De novo design of membrane protein channels

Thesis

by

CRISTINA-MIHAELA CROITORU

Doctor of Philosophy

ASTON UNIVERSITY

March

2024

©Cristina-Mihaela Croitoru, 2024

Cristina-Mihaela Croitoru asserts her moral right to be identified as the author of this thesis.

This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright belongs to its author and that no quotation from the thesis and no information derived from it may be published without appropriate permission or acknowledgement.

Keywords*:* membrane protein, *de novo* rationale protein design, computational biology, molecular dynamics simulations, membrane protein folding, biocompatibility, *E. coli de novo* membrane protein expression, bimolecular fluorescence complementarity, 3D printing, droplet-interface bilayers

Thesis summary

De novo design of membrane protein channels

Cristina-Mihaela Croitoru

Doctor of Philosophy

ASTON UNIVERSITY

2024

Advances in the field of synthetic biology and *de novo* protein design come in the aid of the existing methods, and aim to contribute with providing essential answers referring to the sequence-structure-function relationship problem. Advances in the implementation of computational techniques in these research fields, promoted the speed of research, however, the computer-only based studies cannot provide sufficient data, especially due to insufficient real-life-based training information. For this reason, the need of interdisciplinary studies and combining computational techniques with laboratory-based experiments are required.

De novo protein design aims to contribute to the design of small building blocks, which can self-associate into known or new structures, stabilising existing scaffolds. Thus, it enables the possibility of creating a library of small building blocks and their influence on structure formation and stability.

In this cross-disciplinary study, we aimed to contribute to *de novo* membrane protein design with a novel, repetitive sequence (CC1), able to associate into an antiparallel homotetrameric helical bundle. In this study it is also proposed an additional computational framework, for *de novo* membrane protein design, combining Crick parametrisation tools and simulations in implicit and explicit membranes. The design pathway combined minimal and rational approaches, completed by knowledge-based and statistical studies. Molecular dynamics simulations in GROMACS using explicit POPC, POPE and POPG-containing membranes, showed that the antiparallel homotetrameric bundle formed by the repetitive CC1 sequence, appeared to be stable across all types of lipid compositions tested, showing great stability, as predicted by the initial energy score functions. Moreover, in POPE and POPG- explicit membranes, the CC1 antiparallel tetramer showed structural conformations which guide towards a potential mechanosensitive channel-like function. The experiments were compared to controls represented by the designed REAMP tetramer and poly-leucine antiparallel bundle.

The CC1 antiparallel homotetramer has also proven biocompatibility, when expressed in *E. coli* C41 (DE3) cells, having in the structure the mistic protein and the split-variant of the superfolder yellow fluorescent protein, as “internal chaperone”, to overcome the challenges associated with the translocation machinery. The expression levels have shown to be comparable to the REAMP control, having attached the same tags, and the expression levels have shown temperature and medium composition dependence, essential for future considerations.

CC1 was also successfully cell-free synthesised. The efficiency of insertion into liposomes of the folded state and the fluorescence of the sfYFP have shown liposome composition dependence. CC1 have shown to also stabilise POPC: DPhPC containing bilayers in droplet-interface bilayers assays.

The present work comprises a novel finding in the field of *de novo* design, highlighting a novel, biocompatible sequence, with a special highlight on the importance of the lipid composition in *de novo* protein stability and membrane insertion.

Dedication

I dedicate this thesis to my parents, friends, professors, and everyone else who has impacted my life journey until this point

Funding

This project has received funding from the European Commission Horizon 2020 research and innovation programme under the Marie Curie Sklodowska grant agreement no. 847419

Acknowledgements

Academic acknowledgements

I would like to thank my supervisor, Dr. John Simms, for all the shared knowledge, help, motivation, and patience. All the great conversations made me understand that science is not only about results or the ratio between good and bad science days but also about perseverance and the process of growing as a person.

I would like to thank my associate supervisor, Dr. Alice Rothnie, for her kind guidance and support.

I would like to thank AMPL members for the great conversations during the AMPL meetings.

I would like to thank the members of the first and second cohorts of the MemTrain programme with whom I shared this great research experience.

I would like to thank members of lab MB331, for the help, conversations and co-operance.

I would like to thank Nikki, Keith, Roy and Collin from the HLS orders for their contribution.

I would like to thank the technical team, especially Fatehma and Rachel, for all their help, especially with the technical difficulties.

Personal acknowledgements

Aș dori să le mulțumesc din tot sufletul părinților mei pentru toată susținerea pe care mi-au oferit-o de-a lungul timpului, pentru toate eforturile lor depuse și sacrificiile pe care le-au făcut, pentru a-mi oferi mie șansa de a fi acum, aici. Totodată, și pentru toate orele petrecute pe Skype de la mii de kilometri depărtare, doar pentru a mă face să mă simt ca și când aș fi fost acasă, de oriunde m-aș fi aflat, atât în momentele de glorie, fie ele puține, cât și în cele mai puțin bune, motivându-mă la tot pasul. (I would like to thank my parents for all the support they gave me throughout the time, for all their efforts and sacrifices they made, in order to give me the chance to be now, in this position. Also, I would like to thank them for all the hours spent with me on Skype from thousands kilometres away, just to make me feel like home, despite the place I was finding myself to be in, either during the glorious moments, even a few of them, either in the less good moments, motivating me at each step).

I would like to thank my best friend, colleague, and sister I never had, Olivia. She has shown me immense moral support and made my life easier and better throughout my studies and afterwards. What's all left to say is our inside quote: "…, are you OK?".

I would like to thank my friends, especially Carolina, Divya, Laura, Manoj, Amandeep and Annelise, for the support, laughs, cries, wonderful conversations and quality time I have spent with you all.

I would like to thank Jennifer, for being such a good friend and neighbour, and for supporting me towards the end of my studies and making my days brighter with all her pranks.

To everyone else I know and have not mentioned in this section, I would like to thank you.

Table of contents

List of abbreviations

AASTY	Poly(acrylic acid-co-styrene)
AFM	Atomic force microscopy
APS	Ammonium persulfate
ATP	Adenosine triphosphate
BiFC	Bimolecular fluorescence complementarity
BLM	Black lipid membrane
Boc	Tertbutoxycarbonyl
BSA	Bovine serum albumin
CAD	Computer-aided design
CAM	Computer-aided manufacturing
CASP	Critical Assessment of Structure Prediction
CCCP	Coiled-coils Crick Parametrization
CFE	Cell-free expressed
CFPS	Cell-free protein synthesis
CFTR	Cystic fibrosis transmembrane conductance regulator
CG	Coarse-grained
CHAMP	Computer helical anti-membrane protein
CL	Cardiolipin
CMC	Critical micelle concentration
CPU	Central processing unit
dCTP	Deoxycytidine triphosphate
DDM	Dodecyl maltoside
dGTP	Deoxyguanosine triphosphate
DHA	Docosahexaenoic acid
DHFR	Dihydrofolate reductase
DIB	Droplet interface bilayer
DIBMA	Diisobutylene-maleic acid
DM	Decyl maltoside
DMPC	1,2-Dimyristoyl-*sn*-glycero-3-phosphocholine
dNTP	Deoxynucleoside triphosphate
DPhPC	1,2-Diphytanoyl-sn-glycero-3-phosphocholine
DsbB	Disulfide bond formation protein B
dTTP	Thymidine triphosphate
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
EmrE	Ethidium multidrug resistance protein
EPR	Electron paramagnetic resonance
ER	Endoplasmic reticulum
ERAD	Endoplasmic reticulum-associated degradation
FA	Fatty acid
FAS	Fatty acid synthesis
Fmoc	9-Fluorenylmethoxycarbonyl
FRET	Fluorescence resonance energy transfer
GFP	Green fluorescent protein
GPCR	G protein-coupled receptor

GPU	Graphics processing unit
HEPES	4- (2-Hydroxyethyl)-1-piperazine ethane-sulfonic acid
HIV	Human immunodeficiency virus
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-d-1-thiogalactopyranoside
IVTT	In vitro transcription and translation
KIH	Knobs-into-holes
LacY	Lactose permease
LB	Luria broth
LSB	Laemmli sample buffer
MD	Molecular dynamics
MP	Membrane protein
mRNA	Messenger ribonucleic acid
MscL	Large-conductance mechanosensitive channel
MW	Molecular weight
Ni^{2+}-NTA	Nickel nitrilotriacetic acid
NMR	Nuclear magnetic resonance
O.D.600	Optical density at 600 nm
OG	Octyl glucoside
OmpA	Outer membrane protein A
PBC	Periodical boundary condition
PC	Phosphatidylcholine
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDI	Protein disulfide isomerases
PDMS	Polydimethylsiloxane
PE	Phosphatidylethanolamine
PEG20K	Polyethylene glycol-20K
PG	Phosphatidylglycerol
PI	Phosphatidylinositol
PIP	Phosphatidylinositol phosphate
PME	Particle mesh Ewald
POPC	1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
POPE	1-Hexadecanoyl-2-(9Z-octadecenoyl)-*sn*-glycero-3-phosphoethanolamine
POPG	1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol
POPS	1-Hexadecanoyl-2--(9Z-octadecenoyl)-*sn*-glycero-3-phospho-L-serine
pRMSD	Pairwise root-mean-square deviation
PS	Phosphatidylserine
PVDF	Polyvinylidene fluoride
REAMP	Recombinantly Expressed Artificial Membrane Protein
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
rRMSD	Reference root-mean-square deviation
rSAP	Shrimp alkaline phosphatase
SDS	Sodium dodecyl-sulfate
SDS-PAGE	Sodium dodecyl-sulfate polyacrylamide gel electrophoresis
SEIRAS	Surface-enhanced infrared spectroscopy
SERS	Surface-enhanced Raman spectroscopy

SFA	Saturated fatty acid
sfYFP	Superfolder yellow fluorescent protein
SM	Sphingomyelin
SMA	Styrene maleic acid
SPPS	Solid-phase peptide synthesis
SRP	Signal recognition particle
TAE	Tris-acetate-EDTA
TB	Terrific broth
TBS	Tris-buffered saline
TBS-T	Tris-buffered saline-Tween 20
TEMED	Tetramethylethylenediamine
TM	Transmembrane
TMD	Transmembrane domain
TRPV1	Transient receptor potential vanilloid 1
UFA	Unsaturated fatty acid
VdW	Van der Waals
VLCFA	Very long-chain fatty acid
VMD	Visual molecular dynamics
XylE	D-xylose/proton symporter
YFP	Yellow fluorescent protein

List of Figures

List of Tables

1. General introduction

1.1. Membrane proteins

Membrane proteins are among the most important biomolecules, firstly because they represent the mediators between the cell's external and internal environment (Pan and Vachet, 2022). They perform numerous essential functions, such as facilitating cell signalling, transport, cell adhesion, catalysis, or the regulation of the membrane's electric potential (Chorev and Robinson, 2020).

Their importance is further sustained by the large proportion of existing drugs which target membrane proteins. Almost 70% of the current medicines target membrane proteins, and from those, almost 19% target ion channels (Arinaminpathy *et al.*, 2009).

Having such critical roles within the cell, throughout the years, scientists have been intrigued by membrane proteins precisely because of the challenge of studying them. Compared to soluble proteins, the greatest challenge in studying and understanding membrane proteins' structures relies on their topology in a lipid environment, which imposes certain limitations. Firstly, this refers to the dynamics and heterogeneity properties of the lipid environment, certain lipid compositions are essential for some protein function, and even in artificial systems, it is difficult to recreate the same lipid environment to maintain the structural integrity and functionality. Also, structural studies often involve the isolation of the membrane protein from the lipid environment, often using detergents, which cannot always entirely mimic the natural lipid environment, causing protein aggregation, or altered behaviour. Membrane protein crystallization is another technical aspect associated with the challenges imposed by the close relationship between the membrane protein and the surrounding lipid environment; the stabilisation of the studied MP requires special, expensive techniques and materials, such as the use of lipid cubic phases, or nanodiscs-forming agents.

Currently, in the largest protein database (PDB), most of the experimentally determined protein structures presented are obtained from investigation techniques such as X-ray crystallography or nuclear magnetic resonance (NMR). However, membrane proteins possess a significant feature: they are dynamic entities, and traditional structural analysis techniques do not exploit this feature.

These methods still cannot overcome the high dynamics barrier. Thus, cross-interdisciplinary methods, including expanding computational techniques, aim to decipher the dynamic lipid-embedded folds.

As known from such methods, most MP structures are alpha-helical bundles. The structural maintenance and recognition depend on non-covalent interactions, hydrogen bonds, and van der Waals

forces, which are in close correspondence with the dielectric features of the membrane and which are distributed in such a manner that they satisfy the stability of the structure in both the hydrophobic core formed by the lipid acyl chains, but also in the head group region. Multiple studies in the past have attempted determination of the structural 'hot spots' that drive the folding and stability through site-directed mutagenesis; however, most led to inconclusive results.

Understanding the driving forces is now one of the priorities for protein engineers, who are relying on the balanced application of bioinformatics, molecular simulations, and experimental validation, spotlighting sequence-structure determinants.

1.1.1. From synthesis to folding

The intricate orchestration of synthesising, inserting, and folding membrane proteins represents a pivotal journey that underpins their functional integrity within the lipid bilayer of cellular membranes.

Membrane protein synthesis is initiated in the nucleus, with the transcription of the genes encoding the membrane protein to messenger RNA (mRNA). The mRNA is then exported from the nucleus to the cytoplasm, where it can associate with the ribosomes, which are found either free in the cytosol or bound to the membrane of the endoplasmic reticulum (ER), and which represent the cellular machinery for protein synthesis.

The ribosome translates the mRNA into a polypeptide chain. Numerous membrane proteins bear an N-terminal signal sequence responsible for guiding the ribosome towards the endoplasmic reticulum (ER) membrane. This unique sequence is identified by the signal recognition particle (SRP), which guides the ribosome to the ER membrane, where the SRP is located. Once the ribosome is linked to the ER membrane, the signal sequence integrates into the translocon, a protein-conducting channel within the ER membrane, allowing the translation process to continue (Schulze *et al.*, 2014).

The complex formed by the SRP machinery and the ribosome docks to the SRP receptor and the synthesised polypeptide is then translocated across the ER membrane through a protein-conducting channel known as the translocon. After this step, the signal sequence is typically cleaved by a peptidase. (Peetz *et al.*, 2017). The hydrophobic segments of the polypeptide (transmembrane domains) are recognised by the translocon, which enables the integration into the lipid bilayer.

The orientation of transmembrane domains is determined by the distribution of charged residues flanking the hydrophobic segments, following the positive-inside rule, where positively charged residues are more likely to be found on the cytosolic side of the membrane (Pleiner *et al.*, 2020).

Proteins that span the membrane multiple times thread in and out of the translocon until all

transmembrane domains are properly inserted.

Initial folding events may occur within the translocon or immediately after the release from the translocon, where the protein begins to adopt its three-dimensional structure.

In Figure 1.1., panel A there are illustrated the membrane protein synthesis and membrane targeting processes, completed by the membrane insertion and assembly, as represented in panel B.

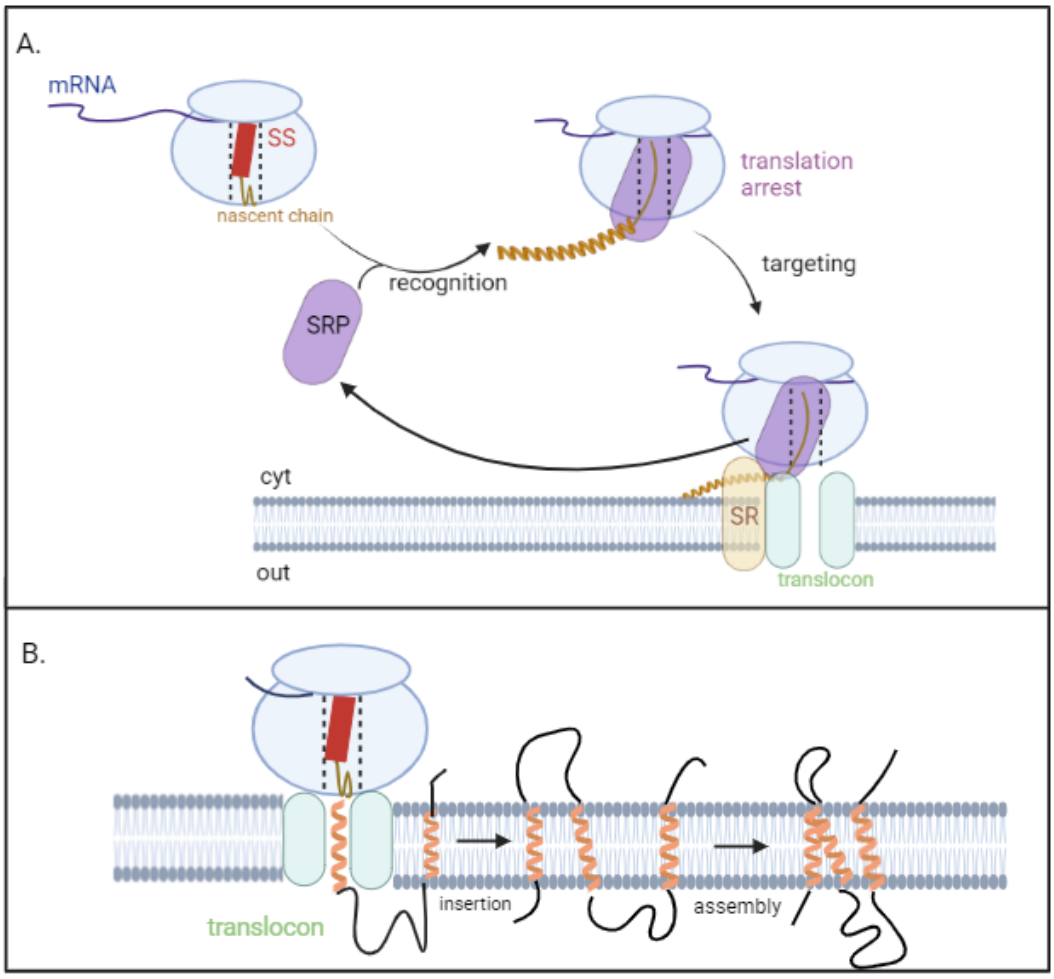


Figure 1.1. Membrane protein synthesis and membrane targeting and insertion. A. The ribosome (light blue) translates the mRNA of a membrane protein. The nascent chain is represented in brown, which contains the signal sequence (SS), represented as the red rectangle. The sequence recognition particle (SRP), represented in purple, targets the hydrophobic sequence of the resulting helix (orange), binds to the ribosome and arrests the nascent polypeptide elongation (translation arrest). The resulting complex binds to the SRP receptor (SR, yellow), which is associated with the translocon (green), after which it dissociates from the SR, causing the transfer of the hydrophobic sequence to the translocon. B. The insertion of the helices (orange) enables the membrane integration of the newly synthesised protein, and can associate to form homo- or heteromeric assemblies. (Created with BioRender)

In prokaryotes, the sequence recognition particle (SRP) has the role to target the newly synthesised polypeptide chain to a membrane-associated receptor, FtsY, and subsequently to the SecYEG translocon.

The insertion in the cytoplasmic membrane, mediated by the Sec translocon is supported by two proteins. The support proteins are the ATPase SecA, is peripherally associated, and it enables the translocation of periplasmic loops, and the YidC, which was studied for its possible implications in the mediation of the lateral transport of the transmembrane helices into the membrane (Freigassner, Pichler and Glieder, 2009).

Molecular chaperones, together with enzymes like protein disulfide isomerases (PDIs), actively contribute to the folding process within molecular biology. The key role lies in ensuring the precise formation of disulfide bonds, thereby averting undesired aggregation of molecules (Posokhov *et al.*, 2008).

