The role of dynamic hydrogen bond networks in protonation coupled dynamics of retinal proteins

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(*doctor rerum naturalium*)

am Fachbereich Physik
der Freien Universität Berlin
vorgelegt von

Michail Lazaratos

Berlin
2022

Erstgutachterin: Prof. Dr. Ana-Nicoleta Bondar

Zweitgutachter: Prof. Dr. Roland Netz

Tag der Disputation: 30.11.2022

Αφιερωμένη στους γονείς μου, που δε σταμάτησαν ποτέ να πιστεύουν σε εμένα.

Dedicated to my parents, who never stopped believing in me.

Abstract

Hydrogen bonds (H-bonds) are an essential interaction in membrane proteins. Embedded in complex hydrated lipid bilayers, intramolecular interactions through the means of hydrogen bonding networks are often crucial for the function of the protein. Internal water molecules that occupy stable sites inside the protein, or water molecules that visit transiently from the bulk, can play an important role in shaping local conformational dynamics forming complex networks that bridge regions of the protein via water-mediated hydrogen bonds that can function as wires for the transferring of protons as a part of the protein's function. For example, the membrane-embedded channelrhodopsins which are found in *archaea* are proteins that couple light induced isomerization of a retinal chromophore with proton transfer reactions and passive flow of cations through their pore. I contributed to the development of a new algorithm package that features a unique approach to H-bond analyses. I performed analyses of long Molecular Dynamics (MD) trajectories of channelrhodopsin variants embedded in hydrated lipid membranes and large data sets of static structures, to detect and dissect dynamic hydrogen-bond networks. The photocycle of channelrhodopsins begins with absorption and isomerization of the retinal from an all-*trans* state to a 13-*cis* state and followed by the deprotonation of the Schiff base. Thus, the retinal is found in the epicenter of the analyses. Through the use of 2-dimensional graphs of the protein H-bond networks I identified protein groups potentially important for the proton transfer activity. Local dynamics are highly affected by point mutations of amino acids important for function. The interior of channelrhodopsin C1C2 hosts extensive networks of protein and H-bonded-water molecules, and a never reported before, network that can bridge transiently the two retinal chromophores in channelrhodopsin dimers.

In a recently identified inward proton pump, AntR, I applied centrality measures on MD trajectories of the homology model I generated, to assess the communication of the amino acid residues within the networks. I detected a frequently sampled long water chain that connects the retinal with a candidate proton acceptor, as well as a conserved serine in the vicinity of the retinal chromophore plays a significant role in the connectivity and communication of the H-bond networks upon isomerization. A similar water bridge is sampled in independent simulations of ChR2, where a participant for the proton donor group connects to the 13-*cis*,15-*anti* retinal. Proton transfer reactions often take place through certain amino acids, forming patterns. I analyzed H-bond patterns or motifs in large hand-curated datasets of static structures of α-transmembrane helix proteins, organized according to the superfamilies they belong, their function and an alternative classification method. The presence of motifs in TM proteins is tightly related to their families/superfamilies of the host protein and their position along the membrane normal.

Zusammenfassung

Wasserstoffbrücken (H-Brücke) sind eine wesentliche Wechselwirkung in Membranproteinen. Eingebettet in komplexe hydratisierte Lipiddoppelschichten sind intramolekulare Wechselwirkungen über Wasserstoffbrückenbindungsnetzwerke oft entscheidend für die Funktion des Proteins. Interne Wassermoleküle, die stabile Stellen im Inneren des Proteins besetzen, oder Wassermoleküle, die vorübergehend aus der Masse zu Besuch kommen, können eine wichtige Rolle bei der Gestaltung der lokalen Konformationsdynamik spielen, indem sie komplexe Netzwerke bilden, die Regionen des Proteins über wasservermittelte Wasserstoffbrückenbindungen überbrücken, die als Drähte für den Transfer von Protonen als Teil der Proteinfunktion funktionieren können. Die in Archaeen vorkommenden, in die Membran eingebetteten Kanalrhodopsine sind beispielsweise Proteine, die die lichtinduzierte Isomerisierung eines Retinachromophors mit Protonentransferreaktionen und dem passiven Fluss von Kationen durch ihre Pore verbinden. Ich habe an der Entwicklung eines neuen Algorithmuspakets mitgewirkt, das einen einzigartigen Ansatz für H-Bindungsanalysen bietet. Ich habe lange Molekulardynamik-Trajektorien von Kanalrhodopsine-Varianten, die in hydratisierte Lipidmembranen eingebettet sind, sowie große Datensätze statischer Strukturen analysiert, um dynamische Wasserstoffbrücken-bindungsnetzwerke zu erkennen und zu zerlegen. Der Photozyklus der Kanalrhodopsine beginnt mit der Absorption und Isomerisierung des Retinals von einem all-*trans*-Zustand zu einem 13-*cis*-Zustand, gefolgt von der Deprotonierung der Schiff-Base. Somit steht das Retinal im Mittelpunkt der Analysen. Durch die Verwendung von 2-dimensionalen Graphen der Protein- H-Brückenetzwerke identifizierte ich Proteingruppen, die für die Protonentransferaktivität wichtig sein könnten. Die lokale Dynamik wird durch Punktmutationen der für die Funktion wichtigen Aminosäuren stark beeinflusst. Das Innere von Kanalrhodopsine C1C2 beherbergt ausgedehnte Netzwerke von Protein- und H-Brücke-Wassermolekülen und ein bisher unbekanntes Netzwerk, das die beiden retinalen Chromophore in Kanalrhodopsine-Dimeren vorübergehend überbrücken kann.
In einer kürzlich identifizierten Protonenpumpe, AntR, wendete ich Zentralitätsmaße auf MD-Trajektorien des von mir erstellten Homologiemodells an, um die Kommunikation der Aminosäurereste innerhalb der Netzwerke zu bewerten. Ich fand, dass eine häufig gesampelte lange Wasserkette, die das Retinal mit einem Protonenakzeptor verbindet, sowie ein konserviertes Serin in der Nähe des Retinal-Chromophors eine wichtige Rolle bei der Konnektivität und Kommunikation der H-Brückesnetzwerke bei der Isomerisierung spielt. Eine ähnliche Wasserbrücke ist in unabhängigen Simulationen von Kanalrhodopsine-2 zu finden, wo ein Teilnehmer für die Protonendonorgruppe mit dem 13-*cis*,15-*anti*-Retinal verbunden ist. Protonenübertragungsreaktionen finden oft über bestimmte Aminosäuren statt und bilden Muster. Ich analysierte H-Brückemuster oder -motive in großen, von Hand kuratierten Datensätzen statischer Strukturen von α-Transmembranhelix-Proteinen, geordnet nach den Superfamilien, zu denen sie gehören, ihrer Funktion und einer alternativen Klassifizierungsmethode. Das Vorhandensein von Motiven in TM-Proteinen steht in engem Zusammenhang mit ihren Familien/Superfamilien des Wirtsproteins und ihrer Position entlang der Membrannormale.

.

Publications

During the course of my PhD studies, I contributed to 7 peer-reviewed publications as a co-author under the supervision of Prof. Dr. Ana-Nicoleta Bondar. This dissertation is based on three publications, which are fully published in peer-reviewed journals.

Siemers, M.†, **Lazaratos, M.**†, Karathanou, K., Guerra, F., Brown, L.S. and Bondar, A.N., 2019. Bridge: A graph-based algorithm to analyze dynamic H-bond networks in membrane proteins. *Journal of chemical theory and computation*, *15*(12), pp.6781-6798. Doi: https://doi.org/10.1021/acs.jctc.9b00697

Harris, A., **Lazaratos, M.**†, Siemers, M.†, Watt, E., Hoang, A., Tomida, S., Schubert, L., Saita, M., Heberle, J., Furutani, Y. and Kandori, H., 2020. Mechanism of inward proton transport in an antarctic microbial rhodopsin. *The Journal of Physical Chemistry B*, *124*(24), pp.4851-4872. Doi: https://doi.org/10.1021/acs.jpcb.0c02767

Lazaratos, M., Siemers, M., Brown, L.S. and Bondar, A.N., 2022. Conserved hydrogen-bond motifs of membrane trans-porters and receptors. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, p.183896. Doi: https://doi.org/10.1016/j.bbamem.2022.183896

A methodological development named "*Unique Shortest Paths – USP*" that was originally described in the following publication will be presented in this dissertation. No other content from this publication will be presented or discussed here.

Karathanou, K.†, **Lazaratos, M.**†, Bertalan, É., Siemers, M., Buzar, K., Schertler, G.F., Del Val, C. and Bondar, A.N., 2020. A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2. *Journal of structural biology*, *212*(2), p.107617. Doi: https://doi.org/10.1016/j.jsb.2020.107617

† Equal contribution.

The article "Harris, A., **Lazaratos, M.**, et al. 2020. Mechanism of inward proton transport in an antarctic microbial rhodopsin. *The Journal of Physical Chemistry B*, *124*(24), pp.4851-4872. Doi: https://doi.org/10.1021/acs.jpcb.0c02767" was selected as the supplementary cover of the issue "June 18, 2020, Volume 124, Issue 24, Pages 4851-5092", for which I prepared the supplementary cover with Andrew Harris.
The supplementary cover is found below.

Supplementary cover image for the Journal of Physical Chemistry B, June 18, 2020, Volume 124, Issue 24, Pages 4851-5092.
doi: https://pubs.acs.org/pb-assets/images/_journalCovers/jpcbfk/jpcbfk_v124i024-2.jpg?0.5635301250446989

Three additional papers were published during my PhD studies and will not be presented in this dissertation:

Lazaratos, M., Karathanou, K. and Bondar, A.N., 2020. Graphs of dynamic H-bond networks: from model proteins to protein complexes in cell signaling. *Current Opinion in Structural Biology*, *64*, pp.79-87. Doi: https://doi.org/10.1016/j.sbi.2020.06.006

Mroginski, M.-A., Adam, S., Amoyal, G.S., Barnoy, A., Bondar, A.-N., Borin, V.A., Church, J.R., Domratcheva, T., Ensing, B., Fanelli, F., Ferré, N., Filiba, O., Pedraza-González, L., González, R., González-Espinoza, C.E., Kar, R.K., Kemmler, L., Kim, S.S., Kongsted, J., Krylov, A.I., Lahav, Y., **Lazaratos, M.**, NasserEddin, Q., Navizet, I., Nemukhin, A., Olivucci, M., Olsen, J.M.H., Pérez de Alba Ortíz, A., Pieri, E., Rao, A.G., Rhee, Y.M., Ricardi, N., Sen, S., Solov'yov, I.A., De Vico, L., Wesolowski, T.A., Wiebeler, C., Yang, X. and Schapiro, I., 2021. Frontiers in Multiscale Modeling of Photoreceptor Proteins. *Photochemistry and Photobiology*, *97*(2), pp.243-269. Doi: https://doi.org/10.1111/php.13372

Qi, C., Lavriha, P., Mehta, V., Khanppnavar, B., Mohammed, I., Li, Y., **Lazaratos, M.**, Schaefer, J.V., Dreier, B., Plückthun, A. and Bondar, A.N., 2022. Structural basis of adenylyl cyclase 9 activation. *Nature Communications*, *13*(1), pp.1-11. Doi: https://doi.org/10.1038/s41467-022-28685-y

Acknowledgements

This dissertation would not be possible without many people that helped me along the way, in their ways. I cannot express my gratitude enough for my supervisor Prof. Dr. Ana-Nicoleta Bondar who gave me the opportunity to work on amazing projects and the freedom to pursue my scientific curiosity. I wholeheartedly thank her for devoting time and energy on my projects to answer my questions, guiding me as well as funding my studies and making this PhD possible. I received enormous amounts of support throughout my PhD studies, from project guidance, group seminars, manuscript writing to writing this dissertation, even during her relocation. This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 221545957 – SFB 1078 / TP C4 (to Ana-Nicoleta Bondar), from February 2018 to May 2022.

I would like to thank my second supervisor Prof. Dr. Roland Netz for his support and guidance, and Prof. Dr. Joachim Heberle for his guidance and for providing me with a solution for funding for the last stages of my studies (January 15th 2022-May 15th 2022). I sincerely thank Prof. Dr. Ana-Nicoleta Bondar for the recommendation letter that helped me to receive the funding mentioned above. Many thanks to Prof. Dr. Ulrike Alexiev for the support in the last stages of my PhD. Besides my supervisors in the FU-Berlin I would like to thank Prof. Dr. Leonid S. Brown from the University of Guelph with whom I had the honor of collaborating in three publications, and I learned a lot during the process. His student at the time, Dr. Andrew Harris was a very pleasant collogue to work with during the project of AntR. We had a very good time when he visited our department as a visiting scientist and we prepared a supplementary cover together, and I thank him for the collaboration. I would like to thank Jens Dreger from the Physics Department, Dr. Loris Bennett, and Dr. Boris Proppe from the Zedat High-Performance Computing center of the ZEDAT and Dr. Christian Tuma of the North-German Super Computing Alliance HLRN for excellent technical support and computational resources. Many thanks to Ms. Sieglinde Endrias and Ms. Sylvia Luther for administrative assistance throughout my stay at the FU Berlin and to Mr. Andreas Heß for his patience and help during the submission process of this dissertation.

I am truly grateful to the SFB 1078 (project C4) for making the project possible and allowing me to participate in seminars, colloquia, talks, conferences, and retreats. They have been instrumental in inspiring me to drive my projects forward and always keep evolving. This project would not be possible without the entire infrastructure of the Freie Universität Berlin, and I am grateful that I was able to be a part of this institution.

A special thanks to my friends and co-workers with whom I shared many experiences during the four years of this project. Oskar Klaja, Lukas Kemmler and I, shared an office together and we some good time there. Lukas has been a great listener and we exchanged a lot of opinions during our time in the office. Krzysztof Buzar and Konstantina Karathanou received plenty of visits by me, but they never complained. We spent a lot of time together and I learned a lot from them. Many thanks to Krzysztof for helping me with programming scripts when I did not have the experience. I learned a lot from our discussions and hands-on sessions. Konstantina was there

from the beginning to show me the ropes, when everything looked so complicated when I first moved here, and I thank her for that. A huge thanks to Malte Siemers, who was the backbone for the design and implementation of Bridge, which allowed me to use it to its full capabilities for data analyses and then use it as a platform to expand it with more features and functions as I progressed in my PhD. Malte was always there to help me with my code when it seemed that I was missing something, and I really thank Malte for that. I wish them well in their future and I never forget the moments we shared. I met Honey Jain and Bhav Kapur during the end of my PhD, but I truly believe they are good people. Honey taught me a lot about India, and I will not forget the fascinating discussions we had in the office. We might never agree on how long we should have the windows open. If it were for me, we would never close them again. I thank Krzysztof Buzar, Federico Baserga, Pit Langer, David Burr and Stefanie Schrottke for being good friends and introducing me to Dungeons & Dragons. Maybe I will be talking more in the next campaign.

I think that I wouldn't have come so far it were not for my friends and family. They have always been there for me, even when some are 2.000km away. I will be looking forward to seeing them in person again.

Table of Contents

List of Equations

List of Figures

List of Tables

List of Appendices

List of Abbreviations

AntR	Antarctic Rhodopsin
Aqy1	Aquaporin-1
BC	Betweenness Centrality
BR	Bacteriorhodopsin
C1C2	Channelrhodopsin-1-Channelrhodopsin-2 chimera
CHARMM	Chemistry at HARvard Macromolecular
ChR2	Channelrhodopsin-2
CP	Cytoplasmic
DC	Degree Centrality
EC	Extracellular
EM	Electron Microscopy
FTIR	Fourier transform infrared
GPCR	G-protein coupled receptor
H-bond	Hydrogen bond
IR	Infrared
LJ	Lennard-Jones
MD	Molecular Dynamics
NAMD	NAnoscale Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NPA motif	Asparagine-Proline-Alanine motif
OPM	Orientations of Proteins in Membranes
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PDB	Protein Data Bank
POPA	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidic acid

POPC	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
POPE	1-palmitoyl-2-oleoyl-sn-glycero-3 phosphoethanolamine
POPI	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoinositol
POPS	1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine
PPM	Positioning of Proteins in Membranes
RMSD	Root-mean-square deviation
RSB	Retinal Schiff base
SB	Schiff base
SF	Selectivity Filter
TCDB	Transporter Classification Database
TIP3P	Transferable Intermolecular Potential 3P
TIP4P	Transferable Intermolecular Potential 4P
TM proteins	Transmembrane proteins
USP	Unique Shortest Paths
UV	Ultraviolet
vdW	Van der Waals
VMD	Visual Molecular Mechanics

Chapter 1 Introduction

The dissertation is focused on hydrogen bonding in transmembrane proteins (TM proteins), which are found embedded in biological cell membranes. A large part of this thesis will be dedicated to a TM protein coming from algae, which has a photoactivatable switch as a co-factor. The retinal chromophore is found to be bound to a conserved lysine amino acid residue, through a Schiff base. It is a light-gated ion channel and undergo a photocycle where light is absorbed, and the retinal chromophore isomerizes from an initial *all-trans* state to a *13-cis* state. Another protein that will be discussed is a newly discovered rhodopsin found in Lake Fryxell, Antarctica, called AntR. In the last part of this thesis, a study on H-bond motifs that could be important for function will be presented. Large data sets of TM proteins were compiled, curated, and analyzed for this purpose.

1.1 Proteins

Proteins are essential biomolecules for life as we know it. They are polymers, comprised by linked L-α-amino acids which are the building blocks. There are hundreds of natural amino acids but only 20 are found abundantly in proteins [1]. Amino acids feature a carboxyl ($-CO_2^-$) and an amino group ($-NH_3^+$), separated by carbon atoms. In α-amino acids there is only one carbon atom separating the carboxyl and amino groups [1]. That carbon atom is α to the carboxyl group [1]. Attached to the α-carbon is a side chain (R), unique to the amino acid, and its chemical nature can characterize the amino acid as polar, nonpolar, acidic, or basic. The α-carbon is a chirality center, provided the side chain is not a hydrogen atom, and naturally occurring amino acids in proteins are characterized as L-amino acids. Amino acids are linked together through peptide bonds formed between the carboxyl and the amino group, forming peptides. Peptides can be characterized according to their length as oligopeptides (2-20 amino acids) and polypeptides. (20-50 amino acids). Larger amino acid chains are referred to as proteins. Protein can have a complex structure in the 3-D space and feature four main levels of structure. The primary structure is the raw amino acid sequence of the protein. The secondary structure is the local three-dimensional structures that are stabilized through hydrogen bonds. Very common are the α-helix, the β-sheet, and the loops. The tertiary structure is the overall shape of the protein in 3-D space and is often referred to as the protein fold. Typically, proteins will fold in a way that they maximize their stability and minimize their energy. In the tertiary structure elements like the disulfide bonds or hydrophobic interactions through the nonpolar amino acids are observed. Lastly the quaternary structure is comprised of multiple folded chains (subunits) that come together

into a single complex. Proteins can transport molecules or ions, catalyze reactions, be involved in cell signaling or provide structure in the cell, among many others.

1.2 Hydrogen Bonds

In 2011 the IUPAC task force recommended a new definition for the hydrogen bond, that would include the insight gathered on the topic over the last half-century [2, 3]. The suggested definition for the hydrogen bond is as follows: "The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation." [2, 3]. Commonly, a hydrogen bond involves oxygen, nitrogen, and fluorine atoms (three of the most electronegative atoms in the periodic table) that act as the donor atom and a hydrogen atom that is covalently bound to the donor. The hydrogen atom is shared with an acceptor atom, which is also electronegative in nature [4]. The typical nomenclature for a hydrogen bond is D-H···A-X, where D-H is the hydrogen bond donor and the hydrogen bond depicted with the three dots. The acceptor can be the atom or anion; A or A^-, respectively, or the fragment A-X, when X is bonded to A [2, 3]. Atoms D and A can be the same. A hydrogen atom can only form one chemical bond, when considering the simple valence theory [5]. This led Linus Pauling to suggest that the nature of the hydrogen bond is of electrostatic (ionic) nature [2, 3]. Even in the 50's, the hydrogen bond energy was decomposed in four components by Coulson [6], and Tsubomura [7] suggested that the hydrogen bond cannot be characterized as primarily electrostatic, even if the other components are practically zero [7, 8]. The notion of a purely electrostatic nature of the hydrogen would also fail to explain experimental findings. For example, infrared (IR) spectroscopy has been a vital technique in the detection of hydrogen bonds. It was initially suggested that the D-H stretching peak disappears from the spectrum, upon hydrogen bond formation [2, 8, 9], but it was later shown that the stretching peak is red-shifted instead [2, 8, 10]. The shift is correlated to the strength of the hydrogen bond D-H···A [2, 8, 10], and in general the intensity of the band and the bandwidth are also increased. The greater the lengthening of the D-H bond upon hydrogen bonding, the stronger the hydrogen bond is, and new vibrational modes associated with the H···A hydrogen bond are generated [2, 3]. In some cased, the length of the D-H bond could decrease, resulting in a blue-shift. It has been shown that blue and red-shifted hydrogen bonds do not have a fundamental difference [2, 11]. The D-H···A hydrogen bond will result in an electronic redistribution around the hydrogen atom, meaning that the proton is now strongly de-shielded [2]. NMR techniques can detect this de-shielding as a shift to a lower magnetic field [2], with the magnitude of the shift being relatable to the hydrogen bond strength [12, 13]. In their review van der Lubbe and Fonseca Guerra note that hydrogen bonds are a complex superposition of different energy terms. Those terms can have varying importance, according to the molecular system at hand [8]. Nowadays it is generally agreed that hydrogen bonds are a complex interplay between the five following contributions: a) electrostatic or coulomb energy, b) exchange repulsion, c)

polarization energy, d) charge-transfer energy or covalent bonding, and e) dispersion forces [6, 7, 13].

Hydrogen bonds are an interaction observed in nature abundantly, from providing water with its properties, such as surface tension and its boiling point, to providing structure to DNA and RNA by holding the base pairs of together. Hydrogen bonds are also found to play an important role in the catalytic active site of enzymes. Those intermolecular and intramolecular interactions can be essential for the folding, stability and function of proteins [14, 15], stabilizing the secondary and tertiary structures. White and Wimley showed in their review that hydrogen bonds between peptide bonds are actually the driving force for having proteinic secondary structure (α-helix, β-sheet) in apolar environments e.g. lipid bilayers, since the transfer free energy of a hydrogen bonded peptide bond in a POPC interface is more than two-fold lower [16, 17] compared to a standard peptide bond partitioned to the POPC interface [18]. High-resolution crystal structures have revealed that protein interiors can host extensive networks of hydrogen bonds that interconnect different regions of the protein [19, 20], and experiments and computations suggest important roles of internal hydrogen bond networks in shaping the protein conformational dynamics [21-24]. In the gas phase, the hydrogen bonds of a peptide environment is ≈4.9 kcal/mol [25-27]. In water, the hydrogen bond energy is significantly reduced to ≈1.5 kcal/mol [28, 29]. MD simulations can reproduce the experimental values and *ab initio* calculations with very good agreement. The isolated α-helix and β-hairpin [30] were simulated in the gas phase and in a water solution. The calculated energy of activation was 5.57 and 4.79 kcal/mol in the gas phase and 1.93 and 1.58 kcal/mol in water, for the α-helix and β-sheet, respectively [27].

Hydrogen bonds depend highly on their local environment [31]. In the interior of bacteriorhodopsin, the hydrogen bond motifs (see following subchapter) T46-D96 and T90-D115 were characterised as the strongest measured interactions, contributing -1.7 ± 0.3 kcal/mol. Considering that the contributions involve two hydrogen bonds per pair, the corresponding contribution was determined about -0.9 kcal/mol per hydrogen bond [31]. The second strongest interaction measured was between the pair T170-S226 at -0.8 ± 0.3 kcal/mol, while other pairs had even weaker stabilizing contributions [31]. Through double-mutant cycles it was shown that the eight hydrogen bonds studied showed a small contribution of 0.6 kcal/mol [31]. In the outer membrane protein A channel, the interaction of a salt bridge between E52-R138 was measured at -5.6 ± 0.4 kcal/mol, while the salt bridge between E128 and R138 showed a much weaker interaction of -0.6 ± 0.5 kcal/mol [32].

During dynamics at room temperature individual hydrogen bonds within hydrogen bond networks can display complex dynamics whereby hydrogen bonds rapidly break and reform [21, 33, 34]. Such hydrogen bonds that can break and reform without significant energy penalties might provide proteins with structural plasticity needed for function [21, 31]. Dynamic hydrogen bond networks appear of particular importance for proteins whose functioning involves proton transfer. Dynamics of hydrogen bond networks might be required to orchestrate formation of proton transfer paths [33, 35, 36]. IUPAC states that hydrogen bonds participate in proton-transfer

reactions and they can be considered the partially activated precursors to such reactions [2, 3].

Another type of interaction that can play a crucial role in protein stability [37, 38] and folding [39] are hydrophobic bonding or hydrophobic interactions [40]. Hydrophobic interactions are formed when non-polar groups associate in an aqueous solution, so that they interact less with the neighboring water molecules [40]. In a study of small globular proteins Pace and co-workers determined that the hydrophobic interactions played the dominant role in the stability of the protein and always have a larger contribution as compared to hydrogen bonds [41].

In the first part of this dissertation, an algorithm package [42] I contributed to the development of, is presented. This algorithm enables efficient analyses of hydrogen bond networks from experimentally determined structures and computer simulations. An expansion [43] of said algorithm was presented one and a half years after its original release with the addition of hydrophobic cluster interactions. Hydrogen bonds will be referred to as H-bonds for simplicity, from this point on in this dissertation.

Hydrogen bonding criteria

An early description of H-bonds in proteins by an electrostatic potential was given in ref. [14]. The authors calculated the electrostatic interaction energy between H-bonding groups. That was possible after assigning partial charges to atoms C, O, N, H. A H-bond was accepted between groups C=O, N-H if the interaction calculated was of smaller value than the threshold given at -0.5 kcal/mol. Observations from accurate crystal structures and computations led to the conclusion that a geometric criterion based on the distance between the hydrogen atom and the oxygen heavy atom suffices to describe key elements of H-bonding [44]. The strength of the H-bond between two amide groups was found to be largely independent of the H-bond angle when the amides were at optimal interaction distance and the steric hindrance was minimal [45]. Consistent with the above notion, the authors [46] proposed an analysis to assess proteinic structure quality and to refine NMR structures. There, they used a geometric criterion for the describing hydrogen bonds. The H-bond distance R_{HO} was considered smaller than 2.5 Å and the hydrogen bond angle was taken in the range of $120° \leq \theta_{NHO} \leq 180°$. A similar approach was used in another study [47] where the distance between oxygen atoms was set to $R_{O\text{-}O} \leq 3.5$ Å and the angle in the range $120° \leq \theta_{OH\text{-}O} \leq 180°$.

Algorithm implementations

Several algorithms have been proposed for the analysis of H-bonds from molecular dynamics simulations and experimentally determined static structures. In what follows a brief overview of some recent developments in the field is presented. MDAnalysis [48, 49] is a toolkit to analyze molecular dynamics trajectories. MDAnalysis is implemented as a Python package and makes use of object-oriented programming. H-bonds are defined according to the default H-bond donors and acceptors

of the CHARMM27 force field, and can be monitored along the simulation trajectory, and lifetimes for H-bonds are calculated from the autocorrelation function [50]. MDAnalysis further includes the "WaterBridgeAnalysis" module, which allows the user to identify water bridges that connect two disjoint sets of amino acid residues. A water bridge is defined as one water molecule being H-bonded to one or more amino acid residues from each of the two sets simultaneously. Wires with more than one water molecules cannot be computed and an overlap between the two sets cannot be handled by the program. HBonanza (Hydrogen-Bond analyzer) is an open source tool implemented in Python [51], designed for the analysis and visualization of H-bond networks of static structures, and analyses of molecular dynamics simulations. It can be used to identify H-bonds between amino acid residues. Analyses results are visualized in the Visual Molecular Dynamics (VMD) [52] program. H-bond networks are identified recursively starting from a source group of interest, such as the ligand of a protein. HBAT [53] is an automated tool for the analyses of non-bonded interactions in PDB files. It is created using Perl and was proposed as a new generation software since it also features a Graphical user Interface (GUI) for ease of use [53]. HBAT is mainly targeted to crystallographers and works best with already predetermined hydrogen coordinates. One of the first programs for detailed analysis of individual structures as well as macromolecules is HBexplore [54]. Targeted for nucleic acid structures, and applicable to protein structures this program written in C offers options for further analysis compared to its predecessors. Since this program is targeted for single structures originating from the PDB [55, 56], the authors have included a procedure for generating hydrogen positions according to the protocol of Cornell et al. [57], before the actual analysis. An exemplary illustration of network analysis was demonstrated by the group of Zaida Luthey-Schulten were they studied allosteric communication effects in tRNA:protein complexes [58-60]. They performed community analysis employing the Girvan-Newman algorithm [61] to calculate the community distributions while visualization of networks was achieved through graph theory elements.

1.3 Channelrhodopsin-2

Coming from the unicellular alga *Chlamydomonas reinhardtii,* Channel-rhodopsin-2 (ChR2) [62, 63] is found in the eyespot region and is involved in the photoperception of the alga, resulting in the mediation of the phototaxis [64, 65] and photophobic response [66-68]. ChR2 belongs to the larger family of microbial rhodopsins, is a light sensitive non-selective cation channel [62, 63], and consists of seven transmembrane helices (TMs) containing a retinal chromophore covalently bound to a conserved lysine ($K257^{ChR2}$) residue via a protonated Schiff base. Upon illumination with blue light a photocycle begins, where the retinal chromophore isomerizes from an all-*trans* to a 13-*cis* state, triggering a series of conformational changes resulting to the opening of the ion pore to a diameter of around 6 Å [62, 69] allowing cations their passage across the membrane. The ion conduction pathway is located between helices 1, 2, 3 and 7 of each monomer [70]. The pore diameter was suggested around 6 Å in

diameter, which constitutes it wider than a pore of a voltage-activated Na^+ channel [62], since methylated ammonium ions (methyl / di-methyl and tetra-methyl-ammonium [62]) and guanidinium could permeate ChR2, whereas tetra-ethyl-ammonium, Mg^{2+} and Zn^{2+} did not measurably permeate [62]. ChR2 will primarily conduct protons across the membrane, along with monovalent and divalent cations, to some extent [70]. The relevant conduction of protons compared to Na^+ is 10^5-10^6 [70], while Ca^{2+} conductance is ≈ 12% that of Na^{2+}, and cation conductance is pH dependent [62, 70, 71]. Increasing atomic radius of the cation would decrease its permeability [62], and it was suggested that the cations would need to be in a mostly dehydrated state in order to permeate the pore [62], but the dehydration is most likely not complete [70]. Gatsby suggested that channelrhodopsins might have evolved from microbial pumps, such as bacteriorhodopsin or halorhodopsin, by degradation of a gate [72]. Although ChR2 is characterized as a light-gated ion channel, the flow rate of ≈ 3 x 10^4 [62] ions per second can be considered slow for a channel, and it is comparable to a fast pump [73], constituting it in a "grey zone" between a pump and a channel [72]. ChR2 shows high homology to bacteriorhodopsin in regards to helices 3 and 7 [74] led to the question if it can function as proton pump [74]. Indeed, the bifunctional character of ChR2 was shown in the absence of any electrochemical gradient, where ChR2 could act as a leaky proton pump [74].

Nagel et al. suggested the four-state model [62] to describe the electrophysiological reactions during the photocycle [70]. In this model ChR2 reaches an excited state from the photo-activation (<1 ns), followed by dark reactions leading to the open state (<1 ms) and closing of the desensitized state (10-400 ms), depending on the pH [62]. The closing of the channel has effective time constants between 10 and 20 ms [70]. In the same study, the three-state model was also proposed [62], later modified by Nikolic et al. [75, 76]. Only the four-state photocycle models with two closed and two open states that were proposed by Nikolic et al. [76] and Hegemann et al. [77] could describe the kinetics and the dark recovery of the peak current [70]. Many spectroscopic studies have contributed to understanding the photocycle of ChR2. UV-visible spectroscopy provides insight on the protonation of the retinal Schiff base, electronic changes [78] in the retinal and interactions between the protein and the retinal [79]. Time-resolved Fourier transform infrared (FTIR) spectroscopy is another widely used technique to access structural dynamics (conformational changes) in retinal proteins and changes in the hydrogen bonding of amino acid residues [79] (see subchapter *"Hydrogen Bonds"* above). The photocycle of ChR2 is comprised of the dark state and four main kinetic intermediate states, distinguished by the absorption maxima (λ_{max}). The dark state exhibits an λ_{max} at ≈ 470 nm, followed by the main kinetic intermediates P_1^{500}, P_2^{390}, P_3^{520}, and P_4^{480} [70, 76, 78-85]. The depolarization of the cell membrane upon light-activation of ChR2 has constituted it as a very widely used and versatile [83] optogenetics tool [86-96], when expressed in mammalian neurons.

Despite its wide usage in optogenetics, a high-resolution structure of the wild-type ChR2 remained unknown until very recently when Volkov et al. solved the three-dimensional structure through x-ray crystallography of the wild-type ChR2 at a resolution of 2.39 Å [97]. Previous attempts to gain insight on the structure of ChR2 were

the presentation of a low-resolution structure of the C128T mutant through cryo- electron microscopy, solved at 6 Å [98]. In 2012 Kato *et al.* [99] solved the crystal structure for a chimeric Channelrhodopsin, consisting of the first five transmembrane helices from Channelrhodopsin-1 and the last two TM helices from Channelrhodopsin-2 [97]. Additionally, homology models have been employed in order to derive three dimensional structures of ChR2 [100-103]. Knowing the structure of Channelrhodopsin-2 is crucial due to its wide and successful applications in optogenetics. Nevertheless, there are still numerous questions to be addressed about the function of protein.

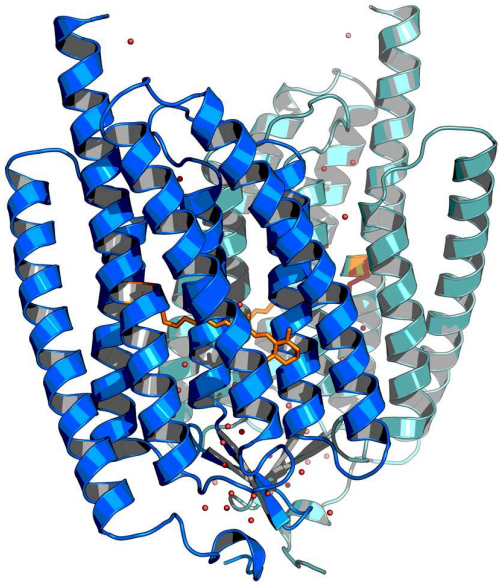

Figure 1.1. Molecular representation of the crystal structure of ChR2. The graphics are based on the coordinates the crystal structure of ChR2 (PDB ID : 6EID [97]). The retinal chromophore is shown as orange lines and the two monomers as blue and cyan ribbons. Red spheres represent the oxygen atoms of co-crystallized water molecules. The graphics were prepared using PyMol [104].

1.4 Functional motif examples in proteins

Motifs are structural or sequence patterns that can have a large impact in the biological function of important molecules, ranging from DNA to lipids and proteins. Specifically in proteins, motifs or patterns of amino acid residues interact in ways that can be the driving force for functionality and biological assembly. A prime example is the 7-amino acid residue motif LIxxGVxxGVxxT that can dimerize effectively any α-transmembrane helices that is introduced into [105]. Lemmon and co-workers were able to quantify this with a series of experiments and computations on a chimera based on the nuclease from *Staphallococcus aureus* (SN) and the transmembrane part of the human glycophorin A (GpATM). With a multi stage mutagenesis, converting all amino acid residues of the GpATM to leucine except the ones that comprise the motif, effectively introducing the motif to poly-leucine, they concluded that the motif is responsible for the

dimerization since there were no differences between the SN/GpA and the poly-leucine mutant through a competition experiment [105].

A more general GxxxG motif is also responsible for dimerization of α-helices. The glycine amino acid residues are found within one helical turn, are aligned and form a flat surface (platform) for the glycines of the other helix to approach and form a homodimer structure [106, 107]. The helix-helix interactions are stabilized by H-bonding between the Cα H atom and a carbonyl group belonging to the backbone. This H-bond provides stability to the helix-helix assembly by 0.88 kcal/mol [108]. In a large scale search on the non-redundant PDB [55] containing 12808 α-helices, Kleiger and co-workers [109] showed that the GxxxG motif's occurrences are 41 ± 9% higher than expected, if the glycine amino acid residues were uniformly distributed. Similarly, the AxxxA motif's occurrences were computed to be 30 ± 3% higher than expected. Structures containing the GxxxG motif showed a statistically significant increase in the helix-helix distance in the area of 6-7 Å, while the AxxxA containing structures did not, implying a biologically significant importance of the motif for function. Even though the GxxxG motif has been thoroughly described as a dimerization signal Arbely and co-workers identified the pattern in the spike protein of SARS-CoV [110]. The SARS-CoV spike protein is known to be a homotrimer and it is the only transmembrane domain out of the ten putative coronaviruses mentioned in the study. The motif is also found in the proximity of highly conserved amino acid residues. Through mutagenesis of the glycine groups to isoleucine in two steps they showed that the motif is vital for the oligomerization of the transmembrane domain of the SARS-CoV, through the means of quantitative CAT assay and electrophoresis. The motifs SxxSSxxT and SxxxSSxxT were identified in twelve TM domains in the search of Dawson et al. [111] for polar motifs, and it was suggested that the combinations of the two or three serine residues contribute to the stability of helical interactions [111]. Due to the low number of identified motifs, it was suggested that it is not common in helical association [111]. The motif was found to conserved in any protein family and it was thus suggested that it would be unlikely to play a role in the function of the proteins it is found in [111]. The truncated motif SxxS was found in the structure of halorhodopsin with a possible involvement in helical association and stabilizing the folded structure [111, 112]. In a bioinformatics analysis of channelrhodopsin sequences [100], it was shown that channelrhodopsins have an increased content of polar groups in helix B, compared to bacteriorhodopsin. A sequence motif was detected in all, except two of the studied sequences as ExxxxxxExxxxxxExxxX. Further analysis indicated that the glutamate motif would be characterized as (GW)EEbYVCxbEbxKVxbEbbxE(F), where *b* represents a hydrophobic group [100].

Patterns are identified in another secondary structure of proteins, in the β-sheets. Understanding the β-sheet connectivity is vital for predicting protein folding and the three-dimensional structure of an amino acid sequence. An early description and classification of β-sheet topologies and connectivity patterns was given by Richardson [113]. Using elements of graph theory, such as the subgraph isomorphism algorithm, Mitchell and co-workers [114] introduced an intensive search for these motifs across the structures of three-dimensional structures of proteins available in the PDB at the time.

Effectively introducing one of the first database analyses of proteinic crystal structures with a known 3-D structure [114].

One of the earliest applications of graph theory in motif search was to provide a methodology for H-bond motifs and patterns in small organic molecules, categorizing how different functional groups H-bond. The motifs are defined based on the H-bond types, i.e., the chemical nature of the H-bond partners. With a formula of four components a graph set can be defined to characterize H-bond patterns between molecules, while higher order networks can be derived by combining basic motifs. The suggested protocol provided empirical rules for future H-bonding analyses of crystals or solutions of simple or more complex systems, directed to H-bond driven aggregation [115, 116]. The methodology was applied on small organic crystals in order to identify patterns in H-bonding between protein-DNA, with a focus around arginine, glutamate, asparagine or molecules that contain fragments of those amino acid residues.

Membrane proteins that feature coupling of proton transfer reactions with conformational changes to perform their biological function often rely on acid residues with carboxylic or imidazole chemical groups in their sidechains. Ser/Thr motifs are suggested to affect local helical dynamics through the means of kink introductions, solvation enhancement and flexibility. In a large-scale study of bioinformatics analyses combined with MD simulations, Ser/Thr structural motifs were identified in key areas bacteriorhodopsin, halorhodopsin and SecY. It was shown that Ser/Thr amino acid residues sidechains compete with their amide groups to H-bond to *i-3* or *i-4* carbonyl oxygen atom [117-119]. In bacteriorhodopsin, an Asp-Thr-Asp motif is found in the vicinity of retinal Schiff base including D85 which is the proton acceptor after illumination in the beginning of the photocycle. It was shown through MD simulations that the D96, D115 sites participate in H-bonding with Thr amino acid residues part of a TT motif. The Thr group was involved in an intrahelical H-bond to the *i-4* carbonyl at both cases [119, 120]. The sites where shown to be of different sensitivities to perturbations depending on the hydrophobicity of their local environment and it was suggested that they are important for the stability of the protein [120].

Besides structural patterns, motifs can also an essential part for the mechanism of function for many proteins originating from different families. Specifically titratable amino acid residues that can H-bond and their protonation changes can alter local helical dynamics [119, 121]. In the first iteration of the algorithm package Bridge [42], the first algorithm to search for Ser/Thr···Asp/Glu motifs as well as the Ser/Thr hydroxyl···backbone carbonyl of the *i*-3, 4, or 5 relative position that were detected as a pattern and described in refs [35, 36, 117-119, 122] were implemented. Such H-bond motifs are believed to be important for proton binding and shaping local dynamics [35, 36]. One of the most well-studied proton pumps is bacteriorhodopsin (BR), that involves 5 distinct proton transfer steps in its photocycle [123]. It begins with the retinal Schiff base transferring the proton to the primary proton acceptor, D85 [124]. T89 is found in the *i*+4 relative position to D85 and it has been shown that it can an intermediate proton carrier during the first proton transfer event in the photocycle [125-127]. Next in sequence of T89, is T90 that forms an interhelical H-bond with D115 and an intrahelical H-bond with W86 [36]. In BR, the interhelical Thr-Asp motifs within proton transfer

groups [36] organize in clusters, which largely respond to mutations and their dynamics is highly affected by the surrounding local environment [36]. Similar H-bond patterns are observed in the channelrhodopsin C1C2 chimera, the SecY protein translocon, *Anabaena* sensory rhodopsin, KR2 sodium pump, the SERCA calcium pump ATPase, and the multidrug transporter AcrB [35, 36, 100, 121, 128].

In an ammonia channel, two highly conserved histidines are found in the center of a hydrophobic pathway (Figure A.7a) [129]. They form a Histidine-Histidine (His-His) motif via H-bonding through Nδ-H···Nδ , they H-bond to the transporting ammonia molecules contributing H-bond donors, effectively not allowing NH_4^+ ions to be transported [129].

In the TM section of the *Vigna radiata* proton translocating pyrophosphatase homodimer [130] a narrow pathway consisting of charged amino acid residues was identified, including E301, R242, D294 and K742 with the latter forming a salt bridge (Figure A.7b). It was speculated that K742 regulates the protonation/deprotonation of D294 and E301 resembling the machinery of BR [131, 132]. The pathway is proposed to be the proton translocating pathway since it contains trapped water molecules, a common feature of proton transferring systems like BR. One of the water molecules is stabilized by N738, which I identified it to be a motif partner in the search, along with D294.

Chapter 2 Methodology

This work was based on a variety of methods, approaches, and techniques. Namely the biological systems are modelled in the computer and simulated. The technique that was used for those simulations is called Molecular Dynamics Simulation or MD Simulation or MD, which will be the term used throughout this work. MD is a computer simulation technique developed in the 1950's and although in the beginning it was limited to simulating small systems of hard spheres [133], with the rapid development of efficient algorithms and more importantly the capabilities of computer hardware, MD is now capable of simulations of macromolecules in the realm of nanoseconds until even milliseconds [134] using specialized hardware [135]. In the simplest case, the equations of motions of Newton are numerically solved for every particle in a system and thus the time evolution of the system, or trajectory, is determined. Physical and structural properties can be calculated from the trajectories with principals of statistical mechanics. In what follows, some essential concepts of MD are presented. More details about the principles of MD can be found in the Appendix, subchapter *"Molecular Dynamics simulation principles"*.

2.1 Force Field

Quantum mechanical methods can describe systems with high accuracy, but they can be performed at relatively small systems due to their computational cost. Molecular mechanics follows a classical approach for describing a system, ignoring electron motion and interactions. Atoms, parts of molecules and even whole molecules are described as spheres connected with springs. In what follows, spheres are used to describe atoms. Force fields are empirical functions that approximate the potential energy of a system. They are based on the Born-Oppenheimer approximation where the electrons' motions in the system are ignored, and the potential energy is formulated in terms of the nuclei. Force fields are based on parameters that are generated through quantum mechanical calculations and they can accurately approximate qualitatively and quantitatively molecular behavior, while their use in through a classical approach offers high computational efficiency. The quality of the simulations largely depends on the underlying quality of the parameters (or the force field consequently). That quality is evaluated by the ability of reproducing and predict molecular properties [136]. When a force field is derived to describe large molecules such as proteins or DNA, the functional is kept at a very simplistic form (eq. 2.1) [137]. Such force fields are classified as "Class-I" or "Harmonic" because harmonic functions are used to describe stretching and bending terms. Class-I force fields no not include any cross terms i.e., coupling between internal

coordinates. A Coulombic potential and a 12-6 Lennard-Jones (LJ) are used for the electrostatic and van der Waals (vdW) interactions, respectively. No cross-terms are being used in that case. The force field can be further simplified by not including an explicit description for the hydrogen atoms. Instead, they are merged into the neighboring heavy atoms, whose vdW radii are increased to compensate that adjustment [137]. This further simplification approach would not be for example a good choice when attempting to describe H-bonding dynamics. But it could serve well when exploring large-scale conformational changes during extended simulation time.
A simple functional that describes a complete system is [137]:

$$\mathcal{V}(\boldsymbol{r}^N) = \sum_{bonds} \frac{k_i}{2}\left(b_i - b_{i,0}\right)^2 + \sum_{angles} \frac{k_i}{2}\left(\theta_i - \theta_{i,0}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left(1 + cos(n\varphi - \gamma)\right) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \quad 2.1$$

The interactions that a force field describes can be categorized in *bonded* and *non-bonded* and are computed as a function of the coordinates.

$$\mathcal{V}_{total}(\boldsymbol{r}^N) = \mathcal{V}_{bonded} + \mathcal{V}_{non-bonded} \quad 2.2$$

In turn, the *bonded* term contains the descriptions for bond stretching, angle bending, and torsion.

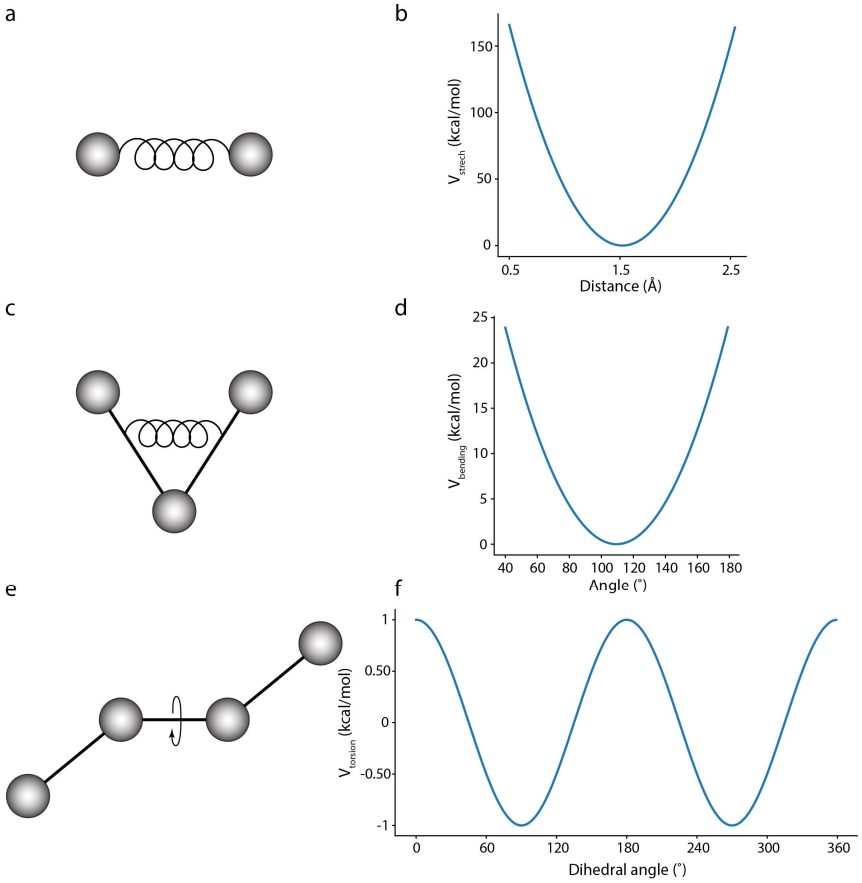

Figure 2.1. Bonded interactions described in a harmonic force field functional. Schematic and graphical representations, respectively, for bond stretching (a, b), angle bending (c, d) and dihedral torsion (e, f). Bond stretching and angle bending are described by a harmonic potential and dihedral torsion Distance is measured in Angstrom (Å), angle in degrees (°) and energy in kcal/mol.

The *non-bonded* term describes the vdW interactions and electrostatics and is calculated for atoms separated by more than two bonds (1-4 interactions).

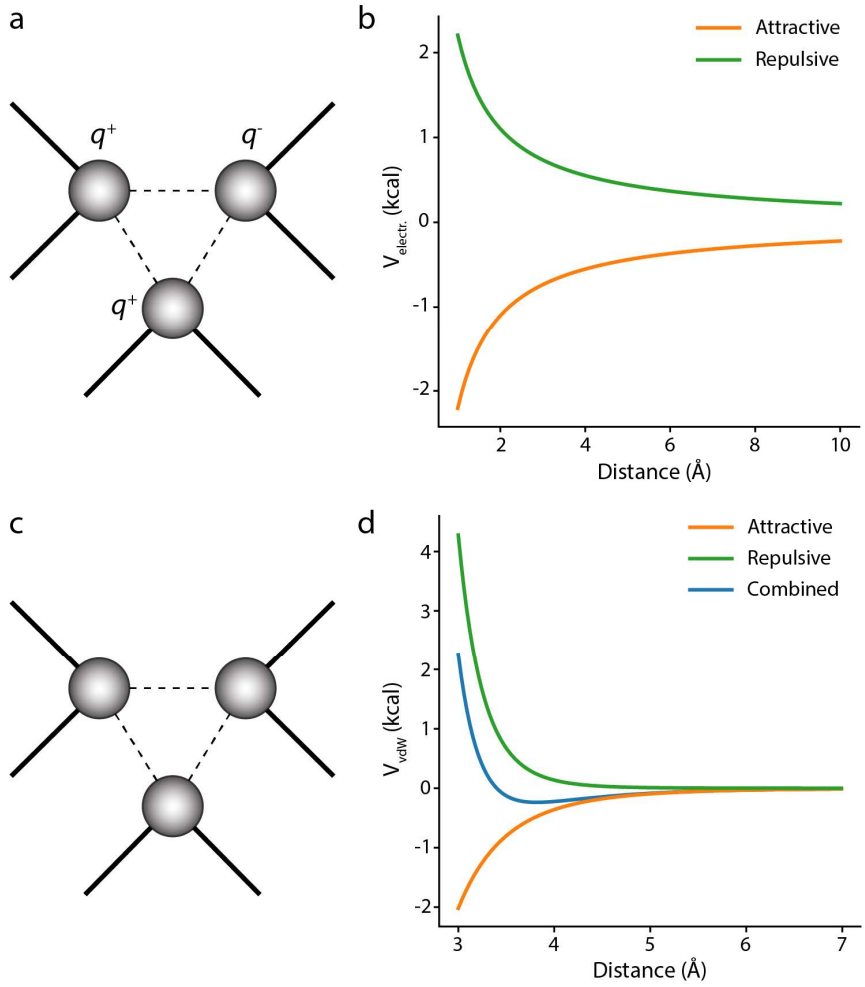

Figure 2.2. Non-bonded interactions described in a harmonic force field functional. Schematic and graphical representations, respectively, for electrostatic interactions (a, b) and vdW interactions (c, d). Electrostatic interactions are described by Coulomb's law and the vdW interactions by a 12-6 Lennard-Jones potential. Distance is measured in Angstrom (Å) and energy in kcal/mol.

In this work, the CHARMM force field [136, 138-147] (eq. 2.3) is used exclusively for the description of the potential energy of the proteins, lipids, water, and ions that will make up the simulation systems. The CHARMM force field is an all-atom, Class-I additive force field, designed for simulations with explicit solvents [145]. It represents all hydrogen-atoms explicitly and does not sum them into neighboring heavy atoms, and provides a good description of proteinic [136, 147] and lipid [146] properties. A Urey-Bradley (UB) term is added to the functional as a harmonic function of the 1-3 distance and an additional harmonic term for the description of the improper dihedrals is introduced, in order to control chirality and out-of-plane motions [145]. The UB term is used rarely, while the improper dihedrals' term is widely used in the CHARMM force field [145]. Both of these terms are included to optimize the fit to vibrational spectra [145]. The CMAP [141, 143] term is a correction for small systematic errors that occur in the description of the protein backbone, from the CHARMM force field [145]. The

result is a highly improved description and agreement with QM maps [141, 143, 145, 148]. The CHARMM force field equation, shown in eq. 2.3, relies on parameters that depend on the atom types. Equilibrium values for the bond distance b, angle θ, UB distance S and improper angle ω are noted with the subscript ($_0$) and the respective force constant K. The dihedral angles are modelled with a cosine function featuring a force constant, the multiplicity or periodicity n and the phase shift γ. The CMAP correction term is a function of the backbone dihedral angles φ, ψ, and concludes the *bonded* term of the function. For the *non-bonded* part, a 12-6 Lennard-Jones potential and Coulomb's law are employed. In the LJ term, for two interacting atoms, i and j, the depth of the potential is found as ε_{ij}^{min}, and the minimum interaction distance R_{ij}^{min}, at which the potential assumes the minimum value. Similarly, the Coulombic interactions are simply described by the partial atomic charges $q_i,\ q_j$ of the interacting atoms and the interatomic distance r_{ij}. The relative dielectric ε constant assumes the value of 1 when an explicit solvent is applied [145].

$$\begin{aligned} \mathcal{V}(\vec{R}) = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{Urey-Bradle} K_{UB}(S - S_0)^2 \\ & + \sum_{dihedral} K_\varphi(1 + cos(n\varphi - \gamma)) + \sum_{impropers} K_\omega(\omega - \omega_0)^2 \\ & + \sum_{residues} U_{CMAP}(\varphi, \psi) \\ & + \sum_{non-bonded} \left\{ \varepsilon_{ij}^{min} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\} \end{aligned} \quad 2.3$$

The CHARMM force field with the harmonic terms and the 12-6 LJ is considered a good compromise between accuracy and computational cost, especially for simulations around room temperature [145], and is suitable for heterogenous systems such as protein-lipid simulations which will be employed throughout this work. Though, it is not limited to only these, it additionally includes descriptions for nucleic acids [149], sugars [150], drugs [151], and drug-like moieties [151]. The accuracy of the force field parameters used for a molecular dynamics simulation will determine how reliable the dynamics of the system are, and thus how reliable the predictions from the simulation are. Especially for co-factors, such as the retinal chromophore, which can be especially challenging to represent in the setting of classical mechanics, due to its complex electronic structure. The importance of reliable parameters was illustrated by Babitzki et al. [152] for various MD studies of bacteriorhodopsin [153-155] using the common widely available force fields and by Hayashi et. al. [156] using the AMBER force field with the TIP3P water model, where a consistent decay of the hydrogen bond between the retinal Schiff base and w402 was observed, followed by the formation of a stable ion pair between the protonated Schiff base and D85 [152]. W402 left the crystallographic position in the vicinity of the retinal Schiff base [152, 155]. Two major parametrization works from Nina et al. [157] and Tajkhorshid et al. [158] were directly compared in MD simulations

of squid rhodopsin for both an all-*trans* and an 11-*cis* retinal in a study by Jardon-Valadez et al. [159]. The parameter sets differ with one another in the quantum mechanical methods they are generated with. In the former, Hartree-Fock in the gas phase was used to derive the bond lengths, angles and partial charges, while water interactions with the Schiff base are accounted for, with two different locations for the water molecule [157, 160, 161]. Simulations with those parameters could reproduce very accurately the free energy the all-*trans* and 13-*cis* models of the retinal, within $\sim k_B T$ from one another, and the experiment [162]. The same parameters were used to correctly characterize the favorable binding energy and interactions between "water A", D85, D212 and the retinal Schiff base [160]. The water molecule notes as "water A" was later identified in crystal structures of bacteriorhodopsin and now referred to as w402 [19, 163]. The latter is performed with B3LYP in the gas phase [164-166] and shows significantly larger torsional barriers, and allowed for a reliable description of the retinal's geometry [158, 161]. It was shown that the parameters used for the retinal chromophore largely affect the dynamics of the system, from the retinal geometry, with the relative orientation of the β-ionone ring to the polyene chain, to the dynamics of the internal water molecules, underlying the strong coupling between the retinal and its local environment [159].

2.2 Water description

In a typical MD simulation, most of the components simulated are water molecules. It is thus vital to have an accurate description of water, that can reproduce its microscopic and macroscopic, bulk properties. Several water models have been proposed and developed over the years with most notable the Single Point Charge - SPC (original [167] and refined [168]), Single Point Charge / Extended - SPC/E [169] , Transferable Intermolecular Potential 3P - TIP3P (original [167] and modified [170]), Transferable Intermolecular Potential 4P - TIP4P [167]. It was decided that CHARMM [171] would be parametrized with the TIP3P water model since it can reproduce the first hydration and shell and the energetics of liquid water very well [138]. Though it lacks in its tetrahedrality and results in fast dynamics through the means of self-diffusion. Its self-diffusion coefficient value is computed to be more than double [172] compared to the experimentally measured one [173, 174], but this can be adjusted with a factor called the "Langevin damping coefficient" which introduces friction to the system. In comparison the SPC/E model offers better tetrahedrality [138] and a self-diffusion coefficient much closer to the experimental one [172], but it introduces inconsistencies in heterogenous systems [138]. It would be a good choice in solvent simulations with an additional energy correction term, but this could lead to overestimated solute-solute interactions and incorrect predictions and computations of the solutes' properties [138]. The TIP4P gives very good results overall but is very expensive computationally. With the introduction of a virtual particle, it would introduce complications in the way the forces would be projected to the real atoms of the simulations [138]. CHARMM used a slightly modified TIP3P where Lennard-Jones parameters are used for both oxygen and hydrogen atoms

[145, 170]. The LJ parameters were introduced to the hydrogen atoms in order to avoid singularities during calculation of integral equations [170, 175].

2.3 Concepts from graph theory

Graph definition

Graphs are mathematical constructs that represent relations between pairs of objects. In principle they are composed of two sets of distinct components. The *vertices* or *nodes* and the *edges*. In this work only the term *nodes* will be used. Edges connect the nodes with one another. Graphs find many applications in various fields [176] since attributes can be assigned to the nodes and edges. The relations between attributes can be then represented and studied through connectivity networks. Namely, applications of graph theory can be found in computer science and computer architecture, biology, physics, chemistry, sociology, linguistics and many more. The concepts of graph theory are used in this work in the context of H-bond networks, where nodes represent H-bond donor/acceptor atoms or whole amino acid residues. Edges represent direct or water-mediated H-bonds.

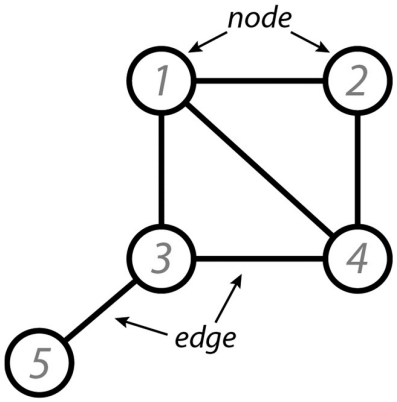

Figure 2.3. Schematic example of a connected graph. The nodes are labelled *1* to *5*. Edges connect the nodes with one another.

Centrality

Centrality is a set of measures used in graph theory and network analysis to assign a rank or a weight on nodes of a graph, based on their importance. Importance can be relatively hard to define because it would be related to the problem at hand. It can be conceptualized as the relation to the flow [177] of the graph or its cohesiveness [178]. There are many types of centrality measures, but in this work only the *betweenness* [179, 180] and the *degree* centralities will be discussed and applied. The betweenness centrality (BC), originally published for psychological analysis and social networks [179] finds many applications to this date, in many fields because it can express the role of nodes in the communication of the network. The BC measure is based on the shortest path problem

of graph theory. In a connected graph, given two random nodes, there will always be a path that bridges them. The shortest path is the path (or paths) that connects the given nodes with the minimum number of edges. In other words, the sum of edges of the shortest path(s) is minimized. For a node to have a high BC value, it means that this node acts as a bridge for many other shortest paths, passing through it. That node contributes highly to the communication of the network, and it is expected that the network will be highly perturbated or even disconnected with its removal. The degree centrality (DC) is a simpler concept of graph connectivity, as compared to BC. It expresses the direct connectivity of every node in a connected graph. The DC of a node is the number of edges that are attached to that node.

For both BC and DC computations, the NetworkX [181] package for the Python programming language was used. The BC of node *i* of a connected graph is given as the fraction of all shortest paths that pass through that node (*i*) for every node pair (*a*, *b*) found in the graph.

$$BC(i) = \sum_{a,b \in V} \frac{\sigma(a, b|i)}{\sigma(a, b)} \qquad 2.4$$

The number of shortest paths from node *a* to node b that pass through node *i* is given by $\sigma(a, b|i)$, the number of all shortest paths from *a* to *b* is given by $\sigma(a, b)$, and the set of nodes is given by *V*. BC can be normalized by a factor of $2/(N-1)(N-2)$, where *N* is the number of nodes in the graph. The algorithm to compute the BC is Brandes' algorithm [182].

The *DC* of a node *i* is the sum of its adjacencies [180, 183].

$$DC(i) = \sum_{j=1}^{N} \alpha(i, j) \qquad 2.5$$

The term $\alpha(i, j)$ assumes the value of 1 only when the nodes $i,\ j$ are connected with an edge, otherwise it assumes the value 0. The formulation above is independent of the graph size. The DC can be normalized for the size of the graph with the factor $N-1$ with the complete expression shown in eq. 2.6:

$$D\acute{C}(\imath) = \sum_{j=1}^{N} \alpha(i, j) \Big/ N - 1 \qquad 2.6$$

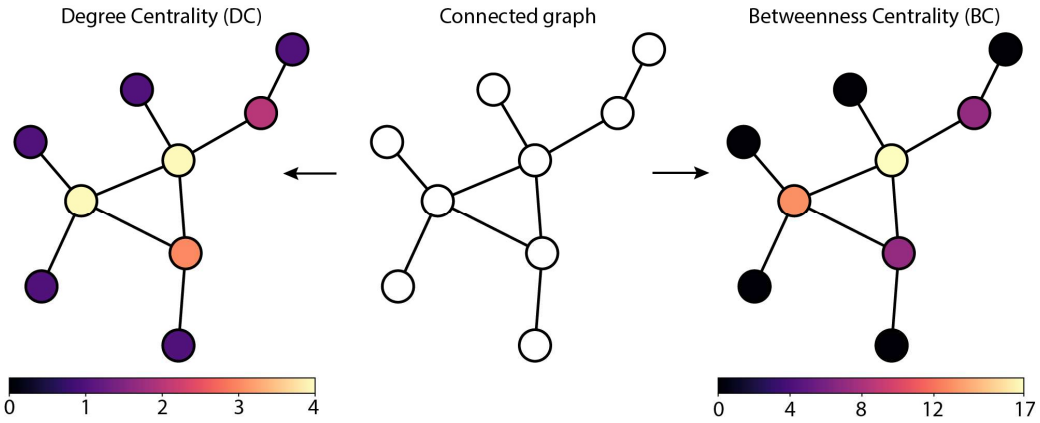

Figure 2.4. Schematic example of the BC and DC measures on an arbitrary connected graph. In the middle is the starting graph. On the left side a DC computation is shown and on the right side a BC respectively. Both computations are unnormalized and shown on a perpetually uniform color scale.

Using a graph-based approach to analyze H-bond networks offers the possibilities of using multiple tools from graph theory to bring new insights and new perspectives to H-bond analyses. Specifically considering the centrality measures where relatively important nodes for the networks can be identified. When amino acid residues are represented by nodes in a graph, by employing the BC and DC measures, critical amino acid residues of the connected H-bond network can be detected. This is especially important in the prediction of possible sites for point mutagenesis. As of this date there is no clear criterion for what could make a good mutation candidate and most of the decisions are met with empirical criteria and are mostly knowledge or experience based. The concept of applying BC and DC measurements in analyses of dynamic H-bond networks was introduced in our laboratory by Konstantina Karathanou [184] in analyzing complex H-bonds sampled in the protein motor SecA of the Sec protein secretion pathway found in bacteria [184]. Inspired by the capabilities of this analysis, BC and DC functions were implemented in H-bond analysis algorithm package, Bridge [42] (discussed in chapter *"Channelrhodopsin C1C2"*) and they were used in analyses of AntR to provide insight on the importance of S74 upon retinal isomerization [185] (discussed in chapter *"Antarctic Rhodopsin"*).

Unique Shortest Paths (USP)

In a previous study which will not presented in this work [186] we introduced a new measure named Unique Shortest Paths (USP) that reports the number of unique shortest paths. Betweenness centrality (BC) computations on graphs of H-bonds of biomolecules can lead to redundancy of shortest pathways. Consequently, the BC values can pose a difficulty when it comes to interpreting them in terms of their biological meaning. The USP computation will measure only the longest shortest paths between pairs of nodes. All measured paths are unique and will never be a part of (or belong to)

another path (Figure 2.5). In Figure 2.5 a schematic representation of a USP calculation is presented on a random H-bond network and an isolated pathway within the graph ($a \rightarrow h$). The USP value of node *e* i.e., the number of unique shortest paths going through node *e* is 1. Subpaths $a \rightarrow f$ (magenta) and $a \rightarrow g$ (cyan) are included in the longer $a \rightarrow h$ (black) and are not considered.

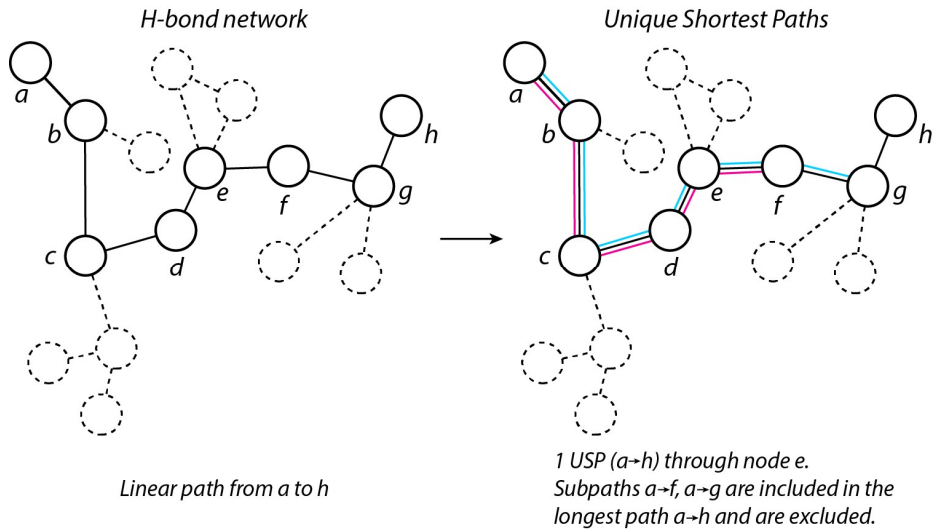

Figure 2.5 Unique shortest paths (USP) definition scheme. Starting from a random H-bond network (left), and considering only the path *a* to *h*, the USP computation will search for the longest shortest path to bridge nodes *a* and *h*. For demonstration purposes the USP(*e*) will be discussed. The USP value of node *e* i.e., the number of unique shortest paths going through node *e* is 1. Subpaths $a \rightarrow f$ (magenta) and $a \rightarrow g$ (cyan) are included in the longer $a \rightarrow h$ and are not considered. In contrast, a BC computation would include subpaths $a \rightarrow f$ (magenta)and $a \rightarrow g$ (cyan) as shortest paths going though node *e*. Adapted from ref. [186].

In comparison the BC measure will count all shortest paths between pairs of nodes in a graph (Figure 2.5, Figure 2.6), thus it would include subpaths $a \rightarrow f$ (magenta) and $a \rightarrow g$ (cyan), as shortest paths going though node *e*. USP allows intuitive interpretation in the role of a node in an H-bond graph. In this work I will be applying the Unique shortest paths algorithm that we developed [186]. Similarly to the BC, the USP is applied to an H-bond graph and returns a value per node which is an amino acid residue in our case (Figure 2.6) [186].

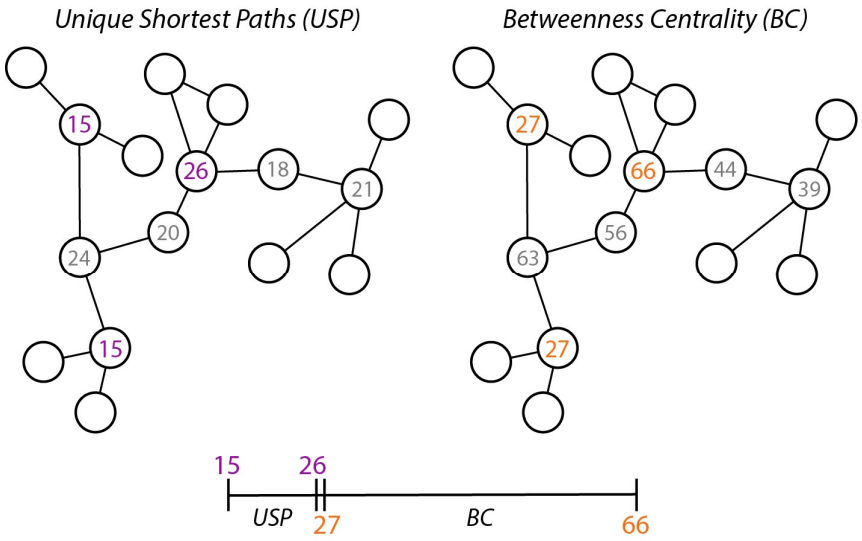

Figure 2.6 Schematic representation of USP vs. BC values for an arbitrary H-bond graph. The USP values range from 15 to 26 (purple), while the BC values range from 27 to 66 (orange), for the same graph.

Length of the shortest paths computations

USP is a valuable tool to understand the connectivity of a node in a graph and the likeliness of high disturbance of that graph when the node is removed. But it does not provide insight on the extent of that connectivity. For the work presented in Chapter 5 *"Conserved H-bond motifs in membrane transporters"*, in the context of H-bond motif detection I have constructed an analysis to compute the lengths of the shortest paths that involve groups of H-bond motifs (Figure 2.7, Figure 2.8). The amino acid residues of interest are isolated in search of the root node, e.g., Asp/Glu groups in Asp/Glu-Ser/Thr motifs. The root nodes are filtered so that only unique entries are analyzed. The topology of the root node in the graph determines the analysis that follows.

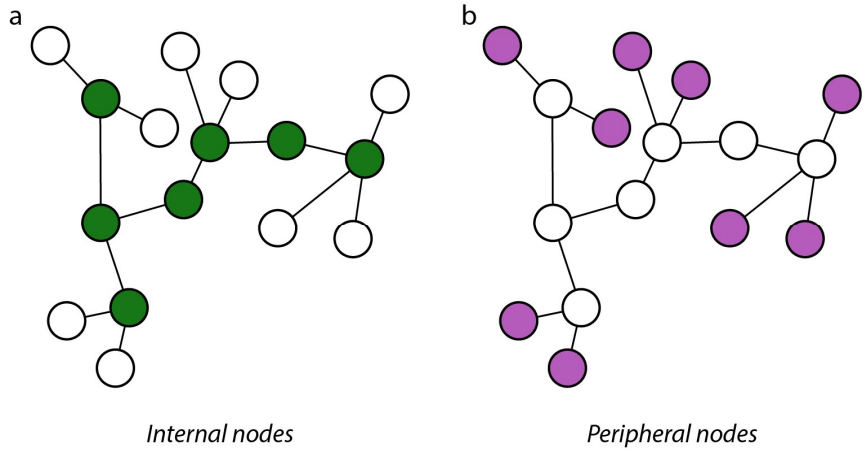

Figure 2.7 Schematic representation of the *Internal* vs *Peripheral* node topologies in an H-bond graph. H-bond paths can pass through internal nodes, but they can only begin/end on peripheral nodes.

If the root is an internal node (Figure 2.7), then the number of shortest paths going through that root node is computed, the length of the longest path(s) is collected and the number of paths with that length are detected (Figure 2.8). If the root is a peripheral node (Figure 2.7, Figure 2.8), no shortest path can pass through it. The path(s) must begin or end in that node. In that case the number of shortest paths beginning from that root node is computed (Figure 2.8). Same as previously, the length of the longest path(s) is collected and the number of paths of that length are detected.

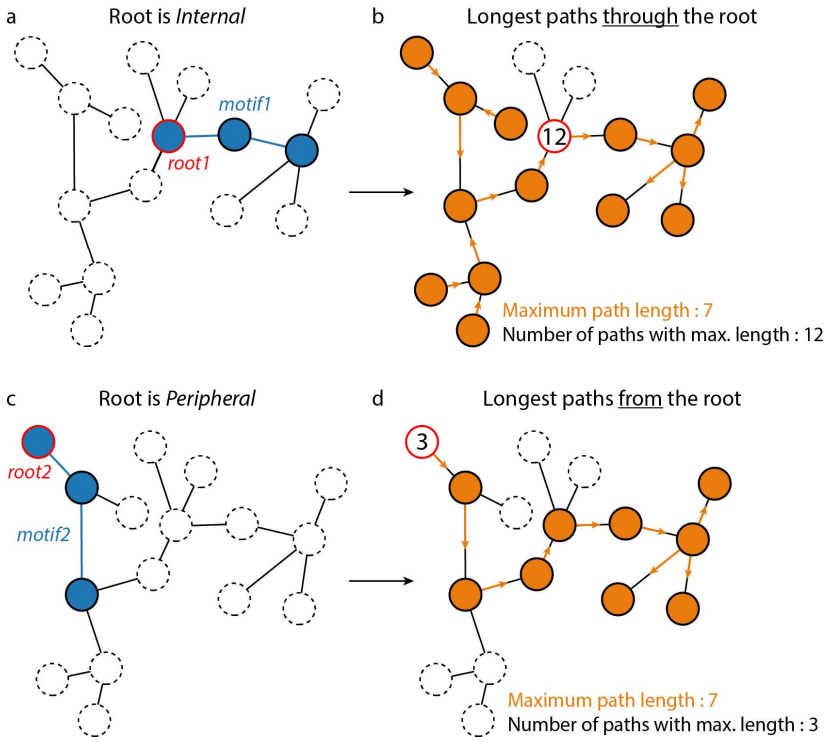

Figure 2.8 Schematic representation of a longest shortest path analysis from motif roots. An H-bond motif is detected in a set of crystal structures or trajectories. The root node is selected and

found to be internal (a) or peripheral (c) to the graph. The shortest paths passing through the root node (b) or beginning from the root node (d) are computed and the length of the longest of shortest paths as well as the number of unique paths of that length are collected. In panels b, d the arrows are shown for illustration purposes only and they should not be confused with directed graph computations.

In MD simulations of Aqy1 and ChR2 the motifs detected for this computation are not filtered for their position along the membrane normal and the root nodes are not filtered to unique entries. It was found that different monomers of oligomeric structures can show different dynamics of H-bond networks, such as the network of H-bond in the retinal vicinity of the C1C2 chimera [42].

2.4 Programs/Software

Chemistry at HARvard Macromolecular Mechanics

Chemistry at HARvard Macromolecular Mechanics (CHARMM) is a one of the most widely known force fields [140] and one of the earliest molecular dynamics simulation software packages [145, 171]. Its first release was in 1983 and was presented as a very flexible, coherent, and efficient program to model, simulate and analyze biological molecules. It is still used to this date after several releases. In this work, the CHARMM program was used to model transmembrane proteins in lipid bilayer environments through its recent, in comparison, graphical interface CHARMM-GUI [187-191], released in 2008. CHARMM was also used to perform dihedral angle isomerization of the retinal chromophore of ChR2 and AntR via adiabatic mapping, followed by energy minimization.

NAnoscale Molecular Dynamics

NAMD [192, 193] was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. NAMD is one of the most widely known programs to perform molecular dynamics simulations. It offers parallelization in order to run large scale simulations efficiently, using supercomputing clusters. It is implemented in the C++ language and is compatible with the CHARMM force-field (potential energy function), parameters and file formats. NAMD uses the stochastic Langevin equation (eq. 2.7) [193, 194] to generate the canonical (NVT) ensemble and a natural extension of the Verlet method to integrate the equations of motion, named the Brünger-Brooks-Karplus method [195].

$$M\dot{v} = F(r) - \gamma v + \sqrt{\frac{2\gamma k_B T}{M}} R(t) \qquad 2.7$$

In the Langevin equation 2.7 above γ is the friction coefficient, M the mass, v the velocity and R an univariate Gaussian random process [193]. To generate the NPT ensemble it uses a modified Nosé-Hoover method, which is a combination of the Nosé-Hoover method for constant pressure [196] and piston fluctuation control [197]. For the electrostatic interactions the smooth particle-mesh Ewald (SPME) [198] method is employed.
MD trajectories in this work were generated with NAMD in the high-performance computing (HPC) cluster Curta [199] of the Zentral Einrichtung Datenverarbeitung (ZEDAT)-Freie Universität Berlin, the Cray system of the Norddeutsche Verbund für Hoch- und Höchstleistungsrechnen HLRN-III in the Zuse-Insitut Berlin (ZIB), as well as the HLRN-IV systems of the Georg-August-Universität Göttingen and Zuse-Institut Berlin.

Visual Molecular Dynamics

Visual Molecular Dynamics (VMD) [52] is a sister program of NAMD [193] for molecular graphics, in which visualization, small simulations, systems setups, and analyses can be performed. It features a powerful atomic selection syntax with Boolean operators, spatial and regular expressions. VMD features a complete graphical interface, as well as an interface with the Tcl scripting language, allowing for powerful scripting capabilities, specifically for analysis purposes. VMD was used in this work to visualize molecules, generate graphics, and perform analyses using Tcl scripts. To render molecular graphics the Tachyon Ray Tracing library [200] was used, within VMD.

