] L.C. Xue, D. Dobbs, A.M. Bonvin, V. Honavar, Computational prediction of protein interfaces: A review of data driven methods, FEBS Lett. 589(23) (2015) 3516-26.

[133] D.R. Caffrey, S. Somaroo, J.D. Hughes, J. Mintseris, E.S. Huang, Are protein–protein interfaces more conserved in sequence than the rest of the protein surface?, Protein Sci. 13(1) (2004) 190-202.

[134] A. Avila-Herrera, 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 16(1) (2015) 268.

[135] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon, Nature 433(7024) (2005) 377-81.

[136] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157(1) (1982) 105-132.

[137] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, Nat. Struct. Mol. Biol. 3(10) (1996) 842-848.

[138] D. Engelman, T. Steitz, A. Goldman, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, Annual review of biophysics and biophysical chemistry 15(1) (1986) 321-353.

[139] D. De Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution, Nat. Rev. Genet. 14(4) (2013) 249.

[140] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. 108(49) (2011) E1293-E1301.

[141] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing, Proc. Natl. Acad. Sci. 106(1) (2009) 67-72.

[142] B. Lunt, H. Szurmant, A. Procaccini, J.A. Hoch, T. Hwa, M. Weigt, Inference of direct residue contacts in two-component signaling, Methods Enzymol., Elsevier2010, pp. 17-41.

[143] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation, PloS one 6(12) (2011) e28766.

[144] L. Kaján, T.A. Hopf, M. Kalaš, D.S. Marks, B. Rost, FreeContact: fast and free software for protein contact prediction from residue co-evolution, BMC Bioinformatics 15(1) (2014) 85.

[145] F.M. Codoñer, M.A. Fares, Why should we care about molecular coevolution?, Evolutionary Bioinformatics 4 (2008) 29-38.

[146] H. Inoue, H. Nojima, H. Okayama, High efficiency transformation of *Escherichia coli* with plasmids, Gene 96(1) (1990) 23-28.

[147] C. Chung, S.L. Niemela, R.H. Miller, One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution, Proc. Natl. Acad. Sci. 86(7) (1989) 2172-2175.

[148] E. Lindner, Identifikation heterotypischer TMD-TMD Interaktionen., Technische Universität München (2006).

[149] S.K. Lee, H.H. Chou, B.F. Pfleger, J.D. Newman, Y. Yoshikuni, J.D. Keasling, Directed evolution of AraC for improved compatibility of arabinose-and lactose-inducible promoters, Appl. Environ. Microbiol. 73(18) (2007) 5711-5715.

[150] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22(13) (2006) 1658-1659.

[151] H. Watson, Biological membranes, Essays Biochem. 59 (2015) 43-69.

[152] S. Unterreitmeier, A. Fuchs, T. Schäffler, R.G. Heym, D. Frishman, D. Langosch, Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs, J. Mol. Biol. 374(3) (2007) 705-718.

[153] M. Guharoy, P. Chakrabarti, Conserved residue clusters at protein-protein interfaces and their use in binding site identification, BMC Bioinformatics 11(1) (2010) 286.

[154] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14(6) (2004) 1188-1190.

[155] A.L. Lomize, H.I. Mosberg, Thermodynamic model of secondary structure for α-helical peptides and proteins, Biopolymers 42(2) (1997) 239-269.

[156] A.L. Lomize, I.D. Pogozheva, H.I. Mosberg, Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes, Journal of chemical information and modeling 51(4) (2011) 930-946.

[157] B. Efron, Bootstrap methods: another look at the jackknife, Breakthroughs in statistics, Springer1992, pp. 569-593.

[158] C.L. Ried, C. Scharnagl, D. Langosch, Entrapment of Water at the Transmembrane Helix–Helix Interface of Quiescin Sulfhydryl Oxidase 2, Biochemistry 55(9) (2016) 1287-1290.