For example, studies on the effect of chaperones have been investigated for *E. coli*, expressing the magnesium transporter (CorA), and it was shown that the co-expression of the cytoplasmic chaperone DnaK increased the expression levels of the protein of interest and also was assigned with the capability of preventing the accumulation of the protein in inclusion bodies (Chen *et al.*, 2003).

In Figure 1.2. it is schematically represented the membrane protein biogenesis, highlighting the implications and localisations of the FtsY, SecA, SecYEG, YidC and DnaK in the processes.

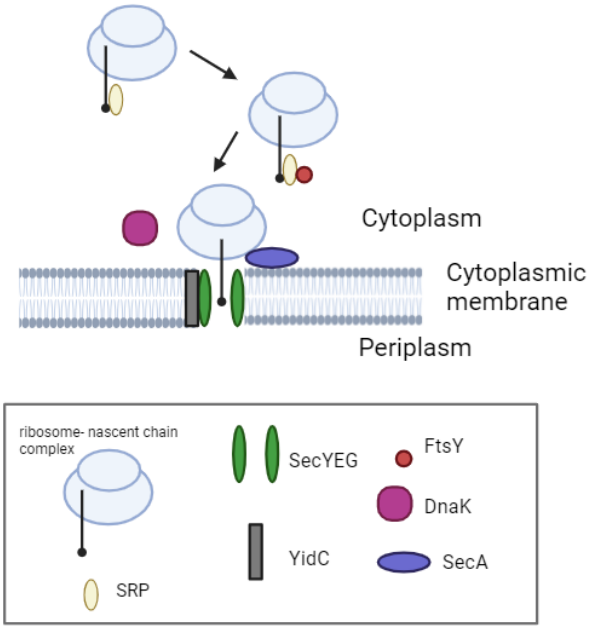


Figure 1.2. Prokaryotic membrane protein biosynthesis. The polypeptide nascent chain exiting the ribosome binds to the SRP (yellow), associated with its receptor, FtsY (red). The resulting complex binds to the SecYEG (green)/YidC (grey) pore. The insertion and translocation of the cytoplasmic and periplasmic loops across the membrane and the assembly are enabled by the SecA (blue) and chaperones (e.g. DnaK, purple). (Created with BioRender)

The quality control mechanism enforced within the ER rigorously sieves out specifically folded and assembled proteins for onward transportation to the Golgi apparatus. In contrast, proteins undergoing

misfolding find themselves sequestered inside the ER and susceptible to potential degradation through the intricate ER-associated degradation (ERAD) avenues (Vembar and Brodsky, 2008).

Many membrane proteins undergo glycosylation in the ER, which can play a role in protein folding and stability.

Some membrane proteins also receive lipid modifications, such as prenylation or palmitoylation, which can affect their membrane localisation and function (Wang and Casey, 2016).

Moreover, additional folding adjustments and modifications can occur in the Golgi apparatus before the protein reaches its final location.

1.2. Lipid bilayer: Structures and influence

The lipid bilayer, a crucial element of all cell membranes, shapes a vibrant and intricate domain for numerous biological processes. Its composition exhibits remarkable diversity, varying significantly across cell types and even within specific cell regions. Comprising phospholipids, cholesterol, and glycolipids primarily each contributes distinct characteristics influencing the membrane's flexibility, thickness, and functionality.

Phospholipids reign as the dominant lipid species within the lipid bilayer, orchestrating a delicate balance of dual characteristics. These molecules, known for their amphipathic nature, possess both hydrophobic fatty acid tails and hydrophilic phosphate heads. Through this unique composition, phospholipids craft the bilayer framework, positioning their hydrophobic tails towards the interior to evade water while directing their hydrophilic heads outward, embracing the aqueous surroundings.

The foundation of a phospholipid is a three-carbon glycerol molecule. Each carbon atom in glycerol can form ester or ether linkages with other molecules.

Attached to the first and second carbons of the glycerol backbone are two fatty acid tails. These tails are hydrophobic and can vary in length and in the presence of double bonds. The variation in fatty acid composition affects the fluidity and permeability of the cell membrane. Conventionally, the indicated positions of the fatty acids (FA) are based on their steric specificity in rapport to the glycerol carbon atoms they are covalently linked to, namely sn-1, sn-2, or sn-3. Saturated fatty acids contain no double bonds, making the membrane more rigid, while unsaturated fatty acids contain one or more double bonds, introducing kinks that increase membrane fluidity. The third carbon of the glycerol backbone is linked to a phosphate group, which is further connected to the polar head group (e.g. choline, ethanolamine, glycerol, etc). In Figure 1.3. it is exemplified a general structure of a phospholipid.

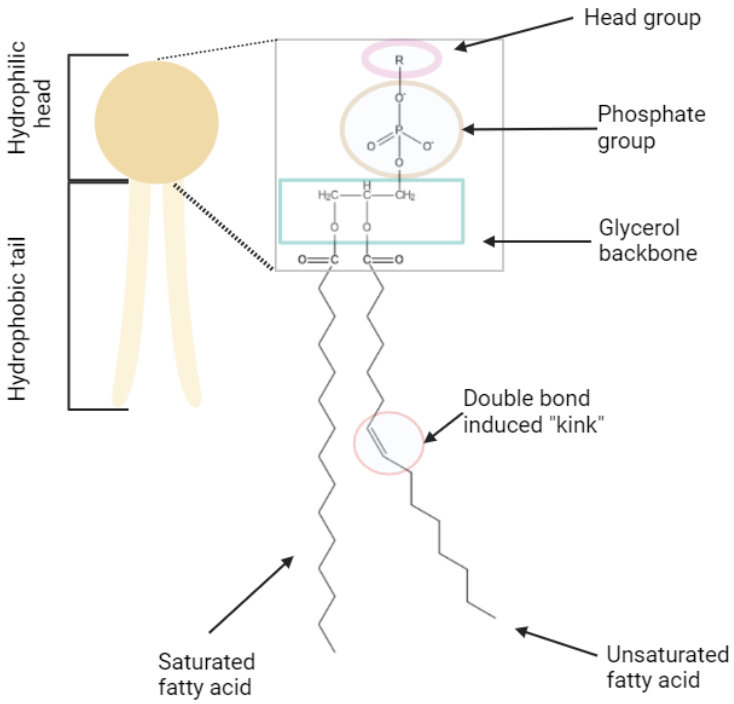


Figure 1.3. General structure of a phospholipid. The structure is divided into two main components: a hydrophilic head and a hydrophobic tail. The head comprises a head group attached to a glycerol backbone via a phosphate group. The hydrophobic tail can contain saturated fatty acids and unsaturated fatty acids, the latter introducing a kink in the chain structure. (Created with BioRender and ChemDrawPro).

1.2.1. Fatty acid composition influences membrane fluidity

The degree of saturation/unsaturation of the lipids in a cell membrane is a highly regulated process determined by a plethora of mechanisms which can influence the degree of membrane fluidity.

The saturated fatty acids (SFAs) have in composition a saturated acyl chain, and the tetrahedral geometry of the sp^3 hybridisation state of the carbon atoms in the structure of the acyl chain determines a straight conformational appearance of the chain. In terms of packing, the straight structures of the acyl chains allow for close packing, increasing the density and rigidity of the membrane. The close packing of SFAs in the lipid bilayer can determine the assembly of a more ordered and less fluid membrane. It was shown in certain studies that the presence of saturated very long-chain fatty acids (VLCFAs) in the composition of a membrane could induce significant hydrophobic mismatch in fluid phospholipid bilayers, which further affected the membrane organisation and interactions between the bilayer's leaflets with a further extended impact on signal transduction mechanisms (Paz Ramos *et al.*, 2016).

In contrast, unsaturated fatty acids (UFAs), contain one or more double bonds. Due to the rigidity imposed by the double bond π (pi) and p-type orbitals quantum mechanical features and sp^2 carbon hybridisation and the possibility of geometrical isomerism cis-trans, the unsaturated chains induce a "kinked" geometry. This results in increased membrane fluidity and flexibility. The presence of UFAs can also influence the membrane's response to environmental stress factors, such as temperature changes. Studies have shown that docosahexaenoic acid (DHA), which is a highly unsaturated fatty acid, can heavily impact the physicochemical properties of the membrane, hence the associated interactions between the membrane lipids and the embedded proteins (Harayama and Shimizu, 2020). For example, it was shown that cyclopropane fatty acids in the cell membrane of *Escherichia coli* provide more rigid kinks in the lipid acyl chain, which determine the enhancement of the membrane fluidity at low temperatures and act as shields for the cell against cold shock (Maiti, Kumar and Daschakraborty, 2023).

In addition to the saturation degree of the fatty acids, also the length represents an impacting factor for the membrane fluidity. Long-chain fatty acids increase the membrane rigidity due to the higher amount of interactions that can occur, by the increase of the contact surface area. In contrast, short-chain fatty acids could increase membrane fluidity. However, they rarely occur in natural membranes (Maulucci *et al.*, 2016).

1.2.2. Fatty acid composition determined by various factors

The synthesis of fatty acids, especially in *E. coli* cells, is known to be mainly regulated by the enzymes involved in the fatty-acids synthesis pathway (FAS). There are known two key enzymes, which are involved in this process: β-ketoacyl-acyl carrier protein synthase I (FabB) and the β-ketoacyl-acyl carrier protein synthase II (FabF). These enzymes determine the types of synthesised fatty acids. FabF is known to be involved in the synthesis of the unsaturated fatty acids. The length of the fatty acids is controlled by FabB (Garwin, Klages and Cronan, 1980; Jeon *et al.*, 2012). As expected, any mutations in the genes encoding these enzymes have a strong impact on the follow-up FA synthesis. It was proven for *E. coli* mutants, which lacked FabF, to have a strong deficiency in the cis-vaccenic acid, known to be the most prevalent unsaturated FA in *E. coli* membranes. Furthermore, studies have shown that temperature plays an important role in the control of fatty acids (Garwin, Klages and Cronan, 1980).

The environmental factors have further been shown to strongly impact the synthesis of fatty acids. The stringent response, which represents a global regulatory mechanism that *E. coli* cells harness, is activated in conditions of nutrient deprivation, such as carbon source starvation or other stress factors

(pH, temperature). This mechanism influences the fatty acids composition and there were observed modifications such as FA elongation, cyclopropanation and increased cardiolipin formation (Wehrli *et al.*, 2016). It was proposed that the stringent response mechanism is mediated by the accumulation of guanosine tetraphosphate (ppGpp), which targets a wide range of cellular modifications, including the lipid (re)composition.

1.2.3. Phospholipid composition variety and impact on membrane shape

The major classes of phospholipids found in membranes include:

Phosphatidylcholine (PC): The most common phospholipid in animal cell membranes, contributing to membrane fluidity and permeability.

Phosphatidylethanolamine (PE): PE is abundant in the inner leaflet of the plasma membrane and in the mitochondrial membranes. It contributes to membrane curvature and fusion.

Phosphatidylserine (PS): PS is mainly located in the inner leaflet of the plasma membrane and plays a role in cell signalling and apoptosis.

Phosphatidylinositol (PI): Found in the inner leaflet of the plasma membrane, PI is involved in cell signalling pathways, particularly through its phosphorylated derivatives.

Noteworthy, the phospholipid composition can vary across various cell types.

For example, in bacterial cell (*E. coli*) membranes, the common phospholipids include PE (75%), PG (10%) and cardiolipin (CL) (7%) (Sohlenkamp and Geiger, 2015). In yeast cells (*Saccharomyces cerevisiae*) the membrane include PE, PC, PI and PS (Kohlwein, 2017), whilst mammalian cells (*HEK293T*) are rich in PC, PE, sphingomyelin (SM) and cholesterol (Cockcroft, 2021) and insect cells include PC, PE and PI (Moreno *et al.*, 2021).

The packing parameter, also known as the lipid shape, directly impacts the curvature of the membrane based on the ratio established between the polar head group and the hydrophobic acyl chain and exhibits lateral pressure in different regions on the bilayer, depending on the structure.

Simply, an almost evenly distributed ratio leads to the formation of planar bilayers, or neutral curvature, as is the case for phosphatidylcholine (PC). A small area ratio head group/tail leads to a cone-shaped lipid arrangement, known as negative membrane curvature, as is the case for phosphatidylethanolamine (PE) and cardiolipin (CL). Inverted-cone-shaped lipids induce however, a positive membrane curvature, such as phosphatidylserine (PS) and less the phosphatidylglycerol (PG) (Golla, Boyd and May, 2024). In Figure 1.4 there is schematically represented the influence of the structure

of the phospholipids on the lateral pressure exhibited.

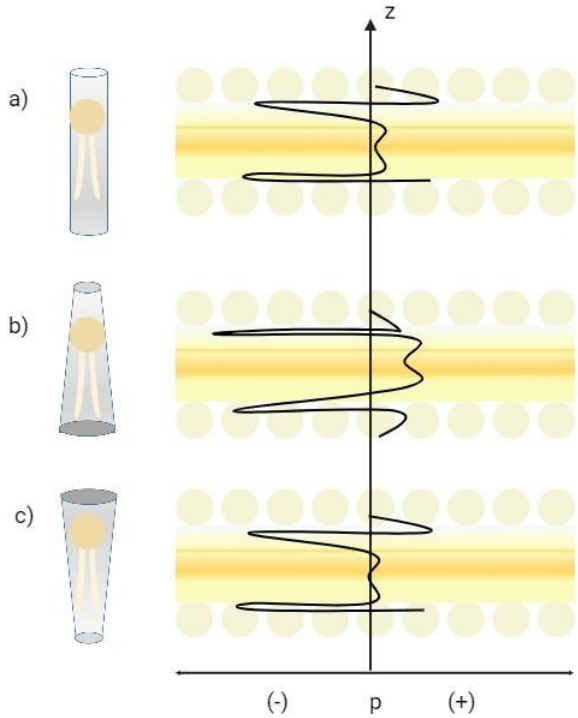


Figure 1.4. Schematic representation of the phospholipid shapes (left) and their impact on the lateral pressure profiles (right) in the formed bilayers. The variation in different regions of the bilayer (head group areas and bilayer core) is represented as a curved line in rapport to the bilayer depth (z). An increase in the pressure is situated in the (+) plan, respectively a decreased lateral pressure in the (-) plan a) In cylindrical-shaped phospholipids, forming planar bilayers (e.g. POPC) exhibits relatively equalised lateral pressure in the head group and tail areas. b) Cone-shaped phospholipids, with small head groups (e.g. POPE), which form bilayers with negative curvature, redistribute the positive lateral pressure from the head groups to the hydrophobic core. c) Inverse-cone-shaped phospholipids (such as lysoPC, or PS), forming positively-curved bilayers, have increased lateral pressure in the head group region and lower pressure in the core. (Created with BioRender).

1.2.4. Interactions with MP

The hydrophobic regions of membrane proteins, often composed of nonpolar amino acid residues, interact with the hydrophobic fatty acid tails of phospholipids. This interaction is fundamental for the integration of proteins into the lipid bilayer and contributes to the stability of the protein's position within the membrane (De Planque and Killian, 2003).

In addition to the hydrophobic interactions, also electrostatic and dipole interactions can be established. Amino acids with positive or negative charges located on the exterior of membrane proteins can engage with the charged head sections of phospholipids (Fernandes *et al.*, 2015). To illustrate, basic amino acid residues could interact with the negatively charged head components of phospholipids such

as phosphatidylserine (PS) or phosphatidylinositol phosphates (PIPs). The interactions between dipole moments of phospholipid head groups and dipoles present in the protein structure intricately mould the orientation and function of the protein within the membrane, spotlighting a profound influence on its behaviour.

Furthermore, hydrogen bonds can establish connections between the polar amino acid residues within membrane proteins and the polar head groups existing in phospholipids. These interactions are known for being specific to particular phospholipids, consequently influencing the localised lipid composition around the protein and potentially altering its functional behaviour (Lee, 2005).

Weak van der Waals forces might arise between the membrane protein and the fatty acid tails or head groups of phospholipids. These intricate engagements significantly impact the protein's overall stability and dynamics within the membrane; however, they add complexity to its behaviour (Yeagle, 2014). In Figure 1.5, the general types of interactions between proteins and lipids are schematically represented.

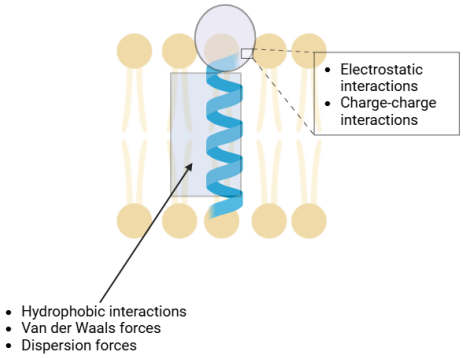


Figure 1.5. Schematic representation of general interaction types between transmembrane helices and phospholipids. In the bilayer core, hydrophobic interactions, van der Waals and dispersion forces between the TM region of the helix and the tails of the phospholipids. In the helix limit surrounded by the phospholipids head groups, there can occur electrostatic, charge-charge interactions and hydrogen bonds. (Created with BioRender and ChemDrawPro).

The interaction among membrane proteins and surrounding phospholipids has the potential to induce noteworthy structural changes in the protein, thereby influencing its operational effectiveness. Moreover, the existence of specific proteins may alter the local lipid environment, consequently impacting membrane curvature or thickness, which in turn can significantly affect protein functionality.

1.3. Amino acid distribution and bias in integral membrane proteins

1.3.1. Amino acid distribution

The amino acid distribution and bias in integral membrane proteins have been extensively studied and there have been established several structural distribution patterns.

In terms of amino acid distribution patterns, it has been shown that large hydrophobic residues (e.g. leucine, isoleucine, valine) are more prevalent in the transmembrane regions, as studies have shown and decrease in frequency near the interface with water, and patterns, where leucine was prevalent, were also associated with the alpha-helical transmembrane domain, for which leucine appeared to be favoured, as early studies on the amino acids' alpha-helical formation predisposition have shown (Mbaye *et al.*, 2019).

Aromatic residues, such as tyrosine, tryptophan, and phenylalanine, are associated with a localisation preference near the membrane-water and protein-membrane interfaces and, through the cation-π interactions, being able to anchor the protein domains within the membrane (Ulmschneider and Sansom, 2001).

The extremities of the transmembrane domains were shown to be predominantly occupied by charged residues, which, depending on the charge they possess, exhibit a preference for a certain side of the membrane. In this sense, studies of von Heijne and colleagues (Von Heijne, 2006) described that positively charged residues, such as arginine or lysine, are enriched in the cytoplasmic regions, also describing the 'positive-inside' rule, which contributes to the topology of membrane proteins. On the other side, the negatively charged residues are rare within transmembrane helices, with a bias against being located in the cytosolic interfacial region, preferentially on the non-cytoplasmic side of the membrane, following the 'negative-outside' rule (Baker *et al.*, 2017).

1.3.2. Amino acid distribution bias across monotopic and polytopic integral membrane proteins

Structural studies have indicated certain subtle, however essential amino acid distribution differences between monotopic and polytopic integral membrane proteins despite the generally defined distribution patterns. These include the identification of a higher presence of polar/charged amino acids within the TMs of polytopic membrane proteins (Baker *et al.*, 2017), which are involved for example in channel/pore formation, or transport functions. Also, a larger dispersion of aromatic residues, also including at the membrane-protein interfaces for the same protein subclass. In addition, proline, as a helix breaker has been found as a common residue for polytopic integral MP, in flexible or kinked helices of GPCRs. Such examples can be found in the TM6 and TM7 or rhodopsin (Pope *et al.*, 2020), or in the TM5, TM6 and TM7 of the beta2-adrenergic receptor, contributing to the conformational changes required for

ligand binding (Shi *et al.*, 2002).

1.4. Alpha-helix is the predominant secondary structure

The predominance of α-helices as the secondary structure of membrane proteins is a result of their capacity to satisfy hydrogen bonding requirements, interact in a favourable manner with the hydrophobic lipid environment and form stable structures.