OPM database/PPM server

In this study the OPM database and the PPM webserver are used to orient proteins along the membrane normal, in addition to the VMD "Alignment to principal axes" package. The OPM/PPM are freely accessible resources that provide spatial coordinates for known, solved three-dimensional structures of proteins, positioned in a lipid bilayer environment and orientation in regard to their cellular localization [201]. A short summary of the principles of the placement of proteins in a lipid environment is given below. In their study, Lomize and co-workers presented a methodologically improved model [202] that was implemented in their program for predicting energetically favorable orientation of various macromolecules in model membranes, ranging from integral to peripheral proteins and from peptides to small organic molecules. The new implicit solvent model treats a solute (macromolecule) as a rigid body and the lipid bilayer as an anisotropic fluid / binary mixture of a polar aqueous and a non-polar lipid phase and is

based on the energy transfer profiles from solvent to an arbitrary position in the model membrane, of amino acid residues in regard to their position along the membrane normal [202]. The membrane itself is modelled to have varying parameter profiles along the membrane normal. For every amino acid-residue the $\Delta G_{transfer}$ is computed by placing it in an a-helix or b-barrel consisting by Ala residues, perpendicular to the model membrane moving along the membrane normal in a 1 Å step, while adjusting the bilayer thickness for a-helical and b-barrel integral proteins, respectively [202]. The $\Delta G_{transfer}$ for the standard amino acid residues after normalization to account for the over-exposure of simple α-helical model compared to a polytopic protein, also follows the biological hydrophobicity scale [203, 204]. The computational method was applied to over 1000 proteins from the OPM database, after exhaustive testing and calibrations against experimental energy profiles. Additionally, it offers significant improvements compared to the older version PPM 1.0 when it comes to the correlation of the experimental vs. computed free energy of membrane insertion for peripheral proteins. The correlation between the experimental and calculated free energy of transfer was increased from (R^2 = 0.47 / RMSE = 2.73 kcal/mol in PPM version 1.0 to R^2 = 0.78 / RMSE = 1.13 kcal/mol in the PPM version 2.0.

Chapter 3 Channelrhodopsin C1C2

This chapter is based on the following publication where Malte Siemers and I shared equal contribution as first authors:

Siemers, M., Lazaratos, M., Karathanou, K., Guerra, F., Brown, L.S. and Bondar, A.N., 2019. Bridge: A graph-based algorithm to analyze dynamic H-bond networks in membrane proteins. *Journal of chemical theory and computation*, *15*(12), pp.6781-6798.

I collaborated with Malte Siemers during this work and contributed to the development of Bridge under the guidance and supervision of Prof. Dr. Ana-Nicoleta Bondar. It should be disclosed that Malte Siemers wrote the Bridge analysis algorithm package in the Python programming language, following previous work where he collaborated with Federico Guerra [33], to develop the algorithm which detects water wires. Analysis functions of Bridge and basic concepts of its operation are presented here for consistency and clarity, but it should be disclosed they were implemented by Malte Siemers. I contributed to concepts of the analysis functions and how they should be implemented within the code and presented by testing code versions, bug reporting and providing feedback. I also helped in the design of the PyMol interface through feedback and reporting of bugs. The trajectories in the above publication and in this chapter are set up and ran by Prof. Dr. Ana-Nicoleta Bondar and they were then passed on to me to analyze. Konstantina Karathanou contributed to the above publication with the section *"Lipid-Protein H Bonds Connect the Membrane to the Internal H-Bond Network"*, using Bridge for the protein-lipid analysis. This section will not be discussed in this chapter. I performed all other analyses presented in the above publication, and subsequently, all analyses presented in this chapter, I contributed to the writing of the text and prepared the majority of the published figures under the close guidance and supervision of Prof. Dr. Ana-Nicoleta Bondar. Lukas Kemmler wrote the script I used to perform the STRIDE analysis for the trajectories.

Parts of the work presented in this chapter are originally published in the Journal of Chemical Theory and Computation. Published figures contain input from other co-authors with the most important input being that of Prof. Dr. Ana-Nicoleta Bondar. Figures and text originally published in the journal are modified in order to be presented in this chapter and will be noted with "Adapted from ref. [42]."

Doi: https://doi.org/10.1021/acs.jctc.9b00697
Author-directed link: http://pubs.acs.org/articlesonrequest/AOR-KP4PD4QI6BqCA6tN82wg

Existing software packages and programs for H-bond analyses (as described in subschapter *"Hydrogen Bonds"*, section *"Algorithm implementations"*), will only work with single frame structures using the PDB file format, with the exception of MDAnalysis. HBonanza can still perform an analysis with a trajectory in PDB format. In this chapter a new generation software that performs dynamic H-bond analysis and water mediated H-bond analysis for molecular dynamics simulations & static structures is described. The algorithm package features a wide variety of algorithm implementations, visualization routines and a graphical interface for one the most widely used molecular graphics platforms, PyMol [42, 104]. To illustrate the capabilities of the new algorithm package on identifying and characterizing dynamic H-bond networks in complex protein environments, I used trajectories from molecular dynamics simulations. As protein model systems the membrane protein, channelrhodopsin chimera C1C2 was chosen. Channelrhodopsins use a retinal moiety as a chromophore and are well studied through experiments and simulations. Their biological function involves changes in H-bonding and H-bond networks of selected protein groups. In Channelrhodopsin-2, the reaction cycle involves protonation-coupled protein conformational changes and opening of a pore that transports cations across the membrane [62].

3.1 Algorithms to detect H bonds and water wires

An H-bond between two groups is registered when the distance between the heavy atoms of the H-bond donor and acceptor, and the angle formed by the acceptor heavy atom, the hydrogen atom, and the donor heavy atom, are smaller than the user-defined value. Here the H-bond distance threshold was set at 3.5 Å and the H-bond angle threshold at 60°. Computing H-bonds based on geometric criteria involves computations of nearest neighbors, i.e., atoms within a certain distance from each other, and subsequently computations of the H-bond angles for nearest neighbor atoms. These computations are generally costly when the biomolecule is large and water-mediated H-bonds are included. Bridge makes efficient the computations of H-bonds by using k-d trees [205] to partition the spatial points and for the detection of nearest neighbors. Bridge is implemented in Python, which is generally slow for direct computations, and it relies on compiled libraries such as Numpy for high performance computations. For angle computations, the implementation of the Einstein sum convention from the Scipy python package is used. The angle criterion is an optional argument in the function, so that structures that do not contain any hydrogen atoms can still be analyzed without further processing. Selections are on which the analysis will be performed on, are translated from atom-parsing language into sets of coordinates using the MDAnalysis package [48, 49]. Static selections are used for the H-bond analysis since they are computationally efficient. The algorithm described in [33] is used to compute the water mediated connections or water wires. Water wires are defined as shortest connections between two protein groups that are mediated via H-bonded water molecules and the Dijkstra's algorithm [206] is used to find the shortest connections. The length of a water wire (L) is defined as the number of H-bonded water molecules that bridge two protein groups. The

maximum value of the wire length *L*=5 is chosen, since an H-bonded chain of five waters corresponds to an essentially zero probability of proton transfer between an acid and a base in aqueous solution [207]. The internal data structure of Bridge for the results of the H-bond or the water wire computation is Python dictionaries. Elements of graph theory are employed in order to visualize an interconnected 3-D structure, which is the protein, on a 2-D surface. Bridge does that by a principal component analysis (PCA) and uses the projection with the highest variance along the bases axes, resulting into the least number of overlaps of nodes in the graph.

3.2 Bridge functions and algorithm implementation

A unique characteristic of Bridge is that it uses the graph element as the foundation for further analyses of the H-bond networks in the system of interest. Below, key functions that were used to analyze H-bond networks C1C2 are summarized. Graphs can be curated in order to extract valuable information for the analyses and further analyze the networks they represent. Such curations will be referred to as *filters*. The first filter that is applies in the following work is the occupancy filter. H-bonds are sampled for a percentage of time during the trajectory and are represented via edges in a graph. The mean occupancy is stored in a Python dictionary and using the *filter_occupancy* function with the specified threshold value, H-bonds whose mean occurrence is higher than the threshold value are returned. In this way, stable H-bond connections can be easily identified in highly complex graphs. The next step in dissecting H-bond graphs is search for the sub-graphs of all connections starting from a single node. In proton transfer systems it is often shown that proton transfer processes take place via continuous pathways of H-bonds. Two protein groups that are located within distance significantly longer than of one H-bond might bridge via continuous pathways of several intermediate components. Each component can consist of a protein-protein H-bond, a protein-water H-bond, or a water-water H-bonded chain. In retinal proteins the first proton transfer reaction of the photocycle involves the retinal Schiff base, where the proton will be transferred to the counterion (another amino acid residue). Later in the photocycle, proton transfer reactions are observed between amino acid residues and such reactions can involve water molecules, for example in the case of bacteriorhodopsin [125, 208-210]. With this in mind, the *connected component* search was implemented in Bridge to identify all connections starting from a *root* node, which would be passed as an argument in the function. Nodes and edges that are not part of any connected component are excluded. A function to extract the shortest paths between a *root node* and an *end node* was also implemented, that searches for the shortest paths employing Dijkstra' algorithm [206]. Root and end nodes are connected via intermediate nodes that make up the pathways and that can be the case that there are multiple shortest paths can connect the root and end nodes. The functionality to extract single paths was introduced to isolate individual pathways of amino acid residues from complex networks, for the purpose of processing them further. When a path, or paths is the result of a filter, the timeseries of its individual components can be overlapped to determine the joint occupancy of the

complete pathway. This value, denoted as JO shows the percentage of time that the whole pathway is sampled for. A schematic example of root and intermediate nodes as well as pathway exampled in a 2-D graph form and molecular graphics is shown in Figure 3.1.

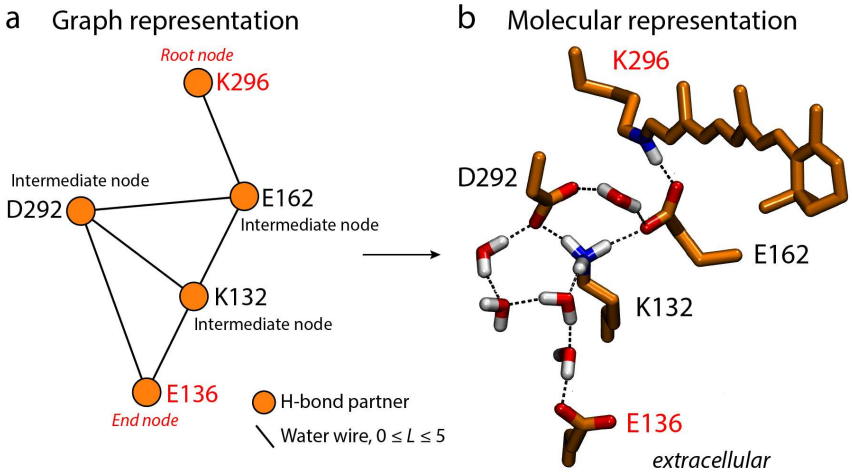

Figure 3.1. Schematic representation of a dissected H-bond network. (a) Graph representation of an arbitrary network, consisting of H-bond pathways to connect the root node K296 and the end node E136. Amino acid residues are represented as nodes and are connected via edges, which represent the H-bonds. Edges are water-mediated H-bonds (water-wires) with its length varying from $0 \leq L \leq 5$. The shortest paths connecting K296 and E136 feature the intermediate nodes E162, D292 and K132. (b) Molecular graphics of the graph shown in panel a. The root and end nodes are labelled with red font in both panels. Adapted from ref. [42].

The analysis protocols for an H-bond graph would include a prefilter of occupancy rates to 1% that is being applied to an unfiltered graph, to eliminate infrequent connections with no significant sampling importance, that could influence the results of the following filters. Alternatively, the occupancy threshold can be raised to values around ~50% or higher to analyze the frequently sampled H-bonds. A connected component analysis follows in order to determine connections that begin from a root node. In this chapter the root node would be the retinal Schiff base. Otherwise, a filter to determine the shortest paths between the root node and a target node would be applied instead. In the case that multiple H-bond paths are retuned, or there is a specific pathway of interest, single linear paths can be further isolated and analyzed independently. The order of the filters applied to a graph that represents an H-bond network plays a role in the output. The three primary filters that were applied in this chapter are summarized in Figure 3.2 below.

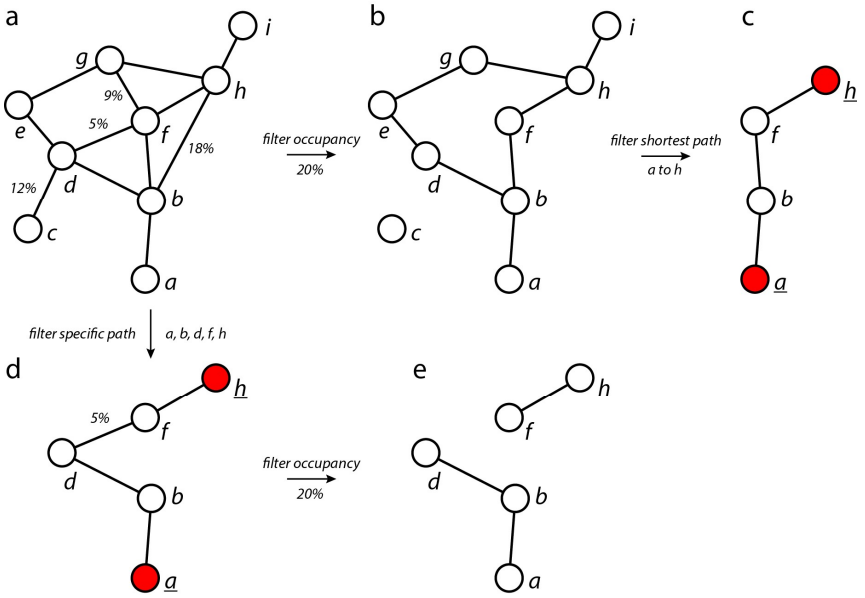

Figure 3.2. Schematic overview of a graph dissection procedure using the Bridge analysis algorithm package, showcasing three independent filters on an arbitrary H-bond network. (a) Arbitrary H-bond network with nodes *a*-*i*. Occurrence rates for the connections *b*-*h*, *c*-*d*, *d*-*f* and *f*-*g* are shown for the occupancy filter demonstration. (b) H-bond network originating from panel *a* after applying an occupancy filter of 20%. Node *c* is no longer part of the connected graph. (c) H-bond network showing the shortest paths between root node *a* and end node *h*. (d) H-bond network originating from panel *a* after applying a specific pathway filter, filtering for the path *a*-*b*-*d*-*f*-*h*. (e) H-bond networks originating from panel *d* after applying an occupancy filter of 20%. The connection *d*-*f* is sampled 5% of the time and is filtered out. There are no continuous pathways between *a* and *h* using the procedure of panels (*d*) and (*e*), while there is one pathway following the procedure of panels (b) and (c). Adapted from ref. [42].

I applied the newly developed capabilities of Bridge in simulations of C1C2 to explore the dynamic H-bond networks it hosts in its interior. Four simulation systems of C1C2 were prepared using the chimeric structure as the starting coordinates [99]. One simulation was prepared for the wild-type and three simulations for three point-mutations of functionally important amino acid residues. Namely, the mutants R159A, E162T and H288A were prepared, using the wild-type as the template. The missing atoms and amino acid residues of the crystal structure were generated using Modeller [211-213]. The crystal structure [99] contains 43 oxygen atoms of water molecules, and they were included in the starting coordinates of the setup. Hydrogen atoms were set up using the HBUILD command of CHARMM [171]. Three disulfide bridges link the dimer involving the amino acid residues C66, C73 and C75 [99]. Standard protonation states were applied for titratable amino acid residues, except for E162 which was considered deprotonated as suggested in refs. [214] and E122, E129 and D195 which were considered protonated as suggested in refs. [215-217]. Histidine amino acid residues were set up with the hydrogen atom on the Nε. A hydrated lipid bilayer of 576 POPC lipid molecules was employed to embed the protein in. Chloride ions added to neutralize the system's charge. The summary of the MD simulations performed for this chapter is presented in Table 3.1 and the complete system setup is shown in Figure 3.3.

Table 3.1. Summary of the MD simulations prepared and performed in this chapter. The simulation system, the mutation applied, number of atoms, temperature, the lipid bilayer environment, length of the simulation are reported in the table. Adapted from ref. [42].

Simulation	Mutation	Number of atoms	Temperature (K)	Lipids	Length (ns)
S1	Wild type	288,831	300	POPC	198.4
S2	E162T	278,831			215.4
S3	R159A	278,805			215.4
S4	H288A	278,817			215.0
Total					844.2 ns

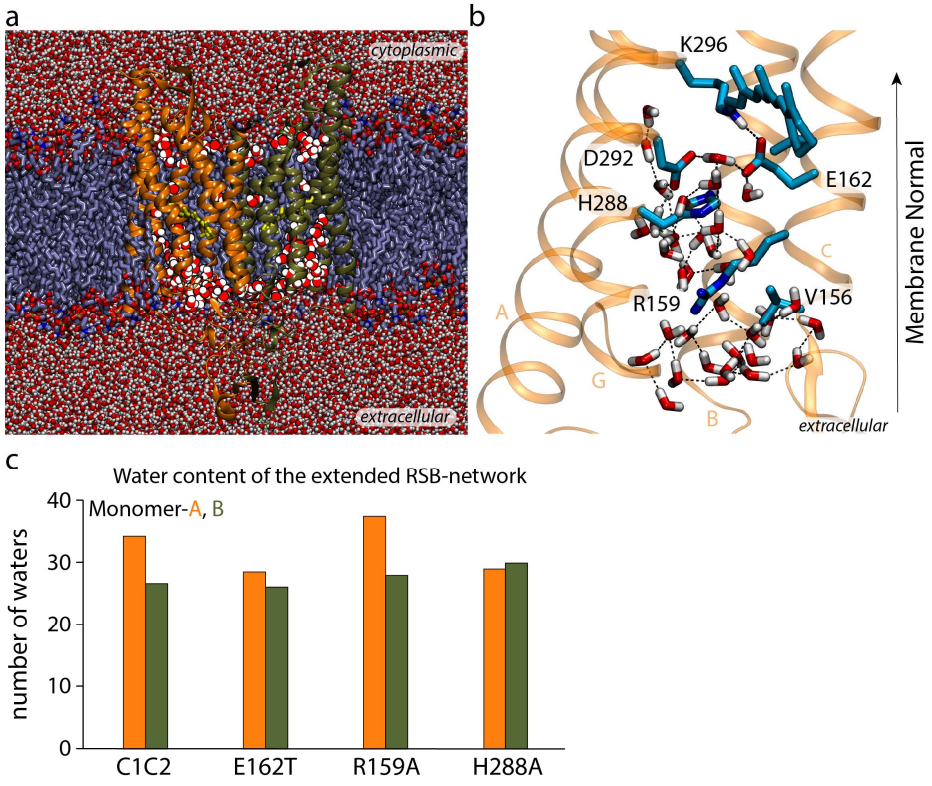

Figure 3.3 Molecular architecture of the C1C2 chimera. (a) Molecular graphics of the C1C2 chimera embedded in a hydrated POPC lipid bilayer. The protein in shown in Ribbons, colored per monomer. Monomer-A is shown in orange and Monomer-B in tan. The retinal chromophore is shown as yellow bonds. Internal water molecules are highlighted as compared to the bulk water shown in a vdW representation. (b) Extended H-bond network of the RSB. The network is shown for Monomer-A of the wild-type C1C2 simulation. The extended RSB network includes E162, D292, H288, R159 and V156, which is located within 1 helical turn from R159. V156 was included in the definition of the network to accommodate for the Arg to Ala mutation in the position 159. Water molecules within 6 Å of the protein groups of the extended RSB network are depicted. (c) Average number of water molecules of the extended RSB network across four

simulations of the Channelrhodopsin C1C2. Results for Monomer-A and Monomer-B are shown in orange and tan respectively. Additional data analyses for internal water molecules in C1C2 simulations are presented in Figure A.2. For clarity, only the H atom of the RSB is shown. Molecular graphics were prepared with VMD. Adapted from ref. [42].

3.3 H-bond network analyses of C1C2

Graphs of the retinal networks in simulations of wild-type C1C2

During the MD simulations of the C1C2 dimer at room temperature the protein structure remains stable, with an average Cα root-mean-squared deviation value of 1.85/1.81 Å (Monomer-A & B respectively) (Table 3.1, Figure A.1). Numerous water molecules visit transiently the interior of the protein: there are, on the average, 64.6/58.4 water molecules (Monomer-A & B respectively) in the interior of the protein (Table 3.2, Figure A.2). In the extracellular side specifically, the channel hosts between ~25 to 35 water molecules in the extended RSB network (Figure 3.3b, c). Studying C1C2 in this chapter, a special focus on the retinal Schiff base and its connectivity will be given. In C1C2 and other retinal proteins, the retinal Schiff base binds a proton in the inactive state of the protein, and functions as the primary proton donor group in the beginning of the photocycle. Particularly important protein groups close to the retinal Schiff base are D292, which corresponds to the primary proton acceptor of ChR2 [218] and, E162 which interacts directly with the RSB as indicated by FTIR [214]. E129 ($E90^{ChR2}$) is thought to undergo proton transfer reactions during the reaction cycle [215, 218].

Table 3.2. Internal water molecules in the inter-helical region of C1C2 trajectories. The number of water molecules is reported separately for Monomer-A and Monomer-B. Timeseries of internal water molecules are presented in Figure A.2. Adapted from ref. [42].

Simulation	Number of internal water molecules	
	Monomer-A	Monomer-B
S1	64.4±5.2	58.4±4.9
S2	58.8±5.1	62.9±5.7
S3	69.7±9.3	61.1±5.7
S4	61.1±5	59.5±5.7

Each of the C1C2 monomers contains numerous H-bonds, and they extend in complex networks (Figure 3.4a). Decomposing the initial graph using a connected component analysis with the RSB as the root node, pathways that connect the retinal Schiff base to E129 are identified. The direct H-bond network of the RSB in Monomer-A is rather local and disconnected from the rest of the protein, albeit rather stable (Figure 3.4b, c). The shortest path connecting the retinal Schiff base to E129 (Figure 3.4b, c, orange lines) is

found to pass via E162 and includes two segments: one segment connects the retinal Schiff base to E162, and the second segment connects E162 to E129 via K132, which shapes the energetics of SB deprotonation [219]. The joint occupancy for the path via the direct K132-E129 H-bond is sampled 4.7% of the time. The alternative pathway that is no longer the shortest, bridges K132 to the counterion D292 which in turn H-bonds to E129 with a joint occurrence rate of ~99%. On the other side of the network, T166 of helix C H-bonds for 98.5% of the time with the sidechain of the suggested [214] proton acceptor E162 (Figure 3.4b, c). Repeating the analysis for Monomer-B, the network is found to differ significantly from the one generated for Monomer-A. The RSB maintains a direct connection with E162, as indicated by FTIR [214]. Here, a connection from the RSB to E129 is not sampled, but another pathway to the extracellular side of the membrane is revealed. E274 is the equivalent of $E194^{BR}$ [99] which a part of the proton release group in BR [220]. The disconnecting point of the network is the H-bond between D292 and H288 which is sampled only 1.2% of the time. The shortest path would in turn continue to the gating arginine R159 with a 7.5% sampling rate and then directly to E274 at 22.5% (Figure 3.4c). Visualization of the pathways suggests that H-bonding would be possible (Figure 3.4d). The complete pathway RSB-E274 is rarely sampled at 0.2% of the time through direct H-bonds only (Figure 3.4c). The inclusion of water molecules is vital if the frequency that the RSB connects to EC side is to be quantified, through the ~25-35 water molecule it hosts (Figure 3.3c, Figure A.2e-h).

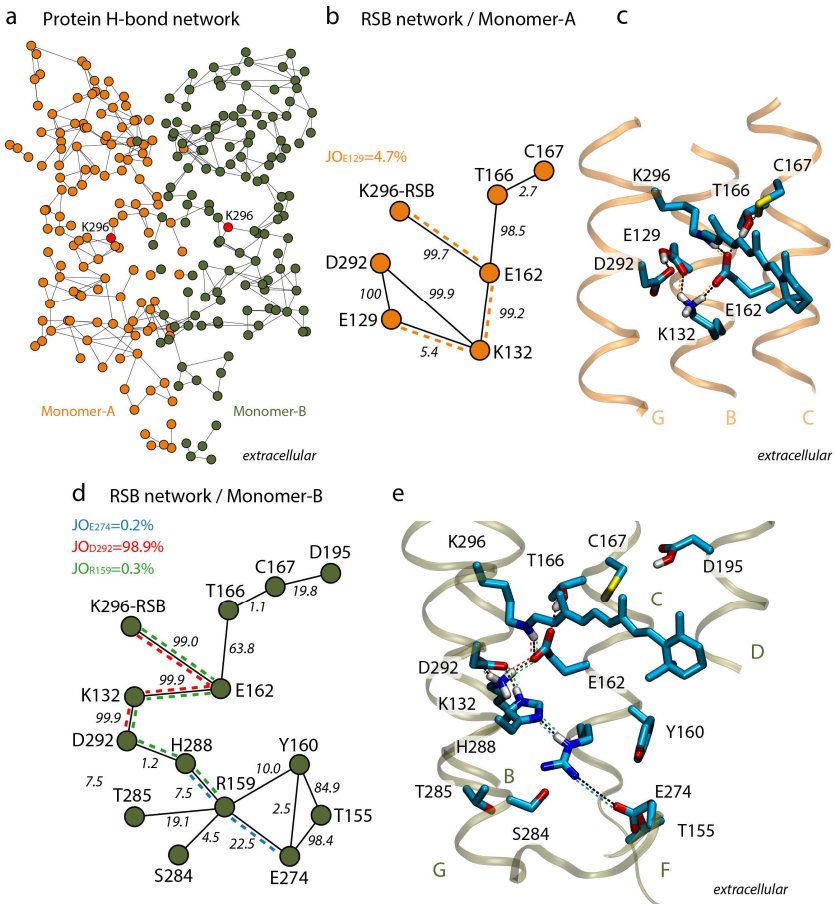

Figure 3.4. Generating H-bond graphs and identifying H-bond pathways using Bridge in C1C2. (a) H-bond graph of direct protein-protein H-bond network of the wild-type C1C2 simulation. Nodes are colored according to the monomer they belong to, orange for Monomer-A and tan for Monomer-B respectively. The node of K296-RSB is highlighted in red for both monomers. (b) Subgraphs of the direct protein-protein H-bonds network, featuring K296-RSB as the root note in Monomer-A. (c) Molecular graphics of the networks shown in panel *b*. The highlighted path K296-E129 is shown with orange dotted lines. (d) Subgraphs of the direct protein-protein H-bonds network, featuring K296-RSB as the root note in Monomer-B. (e) Molecular graphics of the networks shown in panel *d*. The highlighted paths K296-D292, K296-R159 and H288-E274 are shown in blue, red, and green dotted lines respectively. The occupancy (%) of each H-bond is annotated along every edge. The Joint Occupancy of the indicated paths in panels *b* and *d* are annotated in the respective color as dotted pathways. Adapted from ref. [42].

Water wires of C1C2

Allowing water molecules to mediate H-bonds between amino acid residues reveals that the dimer can host extensive connections that span through the length of the two monomers of C1C2. There are in total 2908 (1188 for Monomer-A, 1186 for Monomer-B & 534 wires connecting Monomer-A and -B, respectively) water-mediated bridges (Figure 3.5a), as compared to 318 (146 for Monomer-A, 145 for Monomer-B & 27 between Monomer-A and -B, respectively) connections in which protein groups interact directly with each other (Figure 3.4a). Most of the water-mediated bridges between protein groups are short lived, with occupancies of 0-10% (1710 connections). There are, however, 163 water wires with occupancies ≥ 90%. These water wires are spread within the protein and split into 3 zones. High-occurence wires are found in the extracellular part of the protein that is heavily exposed to the bulk, the extracellular part of the protein that is in the general vicinity of retinal Schiff base and lastly, the cytoplasmic region of the protein that is in contact with the bulk.

Table 3.3. Protein-protein and water-mediated H-bonds in C1C2 trajectories. The number of H-bonds and water wires are reported separately for Monomer-A and Monomer-B.

Simulation	Number of H-bonds		Number of water wires	
	Monomer-A	Monomer-B	Monomer-A	Monomer-B
S1	146	145	1188	1186
S2	135	142	1149	1217
S3	146	136	199	1150
S4	148	126	1149	1217

Water molecules can extend the H-bond networks and the long-distance bridging of the RSB to the EC side, effectively allowing the RSB to participate in stable and frequently sampled H-bond networks with amino acid residues important for proton transfer.

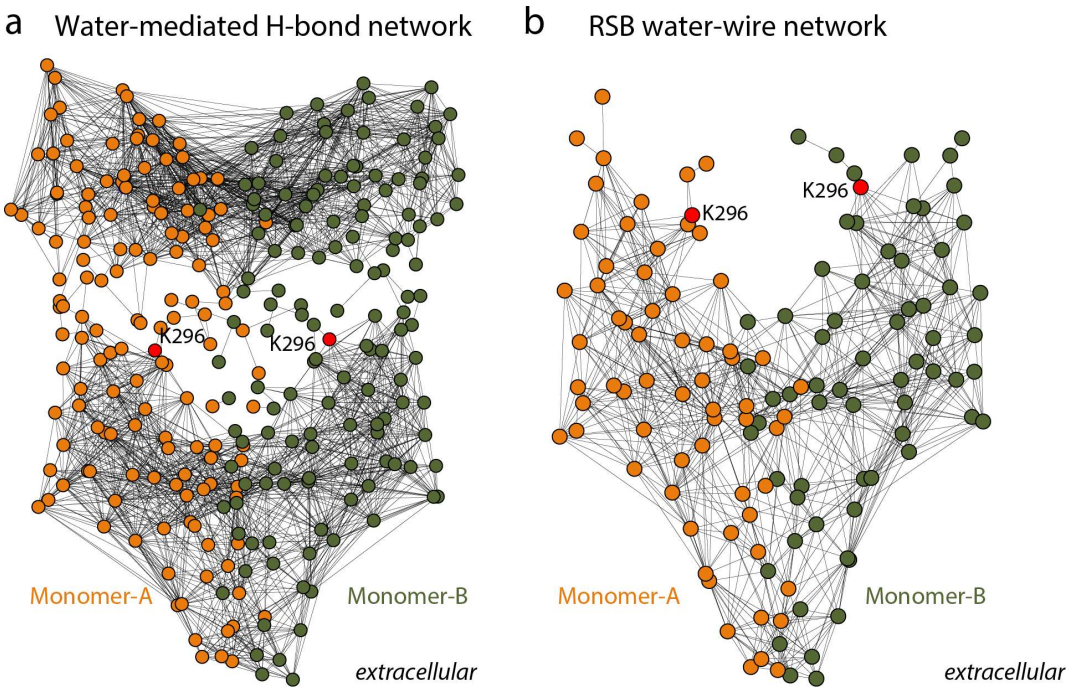

Figure 3.5. Computing water-wires and generating their graph representations using Bridge in C1C2. (a) 2-D graph of the water-mediated H-bonds of the wild-type C1C2 simulation. The graph is shown prior to any filter applied. (b) Subgraph of the water-mediated H-bonds, featuring K296-RSB as the root note in Monomer-A in the search. The RSB of Monomer-B is found in that subgraph as part of the network. The *connected component* search shown in this panel was applied to a pre-filtered graph of a minimum occurrence of 1%. Nodes are colored according to the monomer they belong to, orange for Monomer-A and tan for Monomer-B respectively. The node of K296-RSB is highlighted in red for both monomers. Adapted from ref. [42].

Using the shortest path computation with the RSB of Monomer-A as the root node, and E129 as end node, the RSB is found to bridge to E129 via two networks, each comprised of three segments (Figure 3.6). Both include a direct, protein-protein connection of the Schiff base to E162 as the first segment, and two water-mediated chains. E162 H-bonds to D292 and in turn to D292 H-bonds to E129. A noticeable difference is that in the water-mediated analysis the two counterions are connected via a water molecule, allowing for more stable networks and more connectivity options for the RSB. Allowing for water to mediate the H-bonds the E162-D292 connection is sampled 99.2% of the time. The calculated joint occupancy of the RSB-E162-D292-E129 pathway is computed to be 99.2%, thus, the retinal Schiff base bridges to E129 at all times via direct protein-protein and protein-water H-bonds. The alternative pathway bridges E162 to K132 and in turn to E129 (Figure 3.6a, b). The H-bond K132-E129 is sampled slightly more with the introduction of water molecules in the calculations from 5.4% to 14.4%. The pathway via K132 is sample significantly less frequently sampled at 14.3% of the time. In Monomer-B, the shortest bridges are composed of short-lived individual wires, in comparison to Monomer-A. Although, not being the shortest pathway, the same pathway to connect the RSB to E129 was identified (Figure 3.6a, b), with the individual

connection of D292-E129 (18.9%) being the determining factor for the joint occupancy of the whole path (17.2% vs 99.2% for Monomer-A).
Comparing the orientation of E129 in Monomer-A vs Monomer-B (Figure 3.6) provides insight for this significant difference between the connectivity of the monomers. E129 can further extend through water-mediated H-bonds to form two bistable H-bonds E129-N297 (47.9%) and D292-N297 (61.9%) in Monomer-A. The corresponding connections in Monomer-B show significant differences compared to Monomer-A, where E129 is constantly connected to N297, making the bridging of E129-D292 far less frequently sampled (E129-N297 99.9% and D292-N297 1%). E129 and N297 are directly H-bonded (Figure 3.6c, d) and this has an effect on the orientation of E129 in relation to D292. N297 is a conserved amino acid residue [100] of the central gate [70, 99] which is formed together with E129 and S102. Point mutations of the equivalent asparagine in ChR2 ($N258^{ChR2}$) shows that the photocurrent is largely abolished when mutated to N258V, and the conductance of sodium/potassium ions is reduced when mutated to N258Q [221]. The local H-bonding cluster of the SB further extends to E136 via protein-water H bonds (Figure 3.7). The E136A mutant shows a reduced photocurrent [99], and E136 is thought to interact with hydrated cations that are passing through the channel [222].

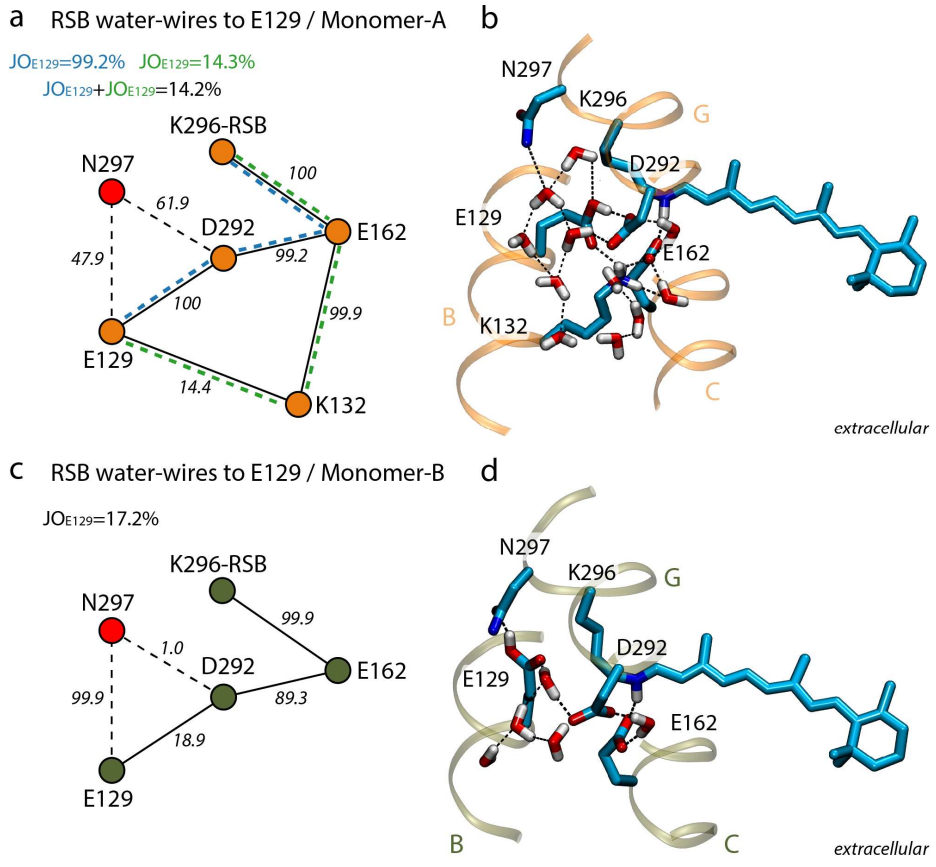

Figure 3.6. Water-mediated connections of the RSB to E129 in simulations of wild-type C1C2. (a) Shortest-paths connecting the RSB to E129 in Monomer-A. Pathways to E129 via D292 and K132 are colored in blue and green dotted lines respectively. Their JO values are annotated in the same colors, along with the JO value of the combined blue and green paths. The node of N297

is shown in the graph as a branched connection, since is not part of the shortest-path computation. (b) Molecular graphics for the H-bond pathways shown in panel *a*. The orientation of E129 in relation to N297 favors water-mediated connections between N297-E129 and N297-D292. (c) Shortest-paths connecting the RSB to E129 in Monomer-B. The shortest paths computation shows a consecutive series of H-bonds from K296 to E129. The node of N297 is shown in the graph as a branched connection, since is not part of the shortest-path computation. (d) Molecular graphics for the H-bond pathways shown in panel *c*. The orientation of E129 in relation to N297 favors direct connections between N297-E129. Graphs were pre-filtered to a minimum occupancy of at least 10% before being subjected to the *all-paths* filter. The occupancy (%) of each H-bond is annotated along every edge. Adapted from ref. [42].

With the addition of H-bonded water molecules in the network computations the frequency that the retinal Schiff base connects to carboxylate groups on the extracellular side of the protein increased significantly (Figure 3.7, Figure 3.8). Inspired from bacteriorhodopsin, a well-studied and model system for retinal proteins, H-bond pathways to carboxylate groups on the extracellular side of the membrane were explored, that could serve as potential proton release groups. Both E136 and E274, are located in the extracellular part of the protein in close contact with the bulk and the former's position is the equivalent of $E194^{BR}$. Thus, it could be suggested that they could act as possible groups that could release the proton to the bulk.

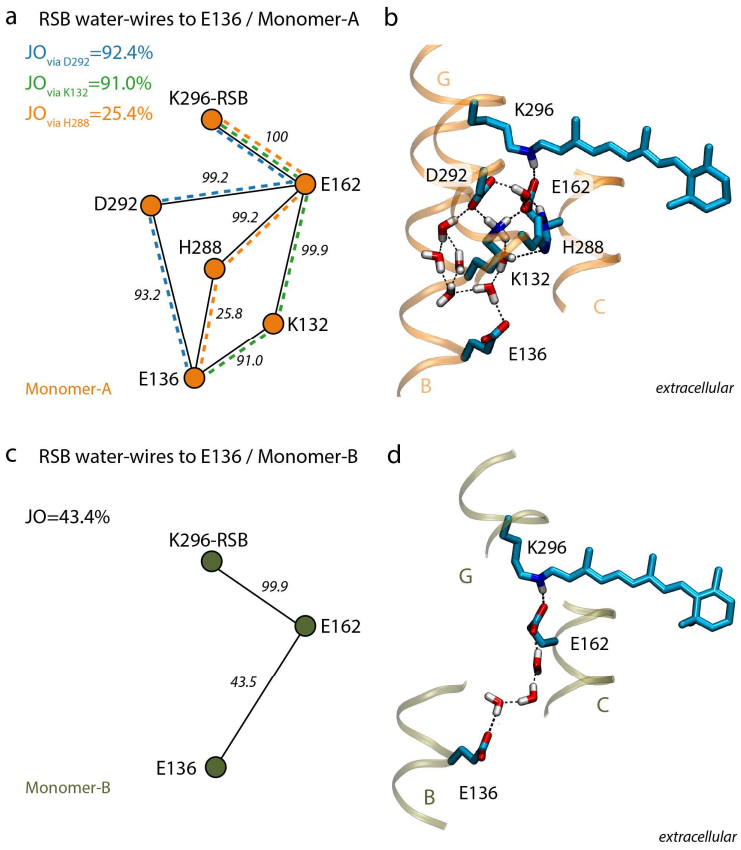

Figure 3.7. Water-mediated connections of the RSB to the extracellular side of the membrane in simulations of wild-type C1C2. (a) Shortest-paths connecting the RSB to E136 in Monomer-A. Pathways to E136 via D292, K132 and H288 are colored in blue, green, and orange dotted lines

respectively. Their JO values are annotated in the same colors. (b) Molecular graphics for the H-bond pathways shown in panel *a*. (c) Shortest-paths connecting the RSB to E136 in Monomer-B. The shortest paths computation shows a consecutive series of H-bonds from K296 to E136 via E162. (d) Molecular graphics for the H-bond pathways shown in panel *b*. Graphs were pre-filtered to a minimum occupancy of at least 10% before being subjected to the *all-paths* filter. The occupancy (%) of each H-bond is annotated along every edge. Adapted from ref. [42].

In Monomer-A the H-bond networks are extended to connect to E136 via 3 alternative shortest paths. The pathway branches after the connection RSB-E162 into 3 subpaths, via D292, or H288 or K132 to end up on E136 (Figure 3.7a, b). The pathways via K132 and D292 are sampled in very high occurrence rates (91-92.4%, green and blue highlights, Figure 3.7a, b). The determining factor for the lower sampling rate of the pathway via H288 is the H-bond H288-E136 which sampled only 25.8% of time. In Monomer-B, E162 can bridge to E136 through H-bonded water molecules, effectively enabling a shorter pathway to be sampled, albeit at a lower rate - 43.4% (Figure 3.7c, d).

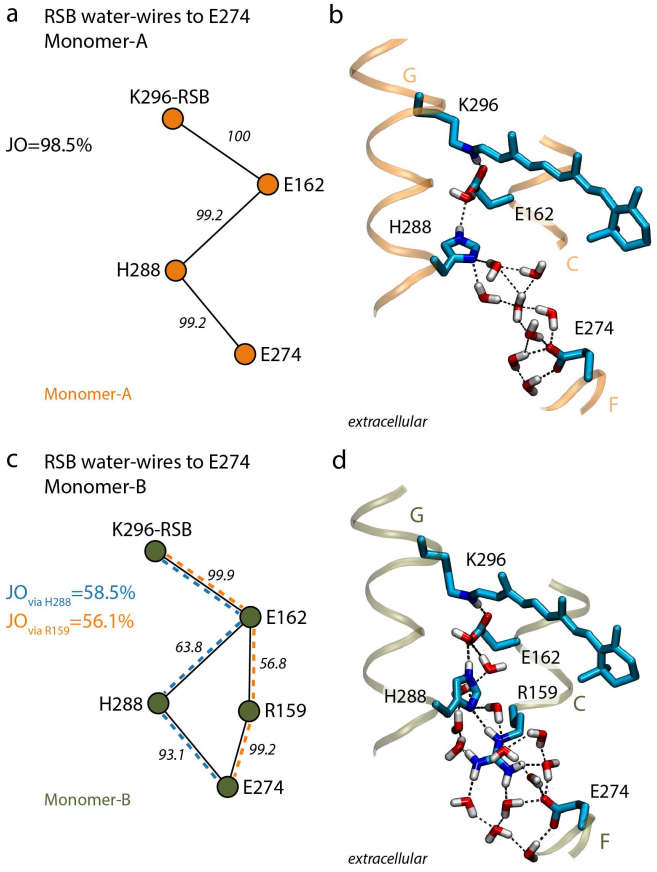

Figure 3.8. Water-mediated connections of the RSB to the extracellular side of the membrane in simulations of wild-type C1C2. (a) Shortest-paths connecting the RSB to E274 in Monomer-A. The shortest paths computation shows a consecutive series of H-bonds from K296 to E136 via E162 and H288. (b) Molecular graphics for the H-bond pathways shown in panel *a*. (c) Shortest-paths connecting the RSB to E274 in Monomer-B. Pathways to E274 via R159, and H288 are colored in blue, and orange dotted lines respectively. Their JO values are annotated in the same colors. (d) Molecular graphics for the H-bond pathways shown in panel *b*. Graphs were pre-

filtered to a minimum occupancy of at least 10% before being subjected to the *all-paths* filter. The occupancy (%) of each H-bond is annotated along every edge. Adapted from ref. [42].

The connection of the RSB to E274 in Monomer-A features one shortest pathway through E162-H288-E274 which is very frequently sampled at 98.5% (Figure 3.8a, b). The equivalent connection in Monomer-B features a branching of two sub-paths with the RSB-E162 being a common connection. In turn E162 can connect either to H288 or to R159 and then E274. Both alternative pathways are sampled at similar rates 56.1-58.5% (Figure 3.8c, d), with the determining factor the H-bonds between E162-H288 and E162-R159.

An extended H-bond network to bridge the two proton donors

The decomposition of the water wire graph (Figure 3.5) indicated that the retinal Schiff base of Monomer-A connects to the retinal Schiff base of Monomer-B (Figure 3.9a, b). The extended network features 46 amino acid residues and numerous water molecules (Figure 3.9a). After prefiltering to show only high occupancy H-bonds of at least 50% the network reveals four extended pathways of 12 H-bonds each, that connect the two RSBs (Figure 3.9b). The joint occupancy of the retinal-retinal H-bond networks is relatively low, in the range 9.4-10.5%, but this stands for the complete pathways to be sampled as whole. For this kind of extended network, it is suggested that the likelihood of complete pathways to be sampled at these occurrence rates is a find of great significance [42]. From the timeseries, (Figure 3.9) the different paths show a similar trend in terms of their likelihood to be present for the trajectory segment at hand. The H-bond networks that connect the retinal Schiff base to the extracellular half includes protein groups known to be essential for the functioning of C1C2. The E162T mutant shows faster kinetics and reduced photocurrent amplitude [223] as well as much reduced light sensitivity as reported in [224]. R159A leads to a reduction of the photocurrent produced as reported in [99, 225].

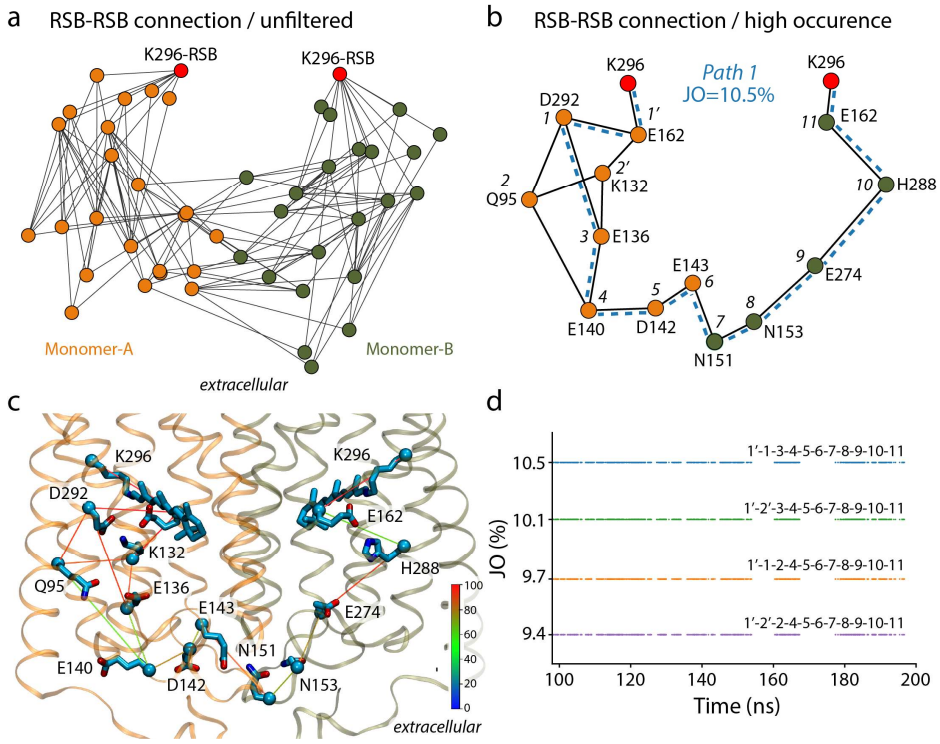

Figure 3.9. The retinal Schiff bases of Monomer-A and Monomer-B are connected in the C1C2 simulations through an extended water-mediated H-bond networks of the extracellular side. (a) Graph representation of the RSB-RSB H-bond network. The nodes of K296-RSB are shown as red nodes for both Monomers. (b) High occurrence RSB-RSB H-bond network where the individual connections are pre-filtered to 60% minimum occurrence. Path-1 that consists of the nodes 1'-1-3-4-5-6-7-8-9-10-11 is highlighted in blue dotted lines. (c) Molecular graphics of Path 1 shown in panel b. Lines represent the water-mediated connections and are colored according to their occupancy using a BGR color scale. For clarity the lines connect the amino acid residues to their backbone C atom. (d) Timeseries of the four pathways (1-4) that bridge the RSB of Monomer-A to the one of Monomer-B in the wild type C1C2 simulations. The components that make up the pathways are annotated above their respective trendline. Adapted from ref. [42].

Prominent contributors to this network are the carboxylate counterions E162 and D292, K132, H288, and carboxylates that face the extracellular bulk (Figure 3.9b, c). Repeating the analysis to search for pathways that bridge the retinal Schiff bases of Monomer-A and Monomer-B across reveals a main difference between C1C2 and the mutants (Figure 3.10Figure 3.9). All the RSB-RSB bridges across the different simulations performed are illustrated in a schematic representation (Figure 3.10aFigure 3.9). C1C2 differentiates from the mutants in the sense that the RSB-RSB bridges include entirely different amino acid residues aside the counter ion E162. On the contrary, all 3 mutations feature the almost the same amino acid residues that comprise the RSB-RSB pathways. In wild-type C1C2, the extended H-bond network that can transiently connect the two retinal Schiff bases (Figure 3.9) includes the three protein groups that were mutated for simulations S2-S4. Each of the three mutations are found to cause significant rearrangements of the retinal-retinal network, even though common components of the networks can appear through the different simulations (Figure 3.10, Figure 3.11). In the case of the long-lived

individual water wires of E162T, it can be deduced that there are more components comprising the pathways, thus multiple combination of nodes to link retinal in Monomer-A to Monomer-B. The number of components is reduced in the R159A model, while the joint occupancy of the pathways is increased (Figure 3.10d, paths 11 to 16, Figure 3.11) and finally in the H288A model there are two pathways, where a branching is observed in node 10 or 10', to link the two RSBs. The joint occupancies of the pathways are summarized in a histogram (Figure 3.10d), where the effect of the mutations is shown. In the case of S4 an increase of joint occupancy of around ~4-fold is observed, where in the cases of S2 and S3 the increase in joint occupancy is significant compared to the wild-type, up to ~3-fold (Figure 3.10d) [42].

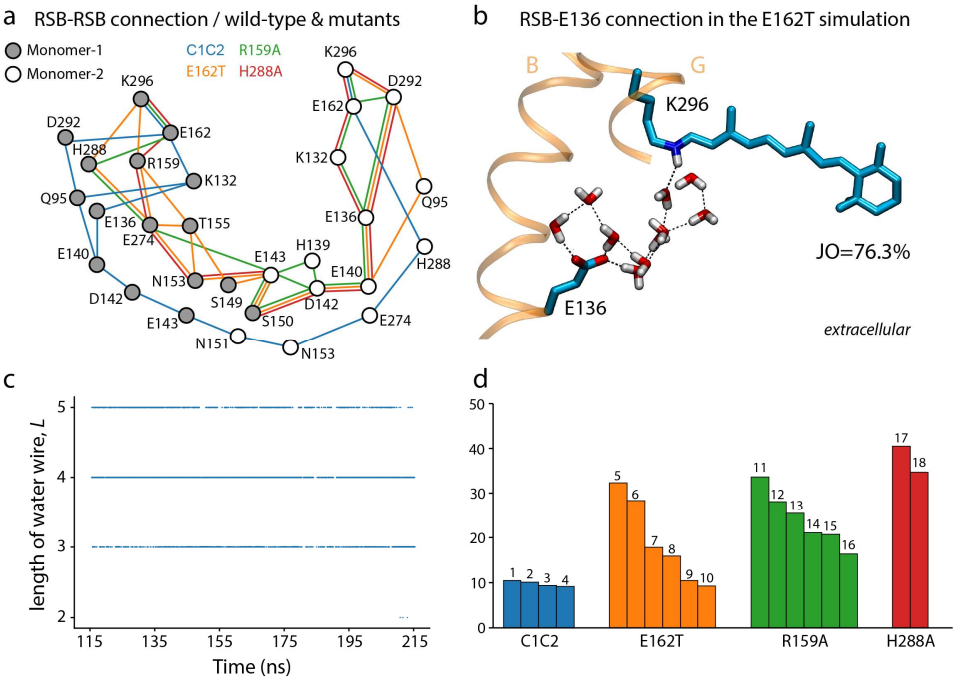

Figure 3.10. Mutations affect the dynamics of the extended RSB-RSB H-bond network in the extracellular side of C1C2. (a) Schematic overview of the extended RSB-RSB network across all four C1C2 simulations. Amino acid residues of Monomer-A are shown as gray nodes and of Monomer-B in white nodes. The nodes of K296-RSB are shown as red nodes for both Monomers. Pathways of the wild-type C1C2 are shown in blue lines, E162T mutant in orange, R159A mutant in green and H288A mutant in red. (b) The water-mediated H-bond bridging the retinal Schiff base to E136 in the E162T. The connection is sampled 76.3% of the time. (c) Timeseries of the water wire length of the RSB-E136 connection. The number of water molecules ranges from 2 to 5. (d) Occupancy histogram for the 18 high-occupancy paths connecting the retinal Schiff bases across four different simulations. The RSB-RSB networks for C1C2 are shown in Figure 3.9b and for the mutants are shown in Figure 3.11. Adapted from ref. [42].

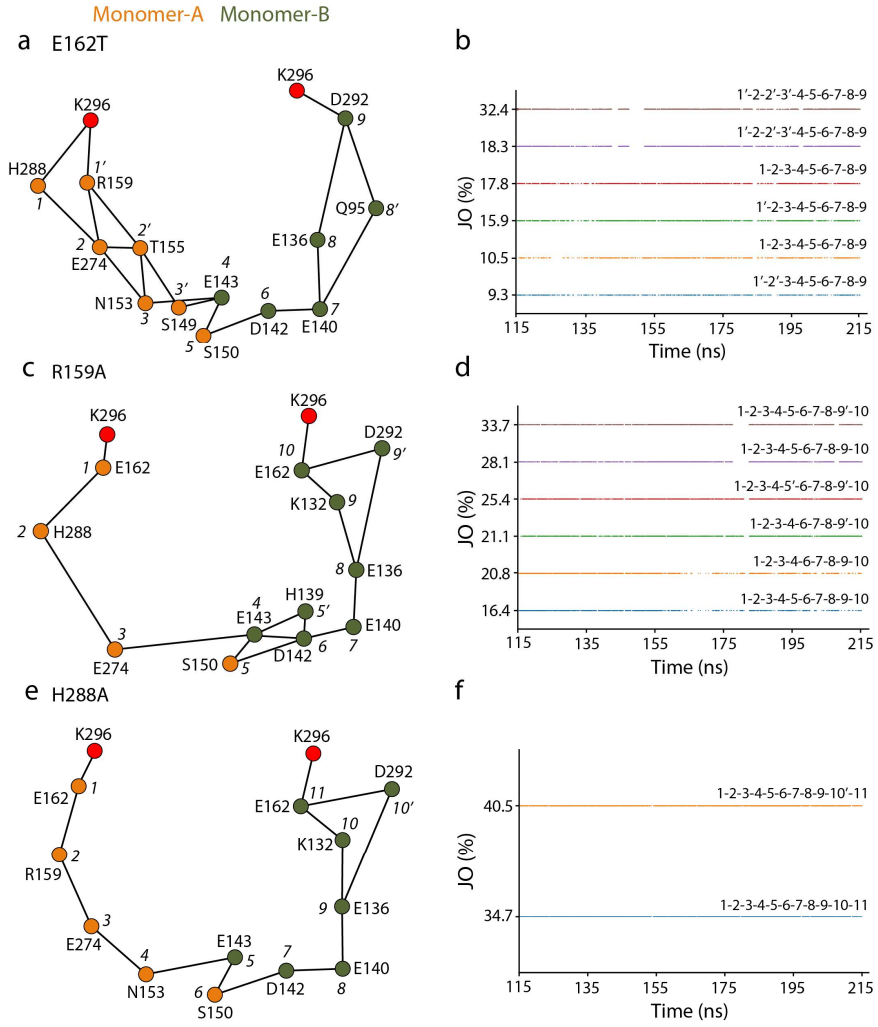

Figure 3.11. Extended water-mediated H-bonds connecting the retinal Schiff bases in C1C2 mutant simulations. Amino acid residues of Monomer-A are shown in orange nodes and of Monomer-B in tan nodes. The nodes of K296-RSB are shown as red nodes for both Monomers. (a, c, e) RSB-RSB H-bond networks sampled for the E162T mutant (a), R159A mutant (c) and H288A mutant (e) simulations of C1C2. For clarity, only high occurrence H-bond pathways are shown. (b, d, f) Occupancy timeseries for the pathways shown in the respective graphs. Every node except the K296-RSB is indexed. Primed numbers indicate an alternative route along the path. Adapted from ref. [42].

Mutations alter the dynamic networks of C1C2

To probe how C1C2 responds to mutations that alter H-bonding in its extracellular half, the dynamics of E162T (S2), R159A (S3) and H288A (S4) were studied. On the timescale of the S2-S4 simulation lengths, the overall structure of C1C2 remains stable, with relatively small values of the backbone RMSD, many water molecules visit the protein transiently and the fractions of secondary structure elements reach a plateau (Figure A.1, Figure A.2, Figure A.3).

To further access the effect of the introduced mutations a specific pathway for all four C1C2 variants was examined. When connecting to the EC, namely to E136 or E274, the RSB strongly H-bonds i.e., 100% occurrence rate, to the counterion E162. When E162 is mutated to threonine (S2), the retinal Schiff base can connect directly to the other counterion D292, and the pathways will shift (Figure 3.11a) towards the two most suitable candidates that make up the shortest paths to the EC, bridging to R159 and H288 via H-bonded water molecules (Figure 3.11). Additionally, the replacement of a glutamate with a threonine shifts the hydrogen bond network to D292, allowing water molecules to mediate a long-lived connection from the retinal Schiff base to the extracellular located E136 (Figure 3.10b). The length of the water chain in terms of water molecules varies from 3 to 5, with an average wire length value L=3.9 (Figure 3.10b). The water-mediated pathway from the retinal Schiff base to E136 reveals a JO=76.3%. In Monomer-B, where the retinal Schiff base bridges directly to D292, water-mediated connections via E136, E140 and D142 can lead to the extracellular side of the membrane where the protein comes in contact with the bulk (Figure 3.11a). Another aspect of different dynamics between E162T vs. wild type could be due the role of K132 in the networks. The E162T variant features more frequent sampling of extended H-bond pathways as compared to the wild-type C1C2.

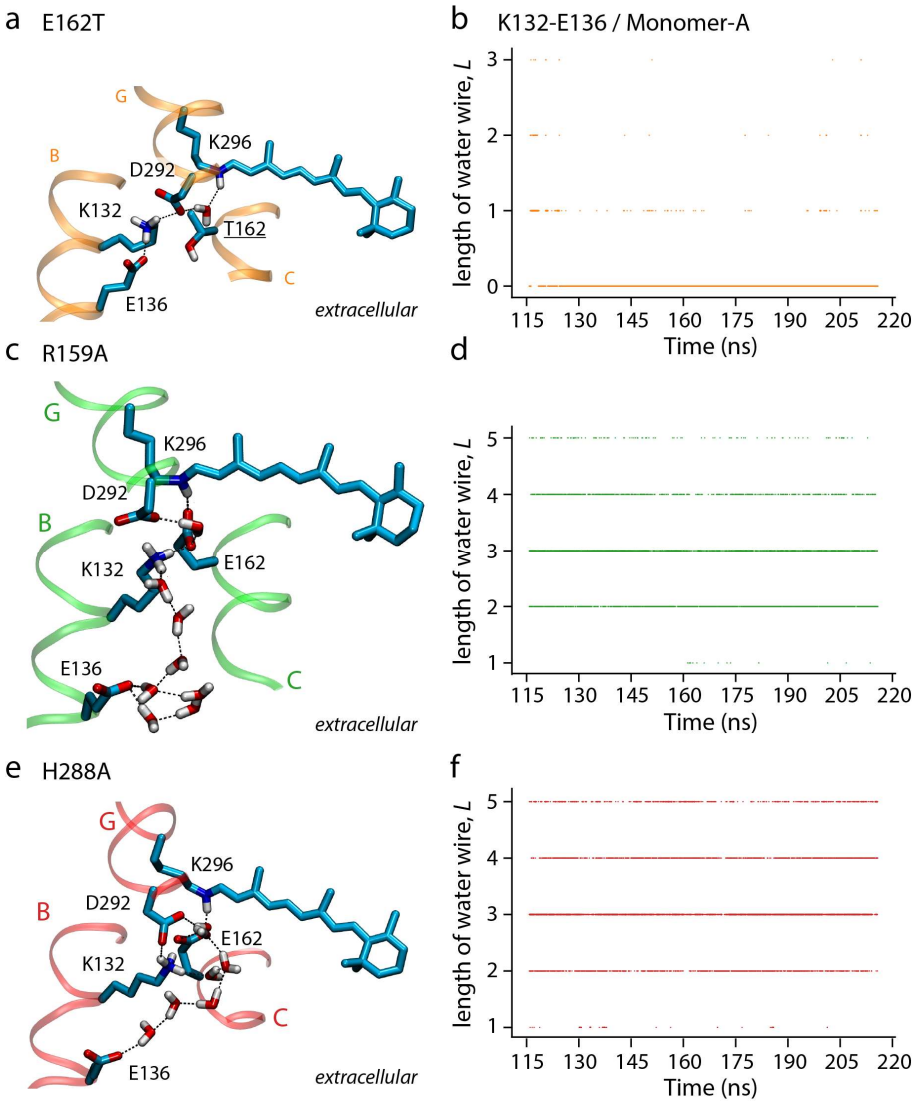

Figure 3.12. Internal communication networks of C1C2 are disturbed upon mutation. (a, c, e) The RSB-E136 connection sampled for the E162T mutant (a), R159A mutant (c) and H288A mutant (e) simulations of C1C2. The pathways are mediated from E162, D292, K132 and water molecules. (b, d, f) Water-wire length timeseries for the K132-E136 connection in the E162T mutant (b), R159A mutant (d) and H288A mutant (f) simulations of C1C2. E162T mutant I shown in orange, R159A mutant in green and H288A mutant in red. Adapted from ref. [42].

The significance of R159 in the studied pathway is more challenging to access through the point mutation. Observing the path in C1C2 and R159 there are no differences in the nodes list (Figure 3.10a, Figure 3.11c). The difference lies in the occupancy of the connection H288-E274. Visual inspection of the trajectory could provide insight for why there is an occupancy difference in the R159A vs. wild type. The substitution of arginine with an alanine allows for greater mobility of H288 [42], thus making the network very dynamic and as a result the H288-E274 occupancy is reduced from 99.2 to 76.9%. K132 can directly connect to E162, thus changing orientation compared to the E162T variant (Figure 3.11, Figure 3.12). Extended water wires are still able to constitute the K132-

E136 connection in the EC side, with increased water content. R159 could be characterized as very a suitable candidate for an interconnecting node in the pathways linking to the EC side, for the wild-type simulations along the simulations S2 and S4 of the mutations. In wild-type C1C2, H288 is part of the H-bond network of the retinal Schiff base (Figure 3.5, Figure 3.7, Figure 3.8, Figure 3.9, Figure 3.10), and it participates in the long-distance H-bond path that can connect transiently the retinal Schiff bases of the two protein monomers (Figure 3.9) and it contributes in all other systems with frequent connections to E162 & E274 (in wild-type, R159A), with K296 & E274 (in E162T).

In the H288A simulation, absence of the His sidechain causes rearrangement of the H-bond network and altered dynamics of specific H-bonds of the network. Direct H-bonding between the retinal Schiff base and the primary proton acceptor D292 is sampled very rarely, just 5.3% for Monomer-A, and 6.9% for Monomer-B respectively, and the Schiff base H-bonds instead to E162 (direct connection) which further bridges to E136 via H-bonding water (wire length L=4.2) with the JO computed at JO=68.9% (54% for Monomer-B). Alternatively, a low occupancy path via D292 can connect the retinal Schiff base to E136 (JO=5.3% for Monomer-A and 6.7% for Monomer-B, respectively). K132 is oriented towards the counterion in the H288A variant. Extended water-wires are still sampled to connect K132 to E136 with the water-wire length value being higher than the R159A equivalent.

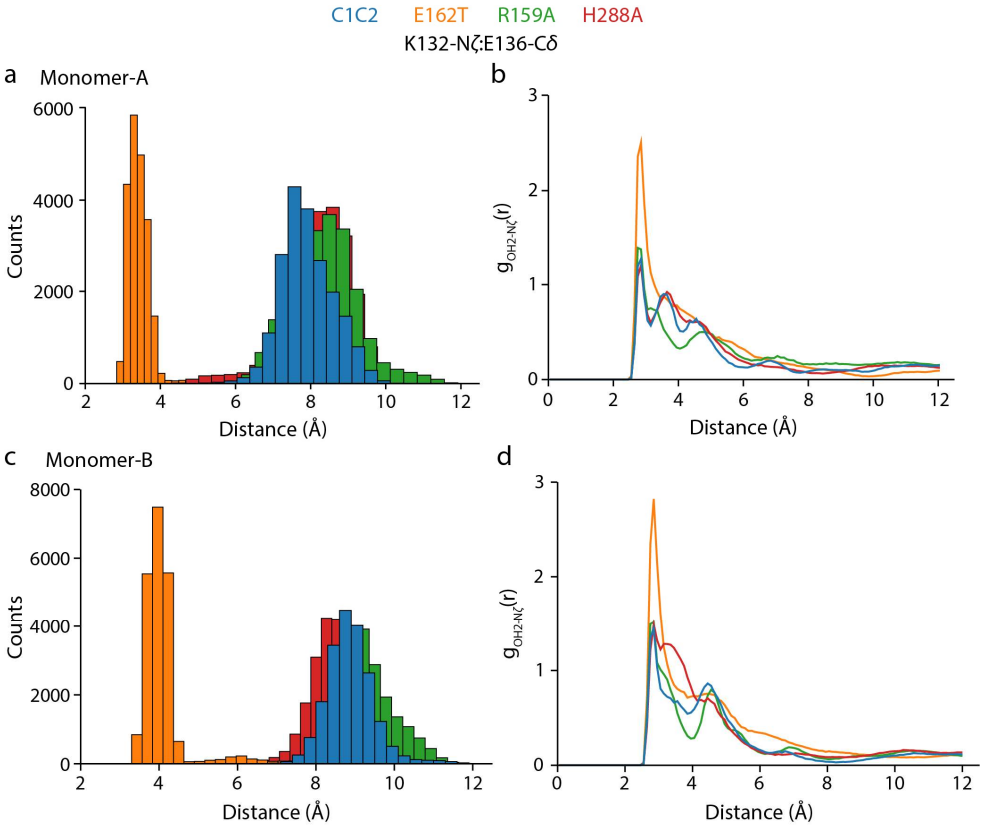

Figure 3.13. The effect of the E162T mutation in the internal communication networks of C1C2. (a, c) Histograms of the interatomic distance between the Nζ of K132 and the Cδ of E136 in

Monomer-A (a) and B (c), respectively. The histograms are computed from timeseries of the entire length of each trajectory. (b, d) Radial distribution function for the distance between water oxygen atoms and the Nζ of K132 in Monomer-A (b) and B (d). Bars for the wild-type C1C2 are shown in blue, E162T mutant in orange, R159A mutant in green and H288A mutant in red. Adapted from ref. [42].

In the wild-type C1C2, E136 bridges to K132 via H-bonded waters. In contrast, in the E162T mutant, E136 salt-bridges directly to K132, and it only rarely connects to K132 via water (Figure 3.12, Figure 3.13). The distance between the Nζ atom of the lysine and the Cδ of the glutamate is computed around ~3.5 Å in Monomer-A and ~4 Å in Monomer-B (Figure 3.13). In comparison, in the wild-type and the two other mutants variants (R159A, H288A), the distance between the functional groups of K132 and E136 is significantly larger (6 Å -11 Å), with the distributions centered around 8-9 Å. Thus, the salt bridge between K132 and E136 reduces the mobility of E136 since the distribution of the K132-Nζ:E136-Cδ distances are significantly wider in the wild-type and R159A/H288A mutants while only in the case of E162T the distributions feature a narrower range of distance values (Figure 3.13).

3.4 Summary

In this chapter, the new-generation algorithm package I contributed to the development of, *Bridge* [42], was described. Bridge was specifically designed with the efficient analyses of complex H-bonds networks in mind. It can be independently applied in static (crystal) structures and trajectories from MD simulations of proteins. In reality, the algorithm is very flexible, as it allows for lipid molecules, DNA or any other entity that features H-bonds to be analyzed, as long as a structure file exists. A key difference compared to other, already available analyses packages is that *Bridge* was designed to employ graph-based approaches to compute [33] and analyze [42, 185, 226] H-bond networks.

Implementations of graph-based algorithms that were employed in *Bridge* as the "Connected Component analysis" and Dijkstra's algorithm allow for an intuitive approach to the analyses. Most of the analyses features of *Bridge* were designed for proton-transferring systems in mind, and especially insightful are proved the graph curation functions and routines of *Bridge* in the case of retinal proteins, since the first step of the proton transfer events observed during the photocycle begins from the protonated retinal Schiff base. The RSB is thus a solid starting point when a network of H-bonds is to be dissected and queried. I analyzed the all-*trans* state of the C1C2 chimera and focused on the extended hydrogen-bond networks and clusters that dynamically form in the interior of the protein. The graph-based approach allows for the conclusion to be drawn, that E162 plays a very significant in linking the retinal to the other amino acids of the retinal vicinity, maintaining direct H-bond with the Schiff base almost 100% of time. Quantifying the local networks of the SB, I found that they can vary between the two monomers. An extended network that is comprised only by direct H-bonds stretches to the EC side, more specifically ending at E274, the equivalent of $E194^{BR}$ [99], which is a part of the proton release group in BR [220, 227-229]. In both monomers K132 is a

central node in the networks and retains stable connections to the two counterions. An H-bond motif between the sidechains of T166 and the counterion E162 is observed in both monomers. An equivalent motif is observed in ChR2 between T127 and E123 [230], only when E123 is unprotonated [230]. A similar motif is also observed in the acidic dark state of BR where T89 participates in an intrahelical H-bond motif with the backbone carbonyl of D85 which is found in the *i*-4 relative position [35].

Water mediated H-bonds increase the complexity of the systems significantly and allow for many more connections and pathways to be sampled. The interactions between E129 and N297, noted in the literature as the inner gate [101, 215] vary between the two monomers, with Monomer-B showing very frequent sampling (100%). Aside the local interactions in the vicinity of the RSB, water-mediated networks extend to carboxylates of the EC side, namely E136 and E274 ($E194^{BR}$) with very high sampling rates for the complete pathways. The pathways detected are shortest paths and the contain the two counterions E162, D292, K132, H288 and R159 in various combinations, as intermediate groups between the root and the end nodes.

Simulations of mutations on key amino acid residues heavily affect the local dynamics of the C1C2 networks. The impact of the mutations was characterized by a complex and extended H-bond network. With the graph-based queries and curation functions of networks I identified a very complex network of 46 amino acid residues bridging the two proton donors, the Schiff bases of the two monomers. Such a network has not been reported in the literature before, to the best of my knowledge.

Chapter 4 *Antarctic Rhodopsin*

This work is based on the following publication where I was second author:

Harris, A., **Lazaratos, M.**, Siemers, M., Watt, E., Hoang, A., Tomida, S., Schubert, L., Saita, M., Heberle, J., Furutani, Y. and Kandori, H., 2020. Mechanism of inward proton transport in an antarctic microbial rhodopsin. *The Journal of Physical Chemistry B*, *124*(24), pp.4851-4872.

I collaborated with Malte Siemers during this work, and it should be disclosed that Malte Siemers and I contributed to the theoretical section of the published material under the supervision of Prof. Dr. Ana-Nicoleta Bondar. I generated the homology model of AntR, using the amino acid sequence that I was trusted with, from Prof. Dr. Leonid S. Brown. I performed all system setups and MD simulations presented in the above publication along with all H-bond and network analyses on those systems. I prepared the figures for the corresponding chapters in the publication, and I contributed to writing of the text for the corresponding sections, under the close guidance and supervision of Prof. Dr. Ana-Nicoleta Bondar. Published figures contain input from other co-authors with the most important input being that of Prof. Dr. Ana-Nicoleta Bondar. Malte Siemers implemented the centrality analysis and the comparative graph analysis functions into the Bridge algorithm package. The former were previously introduced to our group from Konstantina Karathanou [184]. Lukas Kemmler wrote the script I used to perform the STRIDE analysis for the trajectories.

4.1 Background

In Lake Fryxell, Antarctica, under several meters of permanent ice a new group of inward proton-pumping rhodopsins was discovered [185]. The representative member that was studied is referred to as AntR (Antarctic Rhodopsin). The newly formed AntR subgroup is clustered along the recently discovered Schizorhodopsins (SzRs) [231] and is found between Xenorhodopsins (XeRs) [232] and Heliorhodopsins [233] in the phylogenetic tree [185]. AntRs and SzRs share a common characteristic. They both feature a single carboxylate counterion on helix G (D185 in AntR). In contrast, XeRs feature their single carboxylate on helix C. A unique feature of AntRs is that they feature a tyrosine hydrophobic group on helix C, in the position of the missing carboxylate counterion group, creating a strong electrostatic asymmetry or imbalance, in the retinal Schiff base vicinity. This characteristic is suspected to affect the deprotonation path of the Schiff base during the photocycle [185]. AntR is found to be an inward-proton pump, same as XeRs [234], which are the only inward proton transporting microbial rhodopsin group that has been described so far [232]. Members of XeRs show inward proton transport by their wild types, namely PoXeR, NsXeR and RmXeR. Anabaena Sensory Rhodopsin (ASR) does not pump protons or ions during its photocycle and functions as a sensor. Inward proton transport can be observed when a point mutation is introduced to the cytoplasmic proton acceptor D217, from an aspartate to a glutamate [235, 236]. AntR also has a cysteine group, C75, the homologous of a conserved cysteine in Channelrhodopsins that form the DC gate [217, 237]. Nevertheless, AntR is found to be an inward proton pump, that is capable of transporting protons against the membrane gradient [185].

4.2 Homology model and simulation system preparation of AntR

Starting from the raw amino acid sequence, a homology model for AntR was derived, using the Phyre2 webserver [238] through a sequence alignment on multiple retinal proteins whose structures are known. Six templates were used to build the model of AntR:

i) Chain D / Halorhodopsin from *Natronomonas phraonis* (3A7K [239])
ii) Chain A / Halorhodopsin from *Halobacterium salinarium* (1E12 [112])
iii) Chain A / Sensory rhodopsin II from *Natronomonas phraonis* (1H2S [240])
iv) Chain C / Proton pumping rhodopsin AR2 from *Acetabularia acetabulum* (3AM6 [241])
v) Chain A / L1-intermediate of Halorhodopsin T203V from *Halobacterium salinarium* (2JAG [242])
vi) Chain A / Acetabularia rhodopsin 1 from *Acetabularia acetabulum* (3WT9-superseded by 5AX0 [243])

The six templates (i-vi) used in the modelling of AntR share a 18-24% sequence identity with the AntR target sequence. Three amino acid residues, M1, V214 and G215

were modelled *ab initio*. The alignment of the remaining 212 amino acid residues resulted in a confidence score of 100%. The confidence score is given by the HHsearch [244] alignment algorithm for significantly improved alignment and detection that is employed by Phyre2 [238]. The retinal chromophore was docked in the model by coordinate transferring. A coordinate overlap was performed between the model and the first template (i), of Halorhodopsin. In order to avoid steric clashes a partial geometry optimization was performed for the amino acid residues around 4 Å of the retinal Schiff base using CHARMM.

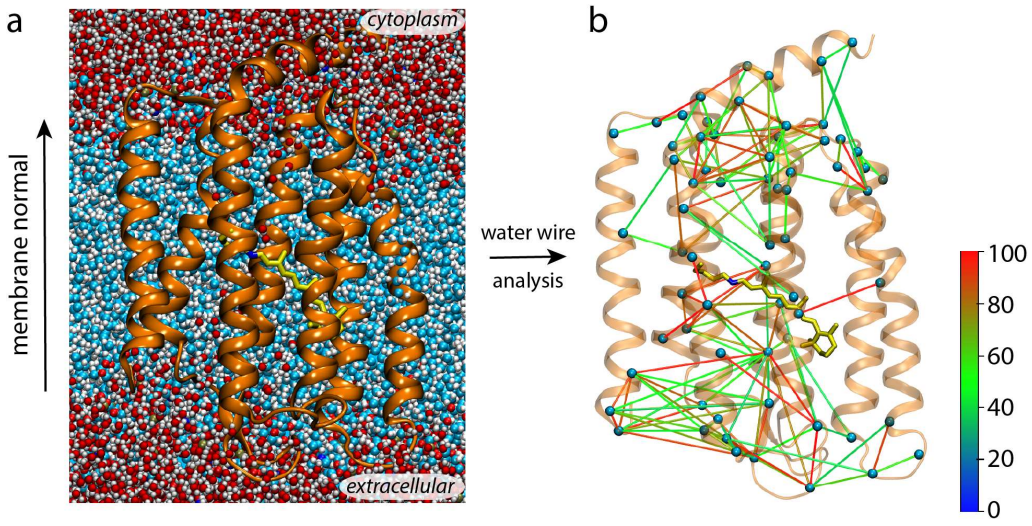

Figure 4.1. Overview of the AntR homology model. (a) System setup of the generated homology model of AntR, with an all-*trans* retinal chromophore, embedded in a hydrated lipid bilayer. (b) Water wire analysis of the all-*trans* AntR with a 35% pre-filter. Lines represent water-mediated connections between amino acid residues and are colored according to their occupancy rates using a blue-green-red (BGR) color scale. For clarity, the lines connect the H-bonded amino acid residues through their Cα-atoms. The protein is shown as orange ribbons using the New Cartoon representation of VMD. Adapted from ref. [185].

Isomeric states of the retinal Schiff base

Three isomeric states of the retinal have been modelled in this study. The retinal chromophore was docked in the all-*trans* conformation. Using CHARMM, the 13-*cis*,15-*anti* conformer was modelled by performing a twist on the all-*trans* using coordinate driving in CHARMM with the C_{12}-C_{13}=C_{14}-C_{15} retinal dihedral angle set to 0º. After 65ns of simulation time on the all-*trans* system, a twist on the C_{14}-C_{15}=N-Cε retinal dihedral angle was performed by posing a 5 kcal/mol force on the dihedral and a target value of 0º, using the "collective variable" option within NAMD, obtaining in the 13-*cis*,15-*syn* conformer (Figure 4.2). In order to perform the twist, the positions for all atoms that are not within 4 Å of the retinal Schiff base were frozen. An optimization of the retinal vicinity was performed in CHARMM to avoid steric clashes. Afterwards the constraints were lifted, and the simulation was performed for an additional 135ns.