[159] S.P. Barwe, A. Skay, R. McSpadden, T.P. Huynh, S.A. Langhans, L.J. Inge, A.K. Rajasekaran, Na,K-ATPase β-subunit cis homo-oligomerization is necessary for epithelial lumen formation in mammalian cells, J. Cell Sci. 125(Pt 23) (2012) 5711-20.

[160] J. Schlessinger, Cell signaling by receptor tyrosine kinases, Cell 103(2) (2000) 211-225.

[161] A. Gschwind, O.M. Fischer, A. Ullrich, The discovery of receptor tyrosine kinases: targets for cancer therapy, Nat. Rev. Cancer 4(5) (2004) 361.

[162] J. Schlessinger, A. Ullrich, Growth factor signaling by receptor tyrosine kinases, Neuron 9(3) (1992) 383-391.

[163] I.N. Maruyama, Mechanisms of activation of receptor tyrosine kinases: monomers or dimers, Cells 3(2) (2014) 304-330.

[164] D.M. Freed, D. Alvarado, M.A. Lemmon, Ligand regulation of a constitutively dimeric EGF receptor, Nat. Comms. 6 (2015) 7380.

[165] T.C. Seegar, B. Eller, D. Tzvetkova-Robev, M.V. Kolev, S.C. Henderson, D.B. Nikolov, W.A. Barton, Tie1-Tie2 interactions mediate functional differences between angiopoietin ligands, Mol. Cell 37(5) (2010) 643-55.

[166] C.D. Kontos, E.H. Cha, J.D. York, K.G. Peters, The Endothelial Receptor Tyrosine Kinase Tie1 Activates Phosphatidylinositol 3-Kinase and Akt To Inhibit Apoptosis, Mol. Cell. Biol. 22(6) (2002) 1704-1713.

[167] A. Tsiamis, P. Hayes, H. Box, A. Goodall, P. Bell, N. Brindle, Characterization and regulation of the receptor tyrosine kinase Tie-1 in platelets, J. Vasc. Res. 37(6) (2000) 437.

[168] M.B. Marron, H. Singh, T.A. Tahir, J. Kavumkal, H.-Z. Kim, G.Y. Koh, N.P. Brindle, Regulated proteolytic processing of Tie1 modulates ligand responsiveness of the receptor-tyrosine kinase Tie2, J. Biol. Chem. 282(42) (2007) 30509-30517.

[169] T. Avril, E.R. Wagner, H.J. Willison, P.R. Crocker, Sialic acid-binding immunoglobulin-like lectin 7 mediates selective recognition of sialylated glycans expressed on *Campylobacter jejuni* lipooligosaccharides, Infect. Immun. 74(7) (2006) 4133-4141.

[170] H. Attrill, H. Takazawa, S. Witt, S. Kelm, R. Isecke, R. Brossmer, T. Ando, H. Ishida, M. Kiso, P.R. Crocker, The structure of siglec-7 in complex with sialosides: leads for rational structure-based inhibitor design, Biochem. J. 397(2) (2006) 271-278.

[171] G. Nicoll, T. Avril, K. Lock, K. Furukawa, N. Bovin, P.R. Crocker, Ganglioside GD3 expression on target cells can modulate NK cell cytotoxicity via siglec-7-dependent and-independent mechanisms, Eur. J. Immunol. 33(6) (2003) 1642-1648.

[172] A.S. Allegretti, G. Ortiz, S. Kalim, J. Wibecan, D. Zhang, H.Y. Shan, D. Xu, R.T. Chung, S.A. Karumanchi, R.I. Thadhani, Siglec-7 as a novel biomarker to predict mortality in decompensated cirrhosis and acute kidney injury, Dig. Dis. Sci. 61(12) (2016) 3609-3620.

[173] R. Biassoni, C. Cantoni, D. Pende, S. Sivori, S. Parolini, M. Vitale, C. Bottino, A. Moretta, Human natural killer cell receptors and co-receptors, Immunol. Rev. 181(1) (2001) 203-214.