In the structure of an α-helix, the intramolecular hydrogen bonds between the backbone carbonyl oxygen and the amide hydrogen from (i+4) residue have the role to stabilise the structure and neutralise the polarity of the backbone atoms, shielding the polar backbone, particularities which make alpha-helices suitable for embedding in nonpolar environments, such as the lipid bilayer, reducing in this way the necessity of satisfaction of further interactions with the surrounding lipids. Also, the hydrophobic amino acid content contributes to the interactions with the lipid bilayer hydrophobic core by the sterical arrangement of the component amino acids (predominantly leucine, isoleucine, valine, phenylalanine) side chains, which are projected into the lipid bilayer core, thus establishing efficient hydrophobic interactions. The compatibility with the lipid bilayer, especially with the interface, is completed by the aromatic residues (tyrosine or tryptophan), often located at the ends of α-helices, and help anchor the helices, satisfying both the interactions with the hydrophobic core and the polar head groups (Sparks *et al.*, 2014).

Also, alpha-helices, unlike beta-barrels, are more compact and flexible, can adapt to conformational changes required for functions more easily and can pack together tightly within a lipid bilayer. Beta-barrels, however, are less common secondary structures across MP, form large, rigid, cylindrical structures, which require a defined number of strands to form a stable (which also induces the size limitation for constructing small, efficient MP), membrane-spanning pore and need to follow a more complex and different hydrogen bonding pattern, compared to alpha-helices; they must fold back on themselves to create a closed barrel with the hydrophobic side chains pointing outwards and a rigid, complex final structure (Wimley, 2003).

Examples of classes of membrane proteins with alpha-helical structure are G-protein coupled receptors (GPCRs), which have seven transmembrane alpha helices which form a helical bundle involved in signal transduction; channels such as potassium channel, or aquaporin also have alpha-helical structure, as well as transporters such as the glucose transporter (GLUT) family, or ABC transporter family, which require conformational changes for their function, satisfied by the flexibility characteristic of alpha helices in their composition

1.5. Forces involved in membrane helix association

Structural and biophysical studies have led to the conclusion that several key forces are involved in the transmembrane helix association, which is a fundamental process throughout the folding and assembly of membrane proteins. The complex mechanism is dictated by an interplay between the hydrophobic interactions, Van der Waals forces, hydrogen bonding, electrostatic interactions, and also completed by the lipid-mediated effects, and below, the role of each is shortly described below.

As described in the general amino acid distribution, the nonpolar amino acids are located preferentially within the hydrophobic membrane core, and the hydrophobic sidechains interact with each other, contributing to the helix packing. Nonetheless, the association between the hydrophobic amino acid chains is less dominant than it was observed for soluble proteins; the localisation of the non-polar amino acids within a transmembrane domain would not exhibit such a strong hydrophobic effect, as would be the case for soluble proteins, due to a less water-accessible position within the lipid core (DeGrado, Gratkowski and Lear, 2003).

Van der Waals interactions would be established between the side chains of adjacent helices, contributing to helix-helix associations. These interactions would be, however, disturbed by structural changes, which could lead to the repositioning of helices at various distances between each other. Thus, the minimum distance requirement for the establishment of Van der Waals could not be met, and hence the inter-helical association could be impacted (Rees and Engelman, 2017).

The hydrogen bonding is considered a strong driving force for transmembrane domain association; this type of force is a major folding/association force for soluble proteins; however, it could be encountered in the regions found within the lipid head group regions or could involve residues that are shielded from the lipid environment (DeGrado, Gratkowski and Lear, 2003).

The electrostatic interactions play essential roles that dictate the behaviour of a transmembrane helix, by determining its orientation, following the ‘positive-inside’ rule, stabilising oligomeric complexes and stabilise the helix through its charged residues placed near the ends of a transmembrane helix within the phospholipids head group regions (Rees and Engelman, 2017).

The effects exhibited by lipids can influence association of the helices through the effects imposed on the helix tilting due to the hydrophobic mismatch and the length of the helix, specific lipid-protein interactions, such as cholesterol-binding, or, through membrane deformation (Corin and Bowie, 2022).

1.6. Three-stage model of membrane protein folding

Almost five decades ago, Popot and Engelman (Popot and Engelman, 1990) were proposing the

'two-stage model' for membrane protein folding and oligomerisation. This model comprised, in the first stage, the formation of individual transmembrane helices that span the membrane, often facilitated by the translocon complex and, in the following stage, the interaction between these helices to form complex, stable and functional structures. Evidence that supports the two-stage model, such as structural analyses that show that transmembrane segments of known MP are hydrophobic alpha-helices, refolding experiments and assembly from fragments, and the existence of very small integral MP, which can fold independently (Engelman *et al.*, 2003) .

Later on, it has been defined the 'three-stage' model of helical membrane protein folding, which highlights the importance of the membrane-water interface, as an intermediate stage during the TM assembly into the membrane. Within this stage are the interactions between the membrane-water interface and the TM. This place acts as a vestibule for stabilisation of the helices before membrane insertion, which represents a crucial step, as demonstrated in molecular dynamics simulation experiments. After the stabilisation step, the helices enter the final stage, involving the insertion and assembly into the final conformation.

In addition to the molecular dynamics simulations, energetic analysis experiments support the second, intermediate stage for the helix stabilisation in the membrane-water interface, which have shown for a series of modelled TMs, the values of the computed insertion stabilisation energies from water to TM domain were half than from TM to water, which suggested the potential dominant role of the membrane-water interface, as a stabilisation vestibule during the TM release from the translocon (Kawamala and Abrol, 2022).

1.7. Forces involved in the folding mechanism

It is known that interactions within the polypeptide chain can contribute to stabilising the structure and also encourage folding, such as the hydrophobic effect, hydrogen bonds, Coulombic or electrostatic interactions, and van der Waals interactions (Dill, 1990; Nick Pace, Scholtz and Grimsley, 2014). However, it is still not understood to which extent the balance is inclined to membrane protein folding, due to the associated challenges such as the necessity of MP overexpression, isolation of high-purity MP and the limitations of the commonly used biophysical analysis techniques such as circular dichroism (CD), or Fourier-transformed force infrared spectroscopy (FTIR).

Furthermore, those forces were analysed separately as interactions that occur within the backbone atoms and interactions established by the amino acid sidechains. Despite being long disputed,

the noncanonical C-H…O hydrogen bonds, as secondary interactions between backbone atoms, were accepted as contributors to the stability of the structures (Newberry and Raines, 2019).

Although the carbon atoms within the protein are weak acids, some amino acid residues have associated sufficiently acidic protons, thus able to engage in hydrogen bonding. A relevant example is histidine, which is able to establish hydrogen bonds from Cε. Indeed, this can be considered a special case because it is known that, commonly, the donors for this type of bond are the Cα-H protons (Horowitz and Trievel, 2012).

The interesting roles of this type of interaction can be further explained by the unusual behaviour of the proline in an α-helix. If the proton is donated from an α-C to the carbonyl acceptors, then it was shown that in this way, the negative effect of proline within the structure, such as helix breaking, is diminished (Chakrabarti and Chakrabarti, 1998).

An additional type of interaction, which has been postulated to have an impact on protein folding, is the n-π* interaction (Newberry *et al.*, 2013). This is a weak interaction which occurs between two adjacent carbonyl groups, one contributing with the n pair or electrons, donating them to the π* orbital of the carbon involved in the peptide bond from the other carbonyl group (Newberry and Raines, 2019). Thus, within a protein, due to the large number of peptide bonds, this type of interaction occurs frequently, and it is believed that the formed resonance contributes to the overall stability of the folded protein.

It was also suggested that a classical dipole-dipole interaction is rather responsible for the chemical shielding corresponding to the carbonyl groups (Worley *et al.*, 2012). The results, however, are supported theoretically by a set of calculations having as subjects different X-ray structures and experimentally by assessing a formamide trimer model, thus drastically limiting the possibility of diversity in terms of other contributing factors, like dihedral angles or torsion angles determined by other scaffolds, which could influence the chemical shifts.

Furthermore, the identification of this type of interaction can help identify the protein's secondary structure because experimental studies have shown that they favour the formation of α-helices (Newberry *et al.*, 2013).

In addition to the interactions that take place within the backbone, secondary interactions engage the side chains of the amino acids, of which phenylalanine, tyrosine, and tryptophan are representatives (Meyer, Castellano and Diederich, 2003).

The special charge distribution determined by the aromatic rings of those sidechains allows for further interactions with anions and cations, the latter of which are recognised as important for folding (Newberry

and Raines, 2019).

Regarding the computational tools which were developed following the thorough study of those forces, they have significantly enhanced the ability to predict a structure and the stability of a protein. However, the process is still not completely accurate. As described, apart from the hydrogen bonds that are contributing massively to the stability of the α-helix, the n-π* interactions are also primordial. However, the force fields developed so far in computational predictions account only for the hydrogen bonds, indeed having absorbed the energy potential, which should be attributed to the n-π* interaction. Still, they don't treat them separately (Newberry *et al.*, 2013).

1.7.1. Some landmarks in membrane protein folding

As it was demonstrated in the initial comparative studies, that the folding process of membrane proteins is much more complex than for soluble proteins, further scientific questions represented the focus of subsequent studies, especially related to the factors that drive, or contribute to the membrane protein folding. The mechanisms of the integral membrane protein insertion and folding were assessed in studies (Cymer, Von Heijne and White, 2015), highlighting the importance of thermodynamics and of the translocon machinery (Sec61 translocon, SecYEG) (Skach, 2009), which was concluded essential for providing a dynamic pathway for the polypeptide synthesis and delivery into the cytosol, lumen and lipid bilayer.

Insertion and folding studies were performed for the OmpA and it was observed that it was transported across the cytoplasmic membrane and kept in an unfolded state in the periplasm, with the aid of a chaperone. The mechanism of insertion was spontaneous and the folding was dependent on the presence and binding to a lipopolysachharide (Kleinschmidt, 2003). Not only for OmpA was observed in studies that the folding process was dependent on the presence of other molecules, but also for LacY (Dowhan and Bogdanov, 2009), the topogenesis being dependent on phophatidylethanolamine.

Apart from the need of other molecules for the proper folding and insertion, another key aspect for protein folding and successful protein engineering was the application of phi-value analysis, which enabled the studying of the folding transition states of membrane proteins and to distinguish between folding mechanisms of alpha-helices and beta-barrels (Booth and Clarke, 2010), (Booth, 2012), or to re-emphasize the folding differences against the soluble proteins (Curnow and Booth, 2009).

1.8. Protein *de novo* design

The upcoming era of *de novo* protein design has a significant impact, as it enables the

understanding of protein particularities, such as the principles of folding, to the extent of their associated already determined functions, such as catalysis, binding properties, providing the chance also to address different biomedical needs, such as the personalised therapies (DeGrado and Korendovych, 2020). Following decades of intensive study and optimisation of strategies for protein design, it is now possible to design proteins with pre-determined structures and functions –"*de novo*". If a few decades ago, the protein-based therapies were isolated from natural sources or if later, when the interest in protein synthesis has risen, and the sequences and the structures chosen as targets were the ones obtained from classical NMR and X-ray analysis, now with substantial aid of computational tools, this has significantly changed and there are multiple approaches. The sequences of existing structures can be rationally varied to investigate the interactions at the atomic level or to sample/landscape different structures and functions, either by assembling local fragments, using human intuition or with the aid of artificial intelligence. Recent studies reveal successfully designed proteins, having attributed emerging functions like modulating the immune system (Silva *et al.*, 2019) , inhibiting viral infections (Chevalier, D. A. Silva, *et al.*, 2017), acting as signalling systems (Feng *et al.*, 2015), protein logic gates (Chen *et al.*, 2020) and even self-assembling biomaterials (Gonen *et al.*, 2015), serving as prototypes for artificial tissues.

One of the major challenges in the study of protein structures, which is one of the aims of protein engineering, is the identification of the structures, motifs, or sequence patterns which are responsible for determining certain functions for a specific protein.

In this regard, a set of rules that can be applied when designing a protein has been postulated, which represents the foundation of the current computational tools.

1.8.1. Minimal, rational, computational protein design

The field of protein design has resulted in the recognition that there is no universal approach to this domain; instead, a variety of methods exist primarily falling into the categories of minimal design and rational design. In the realm of protein design the concept of minimalism involves crafting simplified model proteins that encapsulate the necessary core elements for specific functions. These innovative designs are crafted to unravel proteins to their essential components emphasising the underlying principles dictating protein structure and operation. Within this minimalistic approach the primary objective revolves around grasping how subtle alterations in amino acid sequences can influence the folding and functionality of proteins (Woolfson, 2021). This methodology aids scientists in revealing the fundamental characteristics pivotal for proteins to effectively execute their biological functions. In a

minimal design context the utilisation of hydrophobic and polar residue patterns in the helical wheel which defines the target structure. By contrast, rational design involves utilising specific sequence motifs from naturally occurring proteins to stabilise particular structures. In rational design, experts strategically place each amino acid in a protein by applying chemical principles and computational tools. This method utilizes an understanding of protein characteristics interactions and structure-function connections to create proteins with specific desired properties. Rational design often entails modifying existing proteins or creating new ones, based on a logical comprehension of how molecular changes can affect protein behaviour (Hellinga, 1997). The inaugural rational design methodologies relied on qualitative principles of protein structure that were manually inspected (Hecht *et al.*, 1988). These experiments unequivocally showcased the capability of developing novel sequences from scratch that conform to distinct structures (Bryson *et al.*, 1995; Richardson *et al.*, 1992). Moreover, they elucidated that through a systematic design strategy referred to as "hierarchic design" (Bryson *et al.*, 1995), wherein, ascending levels of complexity are methodically introduced, fresh insights into the core aspects of protein structure and operation can be gleaned. Both minimal and rational design strategies hinge on the use of stable scaffolds which undergo modifications for functional purposes (Dawson, Rhys and Woolfson, 2019). Exploring a different design avenue involves embracing computational techniques to push boundaries. The majority of computational methods leverage databases filled with sequence fragments to craft target structures or functions (Huang, Boyken and Baker, 2016). Each of these strategies has its value, but if the rational design were combined with parametrised computational methods, it would form an even more powerful *de novo* design approach.

The use of computational techniques in membrane protein design has shown to be successful in a series of studies. Senes and colleagues (Senes, 2011) have focused on using computational design techniques, optimising a set of parameters targeting the van der Waals packing and the canonical and weak Cα-H…O hydrogen bonds, successfully leading to computationally stable oligomeric complexes of single-span transmembrane peptides. The complexity of designing multipass transmembrane proteins was achieved in computational design studies of Lu and co-workers (Lu *et al.*, 2018). They reported the accurate computational design of stable transmembrane protein oligomers (monomers, homodimers, trimers, tetramers) with up to eight transmembrane domains. The successful results were achieved by designing buried hydrogen bond networks to drive the specific interhelical interactions, with an additional focus on the incorporation of amphipathic and charged residues to control the orientation of the domains within the membrane. The obtained designs were not only successful in the computational simulations but also expressed in bacterial (*E. coli*) and mammalian cells and characterised using circular dichroism,

analytical ultracentrifugation, and single-molecule forced unfolding experiments. The accuracy of the dimeric and tetrameric designs was probed by the crystal structures.

De novo beta-barrels, able to insert and fold into lipid membranes, were successfully designed by Vorobieva and colleagues (Vorobieva *et al.*, 2021) from first principles. The study comprised iterative tests to probe geometric and sequence constraints on transmembrane β-barrel structures and their effects on packing and insertion in the membrane. Eight different structures were successfully obtained, introducing glycine residues and modulating the hydrogen networks to both satisfy the requirements for a β-sheet structure and also to balance the hydrophobic requirements for membrane insertion.

1.8.2. Small guide for *de novo* design of a membrane protein sequence

De novo design aims to overcome the challenges associated with MP (low expression, misfolding, aggregation) by developing proteins based on a low sequence complexity, detecting motifs involved in transmembrane helix packing (Walters and DeGrado, 2006). These motifs are involved in different types of interactions (helix-helix, lipid-helix), which would give further insights into the handling of MP, making more accessible the roles of each constituent.

Nevertheless, the use of controlled systems and parametrised structural motifs provides closer insights on the impact of the individual amino acid components in the sequence-structure-function relationship.

Initially proposed and discovered by Crick and Pauling in 1953 in keratin, coiled coils represent the foundation stone for *de novo* design, representing a reliable model and bridge between individual helices and complex native proteins (DeGrado and Korendovych, 2020). Conventionally, a classical coiled coil is a repeat of a heptad labelled "*abcdefg*". It has a left-handed packing orientation, and from a top view, the sidechains of the amino acids in positions a and d are facing the interior of the bundle. They are prevalently hydrophobic, whereas the sidechains of e and g are facing the inter-subunit interface or are involved in distanced inter-helical interactions and are generally more polar (Fraser, MacRae and Miller, 1965). The signature heptad repeat of hydrophilic (h) and polar (p) residues is hpphppp (Woolfson *et al.*, 2015), as schematically represented in Figure 1.6.

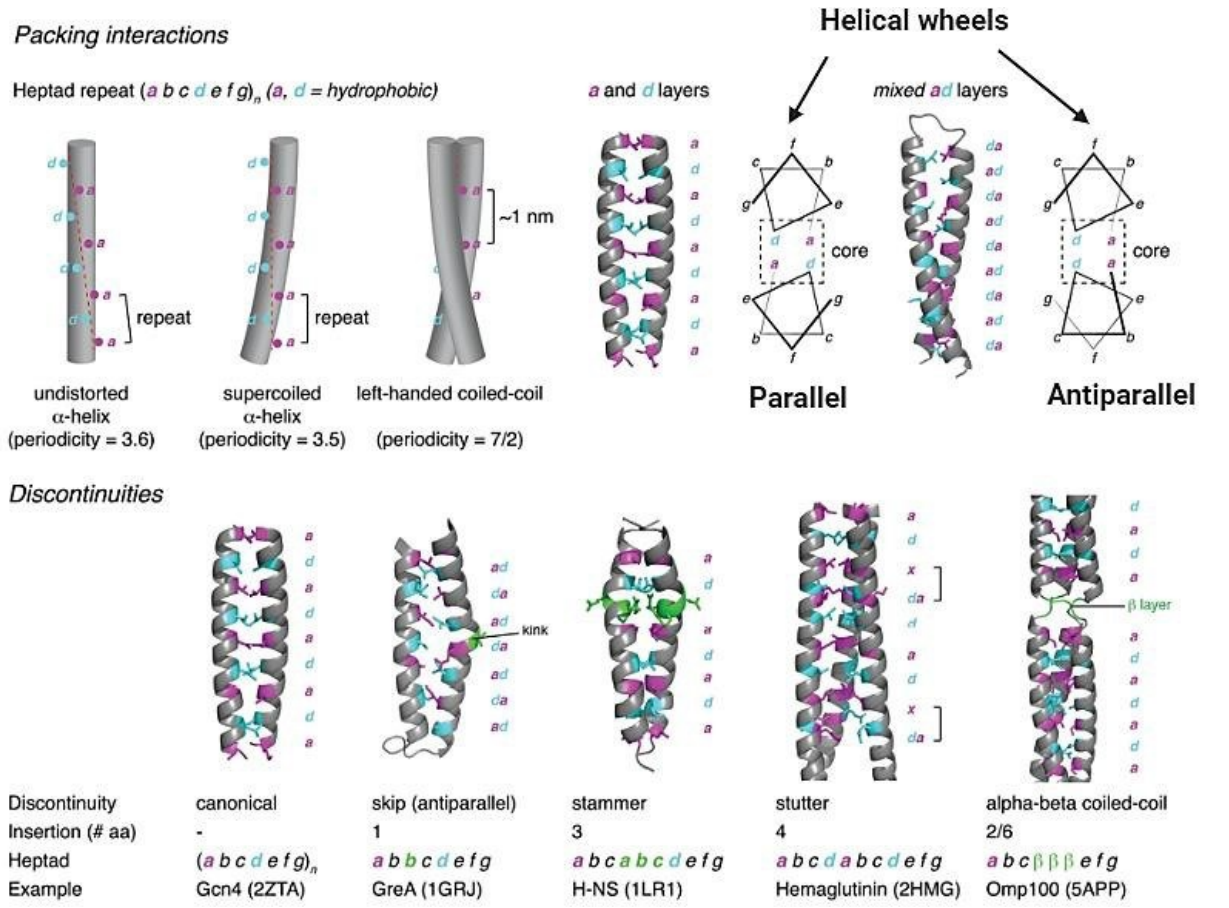


Figure 1.6. Helical wheels in coiled-coil structures and packing interactions. The heptad repeat is labelled as "abcdefg", with the hydrophobic amino acids placed in positions "a" and "d". The parallel and antiparallel helices' interactions differ (a-d for the parallel helix core and aa/dd for the antiparallel helix core). Amino acids insertions from the canonical coiled-coil (such as GCN4) can introduce kinks and structural modifications, as encountered for the structures of GreA, H-NS, Hemaglutinin, or Omp100. (Figure adapted in BioRender from Truebestein and Leonard, 2016).