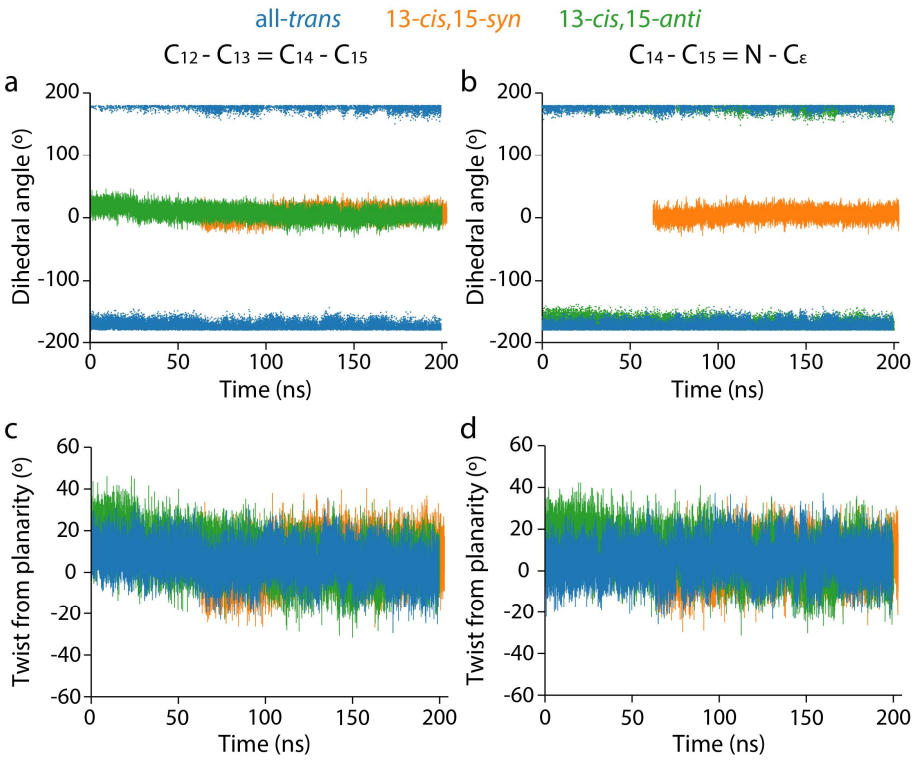

Figure 4.2. Dihedral angle dynamics for the C_{12}-C_{13}=C_{14}-C_{15} and C_{14}-C_{15}=N-Cε in simulations of the all-*trans* (blue), 13-*cis*,15-*syn* (orange) and 13-*cis*,15-*anti* (green) conformations of the RSB. (a, b) The C_{12}-C_{13}=C_{14}-C_{15} (a) and C_{14}-C_{15}=N-Cε (b) dihedral angles' timeseries. Values for the all-*trans* simulation are shown as a scatter for clarity. The 13-*cis*,15-*syn* and 13-*cis*,15-*anti* timeseries are shown as a solid trendline. (c, d) Timeseries for the twist from planarity for the C_{12}-C_{13}=C_{14}-C_{15} (c) and C_{14}-C_{15}=N-Cε (d) dihedral angles. Adapted from ref. [185].

Simulation setup

The "HBUILD" command was used in CHARMM to construct coordinates for H atoms. AntR was oriented along the membrane normal using the PPM webserver of the OPM database [201]. Using CHARMM-GUI interface, the model of AntR was placed in a hydrated lipid bilayer consisting of 496 POPE:POPG lipids in a ratio of 3:1, with 122 sodium ions added for charge neutralization. The total number of atoms in the simulation systems is approximately 165,770. The model of AntR with an all-*trans* retinal embedded in a hydrated lipid bilayer is shown in Figure 4.1.

MD simulation protocol

The systems were equilibrated using the CHARMM-GUI protocol, which features a 6-step relaxation. Three types of restrains are posed on the system and are gradually being reduced during those 6 steps. According to the protocol, harmonic restraints to ions and heavy atoms of the protein are applied, repulsive planar restraints

to prevent water from entering into the membrane hydrophobic region, and planar restraints to hold the position of head groups of membranes along the Z-axis [187]. The TIP3P water model was used, and the CHARMM-36 force field parameters for proteins and lipids.

Comparative H-bond graphs

A function to produce comparative H-bond graphs in the Bridge algorithm package was implemented for this project by Malte Siemers. Having in mind that three identical simulations of the homology model of AntR were generated, with the only parameter changing being the isomerization state of the retinal chromophore I was inspired to propose a new function in order to quantify how the H-bond networks re-arrange upon different isomerization states. In principle, the implemented function is applicable to systems with the same sequence. An additional option in the function was included to allow mutations in the sequence, in order to be more universally applicable for the people interested in using the software. The mutation option requires for the user to input the mutation manually. The function compares two graphs, one being the "*target*" and one being the "*reference*". The comparison between two graphs is displayed through colored edges, using the *reference* graph as the skeleton. Black edges represent the common H-bonds that are featured in both the *reference* and the *target*. The difference between the graphs is displayed using colored edges. Green edges represent H-bonds that are found in the *target* graph and not in the reference, thus being a new H-bond. Red edges represent H-bonds that are found in the *reference* graph but not in the *target*, thus being an H-bond that was lost and is not no longer present. The function is depicted schematically in Figure 4.3.

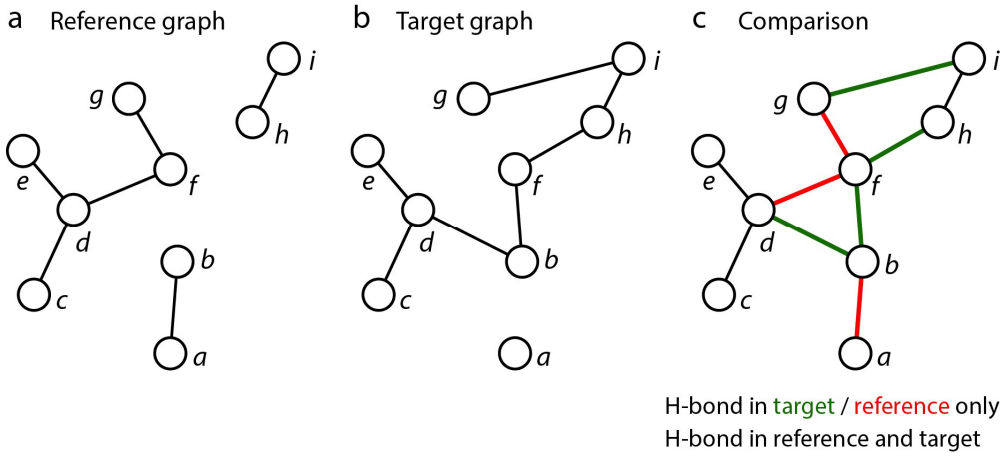

Figure 4.3. Schematic representation of the graph comparison function in Bridge. (a) Graph representation of an arbitrary H-bond network, used as the *reference* graph. (b) Graph representation of the *target* graph. (c) Comparative graph between the *reference* and *target* graph. H-bonds found only in the *target* are shown as green edges, while H-bonds found only in the *reference* graph are shown as red edges. H-bonds that are common between *target* and *reference* are shown as black edges. Adapted from ref. [185].

Structural stability and water content

During MD simulations of the all-*trans* conformation of AntR, the Cα - RSMD values of the helical segments are computed 2.14 ± 0.40 Å, while for the 13-*cis*,15-syn are 1.74 ± 0.13 Å and for the 13-cis, 15-anti, 1.92 ± 0.16 Å respectively (Figure 4.4a,c,e-orange trendline). The turns and coils were decomposed and their respective RMSD values were computed. Those are naturally very dynamic and are in the range between 3-7 Å (Figure 4.4a,c,e-green trendline). The generated homology model does not include any internal water molecules in the starting structure, nor any transferred coordinates of co-crystallized water molecules from any of the templates used. Due to numerous polar amino acid residues in the surface of the protein both in the extracellular and the cytoplasmic sides of the membrane, water molecules from the bulk were observed visiting the internal pore of the protein early in the simulation. The computations of the internal water molecules of AntR showed an average of 24.5 ± 8.5, 24.8 ± 8.7 and 27 ± 9.4 for the all-*trans*, 13-*cis*,15-*syn* and 13-*cis*,15-*anti* simulations respectively (Figure 4.4b,d,f-orange trendline). The positions of water molecules along the membrane normal was additionally computed as time-dependent variable. This was done in order to ensure that the protein in not leaking water molecules in the dark-adapted all-*trans* model (Figure 4.5). Computing the number density profiles for the retinal Schiff base and water around 10 Å within the RSB it was confirmed that water does not visit every coordinate along the membrane normal and a narrow part of the RSB is consistently unvisited by water molecules (Figure 4.5). This effectively creates two separate extended clusters in AntR, located in the extracellular and cytoplasmic sides respectively.

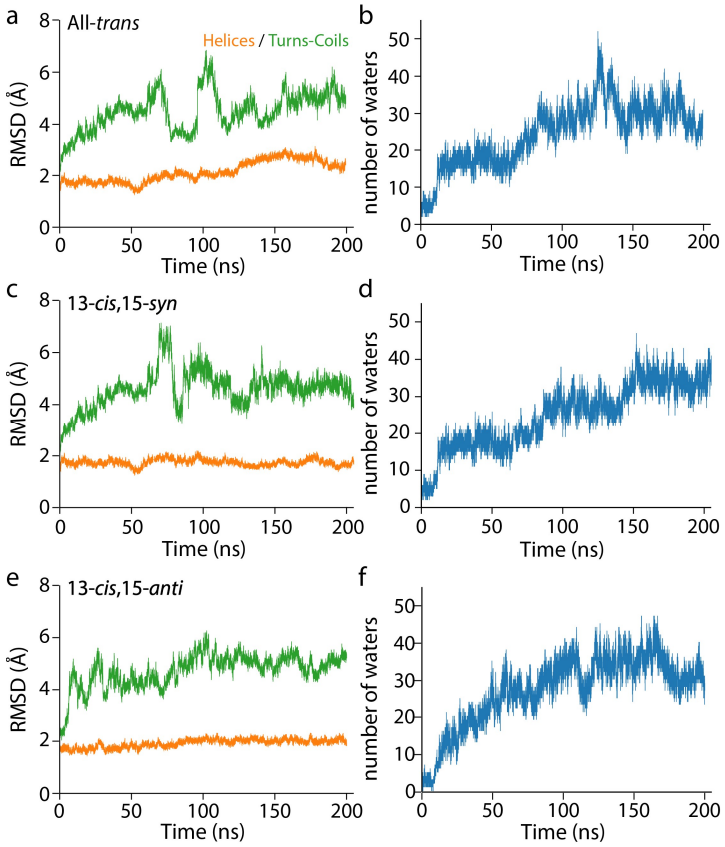

Figure 4.4. Cα - RSMD and number of internal water molecules profiles for the homology model of AntR in three different simulations of different isomeric states of the retinal. (left) Cα - RSMD profiles for the helical segment (orange) and the turns/coils (blue) for simulations in the all-*trans* (a), 13-*cis*,15-*syn* (c) and 13-*cis*,15-*anti* (e) conformations of the RSB. (right) Number of internal water molecules profile for simulations in the all-*trans* (b), 13-*cis*,15-*syn* (d) and 13-*cis*,15-*anti* (f) conformations of the RSB. Adapted from ref. [185].

To further examine the structural stability of the model, a time-dependent secondary structure analysis using STRIDE was performed. The primary type of secondary structure detected in the simulations is the α-helix, comprising about 70% of the protein while the coils comprise ~15% and the turns ~10% (Figure A.5), in all three simulations.

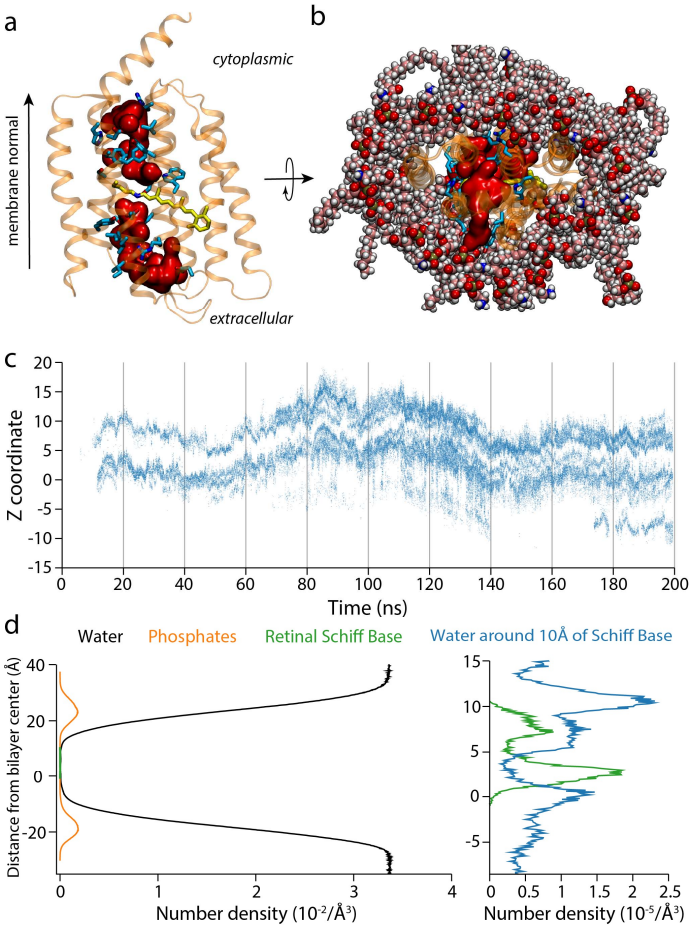

Figure 4.5. Internal water molecule calculations in the simulation of the model of the all-*trans* AntR. (a) Molecular representation of the internal water molecules in the all-*trans* simulation of AntR. The internal water is shown using a volumetric representation, named Quick Surf in VMD. Functional amino acid residues as well as non-polar amino acid residues surrounding the channel are shown. (b) Top view of the molecular graphics shown in panel a. The hydrated lipid molecules around the protein are shown in pink. The protein is shown in transparent for clarity. (c) Timeseries profile of the Z-coordinate e.g., coordinate along the membrane normal, for the water molecules around 5 Å of the RSB during the all-*trans* simulation of AntR. The timeseries was computed for the whole length of the trajectory. (d) Number density profile for water, lipid headgroups and the RSB as a function of the distance from the bilayer center in Å. For the water, only the oxygen atoms were selected (black trendline), for lipid headgroups-the phosphorus atoms (orange trendline), and for the RSB-the N16 nitrogen atom (green trendline). (F) Zoom-in number density profile for the N16 nitrogen atom of the RSB (green trendline) and the water oxygen atoms around 10 Å of the RSB (blue trendline). For the number density profiles the Density Profile Tool [245] of VMD was used, with the last 100 ns of the all-*trans* simulation of AntR and a cell thickness of 0.1 Å. Adapted from ref. [185].

4.3 Extended H-bond pathways in the all-*trans* model

The sequence and in turn, the three-dimensional structure of AntR features numerous homologues from Channelrhodopsin-2 and Bacteriorhodopsin [185]. A prime example is the aspartate group D167 in the extracellular side of AntR. $D167^{AntR}$ is the

equivalent of the proton release group E194 of BR. Since in BR, E194 is part of the extracellular proton release cluster [220], it was examined whether D167^{AntR} bridges to the RSB via protein-protein, protein-water and water-water H-bond pathways. It was found that indeed, D167^{AntR} and three other extracellular carboxylate groups - D2, E6 and E62 bridge to the RSB. The shortest paths from the extracellular carboxylate groups to RSB pass through the central arginine and the counterion D185 (Figure 4.6).

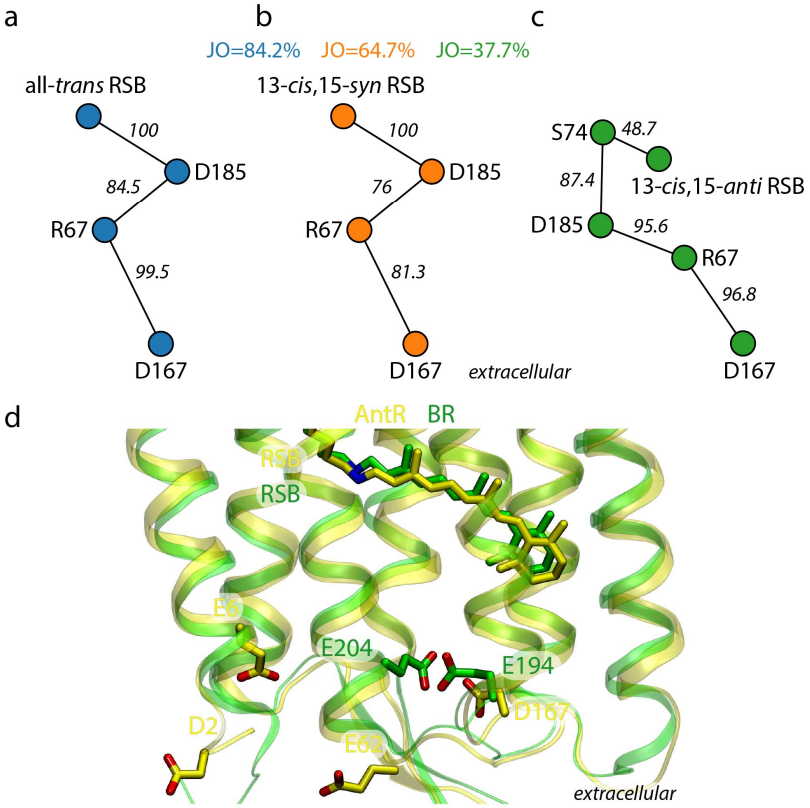

Figure 4.6. H-bond pathways from the extracellular side to the RSB in AntR. (a-c) Graph representations of the shortest paths sampled between D167^{AntR} and the RSB in the all-*trans* (a), 13-*cis*,15-*syn* (b) and 13-*cis*,15-*anti* (c) conformations of the RSB. The joint occupancy for every pathway is shown in its respective color code. (d) Structural alignment between AntR (yellow) and BR (green) (PDB ID: 1C3W). The proton release pair E194-E204 of BR are shown as well as the equivalent of R194, D167 of AntR and three additional carboxylate groups in the extracellular side- D2, E6, E62. Adapted from ref. [185].

For AntR with the RSB in the all-*trans* conformation, the connections to the EC side are very frequent, up to 84% Joint Occupancy for the RSB-D167 path (Figure 4.6), even when considering shortest path computations. The shortest paths connecting the RSB to the carboxylates in the EC side (D2, E6, E62) involve the counterion D185 connected to the Schiff base for 100% of the trajectory analyzed (Figure 4.7). The next step in the pathways for all the cases except D2, is the gating arginine R67 which is very frequently connected to D185 (84.5%). R67 is the key amino acid residue in forming these networks because it is the one that will connect to all the carboxylates on the EC

side, except D2. D2 is reachable though a connection via E6 or E62, for 87.5% and 90.2% of the time, respectively in the all-*trans* simulation (Figure 4.7b,c). An additional feature of D185 and S74 in the all-*trans* model was found. The counterion D185 is the first connection of the RSB and the gateway to the EC side through water-mediated H-bonds. On the other hand, D185 can connect to S74 of helix C, thus allowing for connections to CP side, even with an EC-oriented Retinal Schiff base. Indeed, extended H-bond pathways were computed from the RSB to the CP side, namely two carboxylates. E81 is the equivalent of $D96^{BR}$, which functions as the cytoplasmic donor during the photocycle of the outward proton pump [246-249]. It was suggested that it would be the prime candidate for the role of the proton acceptor during the photocycle of the inward proton pump AntR [185]. Mutagenesis and kinetics data from our collaborators suggest that the E81Q mutant leads to faster deprotonating SB and that E81 could not be the only proton acceptor. Instead, a proton acceptor complex could be in place [185]. D195 is a conserved group among AntRs and is also suggested as a candidate proton acceptor. High-occurrence H-bonds between D195 and E81 and R84 were identified. Although the homologues of L196 function as proton acceptors in ASR and PoXeR [250, 251], kinetics data of the D195N mutant disprove the hypothesis. Instead, it is suggested that a network of highly dynamic hydrogen bonds in CP side provides alternative proton transferring pathways [185].

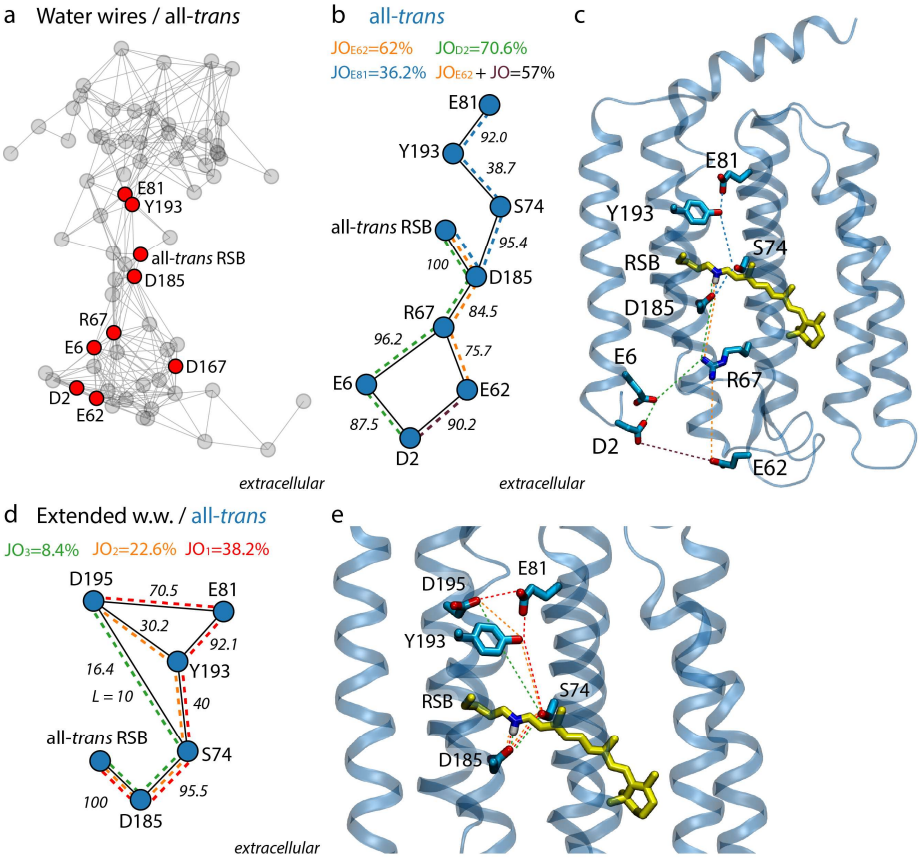

Figure 4.7. Extended H-bond pathways for the all-*trans* AntR. (a) Graph representation of the water-wire analysis on the all-*trans* AntR using a 5% occupancy filter. Functional amino acid

residues and extracellular carboxylates groups that are part of H-bond pathways are highlighted as red nodes. (b) Graph representation of the connected EC and CP pathways in the all-*trans* AntR. (c) Molecular graphics of the networks shown in panel b. (d) Graph representation of the extended water-wire CP pathways in the all-*trans* AntR. For the extended-water wire computation they threshold was increased to 12 water molecules. (e) Molecular graphics of the networks sown in panel d. Pathways are highlighted with colored dotted lines and their Joint Occupancy values are shown in the same color. Adapted from ref. [185].

4.4 H-bond pathways in the 13-*cis* models

13-*cis*,15-*syn*

The 13-*cis*,15-*syn* retinal retains characteristics of the all-*trans* model. Extending the search for pathways bridging the RSB to the extracellular side for the 13-*cis*,15-*syn* retinal it was found that the networks are almost identical to the ones computed for the all-trans retinal. This could be expected since the orientation of the Schiff base remains to be oriented towards the EC side. The major difference is that in 13-*cis*,15-*syn* model, D2 is directly connected with R67 and an extra mediating connection via another carboxylate in the EC side is no longer needed, as compared to the all-*trans*. A slightly reduced Joint Occupancy for the respective pathways is also observed, when compared to the all-*trans* AntR (Figure 4.8a,b).

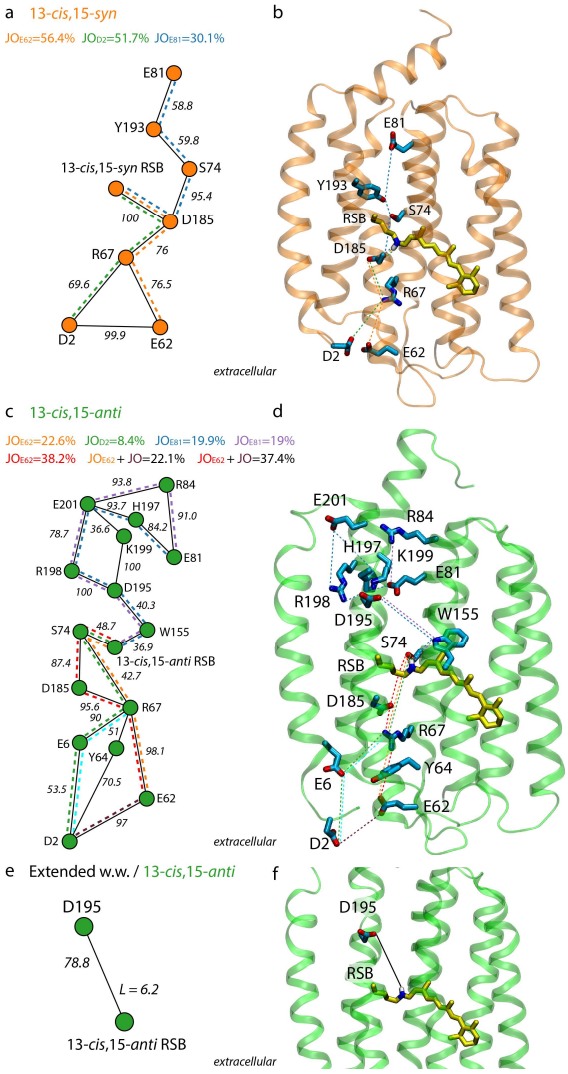

Figure 4.8. Extended H-bond pathways for the 13-*cis*,15-*syn* and 13-*cis*,15-*anti* AntR. (a) Graph representation of the connected EC and CP pathways in the 13-*cis*,15-*syn* AntR. (b) Molecular graphics of the networks shown in panel a. (c) Graph representation of the connected EC and CP pathways in the 13-*cis*,15-*anti* AntR. (d) Molecular graphics of the networks shown in panel c. (e) Extended water-wire computation for the 13-*cis*,15-*anti* AntR. Using a threshold of 8 water molecules. The pathway connects the RSB to D195 trough water-mediated connections for 78.8% of time and an average length of wire L=6.2. Pathways are highlighted with colored dotted lines and their Join Occupancy values are shown in the same color. Adapted from ref. [185].

13-*cis*,15-*anti*

The 13-*cis*,15-*anti* retinal connects to E81 and D195 at the cytoplasmic side of the membrane. An extended water chain that connects the 13-*cis*,15-*anti* AntR to D195 was identified(Figure 4.9). The chain of H-bonded water molecules that is sampled between the RSB and D195 has a range of 4 to 8 water molecules in the wire. The average water-wire length is L=6.2 and is sampled 78.8% of the time (Figure 4.9b). These water

molecules interact with the retinal Schiff base, S74 of helix C and are also supported by H-bonding to the backbone O atoms of Y150 and F192 of helix D and F, respectively.

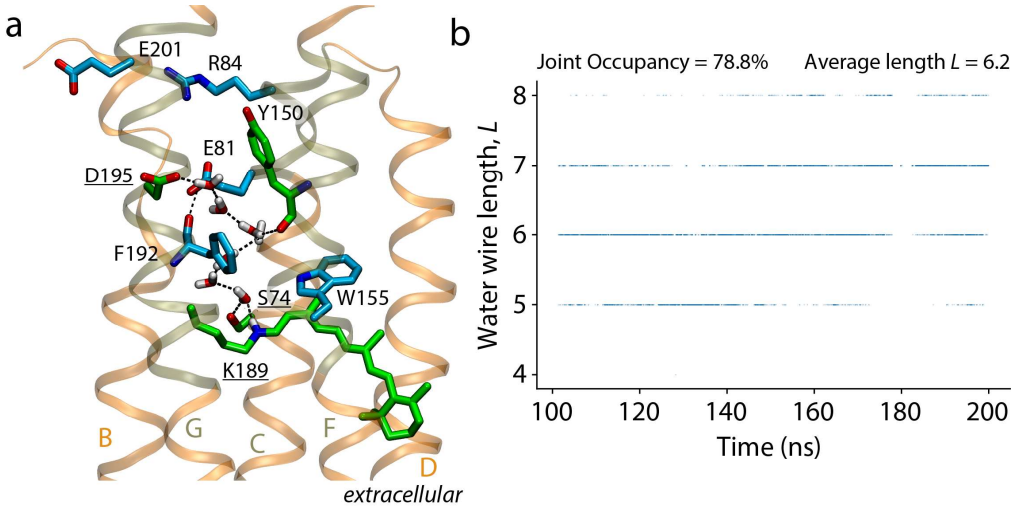

Figure 4.9. An extended water-wire connection in the 13-cis,15-anti AntR model. The wire connects directly the RSB to D195 and is the shortest path sampled between the two amino acid residues. The connection is sampled 78.8% of the time with an average length of 6.2 water molecules. (a) Molecular graphics of the extended water-wire sampled. (b) Water-wire length as a function of time. The water content of the wire ranges from 4 to 8 water molecules. Adapted from ref. [185].

S74 mediates transient bridging of the 13-*cis*,15-*anti* RSB to the extracellular H-bond network. The RSB of 13-*cis*,15-*anti* retinal connects transiently to the extracellular H-bond network via S74, whose hydroxyl group participates in dynamic H-bonds with the Schiff base and with D185 (Figure 4.8). S74 is found to be extremely important for linking the all-*trans* retinal to E81 found in the CP side, as well as linking the 13-*cis*,15-*anti* retinal to the carboxylate groups in the EC side. When the Schiff base is oriented towards the CP side (13-*cis*,15-*anti* model), the Schiff base will connect to S74, which in turn will connect to the counter ion D185 or R67. The JO for the pathways in the 13-*cis*, 15-*anti* conformer is largely reduced when compared to the all-*trans* due to the Schiff base-S74 connection which is sampled 48.7% of the time (Figure 4.8).

4.5 Centrality computations identify “hot-spots”

The treatment of the H-bond networks via graphs is a concept that allows one to analyze them using terms and elements from graph theory. The betweenness centrality shows important nodes that participate in networks. Amino acid residues with high BC values are important for the communication of the network, and their removal from the graph will most likely result in major disruptions. H-bond pathway analyses indicate that S74 plays a critical role in the linking of the two extended H-bond networks sampled in both sides of the membrane. When the RSB is in the all-*trans* and 13-*cis*,15-*syn* conformations e.g., EC-oriented, S74 links the RSB to the CP side of the membrane to

the proposed proton acceptor E81, through the counterion D185. Similarly, in the 13-*cis*,15-*anti* conformer, when the RSB is CP-oriented, S74 mediates the connections between the RSB and the glutamates in the EC side. In all H-bond pathway analyses of the three isomers that were analyzed, R67 is always the link between the RSB and glutamates on the EC side. Computations of centrality measures in terms of BC, DC (Figure 4.10) and USP (Figure A.6) unanimously show that in terms of network connectivity and communication, R67 is the most important node in the graphs. The centrality measure analyses were performed on occupancy-filtered graphs with a minimum occurrence rate of 35%. That criterion was chosen based on the RSB-W155 connection in the 13-*cis*,15-*anti* model, which is sampled 36.9% of the time, thus being the bottleneck. The BC, DC (Figure 4.10) and USP (Figure A.6) graphs show that R67 is the most central node. BC graphs show in addition that S74 is a central node in the communication of the networks, as well as the proposed proton acceptor E81 and Y193 that mediate the connections between RSB and the larger CP clusters. BC and USP values of R67 are relatively similar to one another since the RSB is oriented towards the EC side for both systems and R67 has 3 fewer connections compared to the all-*trans*. In the 13-*cis*,15-*anti* system, the largest disruption in the networks is found, and thus the BC and USP values indicating a major rearrangement of the H-bond networks because of the retinal isomerization. In the 13-*cis*,15-*anti* system, BC and USP values of the key amino acid residues R67, S74, D185, E81 and Y193 are significantly lower compared to the all-*trans* system. In the 13-*cis*,15-*anti* isomeric state the highest centrality value is computed for E201 on the cytoplasmic side of the membrane, located ~25 Å away from the RSB [185]. DC calculations in turn show that number of connections of the key amino acid residues remain similar and comparable to one another. R67 is consistently the most visited nodes with the highest number of connections. The lower *BC* (Figure 4.10) and USP (Figure A.6) values were computed for the 13-*cis*,15-*anti* vs. the all-*trans* state for the important amino acid residues such as R67, S74, E81, D185 and Y193, suggest that the network rearrangement that is sampled upon the retinal isomerization, strongly affects the communication of the network.

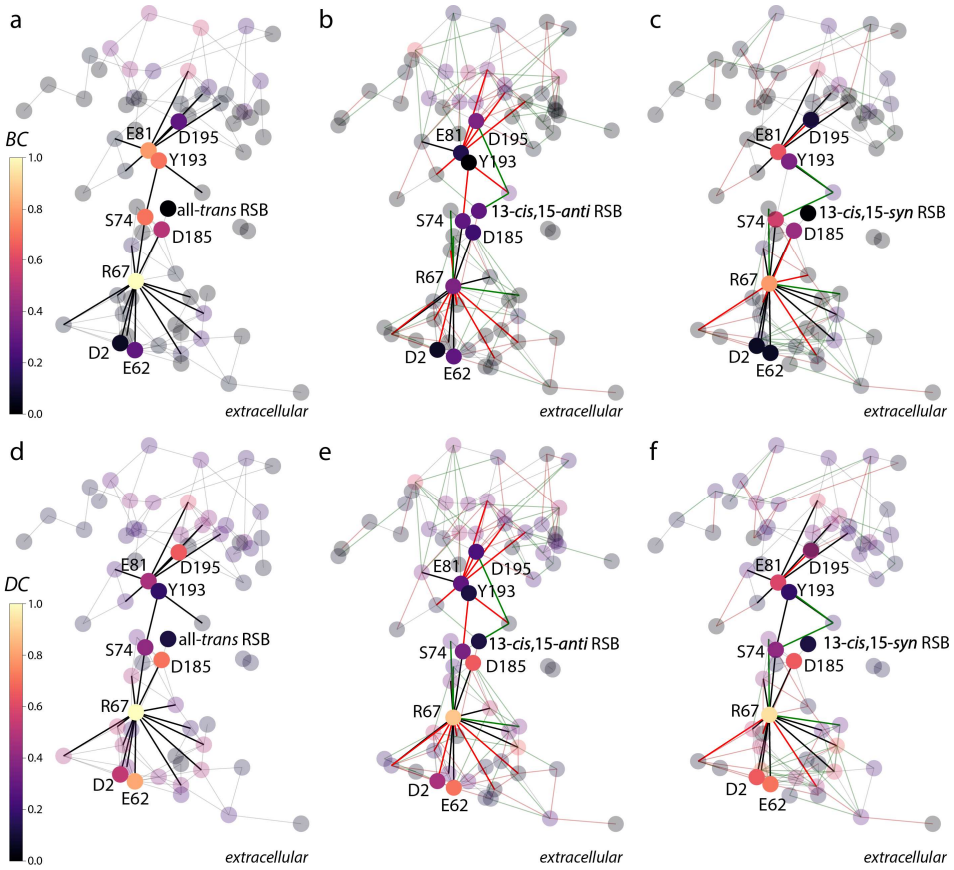

Figure 4.10. Centrality computations in AntR simulations. (a, b, c) BC computations on the all-*trans* (a), 13-*cis*,15-*anti* (b) and 13-*cis*,15-*syn* (c). (d, e, f) DC computations on the all-*trans* (d), 13-*cis*,15-*anti* (e) and 13-*cis*,15-*syn* (f). BC and DC values were computed on water-mediated H-bond graphs of AntR that were pre-filtered to a minimum of 35% occurrence rate. Nodes are colored according to the normalized and weighted BC and DC values. The amino acid residues of interest, and their connections are labeled and shown in full opacity. The remaining nodes and edges are shown in transparency. Panels b, c, e and f show comparative graphs using the all-*trans* graph as the foundation. The nodes shown in those panels are colored with the BC and DC values of the respective system. The edges are colored according to the function principles described in the section "Comparative H-bond graphs". Adapted from ref. [185].

4.6 Summary

In this chapter, presented the modelling and analysis of a newly discovered rhodopsin [185] was presented. AntR was discovered by the group of our collaborator Prof. Dr. Leonid Brown AntR and characterized as a bistable inward proton pump. Its physiological role remains still unclear. With the raw amino acid sequence that was shared with our group, I constructed a homology model of AntR, basing the model on a sequence alignment on 6 retinal proteins whose structures are known. The sequence identity between the template and the target ranges from 18 to 24% and the alignment of the 212 amino acid residues of the sequence received a confidence score of 100%, while

three amino acid residues, were modelled *ab initio.* Those were the first, second to last and last amino acid residues in the sequence.

To study the dynamics of AntR, I prepared three molecular systems, each with a different isomeric state of the retinal and found that AntR retains connections with both side of the membrane for every isomeric state of the retinal, regardless of its orientation. The all-*trans* and 13-*cis*,15-*syn* AntR features stable connections to EC located carboxylates, namely D2, E6, E62 and D167. The latter is the equivalent of one member of the proton release group of BR [220, 227-229], E194. Similar pathways were also found for C1C2 [42]. The pathways that connect the EC-oriented Schiff base to the carboxylates on the EC-side share common intermediate nodes. The counter ion D185 and the gating arginine R67 are common among the pathway which are very frequently sampled. Another aspect of the EC-oriented Schiff base was unraveled, unlike C1C2. An EC-oriented RSB can connect to the CP side through D185 which further connects to S74 and in turn to Y193, E81 ($D96^{BR}$) and D195 (conserved group among AntRs). It was suggested through by experimental evidence by our collaborators that R67 and D167 would be important for the proton uptake because the photocycle is largely perturbed when they are mutated [185], while E81 cannot be the sole proton acceptor. The supported notion, suggested by the combination of the theoretical predictions and experiments is, unlike the well-known example of BR where a proton is released by a proton release group [220, 227-229], that a network of highly dynamic hydrogen bonds in the CP side provides possible alternative proton transferring pathways. An extended water chain between the 13-*cis*,15-*anti* SB and D195 was sampled when the water threshold was increased from 5 to 8 molecules. The water chain is very frequently sampled, at 78.8% of the time, with an average water-wire length of L=6.2. The water molecules of the chain are stabilized by H-bonding to the retinal Schiff base, S74, Y150 and F192. S74 mediates transient bridging of the 13-cis,15-anti RSB to the extracellular H-bond network, allowing the 13-cis,15-anti retinal to connect transiently to the EC side. Unlike C1C2, AntR can maintain connections to both sides of the membrane, regardless of the isomeric state of the retinal, employing key amino acid residues in the process, such as S74 and D185.

With the centrality measures that were introduced for this projected [184, 185] and a newly developed measure of *USP* [186], I was able to quantify R67 as the most important node in the graphs in terms of network connectivity and communication, in addition to S74. The 13-*cis*,15-*anti* state showed smaller centrality values compared to the all-*trans* state for potentially important amino acid residues which would suggest major network rearrangements upon the retinal isomerization, strongly affecting network communications. The S74A mutation results into largely delayed kinetics of the photocycle and weakens the proton transport [185], validating the prediction of the centrality measurements on its importance in the H-bond networks. Being able to experimentally validate predictions from MD analyses, using novel graph theory approaches is very empowering. It shows the important role of theoretical biophysics in the pursuit of pushing the limits on what is known in the field and how experiments can co-exist and harmonically support and complement one another.

Chapter 5 *Conserved H-bond motifs in membrane transporters*

This work is based on the following publication where I was first author:

Lazaratos, M., Siemers, M., Brown, L.S. and Bondar, A.N., 2022. Conserved hydrogen-bond motifs of membrane trans-porters and receptors. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, p.183896.

It should be disclosed that Malte Siemers wrote the original structure mining routine to download and organize protein structures for the web, and loops to analyze multiple structures organized into groups. Original work of Malte Siemers was later modified, adapted and expanded for the purposes of this project as in-house standalone code. New motifs were implemented in the core package of *Bridge* for the purposes of this project. This code was not released as an update to *Bridge*.

I performed all system setups and MD simulations presented in the above publication along with all H-bond and network analyses on those systems. I prepared the figures for the corresponding chapters in the publication, and I contributed to writing of the text for the corresponding sections, under the close guidance and supervision of Prof. Dr. Ana-Nicoleta Bondar. Published figures contain input from other co-authors with the most important input being that of Prof. Dr. Ana-Nicoleta Bondar. Krzysztof Buzar helped me program the Python routine shown in Figure 5.1.

Parts of the work presented in this chapter are originally published in the journal Biochimica et Biophysica Acta (BBA) - Biomembranes. Figures and text originally published in the journal are modified in order to be presented in this chapter. Adapted figures and tables will be noted with “Adapted from ref. [226].”

Doi: https://doi.org/10.1016/j.bbamem.2022.183896
Author-directed link: https://authors.elsevier.com/a/1egZf1ClSNRBq

In the first iteration of Bridge [252], a routine for the detection of three H-bond motifs in crystal structures or trajectories was implemented. In this chapter, large-scale H-bond detections, and analyses for H-bond motif in crystal structures and trajectories are presented. Analyses of large data sets of multiple superfamilies can provide insight for the potential role of a motif in the protein's functions, according to their position along the membrane normal, the superfamily the belong to as well as their occurrence. I expanded Bridge to detect multiple H-bond motifs and I applied the motif detection analyses to large datasets organized according to superfamilies and *protein-groups*. Motifs of high importance in the biological role of the proteins were identified through this method, thus it is suggested that large scale analyses of protein structures and/or simulations can be of high predictive power for new proteins to be discovered or to provide insight in mechanisms of function that are unknown to this day.

5.1 H-bond motif computation

Bridge was used to identify ten H-bond motifs, which were chosen based on typical H-bond interactions in membrane proteins. Motifs i)-iii) were described in the original Bridge release and motifs iv)-x) were implemented during this project specifically. The motifs are as follows, and an overview with specific examples detected in TM proteins is shown in Table 5.1.

i) Intrahelical H-bond of the Ser/Thr hydroxyl group to the i-3, i-4, or i-5 backbone carbonyl (Ser/Thr-O=C)
ii) H-bond between an Asp/Glu carboxylate and a Ser/Thr hydroxyl (Ser/Thr-Asp/Glu)
iii) Combined motif of i) and ii) detected at the same time (C=O-Ser/Thr-Asp/Glu)
iv) H-bond between an Asp carboxylate group and an Asn carboxamide group (Asp-Asn)
v) H-bond between two His imidazole groups (His-His)
vi) H-bond between a Ser/Thr hydroxyl group and an Asn carboxamide group (Ser/Thr -Asn)
vii) H-bond between Asp/Glu carboxyl groups, and Arg guanidinium groups (Asp/Glu -Arg)
viii) H-bond between an Arg guanidinium group and a backbone carbonyl group (Arg- O=C)
ix) H-bond between an Asn carboxamide group and a backbone carbonyl group (Asn- O=C)
x) H-bond between an Asp/Glu carboxyl group and a backbone amide group (Asp/Glu -HN)

Table 5.1. Summary of H-bond motifs used from the original *Bridge* release and newly implemented for this project. The motif search was applied in two datasets of static protein structures (described in the section 5.2 below). For each H-bond motif, one protein is indicated in which the H-bond motif is present, the corresponding unique PDB ID of that structure and its resolution. Adapted from ref. [226].

Motif	H-bond groups		Motif example	PDB ID, reference, resolution (Å)
Ser/Thr-O=C	Ser/Thr hydroxyl	i-3, i-4, i-5 backbone carbonyl	Bacteriorhodopsin T46-F42	5ZIM [253] 1.25 Å

Ser/Thr-Asp/Glu	Ser/Thr hydroxyl	Asp/Glu carboxylate	Aquaporin-1 E51-S107	3ZOJ [130] 0.88 Å
Ser/Thr—Asp/Glu | O=C	Ser/Thr hydroxyl + Asp/Glu carboxyl	Ser/Thr hydroxyl + i-3, i-4, i-5 backbone carbonyl	Channelrhodopsin-2 E101-T246 V242-T246	6EID [97] 2.39 Å
Asp-Asn	Asp carboxyl	Asn carboxamide group	Proton-Translocating pyrophosphatase N738-D294	4A01 [131] 2.35 Å
His-His	His imidazole	His imidazole	Ammonia channel AmtB H168-H318	1U7G [129] 1.40 Å
Ser/Thr-Asn	Ser/Thr hydroxyl	Asn carboxamide	Mep2 Ammonium Transceptor N88-T95	5AEZ [254] 1.47 Å
Asp/Glu-Arg	Asp/Glu carboxyl	Arg guanidinium	Archaerhodopsin-3 D15-R17	6S6C [255] 1.07 Å
Arg-O=C	Arg guanidinium	backbone carbonyl	Sensory rhodopsin II R66-P182	1H2S [256] 1.93 Å
Asn-O=C	Asn carboxamide	backbone carbonyl	Sodium pumping rhodopsin KR2 N112-S70	6YC3 [257] 2.00 Å
Asp/Glu-NH	Asp/Glu carboxyl	backbone amide	Heliorhodopsin D158-L160	6SU3 [258] 1.50 Å

5.2 Compiling the dataset of static structures of TM proteins

For the collective analysis of protein families and superfamilies a dataset containing 1439 structures of alpha-helical polytopic transmembrane proteins was generated. The structure files were downloaded from the Orientations of Proteins in Membranes (OPM) database [201] in ".pdb" format. The OPM database gives protein structures pre-oriented along the membrane normal with the origin of coordinates corresponding to the center of the membrane. To obtain the raw data about families/superfamilies and proteins, an algorithm to retrieve the information about the selected families/superfamilies from the OPM database using the OPM data API was developed. The OPM data API uses a nested dictionary data structure that includes all info provided about a protein. Notable examples are the PDB ID, resolution, family ID, superfamily ID, number of TM subunits and total number of TM helices. Protein superfamilies are also assigned unique ID's, and are organized in a nested format, containing every protein they feature. Most of the structures in the data set (1426) were solved with X-ray crystallography or Electron Microscopy (EM), and 13 structures were solved with Nuclear Magnetic Resonance spectroscopy. The initial and unrefined data set of structures includes proteins from 28 superfamilies (Table 5.2).

Table 5.2. Summary of the dataset of structures compiled prior and after replacements of primary representations with one of their secondary representations of higher resolution. The family/superfamily names, number of members, and the range of resolution values at which they were solved at are presented. Resolution values of "N/A" correspond to the NMR structures. Adapted from ref. [226].

Family/Superfamily name	# Structures	Before replacements	After replacements
		Resolution range (Å)	
Sodium/calcium exchanger	7	1.9–3.5	1.9–3.5
Proton or Sodium translocating F-type, V-type and A-type ATPases	103	1.7–37.0	1.55–37.0
Rhodopsin-like receptors and pumps	358	1.3–60	1.07–60
Proton-translocating pyrophosphatase	5	2.35–4.0	2.18–4.0
Ion channel (VIC) superfamily	405	1.45–12.7	1.2–12.7
Major Intrinsic Protein (MIP)/FNT superfamily	31	1.15–25	0.88–25
Chloride transporter (ClC)	17	2.51–4.34	2.4–4.34
Small conductance mechanosensitive ion channel	15	2.9–4.14	2.9–4.14
Large conductance mechanosensitive ion channel	5	3.5–5.8	3.5–5.8
Ammonia/urea transporters/Na+ exporter	13	1.4–3.5	1.3–3.5
Ion transporter superfamily	6	2.78–3.96	2.78–3.96
Resistance-nodulation-cell division	66	1.9–6.5	1.9–6.5
ABC transporters	161	1.9–7.9	1.9–7.9
Vacuolar iron transporter	2	2.7–3.5	2.7–3.5
P-type ATPase (P-ATPase)	100	2.2–10.0	2.15–10.0
Proton-translocating transhydrogenase	5	2.89– 6.93	2.2–6.93
Copper transporter	1	3.03	3.03
Piezo family	4	3.8–4.5	3.8–4.5
Magnesium ion transporter-E (MgtE)	3	2.2–3.5	2.2–3.5
Cation diffusion facilitator	6	2.9–13.0	2.9–13.0
CorA metal ion transporters (MIT)	8	2.7–4.2	2.7–4.2
Drug/Metabolite Transporter (DMT)	10	2.2–3.5	2.1–3.4
Potassium channel TMEM175	5	2.4–3.3	2.4–3.3
Mercuric ion uptake (Mer) superfamily	1	N/A	N/A
Calcium release-activated calcium (CRAC) channel	6	3.35–6.9	3.35–6.9
Monovalent cation-proton antiporter	17	1.95–3.98	1.95–3.98
Calcium-activated chloride channel	29	3.1–5.14	3.1–5.14
Bestrophin anion channel	17	2.17–3.72	2.17–3.72

To tackle the issue of significant redundancy within the PDB, that a protein can be represented by more than one entries (ID's) OPM defined the term of "*primary representation*" of a protein entry [201]. The primary representation features the most complete protein in terms of protein domains and the least disordered segments [201]. For this specific work case scenario, only the highest quality structures possible were of interest to investigate their H-bond networks and H-bond motifs, so it was deemed vital to include in the dataset the structures solved at the highest resolution (lowest numerical value) possible per protein entry. In some cases, the primary representation contained at least one other representation. Those representation will be referred to as secondary representations in this chapter. I developed an algorithm routine (Figure 5.1) to compare the resolution values and keep the highest quality one (i.e., the lowest absolute value of resolution in Å). Eight occasions where the primary representation's resolution was lower quality than two -or more- alternative entries of the same protein were encountered. In those 8 occasions, the alternative entries' resolutions were the same value in Å. Only 2 out of the 8 occasions featured protein entries that were solved in resolutions ≤ 2.5 Å. In that case, a direct user input in the program was allowed, of ID that will replace the representative entry. In total, 109 primary representations were replaced by one of their secondary representations solved at a higher quality.

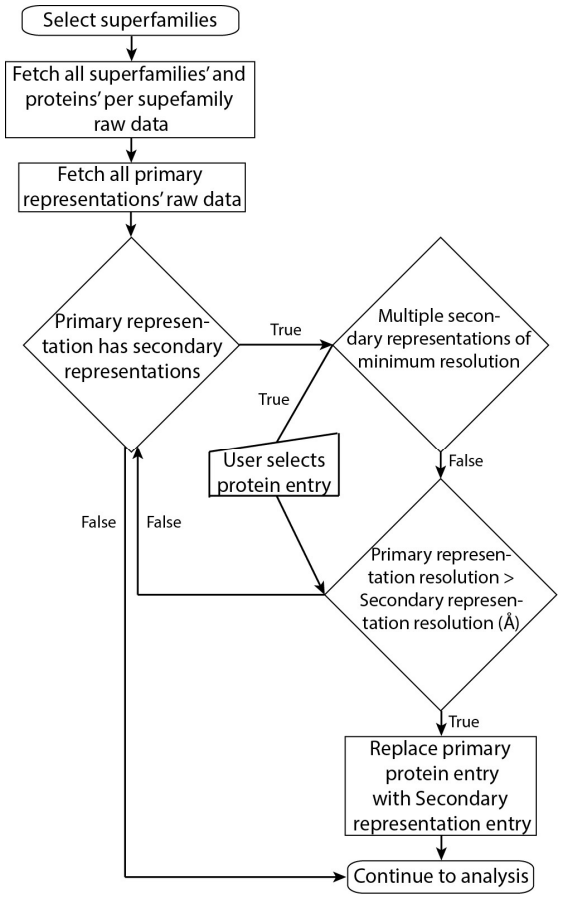

Figure 5.1. Schematic representation of the replacement procedure implemented, as a flow chart. The common symbols for flowcharts are used and they are arrow, rhombus, rounded rectangle, squared rectangle, quadrilateral representing flowchart, decision, terminal, process, and manual input respectively.

An initial resolution filter of 5.0 Å is employed, effectively creating the first resolution curation stage of the dataset that does not include outlier structures of lower quality. Most of the protein structures solved with X-ray crystallography report a resolution of ~3-4 Å with a minimum resolution of 0.88 Å as compared to 1.15 Å prior to replacements (Figure 5.2). The highest resolution structure is the *Pichia pastoris* aquaporin-1 (Aqy1) [130].

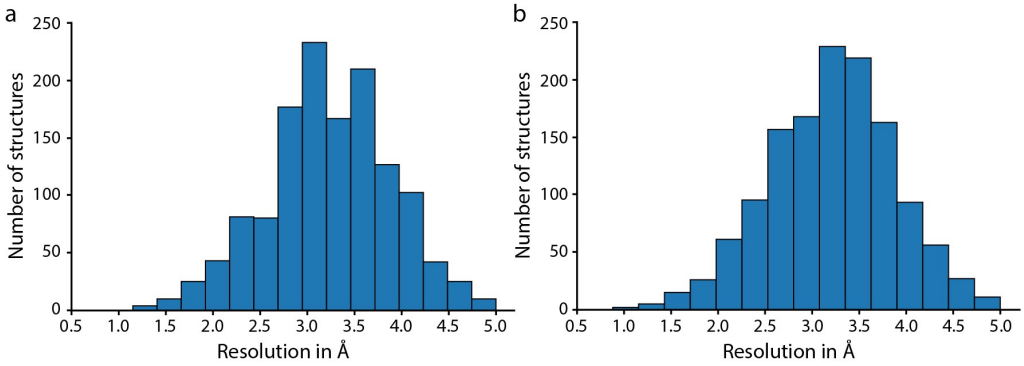

Figure 5.2. Distribution of crystal structures according to their resolution they were solved at. (a) Distribution of protein resolutions of the original dataset, prior to the entry replacements. (b) Distribution of protein resolutions of the dataset after replacements of primary representations with one of their secondary representations of higher resolution. The highest quality structure is solved at 0.88 Å as compared to 1.15 Å prior to replacements, and it is *Pichia pastoris* aquaporin-1 (Aqy1) [130]. The structures' resolutions are distributed in 15 bins.

A handful of high-resolution protein structures could not be analyzed via the automated procedure to identify H-bond motifs and required editing for readability. For these structures, CHARMM-GUI PDB reader [188, 191] was used to generate a formatted protein coordinate file. Ten structures did not contain transmembrane regions and were not considered. For the accuracy of the H-bond networks reported here, two data subsets were generated for computations and analyses. The resulting datasets used for all analyses discussed here include 200 error free protein structures solved at a resolution equal or higher than 2.5 Å (*Set-high*, Table A.1) and 483 error free protein structures solved at a resolution lower than 2.5 Å and equal or higher than 3.5 Å (*Set-low*, Table A.2). The proteins included in *Set-high* and *Set-low* belong to 17 and 26 different protein families/superfamilies respectively (Table 5.3). An automatic mining procedure could result in same protein to be represented by more than one entry. Therefore, a subset of the original *Set-high* was generated to include unique sequences. This subset will be referred to as "*Set-highU*", where "U" stands for "Unique". Every PDB ID of *Set-high* was inspected individually by using the advanced search of the PDB under the "Macromolecules" tab with the "Find similar proteins by:" option. By selecting a high identity cutoff (100% in most cases), similar proteins of the original query were found. Every high-identity match was cross-checked if it was present in *Set-high*. Proteins with more than one entry of the same name e.g. Bacteriorhodopsin, Anabaena Sensory Rhodopsin and others, were checked with identity cutoffs of 50%. A summary of the duplicate structures with high sequence identity matching is shown in Table A.3 and the complete *Set-highU* dataset in Table A.4. One of the most over-represented proteins were bacteriorhodopsin, calcium ATPase and the potassium channel KcsA. Two examples of bacteriorhodopsin originating from *Haloquadratum walsbyi* and *Haloarcula marismortui* shared 54% sequence identity between them and 56-57% with PDB ID:5ZIM from *Halobacterium salinarum*. They were thus not treated as non-unique entries and remained in the *Set-highU* dataset. In Table A.4 the structures with high

sequency identity matches are summarized. For proteins that contain more than one entry with high sequence identity, only the highest resolution one was considered and included in the *Set-highU*.

Table 5.3. Summary of the datasets used for analyses. The family/superfamily names, number of structures in *Set-high*, *Set-highU* and *Set-low* are shown respectively. Adapted from ref. [226].

Family/Superfamily name	**#structures**		
	Set-high	***Set-highU***	***Set-low***
Sodium/calcium exchanger	4	3	3
Proton or Sodium translocating F-type, V-type and A-type ATPases	9	9	27
Rhodopsin-like receptors and pumps	94	65	141
Proton-translocating pyrophosphatase	1	1	3
Ion channel (VIC) superfamily	24	17	118
Major Intrinsic Protein (MIP)/FNT superfamily	20	17	5
Chloride transporter (ClC)	1	1	5
Small conductance mechanosensitive ion channel	0	0	4
Large conductance mechanosensitive ion channel	0	0	2
Ammonia/urea transporters/Na+ exporter	11	10	2
Ion transporter superfamily	0	0	2
Resistance-nodulation-cell division	2	1	26
ABC transporters	7	7	59
Vacuolar iron transporter	0	0	2
P-type ATPase (P-ATPase)	12	3	46
Proton-translocating transhydrogenase	1	1	1
Copper transporter	0	0	1
Magnesium ion transporter-E (MgtE)	1	0	2
Cation diffusion facilitator	0	1	1
CorA metal ion transporters (MIT)	0	0	3
Drug/Metabolite Transporter (DMT)	3	3	4
Potassium channel TMEM175	1	1	1
Mercuric ion uptake (Mer) superfamily	0	0	0
Calcium release-activated calcium (CRAC) channel	0	0	1
Monovalent cation-proton antiporter	5	5	7
Calcium-activated chloride channel	0	0	10
Bestrophin anion channel	4	2	7
Total number of structures	200	147	483

I employed the Transporter Classification Database (TCDB) [259] in order to group the superfamilies included in the datasets according to their biological role. I relied on the first database index/classification number that denotes the transporter class. Since the Rhodopsin-like receptors and pumps and the Voltage gated ion channel superfamilies

are diverge in their respective subfamilies will be treated independently. The Ammonia/urea transporters/ Na^+ exporter family was included in the channels/pores group since all proteins but one belongs to that group. The exception holds the Na^+-translocating NADH-quinone reductase which is grouped as a primary active transporter. It is though solved at 3.5 Å, belonging to upper limit of Set-low, and shows only one D-N motif. The grouped families/superfamilies according to their biological role will be referred here as *protein-groups*. An overview of the protein-groups and their number of members for *Set-high* and *Set-low* is shown in Table 5.4 below.

Although the Rhodopsin-like receptors superfamily was treated as a separate *protein-group*, it was further dissected, starting from *Set-highU*, effectively generating a subset of "Microbial and algal rhodopsins" (*Set-high-mr*) and "A, B, C and F GPCRs" (*Set-high-gpcr*). The Hemolysin-III family and Heliorhodopsin families are distinct as per the TCDB classification as compared to the above distinction. The Human adiponectin receptor 2 (PDB ID: 5LWY [260]) belongs to the Hemolysin-III family and heliorhodopsin (PDB ID: 6SU3 [258]) belongs to the Heliorhodopsin family. Those two proteins will be treated independently and will define *Set-high-hemo*, the subset for the human adiponectin receptor 2 and Set-high-*helio* for heliorhodopsin. The resulting subsets of the Rhodopsin-like receptors and pumps (65) are *Set-high-mr* (28), *Set-high-gpcr* (35), *Set-high-hemo* (1) and *Set-high-helio* (1). An overview of the subsets is shown in Table 5.4 below.

Table 5.4. Summary of the re-organized families/superfamilies into protein-groups using the TCDB [259] classification method. The protein-group names and the number of structures for *Set-high* and *Set-low* are shown respectively. The Rhodopsin-like receptors and pumps and Voltage-gated Ion Channel (VIC) are included are separate superfamilies due to their diversity of subfamilies, as described above. The dissected subsets *Set-high-mr*, *Set-high-gpcr*, *Set-high-hemo* and *Set-high-helio* are shown under the Rhodopsin-like receptors and pumps superfamily. Adapted from ref. [226].

Protein-group names	***#structures***		
	Set-high	***Set-highU***	***Set-low***
Channels/Pores	37	31	38
Electrochemical Potential-driven Transporters	15	13	50
Primary Active Transporters	30	21	136
Rhodopsin-like receptors and pumps superfamily	94	65	141
i. Microbial and algal Rhodopsins (*Set-high-mr*)		28	
ii. A, B, C and F family GPCRs (*Set-high-gpcr*)		35	
iii. Hemolysin III family (*Set-high-hemo*)		1	
iv. Heliorhodopsin family (*Set-high-helio*)		1	
Voltage-gated Ion Channel (VIC) superfamily	24	17	118

Defining the transmembrane domain core in the datasets

The protein dataset compiled so far, consists of only transmembrane α-helical proteins. Previous motif detections [35, 36, 119, 121] found H-bond motifs in TM region of a protein to suggest connection with the biological functional of proteins and events taking place i.e., proton transfer during a photocycle. Thus, the motif detection procedures implemented in this work are applied only to the TM region of every protein in the datasets. An advantage of the OPM is that the structures are pre-aligned [201] along the membrane normal with pseudo or "*dummy*" atoms being used in the downloadable file to denote the limits of the bilayer.

From the members of *Set-high* and *Set-low* the TM region was computed according to the average membrane boundaries indicated by OPM [201]. The average distance from the bilayer center ranges from 12.6 to 19.2 Å with average being between 15.5 Å. The limits of the lipid bilayer are set to ±18 Å, 2.5 Å more than the average distance from the bilayer center in order to account for dynamic fluctuations that take place in fluid membranes as compared to the cryogenic conditions that the crystal structures are determined. Bertalan et al. used a similar approach to define the TM region in large-scale analyses of GPCRs [261]. In further analyses the TM core will be referred to as the area of ±6 Å away from the center of the bilayer.

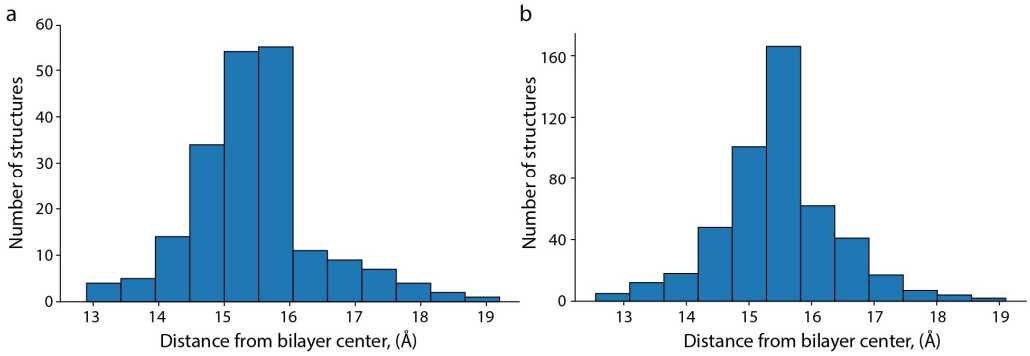

Figure 5.3. Distribution of crystal structures according to their distance from the bilayer center. (a) Distribution of protein of crystal structures of *Set-high*. (b) Distribution of protein of crystal structures of *Set-low*.

5.3 Single structure and MD simulation system setup

Two model systems were chosen to investigate H-bond networks and motifs, outside datasets. The first is a high-resolution structure of aquaporin and the wild-type structure of channelrhodopsin-2.

Aquaporin-1 (Aqy1)

The reason for choosing aquaporin as model system is that the crystal structure of aquaporin was solved at a resolution of 0.88 Å (PDB ID:3ZOJ [130]), which allowed coordinates for H atoms and 221 water oxygen atoms to be solved. Four systems of the

Aqy1 tetramer where generated, with two different lipid bilayer compositions and two different protonation states for the His sidechains (Table 5.5). To prepare the simulation systems, the structure of the aquaporin-1 tetramer was oriented along the membrane normal and was translated to the center of coordinates using the "alignment to principal axes" package of VMD. CHARMM [145, 171] was used to generate the H-atoms for water molecules present in the structure, and CHARMM-GUI to insert the tetramer in the hydrated lipid bilayers.

Simulations denoted as aq1 have the Aquaporin-1 tetramer embedded in a hydrated lipid bilayer of 1-palmitoyl-2-oleoyl-sn-glycero-3 phosphoethanolamine (POPE) lipids for aq1. In simulations denoted as aq2, the lipid bilayer contains a 45%:35%:20% mixture of POPE, 1-palmitoyl-2-oleoylphosphatidylcholine (POPC), and 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (POPS). The ratios were adjusted from ref.17 by adding 10% to POPE and POPC fractions to compensate for the 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidic acid (POPA), 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoinositol (POPI), cardiolipin (CL), lysophospholipids (lyso-PL) and dimethylphosphatidylethanolamine (DMPE) that were not modelled for this bilayer.
Originally, the crystal structure contains H-atoms due to it being solved at a very high resolution of 0.88 Å. Amino acid residues H44, H194 are indicated as Nε2-protonated, H212, H242 indicated as Nδ1-protonated, while all aspartate and glutamate groups are negatively charged. The simulation systems aq1_a and aq2_a have all His groups singly protonated on the Nδ1 atoms. Simulations aq1_b and aq2_b have His groups protonated as indicated in the crystal structure, where H44, H194 are Nε2-protonated, and H212, H242 are Nδ1-protonated (Figure A.8, Figure A.9).

Channelrhodopsin-2 (ChR2)

The simulations of wild-type channelrhodopsin-2 were based on the crystal structure of the dimer PDB ID: 6EID [97], which is dimeric structure from Chain A. The structure features 26 crystallographic waters per monomer (52 total) and they were all included in the starting coordinates of the MD setup. The pre-oriented structure along the membrane normal with an all-*trans* retinal chromophore was used and downloaded from the OPM database [201]. The program MODELLER [211-213] was used to generate the atomic coordinates missing from the crystal structures. Two amino acid residues with a carboxylic sidechain, E90 and D156 were considered protonated according to refs. [79, 80, 101, 215, 217], and cysteines C34/C36′ [97] were disulfide bridged, according to the crystal structure [97]. The channelrhodopsin-2 dimer with the all-*trans* retinal was embedded in a hydrated lipid membrane with 401 POPC lipid molecules and ions added for charge neutrality (Table 5.5) [226]. Originally, the ChR2 contains the retinal Schiff base (RSB) is in an all-*trans*, and 13-*cis,*15-*syn* isomeric states [97, 262, 263]. The model of a ChR2 dimer with a 13-*cis,*15-*anti* retinal was generated using adiabatic mapping in CHARMM, with an all-*trans* retinal as a starting geometry, and twisted the C_{12}-C_{13}=C_{14}-C_{15} dihedral angle from 180° to 0° in intervals of 10°, performing energy minimizations whilst keeping the RSB fixed between intervals. In that way, the compete twist (180º) of the RSB and the nitrogen atom facing the CP side were modelled.

Table 5.5 Summary of the MD simulations prepared and performed in this chapter. The simulation system, the lipid bilayer environment, number of atoms, temperature, protonation and length of the simulation are reported in the table. The "protonation" tab refers to Asp, Glu and His sidechains. Other titratable groups were assigned standard protonation states. For Aqy1, 'Mixed' tab indicates the lipid bilayer composed of 45:35:20 POPE:POPC:POPS". Adapted from ref. [226].

Simulation	Lipids	Number of atoms	Temperature (K)	Protonation	Length (ns)
ChR2, dimer					
all-*trans* 13-*cis*	POPC	177,367	300	E90, D156 all Nε	200
Aqy1, tetramer					
POPE-all HSD (aq1_a)	POPE	261,360	300	all Nδ	261
POPE-HSD/HSE (aq1_b)	POPE			H44-H194 Nε / H212-H242 Nδ	250
Mixed-all HSD (aq2_a)	Mixed	263,094		all Nδ	251
Mixed-HSD/HSE (aq2_b)	Mixed			H44-H194 Nε / H212-H242 Nδ	250
Total					1.4 μs

Reporting the H-bond results

Results for H-bonding presence is crystal structures are based on a single structure. Results for MD simulations are reported as a percentage (%). The percentage represents the time an H-bond is sampled i.e., it meets the geometrical criteria, across the trajectory used for analyses. The presence of an H-bond in a trajectory will be referred to as occurrence/occurrence rate/occupancy and it is normalized by the length of the trajectory segment.

5.4 Conserved motifs in large crystal structure datasets

Bridge was used to compute the H-bond graphs of all high-resolution protein structures included in *Set-high, Set-low, Set-highU*, its subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio*, and their underlying H-bond motifs. H-bonding between Ser/Thr hydroxyl groups and carboxylate groups of Asp/Glu amino acids residues is present in all protein structures of *Set-high* and *Set-highU*, except for Proton-translocating transhydrogenases and Magnesium ion-transporter-E (Figure 5.4a). In the second and fourth chains of the transhydrogenase proton channel (PDBID:5UNI) an aspartate is found barely within H-bonding criteria to a threonine (D80-T79, 3.48 Å). They are located in the loop linking helix 3 and 4 in the cytoplasmic side of the membrane. In the magnesium channel MgtE (PDBID:4U9N), chain A, S334 of helix 2 and E359 of helix 3 are found within 3.93 Å apart, which constitutes them not H-bonded

according to the criteria used. Thought their orientation in chain A suggests that they could be H-bonded during dynamics. In chain B the orientation in the x-ray crystal structure does not appear favorable for H-bonding.

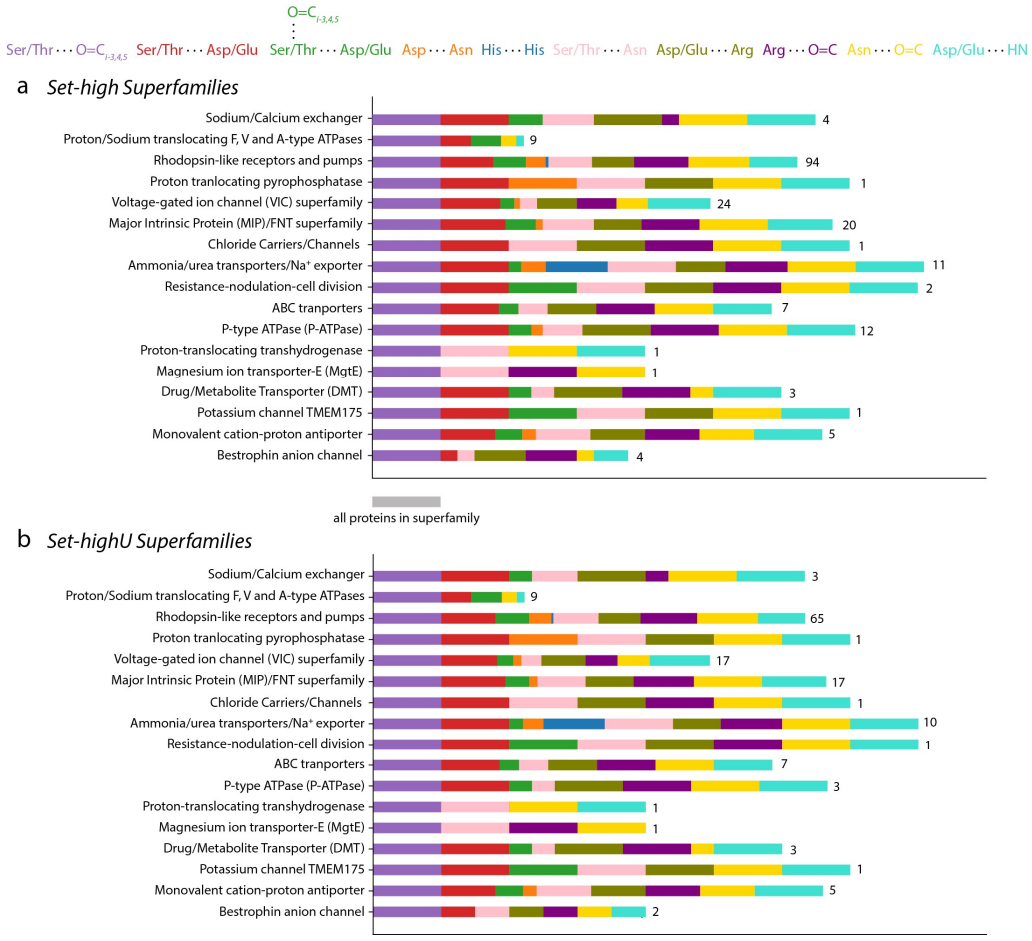

Figure 5.4. Summary of all H-bond motifs identified for protein structures of *Set-high* and *Set-highU*. The distributions of presence percentage of each H-bond motif are presented for each superfamily in sets *Set-high* and *Set-highU*. (a, b) H-bond motif presence for proteins in the superfamilies of *Set-high* (a) and *Set-highU* (b). The length of the gray bar indicates presence of an H-bond motif in all protein structures included in a superfamily. An index of the color-coded H-bond motifs is shown. For every superfamily bar, the number of their protein members is annotated. Adapted from ref. [226].

When using the TCDB classification, each *protein-group* of *Set-high* and *Set-highU* was found to contain at least one carboxylate-hydroxyl motif (Figure 5.5, Figure 5.6, Figure 5.7) with a range of occurrence 88-93%, while the intrahelical carboxylate-backbone carbonyl of the *i*-3,4,5 relative position is found in every protein of every *protein-group*. It is the most widely represented motif in this search (Figure 5.4, Figure 5.5). It is very likely that this is due to the abundance of Serine/Threonine amino acid residues (Figure 5.10), as well as that the second component of the motif is a backbone carbonyl. The combined carboxylate-Ser/Thr & Ser/Thr-backbone carbonyl of the *i*-3,4,5

relative position is a more selective motif and is not sampled as frequently as compared to its individual components. For the *protein-groups* of *Set-high* the range of occurrence is 21-48% (Voltage-gated ion channels and Rhodopsin-like receptors and pumps respectively, Figure 5.5a, Figure 5.6a). In the Rhodopsin-like receptors and pumps *protein-group* there are 45 proteins with the combined motif (Figure 5.6a) in *Set-high* and 32 proteins in *Set-highU*, which makes up for 49% occurrence in the set. When dissecting *Set-highU* into *Set-high-mr* and *Set-high-gpcr* it can be seen that out of the 32 proteins, 21 are microbial/algal rhodopsins and 11 are A, B, C, F family GPCRs (Figure 5.7a, b). Subsets *Set-high-hemo* and *Set-high-helio* do not have any combined motif (Figure 5.7b).

The location of carboxylate-hydroxyl motifs in two superfamilies of structures of *Set-high* was inspected closely. The Rhodopsin-like receptors and pumps (Figure 5.8d), and the super-family Proton or Sodium translocating F-type, V-type and A-type ATPases (Figure A.10). In both superfamilies, structures tend to have one site in which a Serine/Threonine hydroxyl group participates in both inter-helical H-bonding with an Aspartate/Glutamate carboxylate group, and in intra-helical H-bonding with a backbone carbonyl group. In most of these structures, the carboxylate-hydroxyl motif is located lose to the center of the membrane core (Figure A.10).

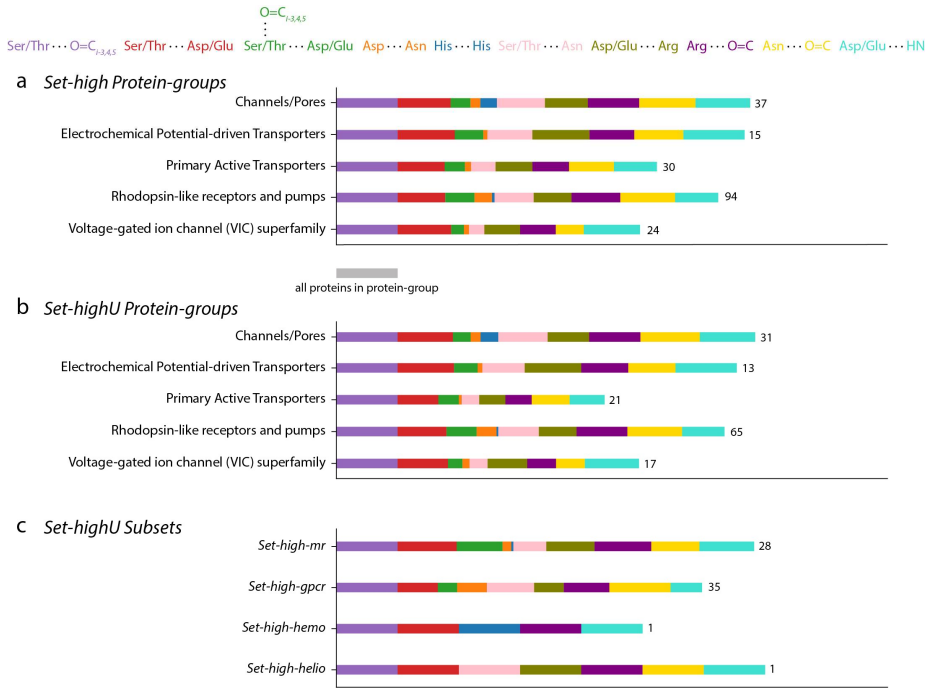

Figure 5.5. Summary of all H-bond motifs identified for protein structures of *Set-high, Set-highU* and the underlying subsets of *Set-highU*. The distributions of presence percentage of each H-bond motif are presented for each *protein-group* in sets *Set-high*, *Set-highU* and the subsets of *Set-highU*. (a-c) H-bond motif presence for proteins in the *protein-groups* of *Set-high* (a), *Set-highU* (b) and the subsets of *Set-highU* (c). The length of the gray bar indicates presence of an H-bond motif in all protein structures included in a *protein-groups*. An index of the color-coded H-bond motifs is shown. For every *protein-group* bar, the number of their protein members is annotated. Adapted from ref. [226].

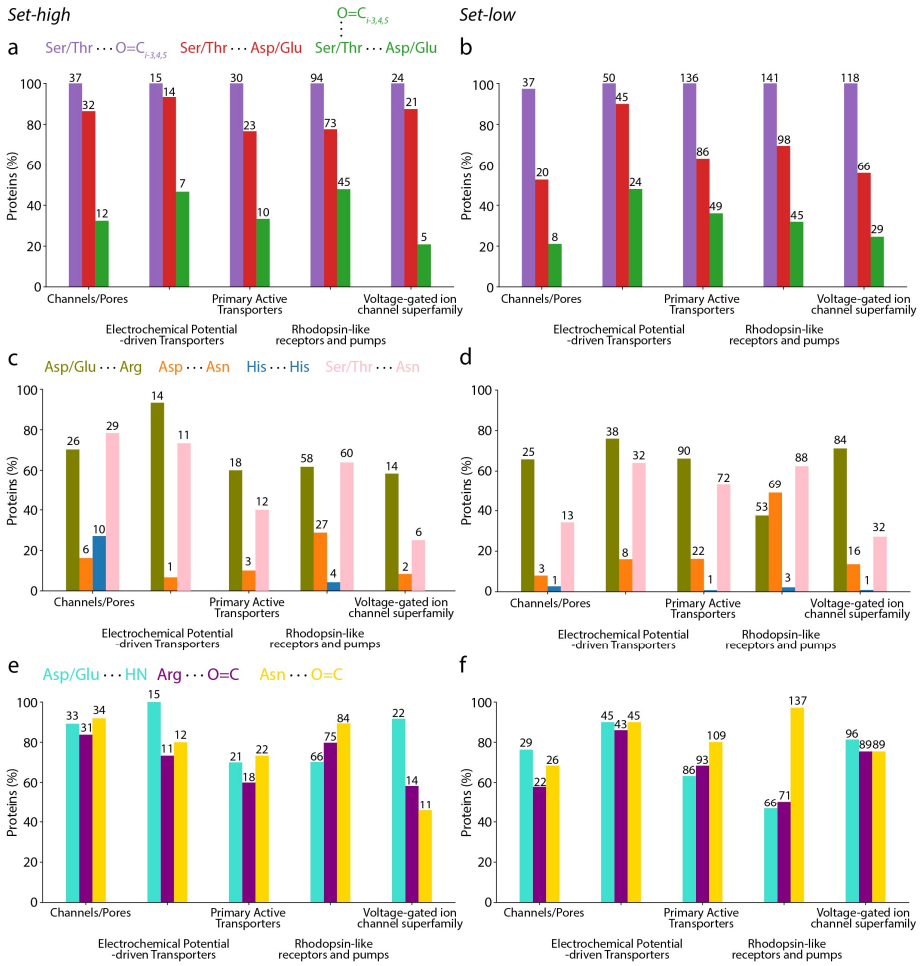

Figure 5.6. Summary of H-bond motif presence in sets *Set-high* and *Set-low*. The percentages of structures that contain an H-bond motif as indicated by the index, is given per *protein-group* for *Set-high* (left) and *Set-low* (right). The number of structures containing each motif is annotated on the corresponding vertical bar, and the total number of structures per *protein-group* in Table 5.4. (a, b) Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position, Asp/Glu-Ser/Thr, and Asp/Glu-Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position H-bonds in *Set-high* (a) and *Set-low* (b). (c, d) Asp/Glu-Arg, Asn-Asp, His-His, Ser/Thr-Asn motifs in *Set-high* (c) and *Set-low* (d). (e, f) Asp/Glu - backbone amide, Arg - backbone carbonyl, and Asn - backbone carbonyl motifs for superfamilies included in *Set-high* (e) and *Set-low* (f). Adapted from ref. [226].