[174] S. Siddiqui, F. Schwarz, S. Springer, Z. Khedri, H. Yu, L. Deng, A. Verhagen, Y. Naito-Matsui, W. Jiang, D. Kim, Studies on the detection, expression, glycosylation, dimerization and ligand binding properties of mouse Siglec-E, J. Biol. Chem. (2016) jbc. M116. 738351.

[175] M.T. Virkki, C. Peters, D. Nilsson, T. Sorensen, S. Cristobal, B. Wallner, A. Elofsson, The positive inside rule is stronger when followed by a transmembrane helix, J. Mol. Biol. 426(16) (2014) 2982-91.

[176] A.J. Barr, E. Ugochukwu, W.H. Lee, O.N. King, P. Filippakopoulos, I. Alfano, P. Savitsky, N.A. Burgess-Brown, S. Muller, S. Knapp, Large-scale structural analysis of the classical human protein tyrosine phosphatome, Cell 136(2) (2009) 352-63.

[177] S.E. Ensslen-Craig, S.M. Brady-Kalnay, Receptor protein tyrosine phosphatases regulate neural development and axon guidance, Dev. Biol. 275(1) (2004) 12-22.

[178] A.K.L. Cheung, J.C.Y. Ip, A.C.H. Chu, Y. Cheng, M.M.L. Leong, J.M.Y. Ko, W.H. Shuen, H.L. Lung, M.L. Lung, PTPRG suppresses tumor growth and invasion via inhibition of Akt signaling in nasopharyngeal carcinoma, Oncotarget 6(15) (2015) 13434.

[179] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, T. Mustelin, Protein tyrosine phosphatases in the human genome, Cell 117(6) (2004) 699-711.

[180] G. Jiang, J. den Hertog, J. Su, J. Noel, J. Sap, T. Hunter, Dimerization inhibits the activity of receptor-like protein-tyrosine phosphatase-α, Nature 401(6753) (1999) 606.

[181] S.T. Shu, Y. Sugimoto, S. Liu, H.-L. Chang, W. Ye, L.-S. Wang, Y.-W. Huang, P. Yan, Y.C. Lin, Function and regulatory mechanisms of the candidate tumor suppressor receptor protein tyrosine phosphatase gamma (PTPRG) in breast cancer cells, Anticancer Res. 30(6) (2010) 1937-1946.

[182] B. Chen, J.L. Bixby, A novel substrate of receptor tyrosine phosphatase PTPRO is required for nerve growth factor-induced process outgrowth, J. Neurosci. 25(4) (2005) 880-888.

[183] A.E. Hower, P.J. Beltran, J.L. Bixby, Dimerization of tyrosine phosphatase PTPRO decreases its activity and ability to inactivate TrkC, J. Neurochem. 110(5) (2009) 1635-1647.

[184] W. Tirasophon, A.A. Welihinda, R.J. Kaufman, A stress response pathway from the endoplasmic reticulum to the nucleus requires a novel bifunctional protein kinase/endoribonuclease (Ire1p) in mammalian cells, Genes Dev. 12(12) (1998) 1812-1824.

[185] A.V. Korennykh, P.F. Egea, A.A. Korostelev, J. Finer-Moore, C. Zhang, K.M. Shokat, R.M. Stroud, P. Walter, The unfolded protein response signals through high-order assembly of Ire1, Nature 457(7230) (2009) 687.

[186] C.Y. Liu, M. Schröder, R.J. Kaufman, Ligand-independent dimerization activates the stress response kinases IRE1 and PERK in the lumen of the endoplasmic reticulum, J. Biol. Chem. 275(32) (2000) 24881-24885.

[187] H. Yoshida, T. Matsui, A. Yamamoto, T. Okada, K. Mori, XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor, Cell 107(7) (2001) 881-891.

[188] S. Casas-Tinto, Y. Zhang, J. Sanchez-Garcia, M. Gomez-Velazquez, D.E. Rincon-Limas, P. Fernandez-Funez, The ER stress factor XBP1s prevents amyloid-β neurotoxicity, Hum. Mol. Genet. 20(11) (2011) 2144-2160.