A canonical coiled-coil is characterised by the conserved hydrophobic amino acids in the positions "a" and "d". Undistorted alpha-helices cannot establish a good packing between the interfaces due to the non-integral periodicity of the helix, which is reduced to 3.5, which corresponds to a 7/2 periodicity of a canonical left-handed coiled coil. The known "knobs-into-holes" packing of two parallel supercoiled helices dispose of "a" and "d" layers, whilst the antiparallel ones exhibit mixed "ad" layers. There can be exemptions from the canonical coiled-coil structures, by the insertion of 1,2,3,4, or 6 amino acids, which can cause structural deformations (Truebestein and Leonard, 2016).

Crick postulated that coiled-coil structures would be defined by the radius and pitch of a superhelical assembly of two or multiple α-helices, in conjunction with the interface angle as a contribution factor to the overall twist between the constituting helices. He developed then equations to

define those parameters, now available to generate structural scaffolds (Crick, 1953; Dawson, Rhys and Woolfson, 2019), and which can be used as parametrised structural templates to further design membrane proteins, as exemplified in Figure 1.7. The versatility of this model has been proven in multiple studies (Zhang *et al.*, 2009; Zhang *et al.*, 2015; Joh *et al.*, 2014), in which it was shown the successful application of a soluble protein-derived structural scaffold for membrane protein design.

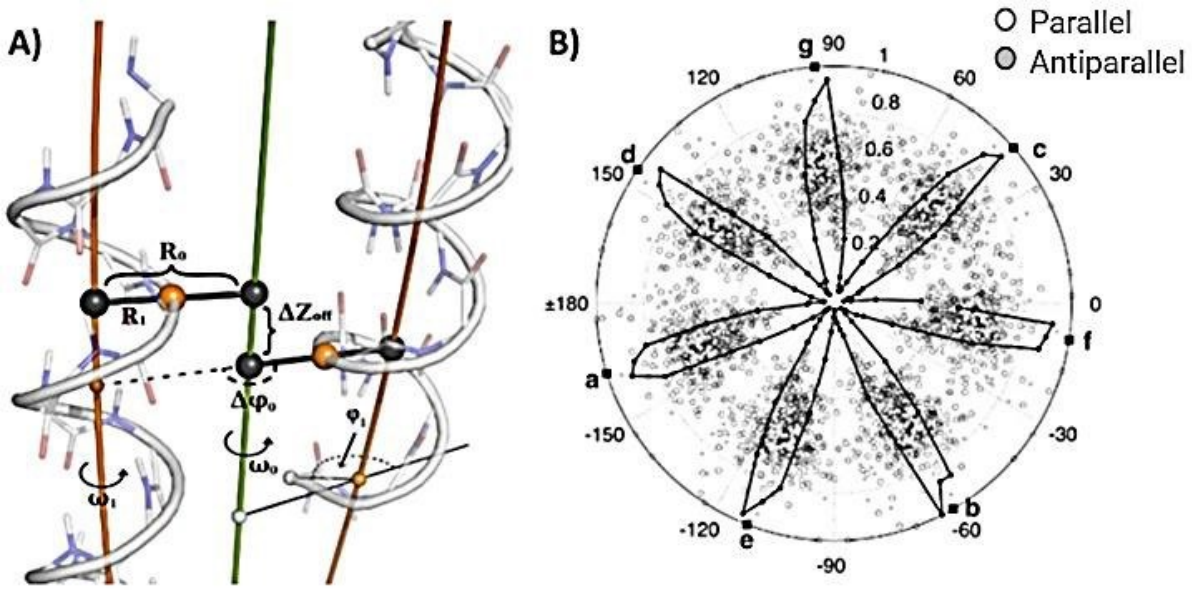


Figure 1.7. Visual representation of the coiled-coils parameters and application to parallel and antiparallel helices. A. Visual representation of the coiled-coil parameters. B. Distribution of the helical phases, namely parameter Φ_1 in parallel (white spheres) and antiparallel (grey spheres) coiled-coils and the imprint trends for the phase angles and heptad positions. (Adapted from CCCP server https://grigoryanlab.org/cccp/). R_0- superhelical radius; R_1- helical radius; ω_0- superhelical frequency; ω_1- helical frequency; ΔZ_{off}- chain axial offset; $\Delta\Phi_0$- chain superhelical offset; $\Delta\Phi_1$- starting helical phase.

1.8.3. Design of four-helix bundles

Four-helix bundles are favoured when creating new membrane protein structures because they possess natural stability, precise configuration, and adaptable functionality. These characteristics present numerous benefits that render them as prime choices for developing proteins tailored for membrane settings.

Four-helix bundles are known for their stable and resilient structure, which makes them ideal for designing membrane proteins, especially in situations where robust stability is essential for their effective performance within the lipid bilayer surroundings (Bilgiçer and Kumar, 2004).

The distinct architecture of four-helix bundles provides the foundation for the precise design and modification of membrane proteins to fulfil specific functions. This structural consistency is vital for formulating functional proteins with the desired characteristics (Lombardi *et al.*, 2019).

Moreover, their modular nature allows for the integration of various domains or motifs to confer specific functions, making them valuable for designing proteins tailored to different biological processes or applications (Merljak, Malovrh and Jerala, 2023).

The successful achievement of the design of four-helix bundles was based on the use (or design) of amphipathic helices with a predicted high tendency of self-association. The relevance of this type of structure can be explained through the importance of ion channels and heme-binding catalytic proteins. The minimalistic approach proved to be efficient for the design four-helix transmembrane bundles, such as a transmembrane dimer of dimers Zn^{2+} transporter (Joh *et al.*, 2014; Joh *et al.*, 2017), a homotetrameric non-biological cofactor binding transmembrane protein PRIME (PoRphyrins In MEmbrane) (Korendovych *et al.*, 2010), and a heme-binding antiparallel tetrahelical protein based on a leucine minimal sequence, REAMP (Recombinantly Expressed Artificial Membrane Protein) (Lalaurie *et al.*, 2018).

The use of four-helix bundles as structural scaffolds for membrane protein design necessitates a comprehensive understanding of the limitations that accompany this approach. Challenges range from accurately predicting ligand binding events to ensuring the proper alignment of transmembrane domains within the membrane environment. Notably, a key limitation pertains to the potential misinterpretation of ligand or ion interactions in the absence of the actual molecules, leading to potential inaccuracies in the projected protein structures (Joh *et al.*, 2017).

Another challenge in formulating four-helix bundles originates from overlooking the orientation of transmembrane domains within the membrane, directly impacting protein domain alignment accuracy. More specifically, the challenge is imposed by the complexity of the possible steric impediments exhibited within the inter-helical interfaces by the structure of the amino acids present in the transmembrane domains, which could impact the protein-protein association, thus, the final structural conformation that needs to be achieved for a specific function. This oversight undermines the portrayal of protein positions within the membrane. The intricate interplay of structural components with the membranes surroundings amplifies the complexity of the task at hand. An in-depth understanding of transmembrane domain spatial arrangement is indispensable to guarantee precise protein structural arrangement and optimise functionality (Sowlati-Hashjin, Gandhi and Garton, 2022).

Some of the strategies to overcome the limitations associated with the use of four-helix bundle scaffolds could include the following presented. The use of computational design principles to stabilise specific structural states, such as the cyclic C2-symmetric states and dihedral D2-symmetric states (Zhou and Lu, 2022). C2-symmetry, known as two-fold rotational symmetry, appears identical upon a 180° rotation around a single axis, and the protein's structure can be divided into two identical halves by a

division plane, and it is observed in homodimeric and heterodimeric proteins. On the other hand, D2 symmetry involves three perpendicular two-fold rotational axes, and the structure appears the same when rotated by 180° around two axes and is often encountered in complex assemblies. Such symmetries would address the challenge associated with the stability and orientation in the membrane of the TM domains. In addition, the experimental (computational) assessment of the helix-helix association can help validate a certain sequence or provide insights referring to de-stabilising sequence motifs (Joh *et al.*, 2017).

1.8.4. Antiparallel helix orientation

The antiparallel orientation of helices is prevalent in the structures of membrane channels. The role of antiparallel helices in membrane channels is multifaceted, contributing to the structural stability, functional specificity, and dynamic behaviour of these essential biological entities.

A computational study on the electrostatic associations of membrane-spanning α-helices modulated by dielectric media underscores the importance behind the antiparallel arrangement present in two adjacent α-helices within the voltage-gated potassium ion channel protein. This specific structural orientation appeared to be significantly impacted by the charged residues interacting within the multi-dielectric environment, thereby giving rise to diverse conformations of the ion-channel protein (Mapder and Adhya, 2013). This result suggests that antiparallel helices contribute to the structural stability of the channel and allow for functional flexibility in response to environmental changes.

1.8.5. Choice and distribution of amino acids- Sequence design

The allocation of amino acids in membrane proteins reveals specific patterns indicative of the proteins' structural and functional needs. This distribution is influenced by diverse factors, such as the hydrophobic nature of amino acids, the imperative for intrahelical and helix-helix structural stability within the lipid bilayer, and the multifaceted functional roles of the proteins encompassing gating transport and signalling mechanisms.

Moreover, as a polypeptide exits the translocon, it can either remain in the aqueous compartment or embed in a lipid bilayer. From a thermodynamic point of view, the insertion of a membrane protein helix is dominated by the hydrophobic effect, and the insertion capacities will reflect how well the designed sequence matches the properties of a lipid membrane (Corin and Bowie, 2022). The hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins are crucial for their structure and stability. Hydrophobic amino acids tend to be located within the transmembrane regions to

interact favourably with the lipid bilayer, while hydrophilic residues are often found in the aqueous parts or at the interfaces of the membrane proteins.

In the intricate structure of membranes, specific amino acid residues, such as charged (e.g. lysine, arginine) or polar (e.g. serine, threonine, asparagine), are commonly situated on the outer regions, known as the head groups. They exhibit a high tendency to interact with the phospholipid head groups and glycerol backbones in this location. Conversely, within the hydrophobic core region, amino acids like isoleucine, leucine, phenylalanine, and valine demonstrate a pronounced affinity for associating with lipids. Conversely, smaller and polar amino acid residues are typically concealed within helical bundles, avoiding lipid interactions due to their strong aversion to lipids. This segregation of amino acids based on their size and polarity contributes significantly to the rational design of stable structures, able to embed in a lipid bilayer (Adamian *et al.*, 2005). These considerations as an approach for the protein rational design led to successful both experimental and computational results. One example of a successfully *de novo*-designed transmembrane protein is a Zn^{2+} -transporting four-helix bundle, reported in the studies of Joh and colleagues (Joh *et al.*, 2014), where it is highlighted that the packing of large apolar amino acids, such as phenylalanine, determined a weaker association between the helices, and conversely, the helices which had in composition small amino acids, such as alanine residues, were more efficiently associated. Other studies that describe the importance of the amino acid size for the packing and structure stability, and also the sufficient driving force for folding provided solely by apolar side chains, was also emphasised in the computational designs of variants of phospholamban by Mravic and colleagues (Mravic *et al.*, 2019). Studies performed for redesigning transmembrane helices and converting them into water-soluble variants represent another proof of how the amino acid propensity can dictate the protein's preferred environment. A successful design in this sense was the conversion of the tetrameric integral membrane region of the bacterial potassium ion channel KcsA into its water-soluble variant, by designing 29 exterior mutations in each of the transmembrane domain units (Slovic *et al.*, 2004).

The achievement of a close helical packing is also another essential factor to take into consideration. This can be achieved through the use of statistically determined amino acid motifs, such as GxxxG, or SmxxxSm, at specific positions within a heptad or generally to follow a knobs-into-holes (KIH) approach for the favourable sterical arrangement and interdigitalisation of the amino acids side chains (Mackenzie, Prestegard and Engelman, 1997; Langosch and Heringa, 1998).

The amino acid bilayer propensity, beyond just the hydrophobicity characteristic, which took into account the biological factors that influence membrane protein topology and insertion, was highlighted in the studies of von Heijne and colleagues. In their studies, they developed an experimental approach to

measure the amino acid propensity within transmembrane helices of a model protein, which would be inserted into biological lipid bilayers, taking into account the influence of the complex interactions with the translocon machinery. The experiments included systematic variations of individual amino acids and resulted in providing quantitative free energy values for membrane insertion of each of the amino acid type, establishing the propensity of amino acids for membrane insertion (Hessa *et al.*, 2005).

Lately, with the aid of coarse-grained molecular dynamics, Sansom and Thomas released the MemProtMD database (Newport, Sansom and Stansfeld, 2019), available online (http://memprotmd.bioch.ox.ac.uk) in which there are represented the specific amino acids propensities in different regions of the membrane, based on the coarse-grained simulations of membrane proteins of known structure.

A successful *de novo* design of membrane proteins should not only consider the type of amino acids that will form the sequence but also the characteristics of the membrane, namely the depth of the membrane. In native proteins, the length of the transmembrane helices ranges between 20-26 amino acids on average, a length which allows the helix to span the hydrophobic region of a typical membrane bilayer. Factors such as helix tilting or lipid bilayer distortion can be correlated to variation in length, with observations describing length in the range between 15-43 residues (Daley *et al.*, 2005), studies even describing for a model peptide that a minimum of 13-14 consecutive hydrophobic residues was necessary for membrane insertions, describing the dependence of the length on the composition (more hydrophobic sequences can insert with shorter lengths) (Baeza-Delgado *et al.*, 2016).

It was also determined that the helix should not only be amphiphilic to provide the desired transmembrane interactions with both regions (polar and hydrophobic) but to have a certain degree of hydrophobicity in order to allow the partitioning in the lipid bilayer in almost a vertical orientation (Lear, Wasserman and Degrado, 1988). The topology can be controlled by the presence of specific amino acids with special characteristics such as lysine, known for its stretching structure and “snorkelling” behaviour towards the polar head groups (Strandberg and Killian, 2003), as well the presence of positive charged amino acids such as lysine and arginine guide the orientation of the helix in a membrane (Von Heijne, 2006), based on the “positive-inside” rule. The inner membrane leaflet is considered to be slightly negative and whilst the statistical analysis on the distribution of amino acids is not definitive, there tends to be a bias of positively charged residues in the inner leaflet.

These factors represented steps to consider during the knowledge-based rational *de novo* design of a membrane protein and the obtained candidates to be further tested for packing, stability, association, and orientation using computational techniques.

1.9. Computational biology

In order to test the accuracy of designs and their ability to exhibit the desired behaviours, throughout the years, the focus has been oriented on the development of computational tools to be able to predict the structure, the interactions, the stability and the function. Those are valuable for the research process, making it time, resource and cost-efficient by significantly enhancing the possibility of sampling thousands of different sequences. Those tools were developed in close relationship to the existing protein data bank, and the sampling is based on short fragments from existing structures, the assessment of different parameters and compatibility in the desired structure. Computational-based techniques highly contribute to the improvement of the speed of the studies of the structure and design of membrane proteins. Despite the ground-breaking advances in the field of artificial intelligence, and development of the well-known AlphaFold, membrane protein design computational workflow cannot rely (solely) on such tool. AlphaFold2 demonstrated high accuracy for the prediction of 3D structures at an unpreceded scale, and it can model multiple domains and interactions, providing insights into complex protein structures (Guo *et al.*, 2022). However, the limitations associated with AlphaFold are crucial, namely the data is not trained for the awareness of a membrane plan and also is not sensitive to point mutations, as communicated by its developers. This makes impossible correct modelling of the orientations of transmembrane (and other) domains within a membrane, essential for a membrane protein. Hence, combined alternative tools can be used to successfully achieve membrane protein design.

1.9.1. Energy function and Rosetta

One of the pioneering structure prediction computational-based tools is called Rosetta, developed by David Baker and co-workers, which was a top performer in the free modelling section of the international CASP competition (Critical Assessment of Structure Prediction) (Gront *et al.*, 2011; Mackenzie and Grigoryan, 2017). Initially, it uses contiguous fragments of 3-9 amino acid residues to analyse the torsion angles from the fragments to replace them in the modelled structure. This sampling method follows a Monte-Carlo simulation, which is an algorithm that relies on repeated random sampling, reducing the bias. Thus, the fragments compatible with the final structure are stochastically determined (Leman *et al.*, 2020). Furthermore, Rosetta's score function has been under continuous improvement over the years. The score function is a combination of scored bonded and non-bonded terms, describing van der Waals energies, hydrogen bonds, electrostatics, disulphide bonds, residue solvation, torsion degree within the backbone, sidechain rotation and reference energy associated with the unfolded state (Leman

et al., 2020).

Numerous innovative approaches have recently emerged within Rosetta, each addressing distinct scientific challenges. Efforts by developers have been dedicated to enhancing the accessibility and user-friendliness of Rosetta, resulting in the creation of intuitive interfaces like Foldit Standalone (Kleffner *et al.*, 2017). This interface assists designers in refining the sequence they've developed through the evaluation of substituted amino acids. Moreover, if the new structure exhibits greater energy levels than the original, even after factoring in temperature probability inputs, it visually identifies localised steric clashes that necessitate modifications to achieve heightened structural stability.

The constraints identified through the scoring function encompass the insufficient estimation of entropy (due to its quick sampling approach sacrificing accuracy), Monte Carlo fragment sampling derived from results of crystal structures, and the employment of an implied solvation model (which enhances the method's speed but limits the impact of explicit ions, water, and lipid models) (Leman *et al.*, 2020).

1.9.2. Implicit membranes and Hippo

For completing the sampling of the successfully designed membrane proteins, especially targeted as channels, the assessment of their orientation in the membrane is another essential step in the computational framework. This part of the study would represent not only the validation test of the distribution preferences of amino acids in the membrane, but it would also give an insight into the interactions interplay that amino acids exhibit on each other regarding the partitioning between aqueous and membrane environments.

Implicit membranes have been developed as valuable tools for the simulation of the orientation of proteins in membranes. These models comprise the essential features of the lipid bilayer, such as the dielectric feature, electrostatic interactions exhibited, and hydrophobic depth. The use of these models enables the enhancement of computational benchmarks against experimental targets and facilitates *de novo* design with amino acid distributions resembling naturally occurring membrane proteins. Based on the essential features and electrostatic interactions exhibited by the implicit membrane, the simulated protein adopts a certain tilt angle, which provides an insight of the possible behaviour *in vivo.* Based on the calculated tilt angle after the simulation completion, the quality of the sequence can be determined and, if needed, improved to adopt a tilt angle close to the relevant range. Such implicit membrane model is comprised in the Hippo software, developed by Ulmschneider and co-workers. The calculations used for the simulations are considered to be reliable, from the comparison between the simulated and

experimentally observed in X-ray and NMR studies of the tilt angles of naturally-occurring proteins (Ulmschneider, Sansom and Di Nola, 2005). Furthermore, knowledge-based sequence simulations in implicit membranes of various depths led to the expected results. The incompatibility of sequence length and membrane depth leads to modifications in the tilt angles against the membrane normal. Based on the amino acids sequence composition, artefacts in the electrostatic interactions between the protein and the implicit membrane would also contribute to modifications in the tilt angles. For example, a 22-amino acids sequence would exhibit a large tilt angle in an implicit membrane of 15 Å depth. Similarly, a sequence with hydrophilic or hydrophobic predominance would exhibit a large tilt angle. In this study, due to the sequence repeat and symmetry in the structure of CC1, the sequence of a single helix was simulated in the implicit membrane model of Hippo software. The sequence design rationale was supported by the obtained tilt angle values. Simulations in Hippo of a single helix provide fast, valuable information regarding the tilt without the need for validating the results for the rest of the composing helices if the helix would be part of a symmetric helical structure. If the structure that needs to be designed is desired to display higher complexity (heteromeric), the simulations of each individual helix would be useful solely to assess the compatibility of the chosen amino acid distribution in the helix with the dielectric features and hydrophobic depth of the implicit membrane.