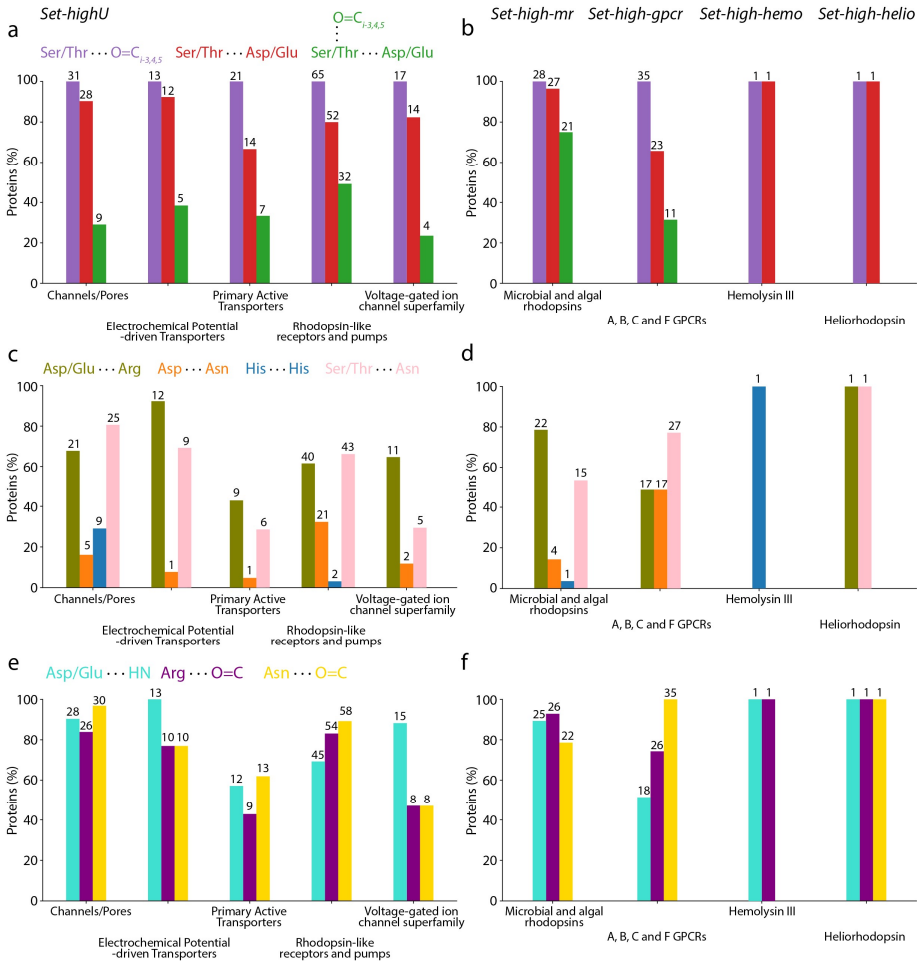

Figure 5.7. Summary of H-bond motif presence in sets *Set-highU* and subsets of *Set-highU*. The percentages of structures that contain an H-bond motif as indicated by the index, is given per *protein-group* for *Set-high* (left) and *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (right). The number of structures containing each motif is annotated on the corresponding vertical bar, and the total number of structures per *protein-group* of *Set-highU* and its subsets in Table 5.4. (a, b) Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position, Asp/Glu-Ser/Thr, and Asp/Glu-Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position H-bonds in *Set-highU* (a) and subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (b). (c, d) Asp/Glu-Arg, Asn-Asp, His-His, Ser/Thr-Asn motifs in *Set-highU* (c) and subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (d). (e, f) Asp/Glu - backbone amide, Arg - backbone carbonyl, and Asn - backbone carbonyl motifs for superfamilies included in *Set-highU* (e) and subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (f). Adapted from ref. [226].

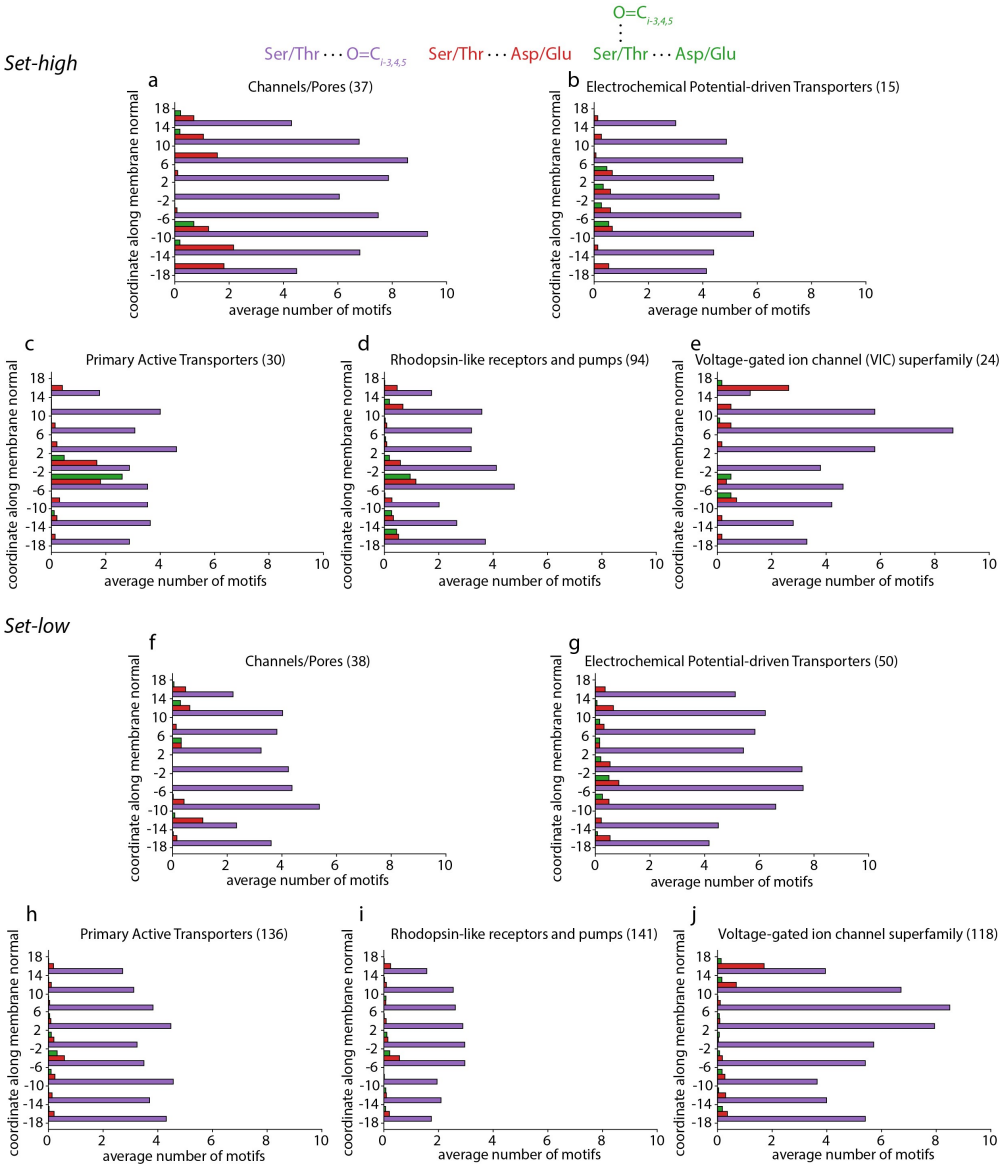

Figure 5.8 Distribution of Asp/Glu-Ser/Thr carboxyl-hydroxyl and Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position and combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs along the membrane normal for protein structures in sets *Set-high* and *Set-low*. H-bond motifs are presented per *protein-groups* for sets *Set-high* and *Set-low*. The number of proteins per *protein-group* is reported in Table 5.4. (a-j) Distribution along the membrane normal for Channels/Pores in *Set-high* (a) vs. *Set-low* (f), Electrochemical Potential-driven Transporters in *Set-high* (b) vs. *Set-low* (g), Primary Active Transporters in *Set-high* (c) vs. *Set-low* (h), Rhodopsin-like receptors and pumps in *Set-high* (d) vs. *Set-low* (i), and Voltage-gated ion channels (VIC) in *Set-high* (e) vs. *Set-low* (j). Adapted from ref. [226].

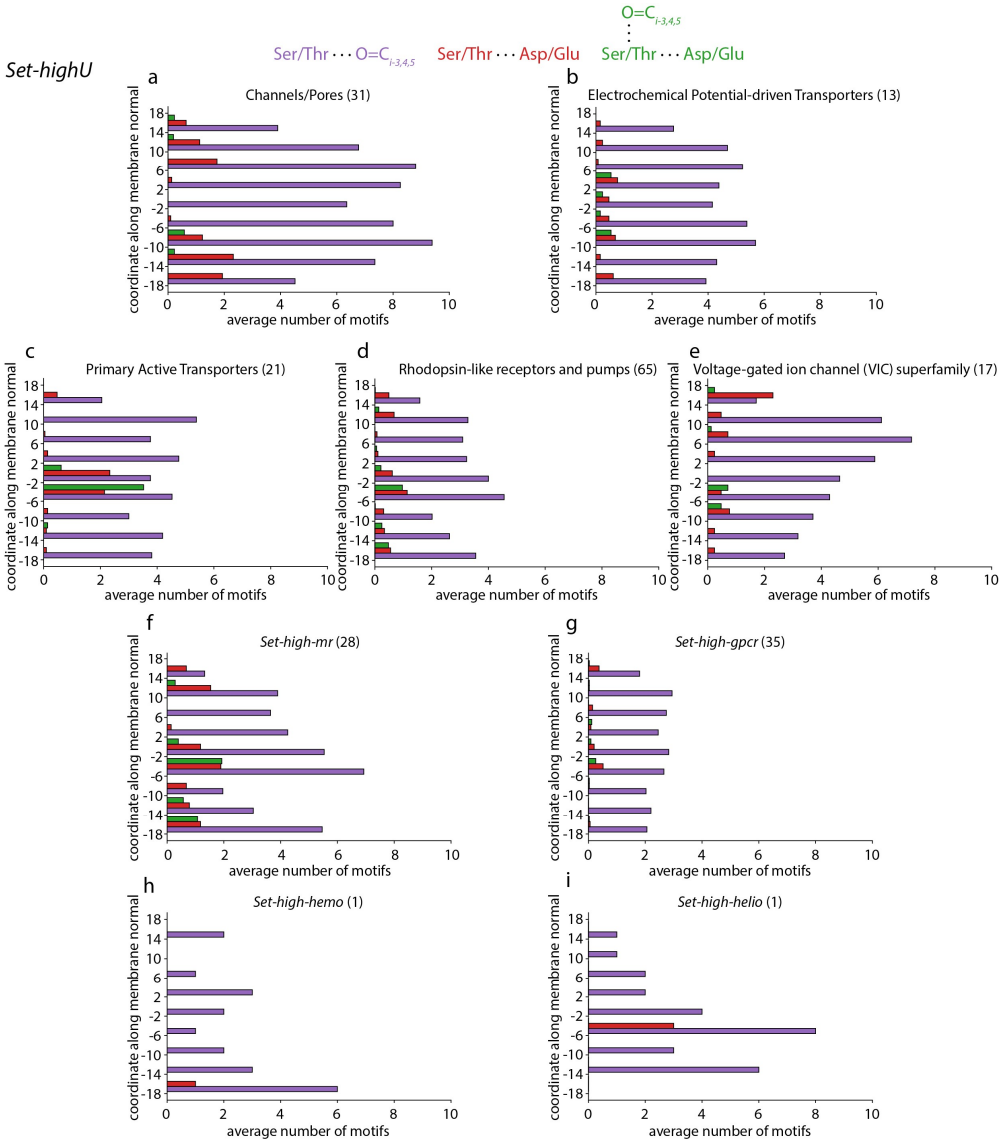

Figure 5.9. Distribution of Asp/Glu-Ser/Thr carboxyl-hydroxyl and Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position and combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs along the membrane normal for protein structures in sets *Set-highU* and its subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio*. (a-e) Distribution along the membrane normal for Channels/Pores (a), Electrochemical Potential-driven Transporters (panel b), Primary Active Transporters (c), Rhodopsin-like receptors and pumps (d), and Voltage-gated ion channels (VIC) (e). (f-i) Distribution along the membrane normal for *Set-high-mr* (f), *Set-high-gpcr* (g), *Set-high-hemo* (h) and *Set-high-helio* (i). Adapted from ref. [226].

Amino acid residues along the membrane normal

A very important aspect of this study is the correlation of the detection of different H-bond motifs according to their position along the membrane normal. The event of finding hydrophilic and bulky amino acid residues in the core of the TM region comes

with a large amount energy tradeoff. This concept is presented as the hydrophobicity scale [18, 204].

From the distributions of the motifs along the membrane normal along with the distributions of functionally important groups, a tight relation between the positions of the inter-helical hydroxyl-carboxylate/intra-helical + inter-helical hydroxyl-carboxylate motif and the positions of the functional groups is observed. In every *protein-group*, specifically in the area of the membrane core (-6 to 6 Å), it is observed that the combined motifs follow exactly the pattern of detection of its individual amino acid residues (Figure 5.8, Figure 5.9, Figure 5.10, Figure A.12). Moreover, the number of detected interhelical carboxylate-hydroxyl motifs are almost identical to the carboxylate groups detected for the same *protein-groups*. With the only exceptions being the Primary active transporters and the Voltage gated ion channel superfamily in position ranges of -2 to 6 Å, the presence of carboxylates in the region is the detection limiting step of finding, or not, those motifs, since an abundance of Serine/Threonine groups is found in every position range compared to the presence of carboxylate amino acid residues (Figure A.11).

Moreover, a direct comparison of the motifs and functional groups detected shows that, specifically for membrane core, the presence of carboxylates almost always results in the presence of H-bond motifs of that nature. This generalization does not appear to apply outside the membrane core where a higher presence of carboxylate groups is computed. Though the number of carboxylate groups are higher where $6 < |z| < 18$ Å, the presence of hydroxyl-carboxylate H-bond motifs is not reflected in the same way as previously. Only in the case of the Rhodopsin-like receptors and pumps superfamily in the position range of -10 to -18 Å, the same detection pattern of number of carboxylates and carboxylate motifs can be observed. It is understandable to observe such divergence, since approaching the edges of the membrane the secondary structure is mostly comprised of loops and turns and not solid helical structure for which those motifs are found. Every *protein-group* shows the presence of carboxylate related motifs in the core of the membrane with the exception of the Channels/Pores (Figure 5.8Aa), where almost no carboxylate groups are found in the whole *protein-group* (Figure 5.10a, Figure A.11). For the remaining groups, the detection pattern could provide insight into the characterization of proteins of yet unknown functionality or provide insight for the direction of the characterization. In the same group, an average of 6 carboxylate groups are found in the range of -14 to -18 Å (cytoplasmic side), but only a ~2 average Interhelical hydroxyl-carboxylate motifs (Figure 5.8) although the presence of hydroxyl groups is very high in the area (~10, Figure A.12a). Thus, the carboxylates in the region might be of other biological importance besides interhelical coupling and possibly proton transfer, also given the fact that the region is most probably consisting of loops and turns. The same pattern show the Rhodopsin-like receptors and pumps and the Voltage-gated ion channels on the extracellular side of the membrane (14 to 18 Å, Figure 5.8d-e, Figure 5.10d-Ee, Figure A.11, Figure A.12). The presence of the carboxylate groups in the region might be a key identifier of their function of the proteins themselves, through pH sensing, proton storing, proton translocating through carboxylate antennas or ion selectivity.

The distribution of carboxylate groups in the Channels/Pores, Electrochemical Potential-driven Transporters and Voltage-gated ion channels is largely symmetric or bell-shaped. The limited -to almost non-detectable- presence in the core of the membrane is those groups is in agreement with the side-chain hydrophobicity scale [18, 204]. Although the Rhodopsin-like receptors and pumps do not follow this exact shape of distribution, the average number of carboxylate groups buried in the membrane core is similar and comparable to the above groups. The Primary active transporters group shown an increased number of carboxylate groups in the range of -2 to -6 Å, which according to the side-chain hydrophobicity scale would result in a large energetic penalty to be paid. Thus, the increased presence of Aspartate/Glutamate groups buried in the membrane core could be crucial for the biological function of the proteins and a main characteristic of the group. Histidine groups show largely a uniform distribution resembling slightly the ones of the carboxylate groups in shape (Figure 5.10, Figure A.11, Figure A.13).

The presence of arginine amino acid residues shows also a most interesting pattern. The average number of Arginine groups is double of Serine/Threonine and quadruple of Aspartate/Glutamate in the Primary active transporters, Electrochemical Potential-driven Transporters and Voltage-gated ion channels in the positions range 14 to 18 Å (Figure 5.10b,c,e, Figure A.11b,c,e, Figure A.14b,c,e). The Rhodopsin-like receptors and pumps and Channels/Pores groups also show an increased presence of Arg groups in the same range (Figure 5.10a,d, Figure A.11a,d, Figure A.14a,d).

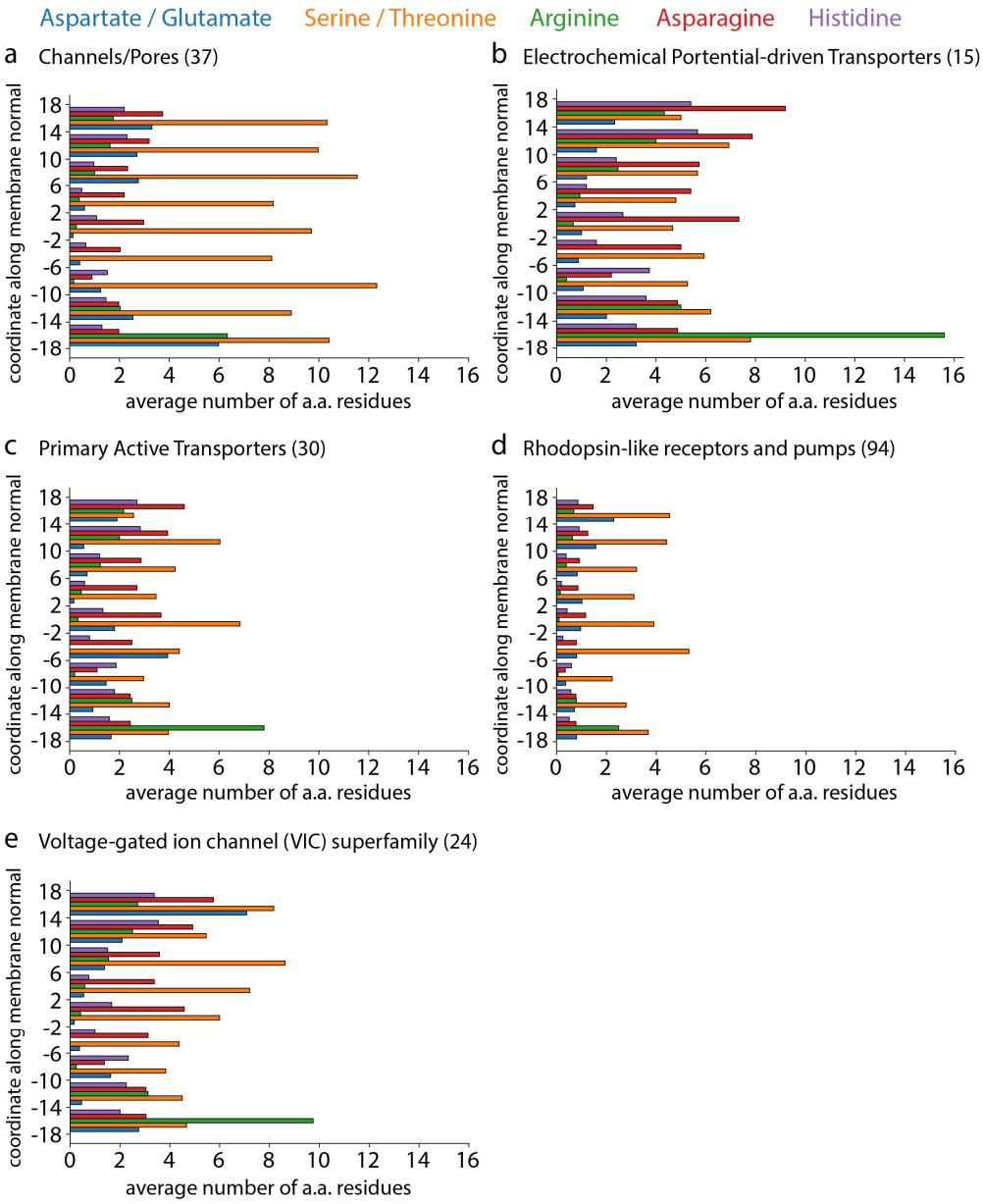

Figure 5.10. Amino acid (a.a) residue location distributions along the membrane normal. The distributions of H-bonding amino acid residues are presented for members of *Set-high*. Amino acid residues are color-coded. Aspartate/Glutamate groups are shown in blue, Serine/Threonine groups in orange, Arginine groups in green, Asparagine in red and Histidine in purple. (a-e) Distribution of a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e).

H-bond motifs of Serine/Threonine amino acid residues

The intra-helical H-bond motif between Serine/Threonine hydroxyl groups and backbone carbonyls of the *i*-3,4,5 relative position is present along the entire length of the membrane protein (Figure 5.8, Figure 5.9). When present, in the in primary active transporters, electrochemical potential-driven transporters, and rhodopsin-like receptors and pumps, H-bonds between Aspartate/Glutamate and Serine/Threonine tend to be close to the center of the membrane core (Figure 5.8b,c,d, Figure 5.9b,c,d), in the channels/pores and voltage-gate ion channels *protein-groups* are relatively close to the membrane interface (Figure 5.8a,e, Figure 5.9a,e). The rhodopsin-like receptors and pumps *protein-group* features many Aspartate/Glutamate and Serine/Threonine motifs in the center of the membrane core as well as towards the cytoplasmic and extracellular sides of the membrane, where prime examples of amino acid residues of critical importance for the mechanism of the protein is found, such as the proton release group of BR [220]. The Archaerhodopsin-2 dimer [264] and trimer [265] share a common motif between T95 and D120 ($D156^{ChR2}$). The Archerorhodopsin-2 trimer [265] has an additional motif between T50-D101 ($D96^{BR}$, $H134^{ChR2}$). A motif that is widely represented in the Rhodopsin-like receptors and pumps protein-group is between T90-D115 and T46-D96 in BR. In the |z|=6 Å along the membrane normal the T90-D115 motif is found across many rhodopsins when considering the *H. salinarium* BR numbering, such as Archaerhodopsin-3 [266]), BR from *H. marismortui* [267], BR from *H. walsbyi* [268], crudorhodopsin-3 [269], halorhodopsin from *N. pharaonic* [270, 271], , halorhodopsin from *N. salinarium* [242], coccomyxarhodopsin [272], *E. sibiricum* rhodopsin [273], Acetabularia rhodopsin I [243], thermophilic rhodopsin from *R. xylanophilus* [274]. The $T46^{BR}$-$D96^{BR}$ motif is found across crudorhodopsin-3 [269], coccomyxarhodopsin [272], thermophilic rhodopsin from *R. xylanophilus* [274] and BR from *H. walsbyi* [268]. The light-driven Na^+ pump KR2 [275, 276], features a motif between S70-D116 ($T90^{BR}$). Similar motifs are sampled for 10 GPCR's, including Jumping spider rhodopsin [277]. The *F. nucleatum* [278] and *I. tartaricus* [279] F_1F_o-ATP synthases c-ring features 11 c-subunits and an ion-binding site forming between adjacent subunit pairs. E32 (Q32 in the *I. tartaricus* [279]), V63, A64, E65, S66 and a water molecule form the Na^+ binding site. E65 is highly conserved and H-bonds to S66 on the adjacent c-subunit, while S66 forms an intrahelical H-bond motif with V63. The same arrangement and interactions are found in the hybrid F / V-rotor ring of the Na^+-coupled ATP synthase from *A. woodii* [280]. In the 9 single-hairpin subunits the motif in the Na^+ binding site is formed between E62-T63-V60. Between the double hairpin subunit and the adjacent single hairpin subunit the motif is formed with E79-T63-V60. In the V-type Na^+-coupled ATP synthase from *E. hirae* [281] the motif is formed between E139-T64-L61.The NavAb voltage-gated Na^+ channel [282] and the voltage-gated Na^+ channel in the activated open conformation NavMs [283] feature a common motif detected between D81 of the voltage sensing domain and S112 in the beginning of the linker. NavMs has an additional motif between E239-T244 in the c-terminus.
A complete list of the proteins discussed is this section, along with their PDB IDs is found in Table A.5.

H-bond motifs of Histidine and Asparagine amino acid residues

In the dataset, each of the three protomers of the ammonia channel AmtB has one inter-helical His-His H-bond located at the center of the transmembrane domain of the protein (see H168 and H318, Figure A.7a). H168 and H318 are thought to H-bond to ammonia substrate molecules and hinder transport of NH_4^+ [129]. The latter was found to be pH dependent [129]. Wang and co-workers [284] presented a mechanism where they suggested that the highly conserved His-His motif acts as a proton relay in the transport of ammonium. The ammonium cation transfers the excess proton to H168, and the ammonia molecule diffuses down the pore. With a rapid proton transfer step from H168 to H318, the ammonia molecule is re-protonated from H318 and diffuses to the bulk as ammonium [284]. Most of the His-His motifs detected in the dataset belong to proteins of the ammonium channel transporter superfamily, where the highly conserved histidines are an essential part of the transport mechanism in ammonia channels [129, 284]. Other proteins in that superfamily include ammonium transporters, ammonium sensor/transducer, Rhesus glycoprotein and urea transporters.

In rhodopsin-like proteins, a Histidine-Histidine H-bond pair is found between H199 and H201 in the Sphingosine 1-phosphate receptor 1 (PDBID:3V2Y [285]) from family A of the GPCRs. In the same family, in the Lysophosphatidic acid receptor 1 (PDBID:4Z35 [286]) another histidine pair is found; H147-H227, between TM-helices 3 and 5. The Calcitonin type 1 (PDBID:6UUS [287]) receptor of the GPCR Secretin B family has a continuous network of three histidines, H149-H334-H370 being H-bonded in series. This could be part of a pH-sensing signaling cascade. The Ammonia channel AmtB features a similar histidine motif between two amino acid residues, and they are involved in proton transfer during the transport of ammonia. Other examples of proteins belonging to the rhodopsin-like receptors and pumps group that feature histidine pairs are the of the Alkaline ceramidase 3 (PDBID:6G7O [288]) with the H81-H221 pair, anabaena sensory rhodopsin (PDBID:1XIO [289]) with two histidine pairs H21-H219 and H8-H69 and a most interesting histidine cluster found in the Adiponectin receptor 2 (PDBID:5LWY [260]) where a triangular H-bond network between H202-H348-H352 is formed. In total, 10 proteins in the *Set-high* belong to ammonium channel transporter superfamily, 4 proteins belong to the Rhodopsin-like receptors and pumps and 1 belongs to the Bestrophin anion channel superfamily.

The proton translocating pyrophosphatase (PDBID:4A01, [131]) has, in each of the two monomers an inter-helical Aspartate/Glutamate H-bond between D294 and N738 motif (Figure A.7b). This H-bond is likely important for the functioning of the transporter, because N738 could stabilize water molecules of the proton transfer pathway [131]. The protonation state of D294 is thought to be influenced by K742 [131]. An ammonium tranceptor that contains the highly conserved his-his motif, is also found to have an Aspartate-Asparagine motif between D365-N246 in the extracellular side of the membrane, close to the N-terminus in the opening of the pore [254]. There is another motif found between D273-N321 that in located in the cytoplasmic side of the membrane, but not in the vicinity of the pore. The crystal structure suggests that this motif is

contributing to helical stability between TM7 and TM9. The majority of the Aspartate-Asparagine motifs though, are found within the Rhodopsin-like receptors and pumps with 25 proteins containing at least one motif. Four of them belong to the microbial and algal rhodopsins family (viral rhodopsin, a chloride pumping MastR, chloride importer and its mutant). Twenty belong to the G-protein-coupled receptor family and one in the GPCR secretin (B) family.

The presence of H-bond motifs in static structures of *Set-highU* and *Set-low* is qualitatively very similar to that discussed above for *Set-high* (Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9). Although the Serine/Threonine-Aspartate/Glutamate + Serine/Threonine - backbone carbonyl of the *i*-3,4,5 relative position is a combination of two motifs, it is found consistently between 21-48% percent of proteins in both *Set-high* and *Set-low* (Figure 5.5, Figure 5.6) and 24-40% in *Set-highU* (Figure 5.4, Figure 5.5, Figure 5.7). The most represented *protein-group* in *Set-high*, 48% of the members of the Rhodopsin-like receptors and pumps feature at least one combined motif. The absolute number of structures is 45 in that *protein-group*. Channels/Pores and the Primary active transporters feature 12 and 10 proteins respectively with the combined motif (32-33% of the *protein-group* respectively).

The Histidine-Histidine motif is not highly sampled in the datasets, with the highest presence found in *Set-high*, in the Channels/Pores group (Figure 5.5, Figure 5.6a), detected at 27% of the members of the *protein-group*. More detailed analysis of the presence of the motif in the individual superfamilies under the *protein-groups* show that all 10 proteins that feature the Histidine-Histidine motif in the Ammonia/urea transporters/ Na^+ exporter superfamily (Figure 5.4a). It has been shown that the Histidine-Histidine motif is vital for the mechanism of action in proteins of that superfamily and participates in proton transfer. Aspartate-Asparagine motifs are also not highly detected except in the case of the Rhodopsin-like receptor and pumps (Figure 5.4a, Figure 5.6a,b, Figure 5.7a), with 27 proteins (29%) in *Set-high,* 21 proteins (32%) in *Set-highU* and 69 proteins (49%) in *Set-low*.

The salt-bridging motif Aspartate/Glutamate-Arginine is highly sampled in all protein-groups (Figure 5.4c,d, Figure 5.5c,d) with a 58-93% presence in *protein-groups* of *Set-high,* 43-92% in *Set-highU* and 38-76% in *Set-low*. The Asparagine-Serine/Threonine motifs follows qualitatively the detection of the salt-bridge motif with 25-78% in *Set-high* (Figure 5.6c), 29-81% in *Set-highU* (Figure 5.7c) and 27-64% in *Set-low* (Figure 5.6d). The backbone containing motifs are the motifs highly sampled in both datasets (Figure 5.6, Figure 5.7), but it could be expected from the abundance of H-bond partners, having one of them being the backbone.

H-bond motifs as a part of long-distance H-bond pathways in static protein structures

The nature of the H-bond motif can provide insight to its potential connectivity and role in the biological function of a protein. Histidine-Histidine motifs are mostly contained in shorter networks in terms of path length. There is the exception of path

length of L=15 that is found, with 12 pathways which feature that length (Figure A.18). All 12 paths are found in one network in the Ammonium sensor/transducer (PDB ID:6EU6 [290]). Asparagine-Serine/Threonine and Arginine-Aspartate/Glutamate motifs have rather skewed distributions of path lengths with most of the samples having an L value between 3 and 6. Both of those examples though show path lengths of 15 or higher, with 27 being the highest length that was computed in the sodium pumping rhodopsin KR2 pentamer, where an extended H-bond network of 44 amino acid residues connecting the 5 monomers. A significant difference between Set-high and Set-low is found in the Histidine-Histidine motif, where in Set-low, only 3 paths of length L=5 are sampled. Combined with the findings that Histidine-Histidine H-bond motifs are very infrequent in α-helical transmembrane proteins (Figure 5.6c,d, Figure 5.7c,d), it could be assumed that the motif is highly dependent on the resolution. I also find that the motif retains short pathways if the pathway is passing through the motif i.e., the motif is internal nodes. Pathways that start from A Histidine-Histidine motif are even less frequently sampled (Figure 5.12b,f, Figure A.17b,f). In general, the results between *Set-high* and *Set-highU* are qualitatively and quantitatively similar. Additional analysis of pathways originating from H-bond motifs where the root note is peripheral (Figure 2.7a), shows that the distributions are qualitatively very similar (Figure 5.12, Figure A.17), but there fewer number of paths. This could be expected since the peripheral nodes that participate in H-bond motifs are numerically fewer than the internal ones.

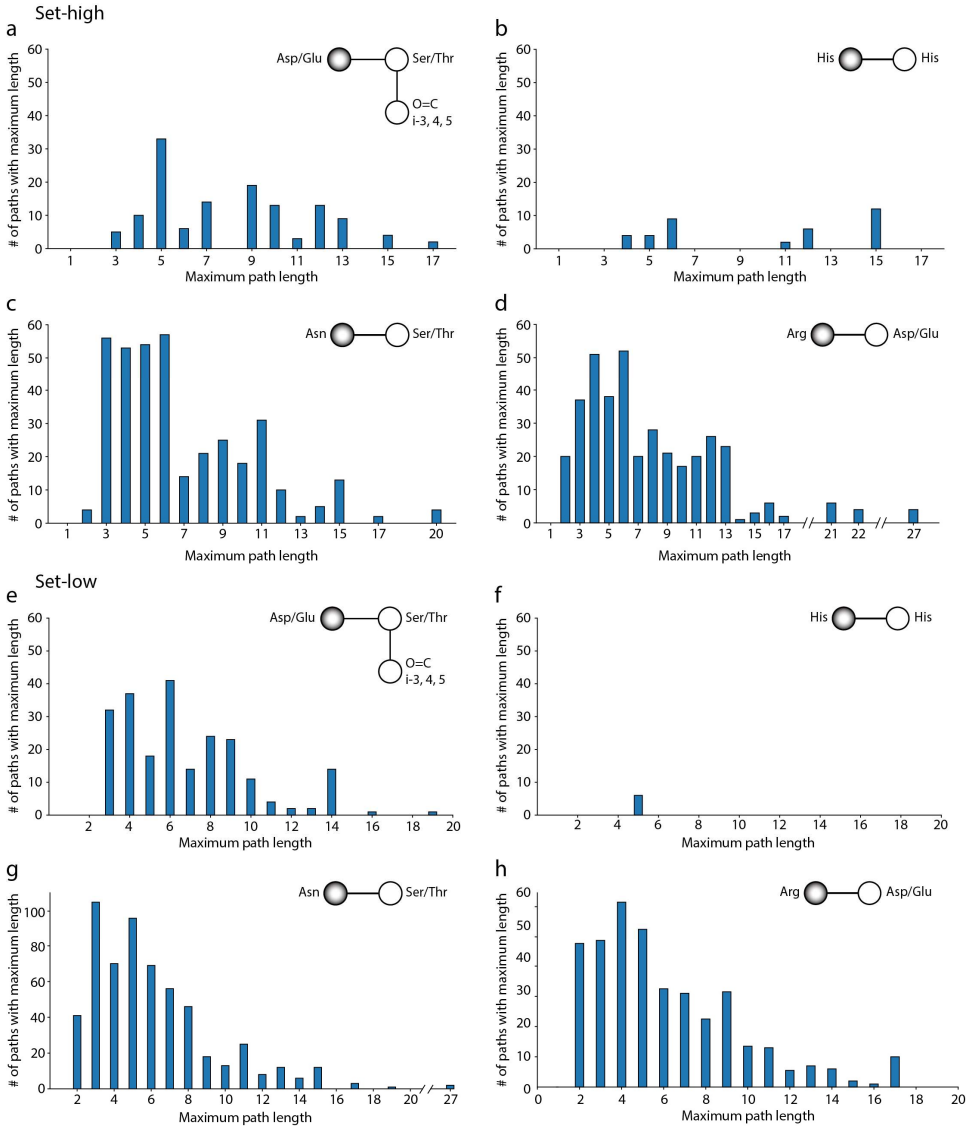

Figure 5.11. Shortest path length computations for H-bond motifs in two crystal structure data sets. The root nodes are unique entries per amino acid residue that participate in H-bond motifs and are *internal* to the H-bond network (Figure 2.7, Figure 2.8). Path length distributions are shown for the carboxylate groups participating in the combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs in *Set-high* vs. *Set-low* (a, d). Histidine amino acid residues participating in His-His motifs in *Set-high* vs. *Set-low* (b, f). Asparagine amino acid residues participating in Asn-Ser/Thr motifs in *Set-high* vs. *Set-low* (c, g). Arginine amino acid residues participating in Arg-Asp/Glu motifs in *Set-high* vs. *Set-low* (d, h). Additional analysis for sets *Set-highU* and subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* is found in Figure A.16. Adapted from ref. [226].

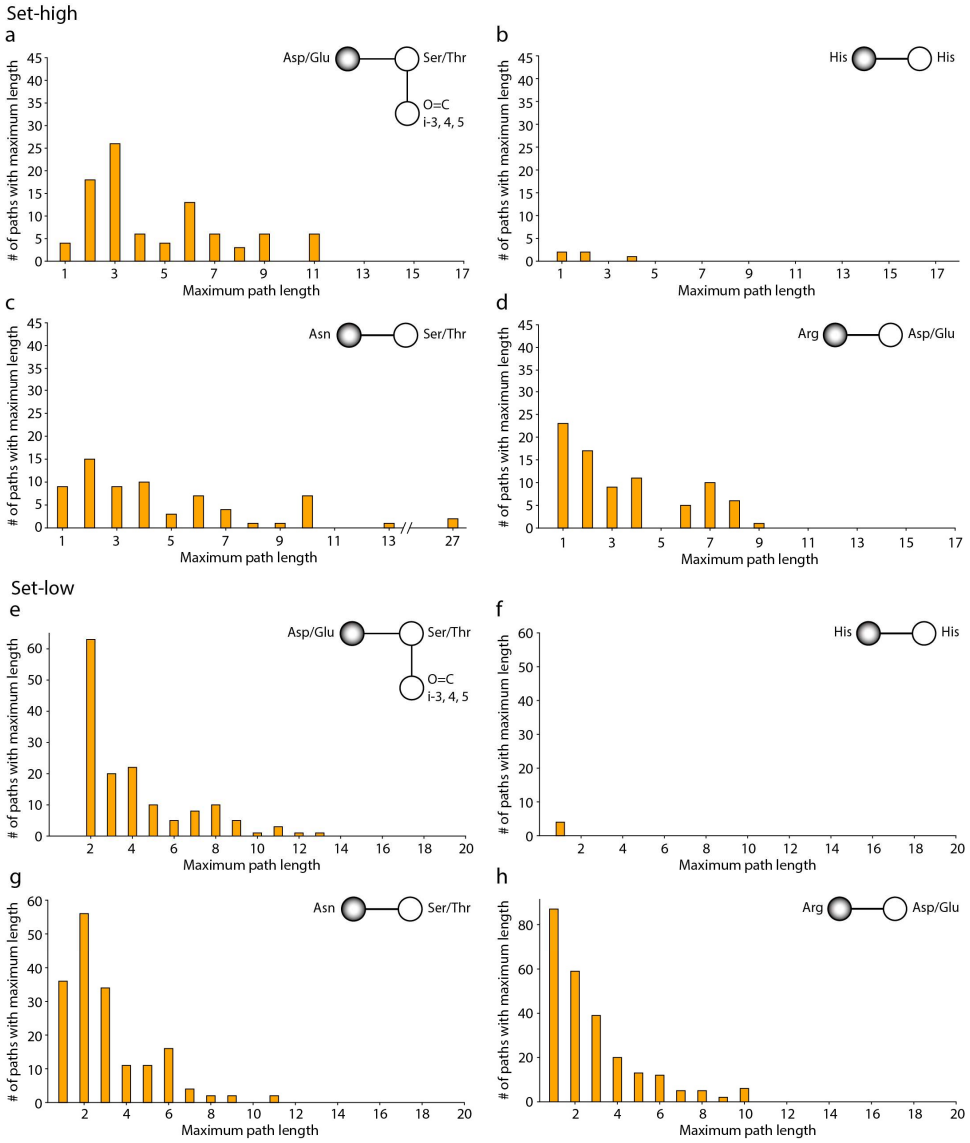

Figure 5.12. Shortest path length computations for H-bond motifs in two crystal structure data sets. The root nodes are unique entries per amino acid residue that participate in H-bond motifs and are *peripheral* to the H-bond network (Figure 2.7, Figure 2.8). Path length distributions are shown for the carboxylate groups participating in the combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs in *Set-high* vs. *Set-low* (a, d). Histidine amino acid residues participating in His-His motifs in *Set-high* vs. *Set-low* (b, f). Asparagine amino acid residues participating in Asn-Ser/Thr motifs in *Set-high* vs. *Set-low* (c, g). Arginine amino acid residues participating in Arg-Asp/Glu motifs in *Set-high* vs. *Set-low* (d, h). Additional analysis for sets *Set-highU* and subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* is found in Figure A.17.

5.5 Highly conserved networks in Aquaporin-1 (Aqy1)

Aquaporins are channels that can be permeated selectively by water molecules. The subfamily of aquaglyceroporins can also transport small organic molecules such as glycerol and urea. Aquaporins have two major structural characteristics, acting as

constricting sites. Firstly, the dual NPA motifs, whose conserved asparagines found in the half-helices are oriented towards the pore. Then, the ar/R motif, which is the tightest constriction point of the pore, located towards the extracellular side of the membrane is a selectivity filter both hydrophobicity and molecule size/steric restraint [291]. The selectivity filter features a highly conserved arginine amino acid residue, but the other components can vary, resulting in various radii for constriction pore. In Aqy1, accompanying the conserved arginine (R227 in Aqy1, PDBID:3ZOJ [292]), is an Nδ-protonated histidine which is believed to strip the water molecules off their solvation shells in order to be transported through the pore [293].

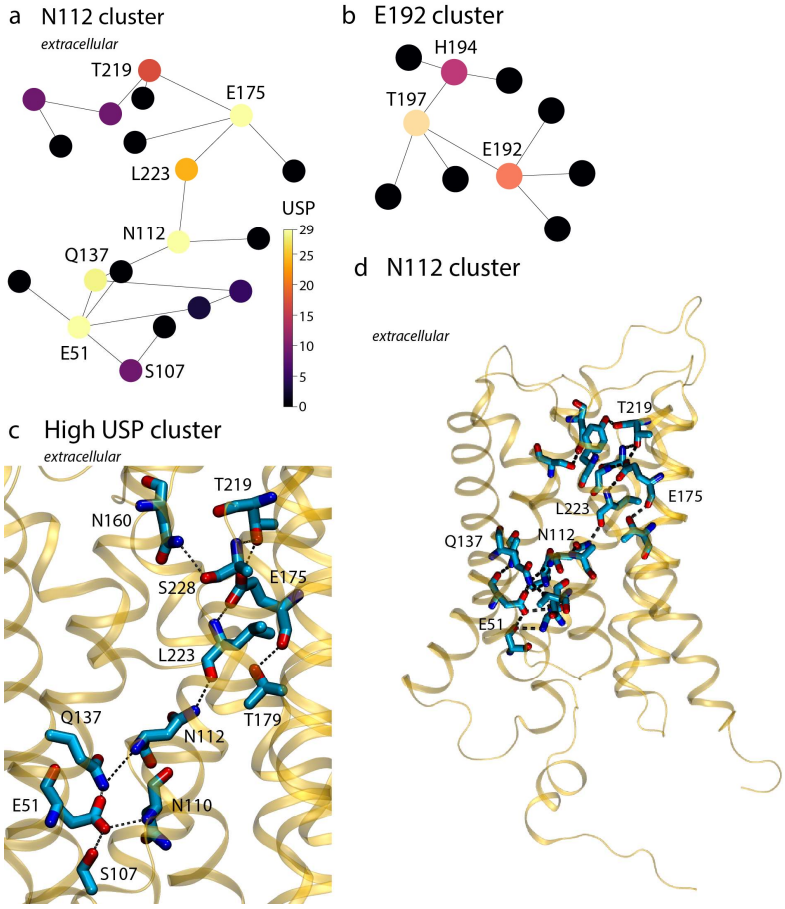

Figure 5.13. High Unique Shortest Paths clusters in the crystal structure of Aqy1. The USP computations are performed in direct protein-protein H-bond graphs. (a) The central N112 cluster is presented in a graph representation. The graph is color coded with a perpetually uniform color scale and it ranges from USP value of 0 to 29 which belongs to N112. (b) The cytoplasmic E192 cluster. (c) Molecular graphics of the H-bond map depicted in panel (a). (d) High-centrality/USP H-bond motif network in the crystal structure of Aqy1. The network depicted is a sub-network of panels a, c. Adapted from ref. [226].

The crystal structure of the aquaporin monomer [292] contains two large H-bond clusters, with 10 and, respectively, 19 protein groups (Figure 5.13a,b,d). The larger H-bond cluster spans through the length of the protein, bridging the extracellular and cytoplasmic

regions, and it contains the high USP groups N112, E51, E175 and the backbone of L223 (Figure 5.13a,d). N112 is one of the two asparagines of the dual-NPA motif, along with N224. L223 plays a crucial role with its backbone in the network connecting H-bonding to N112 with its carbonyl O-atom and to E175 with its amide N-atom. E175 is found in the transmembrane helix 4 and contributes with its second carboxyl-O atom to an Aspartate/Glutamate-Serine/Threonine motif with T219, which is found in the loop before the second half-helix (Figure 5.14). It also participates in a H-bond with its backbone carbonyl to a neighboring threonine, T179 which is found 4 positions further down the TM helix. On the cytoplasmic side of the network, the NPA-asparagine N112 is H-bonded to Q137. Q137 is sequentially H-bonded to E51, resembling an Aspartate/Glutamate-Asparagine motif, since Glutamine and Asparagine differ by a carbon atom in the side chain. The second Aspartate/Glutamate-Serine/Threonine is found between E51 and S107 in the cytoplasmic side of the membrane. Both motifs of this category are found connected indirectly through an H-bond network extending through the membrane normal. The network has a two main clusters in the extracellular and cytoplasmic sides of the membrane and are connected through a linear series of H-bonds, with more notable the amino acid residues E51, Q137, N112, L223 and E175 that span the membrane normal and then branch into the two clusters. Those 5 amino acid residues show important significance in terms of their connectivity in the network, showing the highest centrality BC and USP values.

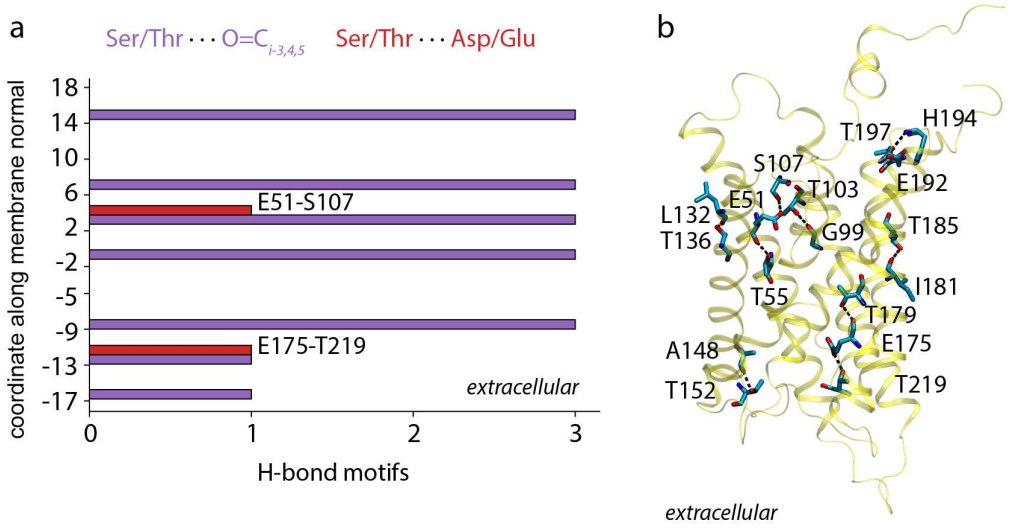

Figure 5.14. Serine/Threonine H-bond motifs in the crystal structure of Aqy1 (PDB ID: 3ZOJ). (a) Distribution of positions along the membrane normal for the Serine/Threonine-backbone carbonyl of the *i*-3,4,5 relative position (purple) and Serine/Threonine-Aspartate/Glutamate (red). (b) Molecular representations of the two motif categories shown in panel a.

Other motifs are found in the extended H-bond network of the crystal structure. Two Asparagine-Serine/Threonine motifs are detected with T116 and S228 located in the half-helices, while the H-bond partner N110 is in the preceding loop before the first half-helix and N160 is the loop that leads to TM-helix 4. N160 also participates in the

Asparagine-backbone carbonyl H-bond motif category with its Nδ atom, H-bonding to the backbone carbonyl of A221. N112 of the dual-NPA motif, participates in the H-bond N112-L223 that connects the two clusters in the extended H-bond network, which is characterized as an Asparagine-backbone carbonyl H-bond motif. The size of the internal H-bond clusters of aquaporin increases drastically when water molecules are included in the computations and the connectivity is expanded. In the selectivity filter R227 can is now bridged to H212 via a water molecule [130]. The USP computations show that the most highly connected nodes are now R105, T197, F101, Y27 and N160 among others. The clusters are now centered around R105 in the cytoplasmic side, where water molecules in the limits of the membrane mediate H-bonds into a circular cluster and N160 in the extracellular side of the membrane. The former N112 cluster limited to protein-protein H-bonds is now expanded in the presence of water and is found in the cluster of N160 in the extracellular side. (Figure A.20).

H-bond networks of the static Aqy1 structure are sampled in two lipid membrane environments

During MD simulations the Aqy1 structure remains stable, with Cα root-mean-squared deviations (RMSD) within 1 Å for all simulations performed (Figure A.21a-d), and with overall similar numbers of internal waters in each monomer (Figure A.21e-h). In the crystal structure the networks of the selectivity filter and the NPA motif are separated when excluding the water molecules from the networks. The connectivity of the amino-acid residues in the Selectivity Filter (SF) were investigated by means of direct H-bonds between sidechain-sidechain and sidechain-backbone. In all 4 MD simulations that were performed for this chapter, they are connected via the connections that R227 and N224 establish (Figure 5.15, Figure A.22, Figure A.23).

Part of the larger BC cluster detected in the crystal structure are the amino acid residues E51 and E175 that are the only ones to participate in very stable interhelical carboxylate-hydroxyl motifs with S107 and T219 respectively, with occurrence rates 100% in most cases. Topologically and structurally, they are found in the outskirts of the network, in opposite sides. They can be connected a relatively short pathway of length L= 4 and Joint Occupancy of ~20%. This pathway includes Q137, one of the two NPA asparagines; N112 [130], and L223. This pathway is found in both the X-ray structure and the MD simulations (Figure 5.15, Figure A.22).

At the selectivity filter of Aqy1, the sidechain of H212 H-bonds to the backbone carbonyl of L208. This H-bond is maintained during dynamics, but at difference occurrence rates (%). In *aq1_b* (POPE lipid bilayer) it is sampled at all four monomers but shows a large difference in the occurrence rate of Monomer-A at 39% of the time, while in monomers B, C and D is sampled at > 98% (Figure 5.15, Table 5.6). In *aq2_b* the H-bond is sampled only at two of the four monomers at 86% and 99% of the time. There are no other connections involving these amino-acid residues, thus it is suspected that the protein-lipid interactions that are posed by the different lipid molecules in *aq2_b* could be an influence on the dynamics of that H-bond.

Nearby a small H-bond cluster of 4 amino-acid residues in the vicinity of the selectivity filter is comprised by a mixture of sidechain-sidechain and sidechain-backbone H-bonds involving F158, N224, A226 and R227 (Figure 5.15). The cluster is maintained during dynamics with the N224-R227 connection being less frequent at 32.5-39.7% average occurrence rates, but it is highly expanded. In *aq1_b* and *aq2_b*, N224 forms a moderately stable H-bond (average of 66.3% and 59.9% respectively) with the backbone of L111, while R227 forms less stable H-bonds with the backbone of A221 (average of 24.7% and 32.8% respectively) and in turn A221 with N160 at 31.7% average occurrence frequency in *aq1_b*, while in *aq2_b* is sampled in three of the four monomers. With the introduction of those H-bonds, the initial 4-amino acid residue cluster in the selectivity filter is merged to the central High-BC cluster that was identified in the crystal structure (Figure 5.15, Figure A.22).

The intrahelical hydroxyl-carbonyl motif S228-N224 H-bond is found in three monomers at moderate occurrence rates (7.4-26.6%), in the case of *aq1_b*, whereas in the *aq2_b* is sampled only in two monomers at lower occurrence rates (8.3-17.2%) (Figure 5.15, Table 5.6, Table A.6, Table A.7).

Table 5.6. Occupancies of selected H-bonds of Aqy1 of the R227-N112 H-bond clusters sampled during MD simulations in a POPE and a mixed bilayer, with H44 and H194 Nε2-protonated. When the H-bond is sampled in all four Aqy1 monomers A-D, the total average occupancy is also reported, computed by averaging the monomer occupancies. For each H-bond the occupancy calculated as an average from the last 100ns of each simulation is reported. Adapted from ref. [226].

H-bond	*H-bond occupancy (%)*									
	POPE - H44/H194 Nε2					Mixed - H44/H194 Nε2				
	A	B	C	D	Average	A	B	C	D	Average
E51-S107	99.9	99.9	99.9	99.9	99.9	100	100	100	99.9	100
S107-N110	51.3	16.1	58.3	41.9	41.9	47.1	51.3	58.6	26.4	45.86
F158-R227	98	95.3	98	99.7	97.8	95.8	96.9	99.9	99.4	98.0
N160-A221	29.5	22.6	29.1	45.6	31.7	19.4	33.4	39.0	-	N/A
E175-T219	100	100	100	100	100	100	100	100	100	100
H212-L208	40	99	99	99	84.3	86	-	98	-	N/A
A221-R227	34.3	22.7	27.6	14.2	24.7	22.8	41.2	16	51.1	32.8
N224-L111	68.4	67.6	61.3	67.8	66.3	65.6	68.9	60.5	44.6	59.9
N224-A226	88.9	91.8	94.6	91	91.6	94.4	93.7	91.4	95.7	93.8
N224-R227	44.1	36.5	39.1	39	39.7	31.4	28.6	41	29	32.5
N224-S228	-	26.6	20.0	7.4	N/A	17.2	8.3	-	-	N/A

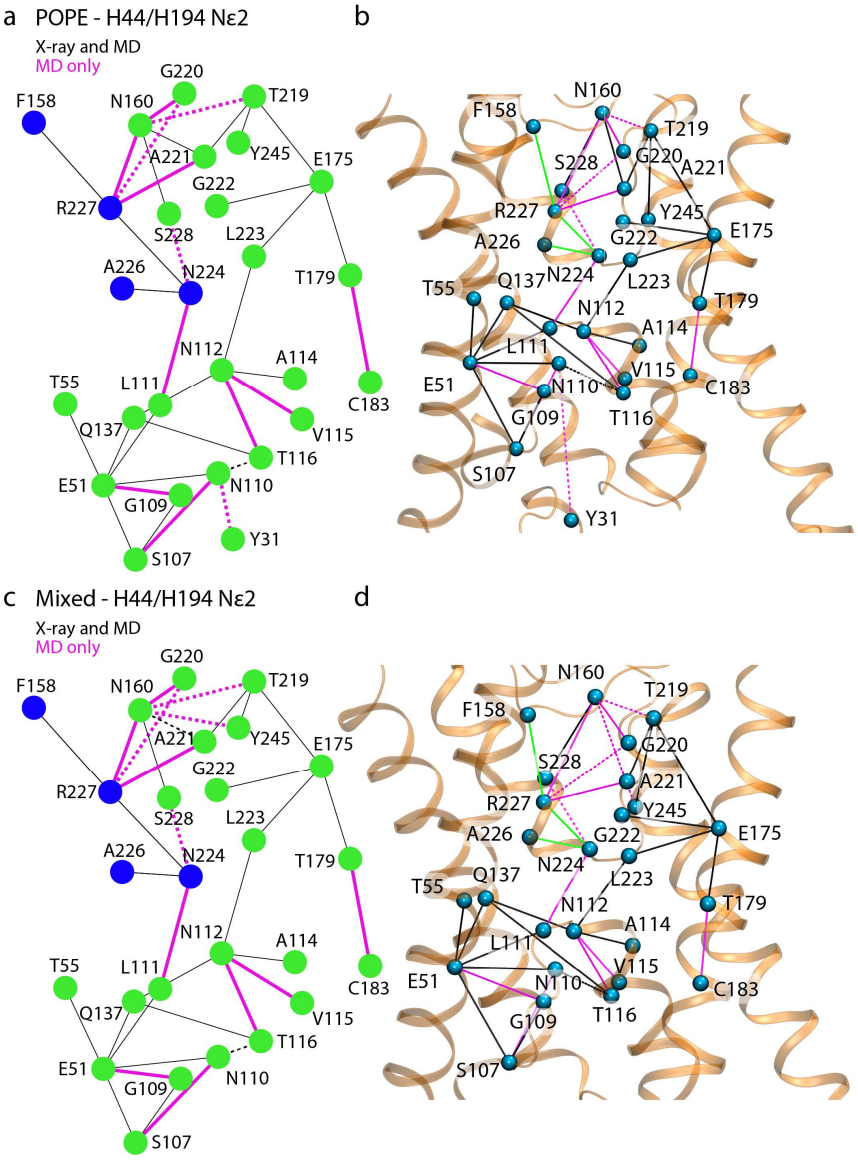

Figure 5.15. H-bond clusters in Aqy1 simulations and crystal structure. The R227-N112 H-bond clusters in the simulations aq1_b, aq2_b and crystal structure are presented in graph representations for simulations of Aqy1 embedded in a POPE (a) and a mixed (c) bilayer. H-bonds found both in the X-Ray and MD simulation are shown with black lines and H-bonds found only in the MD simulation with magenta. Solid lines represent H-bonds found in all four monomers of aquaporin while dotted lines represent H-bonds found in 1 to 3 monomers. Blue nodes represent the R227 cluster in the selectivity filter found in the crystal structure. (b, d) Molecular graphics of the clusters shown in panels a and c respectively. The original R227 cluster is shown with green lines. In panels (b, d) amino acid residues that are H-bonded are connected through their Cα atoms for clarity. The networks are represented on the crystal structure of Aqy1 (PDB ID: 3ZOJ). Molecular graphics with the H-bonded amino acid residues connected through the functional groups are shown in Figure A.22. Additional analysis for simulations aq1_a, aq2_a is shown in Figure A.23. Adapted from ref. [226].

Histidine protonation impacts the internal H-bond network of Aqy1

The simulated systems of Aqy1 are performed in all-Nδ protonated histidines and with H44-H194 Nε / H212-H242 Nδ, as shown in the crystal structure. Path length analyses of the interhelical motifs shows that the all-Nδ protonated systems sampled 7 to 15 more pathways of maximum length as compared to the H44-H194 Nε / H212-H242 Nδ systems (Figure 5.16a,b), while for the Serine/Threonine-Asparagine motifs, a conclusion is still not very clear (Figure 5.16c-f). The extended selectivity filter-NPA network also does not show major differences when sampled in the four monomers of each simulation for a different protonation state, but the histidines are not a part of it. When examining the networks of the affected histidines through the means of direct protein-protein H-bonds, it is found that in most of the cases there is a pattern in which the histidines interact with their surroundings. HD44 prefers to H-bond to S40 and with a linear pathway it can reach R129 via N43 and/or D39. Rarely it can H-bond to R129 directly, but S40 is a much more stable partner. HD44 can H-bond with D39 but the event is sampled very rarely (Figure 5.17a-d). HE44 prefers to H-bond directly to D39 or R129 forming very stable H-bonds. HD194 prefers to strongly H-bond with T197, E192, while HE194 prefers T197, A196, M188, L189 as H-bond partners (Figure 5.17e-h). There are instances of connections such as HD194-A196, but they are rarely sampled, showing that the protonation state of the histidines has an immediate impact of the local H-bond dynamics.

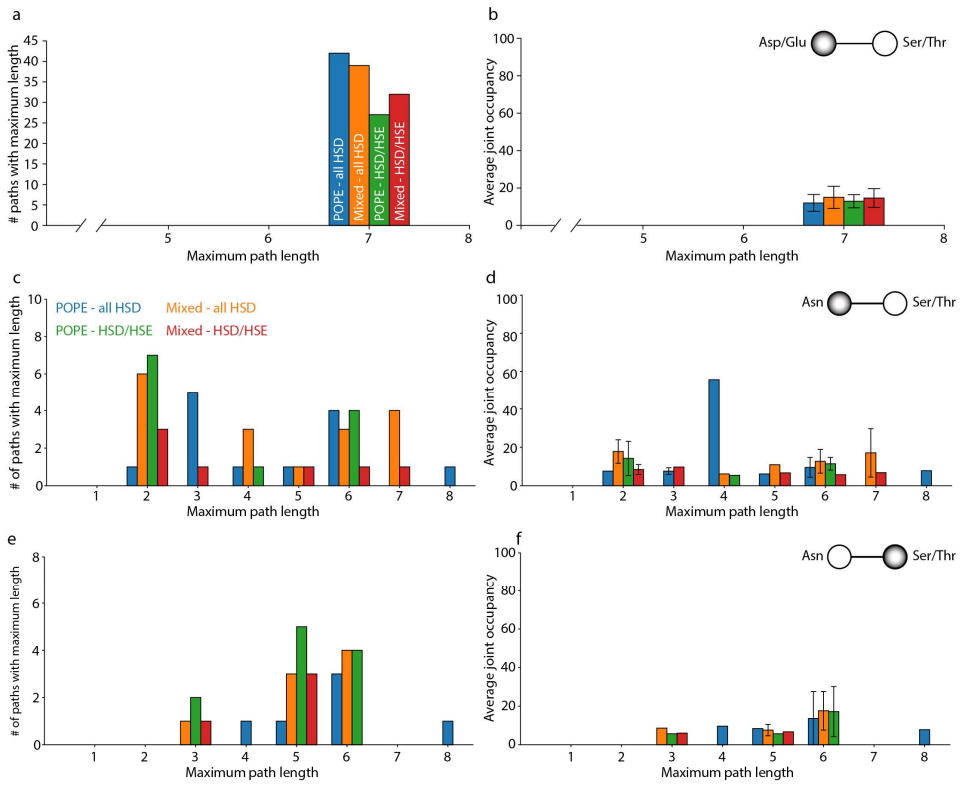

Figure 5.16. Shortest path computation for H-bond motifs in MD trajectories. (a) Path length distributions are shown for the carboxylate groups in the interhelical Asp/Glu-Ser/Thr in the

trajectories of four Aqy1 simulations (Table 5.5). (c) Path length distributions are shown for the asparagine groups in Asn-Ser/Thr motifs for the Aqy1 simulations. (e) Path length distributions are shown for the serine/threonine groups in Asn-Ser/Thr motifs for the Aqy1 simulations. (b, d, f) Average joint occupancy for the paths with the maximum length.

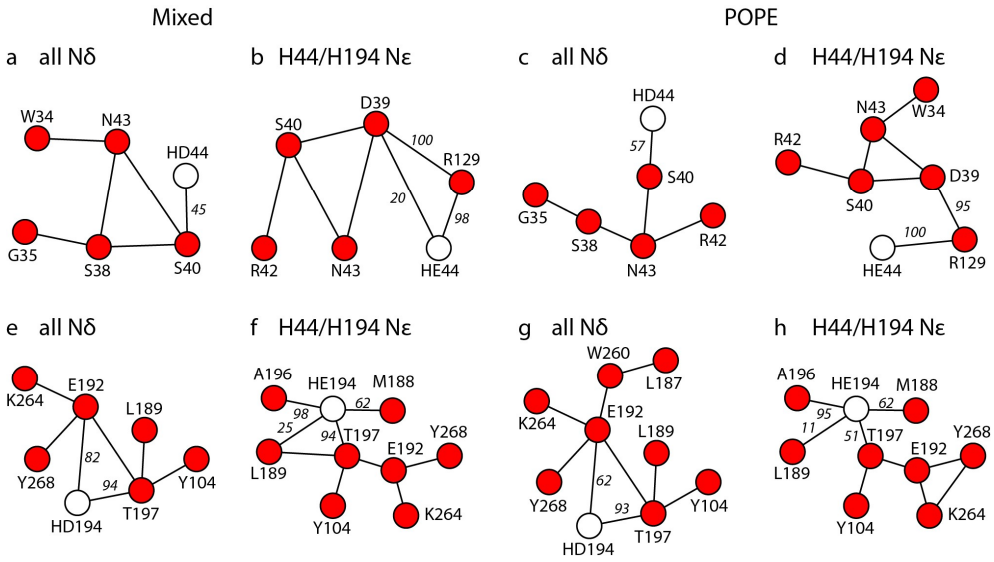

Figure 5.17. Illustration of H-bond clusters of H44 and H194 sampled in four different MD simulations of Aqy1. The nodes representing H44 (a-d) and H194 (e-h) are shown as an empty circle. (a-e) H-bond clusters in Monomer-B from simulations in a mixed membrane with all four histidines of Aqy1 in the Nδ1 protonated (a, e) vs. with H44 and H194 Nε2 protonated (b, f). H-bond clusters in Monomer-B from simulations in a POPE membrane with all four histidines of Aqy1 in the Nδ1 protonated (c, g) vs. with H44 and H194 Nε2 protonated (d, h). Adapted from ref. [226].

5.6 H-bond networks of Channelrhodopsin-2 (ChR2)

The protein-protein H-bond network of ChR2 rearranges upon isomerization of the retinal Schiff base

During MD simulations of the all-*trans* state, the Cα - RSMD values of ChR2 are computed 0.96 ± 0.12 Å for Monomer-A and 0.92 ± 0.08 Å for Monomer-B. Similarly, in the 13-*cis*,15-*anti* conformer the RMSD values are 1.15 ± 0.09 Å for Monomer-A and 1.09 ± 0.10 Å for Monomer-B (Figure A.25). Simulations of ChR2 dimer shown that in the all-*trans* state the RSB H-bond networks differs between the two monomers. Similar behavior has been detected for the chimeric channelrhodopsin C1C2 [42]. In Monomer-A the RSB network is very stable, whilst it does not extend in length (Figure 5.18a). It features the counterion E123 which H-bonds through its side-chain carboxylate to K93, forming a highly stable (99.9%) salt-bridge, and to T127's hydroxyl group, for an equally stable H-bond (Figure 5.18a). This H-bond is very interesting because it is a hybrid motif between the inter-helical Aspartate/Glutamate carboxylate - Serine/Threonine hydroxyl group and the intra-helical Serine/Threonine hydroxyl group to the *i*-3,4,5 backbone

carbonyl (Figure A.26). E123-T127 are H-bonded through their sidechains, carboxylate-hydroxyl but the two amino acid residues are 4 positions in the sequence apart. The orientation and stability of the H-bond could suggest that contributes to the stability of the RSB H-bond network in the dark state. The H-bond remains highly stable during the 13-*cis* simulations with an occurrence rate of 96.2% and 95.6% for monomers A and B respectively (Figure 5.18c). K93 participates in another stable salt-bridge with D253 (Figure 5.18a), and D253 participates with its backbone in a bifurcated intra-helical Serine/Threonine hydroxyl group to the *i*-3/4 backbone carbonyl, with S256 which is found one place before the RSB in sequence. S256-I252 (Figure A.27) is a very stable H-bond (96.2%) while S256-D253 rarely sampled at 16.8% occurrence rate. The motif is sampled evenly and is not directly linked to specific part of the trajectory.

In Monomer-B the network contains the same sidechain-sidechain interactions compared to Monomer-A, but it is further extended, closer to the edge of the membrane. The RSB is still connected to E123 but is also now rarely connected to D253 with occurrence rates of 32.3% and 11.2% respectively (Figure 5.18c,d, Figure A.27). E123 in turn, as detected in the crystal structure, participates in an intrahelical Serine/Threonine hydroxyl to the *i*-4 backbone carbonyl with T127 (Figure 5.18b,c, Figure A.27). The RSB in Monomer-A network extends only until K93, whereas in Monomer-B includes H249 and a tight salt bridge between E97-R120. K93 has one more salt-bridge partner, E90, but the connection is sampled only 19.7% of the time (Figure 5.18b,e). The intrahelical H-bond motifs between I252/D253 and S256 are not sampled in Monomer-B (Figure 5.18a,b,c). A very distinctive difference in connectivity between Monomer-A and Monomer-B regards E123 and T127. In Monomer-B they connect via an intra-helical Ser/Thr hydroxyl to the *i*-4 backbone carbonyl, sampled 53.5% of the time (Figure 5.18b,e). E123 has been shown to maintain its functionality upon mutations [218, 223]. T127 is found to be H-bonded to E123 in different ways in monomers A and B, essentially adapting to the alternations of the RSB network. It can be characterized as hybrid motif between the interhelical Aspartate/Glutamate carboxylate - Serine/Threonine hydroxyl group and the intrahelical Serine/Threonine hydroxyl group to the *i*-3,4,5 backbone carbonyl. E123-T127 are H-bonded through their sidechains, but the two amino acid residues are 4 positions in the sequence apart. It could be suggested that the hybrid/inverted motif in Monomer-A and the intrahelical motif in Monomer-B provides flexibility [35] and is coupled to the RSB network. The H-bond dynamics that are not highly stable in the RSB network of Monomer-B are summarized in Figure 5.18e.

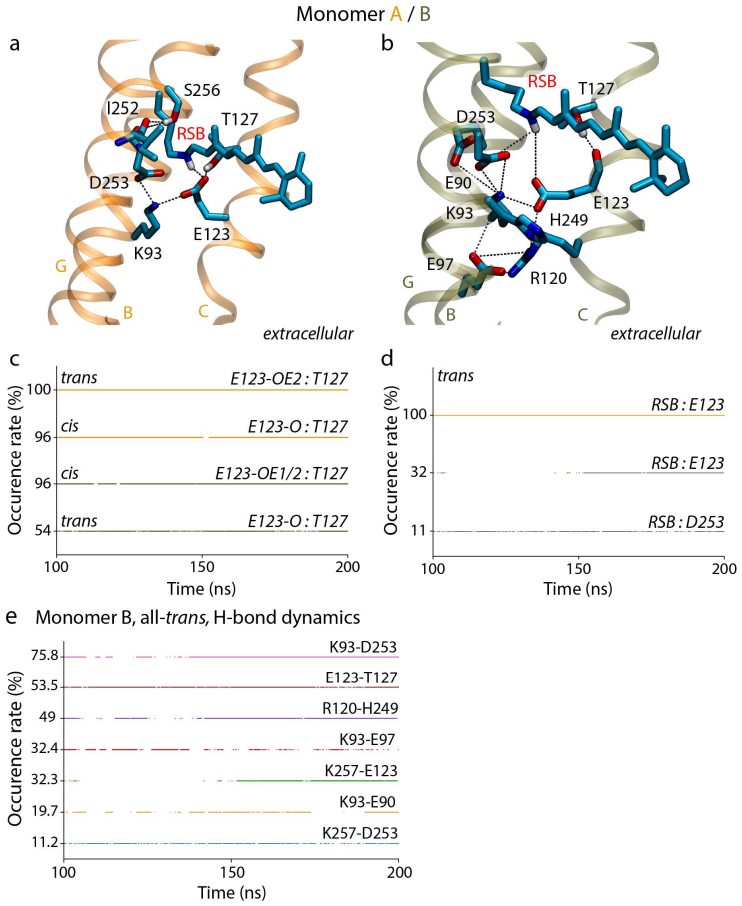

Figure 5.18. H-bond networks of the RSB in MD simulations of ChR2. (a, b) The RSB networks in the all-*trans* isomeric state for Monomer-A (a) and B (b). (c, d) Occupancy timeseries for the E123-T127 connections (c) and the RSB-E123 connections (d) Two connections of E123-T127 are shown for the 13-*cis*,15-*anti* isomeric state. (e) Occurrence rate timeseries of H-bonds sampled in Monomer-B of the all-*trans* simulation of ChR2. The H-bonds correspond to panel (b). Monomers A and B are colored orange and olive respectively. Helices are labeled A-F in the same color of the protein. Helices that do not participate in the networks are omitted for clarity. Adapted from ref. [226].

The H-bond network is radically altered upon isomerization of the RSB to 13-*cis*,15-*anti*. In Monomer-A direct or 1-water mediated H-bonds between the RSB and another H-bond partner are not detected. In Monomer-B, the RSB is found H-bonded to S256 at 74.8% of the trajectory analyzed (Figure 5.21b). S256 is in sequence H-bonded to I252 99.8% of the time forming a S/T-CO motif (Figure 5.21b). A Serine/Threonine-Asparagine motif involving the conserved N258 is consistently sampled across both isomeric states of the retinal and monomers. N258 forms a small interaction network with S63 forming the Serine/Threonine-Asparagine motif, and in turn S63 H-bonds to A59 through an intra-helical Serine/Threonine hydroxyl group to the *i*-4 backbone carbonyl, in both isomeric states and monomers. The two motifs are jointly sampled at higher rates when the retinal is in the all-*trans* isomeric state (85-93%) compared to the 13-*cis* (29-36%), for Monomer-A and B respectively. The effect of the isomerization in the H-bond

dynamics is also reflected in the Arginine-backbone carbonyl, Aspartate/Glutamate-Arginine, Asparagine-backbone carbonyl and Serine/Threonine-Asparagine motifs that originate in the crystal structure and were tracked down in the simulations (Table A.10). The occurrence rates show large differences and specifically for Arginine-backbone carbonyl motifs, it is shown that specific motifs are not detected in Monomer-A or B of the 13-*cis*,15-*anti* simulation, while they are sampled across both monomers in the all-*trans*.

Extracellular H-bond cluster and H-bond motifs

In the EC side of the membrane, two largely extended H-bond networks are detected (Figure A.28). The RSB network is tightly packed around the retinal and it extends to the EC side through two salt bridges. The primary proton acceptor D253 connects to K93 and in turn K93 connects to E97.The network reaches Q56 of helix A (Figure A.28a,c). H-bond motifs dictate the RSB network. Helices D and E feature most of the motifs, which would indicate their role in helical flexibility, as they are located along the helices [35]. In the close vicinity of the retinal, where an intrahelical Serine/Threonine hydroxyl to the *i*-4 backbone carbonyl between E123 and T127 (Figure A.28c). The primary proton acceptor D253 and I252 participate in the same motifs with their backbone to the hydroxyl group of S256. The extended EC cluster of ChR2 features the only combined intrahelical and interhelical motif detected in the crystal structure, between the carboxylate of E101, hydroxyl of T246 and the backbone carbonyl of V242 (Figure A.28b,c).

Water-mediated H-bonds of ChR2

Inspired by the analyses performed and presented in the chapter *"Channelrhodopsin C1C2"* for C1C2 [42], and the results of the direct H-bonds, presented above, the dynamics of the water-mediated H-bond networks of ChR2 were investigated. Unlike C1C2, the N258-E90 interaction of the central gate is sampled in the range of 12.9-17.4%, and in the case of Monomer-A of the all-*trans* model is not sampled at all, resembling the interactions of the K-state [294]. The infrequently sampled interaction of N258 with E90 are water mediated and not direct as it would be in the dark state D^{470} [294]. Supporting the notion of the interactions sampled in the 13-*cis* state to resemble the K-state, is the interaction of E90 to E123, which is sampled 100% in Monomer-A and 74.3% in Monomer-B. In both cases the H-bonds are direct. Similar to C1C2, I investigated connections of the retinal Schiff base to the EC side. The proton release group of ChR2 is not yet known, thus it would be reasonable to suggest that carboxylate groups on the EC side would be a suitable candidate. On the end of helix B, lies E97 and on the end of helix F lies E235. Similar positions of amino acid residues with a carboxylic acid sidechain are found in C1C2 [42] and AntR [185].

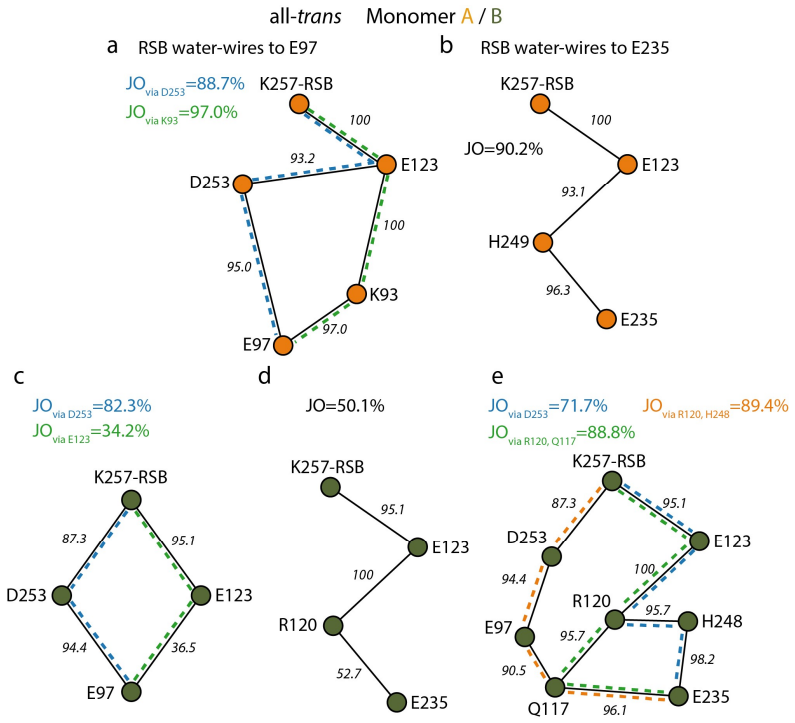

Figure 5.19. Water-mediated H-bonds connecting the Retinal Schiff base to the EC side of the membrane. (a) Shortest-paths connecting the RSB to E97 in Monomer-A. Pathways to E97 via D253 and K93 are colored in blue and green dotted lines respectively. Their JO values are annotated in the same colors. (b) Shortest-paths connecting the RSB to E235 in Monomer-A. The shortest paths computation shows a consecutive series of H-bonds from K257 to E235. (c) Shortest-paths connecting the RSB to E97 in Monomer-B. Pathways to E97 via D253 and E123 are colored in blue and green dotted lines respectively. Their JO values are annotated in the same colors. (d) Shortest-paths connecting the RSB to E235 in Monomer-B. The shortest paths computation shows a consecutive series of H-bonds from K257 to E235, via the counter-ion E123 and R120. (e) Alternative high-occurrence pathways between K257 and E235. The occupancy (%) of each H-bond is annotated along every edge.