[189] C. Duran-Aniotz, V.H. Cornejo, S. Espinoza, Á.O. Ardiles, D.B. Medinas, C. Salazar, A. Foley, I. Gajardo, P. Thielen, T. Iwawaki, IRE1 signaling exacerbates Alzheimer's disease pathogenesis, Acta Neuropathol. 134(3) (2017) 489-506.

[190] B. Leitinger, Molecular analysis of collagen binding by the human discoidin domain receptors, DDR1 and DDR2 identification of collagen binding sites in DDR2, J. Biol. Chem. 278(19) (2003) 16761-16769.

[191] N.A. Noordeen, F. Carafoli, E. Hohenester, M.A. Horton, B. Leitinger, A transmembrane leucine zipper is required for activation of the dimeric receptor tyrosine kinase DDR1, J. Biol. Chem. 281(32) (2006) 22744-22751.

[192] W.F. Vogel, R. Abdulhussein, C.E. Ford, Sensing extracellular matrix: an update on discoidin domain receptor function, Cell. Signal. 18(8) (2006) 1108-1116.

[193] A.D. Konitsiotis, N. Raynal, D. Bihan, E. Hohenester, R.W. Farndale, B. Leitinger, Characterization of high affinity binding motifs for the discoidin domain receptor DDR2 in collagen, J. Biol. Chem. 283(11) (2008) 6861-6868.

[194] B. Leitinger, Transmembrane collagen receptors, Annu. Rev. Cell Dev. Biol. 27 (2011) 265-290.

[195] B. Leitinger, Discoidin domain receptor functions in physiological and pathological conditions, International review of cell and molecular biology, Elsevier2014, pp. 39-87.

[196] Y. Zhang, J. Su, J. Yu, X. Bu, T. Ren, X. Liu, L. Yao, An essential role of discoidin domain receptor 2 (DDR2) in osteoblast differentiation and chondrocyte maturation via modulation of Runx2 activation, J. Bone Miner. Res. 26(3) (2011) 604-617.

[197] C.M. Lawrie, E.S. Sulistijo, 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. 396(4) (2010) 924-936.

[198] I. Mayrose, D. Graur, N. Ben-Tal, T. Pupko, Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior, Mol. Biol. Evol. 21(9) (2004) 1781-1791.

[199] T.A. Hopf, L.J. Colwell, R. Sheridan, B. Rost, C. Sander, D.S. Marks, Three-dimensional structures of membrane proteins from genomic sequencing, Cell 149(7) (2012) 1607-1621.

[200] A. Fuchs, A.J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, D. Frishman, Co-evolving residues in membrane proteins, Bioinformatics 23(24) (2007) 3312-3319.

[201] A. Kauko, K. Illergård, A. Elofsson, Coils in the membrane core are conserved and functionally important, J. Mol. Biol. 380(1) (2008) 170-180.

[202] J.U. Bowie, Membrane protein folding: how important are hydrogen bonds?, Curr. Opin. Struct. Biol. 21(1) (2011) 42-49.

[203] M. Guharoy, P. Chakrabarti, Conservation and relative importance of residues across protein-protein interfaces, Proc. Natl. Acad. Sci. 102(43) (2005) 15447-15452.

[204] D. Talavera, S.C. Lovell, S. Whelan, Covariation is a poor measure of molecular coevolution, Mol. Biol. Evol. 32(9) (2015) 2456-2468.

[205] A.S. Khadria, A. Senes, The transmembrane domains of the bacterial cell division proteins FtsB and FtsL form a stable high-order oligomer, Biochemistry 52(43) (2013) 7542-7550.

[206] J.P. Dawson, R.A. Melnyk, C.M. Deber, D.M. Engelman, Sequence context strongly modulates association of polar residues in transmembrane helices, J. Mol. Biol. 331(1) (2003) 255-262.