Hippo is a user-friendly software and computationally inexpensive method. The sampling of a 22-amino acids sequence in multiple implicit membrane depths can be achieved in less than 30 minutes, with reduced CPU usage.

Recent software alternatives using implicit membranes were developed, such as Franklin2023 (F23). This method also considers the impact of the lipid head group using a mean-field-based approach. It also continues to utilise a depth-dependent dielectric constant in order to characterise the environment (Samanta and Gray, 2024).

1.9.3. Explicit membranes- Dynamics simulations

It is well-known that the folding process is highly influenced by the environment, especially the folding of water-soluble proteins. In the case of membrane proteins, this is still not clearly understood, however, it is believed that the lipid bilayer plays an important role in this regard. *In vitro* studies on outer membrane protein (OMP) tried to elucidate the role that the lipid environment plays in the folding process, and although there could be observed some influences of the large lipid-rich patches composition on the early folding steps, a clear understanding of the *in vivo* mechanism has not been achieved yet

(Horne, Brockwell and Radford, 2020). Similarly, studies performed on OMP in simplified systems, such as liposomes, indicated that the presence of an excess of negative charges on the outer leaflet impedes the insertion of the protein, and the folding is promoted when the inner leaflet is rich in negative charges (Machin *et al.*, 2023). Furthermore, it is of high interest that the dynamics of the proteins within the bilayer, the existing NMR or X-ray crystal structures disclose only snapshots of the structure, the series of conformational changes in relation to the environment, the most important in the understanding of how a protein folds, are lost.

In order to address this problem, further computational tools, to simulate the dynamics, were developed.

One computational technique is represented by the coarse-grained (CG) simulations. The CG simulations represent the bridge between the theoretical models and experimental observations. The CG simulations are used by researchers for various applications, such as the study of the folding free energies (Bell *et al.*, 2017), the study of protein-lipid interactions (Barreto-Ojeda *et al.*, 2018), membrane phase separation in plasma membranes (Liu *et al.*, 2021), etc.

The advantages of these simulations rely in the fact that they represent a computationally inexpensive technique, using simplified systems, by grouping multiple atoms as a single interaction site. Although they can be used to investigate certain trends in different biological phenomena of large energy-containing structures, the accuracy factor is diminished. From the aspect of the study of membrane protein folding, the lack of information assigned to the forces imposed by each atom, bond, possibly essential for the folding process, the use of the coarse grained simulations would need to be completed by more in-depth, all-atom simulations.

Molecular dynamics (MD) simulations represent a computational method used to analyse the physical movements of individual atoms and molecules. It is a more accurate method compared to CG simulations. These simulations involve the calculations of the forces acting on each atom, and then moving the atoms according to Newton's laws of motion. The simulation time is advanced in small increments (time steps) and repeated throughout the simulation time to observe the structural behaviours. A good example of a molecular dynamics package is represented by GROMACS, which can be successfully used in junction with Rosetta. This is able to simulate most atoms explicitly in rapport with the lipid environment and solvent, with a physics-based energy function based on Newton’s equation of motion, thus analysing the structural differences in the trajectory toward an energy minimum (Van Der Spoel *et al.*, 2005) influenced by different environment compositions.

In order to accurately represent the behaviour of the molecules, the energy terms in these

simulations are parameterised based on quantum-mechanical calculations and experimental data. The parameterisation involves defining force fields that describe the atomic forces governing the system. In molecular dynamics simulations, force fields represent essential empirical constructs comprising energy functions and parameters that assess the potential energy within a system according to molecular positions. They play a vital role in elucidating atomic and molecular interactions throughout simulations, encompassing components such as bond elongation, angle flexure, torsional movement, and non-covalent forces, like electrostatics and van der Waals interactions. These force fields are tailored to mirror the geometrical attributes and characteristics of examined structures, facilitating precise depictions of molecular actions within simulations (González, 2011). Some of the most widely used force fields are CHARMM, AMBER, OPLS, or GROMOS, each having individual parametrisation settings, however, they lead to similar force calculations results (Choi *et al.*, 2018). The pitfalls associated with the force fields are represented by the refinement requirements and the high computational demands, which restrict the routine simulations to short scales. However, the latest advances in the field of engineering have enabled access to super-computers in order to overcome these challenges (Hospital *et al.*, 2015), and to start to achieve timescales comparable to biological processes.

Moreover, recent advances in molecular dynamics simulations for protein folding studies include the use of machine learning techniques to refine the force fields, and the use of neural network algorithms for the exploration of kinetics (Musaelian *et al.*, 2023).

Furthermore, another key aspect of molecular dynamics simulations is the use of the periodic boundary conditions (PBC). These allow the simulation of the systems by extending the boundaries infinitely through infinite multiplication of the system's cells copies. This technique ensures that each molecule interacts realistically with its neighbours, even if they are on opposite sides of the simulation box, preserving important thermodynamic properties like temperature, pressure, and density (Kopec and Gapsys, 2022). By employing PBC, atoms which leave the simulation cell are replaced by exact atom copies entering from the opposite cell face, conserving the number of atoms in the cell and removing surface forces. The convention is applied in order to ensure that each atom interacts with at most one image of every other atom in the system, preventing self-interactions or interactions with multiple copies of atoms (Kopec and Gapsys, 2022).

The choice of the pressure coupling in the system can vary depending on the relevance it has for the simulation of the system. In molecular dynamics simulations, the pressure can be semi-isotropic, isotropic, or anisotropic. The semi-isotropic pressure coupling allows the modifications of the pressure along x-y axes, the isotropic system maintains the same pressure across all directions and the anisotropic

coupling system enables independent control of the pressure along each axis (Kandt, Ash and Peter Tieleman, 2007). In the context of the simulation of the dynamics of a membrane protein embedded in a lipid bilayer, the semi-isotropic pressure coupling system enables the study of the conformational changes and is useful in the cases when the membrane is expected to deform differently. The anisotropic pressure coupling system enables the study of highly complex systems that exhibit non-uniform pressure distributions, for example, for the simulation of the dynamics of a membrane protein in a detergent micelle.

Due to the complexity of calculations within a molecular dynamics simulation, to avoid errors, the calculations are often grouped first for the movement of separate structural entities through the imposing of positional restraints.

Positional restraints in molecular dynamics simulations are a technique used to constrain the movement of specific atoms or molecules within a system. These restraints are applied by fixing the positions of selected atoms or groups of atoms, preventing them from adopting different conformations during the simulation. The purpose of using positional restraints is to control the dynamics of certain parts of the system, allowing researchers to focus on specific interactions or behaviours of interest.

Positional restraints are commonly employed in various scenarios in molecular dynamics simulations. For example, in protein-lipid interactions, positional restraints may be applied to maintain the structure or position of proteins within a lipid environment (Jiang, Lacroix and Luo, 2022).

The output of the molecular dynamics simulations is comprised of a trajectory that offers a detailed view of how atoms or molecules move and interact within a system. By analysing these trajectories, researchers can gain insights into the structural changes, interactions, and motions of molecules at an atomic level. Trajectories are crucial for studying the behaviour of complex systems and observing how different components within the system evolve over time (Mohimani *et al.*, 2017).

From the trajectory it can be extracted valuable information, such as the root-mean-square deviation (RMSD). This is a common metric, describing how the positions of atoms deviate from an initial reference structure throughout the time of the simulation, providing insights into conformational changes, stability, and interactions within the system. The observation of a plateau in a RMSD plot often indicates that the system has reached equilibrium or a stable state (Knapp *et al.*, 2011).

The choice of the reference structure can impact the computing efficiency and not necessarily the accuracy of the RMSD analysis. For example, in a study of fast decoy clustering, the choice of reference root-mean-squared distance (rRMSD) over pairwise values (pRMSD) increased the computing efficiency. Researchers determined that the computational efficiency rose markedly without considerably impacting

the precision of selections approximating native structures when utilising rRMSD as the benchmark metric (Li and Zhou, 2005). Notably, this outcome highlighted that whether the choice of the reference structure is based on rRMSD or pRMSD, the accuracy of the results is not compromised.

The assessment of the RMSD does not provide information about the protein's energetic stability by only measuring the differences in the structural positioning and cannot detect the effect of environmental changes on protein stability, so the joint use of multiple computational techniques is essential.

1.10. Membrane protein expression

To date, there are numerous protein design studies for soluble proteins than for the membrane proteins. This could be explained by the introduction of a further challenging consideration point - the complex topological environment of membrane proteins represented by the lipid bilayer. Consequently, apart from the design and sampling of protein-protein interactions, intuitive protein-lipid interactions have to be employed in the experimental design. Despite the prevalent existence of structural motifs for soluble protein design, such as coiled coils, these motifs have become of interest for conversion to be able to span the membrane (Scott *et al.*, 2021), however the synthesis approaches cannot be transferred to *de novo* MP. The challenges in *de novo* membrane protein expression are primarily rooted in the highly hydrophobic character of multispanning transmembrane proteins, which complicates their production and characterisation. The expression of constructs with multiple predicted transmembrane segments in bacterial systems is notably limited, a challenge that mirrors the difficulties encountered with native membrane proteins. This limitation underscores the practical hurdles in the *de novo* design and study of membrane proteins, including issues related to solubility, stability, and functional characterisation (Tulumello *et al.*, 2012). Some of the studies involving *de novo* transmembrane proteins involve solid-phase synthesis, and less on the traditional cell-based expression. Solid-phase peptide synthesis (SPPS) is a versatile technique used for the synthesis of both *de novo*-designed membrane proteins and soluble proteins.

Solid-phase peptide synthesis (SPPS) entails the stepwise addition of amino acids to a developing chain attached to an insoluble resin. These peptides are constructed utilising either tertbutoxycarbonyl (Boc) or 9-fluorenylmethoxycarbonyl (Fmoc) protective strategies to shield the amino groups during the synthetic process (Dolphin, 2006). SPPS is mainly used for the synthesis of soluble proteins. Although it is a powerful technique, the general associated challenges and the nature of membrane proteins can interfere

for the successful application of this technique for the synthesis of *de novo* MP.

SPPS is typically suited for the synthesis of short, linear peptides, whilst MP typically have long domains, that need to insert and fold correctly into lipid bilayers. Furthermore, the TMs of MP are hydrophobic, and SPSS must be carefully managed to prevent the hydrophobic peptide chains from interacting in a non-directed manner, which is a great challenge for the use of SPPS in this case. The complex topology of MP limits the use of SPPS, which may struggle to correctly fold the proteins and additional steps would be required in the synthesis procedure (Niitsu and Sugita, 2023b).

Although there is a rising interest in computational-based protein design, including the prediction of the new protein's behaviour in natural-like simulated membrane bilayers, experimental biocompatibility for *de novo* designs represents a niche, especially for transmembrane helical bundles.

1.10.1. Advantages of cellular expression versus SPSS of de novo membrane proteins

Cellular expression and solid-phase peptide synthesis are efficient methods for producing proteins and peptides, each having advantages, depending of the complexity of the design desired to be expressed/synthesised, size of the design and the function that is intended to be assigned to the specific protein. Compared to cellular expression, SPSS has been encountered more frequently in design studies of small peptides .The size of the design is usually one of the limitation of SPSS, typically up to 50-70 residues in length, the alternatives which could be involved to synthesise larger designs would impose cost limitations, would be laboriously intensive and subject to errors, due to possible incomplete coupling (Mueller *et al.*, 2020).

One of the advantages that cellular expression exhibits in this case, is the possibility of expression of larger and more complex proteins. Lu and colleagues (Lu *et al.*, 2018) have successfully expressed large designs (dimeric and tetrameric designs, of up to 860 residues in length) using bacterial and mammalian cell-based expression.

Another advantage that cell-based expression has is that it confers the natural lipid bilayer, the necessary environment for proper folding, insertion and function, which is not provided by SPSS, which can often lead to non-functional or aggregated protein. In addition to the lipid environment required for folding, the cellular machinery, through its designated compartments (translocon, chaperones), helps reducing the risk of sequence or folding errors, which could take place in the case of SPSS.

1.10.2. Expression challenges and how to overcome them

The cell-based expression and purification of newly created transmembrane peptides require specific techniques designed to overcome the inherent difficulties encountered in producing membrane proteins. One methodology entails utilising cloning methods from molecular biology to generate a collection of partially randomised *de novo* helix-loop-helix transmembrane constructs rich in alanine and isoleucine. These synthetic protein hairpins act as templates for studying the folding of transmembrane structures. The purification strategies utilised for these engineered protein structures can be adjusted for diverse purposes. However, the efficiency of bacterial expression for constructs containing multiple anticipated transmembrane regions remains restricted, akin to natural membrane proteins (Tulumello *et al.*, 2012).

Escherichia coli (*E. coli*) represents the most used expression host system due to its ease of manipulation, cost-efficiency and fast cell growth (Hattab *et al.*, 2015), achieving protein expression in 2-3 days starting from plasmid transformation. It also represented the host target for the expression of some re-designed soluble miniproteins or protein binders (Haven *et al.*, 2022), reflecting the versatility of this cell host. The most widely used *E. coli* strains for protein expression are BL21 (DE3), C41(DE3), or C43 (DE3). These strains are derivatives of *E. coli* B and have been engineered to contain the bacteriophage T7 RNA polymerase gene under the control of the lac UV5 promoter, which allows for the inducible expression of proteins from plasmids carrying a T7 promoter. The BL21(DE3) strain is often chosen as the original strain for protein expression because it tightly regulates the expression of harmful proteins through the T7 expression system. Originating from BL21(DE3), the C41(DE3) and C43(DE3) strains known as the Walker strains were selectively developed for better growth and enhanced protein expression especially for proteins harmful to cells. These strains have mutations that reduce T7 RNA polymerase activity compared to BL21(DE3), helping to mitigate the toxicity associated with overproducing specific proteins, notably membrane proteins. The modifications in C41(DE3) and C43(DE3) allow for the production of higher levels of functional recombinant membrane proteins that would otherwise be toxic to the cell, making them valuable tools for the study of membrane protein structure and function (Heyde and Nørholm, 2021; Baumgarten *et al.*, 2017).

Beyond bacterial systems, other expression hosts, such as insect cells using the baculovirus expression vector system, have been utilised for the production of membrane proteins. This is because inclusion bodies are rarely formed in insect host cells (Jarvis, 2003). However, there are very limited to no studies reporting this for *de novo*-designed proteins.

The growth environment, such as variations in the temperature for the growth of cells expressing different proteins, influenced membrane protein expression. The studies of Bonander and colleagues focused on identifying the reasons which stand behind a successful recombinant membrane protein (Fps1) in yeast. It was observed that temperature, along with pH, had a significant impact on the membrane protein yields, the best growth condition for maximum total expressed protein yield at maximum tested temperature (35°C, from the range between 30-35°C), whereas variations in the pH alter the yields of membrane-bound protein yields (Bonander *et al.*, 2005). Studies on *E. coli* cells have shown that growth temperature influences the lipid composition of membranes, which further impacts the membrane protein insertion and folding, indicating that at temperatures lower than 21°C, due to modifications in the lipid composition, the outer membrane protein biogenesis was affected (Horne, Brockwell and Radford, 2020). The growth conditions could be screened, so that they can be identified the optimum conditions for the best yields for each expressed protein, as indicated in some studies performed on the fungal pathogens (Leach and Cowen, 2014).

If the expression is considered toxic by the cell, it is sent to inclusion bodies. The generation of membrane proteins within the cytoplasmic membrane instead of within inclusion bodies in a misfolded state is generally preferred. However, this preference is often impeded by the saturation of the Sec translocon machinery capacity, leading to reduced production yields. In order to avoid the translocon machinery saturation, alteration to the expression strategy can be performed such as reducing or eliminating the addition of IPTG (Z. Zhang *et al.*, 2015), decreasing the protein expression rate.
Further alternatives can be employed such as modifying the expression plasmids which encode the protein gene, adding extra promoters, proteins, or tags.

Previous studies have shown that, alongside the T7 promoter, the PBAD promoter that responds to arabinose can be utilised to amplify the production of toxic membrane proteins in *E. coli.* The PBAD promoter forms a part of the araBAD operon responsible for arabinose metabolism and is subject to strict control mechanisms. This promoter becomes active in the presence of arabinose but is inhibited by glucose due to catabolite repression. Studies have displayed that a modified version of the PBAD promoter collaborates effectively with engineering methods involving chaperone pathways to boost the expression of membrane proteins that exhibit toxicity upon excessive expression. Utilising the PBAD promoter enables precise regulation of toxic protein expression by adjusting the arabinose concentration in the growth medium. This strategy is especially beneficial for the expression of challenging membrane proteins that have a tendency to aggregate or are detrimental to the host cell in excessive amounts. Through modulation of induction levels, it becomes possible to fine-tune the protein expression and to balance the

protein yield and cell viability (Nannenga and Baneyx, 2011).

Protein fusion tags, such as NT11 have been reported to enhance protein expression in *E. coli* (Nguyen *et al.*, 2019), as well as large fluorescent molecules, such as green fluorescent protein (GFP) (Lalaurie et al., 2018). GFP can also establish the localisation of the expression, quantify the expression level, and track the target protein in downstream processes, such as solubilisation and purification.

Moreover, integrating the gene encoding the mistic protein into the expression plasmid can also contribute to enhancing the expression of *de novo* MP, promoting membrane integration and stability, all while circumventing the translocon machinery, as shown in previous studies (Ananda Chowdhury, Rentian Feng, Qin Tong, Yuxun Zhang, 2014). Mistic naturally integrates into the membrane without the need for a signal sequence, and it has been shown to increase the expression levels of otherwise poorly expressed membrane proteins, when fused to them.

Cell-free protein expression has emerged as a powerful tool for the production of membrane proteins, offering several advantages over traditional cell-based expression systems. This technique involves the use of cell lysates or purified components from cells to carry out transcription and translation processes *in vitro*, allowing for the rapid and efficient synthesis of proteins without the need for living cells. This approach is especially useful to synthetic biology as it has the potential to overcome the challenges associated with expressing membrane proteins in living cells, such as toxicity, low yields, and difficulties in purification (Sachse *et al.*, 2014).

Membrane proteins, in particular, often require specific lipid environments and can cause cell death when overexpressed *in vivo* (Bernhard and Tozawa, 2013). Cell-free systems circumvent this issue by providing a controlled environment for protein synthesis, where the necessary components for transcription and translation are supplied in the reaction mixture. This approach allows for the expression of membrane proteins that might otherwise be impossible to produce in living cells, enabling their study and characterisation (Sachse *et al.*, 2014). An example of a successfully cell-free synthesised *de novo* protein was reported in 2023 in the studies of Fujita and colleagues, describing that the expression of a stable novel β-barrel nanopore, called SVG28, was achieved with the aid of the cell-free technology (PURE system) (Fujita, Kawamura and Kawano, 2023). The protein also maintained its function, as electrophysiology experiments have shown. Another example of successfully cell-free expressed proteins are a set of *de novo* antimicrobial peptides, which were designed with the aid of generative and predictive deep learning algorithms (Pandi *et al.*, 2023).