The water wires connecting the RSB with E97 and E235 resemble the paths found in C1C2 [42] (see subchapter *"H-bond network analyses of C1C2"*). In the all-*trans* model of ChR2, the RSB connects to E97 in both monomers with high-occurrence pathways. In Monomer-A, the pathways branch in the counterion E123 which is the first intermediate node (Figure 5.19a). Then the pathways split to D253, or K93 as the next intermediate node and they both terminate to E97. Both pathways are highly sampled with joint occupancies computed at 88.7% and 97% respectively. On the other side of the pore, E235 can be reached from the Schiff base through the counterion E123 and H249. The single high occurrence shortest path is sampled 90.2% of the time (Figure 5.19b). In Monomer-B, the pathways are similar to the ones sampled in Monomer-A. E97 can be reached from the Schiff base through either one of the counter ions. Th pathway through E123 is sampled significantly less due to the E123-E97 connection, which appears as the bottleneck. The joint occupancies are computed at 82.3% and 34.2% respectively (Figure 5.19c). The pathway bridging the Schiff base and E235 passes through E123 and R120,

as compared to E123 and H249 in Monomer-A (Figure 5.19d). The joint occurrence is computed at 50.1%. If the water wires are pre-filtered to high occurrence rates, an alternative network of water-mediated H-bonds is computed instead (Figure 5.19e). This network features both counterions, E97, R120, H248 and Q117 as intermediate nodes. The three different sub-paths that emerge from the network are sampled 71.7%, 88.8% and 89.4% respectively.

In the simulations of ChR2 with a 13-*cis*,15-*anti* isomeric state, the RSB of Monomer-A participates only in water-mediated H-bonds with the EC side (Figure 5.21a). Monomer-B connects only to S256 via direct H-bonding(Figure 5.21b). The CP oriented Schiff base is connecting to the carboxylates of the EC side through water-mediated connections to H134 or Y70 and N258 of the central gate. Joint occurrences for the pathways are ~10%, which is to be expected from the unfavorable orientation of the amino acid residues (Figure 5.20).

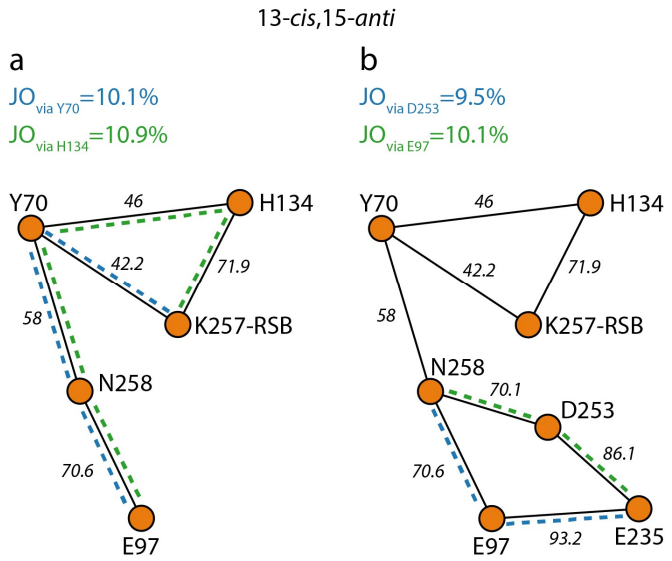

Figure 5.20. Water-mediated H-bonds connecting the Retinal Schiff base to the EC side of the membrane in the 13-*cis*,15-*anti* model of ChR2. (a) Shortest-paths connecting the RSB to E97 in Monomer-A. (b) Shortest-paths connecting the RSB to E235 in Monomer-A. Pathways connecting the Retinal Schiff base to the EC side are not sampled for Monomer-B. The occupancy (%) of each H-bond is annotated along every edge.

Similar to AntR [185], two long range water-mediated connections between the RSB and H134 and E82 were detected. Those connections are not sampled at all in Monomer-B, since it still retains connectivity to S256 (Figure 5.21a,b). The RSB is found connected to H134 and E83 via long water-mediated wires with average number of water molecules in the wires being 3.6-3.9 respectively (Figure 5.21c,d). Both connections are highly stable (81-90%) and can contain 1 to 8 water molecules (Figure 5.21c,d). Such pathways could serve a reprotonation pathways for the RSB. Although D156 is shown to be primary proton donor [80], it is also suggested that H134 should not be excluded as the proton donor [83]. H134 corresponds to the internal proton donor D96 of BR.

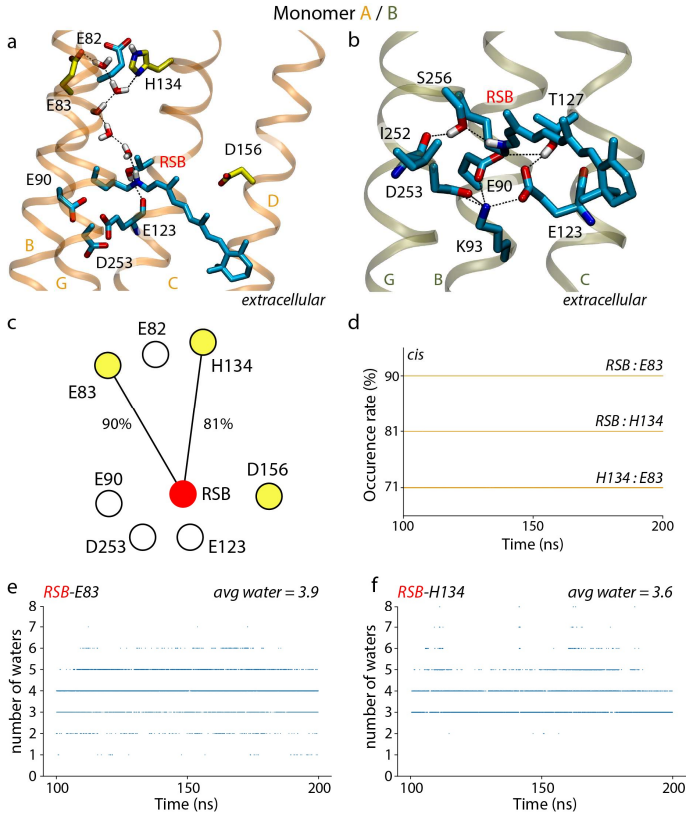

Figure 5.21. H-bond networks of the 13-cis,15-*anti* RSB in MD simulations of ChR2. (a, b) The RSB connections in the 13-*cis*,15-*anti* isomeric state for Monomer-A (a) and B (b). The RSB-H134, RSB-E83 water-mediated connections are not sampled in Monomer-B. (c) Schematic representation of the two water-mediated connections of the RSB with E83 and H134. The primary proton donor D156 [80] and E82 of the cytoplasmic side are shown as well as the E90, E123 and D253 of the RSB network. (d) Occupancy timeseries for the RSB-E83, RSB-H134 and H134-E83 connections. (e, f) Water wire length timeseries for the RSB-E83 connections (e) and the RSB-H134 connections (f). The average water wire length is annotated in both graphs. Monomers A and B are colored orange and olive respectively. Helices are labeled A-F in the same color of the protein. Helices that do not participate in the networks are omitted for clarity. Adapted from ref. [226].

The presence of the extended connections is tightly connected to the conformation of H134 observed during simulations. I sampled two main conformations, *flat* and *up* (Figure 5.22). When in the *flat* conformation (Figure 5.22a,b) H134 is directly bound to E83, and it appears that it functions as blocker for the ion flow passing through the pore (Figure 5.22a,b). Analysis of the C_α-C_β-C_γ-C_δ dihedral angle of H134 shows that the *flat* conformations is observed for values ~100-120° and -60 to -120° (Figure 5.22e). The up conformation is sampled more regularly when the dihedral is between 0-60° (Figure 5.22e). When H134 is in the up conformation it loses the direct H-bond to H134, but it connects to the RSB via the extended water-wire and to E83, but through the presence of water molecules. The *up* conformation appears to provide an easier passage for ion flow (Figure 5.22c,d).

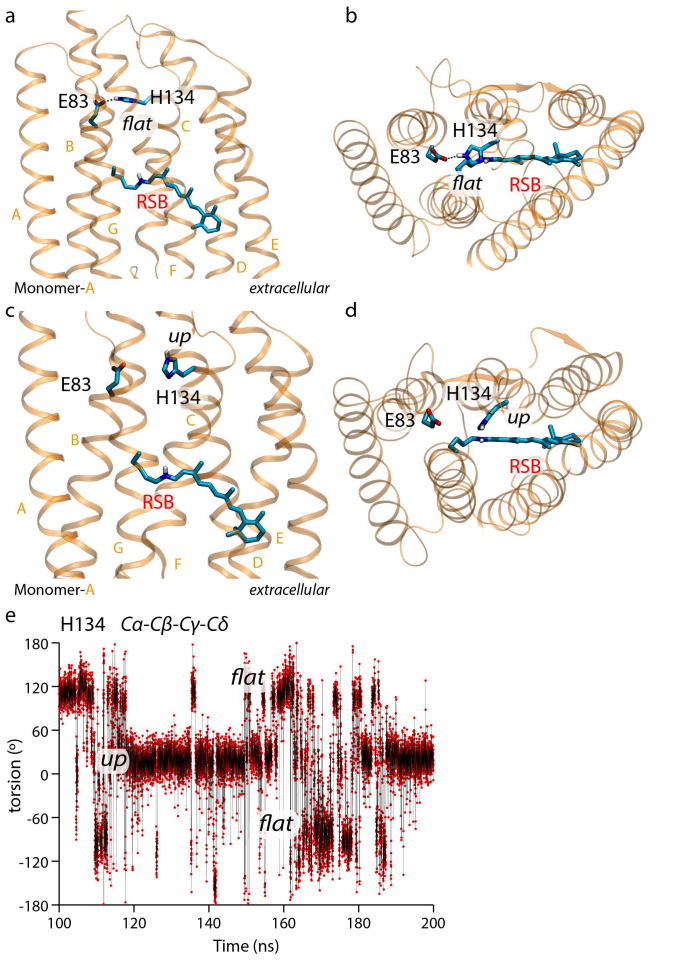

Figure 5.22. Conformations of H134 in Monomer-A of the 13-*cis*,15-*anti* simulation of ChR2. The RSB-H134 water-mediated connection is not sampled in Monomer-B. (a, b) E83 connecting to H134 while in the *flat* conformation in side (a) and top (b) view respectively. (c, d) E83 and H134 are not connected while in the *up* conformation in side (c) and top (d) view respectively. Dihedral angle timeseries for the Cα-Cβ-Cγ-Cδ dihedral angle of H134. *Up* and *flat* conformations are noted in the regions of dihedral angles that they appear on. Adapted from ref. [226].

Network rearrangement and centrality

In the case of water mediated H-bonds in the interior of ChR2, highly complex networks are detected. Allowing water in the computations increased the complexity significantly as compared to direct H-bonds. In the previous subchapter I illustrated that H-bond networks differ depending to the isomeric state of the retinal, as it can affect local dynamics. Using comparative graphs between the two isomeric states of the retinal, the differences in the H-bond networks can be seen. In Figure 5.23 the comparative graphs using a 50% occupancy threshold on the individual connections are shown. Upon retinal isomerization the networks drastically rearrange, as new connections are sampled, and other connections are no longer sampled. The effect of the retinal isomerization is

especially evident in the EC side of the membrane, while fewer connections are affected in the CP side (Figure 5.23), with most significant the long water-mediated connections I detected, connecting the 13-cis RSB to H134 and E83.

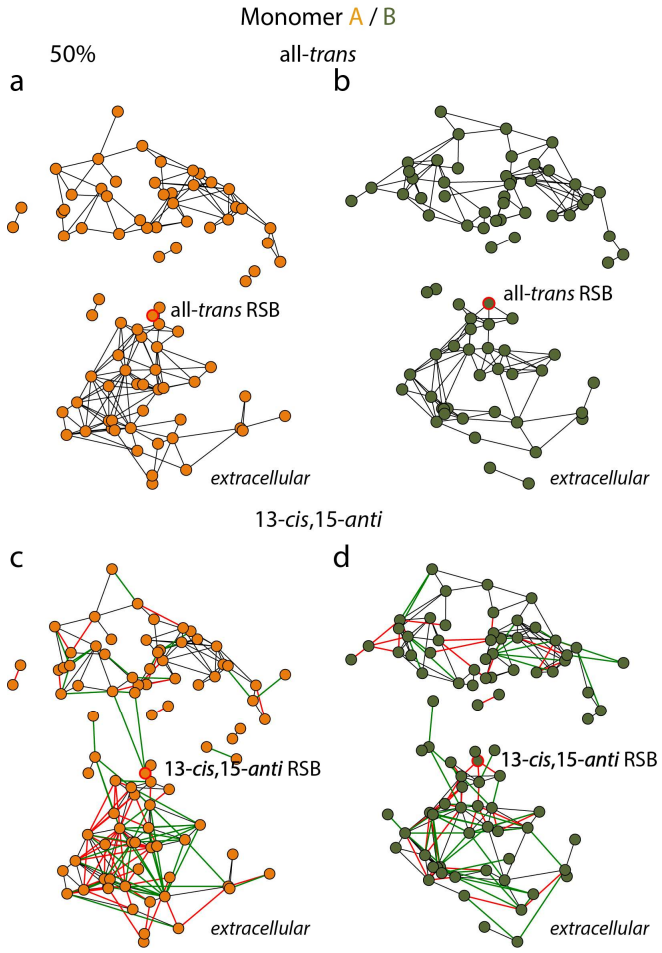

Figure 5.23. Comparative H-bond graphs between the all-*trans* and 13-*cis* models using a 50% occupancy threshold. (a, b) Reference graphs for the water wire networks of ChR2, sampled in Monomer-A (a) and B (b), in the all-*trans* model respectively. (c, d) Difference graphs between the all-*trans* and 13-*cis* models, in Monomer-A (c) and B (d). Additional graphs with 10% and 80% occupancy thresholds are show in the Apendix (Figure A.29 Figure A.30).

In the AntR work (chapter *"Antarctic Rhodopsin"*), I used centrality measures (BC and DC) to predict the importance of S74 in the communication of the network. Such finding was confirmed experimentally [185]. We developed the Unique Shortest Paths (USP) [186] in order to tackle the expected redundancy that BC could lead to, depending on the graph connectivity. In this chapter, the USP measure was used to compute amino acids residues that could be of great importance for the network. A BC computation for the same graph is found in the Appendix. Using graphs of H-bond networks with relatively high individual occupancies (50%) USP values were computed for both monomers and both isomeric states of the retinal. The nodes (amino acid residues) that likely be the most important for the communication of the network are E97,

R43 and E123 in the all-*trans* model of ChR2 (Figure 5.24a, b). E97 was investigated in this subchapter as a suggested proton release group for ChR2, retaining high-occurrence connections with the all-*trans* retinal. Upon isomerization of the retinal, the networks largely rearrange (Figure 5.24c, d). Especially in the EC side of Monomer-A, E97 appears to have lost almost all of its connections, while E235, which is the other carboxylate group that was investigated a possible proton release group, features many new connections, compared to the all-*trans* model (Figure 5.24c, a). Another group on the EC side, R115 has many more connections after the retinal isomerization and a USP very close to the one of E97 (0.21 vs. 0.23). Despite E97 having lost many of its connections, its position in the graph, and therefore the protein is still of great importance, since it retains the second highest USP value, after the isomerization, after D253. The USP values on the CP side do not show significant difference, even after the establishment of the RSB-H134 and RSB-E83 connections. In Monomer-B the amino acid residues with the highest USP values are found in the EC side, with E97 having a reduced USP value compared to the all-*trans* model. On the contrary, D253 has an increased USP upon isomerization of the retinal (Figure 5.24c, d).

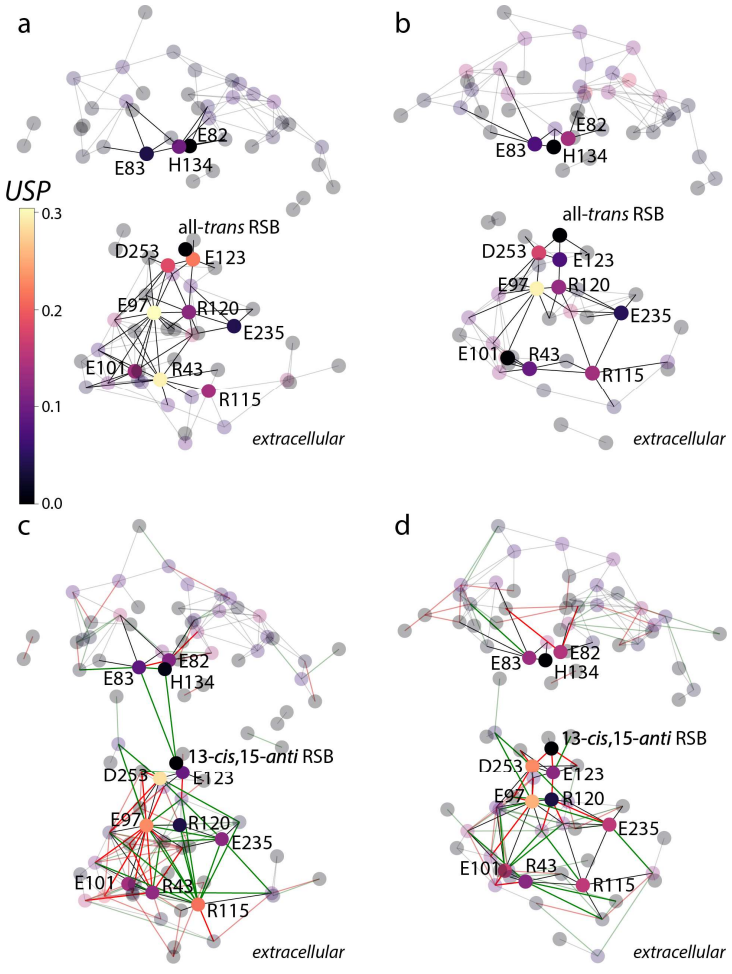

Figure 5.24. Unique Shortest Paths computations for the all-*trans* and 13-*cis*,15-*anti* models of ChR2. Reference graphs computed for the all-*trans* model for Monomer-A (a) and B (b),

respectively. Difference graphs between all-*trans* and 13-*cis* models for Monomer-A (c) and B (d). Graphs were prefiltered to 50% occupancy rates, before the USP computations.

The DC gate

An interaction between C128 of helix C and D156 of helix D is referred to as the DC gate [85, 217]. In the crystal structure of the chimeric C1C2 the distances of between the thiol group and the carboxyl oxygens were measure at 4.4 Å and 4.6 Å [99]. In the wild-type ChR2 structure, the DC gate was detected as a water-mediated H-bond through water *w5* [97]. In the C1C2 trajectories, the DC gate was computed as a direct H-bond for 20% of the time in Monomer-B. For the wild-type ChR2 the gate was sampled at relatively higher occurrence rates, especially in the all-*trans* model (Figure 5.25). In the all-*trans* model the DC gate is sampled for 38.6% and 32.6% of the time, for Monomer-A and B respectively. In the isomerized retinal, the DC gate is sampled half of the time, as compared to the all-*trans*. In Monomer-A, D156 participates in a local interaction network with C128, W223 and T159. The H-bond D156-T159 is sampled for 76.8% of the time, as compared to 18.1% for the DC gate. A similar 17.1% occurrence rate is sampled for Monomer-B. Most of the time that the DC gate is sampled, I find that the connection is mostly water mediated (Figure A.32). In Monomer-A of the 13-cis model, there are instances where the DC gate is a direct H-bond (Figure 5.25c). A water molecule has been suggested to mediate the C128-D156 connection in the literature [97, 103, 230, 295, 296], in line with the above findings.

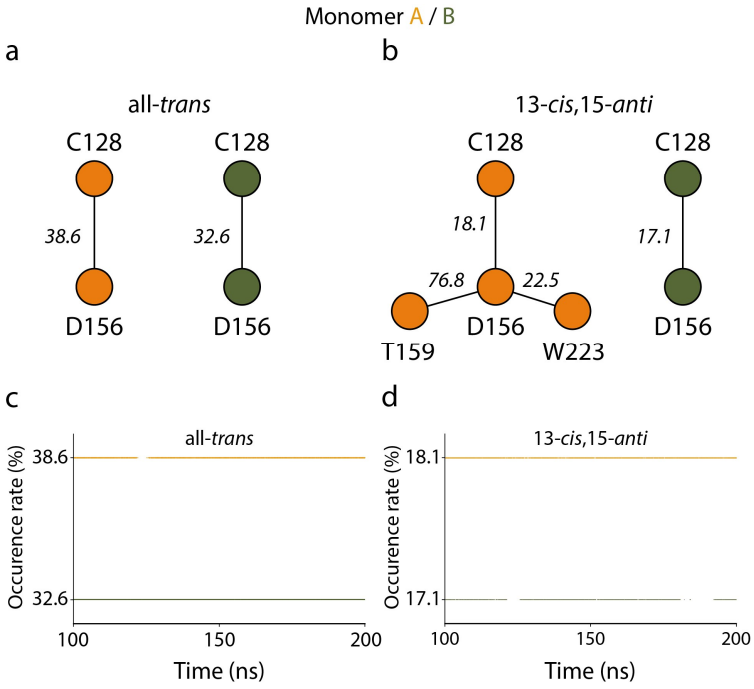

Figure 5.25. The DC gate is sampled in the ChR2 trajectories. (a, b) The DC gate interaction sampled in the all-*trans* (a) and 13-*cis*,15-*anti* (b) models, respectively. (c, d) Occurrence rate

timeseries for the DC gate interaction sampled in the all-*trans* (a) and 13-*cis*,15-*anti* (b) models, respectively. Monomers-A and B are colored orange and green, respectively.

5.7 Summary

In this chapter, large-scale analyses on multiple datasets of hundreds of TM protein in search of conserved H-bond motifs important for function were presented. The static structure analyses were accompanied by MD simulations of an ion channel and a water channel. Two main datasets were initially presented, and more derivatives were generated in the course of this project. The datasets contain α-helical TM proteins belonging to 26 superfamilies. With the refinement procedure I developed and further hand curation to tackle possible redundancy, I presented a high-resolution dataset, a low-resolution dataset and a high-resolution, with unique structures, dataset. I used the Transporter Classification Database (TCDB) [259] to further categorize superfamilies to *protein-groups* to further, according to their biological role. Since the Rhodopsin-like receptors superfamily is very diverse, it was treated as a separate *protein-group*, and was further split into four subsets. Every *protein-group* of *Set-high* and *Set-highU* contains at least one carboxylate-hydroxyl motif, while the intrahelical carboxylate-backbone carbonyl of the *i*-3,4,5 relative position is found in every protein of every protein-group, and it is the most widely represented motif in this search. The combined carboxylate-Ser/Thr & Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position is a more selective motif and is not sampled as frequently compared to its individual components. The combined motif is found more frequently in the Rhodopsin-like receptors and pumps protein-group, which features many Aspartate/Glutamate and Serine/Threonine motifs in the center of the membrane core as well as towards the cytoplasmic and extracellular sides of the membrane, where prime examples of amino acid residues of critical importance for the mechanism of the protein is found, such as the proton release groups of BR [220, 227-229]. Examples of motifs that are known or could possibly be important in the function were found in Archeaorhodopsin-2, Bacteriorhodopsin, Archaerhodopsin-3, Halorhodopsin, Coccomyxarhodopsin, *E. sibiricum* rhodopsin, Scetabularia rhodopsin I, KR2, Jumping spider rhodopsin-1, among others. His-His motifs belong to proteins of the ammonium channel transporter superfamily where the highly conserved histidines are an essential part of the transport mechanism in ammonia channels at the center of the transmembrane domain of the protein. The highly conserved His-His motif acts as a proton relay in the transport of ammonium. Other proteins with His-His motifs are Anabaena sensory rhodopsin and the Calcitonin type 1 receptor of the GPCR Secretin B family with a continuous network of three histidines, similar to the proton-transferring motif in the ammonia channel AmtB. Another motif that is very likely to be involved with protein function is Asp-Asn H-bonds and it is found in the proton translocating pyrophosphatase, where N738 could stabilize water molecules of the proton transfer pathway. The majority of the Asp-Asn motifs are found within the Rhodopsin-like receptors and pumps with 25 proteins containing at least one motif, and its many of its members host proton transfer events.

Aquaporin-1 (Aqy1) is a host for a plethora of structural motifs, such as the ar/R and the dual NPA motif. The crystal structure of the Aqy1 hosts two large H-bond clusters, with 10 and, 19 groups respectively, with former spanning the length of the protein, connecting the extracellular and cytoplasmic regions. In the crystal structure the networks of the selectivity filter and the NPA motif are separated when water molecules are not included in the network computations. In all 4 MD simulations I performed they are connected via the new connections that are established between R227 and N224.

In simulations of the ChR2 dimer, H-bond networks differ between the two monomers, similar to the results of C1C2 [42]. An unprotonated E123 can salt-bridge to K93 and to T127, similar to C1C2 [42]. The second counterion can connect to K93 to form an additional salt-bridge, and S256 to form an H-bond motif. Upon isomerization of the RSB to 13-*cis*,15-*anti*, the H-bond network is heavily rearranging. Similar to AntR [185], I detected two long range water-mediated connections between the RSB and H134 and E82 in monomer-A (81% - L=3.6 and 90% - L=3.9, respectively), that could serve as reprotonation pathways for the RSB, since it is suggested that H134 should not be excluded as the proton donor [83]. The presence of the wire is tightly connected/coupled to the conformation of H134 (*flat* or *up*), and the H-bonding between H134 and E83. ChR2 maintains very frequent water mediated connections to amino acid residues with a carboxylic sidechain located in the EC side of the membrane. Similar to C1C2 and AntR, E97 and E235 were investigated as suggested proton release groups (proton uptake for AntR). USP measures indicate that E97 and E235 are important for the communication of the graphs and removal of those nodes could lead to major disruptions in the networks. Following the model of the rearrangement of D253 upon retinal isomerization, in the K state [294], D253 has and increased USP value in the 13-*cis*,15-*anti* model, as compared to the all-*trans*. The DC gate is sampled in the wild-type ChR2 at higher occurrence rates, as compared to C1C2 [42] and is a water mediated connection most of the time.
A noteworthy aspect of this project was to showcase capabilities of the algorithm package *Bridge*, we made available in 2019. Using the core package, it is possible to implement new standalone code/functions to expand analyses features according to project at hand. Here, I was able to build and curate large datasets of static structures using the API of OPM to collect the structures, with the help of Malte Siemers (see credits in the beginning of the chapter). Additionally, using output from existing *Bridge* functions, I was able to further curate the data and introduce the path-length computations, presented in this chapter. Another example of Bridge adaptation is discussed in the chapter *"Aspects"*.

Chapter 6 Conclusions

Hydrogen-bond networks are complex entities that can shape local and conformational dynamics in proteins. Ion channels or proton pumps such as, channelrhodopsin or bacteriorhodopsin respectively, can host numerous water molecules in their interiors, as shown in multiple crystallographic structures [19, 97, 99, 132, 253, 297-300]. Simulations of said crystal structures in model membranes show that the water molecules detected by X-ray crystallography can form conserved water sites [261]. Water molecules in the interior of the protein, stably bound [160] or dynamically exchanging, in some cases in the microsecond scale [301] , with the bulk increase the connectivity of the amino acid residues and the complexity of the connections [42, 103, 153, 154, 216, 219, 230, 294, 302]. Water-mediated H-bonds can bridge remote areas of the protein and serve as water wires for the transfer of protons. Proton transferring pathways can be suggested from inspection of the crystal structures after refinement of the water positions. A limitation of this approach this that we rely on a single snapshot. MD simulations of TM proteins in hydrated lipid environments can provide insight to the dynamic behavior of the underlying H-bonds. Available analyses algorithms lacked the capabilities of large-scale analyses for long trajectories of MD simulations and more importantly lacked additional features for one to efficiently analyze and represent the raw data.

In the first part of this dissertation, a new-generation algorithm I contributed to the development of, *Bridge*, is described. This new algorithm package is very capable of efficient analyses of complex H-bonds networks in proteinic structures for crystal structures and MD simulations trajectories. The flexibility of the algorithm allows for lipid molecules, DNA or anything other entity that can host H-bonds to be analyzed. *Bridge* employs graph-based approaches to compute [33] and analyze [42, 185, 226] H-bond networks. Implementation of graph-based algorithms such as the "Connected Component analysis" and Dijkstra's algorithm allows for an intuitive approach to the analyses. Especially insightful are proved the graph curation functions and routines of Bridge in the case of Channelrhodopsin, since the starting point of the proton transferring events is known. The proton is transferred from the protonated retinal Schiff base to the negatively charged counter ion. The all-*trans* state of the C1C2 chimera was studied in the first part of this thesis with a strong focus on the extended hydrogen-bond networks and clusters that dynamically form in the interior of the protein. E162 is found to play a very significant role in linking the retinal to the other amino acids of the retinal vicinity, maintaining direct H-bond with the Schiff base almost 100% of time for both monomers. The local networks of the SB varies between the two monomers. In Monomer-A it is focused around the SB itself, extending until E129 on the EC side and to C167 on the CP side. The joint occurrence of the SB connecting to C167 is very, due to the limiting

occurrence of the T166-C167 H-bond sampled 2.7% of the time. In contrast, Monomer-B features an extended H-bond network that stretches to the EC side through only direct H-bonds. The end node in Monomer-B is E274, the equivalent of $E194^{BR}$ [99], which a part of the proton release group in BR [220, 227-229]. The direct pathway between the RSB and E274 is mostly dependent on the infrequent sampling of D292-H288 and H288-R159 (<10%). In both monomers K132 is a central node in the networks and retains stable connections to the two counterions. In both monomers an H-bond motif between the sidechains of T166 and the counterion E162 is observed. An equivalent motif is observed in ChR2 between T127 and E123 [230], but in the case of a protonated E123 the motif is not observed anymore [230]. A similar motif is found in the acidic dark state of BR where T89 participates in an intrahelical H-bond motif with the backbone carbonyl of D85 which is found in the *i*-4 relative position [35]. In Monomer-B a direct interaction between C167 and D195 is sampled 20% of the time, which referred to as the DC gate [85, 217], despite the large distances of 4.4 Å and 4.6 Å between the thiol group and the carboxyl oxygens measured from the crystal structure [99].

Water mediated H-bonds allow for both monomers to connect the RSB to E129, with a considerable difference in the occurrence rates and relative orientations of N297. In Monomer-A, the protonated E129 connects through very stable H-bonds to D292 (100%), while N297 can connect to both E129 and D292 through water-mediated H-bonds. In Monomer-B, E129 is oriented towards N297 forming very stable interactions throughout the duration of trajectory. Such interaction is described in the literature as an inner gate [101, 215] of the EHT [215] model, where a double H-bond between N297 and E129 ($N258^{ChR2}$ and $E90^{ChR2}$) holding helices B and G together and preventing the opening of the pore [303]. Aside the local interactions in the vicinity of the RSB, water-mediated networks extend to carboxylates of the EC side, namely E136 and E274 ($E194^{BR}$) with very high sampling rates for the complete pathways. One of the distinctive analysis capabilities of the algorithm package *Bridge* is, that a joint occurrence of a pathway can be computed, after overlapping the individual timeseries of the underlying H-bonds. If a pathway is sampled as a whole during the trajectory, the exact frames that the path is sampled at can be returned. Knowing when a pathway is connected would be especially insightful for recommending starting coordinates for proton-transfer pathways computations. The pathways detected are shortest paths using Dijkstra's definition [206] feature the two counterions E162, D292, K132, H288 and R159 in various combinations, as intermediate groups between the root and the end nodes. Simulations of mutations on key amino acid residues heavily affect the local dynamics of the C1C2 networks.

Three additional simulations of mutant variants of C1C2, namely E162T, R159A and H288A were prepared. Replacing the glutamate in the position 162 with a threonine forces the RSB to stabilize itself through a connection to the D292 and allows for a long (L ranging from 3 to 5 water molecules) water-mediated connection directly to E136, sampled 76% of the time. This is possible due to E136 orienting towards K132 and forming a stable salt-bridge, thus making the effective distance between the RSB and E136 shorter. The K132-E136 connection rarely features mediating water molecules and is mostly sampled as a salt-bridge. Replacement of R159 or H288 with an alanine sidechain results to an increased K132-E136 distance and long water wires bridging the

two groups. R159 and H288 become the main intermediate node between the Schiff base and the EC side in the absence of another. In the E162T simulation both H288 and R159 bridge the RSB to E274 creating a branched network.

The effect of mutations was clearly quantified in a rather surprising finding. With the graph-based queries and curation functions of networks I identified a very complex network of 46 amino acid residues bridging the two proton donors, the Schiff bases of the two monomers. It was the first time that such finding was reported in the literature, to the best of my knowledge. Initial dissection of the pathways from RSB1 to RSB2 in the wild-type C1C2 resulted in joint occurrence rates around ~10%. Despite the apparent relatively low rate, one would argue that this is a very reasonable result considering that 12 H-bonds have to be sampled simultaneously. The equivalent pathways for the mutant variants follow a completely different network of connections to bridge the two retinals, compared to wild type. Six pathways for E162T and R159A mutants were sampled and just two for the H288A variant. The joint occupancies for the two pathways in the H288A simulations were the highest sampled reaching ~40%, making them four times more likely to be sampled.

In the second part of this dissertation, the modelling and analysis of a newly discovered rhodopsin [185] was presented. AntR was characterized as a bistable inward proton pump with a still unclear physiological role. It features, unlike BR, a single counterion; D185, which is located on helix G. A homology model of AntR was constructed through the raw amino acid sequence using the Phyre2 webserver, basing the model on a sequence alignment on 6 retinal proteins whose structures are known. The six templates share a 18-24% sequence identity with the AntR target sequence and the alignment of the 212 amino acid residues of the sequence received a confidence score of 100%. Three amino acid residues, M1, V214 and G215 were modelled *ab initio.*

Three systems were prepared for this study, each with a different isomeric state of the retinal. AntR is found to retain connections with both side of the membrane for every isomeric state of the retinal, regardless of its orientation. The all-*trans* and 13-*cis*,15-*syn* AntR feature stable connections with carboxylates on the EC side of the membrane, namely D2, E6, E62 and D167. The latter is the equivalent of one member of the proton release group of BR [220], E194.

All shortest pathways connecting the EC carboxylates and the all-*trans* and 13-*cis*,15-*syn* RSBs pass through the counter ion D185 and the gating arginine R67 and are very frequently sampled. D185 plays a crucial role in connecting an EC-oriented RSB to the CP side, connecting to S74 which in turn can connect to Y193, E81 ($D96^{BR}$) and D195 (conserved group among AntRs). Experimental data suggested that R67 and D167 would be important for the proton uptake because the photocycle is largely perturbed when they are mutated [185], while E81 cannot be the sole proton acceptor. Instead, it is suggested that a network of highly dynamic hydrogen bonds in CP side provides alternative proton transferring pathways.

The CP-oriented 13-*cis*,15-*anti* retinal connects to E81 and D195 via a complex network of H-bonds. An extended water chain that between the 13-*cis*,15-*anti* SB and D195 was sampled when the water threshold was increased to 8 molecules. The chain is sampled 78.8% of the time, with an average water-wire length of L=6.2. The water

molecules of the chain are stabilized by with the retinal Schiff base, S74 of helix C and are also supported by H-bonding to the backbone oxygen atoms of Y150 and F192 of helix D and F, respectively. S74 mediates transient bridging of the 13-cis,15-anti RSB to the extracellular H-bond network, allowing the 13-cis,15-anti retinal to connect transiently to the EC side.

Through the means of centrality measures [184, 185] and a newly developed measure of *USP* [186], R67 is found as the most important node in the graphs in terms of network connectivity and communication, in addition to S74. The 13-*cis*,15-*anti* state showed smaller centrality values compared to the all-*trans* state for potentially important amino acid residues such as R67, S74, E81, D185 and Y193 which suggests that the networks rearrange upon the retinal isomerization, strongly affecting network communications. The S74A mutation results into largely delayed kinetics of the photocycle and weakens the proton transport [185], validating the prediction of the centrality measurements on its importance in the H-bond networks.

In the last chapter of this dissertation, large-scale analyses on datasets of hundreds of TM protein in search of conserved H-bond motifs important for function, accompanied by MD simulations of an ion channel and a water channel were presented. Two main datasets of α-helical TM proteins solved with X-ray crystallography or cryo-Electron Microscopy were compiled, with members from up to 26 superfamilies. I developed a refinement technique to include only the highest quality resolution structures per protein entry, in addition to hand curation of proteins with high sequency identity match to tackle possible redundancy. *Set-high* included structures with resolutions $\leq$ 2.5 Å and *Set-low* included structures with resolutions 2.5 Å < res. $\leq$ 3.5 Å. The initial dataset of 1439 structures was split into *Set-high* with 200 members and Set-low with 483 members. *Set-highU* was presented as a refinement of *Set-high* and features 147 unique structures.

Using the Transporter Classification Database (TCDB) [259], the *protein-groups* were introduced to further categorize superfamilies according to their biological role. The diverse Rhodopsin-like receptors superfamily was treated as a separate *protein-group*, and was further dissected, using *Set-highU* as the parent dataset, effectively generating a subset of "Microbial and algal rhodopsins" (*Set-high-mr*) and "A, B, C and F GPCRs" (*Set-high-gpcr*). The Hemolysin-III family and Heliorhodopsin families are distinct as per the TCDB classification and they were treated independently as *Set-high-hemo* and *Set-high-helio* for heliorhodopsin, respectively. The resulting subsets of the *Rhodopsin-like receptors and pumps* (65) are *Set-high-mr* (28), *Set-high-gpcr* (35), *Set-high-hemo* (1) and *Set-high-helio* (1). H-bonding between Ser/Thr hydroxyl groups and carboxylate groups of Asp/Glu amino acids residues is present in all protein structures of *Set-high* and *Set-highU*, except for Proton-translocating transhydrogenases and Magnesium ion-transporter-E.

When using the TCDB classification, each protein-group of *Set-high* and *Set-highU* is found to contain at least one carboxylate-hydroxyl motif with a range of occurrence 88-93%, while the intrahelical carboxylate-backbone carbonyl of the *i*-3,4,5 relative position is found in every protein of every protein-group, and it is the most widely represented motif in this search. The combined carboxylate-Ser/Thr & Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position is a more selective motif and is not sampled as

frequently compared to its individual components. In the Rhodopsin-like receptors and pumps protein-group there are 45 proteins with the combined motif in *Set-high* and 32 proteins in *Set-highU*, which makes up for 49% occurrence in that set. The combined motif is found more frequently in the Rhodopsin-like receptors and pumps protein-group, which features many Aspartate/Glutamate and Serine/Threonine motifs in the center of the membrane core as well as towards the cytoplasmic and extracellular sides of the membrane, where prime examples of amino acid residues of critical importance for the mechanism of the protein is found, such as the proton release groups of BR [220].

Examples of motifs that are known or could possibly be important in the function of their respective proteins were detected. In Archeaorhodopsin-2 the motifs T95 and D120 ($D156^{ChR2}$) and T50-D101 ($D96^{BR}$, $H134^{ChR2}$) are found. In BR the motifs T90-D115 and T46-D96 were detected and their equivalents using the H. salinarium BR numbering, such as Archaerhodopsin-3, Halorhodopsin from *N. pharaonic*, Coccomyxarhodopsin, *E. sibiricum* rhodopsin, Acetabularia rhodopsin I, among others. In the light-driven Na^+ pump KR2 a motif between S70-D116 ($T90^{BR}$) is sampled, with similar motifs sampled in 10 GPCR's, including Jumping spider rhodopsin.

Most of the His-His motifs detected in the datasets belong to proteins of the ammonium channel transporter superfamily where the highly conserved histidines are an essential part of the transport mechanism in ammonia channels at the center of the transmembrane domain of the protein. The highly conserved His-His motif acts as a proton relay in the transport of ammonium. Anabaena sensory rhodopsin features two histidine pairs, H21-H219 and H8-H69. And the Calcitonin type 1 receptor of the GPCR Secretin B family has a continuous network of three histidines, H149-H334-H370 H-bonded in a series, similar to the proton-transferring motif in the ammonia channel AmtB.

In the proton translocating pyrophosphatase a motif between D294 and N738 are sampled. This H-bond is likely important for the functioning of the transporter, because N738 could stabilize water molecules of the proton transfer pathway. The majority of the Asp-Asn motifs are found within the Rhodopsin-like receptors and pumps with 25 proteins containing at least one motif.

Aquaporin-1 (Aqy1) is a host for a plethora of structural motifs, such as the ar/R and the dual NPA motif. The crystal structure of the Aqy1 features two large H-bond clusters, with 10 and, 19 groups respectively. The larger H-bond cluster spans through the length of the protein, bridging the extracellular and cytoplasmic regions, and it contains the high USP groups, and motif members, N112, E51, E175 and the backbone of L223. On the cytoplasmic side of the network, the NPA-asparagine N112 is H-bonded to Q137, which in turn H-bonds to E51, similar to an Asp/Glu-Asn motif. A motif between E51 and S107 in the cytoplasmic side of the membrane is also observed among others.

In the crystal structure the networks of the selectivity filter and the NPA motif are separated when excluding the water molecules from the networks. In all 4 MD simulations they are connected via the new connections that R227 and N224 establish. The initial 4-amino acid residue cluster in the selectivity filter is merged to the central High-BC cluster that was identified in the crystal structure. At the selectivity filter of Aqy1, the sidechain of H212 H-bonds to the backbone carbonyl of L208. The intrahelical

hydroxyl-carbonyl motif S228-N224 H-bond is sampled on at moderate occurrence rates (max. 26.6%). The motifs between E51-S107 and E175-T219 are sampled very frequently, with occurrence rates 100% in most cases.

In simulations of the ChR2 dimer, H-bond networks differ between the two monomers, similar to the results of C1C2. An unprotonated E123 can salt-bridge to K93 and to T127, similar to C1C2. The E123-T127 is additionally sampled above 95% of time in the 13-*cis* simulations in both monomers. The second counterion can connect to K93 to form an additional salt-bridge, and S256 to form an H-bond motif. Through direct H-bonds, the networks can extend to H249, R120 and E97 in the EC side. N258 forms a small interaction network with S63 for a Serine/Threonine-Asparagine motif, and in turn S63 H-bonds to A59 through an intra-helical Serine/Threonine hydroxyl group to the *i*-4 backbone carbonyl. Upon isomerization of the RSB to 13-*cis*,15-*anti*, the H-bond network is heavily rearranging. Similar to AntR [185], I detected two long range water-mediated connections between the RSB and H134 and E82 in monomer-A (81% - L=3.6 and 90% - L=3.9, respectively), that could serve as reprotonation pathways for the RSB, since it is suggested that H134 should not be excluded as the proton donor [83]. The presence of the wire is tightly connected with the conformation of H134 (*flat* or *up*), and the H-bonding between H134 and E83. ChR2 maintains very frequent water mediated connections to amino acid residues with a carboxylic sidechain located in the EC side of the membrane. Similar to C1C2 and AntR, E97 and E235 were investigated as suggested proton release groups (proton uptake for AntR). USP measures indicate that E97 and E235 are important for the communication of the graphs and removal of those nodes could lead to major disruptions in the networks. Following the model of the rearrangement of D253 upon retinal isomerization, in the K state [294], D253 has and increased USP value in the 13-*cis*,15-*anti* model, as compared to the all-*trans*. The DC gate is sampled in the wild-type ChR2 at higher occurrence rates, as compared to C1C2 [42] and is a water mediated connection most of the time.

Chapter 7 Aspects

This work has set a new perspective in the study of membrane proteins. With the implementation of *Bridge*, a next generation-analysis algorithm was made publicly available. Pathways sampled from MD and detected using *Bridge* could the evaluated on their energetics, to answer the questions if they could serve as proton-transferring pathways. Centrality computations based on graphs predicted the importance of S74 and Y193 in AntR [185], which was confirmed by the effect the point mutations have on the function of the protein. This approach could serve as suggestion for site directed mutagenesis, in combination with the knowledge-based criteria that are usually employed. With the ChR2 crystal structure solved in 2017, many questions arise for the function of the protein and the proton-transferring pathways. H-bond pathways similar to C1C2 were found in ChR2. A long range interaction between the RSB and H134 was detected, which poses the question of the primary proton acceptor in ChR2 [80, 83]. Since the technique of Molecular Dynamics is of a classical approach, a limitation has to be acknowledged. Only the motions of the nuclei are considered, while any explicit electronic motion is neglected. Instead, elements of electronic structure and properties are reflected in the parameters used in force field equation to describe the system. Thus, any bond breaking or new bond formation along with polarization effects, for example proton transferring, are not modelled with this approach. The quality of the parameters used in the force-field equation (potential energy functions) can will determine how reliable the dynamics of the system are, and thus how reliable the predictions from the simulation are. This is especially true for co-factors, such as the retinal chromophore, which can be challenging to represent with a classical approach, due to its complex electronic structure. On the other hand, the speed and relative accuracy of the technique allows for large trajectories to be sampled with reasonable use of computation resources and enabling the detection of dynamic H-bonds. Here, treatment of the suggested pathways would require QM techniques or combined QM/MM to further evaluate their validity as far as proton transferring events go.

A similar long water wire connects the RSB to amino acid residues that are thought to participate in a proton acceptor complex in AntR. Mutating the candidate members of the suggested proton acceptor complex could provide insight into the conformational changes and how the networks would re-arrange to accommodate for the mutations. In the unanticipated network between the two retinal Schiff bases, it would not be expected for a proton to transfer from one Schiff base to another. It would be more reasonable to recommend that this extended network couples conformational dynamics in the dimer. Since a network like this has not been suggested in the literature so far, it would need to be validated experimentally. Although, cooperativity in oligomeric

structures has been suggested in the literature for the voltage-gated proton channel Hv1 [304], the potassium channel $K_v7.4$ [305], bacteriorhodopsin [306] and voltage-gated sodium channel $Na_v1.5$ [307]. The aspect of cooperativity of C1C2 in regard to the network identified here was discussed by Bondar in ref. [308]. It could be suggested that MD simulations would be prolonged, and a new set of analysis would be performed in order to evaluate if the H-bonds and H-bond pathways are reproduced. The universal protocol for large datasets could be expanded to analyze even more superfamilies and protein members, since the repositories of the Protein Data Bank and the OPM database are regularly updated with new entries.

Within the context of the Sonderforschungsbereich (SFB) 1078 "Protonation Dynamics in Protein Function", where four main proteins are studied, in three main research areas, this work has provided new insight in the approach of analyses of large datasets of dynamic H-bond networks or static structures from a methodological standpoint, as well as detecting new, and never reported in the literature in the past, H-bond networks that could serve as candidates for proton transferring pathways pushing the forward what is known about the proteins at hand and getting one step closer to answer the fundamental questions of the field.

For example, Photosystem-II [309-320], a very large macromolecular protein complex, comprised of numerous subunits catalyzes a chemical reaction that splits water into molecular oxygen, electrons and protons, using light and is essential for life. A key open aspect of PSII is how the protons generated at the oxygen evolving complex are transported/transferred over very long distances to the bulk. Although numerous crystal structures of PSII are already available, and some specific amino acid residues are thought to be of great importance for the transfer of protons, the exact pathways are still not known. In the proton pump cytochrome *c* oxidase [321-333], a terminal oxidase in the respiratory chain, the proton translocation across the membrane is coupled to a reduction reaction of oxygen to water. Nevertheless, the mechanism of function of the protein is not yet fully understood, along with the individual reaction steps taking place and a still unknown proton loading site.

Similarly, in the light-gated ion channel, Channelrhodopsin-2 [62, 63, 70, 78-80, 82, 83, 101, 215, 334, 335], where the photo-isomerization of the retinal chromophore is coupled with proton transfer steps and passive flow of cations. A key open question is how protons are transferred across long distances in the polar environment of channelrhodopsin with the proton release group(s) being unknown, as of yet. I strongly believe that the wide range of capabilities that the algorithm package, *Bridge*, offers can be of great contribution into tacking the open questions in the main proteins of the SFB.

All of the aforementioned proteins, as well phytochrome, which can be considered conceptually close to channelrhodopsin in terms of the chromophore isomerization and activation of the photocycle, could benefit from the features of *Bridge,* both when studying static structures and/or dynamic trajectories. When considering that those proteins have a chromophore or a cofactor, of a definitive starting point of the proton transfer (oxygen evolving center-PSII) and possible proton transfer pathways (D, K, channels-CcO), identifying stable H-bond networks and pathways through *Bridge* could provide insight. More specifically, using the cofactor or chromophore or the

manganese cluster as the beginning of the pathways in question one can obtain 2-dimensional H-bond maps of the shortest pathways that begin from the molecule of interest. There are cases where specific amino acid residues are suggested to play an important role in proton transfer in the proteins mentioned above. While the idea of a manual search in a crystal structure seems possible in relatively small proteins, such as rhodopsins, it was showcased in this thesis how complex the dynamic H-bond networks of the hydrophilic environment of a retinal protein can be. When considering the case of PSII, the distances are very long, and the combinations of pathways would render non-automated analyses incomplete.

Through *Bridge*, a criteria-controlled pathway search will provide the user with all possible solutions on the respective query. Additionally, through the means of centrality measures, the importance of already suggested amino acid residues could be complemented, as well as new amino acid residues could emerge as important for the communication of the networks. This would enable for one to make predictions that would have to be later experimentally confirmed, as seen for S74 in AntR [185].

Bridge can be a highly adaptable and expandable algorithm package. One and a half years after its original release, *Bridge* was expanded and a standalone graphical interface for H-bond analyses was introduced as *Bridge2* [43]. In addition to the interface, Bridge2 included analyses of hydrophobic interactions using a distance-based criterion between the carbon atoms [43]. In the RBD of the spike protein, a conserved hydrophobic group was detected, as a part of a larger hydrophobic cluster [43]. Hydrophobic contacts between RBDs where characterized as infrequent [43].

Another aspect on the impact this work could have, especially within the SFB, is the possibility for one use *Bridge* as a platform for the analysis of H-bond networks. As shown in chapter "*Conserved H-bond motifs in membrane transporters*" a large dataset was constructed, and *Bridge* was employed to perform large-scale analyses of static structures detecting H-bond motifs important for proton transfer and function. In this way, it is possible for one to construct a new dataset, for example all structures of PSII, and perform comparative analyses between the systems for the sampling frequency pathways with the same components, motif detection and the capability for one to implement any new analysis routine using *Bridge* as a platform or integrating directly to the core package. A prime example is the development of the algorithm "*Conserved (C)-graphs*" [336] by Bertalan et al., two years after the original release of *Bridge*. C-graphs is a standalone Python program that features a graphical interface. *C-graphs* is capable of analyzing datasets of multiple structures at the same time and computing conserved graphs of H-bonds as compared to a separate graph-per-structure approach. Additionally, it uses clustering algorithms to cluster water molecules that are co-crystallized with the proteins and detect conserved water-binding sites [336]. Using the core Bridge algorithm to compute direct and water-mediated H-bonds, C-graphs was able to identify conserved H-bond networks in representative proteins of Class-A GPCRs, such as squid rhodopsin [336]. A common network between squid rhodopsin, adenosine A2A receptor, jumping spider rhodopsin-1 and the dark state of bovine rhodopsin was also identified. Bovine rhodopsin showed large rearrangement of networks between the active and active-like states [336]. It thus appears, that the original release of *Bridge* has already allowed for

further developments in the field of computational biophysics, and one could expect that this is only the beginning.

Appendices

Supplementary Tables

Table A.1. Summary of the dataset *Set-high*. For each structure included in the dataset, PDB ID and the resolution in Å are reported. Adapted from ref. [226].

PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)
3zoj	0.88	4dx5	1.9	6ibb	2.12	5zkc	2.3	2zbd	2.4
6s6c	1.07	6a6m	1.9	6tos	2.13	5dys	2.3	2zbf	2.4
5i32	1.18	1h2s	1.93	6i9k	2.14	6uva	2.3	2zxe	2.4
6ufe	1.2	6rnk	1.94	3ar4	2.15	6wzg	2.3	5i20	2.4
5zim	1.25	4n7w	1.95	6vx9	2.17	4r7c	2.3	6hd8	2.4
3b9w	1.3	5wiu	1.96	6afw	2.18	2q67	2.3	4cbk	2.42
1u7g	1.4	5wqc	1.96	3zk1	2.2	6f7h	2.3	6rz6	2.43
4y9h	1.43	2ns1	1.96	6k6k	2.2	3k3f	2.3	6rz7	2.43
4xtl	1.45	3tds	1.98	6c1r	2.2	1wpg	2.3	4jkv	2.45
3ldc	1.45	6eu6	1.98	5uiw	2.2	5bz3	2.3	5hvx	2.45
5aez	1.47	5jsi	2	5x93	2.2	4wd8	2.3	3spc	2.45
6su3	1.5	6gyh	2	5tzr	2.2	4jq6	2.31	4lp8	2.46
5ax0	1.52	6nwd	2	6li0	2.2	6wk9	2.32	5d5a	2.48
2b2h	1.54	1xio	2	4xnv	2.2	6mwd	2.33	1vgo	2.5
4v1g	1.55	6yc3	2	3vw7	2.2	6vx8	2.33	6eyu	2.5
3ouf	1.55	6igk	2	1u19	2.2	2wgm	2.35	4pxk	2.5
5g28	1.57	5kuk	2	6hlp	2.2	1ap9	2.35	4yzi	2.5
6jo0	1.65	6o9u	2	6ffi	2.2	4mrs	2.35	6h7n	2.5
2f2b	1.68	3d9s	2	1s5h	2.2	6fk6	2.36	3odu	2.5
5b0w	1.7	3llq	2.01	2qks	2.2	4ezc	2.36	4pxz	2.5
2jaf	1.7	2zzl	2.03	1j4n	2.2	6vx7	2.36	6m9t	2.5
5nm4	1.7	3c02	2.05	3rlf	2.2	6by3	2.37	2z73	2.5
2ih3	1.72	3cn5	2.05	3n5k	2.2	6eid	2.39	6iiu	2.5
3wqj	1.8	6m96	2.05	5uni	2.2	5a8e	2.4	6x1a	2.5
4n6h	1.8	6nwf	2.06	4u9n	2.2	6ps2	2.4	5ee7	2.5
3gd8	1.8	5hwy	2.1	3zuy	2.2	2rh1	2.4	6fj3	2.5
2xqu	1.84	4bem	2.1	1ymg	2.24	6uus	2.4	4jta	2.5
6kfq	1.84	2bl2	2.1	3kcu	2.24	6bd4	2.4	6wel	2.5
4qi1	1.85	4l35	2.1	5vk6	2.25	5lwy	2.4	4wfe	2.5
3v5u	1.9	4bvn	2.1	4x1h	2.29	6is6	2.4	4wff	2.5
4f4s	1.9	6qzh	2.1	4kpp	2.3	4uuj	2.4	4g7v	2.5
6k6i	1.9	5c1m	2.1	4k1c	2.3	2r9r	2.4	5jmn	2.5
6lm1	1.9	6x18	2.1	6tqj	2.3	2b6p	2.4	4ymu	2.5

3qap	1.9	6x19	2.1	1iw6	2.3	4fc4	2.4	3ar2	2.5
5jje	1.9	1z98	2.1	3ug9	2.3	4ene	2.4	3fgo	2.5
6sqg	1.9	1ldf	2.1	3vvk	2.3	5ah3	2.4	5zmw	2.5
3ddl	1.9	3kly	2.1	4hyj	2.3	6qd5	2.4	4uu0	2.5
1ors	1.9	3hd6	2.1	4tl3	2.3	2nq2	2.4	6jxh	2.5
6qzi	1.9	5y78	2.1	3x3b	2.3	3wmg	2.4	4n7x	2.5
2o9g	1.9	6tod	2.11	4amj	2.3	1su4	2.4	5a1s	2.5

Table A.2. Summary of the dataset *Set-low*. For each structure included in the dataset, PDB ID and the resolution in Å are reported. Adapted from ref. [226].

PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)
6c1m	2.52	5vkp	2.8	5dhg	3	1l7v	3.2	6nq2	3.4
2zbg	2.55	6a90	2.8	2x72	3	4tqu	3.2	6nt3	3.4
6oh2	2.58	6o1u	2.8	5xez	3	6msm	3.2	6nr3	3.4
6or2	2.59	3q7k	2.8	5yqz	3	6mit	3.2	6u8a	3.4
3n8g	2.59	5yhf	2.8	6nbf	3	4ycm	3.2	6mho	3.4
4knf	2.6	5b57	2.8	5v57	3	5xaa	3.2	6cnm	3.4
5zih	2.6	6bl6	2.8	3fb6	3	3fps	3.2	6rmg	3.4
2ydv	2.6	6rah	2.8	6o72	3	6jju	3.2	5och	3.4
3eml	2.6	3ba6	2.8	6o1p	3	4umv	3.2	4f4c	3.4
5wf5	2.6	2eau	2.8	5t4d	3	5oc9	3.2	4q9h	3.4
4uhr	2.6	5ylv	2.8	3ne2	3	6qp6	3.2	6c0v	3.4
5vbl	2.6	3wgu	2.8	1kpl	3	6p48	3.2	4hyt	3.4
5u09	2.6	6roh	2.8	6d0j	3	4ekw	3.21	6i1z	3.4
5dsg	2.6	6psy	2.8	4k0j	3	5b58	3.21	6qxa	3.41
4xes	2.6	6i1r	2.8	6csx	3	5tzy	3.22	4res	3.41
3oax	2.6	5t1a	2.81	3d31	3	6k7h	3.22	6rdm	3.44
4qin	2.6	5zkp	2.81	2hyd	3	6k7i	3.22	3syq	3.44
4av3	2.6	4kyt	2.83	2c9m	3	5oge	3.22	5khn	3.44
4twk	2.6	6k7l	2.83	3b9r	3	4u2p	3.24	6j5j	3.45
6jpa	2.6	6k7n	2.84	1xp5	3	6qum	3.25	6mvx	3.45
6a95	2.6	5vb8	2.85	5ncq	3	5wo7	3.25	4kfm	3.45
1xl4	2.6	4ayt	2.85	6n24	3	5iwk	3.25	6pw5	3.45
4gx0	2.6	3zdq	2.85	6n26	3	4u1w	3.25	6d7o	3.45
2qi9	2.6	6mjp	2.85	6n27	3	5lj7	3.25	3wkv	3.45
3wmf	2.6	4byg	2.85	2xok	3.01	4ycl	3.25	5ztf	3.45
3ar8	2.6	4rdq	2.85	5w0p	3.01	3uza	3.27	1zcd	3.45
4qim	2.61	6cm4	2.87	4g1u	3.01	6hll	3.27	3t9n	3.46
3spg	2.61	5dqq	2.87	3w5a	3.01	4rwa	3.28	5x5y	3.46
5xap	2.61	6bco	2.88	6a6n	3.02	5z96	3.28	4fi3	3.47
4ntj	2.62	3pbl	2.89	6m97	3.03	5zx5	3.28	5xas	3.47
4x89	2.62	4xt1	2.89	6k7k	3.04	5irz	3.28	6r7x	3.47
6fiz	2.63	6aei	2.89	4kjs	3.05	3j5p	3.28	6qv1	3.48
1gzm	2.65	6rde	2.9	4dxw	3.05	4csk	3.28	4f4l	3.49
3b9b	2.65	6rdg	2.9	3zrs	3.05	6n2d	3.3	6dqj	3.49

6rd5	2.69	6rdj	2.9	4oj2	3.05	3pwh	3.3	5wj9	3.49
5vn7	2.7	6rer	2.9	4myc	3.06	6gps	3.3	4dbl	3.49
6eig	2.7	6csm	2.9	6co7	3.07	6oij	3.3	5uj9	3.49
4fbz	2.7	6csn	2.9	1t9x	3.08	5t04	3.3	3v5s	3.5
4iar	2.7	6a94	2.9	6k7j	3.08	4zwj	3.3	5fl7	3.5
4ib4	2.7	5tvn	2.9	6hbu	3.09	6oya	3.3	6c6l	3.5
6bqh	2.7	3oe0	2.9	6hzm	3.09	6e3y	3.3	5uig	3.5
3vg9	2.7	4z35	2.9	6rdv	3.1	6b3j	3.3	4gpo	3.5
2vt4	2.7	4djh	2.9	6re7	3.1	4l6r	3.3	5wb2	3.5
5o9h	2.7	3cap	2.9	6o7u	3.1	5l7i	3.3	4mqs	3.5
6gpx	2.7	5zkq	2.9	6drz	3.1	3fb7	3.3	6ddf	3.5
6k1q	2.7	6ak3	2.9	4rws	3.1	3f5w	3.3	6nbh	3.5
6igl	2.7	6do1	2.9	3rze	3.1	5wie	3.3	5gpj	3.5
5cxv	2.7	5unh	2.9	6b73	3.1	6ebk	3.3	5tj6	3.5
5zbq	2.7	5v56	2.9	3beh	3.1	4rue	3.3	5h3o	3.5
4rwd	2.7	2a79	2.9	4h33	3.1	6e1m	3.3	5u6o	3.5
5te3	2.7	3lut	2.9	6bqv	3.1	6cud	3.3	4xdl	3.5
4xnw	2.7	3lnm	2.9	6bwj	3.1	6drj	3.3	6nt4	3.5
3ayn	2.7	6rv3	2.9	2onk	3.1	6puo	3.3	6mhv	3.5
6j20	2.7	6oo3	2.9	6eti	3.1	6o20	3.3	6e7p	3.5
5vew	2.7	6o1n	2.9	6s8n	3.1	6jb1	3.3	4j7c	3.5
6g7o	2.7	6rld	2.9	4h1w	3.1	4u1x	3.3	6cnn	3.5
5klb	2.7	3ne5	2.9	6mgv	3.1	4c48	3.3	4u5b	3.5
4pl0	2.7	3puz	2.9	6n23	3.1	3aqp	3.3	3org	3.5
5mkk	2.7	3tui	2.9	3jyc	3.11	4ayw	3.3	4y7k	3.5
6iu3	2.7	4ayx	2.9	6du8	3.11	6hrc	3.3	2oar	3.5
4umw	2.7	5c78	2.9	6fuf	3.12	5ko2	3.3	4p6v	3.5
4i0u	2.7	6rai	2.9	6qv0	3.12	5xu1	3.3	2rdd	3.5
6n25	2.7	5mrw	2.9	5cbg	3.14	5nik	3.3	6r4l	3.5
4mbs	2.71	6roj	2.9	6bcj	3.14	5mpm	3.3	6oeu	3.5
3w9i	2.71	6qti	2.9	6bhu	3.14	6k7g	3.3	6e1h	3.5
6d27	2.74	3h90	2.9	6fkf	3.15	6psx	3.3	6d4j	3.5
4buo	2.75	2iub	2.9	5t0o	3.15	5jrw	3.3	6rvd	3.5
4zj8	2.75	6n28	2.9	5l22	3.15	5vre	3.3	4xwk	3.5
4i9w	2.75	6bpl	2.91	4cz8	3.15	6p49	3.3	5kpj	3.5
6p6x	2.75	6bpp	2.92	4z9g	3.18	4wis	3.3	6ral	3.5
4nef	2.75	5w3s	2.94	6n30	3.2	3um7	3.31	6s8g	3.5
5xan	2.75	2zy9	2.94	6re1	3.2	5wo6	3.31	6iu4	3.5
4bbj	2.75	5zkb	2.95	6o7t	3.2	6dqn	3.33	4nab	3.5
5kw2	2.76	5hk7	2.95	3am6	3.2	6j5i	3.34	5ksd	3.5
5uld	2.78	5irx	2.95	5uen	3.2	6niy	3.34	3b8e	3.5
4lde	2.79	6cvl	2.95	3sn6	3.2	3nog	3.34	3kdp	3.5
4r9u	2.79	6k7m	2.95	5jqh	3.2	5u74	3.34	2yvx	3.5
6rdb	2.8	3tlm	2.95	3oe6	3.2	5uja	3.34	4atv	3.5
6rdp	2.8	6mxt	2.96	5xsz	3.2	6cju	3.35	4czb	3.5
5azd	2.8	4r0c	2.96	5ywy	3.2	5u73	3.35	4cz9	3.5

4wav	2.8	4k5y	2.98	5xjm	3.2	4hkr	3.35	6r65	3.5
4iaq	2.8	3sya	2.98	5l7d	3.2	6coy	3.36	6qpc	3.5
5xra	2.8	6ows	2.98	3f7v	3.2	3udc	3.36	6p46	3.5
5zty	2.8	4bwz	2.98	4bw5	3.2	6d7p	3.37		
6aky	2.8	5aji	2.99	6c9a	3.2	5w81	3.37		
5lwe	2.8	4kjr	3	6agf	3.2	5z1w	3.38		
5glh	2.8	6n2y	3	6j8g	3.2	5gko	3.39		
6me2	2.8	6n2z	3	6j8i	3.2	3v3c	3.4		
6me6	2.8	6re4	3	5vb2	3.2	1uaz	3.4		
4u15	2.8	6red	3	1orq	3.2	5g53	3.4		
4grv	2.8	7c61	3	3vou	3.2	3kj6	3.4		
4dkl	2.8	6a93	3	6bqr	3.2	4daj	3.4		
5ndd	2.8	5tud	3	6o77	3.2	4ej4	3.4		
3v2y	2.8	6bqg	3	6o6r	3.2	3f7y	3.4		
4ww3	2.8	2ycw	3	6u88	3.2	3fb8	3.4		
4zud	2.8	6n4b	3	6mhs	3.2	6gyo	3.4		
5ung	2.8	3uon	3	3nd0	3.2	6gyn	3.4		
6o3c	2.8	5zbh	3	5aex	3.2	3ukm	3.4		
4or2	2.8	6os9	3	5nc5	3.2	4ruf	3.4		
3fb5	2.8	6osa	3	6baj	3.2	6c96	3.4		

Table A.3. Summary of duplicate structures in *Set-high*. The protein name, PDB ID, resolution, sequence match, organism and notes on the protein are shown. Bacteriorhodopsin entries 4QI1 and 4PXK are presented separately since they show 56-57% sequence identity matches with 5ZIM and come from different organisms. Between them, they share a 54% sequence identity match and will be treated as separate entries. Adapted from ref. [226].

Protein	PDB	Res. (Å)	Sequence identity (%)	Organism	Notes
Bacteriorhodopsin	5ZIM	1.3	100	*H. salinarum*	Ground state
	4Y9H	1.4			Ground state
	2ZZL	2.0			M intermediate
	1IW6	2.3			K intermediate
	1AP9	2.4			Ground state
	4QI1	1.9	56*	*Haloquadratum walsbyi*	Ground state
	4PXK	2.5	57*	*Haloarcula marismortui*	D94N
Sodium pump KR2	4XTL	1.5	*97-100*	*Dokdonia eikasta*	Monomer, pH 4.3
	6YC3	2			pH 8.0
	3X3B	2.3			Dimer, pH 4.0
Chloride-pumping MastR	6K6I	1.9	*99-100*	*M. repens*	pH 4.9, 93K
	6NWF (6XL3)	2.1			pH 4-4.9, 93K
	6K6K	2.2			N63A/P118A
Halorhodopsin	5B0W	1.7	*100*	*Natronomonas pharaonis*	red adapted, 11-cis
	3VVK	2.3			M-like state
	2JAF	1.7	*58*	*H. salinarum*	T203V
cysteinyl leukotriene receptor 2	6RZ6	2.4	*100*	*H. sapiens*	pH 8
	6RZ7				pH 7

Heliorhodopsin	6SU3	1.5	*46*	*Acetabularia acetabulum*	pH 8.8
	6IS6	2.4		*Thermoplasmatales archaeon*	pH 8
Archaerhodopsin-2	3WQJ	1.8	*100*	*Halobacterium sp.*	pH 7.0
	1VGO	2.5			pH 5.2
Succinate receptor SUCNR1 (GPR91)	6RNK	1.9	*99*	*Rattus norvegicus*	K18E/K269N, pH 7
	6IBB	2.1			pH 4.8-5.2
Sensory rhodopsin. II	3QAP	1.9	*100*	*N. pharaonis*	Ground state
	5JJE				SRII/HtrII Complex
	1H2S				Ground state
Anabaena sensory rhodopsin	1XIO	2	*99*	*Nostoc sp.*	Ground state
	4TL3	2.3			D217E
Endothelin receptor type B	6IGK	2	*91*	*H. sapiens*	Complex with Endothelin-3
	5X93	2.2			Complex with antagonist K-8794
Beta-2 adrenergic receptor	2RH1	2.4	*99-100*	*H. sapiens*	Inactive state, dimer
	6PS2				Inactive state 2
	5D5A	2.5			Inactive state, in meso, 100 K
Beta-1 adrenergic receptor	4BVN	2.1	*98-99*	*Meleagris gallopavo*	Inactive state
	4AMJ	2.3			Inactive state, bound biased agonist
	5A8E	2.4			Inactive state, inverse agonist bound
	6H7N	2.5			Active state, partial agonist and nanobody bound
Glucagon-like peptide 1 receptor	6X18	2.1	*100*	*H. sapiens*	GLP-1 peptide hormone bound
	6X19				Non peptide agonist bound
	6X1A	2.5			Non peptide agonist bound
Adrenomedullin 2 receptor	6UVA	2.3	*100*	*Homo sapiens*	Complex with adrenomedullin-2 peptide
	6UUS	2.4			Complex with adrenomedullin peptide
Orexin-1 receptor	6TOD	2.1	*100*	*H. sapiens*	Inactive state, complex with EMPA
	6TOS				Inactive state, complex with GSK1059865
Bovine rhodopsin	1U19	2.2	*100*	*Bos taurus*	-
	4X1H	2.3			With transducin peptide
	5DYS				T94I,without transducin peptide
	6FK6	2.4			N2C/D282C
Spinach Aquaporin SoPIP2	3CN5	2.1	*99*	*Spinacia oleracea*	S115E, S274E
	1Z98				Closed conformation
Aquaporin-0	1YMG	2.2	*100*	*B. taurus*	-
	2B6P	2.4			Open pore state
Formate transporter 1, FocA	3KLY	2.1	*56*	*Vibrio cholerae*	-
	3KCU	2.2		*E. coli O157:H7*	
K+ selective NaK ion channel	6UFE	1.2	*100*	*Bacillus cereus*	Open state
	3OUF	1.6			D49Y
Potassium channel KcsA	2IH3	1.7	*90-100*	*Streptomyces lividans*	Conductive conf.
	1S5H	2.2		*chimera*	T75C, closed
	5VK6	2.3		*S. lividans*	Open deep-inactivated conf.

	6BY3	2.4		*S. coelicolor*	Open and conductive conf., T75A
	4UUJ			*S. lividans*	Complex
Kv1.2-kv2.1 paddle chimera Potassium channel	2R9R	2.4	*100*	*Rattus norvegicus*	With beta subunit
	4JTA	2.5		*R. norvegicus*	Complex with Charybdotoxin
TRAAK K+ channel	4WFE	2.5	*100*	*H. sapiens, Mus musculus*	Conductive
	4WFF				Nonconductive
Ammonium transporter Mep2	5AEZ	1.5	*99*	*Candida albicans*	-
	5AH3	2.4			R452D, S453D
Multidrug transporter AcrB	4DX5	1.9	*100*	*E. coli*	Asymmetric complex
	5JMN	2.5			Asymmetric, Fusidic acid bound
Calcium ATPase	3AR4	2.2	*100*	*Oryctolagus cuniculus*	-
	3N5K				E2-Pi state, conf. 7
	1WPG	2.3			E2-Pi state, conf. 1
	1SU4	2.4			E1-2Ca state
	2ZBD				E1P-ADP state
	2ZBF				E2 state
	3AR2	2.5			E1-2Ca state
	3FGO				E2-Pi state, conf. 5
	4UU0				E2 state (Ca-free), conf. 2
	5ZMW				E2 state (Ca-free), conf. 11
Bestrophin-2	6VX9	2.2	*100*	*B. taurus*	Ca^{2+}- unbound state 1
	6VX8	2.3			Ca^{2+}- unbound state 2
	6VX7	2.4			Ca^{2+}-bound state
Sodium/Calcium Exchanger	3V5U	1.9	*100*	*Methanocaldococcus jannaschii*	Structure 1
	5HWY	2.1			Structure 3

*) As compared to the *H. salinarium* bacteriorhodopsin (PDB ID: 5ZIM).

Table A.4. Summary of the dataset *Set-highU*. For each structure included in the dataset, PDB ID and the resolution in Å are reported. Adapted from ref. [226].

PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)	PDB	Res. (Å)
3zoj	0.88	3qap	1.9	4l35	2.1	1ymg	2.24	2nq2	2.4
6s6c	1.07	6sqg	1.9	4bvn	2.1	6tqj	2.3	3wmg	2.4
5i32	1.18	3ddl	1.9	6qzh	2.1	3ug9	2.3	2zxe	2.4
6ufe	1.2	6qzi	1.9	5c1m	2.1	4hyj	2.3	5i20	2.4
5zim	1.25	2o9g	1.9	6x18	2.1	5zkc	2.3	6hd8	2.4
3b9w	1.3	1ors	1.9	1ldf	2.1	6uva	2.3	4cbk	2.42
1u7g	1.4	4dx5	1.9	3kly	2.1	6wzg	2.3	6rz6	2.43
4xtl	1.45	6a6m	1.9	3hd6	2.1	6f7h	2.3	4jkv	2.45
3ldc	1.45	3v5u	1.9	5y78	2.1	4r7c	2.3	5hvx	2.45
5aez	1.47	6rnk	1.94	6tod	2.11	2q67	2.3	3spc	2.45
6su3	1.5	4n7w	1.95	6i9k	2.14	5tja	2.3	4lp8	2.46
5ax0	1.52	5wiu	1.96	3ar4	2.15	3k3f	2.3	6eyu	2.5
2b2h	1.54	5wqc	1.96	6vx9	2.17	5bz3	2.3	4pxk	2.5
4v1g	1.55	2ns1	1.96	6afw	2.18	4wd8	2.3	4yzi	2.5
5g28	1.57	3tds	1.98	3zk1	2.2	4kpp	2.3	3odu	2.5

6jo0	1.65	6eu6	1.98	6c1r	2.2	4k1c	2.3	4pxz	2.5
2f2b	1.68	5jsi	2	5uiw	2.2	4jq6	2.31	6m9t	2.5
5b0w	1.7	6gyh	2	5tzr	2.2	6wk9	2.32	2z73	2.5
2jaf	1.7	6nwd	2	6li0	2.2	6mwd	2.33	6iiu	2.5
5nm4	1.7	1xio	2	4xnv	2.2	2wgm	2.35	5ee7	2.5
2ih3	1.72	6igk	2	3vw7	2.2	4mrs	2.35	6fj3	2.5
3wqj	1.8	3d9s	2	1u19	2.2	4ezc	2.36	6wel	2.5
4n6h	1.8	5kuk	2	6hlp	2.2	6eid	2.39	4wfe	2.5
3gd8	1.8	6o9u	2	6ffi	2.2	2rh1	2.4	4g7v	2.5
2xqu	1.84	3llq	2.01	1j4n	2.2	6bd4	2.4	4ymu	2.5
6kfq	1.84	3c02	2.05	2qks	2.2	5lwy	2.4	6jxh	2.5
4qi1	1.85	3cn5	2.05	3rlf	2.2	4fc4	2.4	4n7x	2.5
4f4s	1.9	6m96	2.05	3zuy	2.2	2r9r	2.4	5a1s	2.5
6k6i	1.9	4bem	2.1	4u9n	2.2	4ene	2.4		
6lm1	1.9	2bl2	2.1	5uni	2.2	6qd5	2.4		

Table A.5. List of proteins with hydrogen bond motifs, discussed in the section “H-bond motifs of Serine/Threonine amino acid residues” of the main text. In the table the protein name, PDB ID and the reference are shown.

Protein	PDB	Reference
Archaerhodopsin-2 dimer	1VGO	[264]
Archaerhodopsin-2 trimer	3WQJ	[265]
Archaerhodopsin-3	6S6C	[266]
H. marismortui BR	4PXK	[267]
H. walsbyi BR	4QI1	[268]
crudorhodopsin-3	4L35	[269]
N. pharaonic halorhodopsin	2JAF	[242]
coccomyxarhodopsin	6GYH	[272]
E. sibiricum rhodopsin	4HYJ	[273]
Acetabularia rhodopsin I	5AX0	[243]
R. xylanophilus thermophilic rhodopsin	6KFQ	[274]
light-driven Na^+ pump KR2	4XTL/3X3B	[275]/[276]
Jumping spider rhodopsin	6I9K	[277]
F. nucleatum F_1F_o-ATP synthase	3ZK1	[278]
I. tartaricus F_1F_o-ATP synthase	2WGM	[279]
A. woodii Na^+-coupled ATP synthase	4BEM	[280]
E. hirae V-type Na^+-coupled ATP synthase	2BL2	[281]
NavAb voltage-gated Na^+ channel	6MWD	[282]
NavMs voltage-gated Na^+ channel	5HVX	[283]

Table A.6. Occupancies of H-bonds of Aqy1 of the R227-N112 H-bond clusters sampled during MD simulations in a POPE bilayer, with H44 and H194 Nε2-protonated. When the H-bond is sampled in all four Aqy1 monomers A-D, the total average occupancy computed by averaging the monomer occupancies is also reported. For each H-bond the occupancy calculated as an average from the last 100ns of each simulation is reported. Adapted from ref. [226].

	Occupancy (%)										
	A	B	C	D	Average		A	B	C	D	Average
Y31-N110	-	13.9	-	-	N/A	N160-T219	20.9	10.5	-	7.9	N/A
E51-T55	99.6	99.6	99.5	99.7	99.6	N160-G220	20.2	11.5	18.7	36.4	21.7
E51-S107	99.9	99.9	99.9	99.9	99.9	N160-A221	29.5	22.6	29.1	45.6	31.7
E51-G109	13.8	18.2	9.3	13.2	13.6	N160-R227	6.5	16.5	21.6	9.5	13.5
E51-N110	99.9	99.9	100	99.8	99.9	N160-S228	70.1	52.1	71.3	58.1	62.9
E51-L111	100	100	100	100	100	E175-T179	99.8	99.8	99.6	99.6	99.7
E51-Q137	99.9	99.8	99.9	99.9	99.9	E175-T219	100	100	100	100	100
S107-G109	57.8	80.4	65.8	72.7	69.2	E175-G222	100	100	99.9	99.9	99.9
S107-N110	51.3	16.1	58.3	41.9	41.9	E175-L223	100	100	100	100	100
N110-T116	7.8	71.3	-	15.9	N/A	T179-C183	68	69.6	64.7	65.4	66.9
L111-N224	68.4	67.6	61.3	67.8	66.3	T219-A221	86.6	88.7	82.7	89.5	86.9
N112-A114	99.6	99.4	99.4	99.3	99.4	T219-Y245	98.9	99.7	99.9	99.7	99.8
N112-V115	8.4	7.4	12.0	10.8	9.7	G220-R227	9.8	-	-	-	N/A
N112-T116	64.6	53.5	74.6	74.5	66.7	A221-R227	34.3	22.7	27.6	14.2	24.7
N112-Q137	21.4	15.8	16.5	19.5	18.3	N224-A226	88.9	91.8	94.6	91	91.6
N112-L223	95.6	95.3	90.4	93.7	93.7	N224-R227	44.1	36.5	39.1	39	39.7
T116-Q137	98.8	98.7	99.3	98.7	98.9	N224-S228	-	26.6	20.0	7.4	N/A
F158-R227	98	95.3	98	99.7	97.8						

Table A.7. Occupancies of H-bonds of Aqy1 of the R227-N112 H-bond clusters sampled during MD simulations in a mixed bilayer, with H44 and H194 Nε2-protonated. When the H-bond is sampled in all four Aqy1 monomers A-D, the total average occupancy computed by averaging the monomer occupancies is also reported. For each H-bond the occupancy calculated as an average from the last 100ns of each simulation is reported. Adapted from ref. [226].