[207] T.H. Walther, A.S. Ulrich, Transmembrane helix assembly and the role of salt bridges, Curr. Opin. Struct. Biol. 27 (2014) 63-8.

[208] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, Nature 450(7172) (2007) 1026-30.

[209] I.T. Arkin, A.T. Brunger, Statistical analysis of predicted transmembrane α-helices, Biochim. Biophys. Acta. 1429(1) (1998) 113-128.

[210] M.M. Javadpour, M. Eilers, M. Groesbeek, S.O. Smith, Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association, Biophys. J. 77(3) (1999) 1609-1618.

[211] J.R. Herrmann, J.C. Panitz, S. Unterreitmeier, A. Fuchs, D. Frishman, D. Langosch, Complex patterns of histidine, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions, J. Mol. Biol. 385(3) (2009) 912-23.

[212] X. Lin, S.M. Tan, S. Law, J. Torres, Two types of transmembrane homomeric interactions in the integrin receptor family are evolutionarily conserved, Proteins 63(1) (2006) 16-23.

[213] F. Paulhe, M. Wehrle-Haller, M.-C. Jacquier, B.A. Imhof, S. Tabone-Eglinger, B. Wehrle-Haller, Dimerization of Kit-ligand and efficient cell-surface presentation requires a conserved Ser-Gly-Gly-Tyr motif in its transmembrane domain, The FASEB Journal 23(9) (2009) 3037-3048.

[214] G. Heijne, Membrane proteins: from sequence to structure, Annu. Rev. Biophys. Biomol. Struct. 23(1) (1994) 167-192.

[215] C. Hunte, Specific protein–lipid interactions in membrane proteins, Portland Press Limited, 2005.

[216] I.C. Dews, K.R. MacKenzie, Transmembrane domains of the syndecan family of growth factor coreceptors display a hierarchy of homotypic and heterotypic interactions, Proc. Natl. Acad. Sci. 104(52) (2007) 20782-20787.

[217] J. Liang, H. Naveed, D. Jimenez-Morales, L. Adamian, M. Lin, Computational studies of membrane proteins: models and predictions for biological understanding, Biochim. Biophys. Acta Biomembr. 1818(4) (2012) 927-941.

[218] H. Hong, Toward understanding driving forces in membrane protein folding, Arch. Biochem. Biophys. 564 (2014) 297-313.

[219] E.S. Manas, Z. Getahun, W.W. Wright, W.F. DeGrado, J.M. Vanderkooi, Infrared spectra of amide groups in α-helical proteins: Evidence for hydrogen bonding between helices and water, J. Am. Chem. Soc. 122(41) (2000) 9883-9890.

[220] R. Walters, W. DeGrado, Helix-packing motifs in membrane proteins, Proc. Natl. Acad. Sci. 103(37) (2006) 13658-13663.

[221] P.W. Hildebrand, S. Günther, A. Goede, L. Forrest, C. Frömmel, R. Preissner, Hydrogen-bonding and packing features of membrane proteins: functional implications, Biophys. J. 94(6) (2008) 1945-1953.

[222] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—Round XII, Proteins 86 (2018) 7-15.

[223] L.M. Munter, P. Voigt, A. Harmeier, D. Kaden, K.E. Gottschalk, C. Weise, R. Pipkorn, M. Schaefer, D. Langosch, G. Multhaup, GxxxG motifs within the amyloid precursor protein transmembrane sequence are critical for the etiology of Aβ42, The EMBO journal 26(6) (2007) 1702-1712.

[224] C.N. Chin, J.N. Sachs, D.M. Engelman, Transmembrane homodimerization of receptor-like protein tyrosine phosphatases, FEBS Lett. 579(17) (2005) 3855-8.

[225] M.A. Lemmon, J.M. Flanagan, H.R. Treutlein, J. Zhang, D.M. Engelman, Sequence specificity in the dimerization of transmembrane .alpha.-helixes, Biochemistry 31(51) (2002) 12719-12725.