1.11. Membrane protein extraction and solubilisation

Most of the downstream processes, such as purification and biophysical studies of MP require the extraction and solubilisation of the membrane protein.

Detergents are historically the most used agents for membrane protein extraction and stabilisation (Stetsenko and Guskov, 2017). They are amphipathic molecules, presenting a polar or charged hydrophilic head group and a hydrophobic tail. Having these structural features, a detergent molecule is capable of disrupting the membrane, by insertion of the tails in the membrane and extracting at higher concentrations.

Generally, the detergent solubilisation mechanism occurs in three steps. The first step comprises the destabilisation of the surrounding lipid environment, and they are able to bind to the membrane protein found in the targeted environment. The detergent structural component of the membrane protein responsible for the binding to the membrane protein is represented by the hydrophobic tails. Hydrophobic interactions are established between the detergent tail and the hydrophobic transmembrane domains. The detergent molecules surround the MP. They are known to form spontaneously pseudo-spherical assemblies, also known as micelles. The spontaneous event occurs as soon as the appropriate concentration of detergent is achieved, known as critical micelle concentration (CMC). There are different classes of detergents used for solubilising membranes, such as non-ionic (e.g. dodecyl maltoside (DDM), decyl maltoside (DM), octyl glucoside (OG)), ionic (e.g. sodium dodecyl-sulfate (SDS)), or zwitterionic (e.g. lauryl dimethylamine oxide, FOS-Choline 12) (Kalipatnapu and Chattopadhyay, 2005). An example of an extensively used ionic detergent in studies on membrane proteins is SDS. It is a harsh, denaturing detergent, disrupting non-covalent bonds and through binding to the protein, it imparts a uniform negative charge, which allows the separation of proteins from a protein-detergent complex based on their molecular weight in techniques such as SDS-PAGE (Winogradoff, John and Aksimentiev, 2020). The micelles formed by SDS can mimic the lipid bilayer, allowing the studies of structures and function, as observed in studies on the stability of the structures of the β-barrel proteins from the bacterial outer membranes, which have been shown to maintain stability in SDS (Horne, Brockwell and Radford, 2020). Ionic detergents have in their structure a charged (positive or negative) polar head group and a hydrophobic tail. Non-ionic detergents are generally mild and non-denaturing. However, they can give low solubilisation yields. When the micelle is formed, the protein is included in a detergent-protein complex. In this complex, lipids can be partially retained, or sometimes completely removed, modifying this way the lateral pressure exhibited on the protein, thus destabilising the structure in some cases (Van Den

Brink-Van Der Laan *et al.*, 2004). If the characteristic CMC of a detergent is high, it represents a cost-ineffective choice for MP solubilisation. In general, a screening of detergents would imply a high financial cost for the experiments, but considering that *de novo* MP would implement unique structural and interaction features, there is no "one fits all" solution. In general, it is desirable are the detergents with a relatively low CMC. The most widely used, DDM, has a CMC of approximately 0.01%, and DM has a CMC of approximately 0.087%. Considering the reported features of the polyoxyethylene (8/9) dodecyl ether, its low CMC (0.003-0.005%), lower than the ones of DDM and DM and the formation of large micelles (Stetsenko and Guskov, 2017), can represent a better solubilisation alternative.

Recently, there have been significant advancements in the realm of creating new detergents and polymers that enhance the solubilisation and stabilisation of membrane proteins. Notably, Calixarene C4C7 and the polymer SMA 2000 have demonstrated greater efficacy in solubilising and stabilising the membrane proteins, such as ABCC4/MRP4, than traditional detergents. Furthermore, styrene-maleic acid (SMA) polymers have been developed for detergent-free extraction of membrane proteins enabling direct extraction from the membrane while preserving their natural lipid environment (Bada Juarez *et al.*, 2019), however exhibiting sensitivity towards divalent cations, causing precipitation. Moreover, cycloalkane-modified amphiphilic polymers (CyclAPol), for example, have been used to directly extract membrane proteins from the membrane, enabling one-step purification and high- resolution structure determination by cryo-electron microscopy (Higgins *et al.*, 2021). So far, the MP structures captured in discs containing partially the surrounding lipid environment, couldn't be used in functional studies, such as ion transport. In Figure 1.8 there are represented some examples from each of class of solubilisation agents mentioned above.

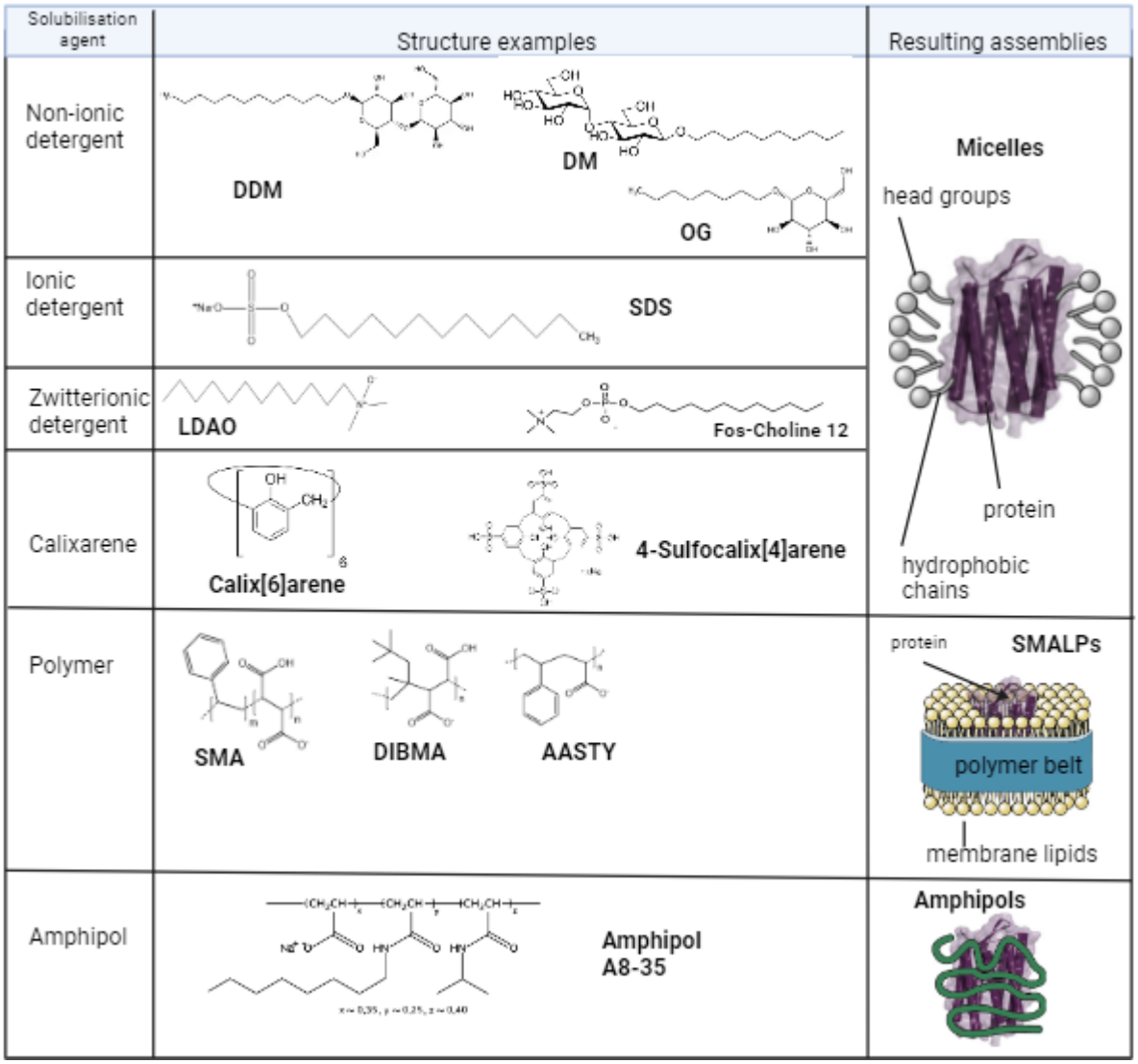


Figure 1.8. Examples of solubilisation agents. Ionic detergents (e.g. DDM, DM, OG), ionic detergents (SDS), zwitterionic detergents (LDAO, Fos-Choline 12), and calixarenes form micelles as assemblies that model the membrane. Polymers (e.g. SMA, DIBMA, AASTY) form SMA lipid particles (SMALPS) and amphipols extract the protein by direct interaction, wrapping around the protein. (Created with BioRender and ChemDrawPro).

1.12. Reconstitution platforms for the study of protein function

Often, the MP study often requires reconstitution into model membrane systems that mimic the natural environment of these proteins. The incorporation of membrane proteins into liposomes is a critical step in studying their structure, function, and interactions in a native-like environment.

The significance of incorporating membrane proteins into liposomes lies in the fact that the lipid bilayer provides a hydrophobic environment that is essential for maintaining the structural integrity and function of these proteins (Lee and Kim, 2018). Also, liposomes allow size-dependent

compartmentalisation, allowing the study of confinement effects on proteins, can maintain proton, electrochemical and solute gradients and can provide selective permeability (Schmitt *et al.*, 2016). The amphipathic nature of membrane proteins makes them inherently difficult to study, especially considering that different membrane proteins may need the allosteric influence of a library of lipids to function or, in some cases, even fold. Whilst liposomes themselves require additional extra for their production, they can provide a detailed, molecular-level platform for understanding membrane protein folding, especially *de novo* designed protein (Carlson *et al.*, 2012). Once incorporated into liposomes, membrane proteins can be studied using a variety of biophysical and biochemical techniques. For example, the activity of ion channels or transporters can be monitored using electrophysiological recordings or fluorescence-based assays (Bazzone, Barthmes and Fendler, 2017). The binding of ligands or inhibitors to receptors or enzymes can be investigated using radioligand binding assays or surface plasmon resonance (Patching, 2014). The conformational dynamics and structural changes of membrane proteins can be studied using spectroscopic techniques, such as fluorescence resonance energy transfer (FRET) or electron paramagnetic resonance (EPR) (McHaourab, Steed and Kazmier, 2011).

1.12.1. Liposome incorporation of membrane proteins

The incorporation of membrane proteins into liposomes is a critical step in studying their structure, function, and interactions in a native-like environment.

The significance of incorporating membrane proteins into liposomes lies in the fact that the lipid bilayer provides a hydrophobic environment that is essential for maintaining the structural integrity and function of these proteins (Lee and Kim, 2018). The amphipathic nature of membrane proteins makes them inherently difficult to study, especially considering that different membrane proteins may need the allosteric influence of a library of lipids to function, or, in some cases, even fold. Whilst liposomes themselves require additional extra for their production, they can provide a detailed, molecular-level platform for understanding membrane protein folding, especially *de novo*-designed protein (Carlson *et al.*, 2012).

Whilst methods have been described for the inclusion of proteins into membranes, e.g. detergent-mediated reconstitution, where the protein is first solubilised in detergent micelles and then mixed with lipids to form proteoliposomes upon removal of the detergent (Seddon et al., 2004), the presence of liposomes using cell-free expression systems has also been developed with great success (Sachse *et al.*, 2014).

Once incorporated into liposomes, membrane proteins can be studied using a variety of

biophysical and biochemical techniques. For example, the activity of ion channels or transporters can be monitored using electrophysiological recordings or fluorescence-based assays (Bazzone, Barthmes and Fendler, 2017). The binding of ligands or inhibitors to receptors or enzymes can be investigated using radioligand binding assays or surface plasmon resonance (Patching, 2014). The conformational dynamics and structural changes of membrane proteins can be studied using spectroscopic techniques, such as fluorescence resonance energy transfer (FRET), or electron paramagnetic resonance (EPR) (McHaourab, Steed and Kazmier, 2011).

1.12.2. Protein insertion efficiency

The efficiency of membrane protein insertion into lipid bilayers is a critical factor in the study of these important biomolecules, as it directly influences the yield, functionality, and stability of the reconstituted proteins (Sachse *et al.*, 2014). The process of membrane protein insertion is governed by a complex interplay of factors, including the physicochemical properties of the lipid bilayer, the structural features of the protein, and the specific lipid-protein interactions that facilitate the insertion and folding of the protein into its native conformation (Cymer, Von Heijne and White, 2015). Understanding the mechanisms underlying membrane protein insertion and the factors that modulate its efficiency is crucial for optimising the production and characterisation of these proteins in both research and biotechnological applications (Rajesh, Knowles and Overduin, 2011).

One of the key determinants of membrane protein insertion efficiency is the composition of the lipid bilayer. The specific lipid species present in the bilayer can significantly influence the ease with which a protein can insert and fold into its native conformation (Dowhan and Bogdanov, 2009). For example, the presence of non-bilayer-forming lipids, such as phosphatidylethanolamine (PE) or cardiolipin (CL), has been shown to facilitate the insertion of some membrane proteins by promoting the formation of transient non-lamellar structures that lower the energy barrier for protein integration (Bowie, 2005). These lipids can also assist in the folding of membrane proteins, by providing a more flexible and dynamic environment that allows for the rearrangement of transmembrane helices and the formation of native contacts (Booth and Curnow, 2009).

In addition to the lipid composition, the structural features of the membrane protein itself can also play a significant role in determining its insertion efficiency. The hydrophobicity, length, and amino acid composition of the transmembrane segments, as well as the presence of specific sequence motifs or topological signals, can all influence the ability of a protein to insert and fold into the lipid bilayer (Hessa

et al., 2007). For instance, the presence of positively charged residues in the flanking regions of transmembrane helices can promote their insertion, by interacting with negatively charged lipid head groups, while the presence of proline residues within the helices can introduce kinks or hinges that facilitate their packing and stabilisation (Cymer, Von Heijne and White, 2015).

The specific lipid-protein interactions that occur during the insertion process are also critical for determining the efficiency and fidelity of membrane protein integration. These interactions can involve both specific binding of lipids to defined sites on the protein surface, as well as more general physicochemical effects arising from the hydrophobic matching between the protein and the bilayer (Killian and Von Heijne, 2000). For example, anionic lipids such as phosphatidylglycerol (PG) or phosphatidylserine (PS) have been shown to interact with positively charged residues on the surface of some membrane proteins, facilitating their insertion and stabilising their native conformation (Dowhan and Bogdanov, 2009). Similarly, the degree of hydrophobic matching between the length of the transmembrane segments and the thickness of the bilayer can influence the tilt angle and packing of the helices, affecting the final structure and stability of the inserted protein (Killian and Nyholm, 2006).

The investigation of membrane protein insertion efficiency requires a multidisciplinary approach, combining biochemical, biophysical, and computational techniques (Rajesh, Knowles and Overduin, 2011). *In vitro* translation and insertion assays, such as those based on cell-free expression systems or reconstituted proteoliposomes, have been widely used to study the factors influencing membrane protein integration and to optimise the conditions for efficient protein production (Sachse *et al.*, 2014). These assays allow for the systematic manipulation of lipid composition, protein sequence, and other experimental variables, enabling the identification of key determinants of insertion efficiency (Cymer, Von Heijne and White, 2015).

Biophysical techniques, such as fluorescence spectroscopy, circular dichroism (CD), and Förster resonance energy transfer (FRET), have also been employed to monitor the kinetics and thermodynamics of membrane protein insertion and folding, providing insights into the underlying molecular mechanisms (Rajesh, Knowles and Overduin, 2011). These techniques can be used to probe the conformational changes and interactions that occur during the insertion process, as well as to assess the stability and functionality of the inserted proteins (Booth and Curnow, 2009).

Computational methods, such as molecular dynamics (MD) simulations and free energy calculations, have also emerged as powerful tools for investigating the energetics and pathways of membrane protein insertion (Cymer, Von Heijne and White, 2015). These methods allow for the atomic-level analysis of lipid-protein interactions and the identification of key residues or sequence motifs that

drive the insertion process, complementing the experimental data and providing mechanistic insights into the factors governing insertion efficiency (Ulmschneider *et al.*, 2014).

1.12.3. Droplet interface bilayers as reconstitution platform

Droplet interface bilayers (DIBs) have emerged as a powerful platform for studying membrane proteins in a controlled and biomimetic environment. The approach uses the formation of a lipid bilayer at the interface between two aqueous droplets submerged in an oil phase, providing a unique system for the reconstitution and characterisation of membrane proteins (Bayley *et al.*, 2008), as illustrated in Figure 1.9.

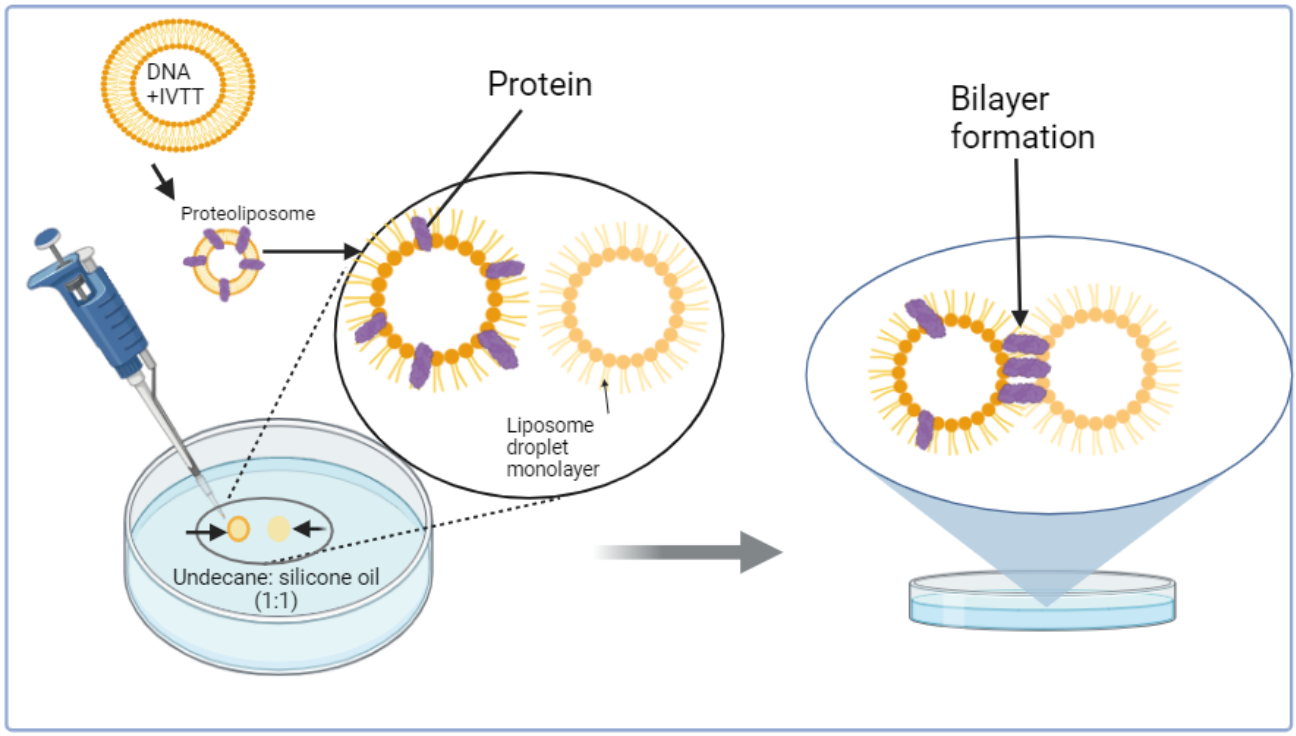


Figure 1.9. Schematic representation of DIB formation and potential application. (Adapted from (Caillon *et al.*, 2020)). The droplets are pipetted in a petri dish containing a mixture of silicone oil and undecane. The droplets can have a lipid composition of various compositions, and they orientate the tails towards the oil mixture and the heads towards the aqueous environment they contain (e.g. protein synthesis mix, liposomes, etc.). The droplets are brought into contact mechanically and the droplets' monolayers form a bilayer, which can influence the distribution of the protein between the bilayer and the monolayer, depending on the lipid composition. A small population (e.g. $1x10^{-4}$ % from total yield) of the cell-free synthesised membrane protein (which could be monomeric) inserts in the droplet interface bilayer, folds and becomes functional (Elfaramawy *et al.*, 2018). (Created with BioRender)

The ability to create stable bilayers with well-defined compositions and the potential for high-throughput screening make DIBs an attractive tool for investigating the structure, function, and

interactions of membrane proteins (Leptihn *et al.*, 2013).