	Occupancy (%)										
	A	B	C	D	Average		A	B	C	D	Average
E51-T55	99.7	99.6	99.2	99.7	99.6	N160-T219	10.3	6.8	11.9	-	N/A
E51-S107	100	100	100	99.9	100	N160-G220	17.9	22.9	31.9	-	N/A
E51-G109	19.2	11.9	12.8	20.4	16.1	N160-A221	19.4	33.4	39.0	-	N/A
E51-N110	99.4	100	99.9	98.4	99.4	N160-R227	13	10.2	5.6	10.9	9.9
E51-L111	100	100	100	99.9	99.9	N160-S228	59.9	35.1	74.6	87.7	64.3
E51-Q137	99.9	99.9	99.9	99.9	99.9	N160-Y245	-	-	-	6.1	N/A
S107-G109	55.6	50.2	48.8	61.9	54.1	E175-T179	99.7	99.7	99.8	99.7	99.7
S107-N110	47.1	51.3	58.6	26.4	45.86	E175-T219	100	100	100	100	100
N110-T116	8.9	-	6.2	17.4	N/A	E175-G222	99.9	100	99.9	99.9	99.9
L111-N224	65.6	68.9	60.5	44.6	59.9	E175-L223	100	100	100	100	100
N112-A114	99.6	99.5	99.6	99.7	99.6	T179-C183	64	65.5	65	67.2	65.4
N112-V115	6	10.9	6.4	6.8	7.5	T219-A221	84.9	74.6	91.7	89.91	85.3
N112-T116	65.3	60.0	64.3	52.4	60.5	T219-Y245	99.6	99.8	99.6	99.9	99.7
N112-Q137	18	25.2	19.4	23.5	21.5	G220-R227	-	8.1	-	-	N/A
N112-L223	93.1	95.5	93.9	93.2	93.9	A221-R227	22.8	41.2	16	51.1	32.8
T116-Q137	98.2	99.2	98.9	99.1	98.8	N224-A226	94.4	93.7	91.4	95.7	93.8
F158-R227	95.8	96.9	99.9	99.4	98.0	N224-R227	31.4	28.6	41	29	32.5
						N224-S228	17.2	8.3	-	-	N/A

Table A.8. Tracking H-bond motifs detected in the crystal structure of Aqy1 (PDB ID:3ZOJ [130]) during MD simulations in a POPE bilayer with H44-H194 Nε2-protonated, vs. Nδ1-protonated. Occurrence rates of the detected motifs are reported for each monomer separately. Adapted from ref. [226].

H-bond motif	H44/H194 Nε2 protonated				H44/H194 Nδ1 protonated			
	A	B	C	D	A	B	C	D
Intra-helical hydroxyl - backbone carbonyl H-bonds								
G35-S38	—	—	—	—	—	35.2	—	—
F45-S49	84.9	80.2	96.4	97.9	96.9	97.4	85.2	48
I46-S49	12.1	29.3	11.1	11.2	9.8	11.7	28.2	—
E51-T55	99.6	99.6	99.5	99.7	99.5	99.6	99.7	99.6
F58-S61	—	—	—	—	—	—	—	—
L85-S89	98.9	99.9	99.7	99.9	98.9	99.9	99.7	99.9
G99-T103	99.9	99.8	99.8	99.8	100	99.9	99.9	100
L132-T136	95.6	99.5	98	98.6	99.1	98.2	98.7	96.7
A148-T152	99.7	99.4	99.6	99.3	99.4	98.4	99.5	99.5
E175-T179	99.8	99.8	99.6	99.6	99.7	99.6	99.6	99.7
I181-T185	99.4	99.6	98.3	99.4	95.2	99.6	98.9	99.2
E192-T197	—	42.8	38.6	44.1	35.9	30	63.6	61.3
F254-S258	99.5	99.6	97	99.6	97.6	84.2	99.3	99.2
Inter-helical carboxylate - hydroxyl								
E51-S107	100	99.9	100	99.9	100	100	99.9	100
E175-T219	100	100	100	100	100	100	12.8	100
Arg sidechain - backbone carbonyl								
R33-R105	62.7	13.5	45.6	92.1	96.1	84.9	34.3	11.1
F158-R227 (Nη1)	98	94.6	97.3	99.7	99	93.5	99.7	99.9
F158-R227 (Nη2)	58.5	40	38.6	52.4	61.1	28.5	55.4	51
R195 (O)-R195 (Nη1)	—	—	—	—	—	7.5	—	—
R227 (O)-R227 (Nη1)	91.3	62	47.3	85.4	94.4	36.9	92.5	77.1
V233-R236	—	—	—	—	—	—	—	—
A234-R236	7.6	86.2	62.5	36.4	—	39.6	29.7	78
Asn carboxamide - backbone carbonyl								
N43-G37	—	—	—	—	—	—	—	—
N112-L223	95.6	95.3	90.4	93.7	94.2	95.3	97.1	94.7
N160-A221	29.5	22.6	29.1	45.6	34.7	17.2	65.4	49.6
Ser/Thr hydroxyl - Asn carboxamide								
N110 (Oδ1)-T116	5.2	—	—	—	11.9	—	—	—
N160 (Oδ1)-S228	66.3	10.2	20.7	55.8	71.4	5.5	77.8	49.7
Asp/Glu carboxyl - backbone amide								
E18 (Oε1)-E18 (N)	—	—	—	—	—	—	—	—
E51 (Oε2)-N110	57.3	99.8	99.6	—	63.6	98.8		45.4
E51 (Oε1)-L111	100	100	100	99.5	31	100	100	100
E175 (Oε1)-G222	12.4	17.9	26.6	100	11	15.6	—	36.2
E175 (Oε1)-L223	100	100	100	100	100	99.9	99.9	99.9
E175 (Oε2)-G222	99.9	99.9	99.9	99.9	99.9	99.9	99.2	99.9

Table A.9. Tracking H-bond motifs detected in the crystal structure of Aqy1 (PDB ID:3ZOJ [130]) during MD simulations in a mixed bilayer with H44-H194 Nε2-protonated, vs. Nδ1-protonated. Occurrence rates of the detected motifs are reported for each monomer separately. Adapted from ref. [226].

H-bond	H44/H194 Nε2 protonated				all Nδ1			
	A	B	C	D	A	B	C	D
Ser/Thr hydroxyl - backbone carbonyl								
G35-S38	45.4	—	—	6	—	23.1	70	28.3
F45-S49	84.5	86.1	92.1	97	92.7	54.6	98.9	44
I46-S49	14.6	23.2	18.2	8.5	20.3	—	9.2	31.5
E51-T55	99.7	99.6	99.2	99.7	99.8	99.7	99.7	99
F58-S61	—	—	—	—	—	—	—	—
L85-S89	99.8	96.1	99.9	99.8	99.9	99.9	99.9	100
G99-T103	99.7	99.8	99.7	99.8	99.8	99.9	99.9	99.8
L132-T136	98.2	93.9	99.4	98.9	98.7	99.4	96.9	96.2
A148-T152	99.8	98.5	95.9	99.8	97.1	99	98.2	98.6
E175-T179	99.7	99.7	99.8	99.6	99.6	99.7	99.2	99.7
I181-T185	98.4	99.6	99.6	99.5	96.7	99.6	97.7	99.4
E192-T197	48.1	52.9	24.8	30.2	73.9	28.6	62.3	7.5
F254-S258	99.6	89.8	99.3	99.6	99.6	95.9	99.4	99.7
Asp/Glu - Ser/Thr hydroxyl								
E51-S107	100	100	100	99.9	99.9	100	100	99.9
E175-T219	100	100	100	100	100	100	100	100
Arg guanidinium - backbone carbonyl								
R33-R105	83.1	27	84.5	68.3	90.6	48.5	91.5	53.9
F158-R227 (Nη1)	95.6	96.4	99.9	99.2	99.9	99.7	95.7	99.8
F158-R227 (Nη2)	38	62.1	55.7	80	56	58	37	60
R195 (O)-R195 (Nη1)	—	—	—	—	—	—	—	—
R227 (O)-R227 (Nη1)	57.7	81.6	93	92.1	87.2	85.2	57.3	93.7
V233-R236	—	16.7	—	—	—	—	—	—
A234-R236	83.1	—	49.5	95.9	71.8	33.2	—	78.4
Asn carboxamide - backbone carbonyl								
N43-G37	—	—	—	—	—	—	—	—
N112-L223	93.1	95.5	93.9	93.2	94.1	95.8	96.4	93
N160-A221	19.4	33.4	39	—	45.1	35.9	27.8	47
Ser/Thr hydroxyl - Asn carboxamide								
N110-T116	8.7	—	—	—	82.7	—	—	56.5
N160-S228	29.6	—	74.5	—	52.6	60.1	39.6	78.5
Asp/Glu carboxyl - backbone amide								
E18 (Oε1)-E18 (N)	—	—	—	—	—	—	—	—
E51 (Oε2)-N110	98.6	99.4	99.2	96.4	63.1	99.4	98.5	97.1
E51 (Oε1)-L111	100	100	100	99.9	100	100	100	99.9
E175 (Oε1)-G222	13.7	37	19.5	10.3	35.4	20.3	20.6	22.2
E175 (Oε1)-L223	100	100	100	100	100	100	100	100
E175 (Oε2)-G222	99.9	99.9	99.9	99.9	100	99.8	99.9	99.9

Table A.10. Tracking of H-bond motifs detected in the crystal structure of ChR2 (PDB ID:6EID [337]) during two MD simulations. For the crystal structure, the distances measured for chain A are reported. Adapted from ref. [226].

*) The H-bond motif between S63-N258 involves Oδ1 atom in the crystal structure and Nε2 atom in the MD simulations.

H-bond motif	All-*trans*		13-*cis*,15-*anti*		Crystal structure (Å)	H-bond motif	All-*trans*		13-*cis*,15-*anti*		Crystal structure (Å)
	A	B	A	B	A		A	B	A	B	A
Ser/Thr hydroxyl - backbone carbonyl						*Arg guanidinium - backbone carbonyl*					
N46-T50	58	—	—	28.6	2.8	W39-R115	17.8	6	—	—	3.5
A59-S63	98	97.3	62.6	88.1	2.7	E41-R115	44.5	45.8	—	—	2.8
Y70-T74	98.1	74.2	91.4	97.7	3.2	S42-R115	84.5	44.3	—	—	3.4
E101-S106	7.6	—	—	15.2	3.2	N137-R268	98.1	96.6	—	69.9	2.7/3.5
K103-S106	—	—	—	—	3.1	Y200-R209	30	24.6	29.5	—	2.8
E123-T127	—	53.5	96.2	—	3	H201-R209	10.9	16.1	13.7	—	3
L132-S136	83.7	46.4	96.2	68.4	3.2	V203-R209 (Nε)	65.1	84.8	70.3	—	3
S146-T149	—	16	—	5.4	3.5	V203-R209 (Nη2)	76.2	88.4	84.6	—	3
G151-S155	73.2	99.9	97.3	99.7	3.1	*Asp/Glu carboxyl - Arg guanidinium*					
L152-S155	—	—	—	—	2.9	E82 (Oε1) – R268 (Nη1)	91.9	96.2	26.3	45.8	2.9
S155-T159	82.7	98.8	—	98.6	2.6	E82 (Oε1) – R268 (Nη2)	49.2	35.2	97.1	51.2	3.5
V161-T165	94.8	99.8	99.2	98.4	2.8	E83 (Oε1) – R268 (Nη1)	99.9	64.6	74.3	71	3.4
Y184-T188	32.5	80.8	—	82.6	3.4	E83 (Oε2) – R268 (Nη1)	21.2	76	74.4	70.8	3.1
E198-T202	76.7	27.8	59.8	35.8	2.8	E198 (Oε1) - R148 (Nη1)	66.4	60.7	66.6	58.4	2.6
R209-T213	46.2	45.8	70.4	9.3	3.1	E198 (Oε1) - R148 (Nη2)	73.8	71	74	56.6	3.1
L218-S222	99.6	99.6	99.4	99.7	2.6	E235 (Oε2) – R120 (Nη1)	—	—	—	—	3.3
F219-S222	—	—	—	—	3.1	*Asn carboxamide - backbone carbonyl*					
V242-T246	—	62.7	26.5	59.9	2.8	N137-S142	68.9	15	83.8	32.5	2.7
T246-T250	46.6	67.4	99.8	6.7	2.7	*Ser/Thr hydroxyl - Asn carboxamide*					
I252-S256	96.2	99.3	89.3	99.8	2.7	S63-N258*	86.5	95.8	55.8	39.5	3.3
D253-S256	16.8	—	—	—	3.4	N137-S142	—	—	—	—	3.1
Asp/Glu carboxyl – Ser/Thr sidechains						S155-N187	39.8	94.8	—	93.7	4.1
E101-T246	—	—	—	—	2.7						
E123 (Oε1)-T127	99.6	—	—	55	5.2						
E123 (Oε2)-T127				40.5	2.9						
Combined Asp/Glu carboxyl-Ser/Thr hydroxyl-backbone carbonyl											
E101-T246	—	—	—	—	2.7						
V242-T246	—	—	—	—	2.8						

Supplementary Figures

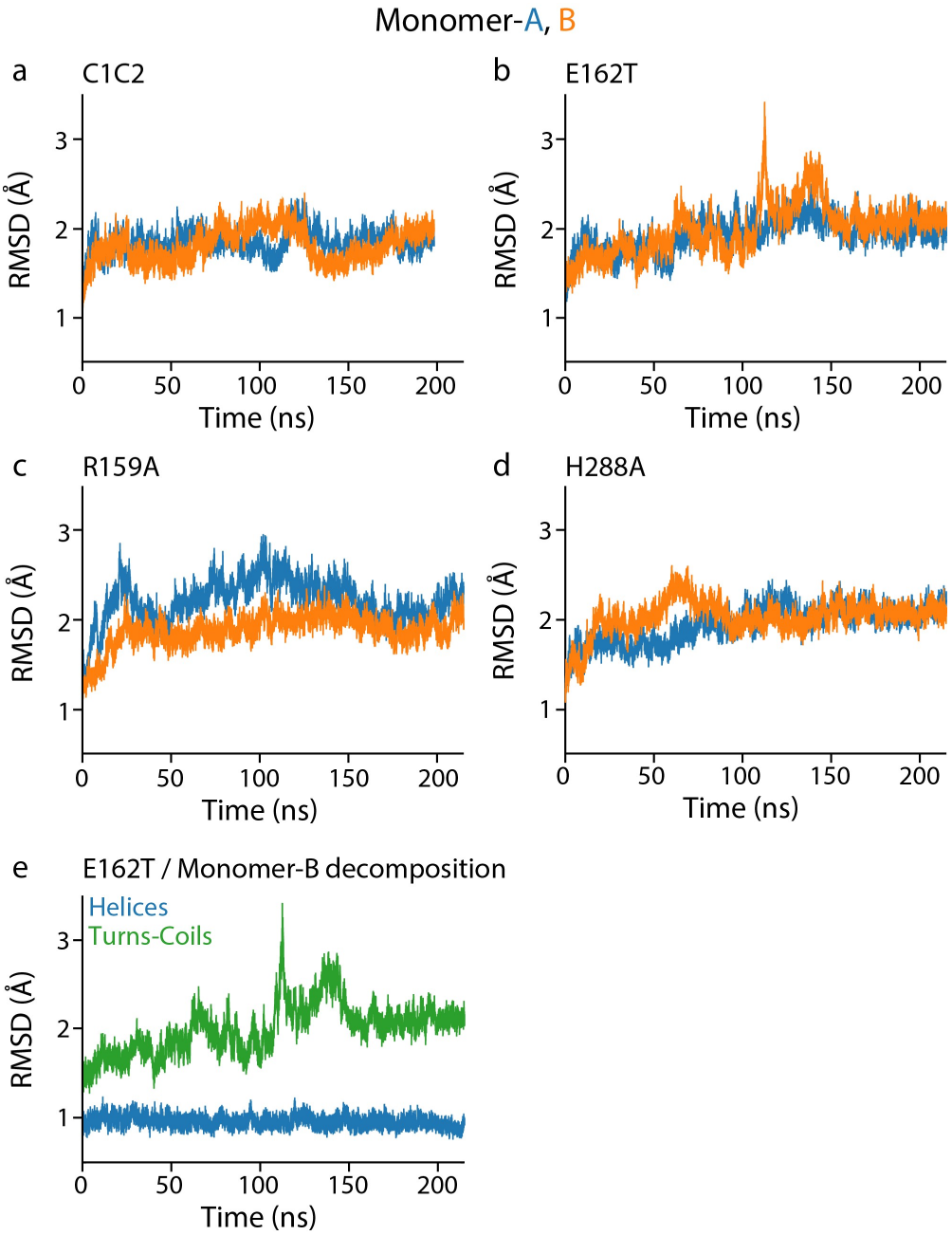

Figure A.1. Cα - RSMD profiles for the Channelrhodopsin C1C2 chimera simulations. Profile timeseries are colored in blue and orange for Monomer-A and Monomer-B respectively. (a-e) Cα - RSMD profiles for the wild-type chimera C1C2 (a), E162T mutant (b) R159A mutant (c) and H288A mutant (d). (e) RMSD profile decomposition for the E162T mutant, where the profile is split in the transmembrane region (blue) and the loops-helices (green). Adapted from ref. [42].

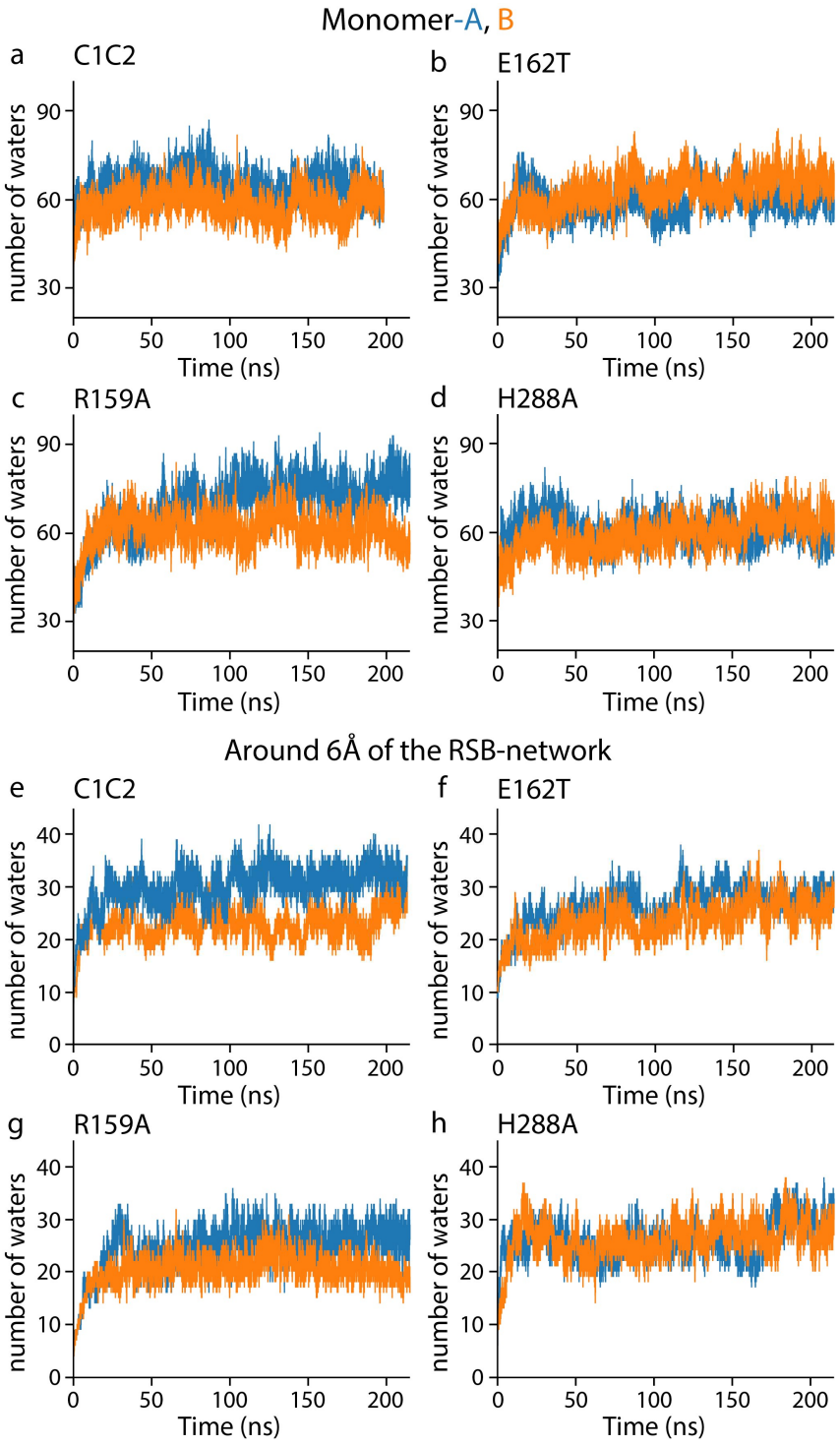

Figure A.2. Internal water molecules in simulations of C1C2 chimera. Internal water molecules profiles are colored in blue and orange for Monomer-A and Monomer-B respectively. (a-d) Internal water molecules profiles for the wild-type C1C2 (a), E162T mutant (b) R159A mutant (c) and H288A mutant (d). Number of water molecules of the extended RSB network. For this computation V156 is considered part of the network to accommodate for the Arg to Ala mutation in the position 159. V156 is located within 1 helical turn from R159. (e-h) Extended RSB-network water molecules profiles for the wild-type C1C2 (e), E162T mutant (f) R159A mutant (g) and H288A mutant (h). Adapted from ref. [42].

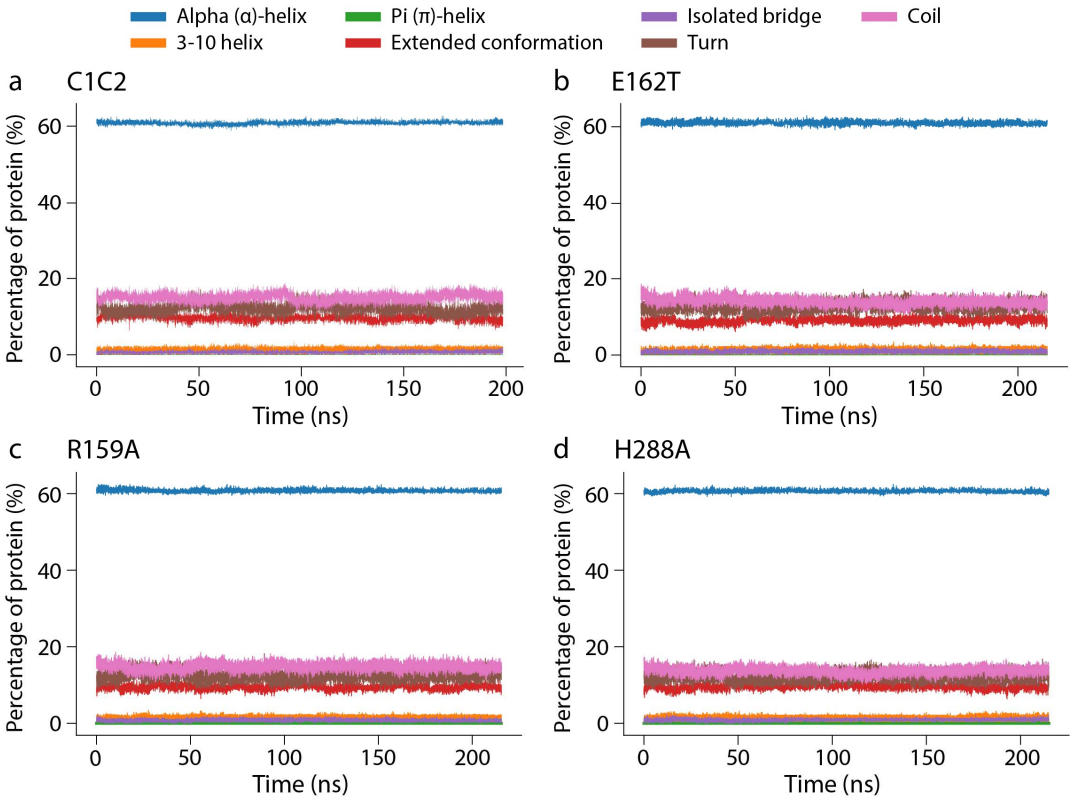

Figure A.3. Time-dependent STRIDE analysis for the Channelrhodopsin C1C2 chimera simulations. (a-d) Secondary structure analyses for the wild-type C1C2 (a), E162T mutant (b) R159A mutant (c) and H288A mutant (d). In the index the different types of secondary structures are noted. Namely, the α-helices shown in blue, the 3-10 helix in orange π-helix in green, extended conformation in red, isolated bridge in purples, turns in brown and coils in pink. Adapted from ref. [42].

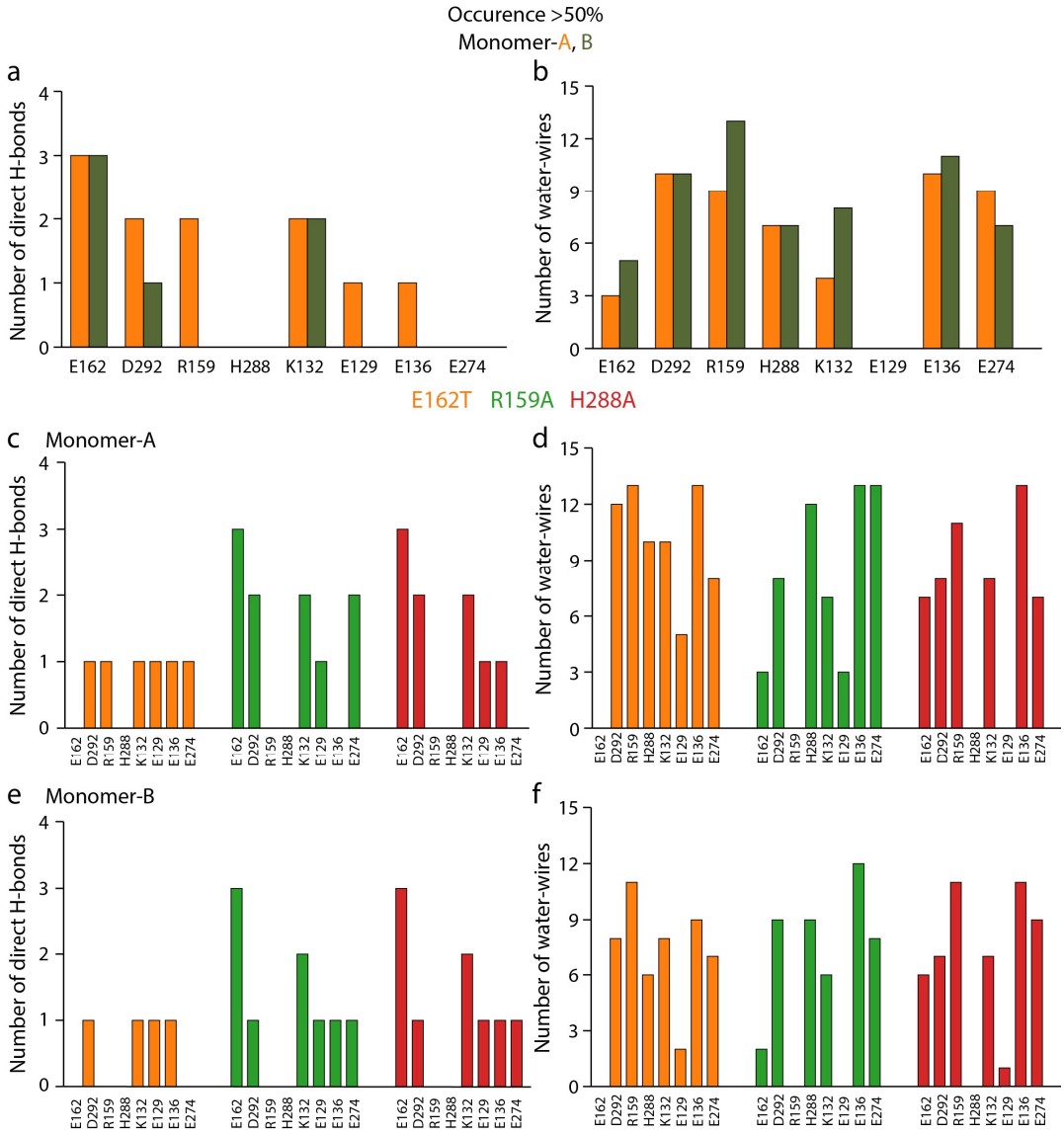

Figure A.4. High occurrence H-bonds in the wild-type and mutants of C1C2. (a, b) High occurrence direct (a) and water-mediated (b) H-bonds of selected amino acid residues in wild-type C1C2. Monomer-A is shown in orange and Monomer-B in tan. (d-e) High occurrence direct (c, e) and water-mediated (d, f) H-bonds of selected amino acid residues in mutants of C1C2 for Monomer-A (c, d) and B (e, f), respectively. E162T mutant is shown in orange, R159A mutant in green and H288A mutant in red. Adapted from ref. [42].

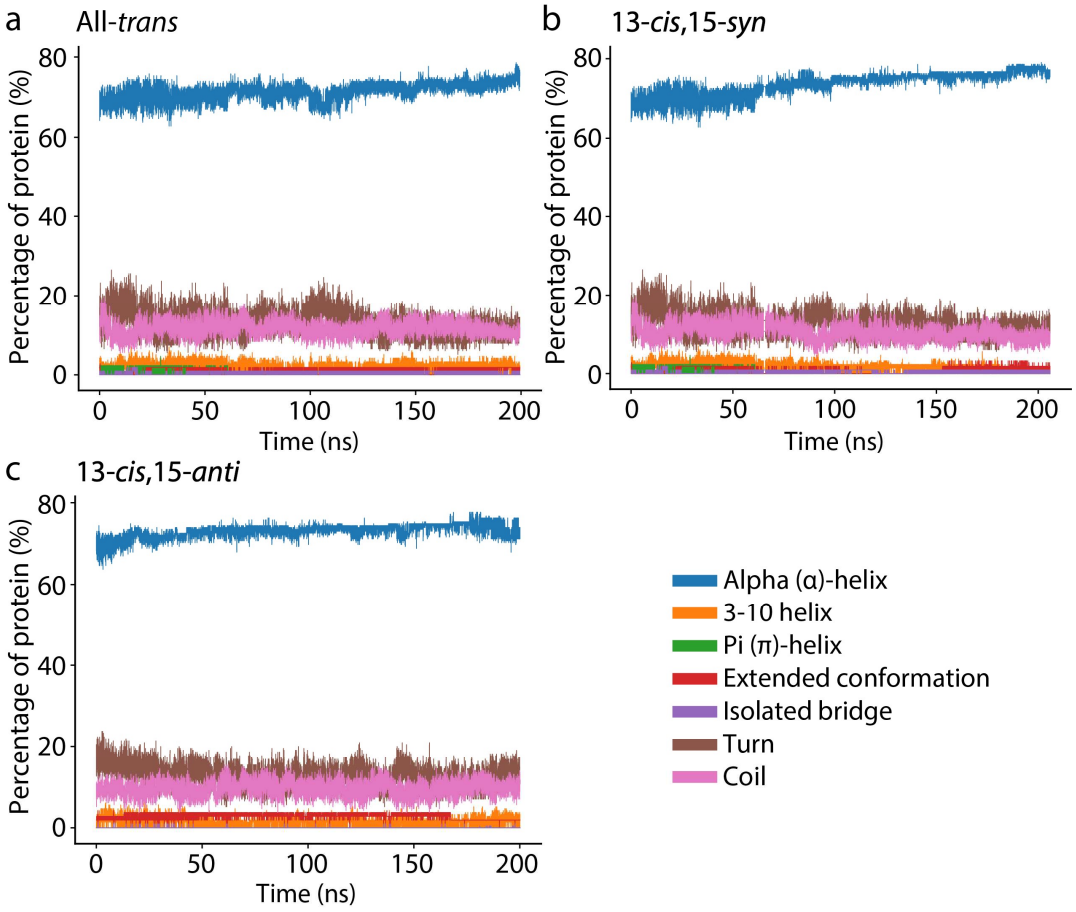

Figure A.5. Time-dependent STRIDE analysis for the homology model of AntR. (a-c) Secondary structure analyses for the all-*trans* (a), 13-*cis*,15-*syn* (b) and 13-*cis*,15-*anti* (c) models of AntR. In the index the different types of secondary structures are noted. Namely, the α-helices shown in blue, the 3-10 helix in orange π-helix in green, extended conformation in red, isolated bridge in purples, turns in brown and coils in pink. Adapted from ref. [185].

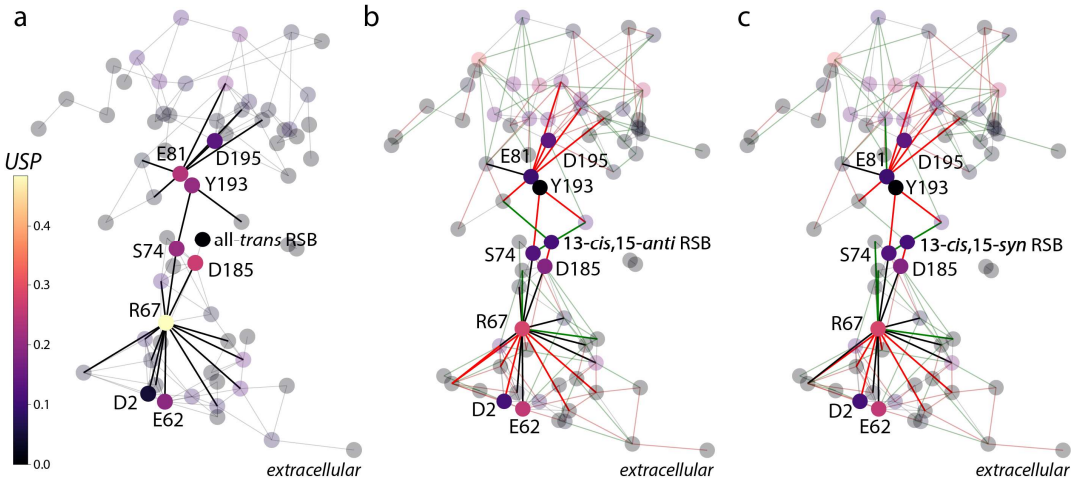

Figure A.6. Unique shortest paths computations in AntR simulations. (a, b, c) USP computations on the all-*trans* (a), 13-*cis*,15-*anti* (b) and 13-*cis*,15-*syn* (c). USP values were computed on water-mediated H-bond graphs of AntR that were pre-filtered to a minimum of 35% occurrence rate. Nodes are colored according to the normalized USP values. The amino acid residues of interest, and their connections are labeled and shown in full opacity. The remaining nodes and edges are shown in transparency. Panels b and c show comparative graphs using the all-*trans* graph as the foundation. The nodes shown in those panels are colored with the USP values of their respective system. The edges are colored according to the function principles described in the section Comparative H-bond graphs. Adapted from ref. [185].

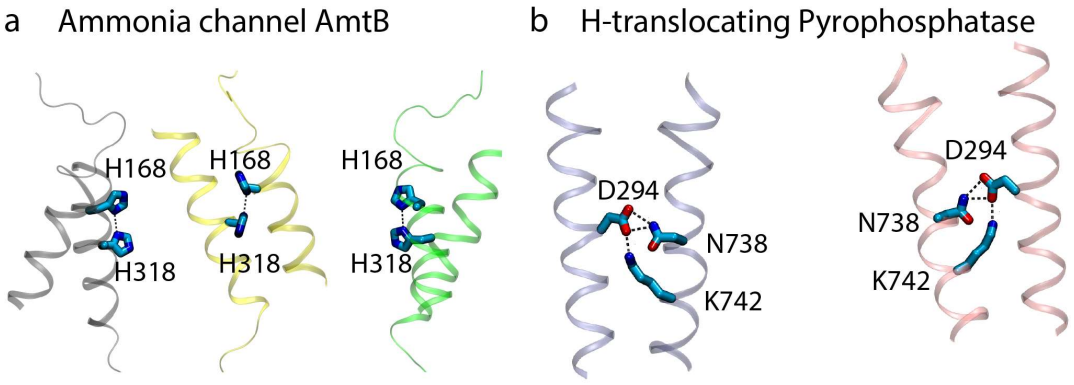

Figure A.7. H-bond motifs as part of a protein's biological function. (A) His-His H-bond motifs found in the Ammonia channel AmtB. (B) Asp-Asn detected in the proton translocating pyrophosphatase. In both cases the motifs are detected in one occurrence per protein monomer. Both motif examples are between two different TM-helices and resemble a gating structure. Every protein monomer is shown in a differently colored ribbon representation. Adapted from ref. [226].

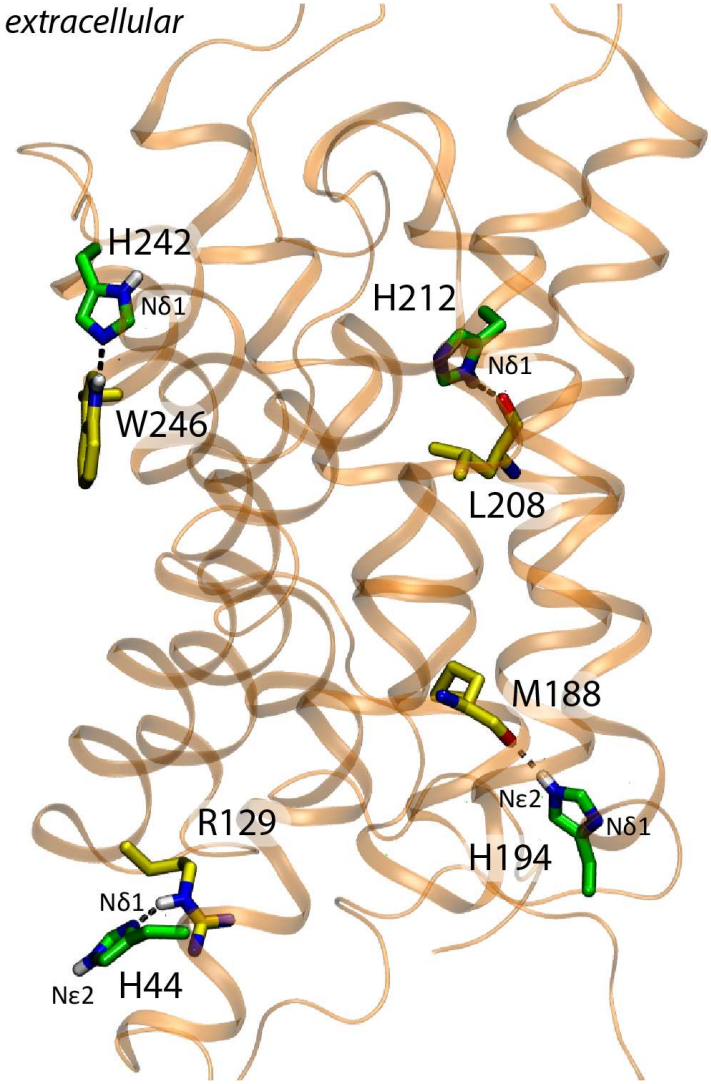

Figure A.8. Location of His amino acid residues in the high-resolution crystal structure of Aqy1. Amino acid residues H44 and H194 are Nε2-protonated, while H212 and H242 are Nδ1-protonated. Molecular graphics are based on the crystal structure of Aqy1 (PDB ID: 3ZOJ) [130]. Adapted from ref. [226].

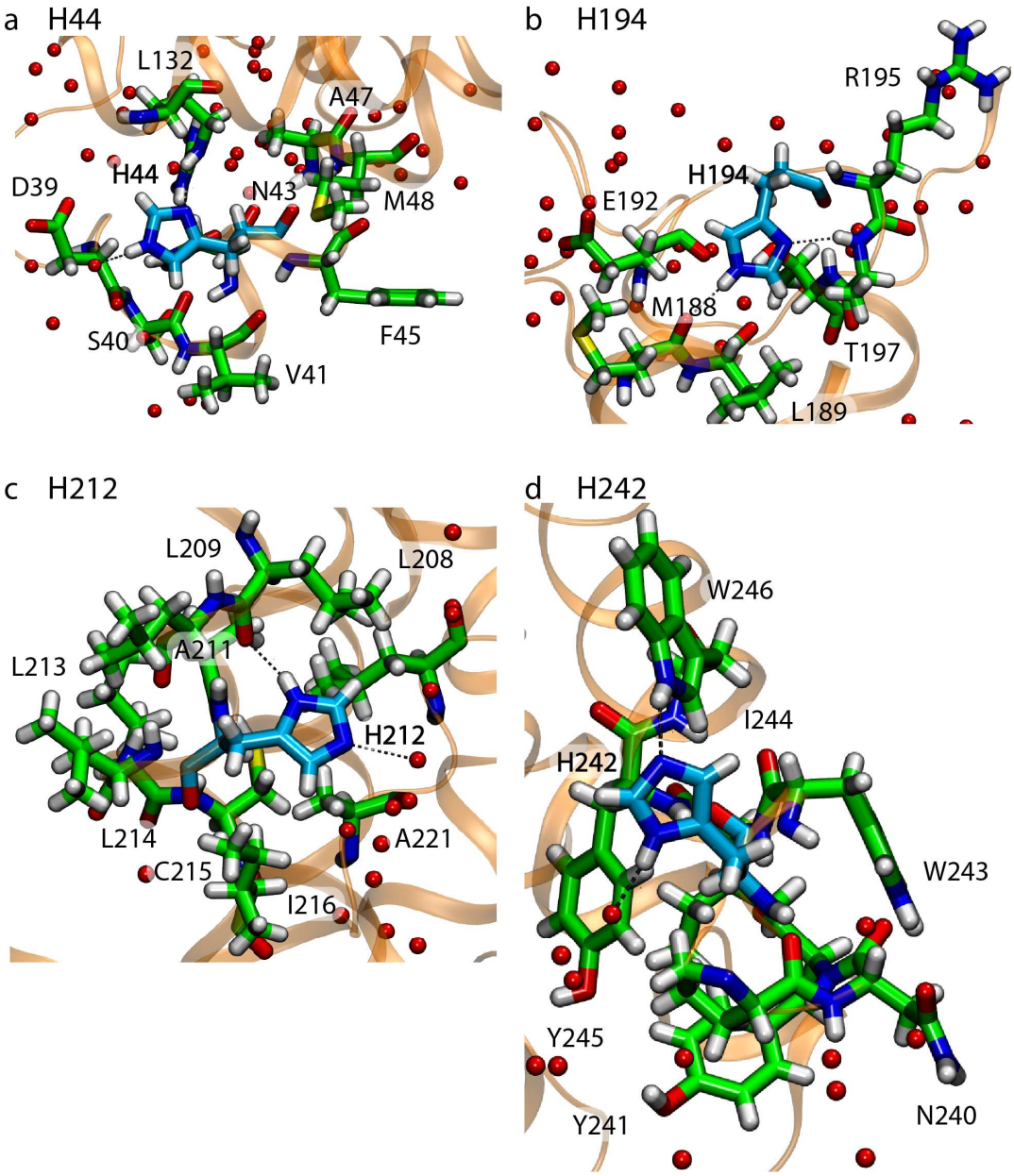

Figure A.9. Local H-bond networks of the His sidechains in the crystal structure of Aqy1. (a-d) H-bond interactions of H44 (panel a), H194 (panel b), H212 (panel c), and H242 (panel d). Water molecules are represented as red spheres. Molecular graphics are based on the crystal structure of Aqy1 (PDB ID: 3ZOJ). Adapted from ref. [226].

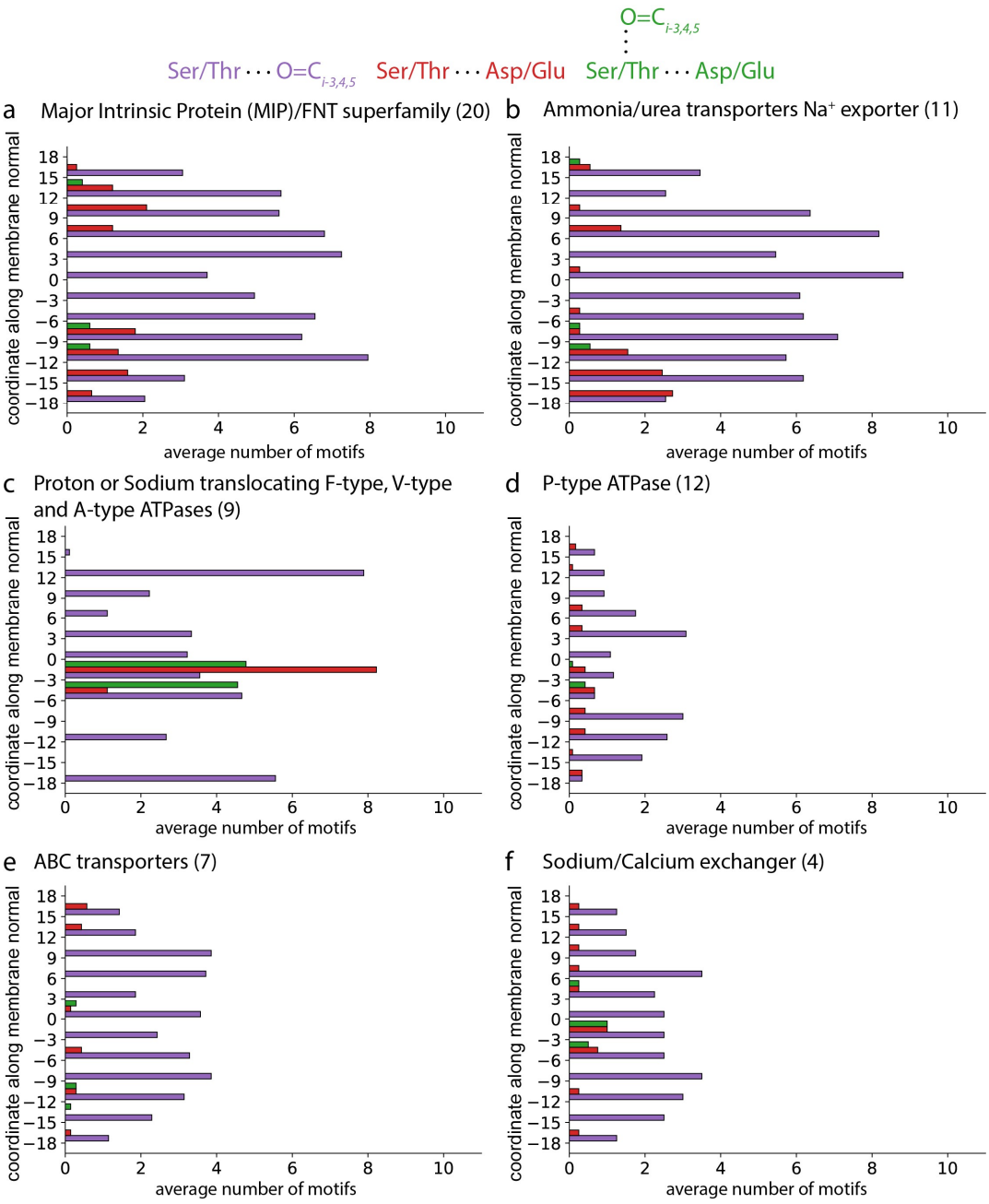

Figure A.10. Asp/Glu-Ser/Thr carboxyl-hydroxyl and Ser/Thr-backbone carbonyl of the *i*-3,4,5 relative position and combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs along the membrane normal for protein structures of superfamilies of *Set-high*. (a) Major Intrinsic Protein (MIP)/FNT superfamily. (b) Ammonia/urea transporters and Na^+ exporter. (c) Proton or Sodium translocating F-type, V-type and A-type ATPases. (d) P-type ATPase. (e) ABC transporters. (f) Sodium/calcium exchanger. Adapted from ref. [226].

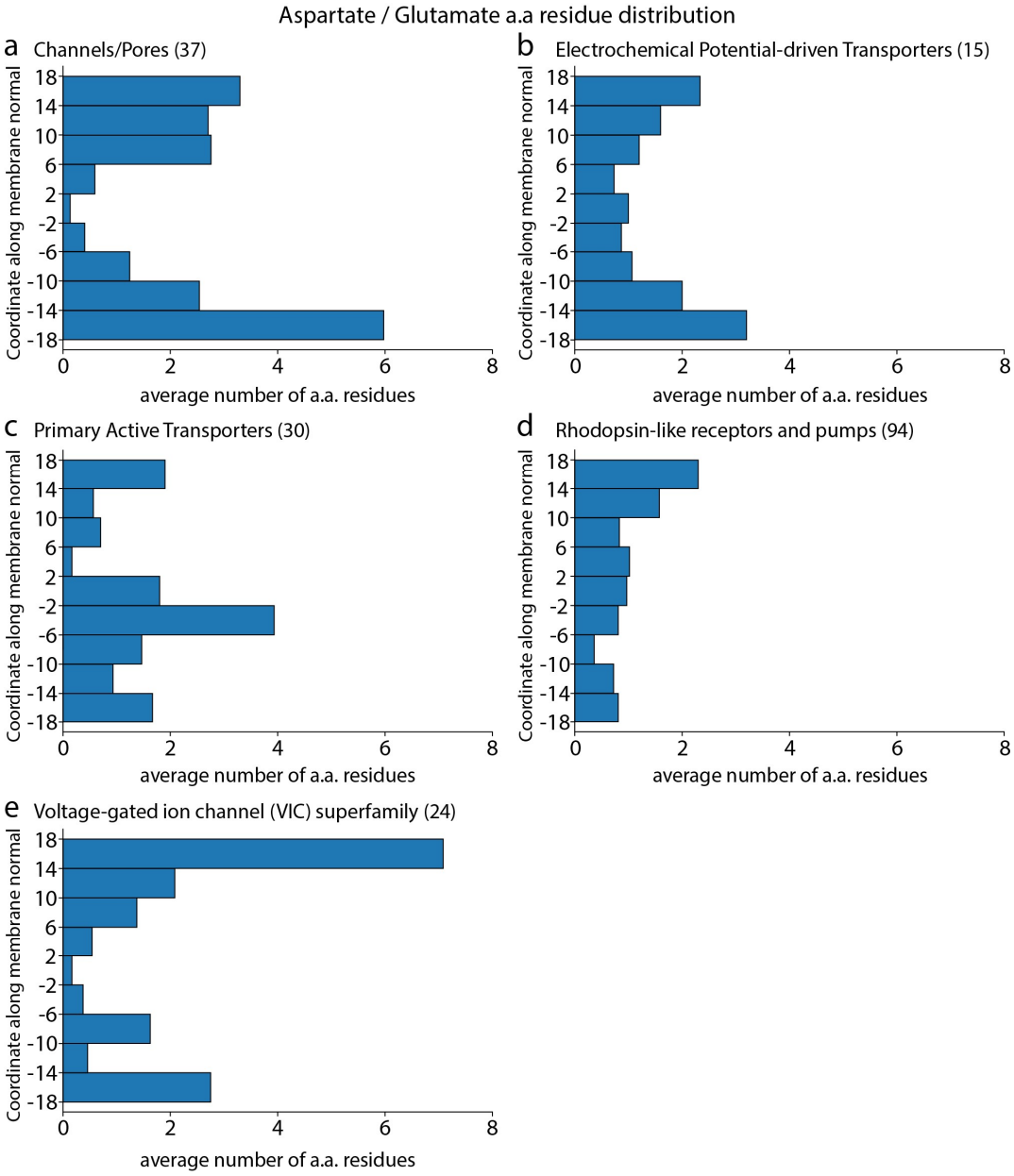

Figure A.11. Dissected amino acid (a.a) residue location distributions along the membrane normal. The distributions of Aspartate/Glutamate amino acid residues are presented for members of *Set-high*. (a-e) Distribution of Aspartate/Glutamate a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e). Adapted from ref. [226].

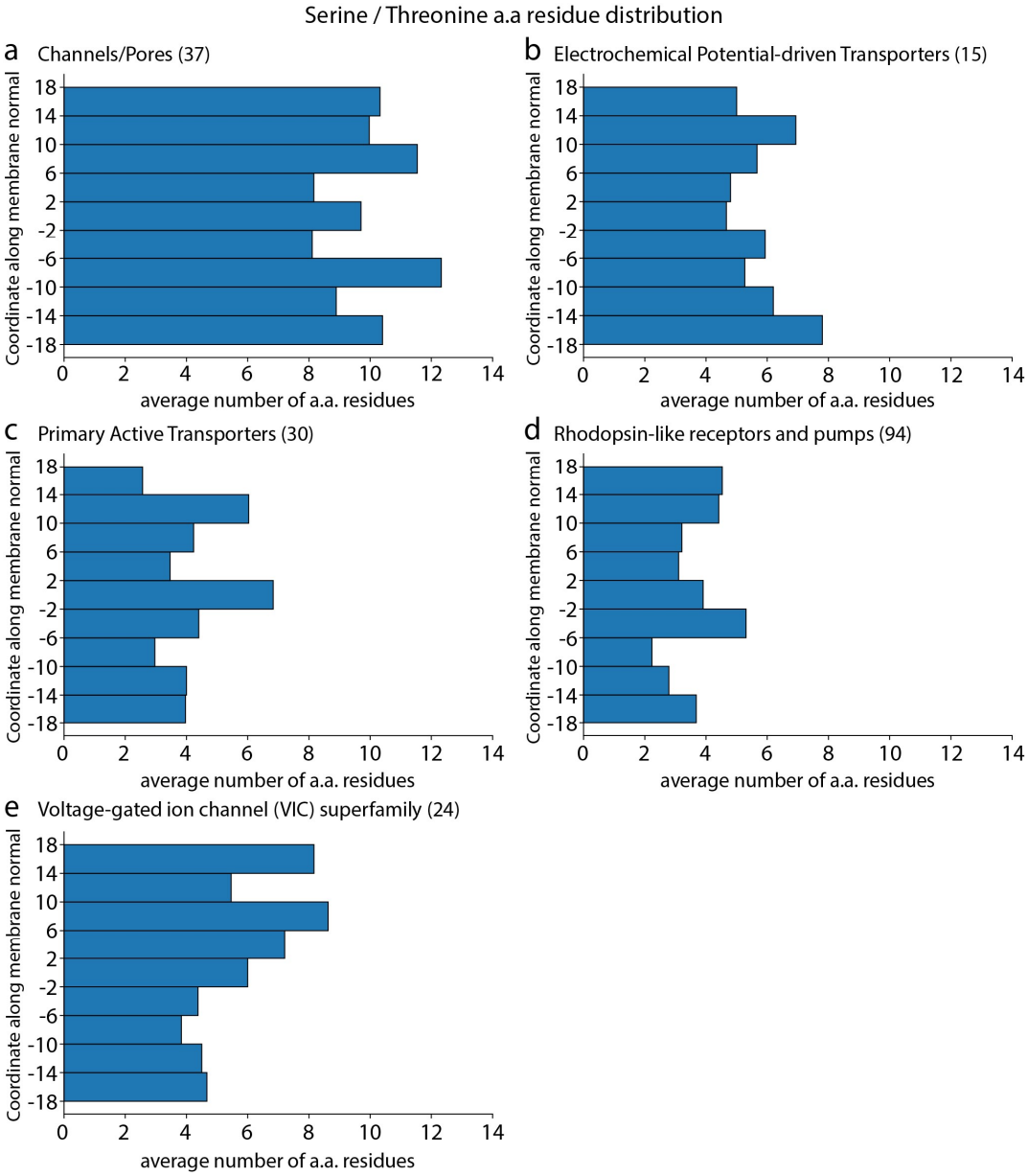

Figure A.12. Dissected amino acid (a.a) residue location distributions along the membrane normal. The distributions of Serine/Threonine amino acid residues are presented for members of *Set-high*. (a-e) Distribution of Serine/Threonine a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e). Adapted from ref. [226].

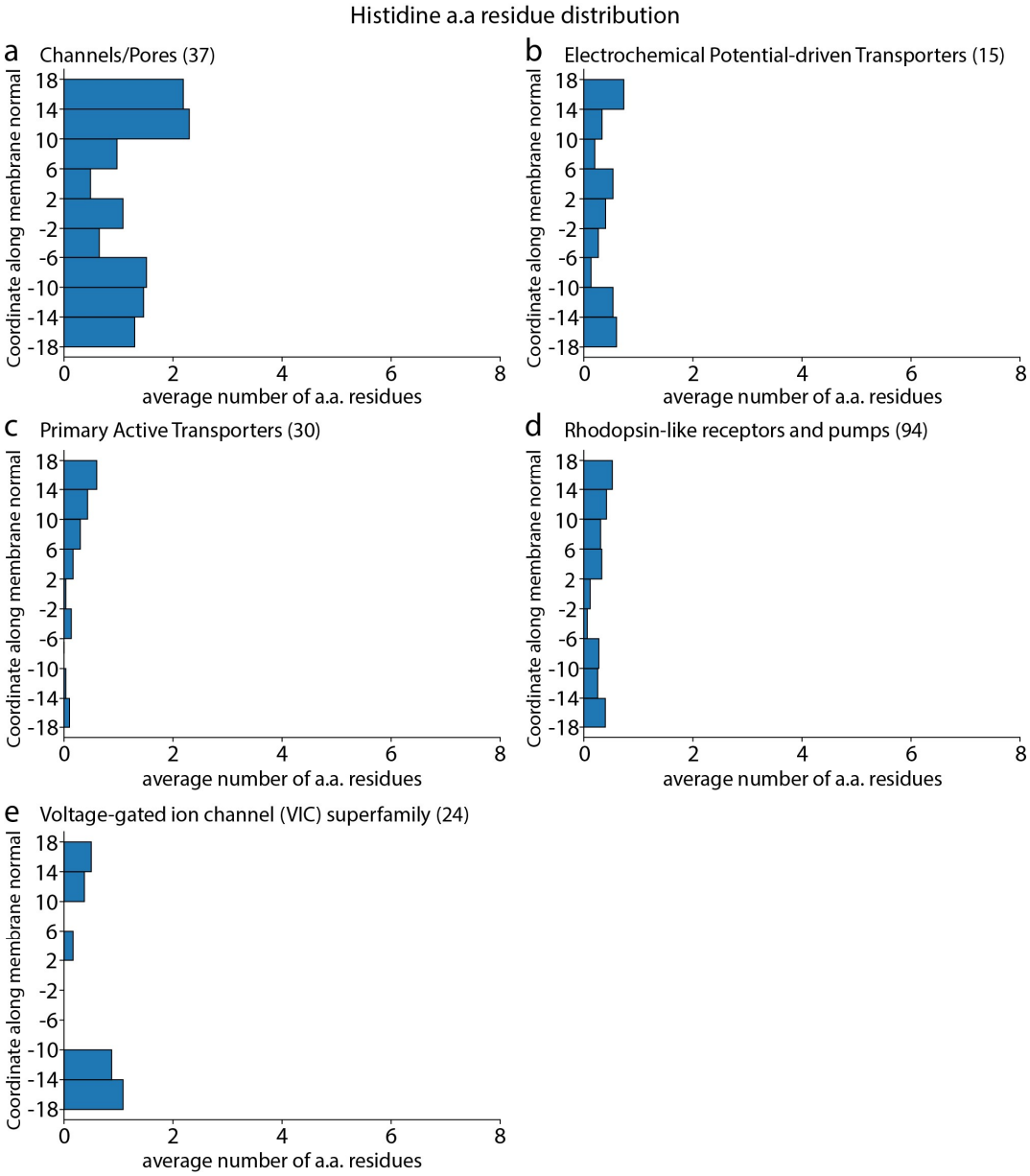

Figure A.13. Dissected amino acid (a.a) residue location distributions along the membrane normal. The distributions of Histidine amino acid residues are presented for members of *Set-high*. (a-e) Distribution of Histidine a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e). Adapted from ref. [226].

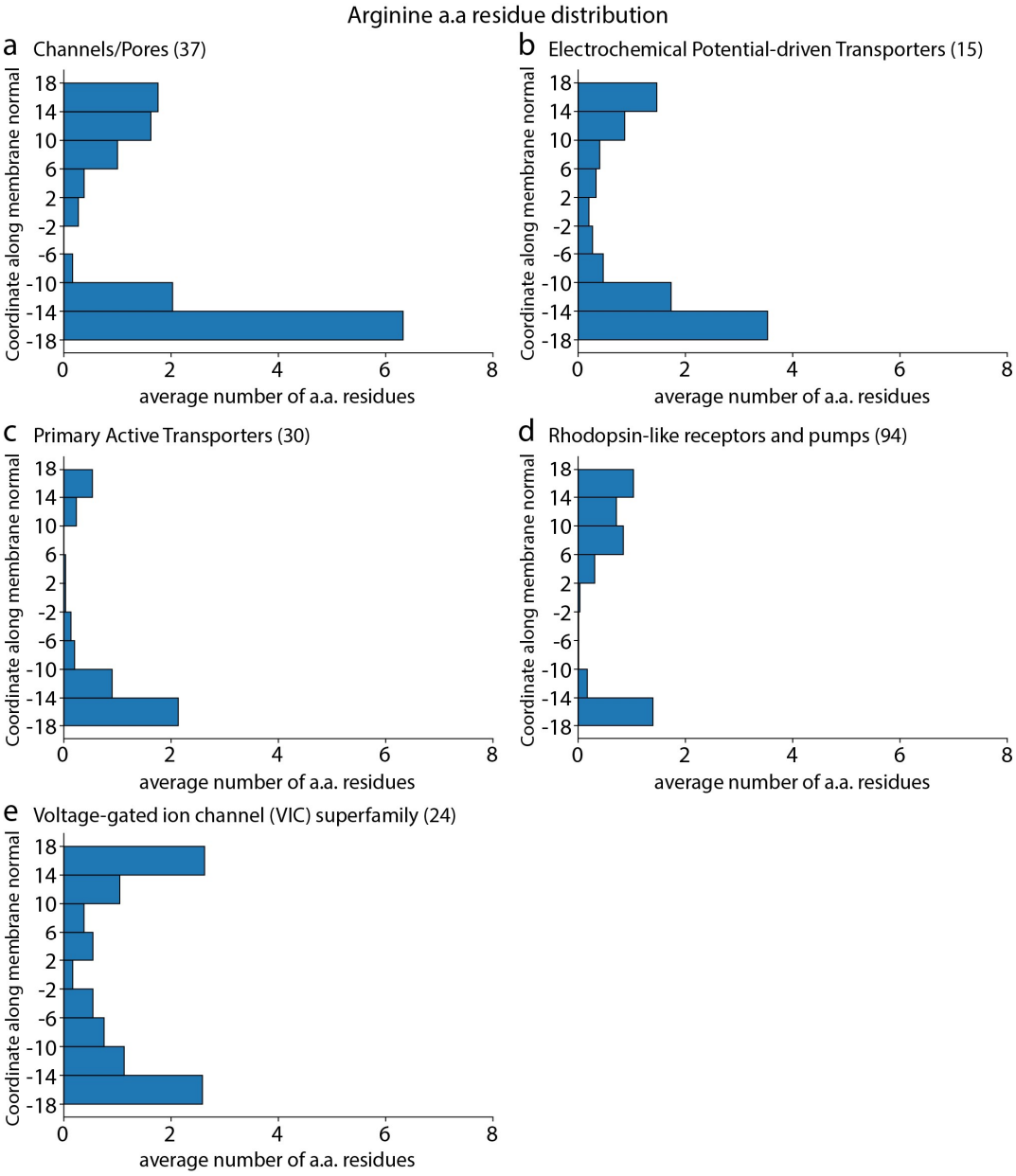

Figure A.14. Dissected amino acid (a.a) residue location distributions along the membrane normal. The distributions of Arginine amino acid residues are presented for members of *Set-high*. (a-e) Distribution of Arginine a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e). Adapted from ref. [226].

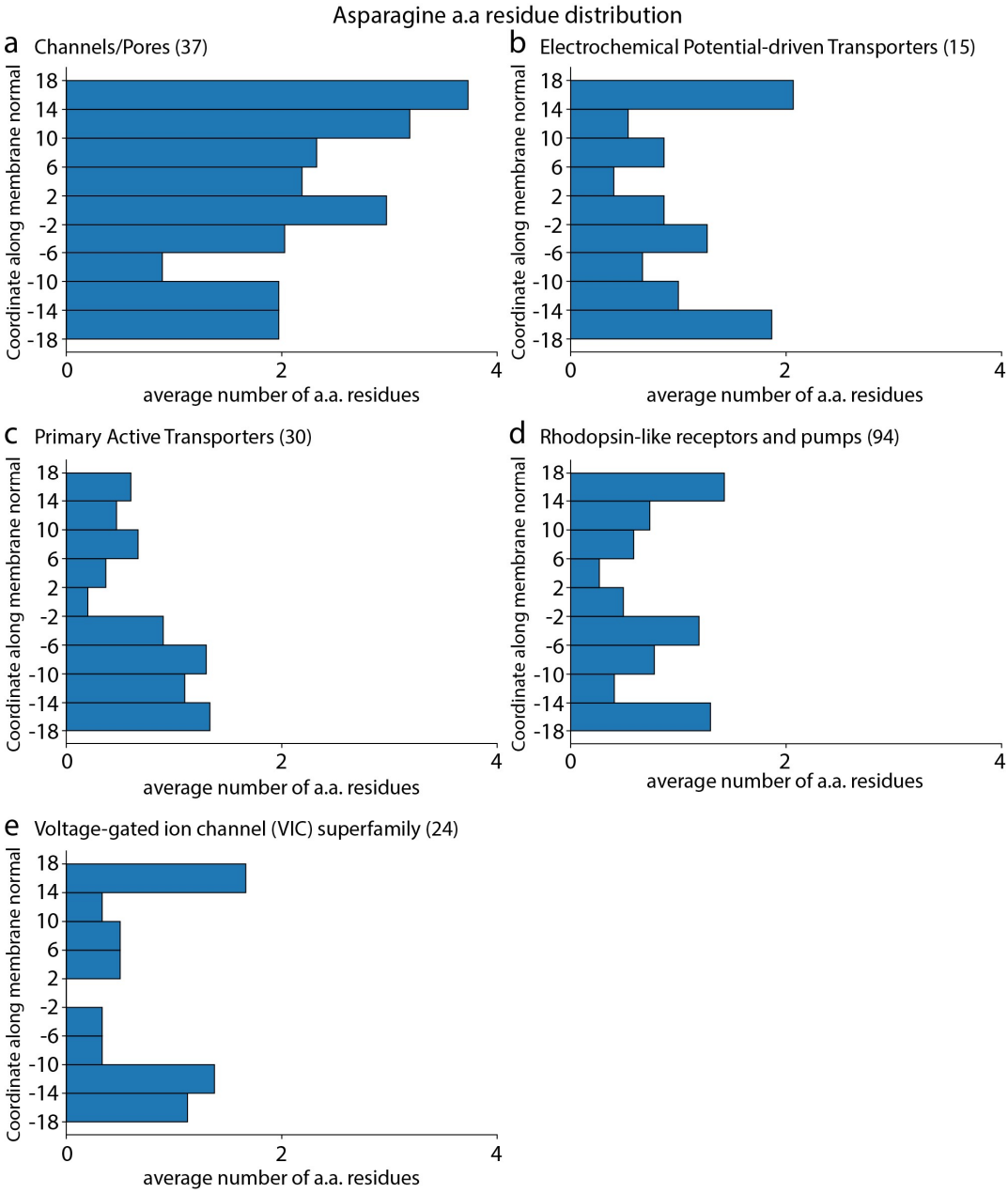

Figure A.15. Dissected amino acid (a.a) residue location distributions along the membrane normal. The distributions of Asparagine amino acid residues are presented for members of *Set-high*. (a-e) Distribution of Asparagine a.a residues in channels and pores (a), in electrochemical potential-driven transporters (b), in primary active transporters (c), in rhodopsin-like receptors and pumps (d) and in voltage-gated ion channels (e). Adapted from ref. [226].

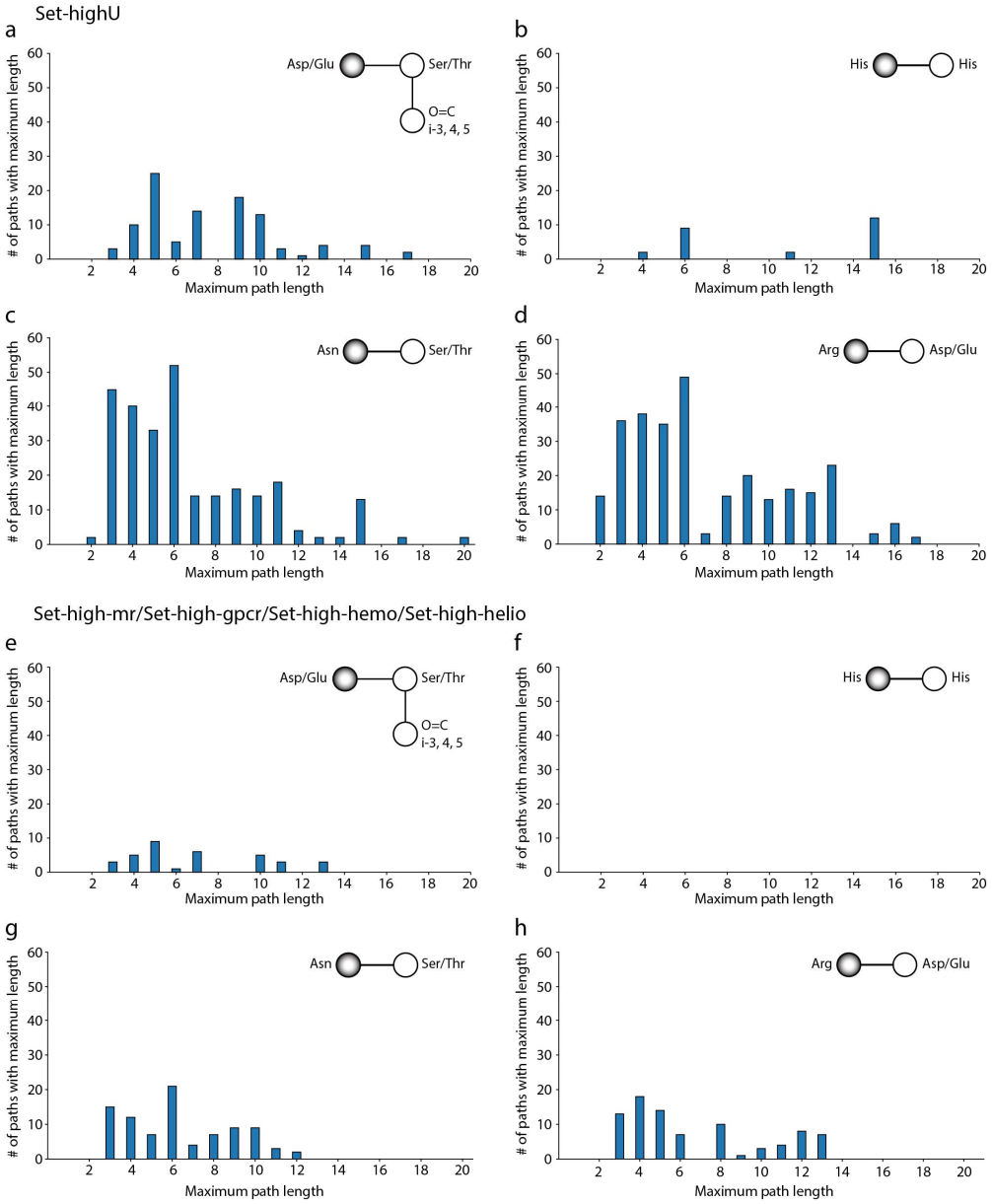

Figure A.16. Shortest path length computations for H-bond motifs in two crystal structure data sets. The root nodes are unique entries per amino acid residue that participate in H-bond motifs and are *internal* to the H-bond network (Figure 2.7, Figure 2.8). Path length distributions are shown for the carboxylate groups participating in the combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (a, d). Histidine amino acid residues participating in His-His motifs *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (b, f). Asparagine amino acid residues participating in Asn-Ser/Thr motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (c, g). Arginine amino acid residues participating in Arg-Asp/Glu motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (d, h). Additional analysis for sets *Set-high* and set *Set-low* is found in Figure 5.11. Adapted from ref. [226].

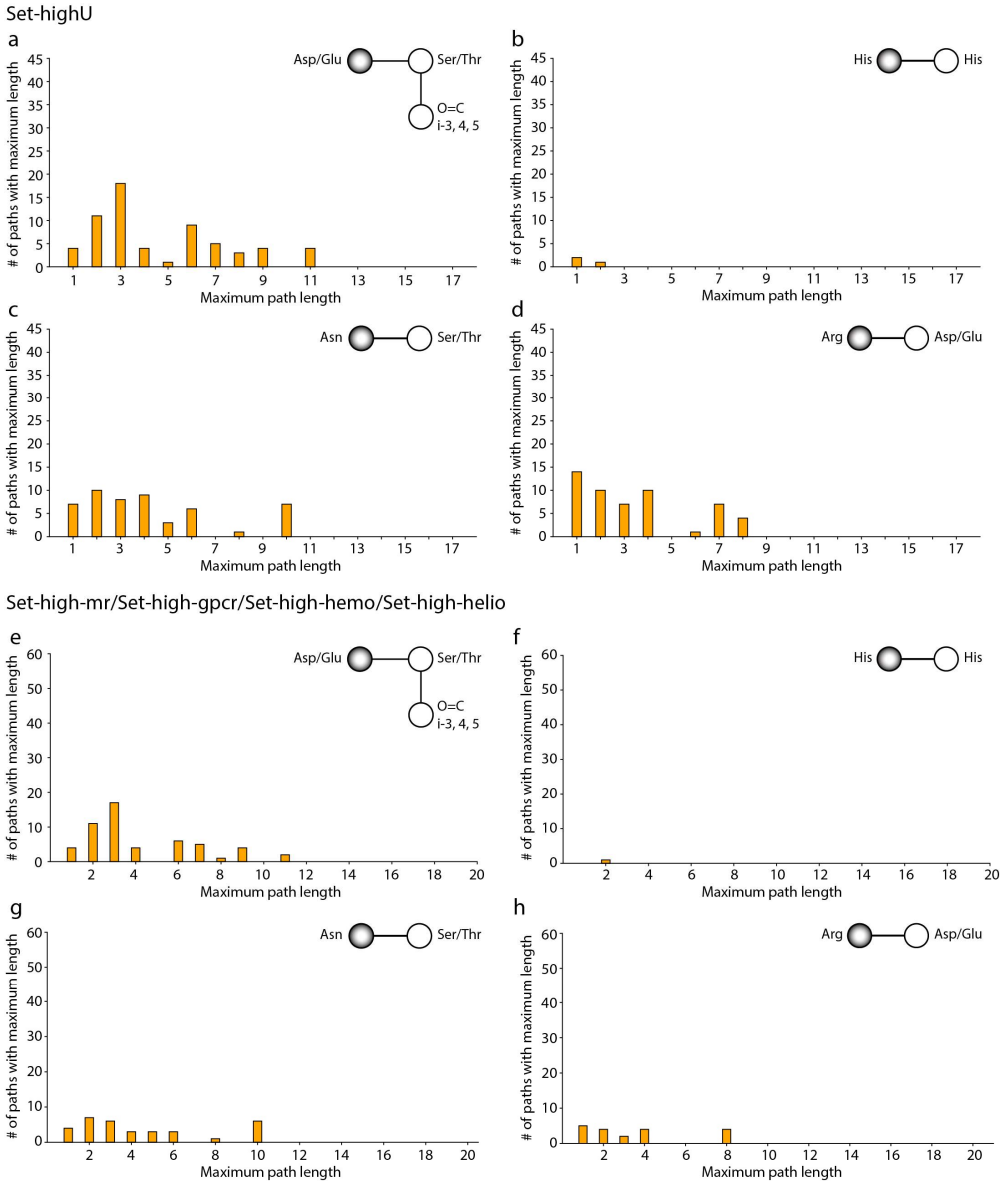

Figure A.17. Shortest path length computations for H-bond motifs in two crystal structure data sets. The root nodes are unique entries per amino acid residue that participate in H-bond motifs and are *peripheral* to the H-bond network (Figure 2.7, Figure 2.8). Path length distributions are shown for the carboxylate groups participating in the combined carboxyl-hydroxyl-carbonyl of the *i*-3,4,5 relative position H-bond motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (a, d). Histidine amino acid residues participating in His-His motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (b, f). Asparagine amino acid residues participating in Asn-Ser/Thr motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (c, g). Arginine amino acid residues participating in Arg-Asp/Glu motifs in *Set-highU* vs. the subsets *Set-high-mr, Set-high-gpcr, Set-high-hemo* and *Set-high-helio* (d, h). Additional analysis for sets *Set-high* and set *Set-low* is found in Figure 5.12.

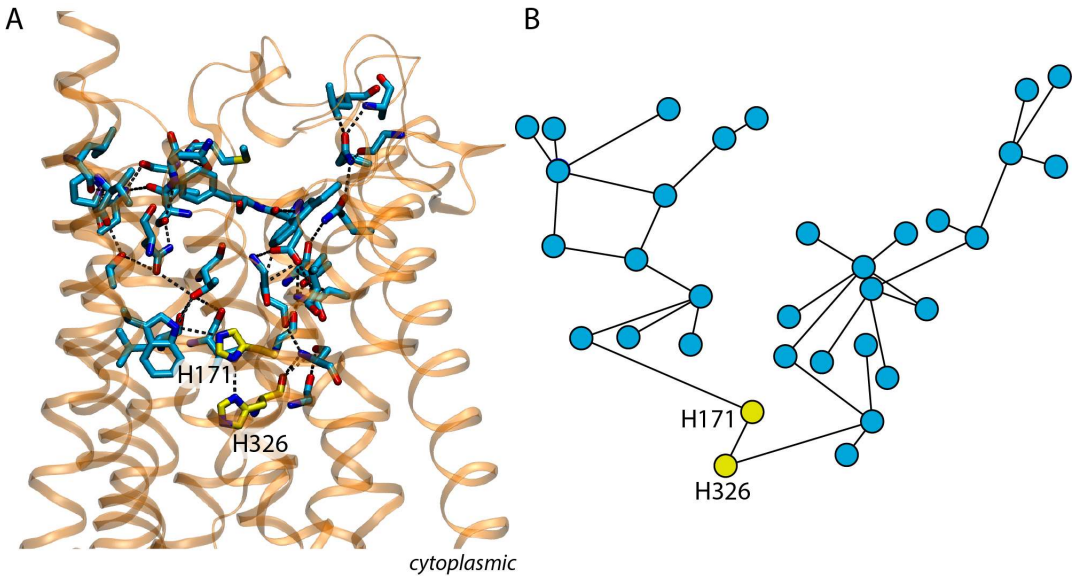

Figure A.18. Illustration of H-bond paths that include the His-His H-bond of the ammonium sensor/transducer. (A) The internal H-bond network that contains the His-His H-bond. The molecular graphics is based on PDB ID:6EU6 [290]. The two H-bonded His sidechains are colored yellow. (B) Graph representation of the H-bond network containing the longest H-bond paths that include the H-bond between H171 and H326. Adapted from ref. [226].