Although similar at first sight to the black lipid membranes (BLMs), they are different. Black membranes are another method for reconstitution of membrane protein, called “black” due to their appearance when viewed under a microscope, due to their thinness, close to the wavelength of visible light. They also confer a stable, planar lipid bilayer. The BLMs, however, are formed by painting a lipid solution across an aperture or by apposition of monolayers from each side of the partition. They can be used in electrophysiology-based experiments for the study of ion channels (Haylock *et al.*, 2020). The key differences between BLMs and DIBs, apart from the formation mechanism, BLMs provide a 2D platform, whist DIBs offer a 3D model. DIBs also allow the construction of asymmetric, complex droplet networks, allowing the high-throughput screening of lipid compositions and providing scalability. Also, DIBs can more easily create compartmentalised structures, acting as artificial cells, or tissue architectures.

One of the key advantages of using DIBs for studying membrane proteins is the ability to form stable bilayers with precisely controlled lipid compositions. By carefully selecting the lipids used to form the monolayers surrounding the aqueous droplets, researchers can create bilayers that mimic the composition of specific cell membranes or investigate the effects of lipid composition on protein function (Venkatesan *et al.*, 2015). This level of control over the bilayer environment is particularly important for membrane proteins, as the lipid composition can significantly influence their stability, folding, and activity (Sezgin and Schwille, 2012). Moreover, the DIB platform allows for the formation of asymmetric bilayers, where each leaflet of the bilayer has a different lipid composition, enabling the study of protein-lipid interactions in more physiologically relevant contexts (Hwang *et al.*, 2008).

1.13. Aims and objectives of the project

The present study aims to contribute to membrane protein design and understand the governing forces of membrane protein folding.

1) Design of building blocks based on a completely novel, repetitive sequence in order to form stable structures, combining a knowledge-based, rational, minimal protein design approach.
2) Validation tests for stability and insertion of the designed sequence

In Chapter 2, it is presented the design process, combining statistical data, knowledge-based information, and the computational methods used for the test of design (CC1) stability and conformational changes in computational membrane models.

3) Testing bioavailability of *de novo* designed membrane protein

In Chapter 3 there are presented the strategies employed for molecular cloning, bacterial expression, solubilisation and purification of *de novo* membrane protein.

4) Testing functionality of the *de novo* membrane model

Chapter 4 presents the testing of reconstitution systems such as DIBs for further tests of functionality of the channel.

2. . *In silico* studies: The pathway to the design of *de novo* transmembrane bundle of helices and molecular dynamics simulations

2.1. Introduction

2.1.1. Synthetic biology for proteins

Synthetic biology approach differs from traditional protein engineering proposals that included repeated mutagenesis (Sezgin and Schwille, 2012). The research field of protein design is currently expanding; however, it is still a debated subject among scientists.

One of the most frequently asked questions related to protein design is: "Why design new protein structures when there are already nearly 200,000 protein structures resolved experimentally (up to the end of 2023 in Protein Data Bank, there are 182,595 protein structures from X-ray experiments) and a plethora of unresolved naturally occurring protein structures?". The fundamental argument of protein designers is based on the fact that nature could not sample all the possible sequences or structures for natural evolution. Furthermore, it is concluded from multiple statistical studies on the natural existing protein structures that the natural evolution process is rather a randomised one and not a directed, selective process. Considering this aspect as well, a vast majority of the possible sequence combinations have not yet been explored by nature. In this sense, *de novo* protein design has come to address some of the unexplored structures and to provide further elucidations regarding protein structure-function relationships (Woolfson *et al.*, 2015).

The repetitive characteristic moiety of proteins is the dipole represented by the amide linkage between amino acids. Initial protein design experiments were based on the mutations of different amino acids of a naturally occurring protein (Shuttleworth, W. A., Hough, C. D., Bertrand, K. P. & Evans, 1992) and on the assessment of any alterations in the protein's behaviour. Considering the diversity of the proteome and the complexity of the sequences and structures of naturally occurring proteins, the amino acid mutation approach would have resulted in a time-consuming cycle, which would hardly answer the scientific question behind the relationship between sequence-structure-function. Furthermore, it was formulated the hypothesis that the behaviour (i.e. folding, representing one of the reasons behind protein design) of a protein is not defined simply as the combination of independent contributions of the constituent amino acids, but it also depends on the environment (Mildvan, 2004; Benner and Sismour, 2005). An alternative approach, extensively used nowadays, is the *rational* production of standard sets of (poly)peptide components, designed and fully characterised (by the relationship between sequence-structure-interactions and function), which can be used as interchangeable building blocks (Jones, 2009;

Fletcher *et al.*, 2012), to solve different problems or test concepts. Recent studies reveal successfully designed proteins, having attributed emerging functions like modulating the immune system (Pan and Kortemme, 2021), inhibiting viral infections (Chevalier, D. A. Silva, *et al.*, 2017), acting as signalling systems (Feng *et al.*, 2015), protein logic gates (Chen *et al.*, 2020), and even self-assembling biomaterials (Gonen *et al.*, 2015), serving as prototypes for artificial tissues.

2.1.2. Membrane proteins- essential and challenging

The main challenge they impose is their topology is a highly complex lipid environment, which is indispensable for the accurate study of MP dynamics, folding and function activity. Traditional methods for structural determination (X-ray crystallography, NMR, cryo-EM) provide different sets of information, that do not always cover all scientific questions about the relationship between the sequence-structure-function. Currently, more structural methods are proposed for the determination of structure, association and potential understanding of folding, such as atomic force microscopy (AFM) (Edwards, LeBlanc and Perkins, 2021), cryo-electron tomography (cryo-ET) (Eisenstein, 2023), however, they also possess their own limitations. Surface limitation, potential sample damage, limited chemical specificity (difficult to identify specific proteins) and slow scanning speed are associated with the disadvantages of the AFM (Pleshakova *et al.*, 2018), whereas cryo-ET is a low-throughput method, it also has the problem of protein identification as AFM, completed by the limitations related to technical specifications (tilt, sample thickness) (Hylton and Swulius, 2021). Thus, cross-interdisciplinary methods, including the expanding computational techniques, aim to decipher the dynamic lipid-embedded folds.

As known from such methods, most MP structures are alpha-helical bundles. The structural maintenance and recognition depend on non-covalent interactions, hydrogen bonds, and van der Waals forces, which are in close correspondence with the dielectric features of the membrane, and are distributed in such a manner that they satisfy the stability of the structure in both the hydrophobic core formed by the lipid acyl chains, but also in the head group region. Multiple studies in the past have attempted determination of the structural 'hot spots' that drive the folding and stability through site-directed mutagenesis; however, most led to inconclusive results.

Understanding the driving forces is now one of the priorities for protein engineers, who are relying on the balanced application of bioinformatics, molecular simulations, and experimental validation, spotlighting sequence-structure determinants.

2.1.3. *De novo* protein design

A systematic assessment of the contribution of individual amino acids for stabilising a certain fold or achieving a certain function, in simplified, basic systems, is one of the direct approaches for understanding the relationship between sequence-structure-function. One of the greatest challenges faced in protein redesign, by substitutions in order to address this scientific question, is represented by the complexity and diversity of the proteins, which could be overcome, if simplified, minimal proteins would be designed and the composing amino acids to be assigned to its contribution within the fold or function given. Thus, *de novo* design is believed to impose more borders and links between the known and unknown; it provides the possibility of exploring new sequences that could stabilise existing or new folds (Woolfson *et al.*, 2015).

To rationally design building blocks, synthetic biologists developed a set of designing rules which derived from parallel studies performed on soluble proteins and membrane proteins.

First, *de novo* design, also known as design ''from scratch,'' allows the full sampling of protein sequence and structure to see whether artificial constructs could replace the natural proteins or could have other useful artificial functions (Curnow, 2019). Membrane proteins were less well-explored, compared to the soluble proteins. This stands behind the numerous challenges associated with MP: low expression levels, misfolding, and non-specific aggregation, which can frustrate the efforts of any biophysical analysis on MP. *De novo* design aims to overcome these challenges by developing proteins based on a low sequence complexity, detecting motifs involved in transmembrane helix packing (Walters and DeGrado, 2006). These motifs are involved in different types of interactions (helix-helix, lipid-helix), which would give further insights into the handling of MP, making more accessible the roles of each constituent.

2.1.3.1. Rational parametrised design- Coiled-coils parametrisation

Initially proposed and discovered by Crick and Pauling in 1953 in keratin, coiled coils represent the foundation stone for *de novo* design, representing a reliable model and bridge between individual helices and complex native proteins (DeGrado and Korendovych, 2020). Conventionally, a classical coiled coil is a repeat of a heptad labelled "*abcdefg*". It has a left-handed packing orientation, and from a top view, the sidechains of the amino acids in positions a and d are facing the interior of the bundle. They are prevalently hydrophobic, whereas the sidechains of e and g are facing the inter-subunit interface, or are involved in distanced inter-helical interactions and are generally more polar (Fraser, MacRae and Miller, 1965). The

signature heptad is a repeat of hydrophilic (h) and polar (p) residues is hpphppp (Woolfson *et al.*, 2015).

Crick postulated that coiled-coil structures would be defined by the radius and pitch of a superhelical assembly of two or multiple α-helices, in conjunction with the interface angle as a contribution factor to the overall twist between the constituting helices. He developed then equations to define those parameters, now available to generate structural scaffolds (Crick, 1953; Dawson, Rhys and Woolfson, 2019).

It was later illustrated that, the propensity of different amino acids (hydrophobic, polar, or charged) could also determine the topology and stability of the coiled-coil (Harbury *et al.*, 1993). They showed that variations in the component amino acids of the GCN4 leucine zipper can determine different oligomeric states, considering that as a wild-type, GCN4 leucine zipper is known to be a parallel homodimer, they could achieve with the mutations parallel dimeric, trimeric and tetrameric variants. Furthermore, the impact of rational, parametrised design was represented in the work of Thomas as colleagues, based on the design rules derived from the GCN4 leucine zipper mutations, and which delivered heterodimers with different dissociation constants, which were about to represent later robust design building blocks for supramolecular assembly, however with no assigned function (Thomas *et al.*, 2013).

In the view of designable channels, the rules derived from Thomson and co-workers' studies determine the formation of α-helical barrels. They obtained *de novo* pentamers, hexamers and heptamers, by combining the Crick computational parameters with the previously adapted coiled-coil design principles (Thomson *et al.*, 2014).

Although they could not assign a specific function, this study proves the tremendous impact of rationalised and parametrised design on protein design and the achievement of stable higher-order protein assemblies.

These are only a few examples of the numerous implementations of Crick's equations that are available for generating designable, stable coiled-coil structure templates.

2.1.3.2. Design strategies

The efforts made in the field of protein design have shown that there is not a single approach towards protein design, and the strategies can differ, being mainly split into minimal design and rational design. Shortly, in *a minimal design sequence,* patterns of hydrophobic and polar residues are used to define a target structure. On the other hand, the *rational design* identifies and uses specific sequence

motifs from naturally occurring proteins, which are known to stabilise certain structures. The minimal and rational design strategies use stable scaffolds, which are applied modifications for functionalisation (Dawson, Rhys and Woolfson, 2019).

Another design strategy is based on computational techniques, which will be further described below in this section. Most computational techniques use databases of sequence fragments to construct target structures or target functions (Huang, Boyken and Baker, 2016). Each of the strategies has its value, but if the rational design were combined with parametrised computational methods, it would form an even more powerful *de novo* design approach.

2.1.3.3. From soluble to membrane protein design

Part of the advances in protein design represents the method which was applied by a series of scientists, namely the redesign of soluble proteins or template proteins and repurposing them for compatibility with the membrane environment (Simms and Booth, 2013). The coiled-coil motif represented the starting point for the design route of membrane-compatible proteins. For example, the MS1 and the subsequent variants were designed based on adapted coiled-coils design principles derived from the GCN4 variants. The MS1 (membrane-soluble) peptide model was used to evaluate the defining forces that stabilise the interhelical interactions of transmembrane helices (Zhang *et al.*, 2009). It showed that the stability of the membrane-soluble (MS1) peptides was dependent on the presence of the small side-chain amino acids at the "a" position in the heptad, which was opposite to the known rule for water-soluble structures, these favouring large hydrophobic side chains at this specific position. Furthermore, a special interest was manifested in the variants having this position with glycine, which have shown the greatest stability among other small amino acids (Gly> Val> Ala> Ile) and preference for the formation of antiparallel dimers, as analysed through analytical ultracentrifugation. The MS1-Val and MS1-Ala, having Val, respectively Ala in the "a" position of the heptad, have shown a preference towards the formation of parallel dimers.

Thus, not only was it shown that structural motifs derived from soluble proteins could be used for membrane protein design, but it also provided valuable information for the design of the sequence for achieving stability and specific orientation of the helices.

Zhang and co-workers have shown with their designed pore-forming protein switches (Y. Zhang *et al.*, 2015) that, not only the redesign of membrane-soluble structural scaffolds can lead to stable membrane-compatible designs, but also the design of soluble scaffolds prone to structural changes due to

pH modifications. In essence, the design followed the coiled coils heptad-based design, for which at the "c" position in the heptad, they have introduced "trigger" residues such as Asp, Glu and His, which were susceptible to changes in the protonation state and hydrophobicity, depending on the imposed pH. The rest of the positions (a, b, d, e, g) comprised small hydrophobic residues (Ala, Leu, Gly, Ser), and Lys and Gln were placed at the f position (solvent exposed in both environments). This design strategy led to structures which were able to associate in a stable coiled-coil bundle at neutral pH (pH=7.4) and to switch to a membrane-inserted channel state at low pH (pH=5.5).

The design of a transmembrane Zn^{2+} transporter reported by Joh and colleagues (Joh *et al.*, 2014) was also based on the Crick parametrised structure models derived from soluble proteins.

2.1.4. *De novo* design of four-helix bundles

The design of four-helix bundles was targeted in early studies, mainly using the minimal and rational design. The successful achievement of the design of four-helix bundles was based on the use (or design) of amphipathic helices with a predicted high tendency of self-association. The relevance of this type of structure can be explained through the importance of ion channels and heme-binding catalytic proteins. The minimalistic approach proved to be efficient for the design four-helix transmembrane bundles such as a transmembrane dimer of dimers Zn^{2+} transporter (Joh *et al.*, 2014; Joh *et al.*, 2017), a homotetrameric non-biological cofactor binding transmembrane protein PRIME (PoRphyrins In MEmbrane) (Korendovych *et al.*, 2010) and a heme-binding antiparallel tetrahelical protein based on a leucine minimal sequence, REAMP (Recombinantly Expressed Artificial Membrane Protein) (Lalaurie *et al.*, 2018).

These are some examples of the designs made possible through the minimalistic approach; however, many other designs have high-resolution structures, thus emphasising the need to consider the steric arrangement of the residues that determine the helix packing.

2.1.5. Principles for membrane protein design

The transmembrane alpha-helical proteins represent around a quarter of the total proteins across multiple genomes. Due to the fact that they rely on a different, highly heterogeneous environment represented by the lipid bilayer, the structural amino acids arrangement also differs compared to water-soluble helical proteins. The residues that project towards the core of the protein must compete with the lipid acyl chains, and the lipid-facing residues adapt to the differences in hydrophobicity depending on the

membrane depth (Joh *et al.*, 2017). According to the two-stage folding paradigm enunciated by Engelman and co-workers (Engelman, Steitz and Goldman, 1986), there should be first considered achieving the features for appropriate insertion of the helices in the membrane and then the association.

2.1.5.1. Hydrophobic depth

A successful *de novo* design of membrane proteins should not only consider the type of amino acids that will form the sequence, but also the characteristics of the membrane, namely the depth of the membrane. It is commonly known that the average end-to-end size of an amino acid is approximately 3.6 Å (or contour length was determined to be 4.0 ± 0.2 Å) (Ainavarapu *et al.*, 2007). The average size of the hydrophobic core of a lipid bilayer is considered to vary between 30-40 Å (Andersen and Koeppe, 2007). Thus, the minimum length of an amino acid sequence sufficient to span the membrane would be of 21 amino acids.

The protein design pioneers Lear and DeGrado have shown in one of their studies for the design of ion channels based on synthetic amphiphilic peptides the relevance of the length of the sequence (Lear, Wasserman and Degrado, 1988). They used a minimalist design approach, having simple, repetitive sequences which were aimed to contribute to the structural interpretation of the channel properties. They had a comparative study between a 21-residue peptide $(LSSLLSL)_3$, which was able to span the membrane and form stable and ion permeable ion channels selective for protons, compared to a shorter (14-residue) peptide version, which was proven not to be able to trespass the lipid bilayer, hence not able to form the desired structural architecture ion an ion channel.

2.1.5.2. Topology rule

It was also determined that the helix should not only be amphiphilic to provide the desired transmembrane interactions with both regions (polar and hydrophobic), but to have a certain degree of hydrophobicity in order to allow the partitioning in the lipid bilayer in almost a vertical orientation (Lear, Wasserman and Degrado, 1988)

Regarding the rationale design and selecting specific amino acids for their characteristics, remarkable and highly encountered in *de novo* sequences are also Trp and Tyr residues. Trp is usually introduced for detection reasons by increasing the absorption signal and Tyr in the proximity of lipid head groups to control the topology, functioning like ‘anchors’ (Sparks *et al.*, 2014; Curnow *et al.*, 2020), and

determining the helical position and dynamics. Lysine also plays an important role in the control of the topology, frequently found close to the polar/apolar region in the membrane. The special feature of lysine is its ability to manifest bond elasticity across the hydrocarbon chain and respect a specific topology, keeping the charged amino group in the polar region of the membrane and the hydrocarbon scaffold in the hydrophobic part of the bilayer, a behaviour known as 'snorkelling'. This is important for the control of precise location in the membrane and loss in the entropy of the system (Strandberg and Killian, 2003).

The importance of charged amino acids such as lysine or arginine residues at the cytoplasmic edge of the transmembrane helices in the orientation (topology) of the helix was enunciated in von Heijne's studies (Von Heijne, 2006), known as the 'positive-inside' rule, and further completed by Baker and colleagues work (Baker *et al.*, 2017), showing the "negative-out" amino acids topological preference, in the context of a biological-relevant designed protein.

It is also true that not only the amino acid composition could determine the topology of a membrane protein but also the lipid environment, as it was reported by Bogdanov *et al.* in the studies on the lactose transporter LacY, elucidating the importance of the phosphatidylethanolamine (PE) in the correct topology in the membrane and function. This finding unlocks further design pathways needed to be tackled for the understanding of the relationship of sequence-structure-function-environment (Dowhan and Bogdanov, 2009).

Due to the incomplete understanding of the biophysical characteristics, statistical approaches have been involved in order to position proteins in the membrane (Schramm *et al.*, 2012). Studies focused on the knowledge-based potential Ez, which derived from the correct determinations of topology based on sequence information of over 80% of the transmembrane proteins of known structures (Ulmschneider and Sansom, 2001; Ulmschneider, Sansom and Di Nola, 2005; Joh *et al.*, 2017). The early studies concluded that if the hydrophobic matching is pivotal, the orientation of the helices in the membrane is determined by the polar, charged and aromatic residues. Using this information, Lomize *et al*. formulated the OPM (oriented membrane proteins) database (Lomize *et al.*, 2012).