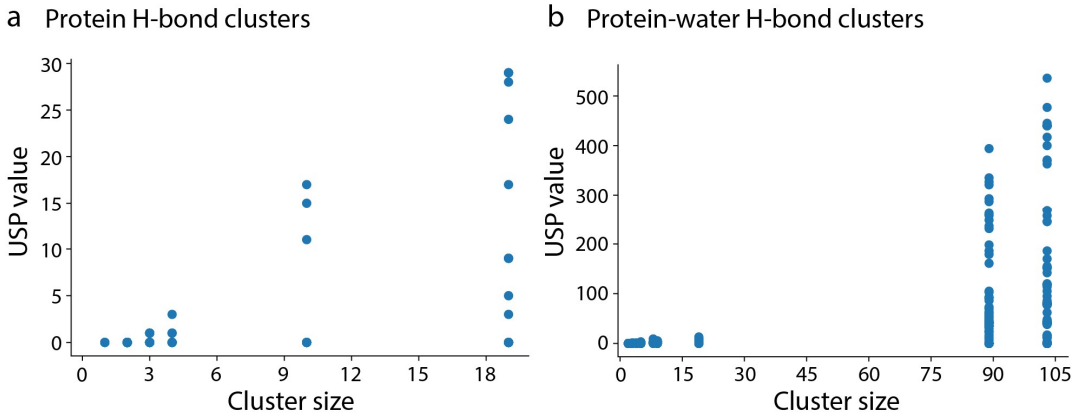

Figure A.19. Cluster size analysis for the Aquaporin1 crystal structure. (a, b) USP values as a function of H-bond cluster size computed in direct protein-protein H-bond graphs (a). USP values as a function of H-bond cluster size computed in direct protein-protein, protein-water, and water-water H-bond graphs (b). Adapted from ref. [226].

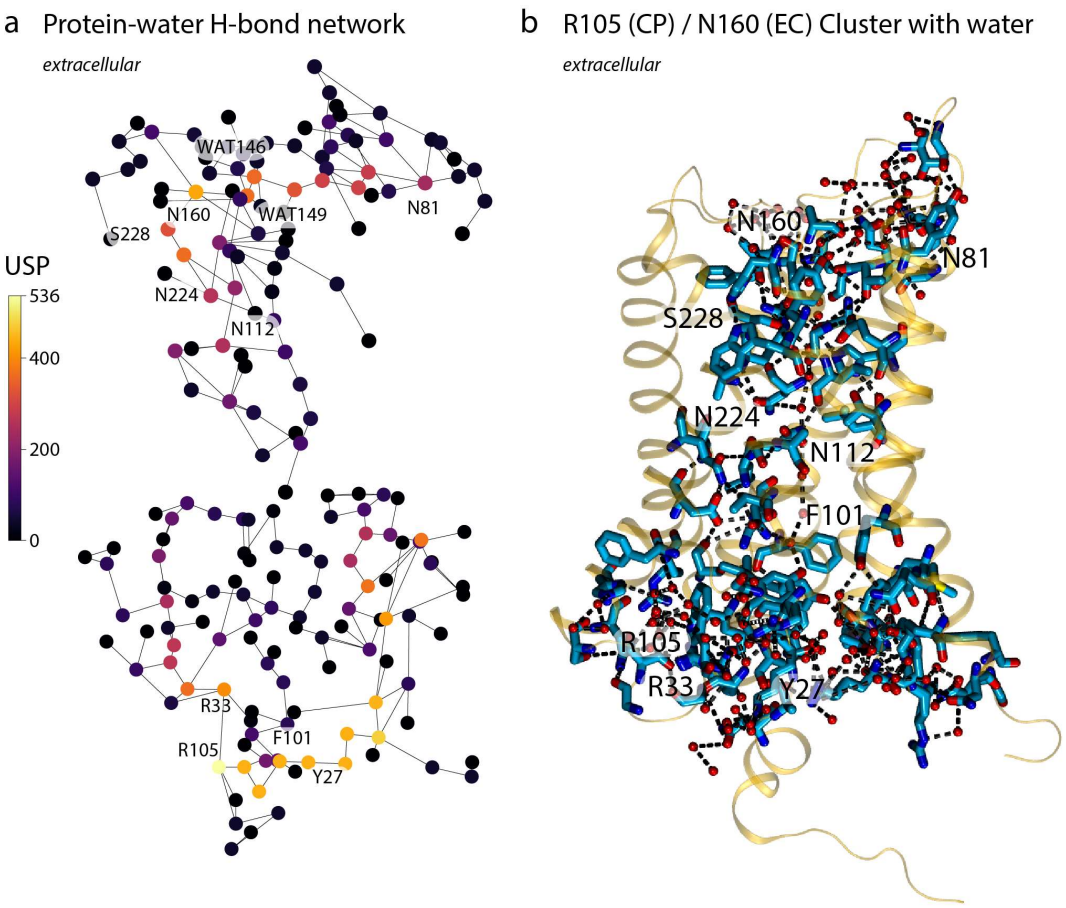

Figure A.20. High Unique Shortest Paths clusters in the crystal structure of Aqy1. The USP computations are performed in direct protein-protein, protein-water, and water-water H-bond graphs. (a) The cytoplasmic R105 cluster and the EC N160 cluster are shown in a graph representation. The graph is color coded with a perpetually uniform color scale and it ranges from USP value of 0 to 536 which belongs to R105. (b) Molecular graphics of the H-bond map depicted in panel (a). Adapted from ref. [226].

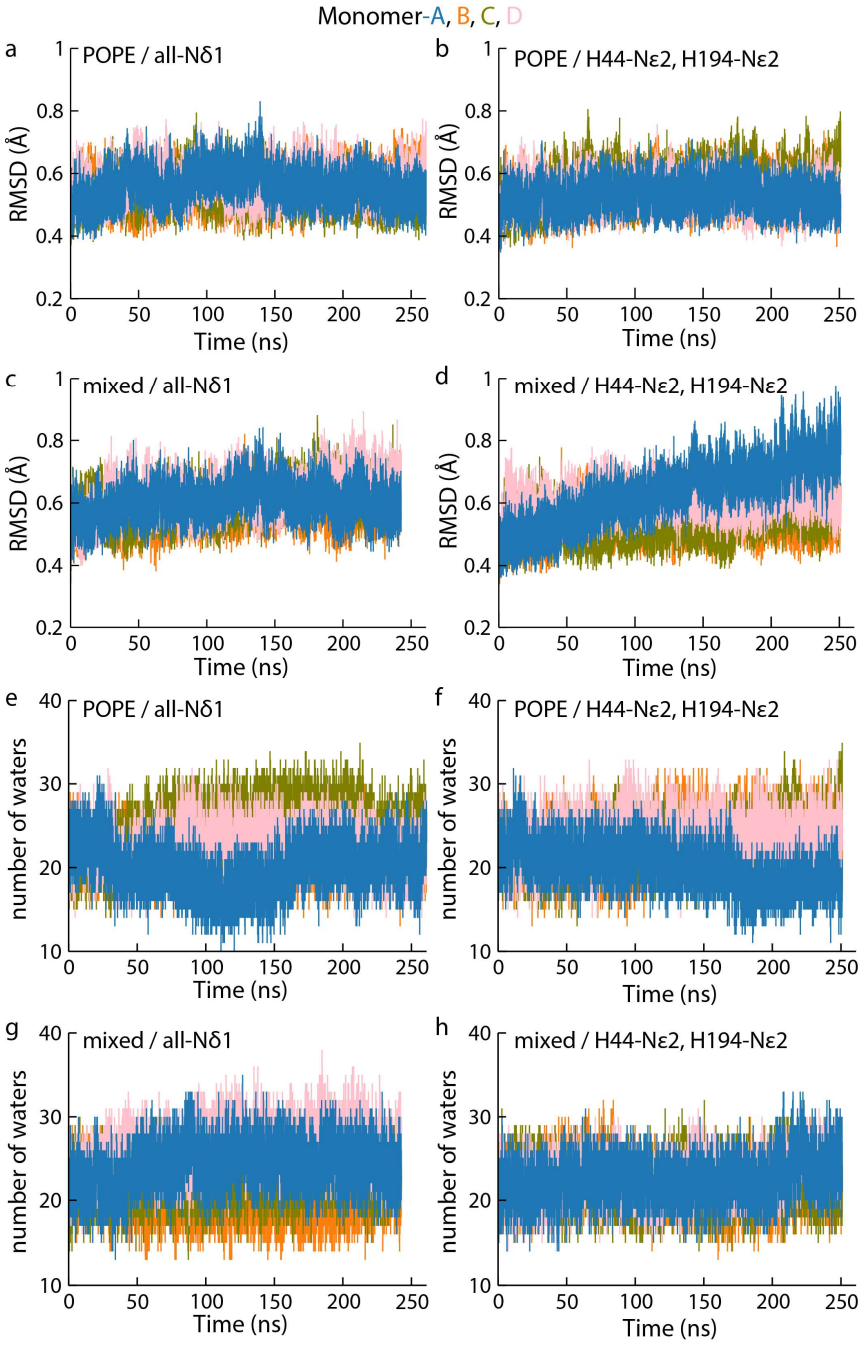

Figure A.21. Structural stability of the Aqy1 tetramer embedded in two lipid bilayer simulations. Cα RMSD profiles (a-d) and number of internal water molecules profiles computed for the four monomers separately. Monomers A-D are colored blue, orange, olive and pink. (A-B) Cα RMSD profiles computed for Aqy1 in POPE membrane, with all four histidine sidechains Nδ1 protonated (a), vs. with H44 and H194 Nε2 protonated (b). (c, d) Cα RMSD profiles computed for Aqy1 in a mixed POPE:POPC:POPS membrane, with all four histidine sidechains Nδ1 protonated (c), vs. with H44 and H194 Nε2 protonated (d). (e-h) Time series for the number of waters that visit the TM region of Aqy1 in simulations with POPE membrane and all four histidine sidechains Nδ1 protonated (e), vs. with H44 and H194 Nε2 protonated (f), and in mixed membrane simulations all four histidine sidechains Nδ1 protonated (f), vs. with H44 and H194 Nε2 protonated (g). Adapted from ref. [226].

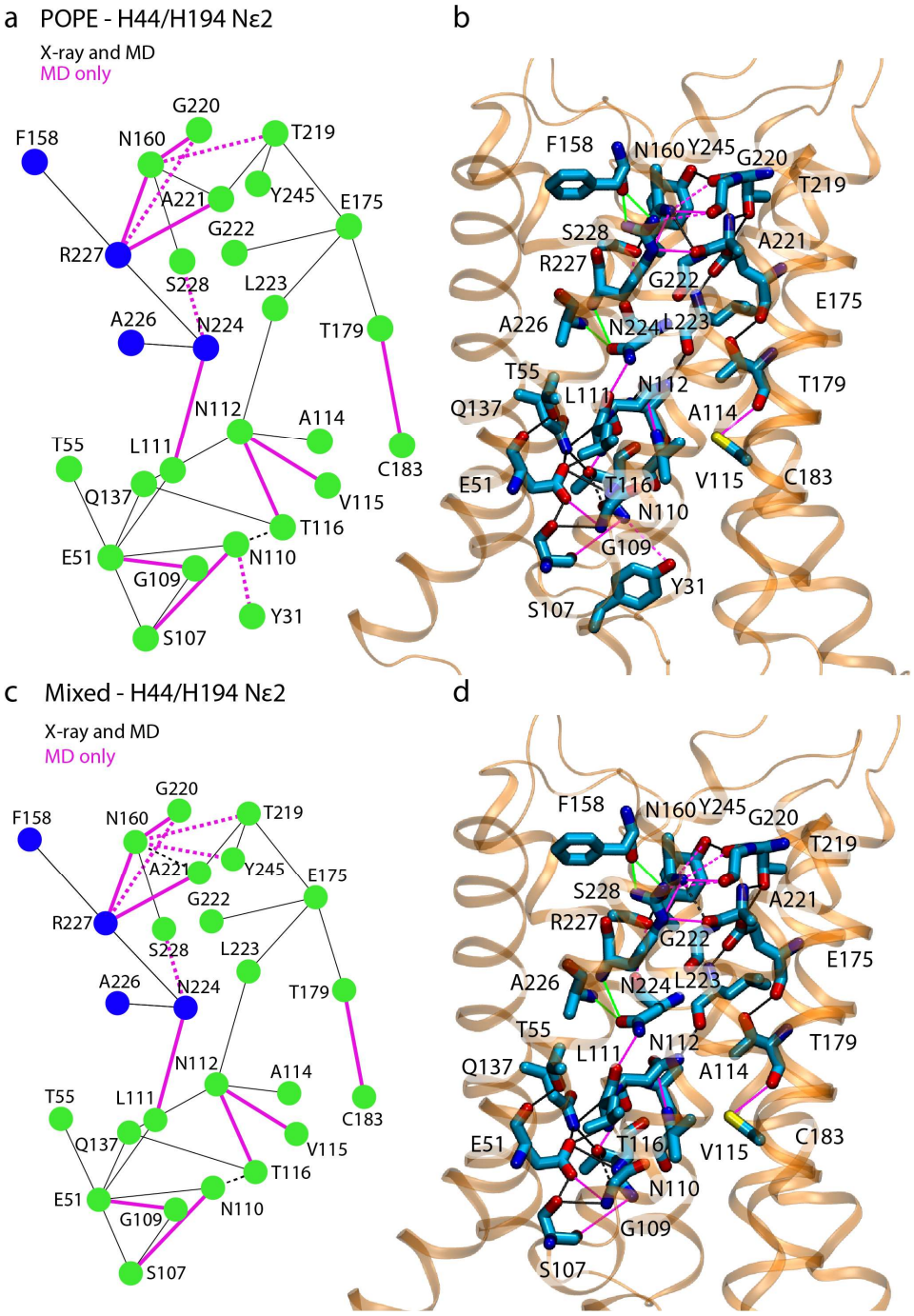

Figure A.22. H-Bond clusters in Aqy1 simulations and crystal structure. The R227-N112 H-bond clusters in the simulations aq1_b, aq2_b and crystal structure are presented in graph representations for simulations of Aqy1 embedded in a POPE (a) and a mixed (c) bilayer. H-bonds found both in the X-Ray and MD simulation are shown with black lines and H-bonds found only in the MD simulation with magenta. Solid lines represent H-bonds found in all four monomers of aquaporin while dotted lines represent H-bonds found in 1 to 3 monomers. Blue nodes represent the R227 cluster in the selectivity filter found in the crystal structure. (b, d) Molecular graphics of the clusters shown in panels A and C respectively. The original R227 cluster is shown with green lines. The networks are represented on the crystal structure of Aqy1 (PDB ID: 3ZOJ). Adapted from ref. [226].

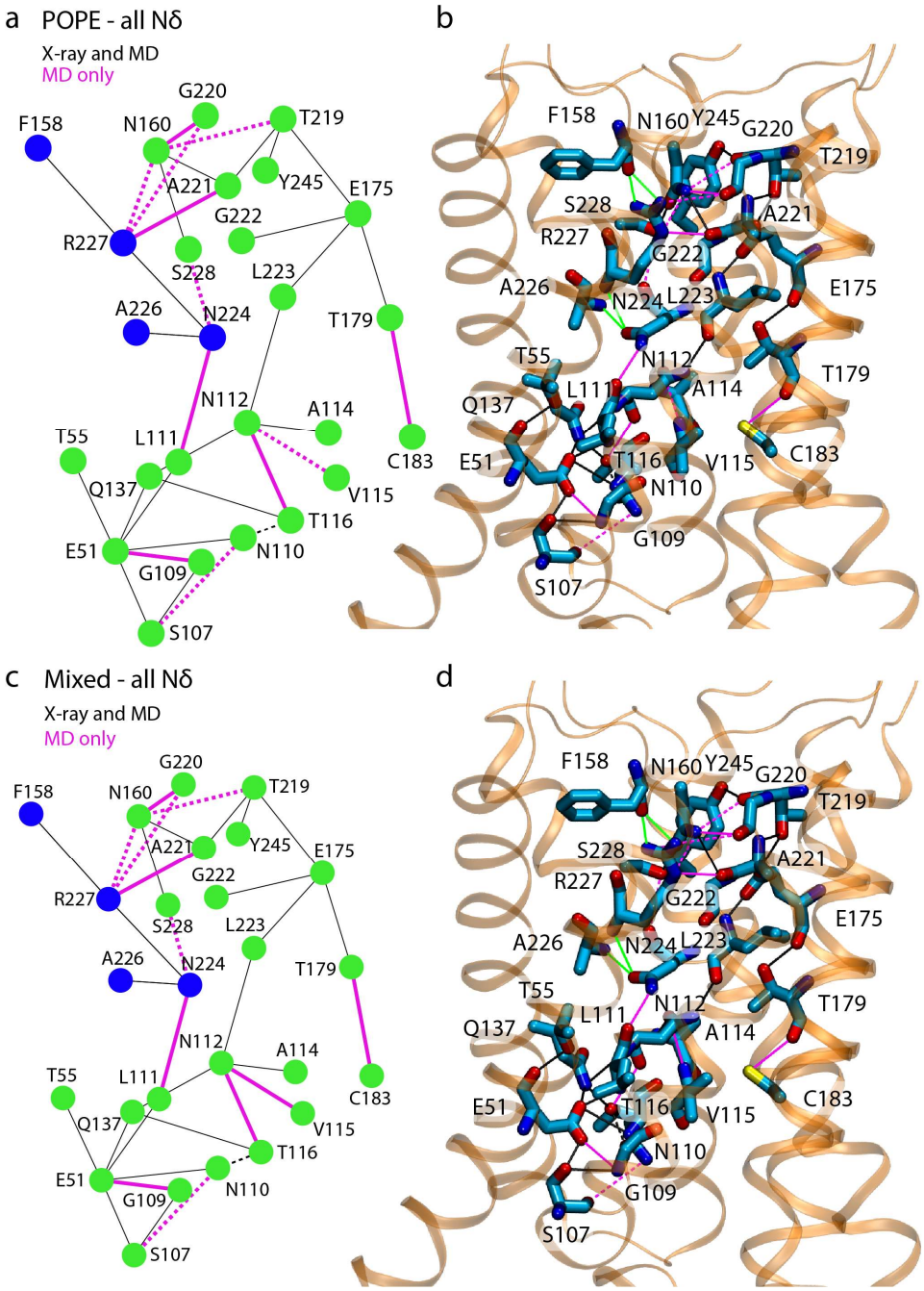

Figure A.23. H-Bond clusters in Aqy1 simulations and crystal structure. The R227-N112 H-bond clusters in the simulations aq1_a, aq2_a and crystal structure are presented in graph representations for simulations of Aqy1 embedded in a POPE (a) and a mixed (c) bilayer. H-bonds found both in the X-Ray and MD simulation are shown with black lines and H-bonds found only in the MD simulation with magenta. Solid lines represent H-bonds found in all four monomers of aquaporin while dotted lines represent H-bonds found in 1 to 3 monomers. Blue nodes represent the R227 cluster in the selectivity filter found in the crystal structure. (b, d) Molecular graphics of the clusters shown in panels A and C respectively. The original R227 cluster is shown with green lines. The networks are represented on the crystal structure of Aqy1 (PDB ID: 3ZOJ). Adapted from ref. [226].

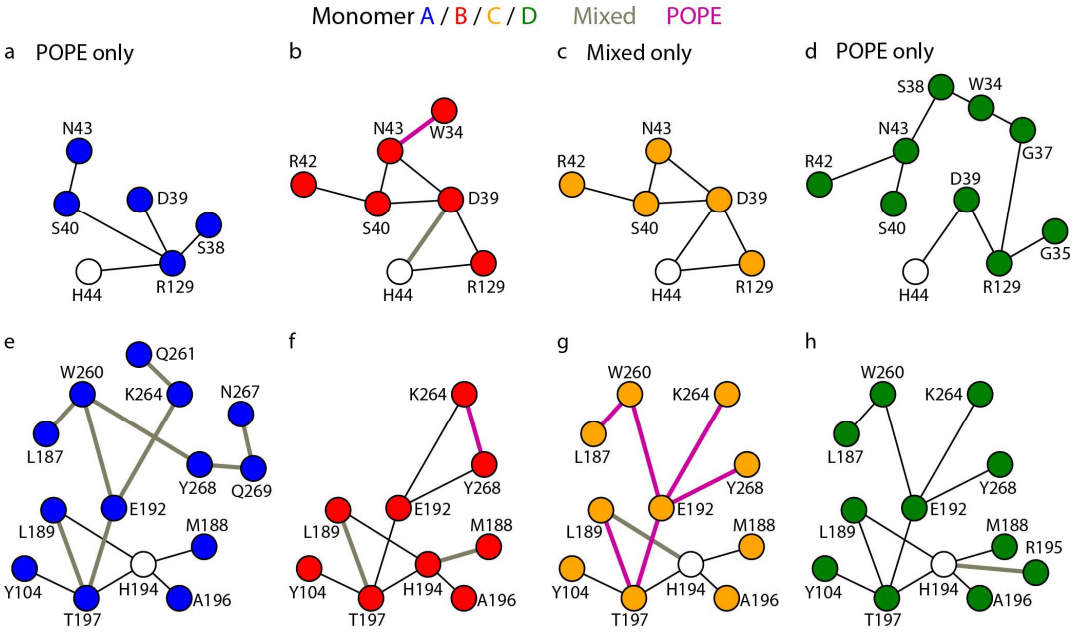

Figure A.24. Comparative graphs of H-bond clusters of H44 and H194 sampled in four different MD simulations of Aqy1. The nodes representing H44 (a-d) and H194 (e-h) are shown as an empty circle. (a-d) H-bond clusters of H44 in monomers A (a), B (b), C (c) and D (d) from simulations in a POPE (purple) vs. a mixed (gray) membrane with H44 and H194 Nε2 protonated. (e-h) H-bond clusters of H194 in monomers A (e), B (f), C (g) and D (h) from simulations in a POPE (purple) vs. a mixed (gray) membrane with H44 and H194 Nε2 protonated. In some cases, the graphs are sampled only in the POPE (a, d) or in the mixed (c) membrane. Adapted from ref. [226].

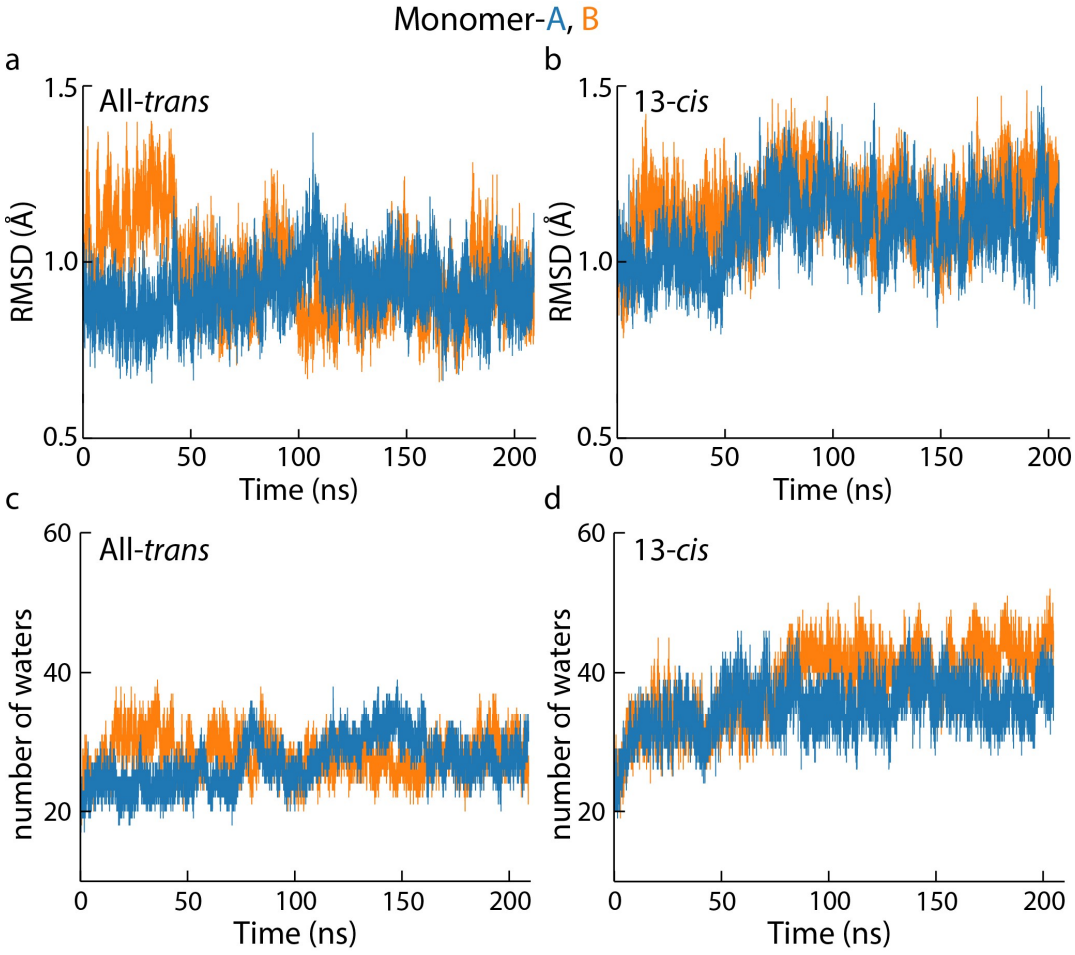

Figure A.25. Cα RMSD profiles and number of internal water molecules profiles in two MD simulations of ChR2. Profiles for Monomers A and B are colored blue and orange respectively. (a, b) RMSD profiles computed for ChR2 with all-*trans* retinal (a) vs. 13-*cis*-15-*anti* retinal (b). (c, d) Internal water molecules profiles of ChR2 with all-*trans* retinal (c) vs. with 13-*cis*-15-*anti* retinal (d). Adapted from ref. [226].

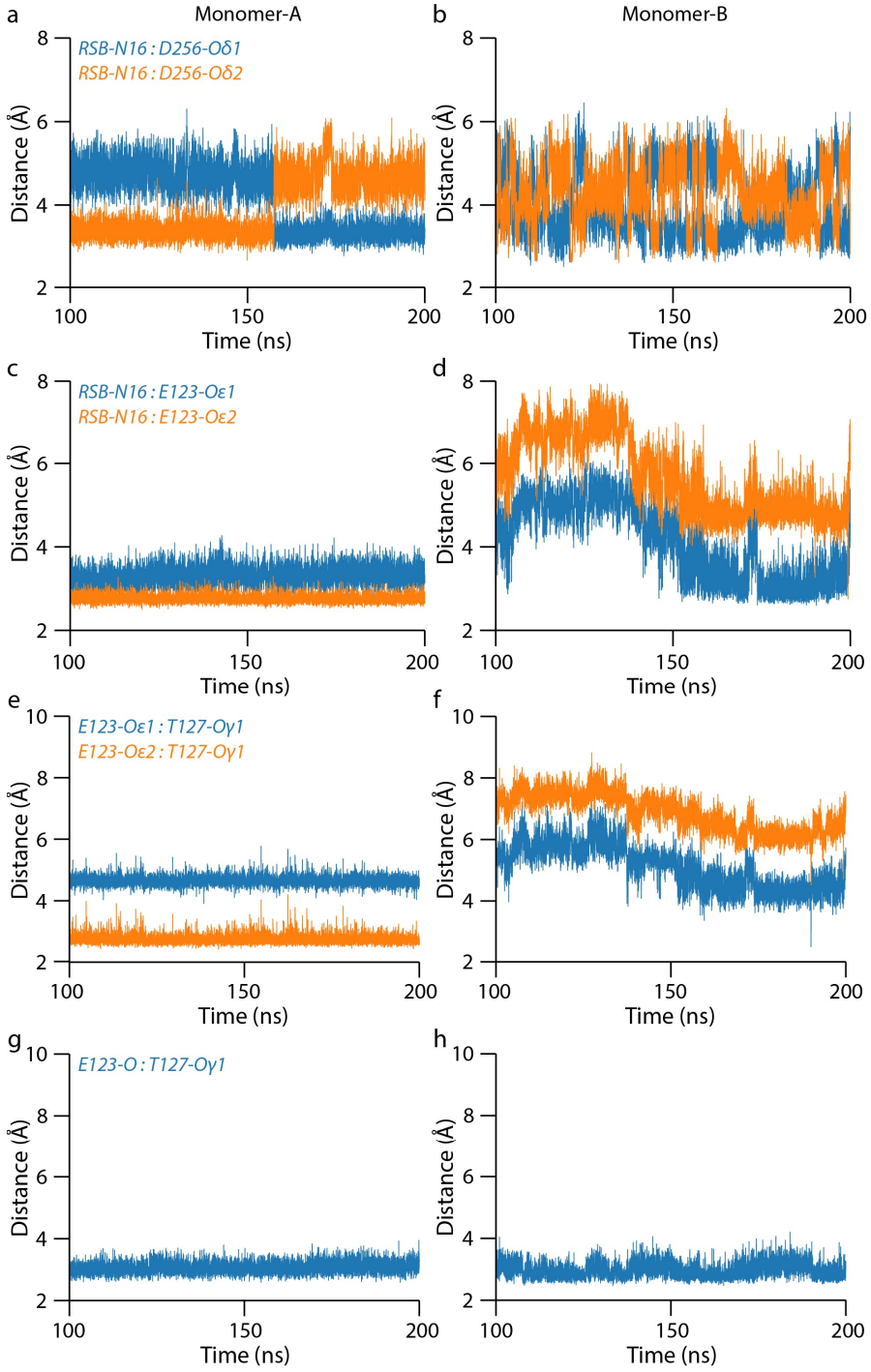

Figure A.26. Distance timeseries of H-bonds at the RSB vicinity in simulations of ChR2 with an all-*trans* retinal. The timeseries are shown for monomers A (left) and B (right). (a, b) Distance timeseries between the protonated nitrogen atom N16 of the RSB and D256-Oδ1 (blue) vs. D256-Oδ2 atoms (orange) in Monomer-A (a) and Monomer-B (b). (c, d) Distance timeseries between the protonated nitrogen atom of the RSB-N16 and E123-Oε1 (blue), and between RSB-N16 and E123-Oε2 (orange) in Monomer-A (c) and Monomer-B (d). (e, f) Distances between T127-Oγ1 and E123-Oε1 (blue), and between T127-Oγ1 and E123-Oε2 (orange) in Monomer-A (e) and Monomer-B (f). (g, h) Distances between T127-Oγ1 and E123-O in Monomer-A (g) and Monomer-B (h). Adapted from ref. [226].

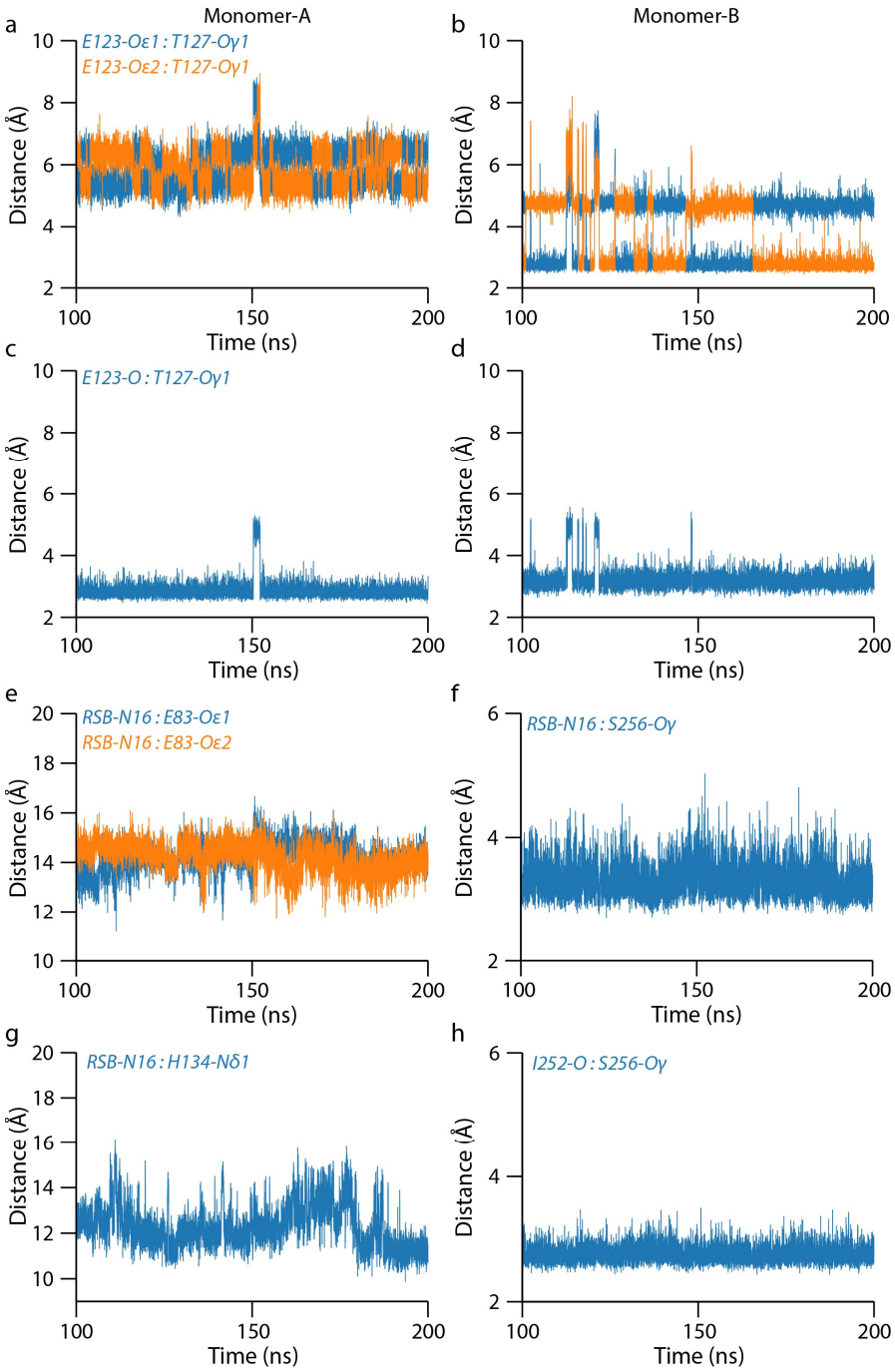

Figure A.27. Distance timeseries of H-bonds at the RSB vicinity in simulations of ChR2 with a 13-*cis*,15-*anti* retinal. The timeseries are shown for monomers A (left) and B (right). (a, b) Distances between T127-Oγ1 and E123-Oε1 (blue), vs. T127-Oγ1 and E123-Oε2 (orange). (c, d) Distances between T127-Oγ1 and E123-O in monomers A (panel C) and B (panel D). (e) Distance between RSB-N16 and E83-Oε1 (blue), and RSB-N16 and E83-Oε2 (orange). (f-h) Distances between RSB-N16 and S256-Oγ in Monomer-B (f), RSB-N16 and H134-Nδ1 in Monomer-A (g), and between I252-O and S256-Oγ in monomer-B (h). Adapted from ref. [226].

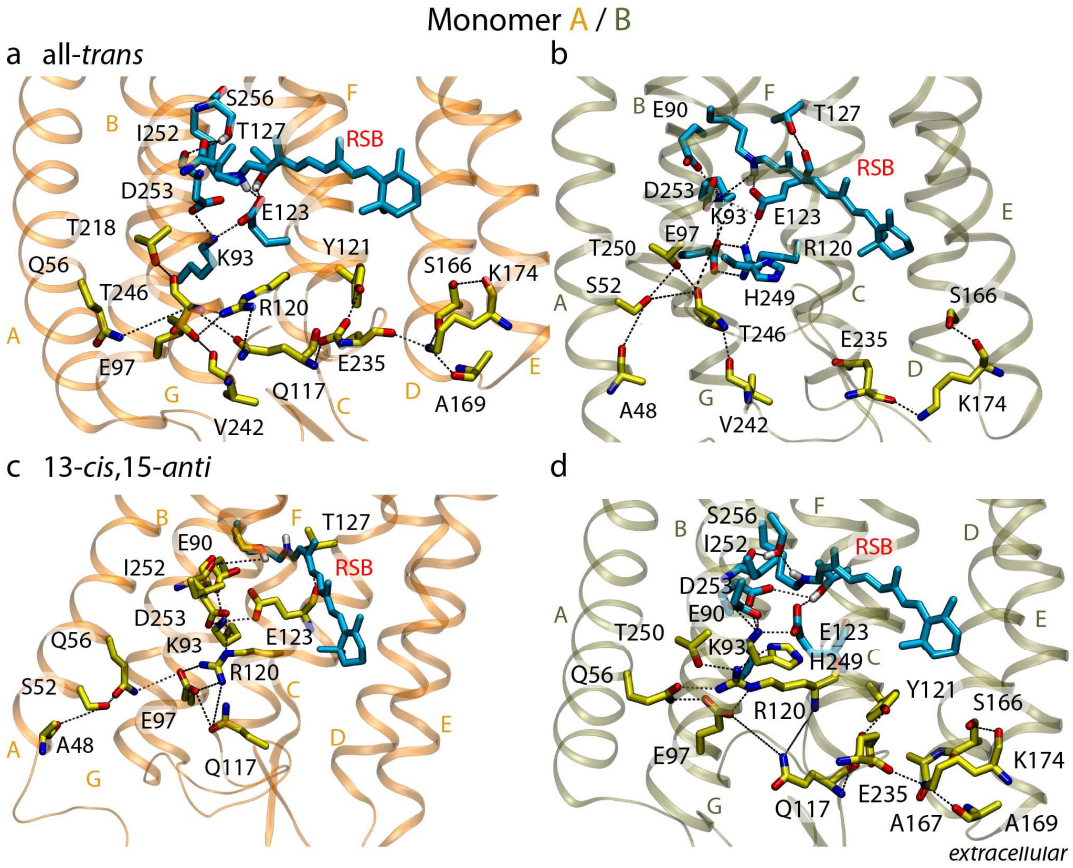

Figure A.28. Extracellular H-bond clusters sampled in MD simulations of ChR2. (a, b) H-bond clusters in the all-*trans* isomeric state sims of ChR2 for Monomer-A (a) and Monomer-B (b). (c, d) H-bond clusters in the 13-*cis*,15-*anti* isomeric state sims of ChR2 for Monomer-A (a) and Monomer-B (b). The EC clusters are presented in yellow color. The direct RSB networks are shown in cyan. The RSB are presented isolated in Figure 5.18 and Figure 5.21. TM-helices are labelled A-F. Adapted from ref. [226].

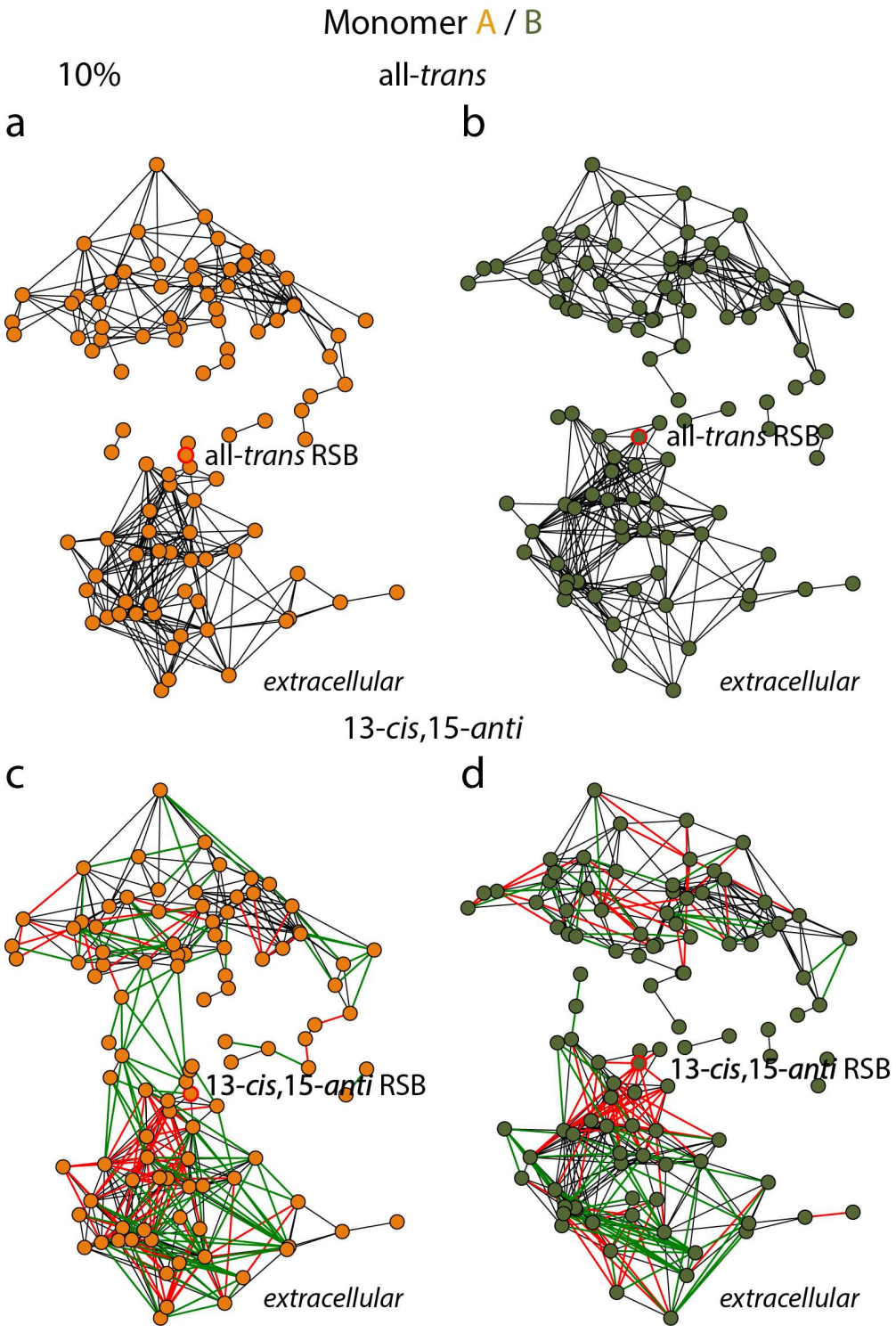

Figure A.29. Comparative H-bond graphs between the all-*trans* and 13-*cis* models using a 10% occupancy threshold. (a, b) Reference graphs for the water wire networks of ChR2, sampled in Monomer-A (a) and B (b), in the all-*trans* model respectively. (c, d) Difference graphs between the all-*trans* and 13-*cis* models, in Monomer-A (c) and B (d).

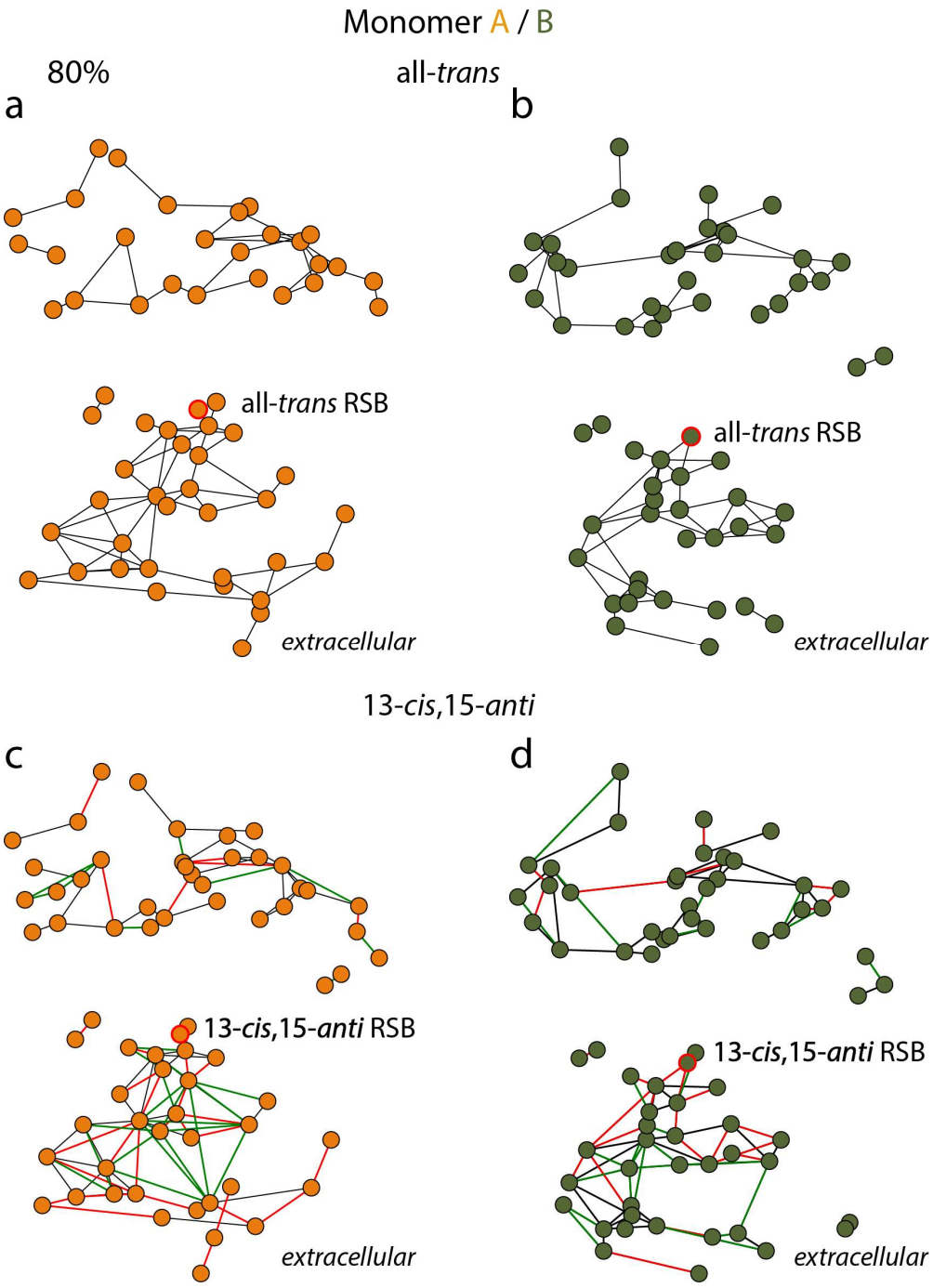

Figure A.30. Comparative H-bond graphs between the all-*trans* and 13-*cis* models using a 80% occupancy threshold. (a, b) Reference graphs for the water wire networks of ChR2, sampled in Monomer-A (a) and B (b), in the all-*trans* model respectively. (c, d) Difference graphs between the all-*trans* and 13-*cis* models, in Monomer-A (c) and B (d).

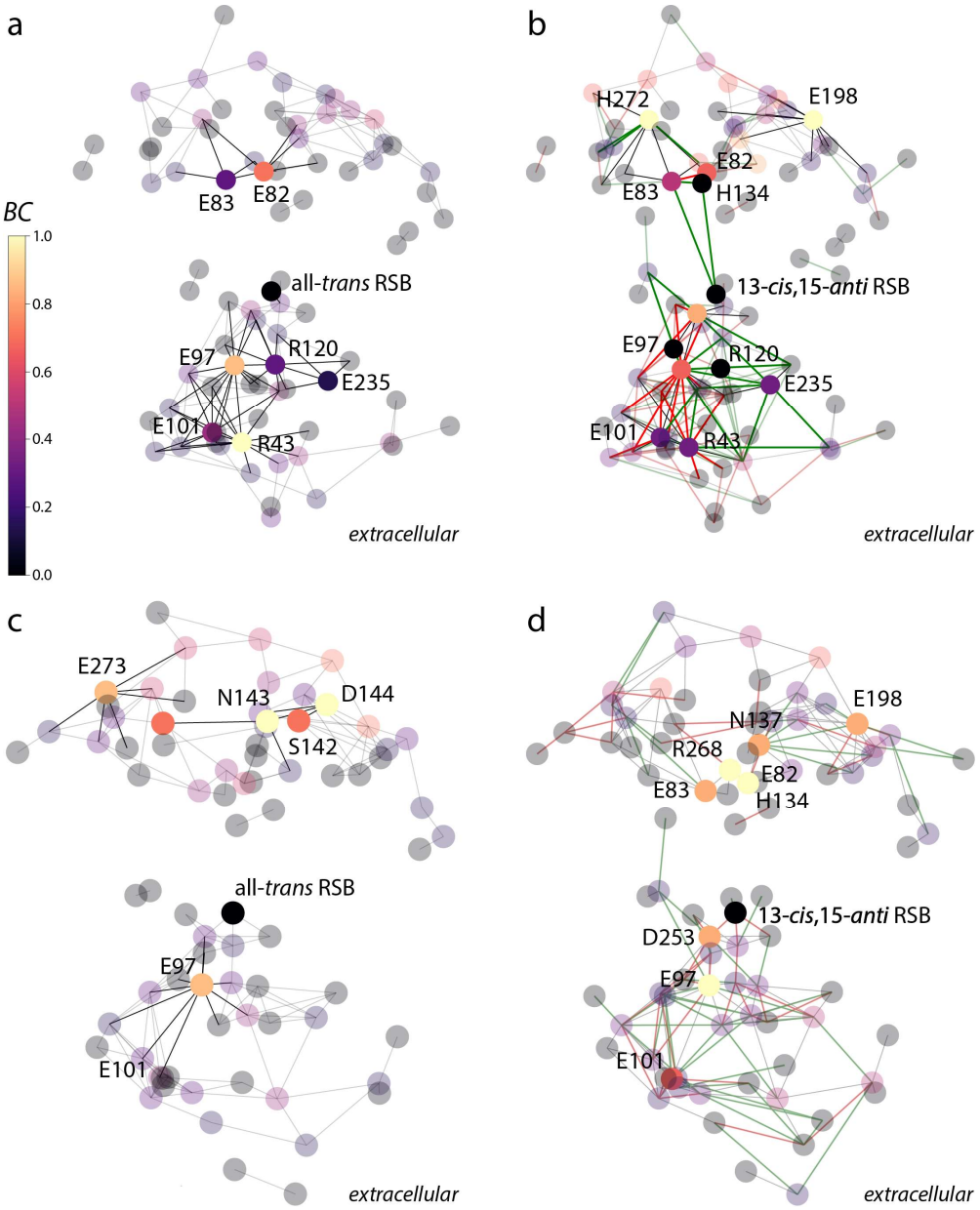

Figure A. 31. Betweenness centrality computations for the all-*trans* and 13-*cis*,15-*anti* models of ChR2. Reference graphs computed for the all-*trans* model for Monomer-A (a) and B (b), respectively. Difference graphs between all-*trans* and 13-*cis* models for Monomer-A (c) and B (d). Graphs were prefiltered to 50% occupancy rates, before the BC computations.

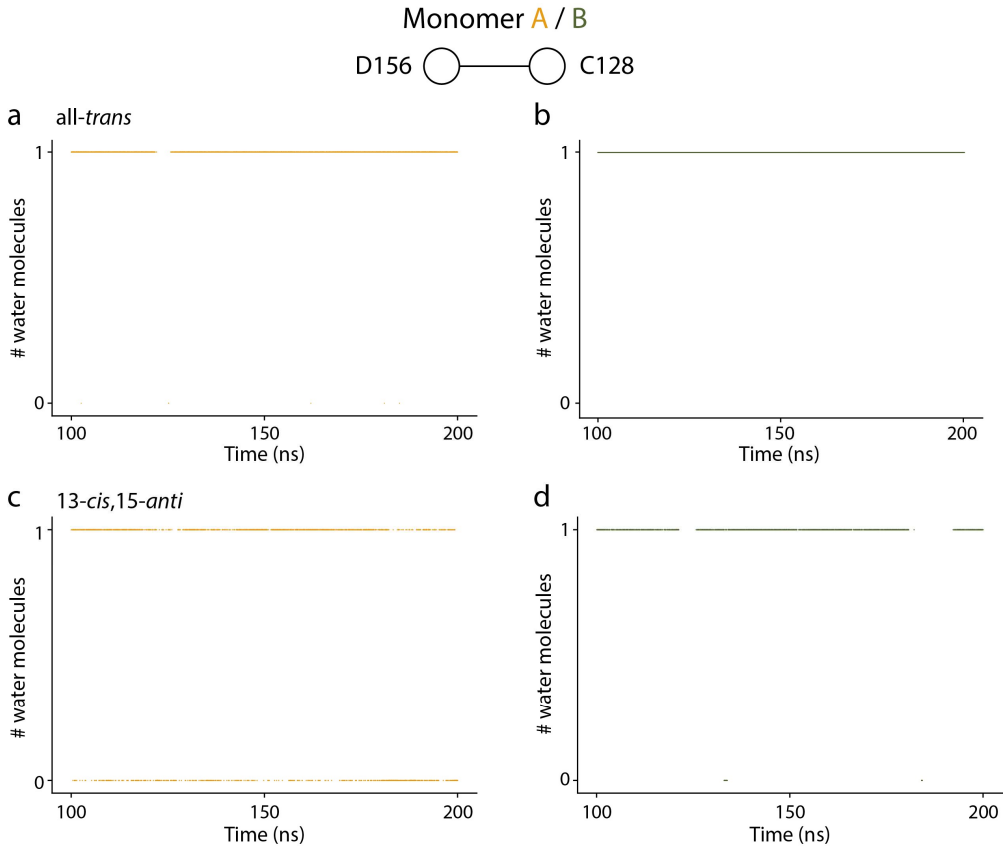

Figure A.32. The DC gate is mostly water mediated. (a, b) Timeseries for the water content of the DC gate sampled in the all-*trans* for Monomer-A (a) and B (b), respectively. (c, d) Timeseries for the water content of the DC gate sampled in the 13-*cis*,15-*anti* for Monomer-A (c) and B (d), respectively. Monomers-A and B are colored orange and green, respectively.

Molecular Dynamics simulation principles

In this appendix I will summarize the basic principles of the widely known and used approach of Molecular Dynamics Simulation. Algorithmic implementation of those principles lies within the individual software available and will not be discussed in this Appendix. The equations and principles presented here are based on classic textbooks of the field [137, 338-341].

Equations of motion

The technique of Molecular dynamics is based on classical mechanic and more precisely the very elegant equation of Newton that describes his second law of motion. According to Newton's second law of motion the force acting on a particle, is expressed by equations A.1 and A.2.

$$\boldsymbol{\mathcal{F}} = m\frac{d^2\boldsymbol{r}}{dt^2} \tag{A.1}$$

$$\boldsymbol{\mathcal{F}} = -\boldsymbol{\nabla}\mathcal{V} \tag{A.2}$$

Combining equations A.1 and A.2 the second law of motion can be written in its differential form.

$$-\frac{d\mathcal{V}}{d\boldsymbol{r}} = m\frac{d^2\boldsymbol{r}}{dt^2} \tag{A.3}$$

The successive configurations that comprise the trajectory are computed by integrating Newton's laws of motion. The integration is performed in very small, finite stages called timesteps using finite difference techniques. At a time *t* the total force acting on each particle is calculated through the vectorial sum of its interactions with other particles [137]. From equation A.1, the accelerations of the particles are computed and their positions at a time $t + \delta t$, considering that the positions and velocities at time *t* are known. All integration schemes employ the Taylor series to expand the expressions of positions, velocities, accelerations, and higher order derivates.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2!}\boldsymbol{a}(t)\delta t^2 + \frac{1}{3!}\boldsymbol{b}(t)\delta t^3 + \frac{1}{4!}\boldsymbol{c}(t)\delta t^4 + \cdots \tag{A.4}$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\delta t + \frac{1}{2!}\boldsymbol{b}(t)\delta t^2 + \frac{1}{3!}\boldsymbol{c}(t)\delta t^3 + \cdots \tag{A.5}$$

$$\boldsymbol{a}(t+\delta t) = \boldsymbol{a}(t) + \boldsymbol{b}(t)\delta t + \frac{1}{2!}\boldsymbol{c}(t)\delta t^2 + \cdots \quad \text{A.6}$$

$$\boldsymbol{b}(t+\delta t) = \boldsymbol{b}(t) + \boldsymbol{c}(t)\delta t + \cdots \quad \text{A.7}$$

Where **r** is the position, **υ** the velocity, **a** the acceleration, **b** is the jerk (or jolt) and **c** is the snap (or jounce) of a particle. Higher order derivates are not shown here for simplicity. There are several algorithms developed for the integration of the equations of motion, and in this work the most widely known ones will be presented. Perhaps the most recognized and used implementation is the Verlet [342] scheme, which expands the expression of the positions around $t \pm \delta t$, as shown in equations A.8 and A.9.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2!}\boldsymbol{a}(t)\delta t^2 + \frac{1}{3!}\boldsymbol{c}(t)\delta t^3 + \mathcal{O}(\delta t^4) \quad \text{A.8}$$

$$\boldsymbol{r}(t-\delta t) = \boldsymbol{r}(t) - \boldsymbol{v}(t)\delta t + \frac{1}{2!}\boldsymbol{a}(t)\delta t^2 - \frac{1}{3!}\boldsymbol{c}(t)\delta t^3 + \mathcal{O}(\delta t^4) \quad \text{A.9}$$

Addition of equations A.8 and A.9 gives:

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \boldsymbol{a}(t)\delta t^2 + \mathcal{O}(\delta t^4) \quad \text{A.10}$$

The Verlet algorithm has two major disadvantages. Numerically it may lead to truncation errors due to the finite precision of computers, since the small term $\mathbf{a}(t)\delta t^2$, of the order of $\mathcal{O}(\delta t^2)$ is added to the larger term $2\mathbf{r}(t) - \mathbf{r}(t-\delta t)$, of the order of $\mathcal{O}(\delta t^0)$ for the calculation of the positions $\mathbf{r}(t+\delta t)$.
Although from eq. A.10 the velocities are not required for the generation of the trajectory, they are required for the kinetic energy, and/or to facilitate for coupling to an external bath. The calculation of the positions does not have an explicit expression for the velocities, but they can be calculated through an additional step, by subtracting equations A.8 and A.9. Equation A.11 shows that the velocities for time t are not actually available until $t + \delta t$.

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t+\delta t) - \boldsymbol{r}(t-\delta t)}{\boldsymbol{2}\delta t} + \mathcal{O}(\delta t^2) \quad \text{A.11}$$

The drawback with the velocities can be addressed using a variation of the Verlet scheme, the intuitively called, “leapfrog algorithm” [343-345] in which half-timesteps $\delta t/2$ are

used for the calculation of velocities and whole timesteps for the positions and accelerations.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right)\delta t + \mathcal{O}(\delta t^2) \quad \text{A.12}$$

$$\boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) = \boldsymbol{v}\left(t-\tfrac{1}{2}\delta t\right) + \boldsymbol{a}(t)\delta t + \mathcal{O}(\delta t^2) \quad \text{A.13}$$

With the leapfrog the velocities are computed explicitly in the integration and numerically is more accurate since there is no subtraction of larger and smaller terms (eq. A.10). Since it is a derivation of the Verlet algorithm, it gives rise to identical trajectories [340], but now the velocities and positions are calculated asynchronously, thus the kinetic and potential energy cannot be calculated at the same time, resulting to the inability of computing the total energy at any time, via the relation $\mathcal{H} = \mathcal{K} + \mathcal{V}$.
An equivalent to the original Verlet algorithm is the "velocity Verlet" [346] which stores the positions velocities and accelerations at the same time t. This minimizes the rounding errors that are present in the original Verlet algorithm [338]. The expression of the positions is yielded from the Taylor expansion for $t + \delta t$ (eq. A.4). Eliminating the velocities, the original Verlet scheme can be yielded.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 + \mathcal{O}(\delta t^3) \quad \text{A.14}$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \frac{1}{2}\delta t[\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)] + \mathcal{O}(\delta t^3) \quad \text{A.15}$$

$$\boldsymbol{v}\left(t+\tfrac{1}{2}\delta t\right) = \boldsymbol{v}(t) + \frac{1}{2}\boldsymbol{a}(t)\delta t \quad \text{A.16}$$

The velocity Verlet scheme is perhaps the most widely known and used integrator for the generation of MD trajectories. It preserves the volume in phase-space (*symplectic* integrator) and the angular momentum, is time-reversible and provides good numerical stability. Namely, other algorithms for integrating the equations of motion that are not going to be discussed here are: Euler algorithm, Beeman's algorithm, velocity-corrected Verlet algorithm, predictor-corrector and Gear predictor-corrector.

Choosing the timestep

The choice of the timestep δt, that will be used in the integration of the equations of motions is a crucial step in the procedure of an MD simulation. It is essentially a compromise between accuracy and computational time. Too large, and there are high risks of numerical instabilities and the simulation "exploding". With a large timestep,

atoms can jump energy barriers, the energy increasing rapidly with time. If two atomic coordinates nearly overlap, the underlying atoms will repel each other with very high velocities and create instabilities in the system. Too small, and the phase space will not be adequately visited, or sampled. The fastest motion of the system described will dictate the value of the timestep, and it is typically an order of magnitude smaller than that. For a flexible molecule, like a protein, the bond stretching that contains a hydrogen atom will undoubtedly be the fastest motion. In an Infrared Spectroscopy (IR) spectrum the C-H vibration is found around 3000-3300 cm^{-1}, and this translates to 10-11 fs. Thus, the timestep is set at 1 fs for most applications. There are ways to speed up a simulation while maintain its accuracy. This can be achieved by employing a multiple timestep integrator scheme.

Periodic Boundary Conditions

A mole of atoms, by definition contains 6.02214076×10^{23} particles. The goal of an MD simulation is to extract macroscopic properties of a system, but it is not possible to simulate the number of atoms that are found in macroscopic systems, on the order of moles. Computer simulations are often performed on a much smaller number of molecules. Historically, the ever so increasing capability of computer hardware and software has enabled the number of simulated atoms to increase. In 1989, Allen and Tildesley report that the system size is between 10 and 10.000 molecules [338]. Today, a simulated system can exceed this quote by two orders of magnitude. The simulation is performed in a confined space which contains all the elements of the system, called the simulation box. The box is usually cubic, but not necessarily. Particles close to the surface (walls of the box) are of the order of $N^{2/3}$, where N is the number of atoms in the box. For a system in the order of moles, this does not pose a big problem, but for a more realistic simulation system size, atoms in the interior of the box are proportionately very few. Such atoms will experience different forces compared to the atoms in the bulk [338], meaning pronounced surfaced effects will predominate, and reliable "bulk" properties will not be able to be derived [137]. To simulate bulk behavior the Periodic Boundary Conditions (*PBC*) have been introduced to mimic the presence of an infinite bulk surrounding the system at hand [340]. The atoms are placed in the box and the box is multiplied in all dimensions. For a three-dimensional system, the central box will now be surrounded by 26 replica images, and those will be surrounded by 98 replica images etc. The positions of the atoms in the replicas are easily computed by adding/subtracting integral multiples of the box sides [137]. The basic principle of the PBC is that when a molecule leaves the central box from any side, it will re-enter through the opposite side from the neighboring box, since atoms in the atoms in the replicas are moving exactly the same way as in the central box [338]. The number density of the central box is conserved in this way, and the only coordinates that need to be stored are the ones of the central image and not of all the periodic images. This would result in an infinite number of coordinates. With the introduction the PBC, surface effects are now eliminated because there are no boundaries between the simulation boxes.

Minimum Image Convention

One aspect that should not be neglected is that particles from the central box will not only interact with other particles from the same box. They will also interact with particles from neighboring images. As a result, the number of pair-wise interactions will vastly increase. The pairwise interactions between an arbitrary molecule i, and every other molecule in the central box are $N-1$. If we were to include the interactions between molecule i and every other molecule amongst the periodic images this would result in infinite terms and deem the simulation impossible in practice. An approximation can be constructed for the short-range interactions. According to the minimum image convention a particle is only allowed to interact with at most one image of every other atom in the system. Only the closest image is considered for the calculation of non-bonded interactions. If a non-bonded cut-off is employed, it should not allow for a particle to interact with itself. Thus, its maximum value should be half the length of the simulation box, assuming a cubic cell. Particles that are found further apart than r_{cutoff} will have interaction values of 0.

Neighbor List

The computation of the non-bonded interactions is perhaps the most time-consuming step in the simulation. The introduction of the cutoff distance r_{cutoff} is an efficient way to reduce computational cost. But, in order to evaluate if two particles are within the cutoff distance, their distance must be first calculated. The time to calculate $\frac{N(N-1)}{2}$ distances is almost as much as calculating the energies themselves, and scales as N^2 [137, 340]. The *neighbor lists* are based on the idea that the neighbors of an atom will not drastically change every 10-20 steps. It is then known for which pair of atoms; the non-bonded interactions will be calculated. To do so, the list includes the neighboring atoms of an atom i that are found slightly further away than the cutoff distance. This is due to energy conservation reasons. Assuming two atoms i,j being 10.5 Å apart, in time t, with the cutoff distance set at $r_{cutoff} = 10$ Å, the neighbor list distance set at $r_{list} = 10$ Å and the neighbor list updating every 10 steps. At that time $r_{ij} > r_{cutoff}$ and the non-bonded interactions between i,j will not be computed. After 5 timesteps $(t + 5\delta t)$ the distance $r_{ij} = 9.7$ Å and the non-bonded interactions between i,j should now be computed because $r_{ij} < r_{cutoff}$. But it will not because the atoms i,j are not included in the and the neighbor list will not be updated for another 5 timesteps. This will result in the energy of the system not being conserved and cause instabilities. If the neighbor list distance was set to $r_{list} = 12$ Å, then atoms i,j would be included as neighbor atoms for the whole update cycle of 10 timesteps and the decisive factor of calculating their non-bonded interactions would be if r_{ij} was smaller than r_{cutoff} or not. Atom pairs that are included in the neighbor list but are outside the cutoff distance are simply ignored. For this reason, the r_{list} is set slightly larger than the r_{cutoff} to ensure energy conservation and avoid instabilities like in the example above.

Cutoff Schemes

In the section "Neighboor list", the tight relationship between the list distance r_{list} and the cutoff distance r_{cutoff} was shown. In this section, three widely known cutoff schemes will be discussed.

Truncation

The most simplistic cutoff scheme is the simple truncation where the interaction energy is set to 0 whenever the distance between two atoms i, j , becomes larger than the cutoff distance $r_{ij} \geq r_{cutoff}$.

$$\mathcal{V}_{vdW}(r_{ij}) = \begin{cases} 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right), & r_{ij} < r_{cutoff} \\ 0 \quad , & r_{ij} \geq r_{cutoff} \end{cases} \qquad \text{A.17}$$

The truncation scheme will introduce discontinuity in the energy at r_{cutoff}, with an infinite force and acceleration.

Potential Shift

With the addition of a constant term and a small linear term to the potential, the derivative at r_{cutoff} becomes 0, and the discontinuity appears in the gradient of the force, and not the force itself [338].

$$\mathcal{V}_{vdW}(r_{ij}) = \begin{cases} \mathcal{V}_{r_{ij}} - \mathcal{V}_{cutof} - (r_{ij} - r_{cutoff})\left(\frac{d\mathcal{V}(r_{ij})}{dr_{ij}}\right)_{r_{ij}=r_{cutoff}}, & r_{ij} \leq r_{cutoff} \\ 0, & r_{ij} > r_{cutoff} \end{cases} \qquad \text{A.18}$$

Although the force is going smoothly to zero, the potential itself differs from the original and consequently the model is essentially not the same. The properties computed for the model will differ, but with a perturbation scheme the properties of the atoms interacting with the "true" (original) potential can be retrieved [137, 338]. Nevertheless, the energy is conserved with the shifted-force potential.

Potential Switch

A method to avoid discontinuity in the energy and force is via a *switch function*. A switch function can be a smoothly decaying polynomial function of distance $S(r_{ij})$ that is multiplied to the original potential with the relation:

$$\mathcal{V}'(r_{ij}) = \mathcal{V}(r_{ij})S(r_{ij}) \qquad \text{A.19}$$

The switching function can be applied for the entire range of values of r_{ij}, but this affects the energy. Alternatively, the switch function is applied only to a small range of values of r_{ij}. The potential will have its original expression until $r_{ij} = r_{low}$. When $r_{low} \leq r_{ij} \leq r_{high}$ the potential switch will be applied to the original potential. With a careful choice of the switch function, meaning that is smoothly changes its value from 1 to 0 and its first and second derivatives at $r_{ij} = r_{low}$ and $r_{ij} = r_{high}$ are also 0, ensures that the force will also approach smoothly to 0 and the integration algorithm will perform the integration correctly [137].

Long-Range Interactions

The usage of cutoff distances to speed up the calculations of non-bonded interactions is good approximation for the short-range interactions, described by the Lennard-Jones potential, but it introduces numerical instabilities for the long-range interactions and serious inaccuracies, and it generally not recommended. The reason is that the Lennard-Jones decays as r^{-6} and the charge-charge interactions decays as r^{-1}. Interactions that decay no faster than r^{-n} can pose a problem in the computations since their range can often be greater than half the box length and scales as $\mathcal{O}(N^2)$. Ewald [347] developed a sum to study ionic crystals, taking advantage of their natural periodicity. His method, *"Ewald Summation Method"* can be also applied to quasi periodic systems that are generated through the Periodic Boundary Conditions. Assuming a cubic simulation box with an edge length L, and N charges. The system is periodic, and a particle will interact with every other particle in the central box, and all the other images, but not with itself. The charge-charge interactions can be expressed by:

$$\mathcal{V}_{el} = \frac{1}{2}\sum_{\boldsymbol{n}=0}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{4\pi\varepsilon_0\left|r_{ij}+\boldsymbol{n}\right|}\right) \qquad \text{A.20}$$

where $\boldsymbol{n}$, the cubic lattice points, $\boldsymbol{n} = \left(n_x L, n_y L, n_z L\right)$. When $|\boldsymbol{n}| = \boldsymbol{0}$, the terms with $i = j$ are omitted to avoid self-interactions. The principle behind the Ewald summation is that the Coulombic interactions are split in a "near"-field and a "far"-field term and the choice of $f(r)$ is crucial to deal large variations at small values of r, and slow decay at large values of r [137].

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r} \qquad \text{A.21}$$

Every charge is represented by a δ function and for every δ function, a Gaussian function of the opposite charge, and centered at the position of the point charge is introduced [341]. The Gaussian is of the form:

$$\rho_i(r) = \frac{q_i a^3}{\sqrt{\pi^3}} e^{-a^2 r^2} \tag{A.22}$$

where a is an arbitrary parameter that determines the width of the distribution. This distribution acts like an ionic cloud, screening the interactions between neighboring point charges [338]. The "near"-field term can be evaluated from the interaction between the screened point charges, by summing over all molecules in the central box and all images in the real space. The original point charge interactions can be recovered by compensating for the addition of those Gaussians, by adding a same shaped Gaussian of the opposite charge (same sign as the original point charge). The cancelling Gaussians are summed in the reciprocal space [338]. Below the final Ewald formula is shown. A detailed derivation of the final expression can be found in refs. [137, 338, 339].

$$\mathcal{V}_{el} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\begin{array}{l} \sum_{|\boldsymbol{n}|=0}^{\infty} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{erfc\left(a|\boldsymbol{r}_{ij}+\boldsymbol{n}|\right)}{|\boldsymbol{r}_{ij}+\boldsymbol{n}|} \\ + \sum_{k\neq 0} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{4\pi^2}{k^2} e^{-\left(\frac{k^2}{4a^2}\right)} cos\left(\boldsymbol{k}\cdot\boldsymbol{r}_{ij}\right) \\ - \frac{a}{\sqrt{\pi}} \sum_{k=1}^{N} \frac{q_k^2}{4\pi\varepsilon_0} + \frac{2\pi}{3L^3} \left|\sum_{k=1}^{N} \frac{q_k}{4\pi\varepsilon_0} r_k\right|^2 \end{array}\right. \tag{A.23}$$

The error function is given by the expression:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2}\, dt \tag{A.24}$$

The last correction term in the final expression is conditional and it depends on the medium that the simulation is performed at. It is needed of the medium is a vacuum and it is omitted if the medium is a conductor [137]. The Ewald summation method is quite an expensive implementation for the computation of long-range interactions, due to the reciprocal space part of the computation. It scales as $\mathcal{O}(N^2)$ and it can be modified with a varying value of a to scale as $\mathcal{O}(\sqrt{N^3})$, but this could introduce incompatibilities between the Coulomb and Van der Waals interaction ranges [137]. A method to speed up the computation was developed by Darden et. al [348], among others, which scales with $\mathcal{O}(N\, log\, N)$ and uses the fast Fourier transform for the summation of the reciprocal space terms. This is possible when the point charges are replaced by a charge-based distribution system (mesh) [137].

Statistical thermodynamics and MD

A Molecular Dynamics simulation generates information in the microscopic scale, such as atomic positions, velocities, and forces. The goal though is to extract macroscopic properties of the underlying system, such as temperature, pressure, compressibility, and many others. The transitions from the microscopic to the macroscopic scale is mediated via the principles of statistical thermodynamics. A very big difference between simulations and experiments is the definition of the average value of a property. A system of N particles can be described by 6N coordinates considering a three-dimensional space.

$$\boldsymbol{p}^N = (p_{1x}, p_{1y}, p_{1z}, p_{2x}, \dots, p_{Nz}) \quad \text{A.25}$$

$$\boldsymbol{r}^N = (r_{1x}, r_{1y}, r_{1z}, r_{2x}, \dots, r_{Nz}) \quad \text{A.26}$$

Assuming a property X, its instantaneous value can be written as $X\big(\boldsymbol{p}^N(t), \boldsymbol{r}^N(t)\big)$. The value of X fluctuates in time due to interactions between the components of the system. In experiments, the observable value of X that is measured, is an average of the instantaneous value of X during the time that the measurement takes place and is expressed by equation A.27 below [137].

$$X_{obs} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_{t=0}^{\tau} X\big(\boldsymbol{p}^N(t), \boldsymbol{r}^N(t)\big)\, dt \quad \text{A.27}$$

For a simulation then, it would be necessary to calculate the dynamic behavior of the system with a satisfying sampling over the phase space, to compute average values of properties. The integration of the equations of motion, using the force field potential energy to describe the interatomic interactions yields the trajectory with the time evolution of the positions, velocities and accelerations [137]. The limitation of the simulation's approach is that it cannot generate trajectories for systems in the macroscopic scale, in the order of moles. Even for a more manageable simulation system, equation A.27 cannot be truly extended to infinite time, but it can be satisfied with long enough sampling over a finite time t, represented by n_t number of time steps with the relation $t = n_t \times \delta t$ [338].

$$\langle X \rangle_{time} = \frac{1}{n_t} \sum_{t=0}^{n_t} X\big(\boldsymbol{p}^N(t), \boldsymbol{r}^N(t)\big) \quad \text{A.28}$$

Through statistical thermodynamics the expected value $\langle X \rangle$ through the time average, is replaced by the equivalent ensemble average. An MD simulation will generate a

thermodynamic ensemble of predetermined variables (N, V, T, P, E, μ, a) that will characterize the thermodynamic state of the particles comprising the system.
The ensemble average is given by equation A.29:

$$\langle X \rangle = \iint X(\boldsymbol{p}^N, \boldsymbol{r}^N)\rho(\boldsymbol{p}^N, \boldsymbol{r}^N)\, d\boldsymbol{p}^N d\boldsymbol{r}^N \qquad \text{A.29}$$

Microscopic coordinates and momenta of the particles are now coordinates in the phase space of *6N* dimensions, and are distributed by a probability density ρ [338]. Through the *ergodic hypothesis*, averages over a trajectory are equivalent to averages over the thermodynamic ensemble [349]. In this work molecular systems will be simulated in the NVT and NPT ensembles. Those are expressed by the partition functions below [338]:

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int e^{-\frac{\mathcal{H}(\boldsymbol{p}^N, \boldsymbol{r}^N)}{k_B T}}\, d\boldsymbol{p}^N d\boldsymbol{r}^N \qquad \text{A.30}$$

$$Q_{NPT} = \frac{1}{N!}\frac{1}{h^{3N}}\frac{1}{V_0} \int dV \int e^{-\frac{\mathcal{H}(\boldsymbol{p}^N, \boldsymbol{r}^N)+PV}{k_B T}}\, d\boldsymbol{p}^N d\boldsymbol{r}^N \qquad \text{A.31}$$

References

1. Klein, D.R., *Organic chemistry*. 2nd edition ed. 2013: Wiley.
2. Arunan, E., G.R. Desiraju, R.A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D.C. Clary, R.H. Crabtree, J.J. Dannenberg, and P. Hobza, *Defining the hydrogen bond: An account (IUPAC Technical Report).* Pure and Applied Chemistry, 2011. **83**(8): p. 1619-1636.
3. Arunan, E., G.R. Desiraju, R.A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D.C. Clary, R.H. Crabtree, J.J. Dannenberg, and P. Hobza, *Definition of the hydrogen bond (IUPAC Recommendations 2011).* Pure and applied chemistry, 2011. **83**(8): p. 1637-1641.
4. Hubbard, R.E. and M.K. Haider, *Hydrogen bonds in proteins: role and strength.* eLS, 2010.
5. Kollman, P.A. and L.C. Allen, *Theory of the hydrogen bond.* Chemical Reviews, 1972. **72**(3): p. 283-303.
6. Coulson, C., *THE HYDROGEN BOND: A review of the present interpretations and an introduction to the theoretical papers presented at the Congress.* Hydrogen bonding, 1959: p. 339-360.
7. Tsubomura, H., *The nature of the hydrogen-bond. I. The delocalization energy in the hydrogen-bond as calculated by the atomic-orbital method.* Bulletin of the Chemical Society of Japan, 1954. **27**(7): p. 445-450.
8. van der Lubbe, S.C. and C. Fonseca Guerra, *The nature of hydrogen bonds: A delineation of the role of different energy components on hydrogen bond strengths and lengths.* Chemistry–An Asian Journal, 2019. **14**(16): p. 2760-2769.
9. Hilbert, G., O. Wulf, S. Hendricks, and U. Liddel, *The hydrogen bond between oxygen atoms in some organic compounds.* Journal of the American Chemical Society, 1936. **58**(4): p. 548-555.
10. Badger, R.M. and S.H. Bauer, *Spectroscopic studies of the hydrogen bond. II. The shift of the O–H vibrational frequency in the formation of the hydrogen bond.* The Journal of Chemical Physics, 1937. **5**(11): p. 839-851.
11. Scheiner, S. and T. Kar, *Red-versus blue-shifting hydrogen bonds: are there fundamental distinctions?* The Journal of Physical Chemistry A, 2002. **106**(9): p. 1784-1789.
12. Konrat, R., M. Tollinger, G. Kontaxis, and B. Kräutler, *NMR techniques to study hydrogen bondingin aqueous solution.* Monatshefte für Chemie/Chemical Monthly, 1999. **130**(8): p. 961-982.

13. Prins, L.J., D.N. Reinhoudt, and P. Timmerman, *Noncovalent synthesis using hydrogen bonding.* Angewandte Chemie International Edition, 2001. **40**(13): p. 2382-2426.

14. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers: Original Research on Biomolecules, 1983. **22**(12): p. 2577-2637.

15. White, S.H. and W.C. Wimley, *Membrane protein folding and stability: physical principles.* Annu. Rev. Biomol. Struct., 1999. **28**: p. 319-365.

16. Wimley, W.C., K. Hristova, A.S. Ladokhin, L. Silvestro, P.H. Axelsen, and S.H.J.J.o.m.b. White, *Folding of β-sheet membrane proteins: a hydrophobic hexapeptide model1.* 1998. **277**(5): p. 1091-1110.

17. Ladokhin, A.S. and S.H.J.J.o.m.b. White, *Folding of amphipathic α-helices on membranes: energetics of helix formation by melittin1.* 1999. **285**(4): p. 1363-1369.

18. Wimley, W.C., S.H.J.N.S. White, and M. Biology, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* 1996. **3**(10): p. 842.

19. Luecke, H., B. Schobert, H.-T. Richter, J.-P. Cartailler, and J.K. Lanyi, *Structure of bacteriorhdopsin at 1.55 Å resolution.* J. Mol. Biol., 1999. **291**: p. 899-911.

20. Umena, Y., K. Kawakami, J.-R. Shen, and N. Kamiya, *Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å.* Nature, 2011. **473**: p. 55-60.

21. Bondar, A.-N. and S.H. White, *Hydrogen bond dynamics in membrane protein function.* Biochim. Biophys. Acta, 2012. **1818**: p. 942-950.

22. Bondar, A.-N. and H. Dau, *Extended protein/water H-bond networks in photosynthetic water oxidation.* Biochim. Biophys. Acta, 2012. **1817**: p. 1177-1190.

23. Trueman, S.F., E.C. Mandon, and R. Gilmore, *A gating motif in the translocation channel sets the hydrophobicity threshold for signal sequence function.* J. Cell Biol., 2012. **199**: p. 907-918.

24. Kretchmer, J.S., N. Boekelheide, J.J. Warren, J.R. Winkler, H.B. Gray, and T.F.J.P.o.t.N.A.o.S. Miller, *Fluctuating hydrogen-bond networks govern anomalous electron transfer kinetics in a blue copper protein.* 2018. **115**(24): p. 6129-6134.

25. Mitchell, J.B. and S.L. Price, *The nature of the N□ H... O□ C hydrogen bond: An intermolecular perturbation theory study of the formamide/formaldehyde complex.* Journal of computational chemistry, 1990. **11**(10): p. 1217-1233.

26. Ben-Tal, N., D. Sitkoff, I.A. Topol, A.-S. Yang, S.K. Burt, and B. Honig, *Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution.* The Journal of Physical Chemistry B, 1997. **101**(3): p. 450-457.

27. Sheu, S.-Y., D.-Y. Yang, H. Selzle, and E. Schlag, *Energetics of hydrogen bonds in peptides.* Proceedings of the National Academy of Sciences, 2003. **100**(22): p. 12683-12687.

28. Fersht, A.R., J.-P. Shi, J. Knill-Jones, D.M. Lowe, A.J. Wilkinson, D.M. Blow, P. Brick, P. Carter, M.M. Waye, and G. Winter, *Hydrogen bonding and biological specificity analysed by protein engineering.* Nature, 1985. **314**(6008): p. 235-238.

29. Davis, A.M. and S.J. Teague, *Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis.* Angewandte Chemie International Edition, 1999. **38**(6): p. 736-749.

30. Dinner, A.R., T. Lazaridis, and M. Karplus, *Understanding β-hairpin formation.* Proceedings of the National Academy of Sciences, 1999. **96**(16): p. 9068-9073.

31. Joh, N.H., A. Min, S. Faham, J.P. Whitelegge, D. Yang, V.L. Woods Jr., and J.U. Bowie, *Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins* Nature, 2008. **453**: p. 1266-1270.

32. Hong, H., G. Szabo, and L.K. Tamm, *Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening.* Nature chemical biology, 2006. **2**(11): p. 627-635.

33. Guerra, F., M. Siemers, C. Mielack, and A.-N. Bondar, *Dynamics of long-distance hydrogen-bond networks in photosystem II.* J. Phys. Chem. B, 2018. **122**: p. 4625-4641.

34. Lazaratos, M., K. Karathanou, and A.-N. Bondar, *Graphs of dynamic H-bond networks: from model proteins to protein complexes in cell signaling.* Current Opinion in Structural Biology, 2020. **64**: p. 79-87.

35. Bondar, A.-N. and J.C. Smith, *Protonation-state coupled conformational dynamics in reaction mechanisms of channel and pump rhodopsins.* Photochem. Photobiol., 2017. **93**: p. 1336-1344.

36. del Val, C., L. Bondar, and A.-N. Bondar, *Coupling between inter-helical hydrogen bonding and water dynamics in a proton transporter.* Journal of structural biology, 2014. **186**(1): p. 95-111.

37. Tanford, C., *Contribution of hydrophobic interactions to the stability of the globular conformation of proteins.* Journal of the American Chemical Society, 1962. **84**(22): p. 4240-4247.

38. Kellis, J.T., K. Nyberg, and A.R. Fersht, *Contribution of hydrophobic interactions to protein stability.* Nature, 1988. **333**(6175): p. 784-786.

39. Dill, K.A., *Dominant forces in protein folding.* Biochemistry, 1990. **29**(31): p. 7133-7155.

40. Nemethy, G., *Hydrophobic interactions.* Angewandte Chemie International Edition in English, 1967. **6**(3): p. 195-206.

41. Pace, C.N., H. Fu, K.L. Fryar, J. Landua, S.R. Trevino, B.A. Shirley, M.M. Hendricks, S. Iimura, K. Gajiwala, and J.M. Scholtz, *Contribution of hydrophobic interactions to protein stability.* Journal of molecular biology, 2011. **408**(3): p. 514-528.

42. Siemers, M., M. Lazaratos, K. Karathanou, F. Guerra, L.S. Brown, and A.-N. Bondar, *Bridge: A graph-based algorithm to analyze dynamic H-bond networks*

in membrane proteins. Journal of chemical theory and computation, 2019. **15**(12): p. 6781-6798.

43. Siemers, M. and A.-N. Bondar, *Interactive Interface for Graph-Based Analyses of Dynamic H-Bond Networks: Application to Spike Protein S.* Journal of Chemical Information and Modeling, 2021. **61**(6): p. 2998-3014.

44. Espinosa, E., E. Molins, and C. Lecomte, *Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities.* Chem. Phys. Lett., 1998. **285**(3-4): p. 170-173.

45. Adalsteinsson, H., A.H. Maulitz, and T.C.J.J.o.t.A.C.S. Bruice, *Calculation of the potential energy surface for intermolecular amide hydrogen bonds using semiempirical and ab initio methods.* 1996. **118**(33): p. 7689-7693.

46. Lipsitz, R.S., Y. Sharma, B.R. Brooks, and N.J.J.o.t.A.C.S. Tjandra, *Hydrogen bonding in high-resolution protein structures: a new method to assess NMR protein geometry.* 2002. **124**(35): p. 10621-10626.

47. Araya-Secchi, R., T. Perez-Acle, S.-g. Kang, T. Huynh, A. Bernardin, Y. Escalona, J.-A. Garate, A.D. Martínez, I.E. García, and J.C.J.B.j. Sáez, *Characterization of a novel water pocket inside the human Cx26 hemichannel structure.* 2014. **107**(3): p. 599-612.

48. Gowers, R.J., M. Linke, J. Barnoud, T.J. Reddy, M.N. Melo, S.L. Seyler, D.L. Dotson, J. Domanski, S. Buchoux, and I.M. Kenney. *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations*. in *Proceedings of the 15th Python in Science Conference*. 2016. SciPy.

49. Michaud-Agrawal, N., E.J. Denning, T.B. Woolf, and O. Beckstein, *MDAnalysis: a toolkit for the analysis of molecular dynamics simulations.* J. Comput. Chem., 2011. **32**(10): p. 2319-2327.

50. Gowers, R.J. and P. Carbone, *A multiscale approach to model hydrogen bonding: The case of polyamide.* J. Chem. Phys., 2015. **142**(22): p. 224907.

51. Durrant, J.D. and J.A. McCammon, *HBonanza: A computer algorithm for molecular-dynamics-trajectory hydrogen-bond analysis.* Journal of Molecular Graphics and Modelling, 2011. **31**: p. 5-9.

52. Humphrey, W., W. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J. Mol. Graph., 1996. **14**: p. 33-38.

53. Tiwari, A. and S.K. Panigrahi, *HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures.* In silico biology, 2007. **7**(6): p. 651-661.

54. Lindauer, K., C. Bendic, and J.J.B. Sühnel, *HBexplore—a new tool for identifying and analysing hydrogen bonding patterns in biological macromolecules.* 1996. **12**(4): p. 281-289.

55. Berman, H.M., J. Westbrook, G. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank.* Nucleic Acid Res., 2000. **28**: p. 235-242.

56. Bernstein, F.C., T.F. Koetzle, G.J. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *The Protein Data Bank: A computer-based archival file for macromolecular structures.* Eur. J. Biochem., 1977. **80**(2): p. 319-324.

57. Cornell, W.D., P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* J. Am. Chem. Soc., 1995. **117**(19): p. 5179-5197.

58. Alexander, R.W., J. Eargle, and Z.J.F.l. Luthey-Schulten, *Experimental and computational determination of tRNA dynamics.* 2010. **584**(2): p. 376-386.

59. Pyrkosz, A.B., J. Eargle, A. Sethi, and Z.J.J.o.m.b. Luthey-Schulten, *Exit strategies for charged tRNA from GluRS.* 2010. **397**(5): p. 1350-1371.

60. Sethi, A., J. Eargle, A.A. Black, and Z. Luthey-Schulten, *Dynamical network in tRNA:protein complexes.* Proc. Natl. Acad. Sci., 2009. **106**: p. 6620-6625.

61. Girvan, M. and M.E.J.P.o.t.n.a.o.s. Newman, *Community structure in social and biological networks.* 2002. **99**(12): p. 7821-7826.

62. Nagel, G., T. Szelles, W. Huhn, S. Kateriya, N. Adeishvili, P. Berthold, D. Ollig, P. Hegemann, and E. Bamberg, *Channelrhodopsin-2, a directly light-gated cation-selective membrane channel.* Proc. Natl. Acad. Sci., 2003. **100**: p. 13940-13945.

63. Nagel, G., T. Szelles, S. Kateriya, N. Adeishvili, P. Hegemann, and E. Bamberg, *Channelrhodopsins: directly light-gated cation channels.* Biochem. Soc. Trans., 2005. **33**: p. 863-866.

64. Foster, K.W., J. Saranak, N. Patel, G. Zarilli, M. Okabe, T. Kline, and K. Nakanishi, *A rhodopsin is the functional photoreceptor for phototaxis in the unicellular eukaryote Chlamydomonas.* Nature, 1984. **311**(5988): p. 756-759.

65. Sineshchekov, O.A., K.-H. Jung, and J.L. Spudich, *Two rhodopsins mediate phototaxis to low-and high-intensity light in Chlamydomonas reinhardtii.* Proceedings of the National Academy of Sciences, 2002. **99**(13): p. 8689-8694.

66. Hegemann, P., W. Gärtner, and R. Uhl, *All-trans retinal constitutes the functional chromophore in Chlamydomonas rhodopsin.* Biophysical journal, 1991. **60**(6): p. 1477-1489.

67. Lawson, M., D. Zacks, F. Derguini, K. Nakanishi, and J. Spudich, *Retinal analog restoration of photophobic responses in a blind Chlamydomonas reinhardtii mutant. Evidence for an archaebacterial like chromophore in a eukaryotic rhodopsin.* Biophysical journal, 1991. **60**(6): p. 1490-1498.

68. Takahashi, T., K. Yoshihara, M. Watanabe, M. Kubota, R. Johnson, F. Derguini, and K. Nakanishi, *Photoisomerization of retinal at 13-ene is important for phototaxis of Chlamydomonas reinhardtii: simultaneous measurements of phototactic and photophobic responses.* Biochemical and biophysical research communications, 1991. **178**(3): p. 1273-1279.

69. Richards, R. and R.E. Dempski, *Re-introduction of transmembrane serine residues reduce the minimum pore diameter of channelrhodopsin-2.* PloS one, 2012. **7**(11): p. e50018.

70. Schneider, F., C. Grimm, and P. Hegemann, *Biophysics of channelrhodopsin.* Annu. Rev. Biophys., 2015. **44**: p. 166-186.

71. Lin, J.Y., M.Z. Lin, P. Steinbach, and R.Y. Tsien, *Characterization of engineered channelrhodopsin variants with improved properties and kinetics.* Biophysical journal, 2009. **96**(5): p. 1803-1814.

72. Gadsby, D.C., *Ion channels versus ion pumps: the principal difference, in principle.* Nature Rev. Mol. Cell Biol., 2009. **10**: p. 344-352.

73. Jayaram, H., A. Accardi, F. Wu, C. Williams, and C. Miller, *Ion permeation through a Cl−-selective channel designed from a CLC Cl−/H+ exchanger.* Proceedings of the National Academy of Sciences, 2008. **105**(32): p. 11194-11199.

74. Feldbauer, K., D. Zimmermann, V. Pintschovius, J. Spitz, C. Bamann, and C. Bamann, *Channelrhodopsin-2 is a leaky proton pump.* Proc. Natl. Acad. Sci. USA, 2009. **106**: p. 12317-12322.

75. Nikolic, K., P. Degenaar, and C. Toumazou. *Modeling and engineering aspects of channelrhodopsin2 system for neural photostimulation*. in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. 2006. IEEE.

76. Nikolic, K., N. Grossman, M.S. Grubb, J. Burrone, C. Toumazou, and P. Degenaar, *Photocycles of channelrhodopsin-2.* Photochemistry and photobiology, 2009. **85**(1): p. 400-411.

77. Hegemann, P., S. Ehlenbeck, and D. Gradmann, *Multiple photocycles of channelrhodopsin.* Biophysical journal, 2005. **89**(6): p. 3911-3918.

78. Radu, I., C. Bamann, M. Nack, G. Nagel, E. Bamberg, and J. Heberle, *Conformational changes of channelrhodopsin-2.* J. Am. Chem. Soc., 2009. **131**: p. 7313-7319.

79. Ritter, E., K. Stehfest, A. Berndt, P. Hegemann, and F.J. Bartl, *Monitoring light-induced structural changes of channlerhodopsin-2 by UV-visible and Fourier transform infrared spectroscopy.* J. Biol. Chem., 2008. **283**: p. 35033-35041.

80. Lórenz-Fonfría, V.A., T. Resler, N. Krause, M. Nack, M. Gossing, G.F. von Mollard, C. Bamann, E. Bamberg, R. Schlesinger, and J. Heberle, *Transient protonation changes in channelrhodopsin-2 and their relevance to channel gating.* Proceedings of the National Academy of Sciences, 2013. **110**(14): p. E1273-E1281.

81. Bamann, C., T. Kirsch, G. Nagel, and E. Bamberg, *Spectral characteristics of the photocycle of channelrhodopsin-2 and its implication for channel function.* J. Mol. Biol., 2008. **375**: p. 986-994.

82. Nack, M., I. Radu, B.-J. Schultz, T. Resler, R. Schlessinger, A.-N. Bondar, C. del Val, S. Abbruzzetti, C. Viappiani, C. Bamann, E. Bamberg, and J. Heberle,

Kinetics of proton release and uptake by channelrhodopsin-2. FEBS Lett., 2012. **586**: p. 1344-1348.

83. Lórenz-Fonfría, V.A. and J. Heberle, *Channelrhodopsin unchained: structure and mechanism of a light-gated cation channel.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2014. **1837**(5): p. 626-642.

84. Verhoefen, M.-K., C. Bamann, R. Blöcher, U. Förster, E. Bamberg, and J. Wachtveitl, *The photocycle of channelrhodopsin-2: Ultrafast reaction dynamics and subsequent reaction steps.* ChemPhysChem, 2010. **11**: p. 3113-3122.

85. Bamann, C., R. Gueta, S. Kleinlogel, G. Nagel, and E. Bamberg, *Structural guidance of the photocycle of channelrhodopsin-2 by an interhelical hydrogen bond.* Biochemistry, 2010. **49**: p. 267-278.

86. Boyden, E.S., F. Zhang, E. Bamberg, G. Nagel, and K. Deisseroth, *Millisecond-timescale, genetically targeted optical control of neural activity.* Nature neuroscience, 2005. **8**(9): p. 1263-1268.

87. Nagel, G., M. Brauner, J.F. Liewald, N. Adeishvili, E. Bamberg, and A. Gottschalk, *Light activation of channelrhodopsin-2 in excitable cells of Caenorhabditis elegans triggers rapid behavioral responses.* Current Biology, 2005. **15**(24): p. 2279-2284.

88. Li, X., D.V. Gutierrez, M.G. Hanson, J. Han, M.D. Mark, H. Chiel, P. Hegemann, L.T. Landmesser, and S. Herlitze, *Fast noninvasive activation and inhibition of neural and network activity by vertebrate rhodopsin and green algae channelrhodopsin.* Proceedings of the National Academy of Sciences, 2005. **102**(49): p. 17816-17821.

89. Bi, A., J. Cui, Y.-P. Ma, E. Olshevskaya, M. Pu, A.M. Dizhoor, and Z.-H. Pan, *Ectopic expression of a microbial-type rhodopsin restores visual responses in mice with photoreceptor degeneration.* Neuron, 2006. **50**(1): p. 23-33.

90. Zhang, F., L.-P. Wang, E.S. Boyden, and K. Deisseroth, *Channelrhodopsin-2 and optical control of excitable cells.* Nature methods, 2006. **3**(10): p. 785-792.

91. Zhang, Y.-P. and T.G. Oertner, *Optical induction of synaptic plasticity using a light-sensitive channel.* Nature methods, 2007. **4**(2): p. 139-141.

92. Zhang, F., L.-P. Wang, M. Brauner, J.F. Liewald, K. Kay, N. Watzke, P.G. Wood, E. Bamberg, G. Nagel, and A. Gottschalk, *Multimodal fast optical interrogation of neural circuitry.* Nature, 2007. **446**(7136): p. 633-639.

93. Zhang, F., M. Prigge, F. Beyrière, S.P. Tsunoda, J. Mattis, O. Yizhar, P. Hegemann, and K. Deisseroth, *Red-shifted optogenetic excitation: a tool for fast neural control derived from Volvox carteri.* Nature neuroscience, 2008. **11**(6): p. 631-633.

94. Lagali, P.S., D. Balya, G.B. Awatramani, T.A. Münch, D.S. Kim, V. Busskamp, C.L. Cepko, and B. Roska, *Light-activated channels targeted to ON bipolar cells restore visual function in retinal degeneration.* Nature neuroscience, 2008. **11**(6): p. 667-675.

95. Gradinaru, V., M. Mogri, K.R. Thompson, J.M. Henderson, and K. Deisseroth, *Optical deconstruction of parkinsonian neural circuitry.* science, 2009. **324**(5925): p. 354-359.

96. Lief Fenno, O.Y. and K. Deisseroth, *The development and application of optogenetics.* Annual review of neuroscience, 2011. **34**: p. 389.

97. Volkov, O., K. Kovalev, V. Polovinkin, V. Borschchevskiy, C. Bamann, R. Astashkin, E. Marin, A. Popov, T. Balandin, D. Willbold, G. Büldt, E. Bamberg, and V. Gordeliy, *Structural insights into ion conduction by channelrhodopsin 2.* Science, 2017. **358**: p. eaaan8862.

98. Müller, M., C. Bamann, E. Bamberg, and W. Kühlbrandt, *Projection structure of channelrhodopsin-2 at 6Å resolution by electron crystallography.* J. Mol. Biol., 2011. **414**: p. 86-95.

99. Kato, H.E., F. Zhang, O. Yizhar, C. Ramakhrishnan, T. Nishizawa, K. Hirata, J. Ito, Y. Aita, T. Tsukazaki, S. Hayashi, P. Hegemann, A.D. Maturana, R. Ishitani, K. Deisseroth, and O. Nureki, *Crystal structure of the channelrhodopsin light-gated cation channel.* Nature, 2012. **482**: p. 369-374.

100. del Val, C., J. Royuela-Flor, S. Milenkovic, and A.-N. Bondar, *Channelrhodopsins: a bioinformatics perspective.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2014. **1837**(5): p. 643-655.

101. Eisenhauer, K., J. Kuhne, E. Ritter, A. Berndt, S. Wolf, E. Freier, F. Bartl, P. Hegemann, and K. Gerwert, *In channelrhodopsin-2 Glu-90 is crucial for ion selectivity and is deprotonated during the photocycle.* J. Biol. Chem., 2012. **287**: p. 6904-6911.

102. Plazzo, A.P., N. de Franceschi, F. da Broi, F. Zonta, M.F. Sanasi, F. Filippini, and M. Mongillo, *Bioinformatic and mutational analysis of channelrhodopsin-2 cation conductance pathway.* J. Biol. Chem. , 2012. **287**: p. 4818-4825.

103. Watanabe, H., K. Welke, F. Schneider, S. Tsunoda, F. Zhang, K. Deisseroth, P. Hegemann, and M. Elstner, *Structural model of channelrhodopsin.* J. Biol. Chem., 2012. **287**: p. 7456-7466.

104. Schrödinger, L., *The PyMol Molecular Graphics System, Version 1.8.* 2015.

105. Lemmon, M.A., H.R. Treutlein, P.D. Adams, A.T. Brünger, and D.M. Engelman, *A dimerization motif for transmembrane α–helices.* Nature structural biology, 1994. **1**(3): p. 157-163.

106. Senes, A., M. Gerstein, and D.M. Engelman, *Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions.* Journal of molecular biology, 2000. **296**(3): p. 921-936.

107. Russ, W.P. and D.M. Engelman, *The GxxxG motif: a framework for transmembrane helix-helix association.* Journal of molecular biology, 2000. **296**(3): p. 911-919.

108. Arbely, E. and I.T. Arkin, *Experimental Measurement of the Strength of a Cα−H⊙⊙⊙ O Bond in a Lipid Bilayer.* Journal of the American Chemical Society, 2004. **126**(17): p. 5362-5363.

109. Kleiger, G., R. Grothe, P. Mallick, and D. Eisenberg, *GXXXG and AXXXA: common α-helical interaction motifs in proteins, particularly in extremophiles.* Biochemistry, 2002. **41**(19): p. 5990-5997.

110. Arbely, E., Z. Granot, I. Kass, J. Orly, and I.T. Arkin, *A trimerizing GxxxG motif is uniquely inserted in the severe acute respiratory syndrome (SARS) coronavirus spike protein transmembrane domain.* Biochemistry, 2006. **45**(38): p. 11349-11356.

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

112. Kolbe, M., H. Besir, L.-O. Essen, and D. Oesterhelt, *Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution.* Science, 2000. **288**(5470): p. 1390-1396.

113. Richardson, J.S., *β-Sheet topology and the relatedness of proteins.* Nature, 1977. **268**(5620): p. 495-500.

114. Mitchell, E.M., P.J. Artymiuk, D.W. Rice, and P. Willett, *Use of techniques derived from graph theory to compare secondary structure motifs in proteins.* Journal of Molecular Biology, 1990. **212**(1): p. 151-166.

115. Etter, M.C., J.C. MacDonald, and J. Bernstein, *Graph-set analysis of hydrogen-bond patterns in organic crystals.* Acta Crystallographica Section B: Structural Science, 1990. **46**(2): p. 256-262.

116. Etter, M.C., *Encoding and decoding hydrogen-bond patterns of organic compounds.* Accounts of Chemical Research, 1990. **23**(4): p. 120-126.

117. Gray, T.M. and B.W. Matthews, *Intrahelical hydrogen bonding of serine, threonine and cysteine residues within a-helices and its relevance to membrane-bound proteins.* J. Mol. Biol., 1984. **175**: p. 75-81.

118. Ballesteros, J.A., X. Deupi, M. Olivella, E.E. Haaksma, and L. Pardo, *Serine and threonine residues bend α-helices in the χ1= g− conformation.* Biophysical journal, 2000. **79**(5): p. 2754-2760.

119. del Val, C., S.H. White, and A.-N. Bondar, *Ser/Thr motifs in transmembrane proteins: conservation patterns and effects on local protein structure and dynamics.* J. Membr. Biol., 2012. **245**: p. 717-730.

120. del Val, C., L. Bondar, and A.-N. Bondar, *Coupling between inter-helical hydrogen bonding and water dynamics in a proton transporter.* J. Struct. Biol., 2014. **186**: p. 95-111.

121. Bondar, A.-N., *Proton-Binding Motifs of Membrane-Bound Proteins: From Bacteriorhodopsin to Spike Protein S.* Frontiers in Chemistry, 2021. **9**.

122. Blundell, T., D. Barlow, N. Borkakoti, and J. Thornton, *Solvent-induced distortions and the curvature of α-helices.* Nature, 1983. **306**(5940): p. 281-283.

123. Lanyi, J.K., *Bacteriorhodopsin.* Internat. Rev. Cytol., 1999. **187**: p. 161-202.

124. Metz, G., F. Siebert, and M. Engelhardt, *Asp85 is the only internal aspartic acid that gets protonated in the M intermediate and the purple-to-blue transition of bacteriorhodopsin. A solid-state 13C CP-MAS NMR investigation.* FEBS Lett., 1992. **303**: p. 237-241.

125. Bondar, A.-N., M. Elstner, S. Suhai, J.C. Smith, and S. Fischer, *Mechanism of primary proton transfer in bacteriorhodopsin.* Structure, 2004. **12**: p. 1281-1288.

126. Nango, E., A. Royant, M. Kubo, T. Nakane, C. Wickstrand, T. Kimura, T. Tanaka, K. Tono, C. Song, R. Tanaka, T. Arima, A. Yamashita, J. Kobayashi, T. Hosaka, E. Mizohata, P. Nogly, M. Sugahara, D. Nam, T. Nomura, T. Shimamura, D. Im, T. Fujiwara, Y. Yamanaka, B. Jeon, T. Nishizawa, K. Oda, M. Fukuda, R. Andersson, P. Båth, P. Dods, J. Davidsson, S. Matsuoka, S. Kawatake, M. Murata, O. Nureki, S. Owada, T. Kameshima, T. Hatsui, Y. Joti, G. Schertler, M. Yabashi, A.-N. Bondar, J. Standfuss, R. Neutze, and S. Iwata, *A three-dimensional movie of structural changes in bacteriorhodopsin.* Science, 2016. **354**: p. 1552-1557.

127. Ni, Q.Z., T.V. Can, D. E., M. Belenky, R.G. Griffin, and J. Herzfeld, *primary transfer step in the light-driven ion pump bacteriorhodopsin: An irreversible U-turn revealed by dynamic nuclear polarization-enhanced magic angle spinning NMR.* J. Am. Chem. Soc., 2018. **140**: p. 4085-4091.

128. Bondar, A.-N. and M.J. Lemieux, *Reactions at biomembrane interfaces.* Chemical reviews, 2019. **119**(9): p. 6162-6183.

129. Khademi, S., J. O'Connell, J. Remis, Y. Robles-Colmenares, L.J. Miercke, and R.M. Stroud, *Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å.* Science, 2004. **305**(5690): p. 1587-1594.

130. Eriksson, U.K., G. Fischer, R. Friemann, G. Enkavi, E. Tajkhorshid, and R. Neutze, *Subangstrom resolution X-ray structure details aquaporin-water interactions.* Science, 2013. **340**(6138): p. 1346-1349.

131. Lin, S.-M., J.-Y. Tsai, C.-D. Hsiao, Y.-T. Huang, C.-L. Chiu, M.-H. Liu, J.-Y. Tung, T.-H. Liu, R.-L. Pan, and Y.-J. Sun, *Crystal structure of a membrane-embedded H+-translocating pyrophosphatase.* Nature, 2012. **484**(7394): p. 399-403.

132. Luecke, H., H.-T. Richter, and J.K. Lanyi, *Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution.* Science, 1998. **280**(5371): p. 1934-1937.

133. Alder, B.J. and T.E. Wainwright, *Studies in molecular dynamics. I. General method.* The Journal of Chemical Physics, 1959. **31**(2): p. 459-466.

134. Lindorff-Larsen, K., P. Maragakis, S. Piana, and D.E. Shaw, *Picosecond to millisecond structural dynamics in human ubiquitin.* The Journal of Physical Chemistry B, 2016. **120**(33): p. 8313-8320.

135. Shaw, D.E., M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, and J.C. Chao, *Anton, a special-purpose machine for molecular dynamics simulation.* Communications of the ACM, 2008. **51**(7): p. 91-97.

136. Huang, J. and A.D. MacKerell Jr, *CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data.* Journal of computational chemistry, 2013. **34**(25): p. 2135-2145.

137. Leach, A.R. and A.R. Leach, *Molecular modelling: principles and applications*. 2001: Pearson education.

138. MacKerell Jr., A.D., D. Bashford, M. Bellot, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E.I. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J. Phys. Chem. B, 1998. **102**: p. 3586-3616.

139. Feller, S.E. and A.D. MacKerell Jr., *An improved empirical potential energy function for molecular simulations of phospholipids.* J. Phys. Chem. B, 2000. **104**: p. 7510-7515.

140. MacKerell Jr, A.D., B. Brooks, C.L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, *CHARMM: the energy function and its parameterization.* Encyclopedia of computational chemistry, 2002. **1**.

141. MacKerell Jr., A.D., M. Feig, and C.L.I. Brooks, *Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.* J. Comput. Chem, 2004. **25**: p. 1400-1415.

142. MacKerell Jr., A.D., *Empirical force fields for biological molecules: overviews and issues.* J. Comput. Chem, 2004. **25**: p. 1584-1604.

143. MacKerell Jr, A.D., M. Feig, and C.L. Brooks, *Improved treatment of the protein backbone in empirical force fields.* Journal of the American Chemical Society, 2004. **126**(3): p. 698-699.

144. Buck, M., S. Bouguet-Bonnet, R.W. Pastor, and A.D. MacKerell Jr., *Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lyzozyme.* Biophys. J., 2006. **15**: p. L36-L38.

145. Brooks, B.R., C.L. Brooks III, A.D. Mackerell Jr, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, and S. Boresch, *CHARMM: the biomolecular simulation program.* Journal of computational chemistry, 2009. **30**(10): p. 1545-1614.

146. Klauda, J.B., R.M. Venable, J.A. Freites, J.W. O'Connor, D.J. Tobias, C. Mondragon-Ramirez, I. Votrobyov, A.D. MacKerell Jr., and R.W. Pastor, *Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types.* J. Phys. Chem. B, 2010. **114**: p. 7830-7843.

147. Best, R.B., X. Zhu, J. Shim, P.E. Lopes, J. Mittal, M. Feig, and A.D. MacKerell Jr, *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles.* Journal of chemical theory and computation, 2012. **8**(9): p. 3257-3273.

148. Li, X., S.A. Hassan, and E.L. Mehler, *Long dynamics simulations of proteins using atomistic force fields and a continuum representation of solvent effects: calculation of structural and dynamic properties.* Proteins: Structure, Function, and Bioinformatics, 2005. **60**(3): p. 464-484.

149. MacKerell Jr, A.D., N. Banavali, and N. Foloppe, *Development and current status of the CHARMM force field for nucleic acids.* Biopolymers: Original Research on Biomolecules, 2000. **56**(4): p. 257-265.

150. Guvench, O., S.S. Mallajosyula, E.P. Raman, E. Hatcher, K. Vanommeslaeghe, T.J. Foster, F.W. Jamison, and A.D. MacKerell Jr, *CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate–protein modeling.* Journal of chemical theory and computation, 2011. **7**(10): p. 3162-3180.

151. Vanommeslaeghe, K., E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, and I. Vorobyov, *CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields.* Journal of computational chemistry, 2010. **31**(4): p. 671-690.

152. Babitzki, G., R. Denschlag, and P. Tavan, *Polarization effects stabilize bacteriorhodopsin's chromophore binding pocket: a molecular dynamics study.* J. Phys. Chem. B, 2009. **113**: p. 10483-10495.

153. Kandt, C., J. Schlitter, and K. Gerwert, *Dynamics of water molecules in the bacteriorhodopsin trimer in explicit lipid/water environment.* Biophysical journal, 2004. **86**(2): p. 705-717.

154. Baudry, J., E. Tajkhorshid, F. Molnar, J. Phillips, and K. Schulten, *Molecular dynamics study of bacteriorhodopsin and the purple membrane*. 2001, ACS Publications. p. 905-918.

155. Grudinin, S., G. Büldt, V.I. Gordeliy, and A. Baumgartner, *Water molecules and hydrogen-bonded networks in bacteriorhodopsin - Molecular dynamics simulations of the ground state and the M-intermediate.* Biophys. J., 2005. **88**: p. 3252-3261.

156. Hayashi, S.E., E. Tajkhorshid, and K. Schulten, *Structural changes during the formation of early intermediates in the bacteriorhodopsin photocycle.* Biophys. J., 2002. **83**: p. 1281-1297.

157. Nina, M., B. Roux, and J.C. Smith, *Functional interactions in bacteriorhodopsin: a theoretical analysis of retinal hydrogen bonding with water.* Biophys. J., 1995. **68**: p. 25-39.

158. Tajkhorshid, E., J. Baudry, K. Schulten, and S. Suhai, *Molecular dynamics study of the nature and origin of retinal's twisted structure in bacteriorhodopsin.* Biophys. J., 2000. **78**: p. 683-693.

159. Jardon-Valadez, E., A.-N. Bondar, and D.J. Tobias, *Coupling of retinal, protein, and water dynamics in squid rhodopsin.* Biophys. J., 2010. **99**: p. 2200-2207.

160. Roux, B., M. Nina, R. Pomes, and J.C. Smith, *Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study.* Biophys. J., 1996. **71**: p. 670-681.

161. Bondar, A.-N., M. Knapp-Mohammady, S. Suhai, S. Fischer, and J.C. Smith, *Ground-state properties of the retinal molecule: from quantum mechanical to classical mechanical computations of retinal proteins.* Theor. Chem. Acc., 2011. **130**: p. 1169-1183.

162. Baudry, J., S. Crouzy, B. Roux, and J.C. Smith, *Simulation analysis of the retinal conformational equilibrium in dark-adapted bacteriorhodopsin.* Biophysical journal, 1999. **76**(4): p. 1909-1917.

163. Luecke, H., B. Schobert, H.-T. Richter, J.-P. Cartailler, and J.K. Lanyi, *Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution.* Science, 1999. **286**(5438): p. 255-260.

164. Tajkhorshid, E., B. Paizs, and S. Suhai, *Role of isomerization barriers in the pKa control of the retinal Schiff base: a density functional study.* J. Phys. Chem. B, 1999. **103**: p. 4518-4527.

165. Tajkhorshid, E. and S. Suhai, *The effect of the protein environment on the structure and charge distribution of the retinal Schiff base in bacteriorhodopsin.* Theoretical Chemistry Accounts, 1999. **101**(1): p. 180-185.

166. Paizs, B., E. Tajkhorshid, and S. Suhai, *Electronic effects on the ground-state rotational barrier of polyene Schiff bases: a molecular orbital study.* The Journal of Physical Chemistry B, 1999. **103**(25): p. 5388-5395.

167. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein, *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**: p. 926-935.

168. Berweger, C.D., W.F. van Gunsteren, and F. Müller-Plathe, *Force field parametrization by weak coupling. Re-engineering SPC water.* Chemical physics letters, 1995. **232**(5-6): p. 429-436.

169. Berendsen, H., J. Grigera, and T. Straatsma, *The missing term in effective pair potentials.* Journal of Physical Chemistry, 1987. **91**(24): p. 6269-6271.

170. Neria, E., S. Fischer, and M. Karplus, *Simulation of activation free energies in molecular systems.* J. Chem. Phys., 1996. **105**: p. 1902-1921.

171. Brooks, B.R., R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem, 1983. **4**: p. 187-217.

172. Mark, P. and L. Nilsson, *Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K.* J. Phys. Chem. A, 2001. **105**(43): p. 9954-9960.

173. Mills, R., *Self-diffusion in normal and heavy water in the range 1-45.deg.* J. Phys. Chem., 1973. **77**(5): p. 685-688.

174. Price, W.S., H. Ide, and Y. Arata, *Self-diffusion of supercooled water to 238 K using PGSE NMR diffusion measurements.* The Journal of Physical Chemistry A, 1999. **103**(4): p. 448-450.

175. Pettitt, B.M. and M. Karplus, *The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach.* Chemical physics letters, 1985. **121**(3): p. 194-201.

176. Dehmer, M. and F. Emmert-Streib, *Analysis of complex networks: from biology to linguistics*. 2009: John Wiley & Sons.

177. Borgatti, S.P., *Centrality and network flow.* Social networks, 2005. **27**(1): p. 55-71.

178. Borgatti, S.P. and M.G. Everett, *A graph-theoretic perspective on centrality.* Social networks, 2006. **28**(4): p. 466-484.

179. Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, 1977: p. 35-41.

180. Freeman, L.C., *Centrality in social networks conceptual clarification.* Social networks, 1978. **1**(3): p. 215-239.

181. Hagberg, A., P. Swart, and D. S Chult, *Exploring network structure, dynamics, and function using NetworkX*. 2008, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

182. Brandes, U., *A faster algorithm for betweenness centrality.* Journal of mathematical sociology, 2001. **25**(2): p. 163-177.

183. Nieminen, J., *On the centrality in a graph.* Scandinavian journal of psychology, 1974. **15**(1): p. 332-336.

184. Karathanou, K. and A.-N. Bondar, *Using Graphs of Dynamic Hydrogen-Bond Networks To Dissect Conformational Coupling in a Protein Motor.* J. Chem. Inf. Model., 2019.

185. Harris, A., M. Lazaratos, M. Siemers, E. Watt, A. Hoang, S. Tomida, L. Schubert, M. Saita, J. Heberle, Y. Furutani, H. Kandori, A.-N. Bondar, and L.S. Brown, *Mechanism of inward proton transport in an antarctic microbial rhodopsin.* The Journal of Physical Chemistry B, 2020. **124**(24): p. 4851-4872.

186. Karathanou, K., M. Lazaratos, É. Bertalan, M. Siemers, K. Buzar, G.F. Schertler, C. Del Val, and A.-N. Bondar, *A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2.* Journal of structural biology, 2020. **212**(2): p. 107617.

187. Jo, S., T. Kim, and W. Im, *Automated builder and database of protein/membrane complexes for molecular dynamics simulations.* PloS one, 2007. **2**(9): p. e880.

188. Jo, S., T. Kim, V.G. Iyer, and W. Im, *CHARMM-GUI: a web-based graphical user interface for CHARMM.* Journal of Computational Chemistry, 2008. **29**: p. 1859-1865.

189. Jo, S., J.B. Lim, J.B. Klauda, and W. Im, *CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes.* Biophysical journal, 2009. **97**(1): p. 50-58.

190. Wu, E.L., X. Cheng, S. Jo, H. Rui, K.C. Song, E.M. Dávila-Contreras, Y. Qi, J. Lee, V. Monje-Galvan, R.M. Venable, J.B. Klauda, and W. Im, *CHARMM-GUI*

Membrane Builder toward realistic biological membrane simulations. J. Comput. Chem, 2014. **35**: p. 1997-2004.

191. Jo, S., X. Cheng, S.M. Islam, L. Huang, H. Rui, A. Zhu, H.S. Lee, Y. Qi, W. Han, and K. Vanommeslaeghe, *CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues.* Advances in protein chemistry and structural biology, 2014. **96**: p. 235-265.

192. Kalé, L., R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten, *NAMD2: greater scalability for parallel molecular dynamics.* J. Comput. Phys., 1999. **151**: p. 283-312.

193. Phillips, J.C., B. Braun, W. Wang, J. Gumbart, E. Takjkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten, *Scalable molecular dynamics with NAMD.* J. Comput. Chem, 2005. **26**: p. 1781-1802.

194. Kubo, R., M. Toda, and N. Hashitsume, *Statistical physics II: nonequilibrium statistical mechanics*. Vol. 31. 2012: Springer Science & Business Media.

195. Brünger, A., C.L. Brooks III, and M. Karplus, *Stochastic boundary conditions for molecular dynamics simulations of ST2 water.* Chemical physics letters, 1984. **105**(5): p. 495-500.

196. Martyna, G.J., D.J. Tobias, and M.L. Klein, *Constant-pressure molecular-dynamics algorithms.* J. Chem. Phys., 1994. **101**: p. 4177-4189.

197. Feller, S.E., Y. Zhang, R.W. Pastor, and B. Brooks, *Constant pressure molecular dynamics simulation: The Langevin piston method.* J. Chem. Phys. , 1995. **103**: p. 4613-4621.

198. Essmann, U., L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen, *A smooth particle mesh Ewald method.* The Journal of chemical physics, 1995. **103**(19): p. 8577-8593.

199. Bennett, L., B. Melchers, and B. Proppe, *Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin.* 2020.

200. Stone, J.E., *An Efficient Library For Parallel Ray Tracing And Animation*. 1998, University of Missouri.

201. Lomize, M., I.D. Pogozheva, H. Joo, H.I. Mosberg, and A.L. Lomize, *OPM database and PPM server: resources for positioning of proteins in membranes.* Nucleic Acid Research, 2011. **40**: p. D370-D376.

202. Lomize, A.L., I.D. Pogozheva, and 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, 2011. **51**(4): p. 930-946.

203. Hessa, T., H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, and G. von Heijne, *Recognition of transmembrane helices by the endoplasmic reticulum translocon.* Nature, 2005. **433**: p. 377-381.

204. Moon, C.P. and K.G. Fleming, *Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers.* Proc. Natl. Acad. Sci. USA, 2011. **108**: p. 10174-10177.

205. Maneewongvatana, S. and D.M. Mount, *Analysis of approximate nearest neighbor searching with clustered point sets.* Data Structures, Near Neighbor Searches, and Methodology, 2002. **59**: p. 105-123.

206. Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.

207. Siwick, B.J., M.J. Cox, and H.J. Bakker, *Long-range proton transfer in aqueous acid-base reactions.* J. Phys. Chem. B, 2008. **112**: p. 378-389.

208. Bondar, A.-N., S. Fischer, J.C. Smith, M. Elstner, and S. Suhai, *Key role of electrostatic interactions in bacteriorhodopsin proton transfer.* J. Am. Chem. Soc., 2004. **126**: p. 14668-14677.

209. Bondar, A.-N., J.C. Smith, and S. Fischer, *Structural and energetic determinants of primary proton transfer in bacteriorhodopsin.* Photochem. Photobiol. Sci., 2006. **5**: p. 547-552.

210. Bondar, A.-N., J. Baudry, S. Suhai, S. Fischer, and J.C. Smith, *Key role of active-site water molecules in bacteriorhodopsin proton-transfer reactions.* J. Phys. Chem. B, 2008. **112**: p. 14729-14741.

211. Fiser, A., R.K. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Science, 2000. **9**: p. 1753-1773.

212. Marti-Renom, M.A., A. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali, *Comparative protein structure modeling of genes and genomes.* Annu. Rev. Biomol. Struct., 2000. **29**: p. 291-325.

213. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J. Mol. Biol., 1993. **234**: p. 779-815.

214. Ito, S., H.E. Kato, R. Taniguchi, T. Iwata, O. Nureki, and H. Kandori, *Water-containing hydrogen-bonding network in the active center of channelrhodopsin.* J. Am. Chem. Soc., 2014. **136**: p. 3475-3482.

215. Kuhne, J., K. Eisenhauer, E. Ritter, P. Hegemann, K. Gerwert, and F. Bartl, *Early formation of the ion-conducting pore in channelrhodopsin-2.* Angew. Chem. Int. Ed., 2015. **54**: p. 4953-4957.

216. Takemoto, M., H.E. Kato, M. Koyama, J. Ito, M. Kamiya, S. Hayashi, A.D. Maturana, K. Deisseroth, R. Ishitani, and O. Nureki, *Molecular dynamics of channelrhodopsin at the early stages of channel opening.* Plos One, 2015: p. 1-15.

217. Nack, M., I. Radu, M. Gossing, C. Bamann, E. Bamberg, G.F. von Mollard, and J. Heberle, *The DC gate in channelrhodopsin-2: crucial hydrogen bonding interaction between C128 and D156.* Photochem. Photobiol. Sci., 2010. **9**: p. 194-198.

218. Lorenz-Fonfria, V.A., T. Resler, N. Krause, M. Nack, M. Gossing, G.F. von Mollard, C. Bamann, E. Bamberg, R. Schlesinger, and J. Heberle, *Transient*

protonation changes in channelrhodpsin-2 and their relevance to channel gating. Proc. Natl. Acad. Sci. USA, 2013. **110**: p. 1273-1281.

219. Adam, S. and A.-N. Bondar, *Mechanism by which water and protein electrostatic interactions control proton transfer at the active site of channelrhodopsin.* PLOS ONE, 2018. **13**(8): p. e0201298.

220. Dioumaev, A.K., H.-T. Richter, L.S. Brown, M. Tanio, S. Tuzi, H. Saito, Y. Limura, R. Needleman, and J.K. Lanyi, *Existence of a proton transfer chain in bacteriorhodopsin: participation of Glu-194 in the release of protons to the extracellular side.* Biochemistry, 1998. **37**: p. 2496-2506.

221. Richards, R. and R.E. Dempski, *Adjacent channelrhodopsin-2 residues within transmembranes 2 and 7 regulate cation selectivity and distribution of the two open states.* J. Biol. Chem., 2017. **292**: p. 7314-7326.

222. Watanabe, S., T. Ishizuka, S. Hososhima, A. Zamani, M.R. Hoque, and H. Yawo, *The regulatory mechanism of ion permeation through a channelrhodopsin derived from Mesostigma viride (MvChR1).* Photochem. Photobiol. Sci., 2016. **15**: p. 365-374.

223. Gunaydin, L.A., O. Yizhar, A. Berndt, V.S. Sohal, K. Deisseroth, and P. Hegemann, *Ultrafast optogenetic control.* Nature Neurosci., 2010. **13**: p. 387-392.

224. Lin, J.Y.J.E.p., *A user's guide to channelrhodopsin variants: features, limitations and future developments.* 2011. **96**(1): p. 19-25.

225. Li, H., E.G. Govorunova, O.A. Sineschchekov, and E.N. Spudich, *Role of a helix B lysine residue in the photoactive site in channelrhodopsins.* Biophys. J., 2014. **106**: p. 1607-1617.

226. Lazaratos, M., M. Siemers, L.S. Brown, and A.-N. Bondar, *Conserved hydrogen-bond motifs of membrane transporters and receptors.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2022. **1864**(6): p. 183896.

227. Brown, L.S., J. Sasaki, H. Kandori, A. Maeda, R. Needleman, and J.K. Lanyi, *Glutamic acid 204 is the terminal release group at the extracellular surface of bacteriorhodopsin.* J. Biol. Chem., 1995. **270**: p. 27122-27126.

228. Garczareck, F., L.S. Brown, J.K. Lanyi, and K. Gerwert, *Proton binding within a membrane protein by a protonated water cluster.* Proc. Natl. Acad. Sci. USA, 2005. **102**: p. 3633-3638.

229. Bondar, A.-N. and J.C. Smith, *Water molecules in short- and long-distance proton transfer steps of bacteriorhodopsin proton pumping.* 2009.

230. Watanabe, H., K. Welke, D.J. Sindhikara, P. Hegemann, and M. Elstner, *Towards an understanding of channelrhodopsin-2 function: simulations lead to novel insights of the channel mechanism.* J. Mol. Biol., 2013. **425**: p. 1795-1814.

231. Inoue, K., S.P. Tsunoda, M. Singh, S. Tomida, S. Hososhima, M. Konno, R. Nakamura, H. Watanabe, P.-A. Bulzu, and H.L. Banciu, *Schizorhodopsins: A family of rhodopsins from Asgard archaea that function as light-driven inward H+ pumps.* Science advances, 2020. **6**(15): p. eaaz2441.

232. Ugalde, J.A., S. Podell, P. Narasingarao, and E.E. Allen, *Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria.* Biology direct, 2011. **6**(1): p. 1-8.

233. Pushkarev, A., K. Inoue, S. Larom, J. Flores-Uribe, M. Singh, M. Konno, S. Tomida, S. Ito, R. Nakamura, and S.P. Tsunoda, *A distinct abundant group of microbial rhodopsins discovered using functional metagenomics.* Nature, 2018. **558**(7711): p. 595-599.

234. Inoue, S., S. Yoshizawa, Y. Nakajima, K. Kojima, T. Tsukamoto, T. Kikukawa, and Y. Sudo, *Spectroscopic characteristics of Rubricoccus marinus xenorhodopsin (Rm XeR) and a putative model for its inward H+ transport mechanism.* Physical Chemistry Chemical Physics, 2018. **20**(5): p. 3172-3183.

235. Kawanabe, A., Y. Furutani, K.-H. Jung, and H. Kandori, *Engineering an inward proton transport from a bacterial sensor rhodopsin.* J. Am. Chem. Soc., 2009. **131**: p. 16439-16444.

236. Dong, B., L. Sánchez-Magraner, and H. Luecke, *Structure of an inward proton-transporting Anabaena sensory rhodopsin mutant: mechanistic insights.* Biophys. J., 2016. **111**: p. 963-972.

237. Govorunova, E.G., O.A. Sineshchekov, H. Li, and J.L. Spudich, *Microbial rhodopsins: diversity, mechanisms, and optogenetic applications.* Annual review of biochemistry, 2017. **86**: p. 845-872.

238. Kelley, L.A., S. Mezulis, C.M. Yates, M.N. Wass, and M.J.E. Sternberg, *The Phyre2 web portal for protein modeling, prediction and analysis.* Nature Protocols, 2015. **10**: p. 845.

239. Kouyama, T., S. Kanada, Y. Takeguchi, A. Narusawa, M. Murakami, and K. Ihara, *Crystal structure of the light-driven chloride pump halorhodopsin from Natronomonas pharaonis.* Journal of molecular biology, 2010. **396**(3): p. 564-579.

240. Gordeliy, V.I., J. Labahn, R. Moukhametzianov, R. Efremov, J. Granzin, R. Schlesinger, G. Büldt, T. Savopol, A.J. Scheidig, and J.P. Klare, *Molecular basis of transmembrane signalling by sensory rhodopsin II–transducer complex.* Nature, 2002. **419**(6906): p. 484-487.

241. Wada, T., K. Shimono, T. Kikukawa, M. Hato, N. Shinya, S.Y. Kim, T. Kimura-Someya, M. Shirouzu, J. Tamogami, and S. Miyauchi, *Crystal structure of the eukaryotic light-driven proton-pumping rhodopsin, Acetabularia rhodopsin II, from marine alga.* Journal of molecular biology, 2011. **411**(5): p. 986-998.

242. Gmelin, W., K. Zeth, R. Efremov, J. Heberle, J. Tittor, and D. Oesterhelt, *The crystal structure of the L1 intermediate of halorhodopsin at 1.9 Å resolution.* Photochemistry and photobiology, 2007. **83**(2): p. 369-377.

243. Furuse, M., J. Tamogami, T. Hosaka, T. Kikukawa, N. Shinya, M. Hato, N. Ohsawa, S.Y. Kim, K.-H. Jung, and M. Demura, *Structural basis for the slow photocycle and late proton release in Acetabularia rhodopsin I from the marine plant Acetabularia acetabulum.* Acta Crystallographica Section D: Biological Crystallography, 2015. **71**(11): p. 2203-2216.

244. Söding, J., *Protein homology detection by HMM–HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-960.

245. Giorgino, T., *Computing 1-D atomic densities in macromolecular simulations: The density profile tool for VMD.* Comput. Phys. Commun., 2014. **185**(1): p. 317-322.

246. Otto, H., T. Marti, M. Holz, T. Mogi, M. Lindau, H.G. Khorana, and M.P. Heyn, *Aspartic acid-96 is the internal proton donor in the reprotonation of the Schiff base of bacteriorhodopsin.* Proc. Natl. Acad. Sci. USA, 1989. **86**: p. 9228-9232.

247. Butt, H.J., K. Fendler, E. Bamberg, J. Tittor, and D. Oesterhelt, *Aspartic acids 96 and 85 play a central role in the function of bacteriorhodopsin as a proton pump.* EMBO J., 1989. **8**: p. 1657-1663.

248. Gerwert, K., B. Hess, J. Soppa, and D. Oesterhelt, *Role of aspartate-96 in proton translocation by bacteriorhodopsin.* Proc. Natl. Acad. Sci. USA, 1989. **86**: p. 4943-4947.

249. Holz, M., L.A. Drachev, T. Mogi, H. Otto, A.D. Kaulen, M.P. Heyn, V.P. Skulachev, and H.G. Khorana, *Replacement of aspartic acid-96 by asparagine in bacteriorhodopsin slows both the decay of the M intermediate and the associated proton movement.* Proceedings of the National Academy of Sciences, 1989. **86**(7): p. 2167-2171.

250. Inoue, K., S. Ito, Y. Kato, Y. Nomura, M. Shibata, T. Uchihashi, S.P. Tsunoda, and H. Kandori, *A natural light-driven inward proton pump.* Nature Communications, 2016. **7**(1): p. 1-10.

251. Shi, L., S.R. Yoon, A.G. Bezerra Jr, K.-H. Jung, and L.S. Brown, *Cytoplasmic shuttling of protons in Anabaena sensory rhodopsin: implications for signaling mechanism.* Journal of molecular biology, 2006. **358**(3): p. 686-700.

252. Siemers, M., M. Lazaratos, K. Karathanou, F. Guerra, L.S. Brown, and A.N. Bondar, *Bridge: A Graph-Based Algorithm to Analyze Dynamic H-Bond Networks in Membrane Proteins.* J Chem Theory Comput, 2019. **15**(12): p. 6781-6798.

253. Hasegawa, N., H. Jonotsuka, K. Miki, and K. Takeda, *X-ray structure analysis of bacteriorhodopsin at 1.3 Å resolution.* Scientific reports, 2018. **8**(1): p. 1-8.

254. Van Den Berg, B., A. Chembath, D. Jefferies, A. Basle, S. Khalid, and J.C. Rutherford, *Structural basis for Mep2 ammonium transceptor activation by phosphorylation.* Nature communications, 2016. **7**(1): p. 1-11.

255. Juarez, J.F.B., P.J. Judge, S. Adam, D. Axford, J. Vinals, J. Birch, T.O. Kwan, K.K. Hoi, H.-Y. Yen, and A. Vial, *Structures of the archaerhodopsin-3 transporter reveal that disordering of internal water networks underpins receptor sensitization.* Nature communications, 2021. **12**(1): p. 1-10.

256. Gordeliy, V.I., J. Labahn, R. Moukhametzianov, R. Efremov, J. Granzin, R. Schlesinger, G. Büldt, A.J. Scheidig, J.P. Klare, and M. Engelhardt, *Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex.* Nature, 2002. **419**: p. 484-487.

257. Kovalev, K., R. Astashkin, I. Gushchin, P. Orekhov, D. Volkov, E. Zinovev, E. Marin, M. Rulev, A. Alekseev, and A. Royant, *Molecular mechanism of light-driven sodium pumping.* Nature communications, 2020. **11**(1): p. 1-11.

258. Kovalev, K., D. Volkov, R. Astashkin, A. Alekseev, I. Gushchin, J.M. Haro-Moreno, I. Chizhov, S. Siletsky, M. Mamedov, and A. Rogachev, *High-resolution structural insights into the heliorhodopsin family.* Proceedings of the National Academy of Sciences, 2020. **117**(8): p. 4131-4141.

259. Saier Jr, M.H., C.V. Tran, and R.D. Barabote, *TCDB: the Transporter Classification Database for membrane transport protein analyses and information.* Nucleic acids research, 2006. **34**(suppl_1): p. D181-D186.

260. Vasiliauskaité-Brooks, I., R. Sounier, P. Rochaix, G. Bellot, M. Fortier, F. Hoh, L. De Colibus, C. Bechara, E.M. Saied, and C. Arenz, *Structural insights into adiponectin receptors suggest ceramidase activity.* Nature, 2017. **544**(7648): p. 120-123.

261. Bertalan, É., S. Lešnik, U. Bren, and A.-N. Bondar, *Protein-water hydrogen-bond networks of G protein-coupled receptors: Graph-based analyses of static structures and molecular dynamics.* Journal of Structural Biology, 2020. **212**(3): p. 107634.

262. Nack, M., I. Radu, C. Bamann, E. Bamberg, and J. Heberle, *The retinal structure of channelrhodopsin-2 assessed by resonance Raman spectroscopy.* FEBS letters, 2009. **583**(22): p. 3676-3680.

263. Ritter, E., P. Piwowarski, P. Hegemann, and F.J. Bartl, *Light-dark adaptation of channelrhodopsin C128T mutant.* Journal of Biological Chemistry, 2013. **288**(15): p. 10451-10458.

264. Enami, N., K. Yoshimura, M. Murakami, H. Okumura, K. Ihara, and T. Kouyama, *Crystal Structures of Archaerhodopsin-1 and -2: Common Structural Motif in Archaeal Light-driven Proton Pumps.* Journal of Molecular Biology, 2006. **358**(3): p. 675-685.

265. Kouyama, T., R. Fujii, S. Kanada, T. Nakanishi, S.K. Chan, and M. Murakami, *Structure of archaerhodopsin-2 at 1.8 A resolution.* Acta Crystallographica Section D, 2014. **70**(10): p. 2692-2701.

266. Bada Juarez, J.F., P.J. Judge, S. Adam, D. Axford, J. Vinals, J. Birch, T.O.C. Kwan, K.K. Hoi, H.-Y. Yen, A. Vial, P.-E. Milhiet, C.V. Robinson, I. Schapiro, I. Moraes, and A. Watts, *Structures of the archaerhodopsin-3 transporter reveal that disordering of internal water networks underpins receptor sensitization.* Nature Communications, 2021. **12**(1): p. 629.

267. Shevchenko, V., I. Gushchin, V. Polovinkin, E. Round, V. Borshchevskiy, P. Utrobin, A. Popov, T. Balandin, G. Bueldt, and V. Gordeliy, *Crystal structure of Escherichia coli-expressed Haloarcula marismortui bacteriorhodopsin I in the trimeric form.* PloS one, 2014. **9**(12): p. e112873.

268. Hsu, M.-F., H.-Y. Fu, C.-J. Cai, H.-P. Yi, C.-S. Yang, and A.H.-J. Wang, *Structural and functional studies of a newly grouped Haloquadratum walsbyi bacteriorhodopsin reveal the acid-resistant light-driven proton pumping activity.* Journal of Biological Chemistry, 2015. **290**(49): p. 29567-29577.

269. Chan, S.K., T. Kitajima-Ihara, R. Fujii, T. Gotoh, M. Murakami, K. Ihara, and T. Kouyama, *Crystal structure of cruxrhodopsin-3 from Haloarcula vallismortis.* PloS one, 2014. **9**(9): p. e108362.

270. Chan, S.K., H. Kawaguchi, H. Kubo, M. Murakami, K. Ihara, K. Maki, and T. Kouyama, *Crystal structure of the 11-cis isomer of Pharaonis halorhodopsin: structural constraints on interconversions among different isomeric states.* Biochemistry, 2016. **55**(29): p. 4092-4104.

271. Nakanishi, T., S. Kanada, M. Murakami, K. Ihara, and T. Kouyama, *Large deformation of helix F during the photoreaction cycle of Pharaonis halorhodopsin in complex with azide.* Biophysical journal, 2013. **104**(2): p. 377-385.

272. Fudim, R., M. Szczepek, J. Vierock, A. Vogt, A. Schmidt, G. Kleinau, P. Fischer, F. Bartl, P. Scheerer, and P. Hegemann, *Design of a light-gated proton channel based on the crystal structure of Coccomyxa rhodopsin.* Science signaling, 2019. **12**(573).

273. Gushchin, I., P. Chervakov, P. Kuzmichev, A.N. Popov, E. Round, V. Borshchevskiy, A. Ishchenko, L. Petrovskaya, V. Chupin, and D.A. Dolgikh, *Structural insights into the proton pumping by unusual proteorhodopsin from nonmarine bacteria.* Proceedings of the National Academy of Sciences, 2013. **110**(31): p. 12631-12636.

274. Hayashi, T., S. Yasuda, K. Suzuki, T. Akiyama, K. Kanehara, K. Kojima, M. Tanabe, R. Kato, T. Senda, and Y. Sudo, *How does a microbial rhodopsin RxR realize its exceptionally high thermostability with the proton-pumping function being retained?* The Journal of Physical Chemistry B, 2020. **124**(6): p. 990-1000.

275. Gushchin, I., V. Shevchenko, V. Polovinkin, K. Kovalev, A. Alekseev, E. Round, V. Borschchevskiy, T. Balandin, A. Popov, T. Gensch, C. Fahlke, C. Bamann, D. Willbold, G. Büldt, E. Bamberg, and V. Gordeliy, *Crystal structure of a light-driven sodium pump.* Nature Struct. Mol. Biol., 2015. **22**: p. 390-395.

276. Kato, H.E., K. Inoue, R. Abe-Yoshizumi, Y. Kato, H. Ono, M. Konno, S. Hososhima, T. Ishizuka, M.R. Hoque, H. Kunimoto, J. Ito, S. Yoshizawa, K. Yamashita, M. Takemoto, T. Nishizawa, R. Taniguchi, K. Kogure, A.D. Maturana, Y. Iino, H. Yawo, R. Ishitani, H. Kandori, and O. Nureki, *Structural basis for Na^+ transport mechanism by a light-driven Na^+ pump.* Nature, 2015. **521**: p. 48-53.

277. Varma, N., E. Mutt, J. Mühle, V. Panneels, A. Terakita, X. Deupi, P. Nogly, G.F. Schertler, and E. Lesca, *Crystal structure of jumping spider rhodopsin-1 as a light sensitive GPCR.* Proceedings of the National Academy of Sciences, 2019. **116**(29): p. 14547-14556.

278. Schulz, S., M. Iglesias-Cans, A. Krah, Ö. Yildiz, V. Leone, D. Matthies, G.M. Cook, J.D. Faraldo-Gómez, and T. Meier, *A New Type of Na+-Driven ATP Synthase Membrane Rotor with a Two-Carboxylate Ion-Coupling Motif.* PLOS Biology, 2013. **11**(6): p. e1001596.

279. Meier, T., A. Krah, P.J. Bond, D. Pogoryelov, K. Diederichs, and J.D. Faraldo-Gómez, *Complete Ion-Coordination Structure in the Rotor Ring of Na+-*

Dependent F-ATP Synthases. Journal of Molecular Biology, 2009. **391**(2): p. 498-507.

280. Matthies, D., W. Zhou, A.L. Klyszejko, C. Anselmi, Ö. Yildiz, K. Brandt, V. Müller, J.D. Faraldo-Gómez, and T. Meier, *High-resolution structure and mechanism of an F/V-hybrid rotor ring in a Na+-coupled ATP synthase.* Nature Communications, 2014. **5**(1): p. 5286.

281. Murata, T., I. Yamato, Y. Kakinuma, A.G.W. Leslie, and J.E. Walker, *Structure of the Rotor of the V-Type Na⁺-ATPase from Enterococcus hirae.* Science, 2005. **308**(5722): p. 654-659.

282. Gamal El-Din, T.M., M.J. Lenaeus, K. Ramanadane, N. Zheng, and W.A. Catterall, *Molecular dissection of multiphase inactivation of the bacterial sodium channel NaVAb.* Journal of General Physiology, 2019. **151**(2): p. 174-185.

283. Sula, A., J. Booker, L.C. Ng, C.E. Naylor, P.G. DeCaen, and B.A. Wallace, *The complete structure of an activated open sodium channel.* Nature communications, 2017. **8**(1): p. 1-9.

284. Wang, S., E.A. Orabi, S. Baday, S. Bernèche, and G. Lamoureux, *Ammonium transporters achieve charge transfer by fragmenting their substrate.* Journal of the American Chemical Society, 2012. **134**(25): p. 10419-10427.

285. Hanson, M.A., C.B. Roth, E. Jo, M.T. Griffith, F.L. Scott, G. Reinhart, H. Desale, B. Clemons, S.M. Cahalan, and S.C. Schuerer, *Crystal structure of a lipid G protein–coupled receptor.* Science, 2012. **335**(6070): p. 851-855.

286. Chrencik, J.E., C.B. Roth, M. Terakado, H. Kurata, R. Omi, Y. Kihara, D. Warshaviak, S. Nakade, G. Asmar-Rovira, and M. Mileni, *Crystal structure of antagonist bound human lysophosphatidic acid receptor 1.* Cell, 2015. **161**(7): p. 1633-1643.

287. Liang, Y.-L., M.J. Belousoff, M.M. Fletcher, X. Zhang, M. Khoshouei, G. Deganutti, C. Koole, S.G. Furness, L.J. Miller, and D.L. Hay, *Structure and dynamics of adrenomedullin receptors AM1 and AM2 reveal key mechanisms in the control of receptor phenotype by receptor activity-modifying proteins.* ACS pharmacology & translational science, 2020. **3**(2): p. 263-284.

288. Vasiliauskaité-Brooks, I., R.D. Healey, P. Rochaix, J. Saint-Paul, R. Sounier, C. Grison, T. Waltrich-Augusto, M. Fortier, F. Hoh, and E.M. Saied, *Structure of a human intramembrane ceramidase explains enzymatic dysfunction found in leukodystrophy.* Nature communications, 2018. **9**(1): p. 1-13.

289. Vogeley, L., O.A. Sineshchekov, V.D. Trivedi, J. Sasaki, J.L. Spudich, and H. Luecke, *Anabaena sensory rhodopsin: a photochromic color sensor at 2.0 Å* Science, 2004. **306**: p. 1390-1393.

290. Pflüger, T., C.F. Hernández, P. Lewe, F. Frank, H. Mertens, D. Svergun, M.W. Baumstark, V.Y. Lunin, M.S. Jetten, and S.L. Andrade, *Signaling ammonium across membranes through an ammonium sensor histidine kinase.* Nature communications, 2018. **9**(1): p. 1-11.

291. Hub, J.S. and B.L. De Groot, *Mechanism of selectivity in aquaporins and aquaglyceroporins.* Proceedings of the National Academy of Sciences, 2008. **105**(4): p. 1198-1203.

292. Eriksson, U.K., G. Fischer, R. Friemann, G. Enkavi, E. Tajkhorshid, and R. Neutze, *Subangstrom resolution X-ray structure details aquaporin-water interactions.* Science, 2013. **340**: p. 1346-1349.

293. Sui, H., B.-G. Han, J.K. Lee, P. Walian, and B.K. Jap, *Structural basis of water-specific transport through the AQP1 water channel.* Nature, 2001. **414**(6866): p. 872-878.

294. Ardevol, A. and G. Hummer, *Retinal isomerization and water-pore formation in channelrhodopsin-2.* Proceedings of the National Academy of Sciences, 2018. **115**(14): p. 3557-3562.

295. Welke, K., H.C. Watanabe, T. Wolter, M. Gaus, and M. Elstner, *QM/MM simulations of vibrational spectra of bacteriorhodopsin and channelrhodopsin-2.* Phys. Chem. Chem. Phys., 2013. **15**: p. 6651-6659.

296. Guo, Y., F.E. Beyle, B.M. Bold, H.C. Watanabe, A. Koslowski, W. Thiel, P. Hegemann, M. Marazzi, and M. Elstner, *Active site structure and absorption spectrum of channelrhodopsin-2 wild-type and C128T mutant.* Chemical Science, 2016. **7**(6): p. 3879-3891.

297. Schobert, B., J. Cupp-Vickery, V. Hornak, S.O. Smith, and J.K. Lanyi, *Crystallographic structure of the K intermediate of bacteriorhodopsin: conservation of free energy after photoisomerization of the retinal.* J. Mol. Biol., 2002. **321**: p. 715-726.

298. Belrhali, H., P. Nollert, A. Royant, C. Menzel, J.P. Rosenbusch, E.M. Landau, and E. Pebay-Peyroula, *Protein, lipid and water organization in bacteriorhodopsin crystals: a molecular view of the purple membrane at 1.9Å resolution.* Structure, 1999. **7**: p. 909-917.

299. Matsui, Y., K. Sakai, M. Murakami, Y. Shiro, S. Adachi, H. Okumura, and T. Kouyama, *Specific damage induced by X-ray radiation and structural changes in the primary photoreaction of bacteriorhodopsin.* J. Mol. Biol., 2002. **324**: p. 469-481.

300. Kouyama, T., T. Nishikawa, T. Tokuhisa, and H. Okumura, *Crystal structure of the L intermediate of bacteriorhodopsin: evidence for vertical translocation of a water molecule during the proton pumping cycle.* J. Mol. Biol., 2004. **335**: p. 531-546.

301. Gottschalk, M., N.A. Dencher, and B. Halle, *Microsecond exchange of internal water molecules in bacteriorhodopsin.* Journal of Molecular Biology, 2001. **311**: p. 605-621.

302. Wietek, J., S. Wiegert, N. Adeishvili, F. Schneider, H. Watanabe, S.P. Tsunoda, A. Vogt, M. Elstner, T.G. Oertner, and P. Hegemann, *Conversion of channelrhodopsin into a light-gated chloride channel.* Science, 2014. **344**: p. 409-412.

303. VanGordon, M.R., G. Gayawali, S.W. Rick, and S.B. Rempe, *Atomistic study of intramolecular interactions in the closed-state channelrhodopsin chimera, C1C2.* Biophys. J., 2017. **112**: p. 943-952.

304. Qiu, F., S. Rebolledo, C. Gonzalez, and H.P. Larsson, *Subunit interactions during cooperative opening of voltage-gated proton channels.* Neuron, 2013. **77**(2): p. 288-298.

305. Perez-Flores, M.C., J.H. Lee, S. Park, X.-D. Zhang, C.-R. Sihn, H.A. Ledford, W. Wang, H.J. Kim, V. Timofeyev, and V. Yarov-Yarovoy, *Cooperativity of Kv7. 4 channels confers ultrafast electromechanical sensitivity and emergent properties in cochlear outer hair cells.* Science advances, 2020. **6**(15): p. eaba1104.

306. Tokaji, Z., *Dimeric-like kinetic cooperativity of the bacteriorhodopsin molecules in purple membranes.* Biophysical journal, 1993. **65**(3): p. 1130-1134.

307. Clatot, J., M. Hoshi, X. Wan, H. Liu, A. Jain, K. Shinlapawittayatorn, C. Marionneau, E. Ficker, T. Ha, and I. Deschênes, *Voltage-gated sodium channels assemble and gate as dimers.* Nature communications, 2017. **8**(1): p. 1-14.

308. Bondar, A.-N., *Mechanisms of long-distance allosteric couplings in proton-binding membrane transporters.* Advances in Protein Chemistry and Structural Biology, 2022. **128**: p. 199-239.

309. Horton, P. and A. Ruban, *Regulation of photosystem II.* Photosynthesis Research, 1992. **34**(3): p. 375-385.

310. Barber, J., *Photosystem II: the engine of life.* Quarterly reviews of biophysics, 2003. **36**(1): p. 71-89.

311. McEvoy, J.P. and G.W. Brudvig, *Water-splitting chemistry of photosystem II.* Chemical reviews, 2006. **106**(11): p. 4455-4483.

312. Dau, H. and M. Haumann, *The manganese complex of photosystem II in its reaction cycle—basic framework and possible realization at the atomic level.* Coordination Chemistry Reviews, 2008. **252**(3-4): p. 273-295.

313. Umena, Y., K. Kawakami, J.-R. Shen, and N. Kamiya, *Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å.* Nature, 2011. **473**(7345): p. 55-60.

314. Dau, H., I. Zaharieva, and M. Haumann, *Recent developments in research on water oxidation by photosystem II.* Current opinion in chemical biology, 2012. **16**(1-2): p. 3-10.

315. Klauss, A., M. Haumann, and H. Dau, *Alternating electron and proton transfer steps in photosynthetic water oxidation.* Proceedings of the National Academy of Sciences, 2012. **109**(40): p. 16035-16040.

316. Ibrahim, M., R. Chatterjee, J. Hellmich, R. Tran, M. Bommer, V.K. Yachandra, J. Yano, J. Kern, and A. Zouni, *Improvements in serial femtosecond crystallography of photosystem II by optimizing crystal uniformity using microseeding procedures.* Structural Dynamics, 2015. **2**(4): p. 041705.

317. Bommer, M., A.-N. Bondar, A. Zouni, H. Dobbek, and H. Dau, *Crystallographic and computational analysis of the barrel part of the PsbO protein of photosystem II: Carboxylate–water clusters as putative proton transfer relays and structural switches.* Biochemistry, 2016. **55**(33): p. 4626-4635.

318. Hussein, R., M. Ibrahim, R. Chatterjee, L. Coates, F. Müh, V.K. Yachandra, J. Yano, J. Kern, H. Dobbek, and A. Zouni, *Optimizing crystal size of photosystem II by macroseeding: Toward neutron protein crystallography.* Crystal growth & design, 2018. **18**(1): p. 85-94.

319. Zaharieva, I. and H. Dau, *Energetics and kinetics of S-state transitions monitored by delayed chlorophyll fluorescence.* Frontiers in plant science, 2019. **10**: p. 386.

320. Kemmler, L., M. Ibrahim, H. Dobbek, A. Zouni, and A.-N. Bondar, *Dynamic water bridging and proton transfer at a surface carboxylate cluster of photosystem II.* Physical Chemistry Chemical Physics, 2019. **21**(45): p. 25449-25466.

321. Capaldi, R.A., F. Malatesta, and V. Darley-Usmar, *Structure of cytochrome c oxidase.* Biochimica et Biophysica Acta (BBA)-Reviews on Bioenergetics, 1983. **726**(2): p. 135-148.

322. Capaldi, R.A., *Structure and function of cytochrome c oxidase.* Annual review of biochemistry, 1990. **59**(1): p. 569-596.

323. Michel, H., J. Behr, A. Harrenga, and A. Kannt, *Cytochrome c oxidase: structure and spectroscopy.* Annual review of biophysics and biomolecular structure, 1998. **27**(1): p. 329-356.

324. Heitbrink, D., H. Sigurdson, C. Bolwien, P. Brzezinski, and J. Heberle, *Transient binding of CO to CuB in cytochrome c oxidase is dynamically linked to structural changes around a carboxyl group: a time-resolved step-scan Fourier transform infrared investigation.* Biophysical journal, 2002. **82**(1): p. 1-10.

325. Ataka, K., F. Giess, W. Knoll, R. Naumann, S. Haber-Pohlmeier, B. Richter, and J. Heberle, *Oriented attachment and membrane reconstitution of His-tagged cytochrome c oxidase to a gold electrode: in situ monitoring by surface-enhanced infrared absorption spectroscopy.* Journal of the American Chemical Society, 2004. **126**(49): p. 16199-16206.

326. Ataka, K., B. Richter, and J. Heberle, *Orientational control of the physiological reaction of cytochrome c oxidase tethered to a gold electrode.* The Journal of Physical Chemistry B, 2006. **110**(18): p. 9339-9347.

327. Hrabakova, J., K. Ataka, J. Heberle, P. Hildebrandt, and D.H. Murgida, *Long distance electron transfer in cytochrome c oxidase immobilised on electrodes. A surface enhanced resonance Raman spectroscopic study.* Physical Chemistry Chemical Physics, 2006. **8**(6): p. 759-766.

328. Kim, Y.C. and G. Hummer, *Proton-pumping mechanism of cytochrome c oxidase: A kinetic master-equation approach.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2012. **1817**(4): p. 526-536.

329. Kirchberg, K., H. Michel, and U. Alexiev, *Net proton uptake is preceded by multiple proton transfer steps upon electron injection into cytochrome c oxidase.* Journal of Biological Chemistry, 2012. **287**(11): p. 8187-8193.

330. Kirchberg, K., H. Michel, and U. Alexiev, *Exploring the entrance of proton pathways in cytochrome c oxidase from Paracoccus denitrificans: Surface charge, buffer capacity and redox-dependent polarity changes at the internal surface.* Biochimica et Biophysica Acta (BBA)-Bioenergetics, 2013. **1827**(3): p. 276-284.

331. Sezer, M., P. Kielb, U. Kuhlmann, H. Mohrmann, C. Schulz, D. Heinrich, R. Schlesinger, J. Heberle, and I.M. Weidinger, *Surface enhanced resonance Raman spectroscopy reveals potential induced redox and conformational changes of cytochrome c oxidase on electrodes.* The Journal of Physical Chemistry B, 2015. **119**(30): p. 9586-9591.

332. Wolf, A., C. Schneider, T.-Y. Kim, K. Kirchberg, P. Volz, and U. Alexiev, *A simulation-guided fluorescence correlation spectroscopy tool to investigate the protonation dynamics of cytochrome c oxidase.* Physical Chemistry Chemical Physics, 2016. **18**(18): p. 12877-12885.

333. Wolf, A., J. Wonneberg, J. Balke, and U. Alexiev, *Electronation-dependent structural change at the proton exit side of cytochrome c oxidase as revealed by site-directed fluorescence labeling.* The FEBS Journal, 2020. **287**(6): p. 1232-1246.

334. Bruun, S., H. Naumann, U. Kuhlmann, C. Schulz, K. Stehfest, and P. Hegemann, *The chromophore structure of the long-lived intermediate of the C128T channelrhodopsin-2 variant.* FEBS Lett., 2011. **585**: p. 3998-4001.

335. Krause, N., C. Engelhard, J. Heberle, R. Schlesinger, and R. Bittl, *Structural differences between the closed and open states of channelrhodopsin-2 as observed by EPR spectroscopy.* FEBS Lett., 2013. **587**: p. 3309-3313.

336. Bertalan, E.v., E. Lesca, G.F. Schertler, and A.-N. Bondar, *C-Graphs tool with graphical user interface to dissect conserved hydrogen-bond networks: Applications to visual rhodopsins.* Journal of Chemical Information and Modeling, 2021. **61**(11): p. 5692-5707.

337. Borschchevskiy, V., d.E.S. Roun, A.N. Popov, G. Büldt, and V. Gordelyi, *X-ray-radioation-induced changes in bacteriorhodopsin structure.* J. Mol. Biol., 2011. **409**: p. 813-825.

338. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids*. 2017: Oxford university press.

339. Rapaport, D.C., *The art of molecular dynamics simulation*. 2004: Cambridge university press.

340. Frenkel, D. and B. Smit, *Understanding molecular simulation: from algorithms to applications*. Vol. 1. 2001: Elsevier.

341. Jensen, F., *Introduction to computational chemistry*. 2017: John wiley & sons.

342. Verlet, L., *Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules.* Physical review, 1967. **159**(1): p. 98.

343. Hockney, R.W., *The potential calculation and some applications.* Methods Comput. Phys., 1970. **9**: p. 136.

344. Potter, D., *Computational physics.* 1973.

345. Van Gunsteren, W.F. and H.J. Berendsen, *A leap-frog algorithm for stochastic dynamics.* Molecular Simulation, 1988. **1**(3): p. 173-185.

346. Swope, W.C., H.C. Andersen, P.H. Berens, and K.R. Wilson, *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters.* The Journal of chemical physics, 1982. **76**(1): p. 637-649.

347. Ewald, P.P., *Die Berechnung optischer und elektrostatischer Gitterpotentiale.* Annalen der physik, 1921. **369**(3): p. 253-287.

348. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems.* The Journal of chemical physics, 1993. **98**(12): p. 10089-10092.

349. Tuckerman, M.E. and G.J. Martyna, *Understanding modern molecular dynamics: Techniques and applications.* The Journal of Physical Chemistry B, 2000. **104**(2): p. 159-178.

Selbstständigkeitserklärung

Name: Lazaratos
Vorname: Michail

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht. Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Datum: 14.06.2022 Unterschrift: _____