Lately, with the aid of coarse-grained molecular dynamics, Sansom and Thomas released the MemProtMD database (Newport, Sansom and Stansfeld, 2019), available online (http://memprotmd.bioch.ox.ac.uk) in which there are represented the amino acids propensities in different regions of the membrane, based on the coarse-grained simulations of membrane proteins of known structure.

2.1.5.3. Helix packing

An additional and substantial contribution to the understanding of helix packing is attributed to Honig and colleagues, who used a structure alignment algorithm to sample similar pairs of interacting helices in soluble, respectively, membrane proteins in order to find a helical packing pattern and other relevant information behind this process (Yang and Honig, 1999). The glycine motif, known as GxxxG, which was first found in glycophorin A, has been extensively studied due to the fact that it was shown to promote the close association between the helices, forming a closely packed parallel dimer in the regions where the motif was present (Mackenzie, Prestegard and Engelman, 1997).

Their work identified specific motifs generally referred to as SmXXXSm, where Sm could be glycine, alanine, and serine. These motifs could be found in the structure of transmembrane helices forming dimers in some members of the receptor tyrosine kinase protein family (Bocharov *et al.*, 2008). By comparing the propensity of these motifs in soluble and membrane protein helices, they have concluded that glycine-containing motifs are the most frequent in TM helix structures, followed by alanine motifs. In contrast, serine is mostly contained in soluble proteins. They also concluded on the residue composition of TMs: Gly, Leu, Phe, Tyr, and Val are the most frequent, whereas the polar and charged amino acids are present in soluble proteins.

The distinctive feature between soluble and membrane helices was noticed to be the higher stability of membrane proteins compared to soluble ones; this also represents another key factor to consider when designing a protein from scratch. The small amino acids motif was proven to be involved in intramolecular folding and overall stability of the protein based on their intrinsic characteristics (Senes, Gerstein and Engelman, 2000). A first explanation for the contribution to the folding and stability is the reduced dimension of the amino acids, thus mediating the close approach and packing of the helices. Moreover, when the Smx3Sm motif is subject to coiled-coil orientation in the heptad, it is observed an interdigitalization of the sidechains of the amino acids, hence, contributing to the stability.

2.1.5.4. Stability, association and orientation by design

It was shown in later studies that the glycine-zipper motifs were found necessary, but not always sufficient for the dimerisation (Barth, 2007) of glycophorin A (GpA) (Schneider and Engelman, 2004), major coat protein (MCP) (Melnyk *et al.*, 2004), BNIP3 (Sulistijo and MacKenzie, 2006) and CCK4 (Kobus and Fleming, 2005) to the formation of dimers. It was further discussed the importance of the small amino acids for the stabilisation of antiparallel transmembrane helices. North and colleagues have shown that designed

peptides containing a serine-zipper in the structure were likely to form antiparallel dimers (North *et al.*, 2006). Further approaches for the design of antiparallel heterodimers comprised alternations of small residues (Val, Ala, Gly) in the "a" position of the heptad, in combination with static Leu in the position "d" of the heptad. This study led to the conclusion that the fact that preference for synthetic peptides in forming heterodimers increased from Val to Gly substitution (Zhang *et al.*, 2009), re-emphasising the importance of glycine of helical association and attributing the antiparallel helical orientation preference. The close packing of the helices determines the formation of hydrogen bonds between the alpha carbons of the backbone of one helix and carbonyl oxygen in proximity, as well as of van der Waals interactions, the latter considered to be the main forces which drive the folding process (Senes,Ubarretxena-Belandia and Engelman, 2001) with possibly the contribution of the lateral pressure and solvophobic exclusion exerted by the surrounding lipids (Mravic *et al.*, 2019). Hydrogen bonds can also stabilise the designs; thus, strongly polar amino acids such as Asp, Asn, Glu and Gln, and Thr can contribute to the association of TM helices (Zhou *et al.*, 2001) and also to the stability by decreasing the energy score of the system (DeGrado and Korendovych, 2020), if the localisation in the membrane is favourable to the interhelical hydrogen bonded interactions. There is less exposure to the regions water-accessible (Lear *et al.*, 2003).

An interesting design application of helical packing motifs was represented by the development of the CHAMP (computer helical anti-membrane protein) method, used to design peptides that bind transmembrane helical targets. The method workflow comprises the identification of sequence motifs, depending on the amino acid propensities of the common transmembrane interaction motifs, in order to predict the favoured geometry for interaction with further helices. Once the geometry is identified, the CHAMP method uses algorithms for the side-chain repacking in the identified motif, in order to enhance the interaction with target helices (Joh *et al.*, 2017).

To sustain the importance of polar interactions in transmembrane helix association, it was experimentally shown that polar residues drive the association of poly-leucine transmembrane helices, which are normally known to not self-associate (Zhou *et al.*, 2001). The reasons for the limited self-association of poly-leucine helices are explained by the lack of hydrogen bonding potential and electrostatic interactions (due to the lack of polar residues), which lead to lower stability, also combined with the conformational flexibility, which could hinder the formation of stable higher order structures.

2.1.6. Design of overall structure - Rationale for the choice of specific amino acid sequence

Amino acids were selected according to 3 key criteria: 1) small residues like glycine, alanine and

valine to enable close helical packing; 2) long hydrophobic stretches to match the membrane core; and 3) specifically positioned charged residues for membrane positioning and dissolved salt bridging. For instance, glycine was included for its conformational flexibility despite helix-breaking tendencies known in soluble proteins, due to destabilisation of regular hydrogen bonding. It was used however as part of the amino acid composition of transmembrane domains, where does not always as a helix breaker. The hydrophobic membrane environment limits the sterical freedom of the glycine structure, the absence of water forces the participation in internal hydrogen bonding, hence, the hydrogen bonds within the backbone become stabilised and the helical structure. Moreover, the small size of the amino acids enables the close packing of the helices, favouring the formation of stable structures, thus, reducing considerably the possibility of helical breaking through the flexibility of glycine. Additionally, lysine residues were intentionally positioned to take advantage of snorkelling behaviours. The chosen antiparallel helical orientation within the bundle as a starting topology towards designing the amino acid sequence represented a challenge, especially due to the fact that it was reported in early studies on both membrane proteins (Bowie, 1997) and soluble proteins (Walther, Eisenhaber and Argos, 1996) that almost all tested naturally-occurring proteins presented a preference for the parallel packing of the helices (Gimpelev *et al.*, 2004). Later studies have shown the possibility of the *de novo* design of antiparallel structures, such as the ROCKER reported by Joh and colleagues (Joh *et al.*, 2017). However, this represented a tetramer formed as a dimer of dimers. Structures like antiparallel transmembrane dimers of dimers have also been established to be designed by Zhang and co-workers (Zhang *et al.*, 2018). These studies not only supported the present design strategy but also highlighted the potential that can be harnessed through *de novo* design.

The rational design approach used in the present study can be divided into several parts.

2.1.7. Design template

Due to the fact that after translation and during insertion, the protein inserts one helix at a time, the symmetry in the sequence represented the rationale for the possibility of structure self-assembling and disposition of helices. Computer modelling results of different designs (Sparr *et al.*, 2005) have shown that the antiparallel association is a result of favourable electrostatic interactions between the dipoles of the helix backbone in an antiparallel dimer. *De novo* designed membrane protein tetrameric antiparallel structures are dimers of dimers with variable sequences between the helix pairs. Hence, it represented

an interest in the design based on a single repetitive sequence, in the view of enhancing the opportunities for the self-association of the helices. This represented a valid rationale, which was proven by the obtained values for the computed helix association energy.

The incipient point was the generation of a backbone structure template, which was about to represent the blank 'canvas' for the next design steps. The generated template for the antiparallel tetramer followed the Crick parametrisation for coiled coils with pre-defined inter-helical angles and radii. This type of template generation, based on the Crick parametrisation (Crick, 1953; Grigoryan and Degrado, 2011), also represented an input/starting point for the computational design of other transmembrane pores (Xu *et al.*, 2020).

For the choice of the amino acids and sequence, although the Crick parametrisation for coiled coils was used to generate a structure template with an idealised helical backbone, the sequence did not follow the heptad periodicity of a typical coiled-coil structure. However, for the design, the possible occurring short-range and long-range helical interactions, the latter reported in different studies as being an important factor for protein folding (L. H. Wang *et al.*, 2004), were rather taken into consideration when the positions of the amino acids were chosen. The short-range interactions can be explained by the ones exhibited by the amino acids within a helix. They can determine the secondary structure formation (Sengupta and Kundu, 2012), whereas the long-range refer to any occurring interactions between adjacent helices. Considering the long and short-range interactions and the morphology of the amino acids, the choice of the sequence was also based on their reported propensity within the membrane. The localisation in the membrane was determined in statistical studies based on the crystal structures of the naturally occurring membrane proteins (Newport, Sansom and Stansfeld, 2019).

2.1.8. Amino acid composition

The design was chosen to be comprised of predominantly small-size amino acids (glycine, alanine and valine), which would provide a close intrahelical and interhelical packing. The close packing served as a further constraining factor to the allowed degree of freedom of the overall structure, allowing the study of the influence of the surrounding lipid bilayer rather for the adoption of conformational changes/dynamics than for the stability or the lipid-mediated secondary structure formation (Gimpelev *et al.*, 2004). The small amino acids used are also part of the identified sequence motifs (Liu *et al.*, 2004), such as the glycine-based motif (GxxxG), which was identified in early studies on glycophorin A (MacKenzie, Prestegard and Engelman, 1997) or generalised as SmxxxSm (with Sm defined as small and

represented by either Gly, Ala, or Ser). The use of this type of amino acid mediates a close approach between the backbones, as it was shown experimentally, and also contributes to a right-handed crossing angle of the interacting helices of approximately 40° (MacKenzie, Prestegard and Engelman, 1997). In this study, the design boundaries were tested, thus, the well-characterised motifs were circumvented. Furthermore, glycine has a flexible structure, often considered a helix breaker, introducing kinks in the structure; however, in conjunction with the rest of the amino acids, the Ramachandran plots have shown that the presence of these residues did not disrupt the overall angle-determined (dihedral phi/psi angles) helical features.

The sequence (CC1) also comprised phenylalanine, lysine and tyrosine residues. The choice of these amino acids relied on the experimentally determined distribution of amino acids in the lipid membrane. The results of a thorough statistical analysis of the natural-existing membrane protein structures performed by Ulmschneider and Sansom (Ulmschneider and Sansom, 2001) included the distribution of phenylalanine in the transmembrane region, as well as the lysine and tyrosine residues towards the limit region between the lipids head group and tail.

Although in the computationally-generated structures of CC1, it was shown to have the phenylalanine residues placed on the interface between the protein and the lipid, and it was shown previously (Ulmschneider and Sansom, 2001) that this specific orientation was prevalent in β-barrel protein structures and no preference was observed in helices, all the tested computational secondary structure analysis methods in this study showed the presence of a helical structure. This would further support the idea of the existence of complementary helical stabilising factors within the sequence.

Lysine residues were placed at the polar/ apolar limits of the membrane in the structure. As a charged residue, it was introduced based on the 'positive inside' rule enunciated by von Heijne, as a result of the study of the recognition by the translocation machinery (von Heijne, 1992) and re-emphasised in recent structure statistical studies (Baker *et al.*, 2017). Also, it was included for the control of the orientation in the membrane mediated by the known 'snorkelling' or anchoring behaviour, as it was previously demonstrated in studies on the KALP and WALP peptides (Strandberg and Killian, 2003). From the provision of membrane anchoring, the tilt in the membrane could be controlled alongside the correct membrane distribution of the rest of the amino acids in the sequence in the hydrophobic and polar regions of the membrane. Tyrosine residues were also placed at the interfacial region, being based on previously reported structural data, describing this distribution preference in membrane proteins (Wallin *et al.*, 1997; Nyholm, Özdirekcan and Antoinette Killian, 2007). Furthermore, it was extensively described the 'aromatic belt', from which tyrosine residues take part, as a common structural motif in membrane proteins, placed

at the limit between the polar and hydrophobic regions in the lipid bilayer.

2.1.9. Computational input in protein design

The tendency in *de novo* protein design is to simplify the design, thus leading to the identification of basic requirements for conformations and functions. Diminishing the complexity of the sequence, the influence of a specific amino acid mutation on the entire protein structure and potential function can be clearly understood.

Nonetheless, an accurate design wouldn't be possible without the involvement of exciting innovations in the field of computational predictions (Curnow, 2019).

Until the great expansion of computational methods, the majority of *de novo*-designed membrane proteins tended to follow the structural patterns of naturally occurring proteins, imposing limits on topological innovation. Evidence of this is based on the fact that both engineered and natural protein motifs targeting functionality still rely on pre-existing natural frameworks.

The computational developments underscore the challenges in precisely predicting the dynamic interaction between proteins and the lipid environment and effectively exploring the vast space of protein structures to identify highly stable, low-energy configurations among the thousands of possible structures classified by significant energy barriers.

The improvement of access to advanced sampling methods and the extension of the duration of molecular dynamics simulations are the key factors to express the potential of the computationally-aided protein design. The coarse-grained simulations have proven to enable the probing of slow reconfigurations, which are beyond the capabilities of atomic-level simulations. The advanced methods targeting the free energy represent reliable methods that are able to identify different transition states, intermediate metastable states and various competing structural pitfalls. Moreover, machine-learning-based methods, such as neural network algorithms, represent reliable tools for the prediction of favoured mutations that enable various functionalities. At the same time, the overall structural stability is maintained. In addition, an in-depth biophysical assessment and structural validation would be enabled if environments that mimic the natural membranes were integrated into high-throughput screening platforms.

Despite the tremendous challenges for the design of tailored *de novo* membrane proteins, the potential merge between current models and emerging techniques for the stability engineering would represent a reliable design pathway for the transformation of designed complexes into adaptable

instruments, which would push the boundaries in the fields of design of molecular diagnostics for example.

The pioneering computational studies relied on limited computing power, which resulted in weeks or days-long computational simulations. The continuous improvements in molecular dynamics force fields and the recent advances targeting computational power are making possible the reach of more biophysically-relevant molecular dynamics simulation timescales through the increase of capability of modelling dynamic structural rearrangements with high atomic resolutions. While the crystallographic and spectroscopic experimental studies remain crucial for the validation of the accuracy of computer-aided predictions, the computational studies enable the chance to guide specific optimisations in shorter timescales (optimisation achieved within weeks instead of months). Rationally harnessing such emerging powers authorises the acceleration of membrane protein design beyond the reliance upon the constraints imposed by naturally evolved structural scaffolds alone.

Successful design strategies would comprise the combined use of parametrised, coarse-grained, sequence-based and dynamics-focused computational techniques.

The coarse-grained models are known to provide computationally inexpensive initial approaches by simplifying atomic constituents into simplified interaction centres. The reduction in complexity, thus the reduction of the degrees of freedom, enables the exploration of a restrained frame of protein and lipid reconfigurations, which cannot be described by unrefined all-atom computational simulations.
Rigid, fixed initial structural templates based on basic secondary structure serve as powerful initial framework blueprints for optimisation of different amino acid sequences. Later, flexibility of design, known to be essential for the exhibition of various functions, is incorporated.

The mapping of the free energy (according to the energy folding energy landscape) enables the differentiation between relative local stable states, with high accuracy and reduced sampling burdens. Current advanced computational methods allow manipulations of the topology, thus enabling further estimations on the transition state barriers. These estimations would further impact the analysis of the protein's dynamics, the transition states being critical in the reaction processes (such as enzymes, catalysts) and conformational changes (with effect on the study, for example, of the ligand binding phenomenon). The determination of the transition energy barriers enables the prediction of the spontaneity of this kind of process and target localisation of improvements (for proteins designed for a specific function). For example, the Markov state models, combined with principal component analysis, can be used to create maps of networks between the kinetics and different relative configurations found along folding pathways of proteins.

Additionally the free-energy mapping methods allow for the evaluation of alternative

morphologies, or shapes, by exploration of different structural configurations and determination of their effect on stability and interactions with other molecules. Artificial models created with the aid of convolutional neural network algorithms are able to sample and predict suitable sequences and map them in accordance with sequence-function relationships, which significantly simplifies design optimisation processes compared to purely iterative approaches.

Although these methods prove to significantly ease the high-throughput screening of potential designs, they might not be easily accessible or user-friendly. Instead, other computational methods, such as Rosetta and molecular dynamics simulations (such as Gromacs), combined with the use of parametrised templates, would be more suitable and (computational) inexpensive for achieving a computational-based rational protein design.

2.1.9.1. Crick-parametrisation-based templates

At the time it was originally postulated, the Crick parametrisation was mainly focused on the geometry of the helices and the helical packing and less, or possibly not at all focused on the sequence patterns which would promote helix-helix interactions (Woolfson *et al.*, 2015). The parametric equations rely only on four parameters. However, their high accuracy was reconfirmed in multiple subsequent experimental structure studies of coiled coils. Thus, due to the preciseness, the Crick parameters represented an excellent basis for the computational-aided design of proteins. In the trend of the expansion of development of computational tools, there were also released in the past years, the computational platforms based on Crick parameters for coiled-coils modelling are the CCCP (Grigoryan and Degrado, 2011) and CCBuilder (Wood *et al.*, 2014). The structural templates generated with the aid of these platforms are great starting points for computational protein design. Such a template was used for the design of the Zn-Rocker by Joh and colleagues when they chose to design a homotetrameric transmembrane bundle. They sampled the backbone generated based on the Crick parameters, in order to find a suitable conformation for the Zn^{2+} binding, but also to stabilise the overall geometry (Joh *et al.*, 2014).

Both platforms are user-friendly and online accessible. In this study, we used the structure generator feature of the CCCP platform. The structure generator tool enabled the customisation of the number of the helices (chains) desired for the template, sequence length and orientation of the helices in the bundle (parallel/antiparallel), as well as the type of the backbone (poly-alanine template over poly-

glycine backbone). It also gives the possibility of constraining or varying some of the Crick parameters (rise per residue, superhelical radius, superhelical frequency and pitch angle). However, the pre-set selections were kept as suggested (constrained rise per residue and adjustable superhelical frequency).

There are also other computational tools for the generation of symmetric protein helical bundles based on the Crick parameters, which are not web-accessible interfaces (they are script-based), such as BundleGridSampler, PerturbBundle, or MakeBundle (Leman *et al.*, 2020). In the case of MakeBundle, for example, as described by the developers (http://rosettacommons.org/), the secondary structure generated is disconnected, as is the case for the CCCP platform, and in order to connect the component helices with loops; further computational method needs to be invoked, such as GeneralizedKIC (GenKIC) (Coutsias *et al.*, 2004; Bhardwaj *et al.*, 2016).

2.1.9.2. Rosetta-based protocols

Rosetta was initially developed nearly three decades ago, when it was aimed to contribute to the prediction of the protein's folding from a given sequence. Studies over the time have shown the applicability of the software also for the design of a sequence from a given fold, as well as the capacity of addressing multiple scientific problems, such as protein-protein, protein-small molecule docking, loop modelling. In terms of accuracy, it scored the highest among other Critical Assessment Structure Prediction (CASP) competitions. The general idea of a Rosetta sequence-based protocol is based on the conformational modification of a given biomolecule (protein for example, recognised in the program as a "Pose"), either deterministically or stochastically through amino acids substitutions (known as "Movers"), and the resulting conformation is then evaluated based on the calculated score function, depending on the energetic contributions imposed by the substituting amino acids (Leman *et al.*, 2020).

The energy function of the newly modified structure is considered to be accepted, if the calculated overall energy is lower than the energy of the original structure (lower energy, meaning higher stability). In the opposite case, it is only accepted in regard to certain mathematical probabilities, which take into account the eventual influence of the system's temperature on the newly exhibited energy.

It is considered a score function of the linear combination of physical-based and statistically derived potentials (as seen in the libraries of 3-9 amino acids long fragments used in Monte Carlo sampling). The potentials described van der Waals (vdW) energies, hydrogen bonds, electrostatics, disulphide bonds, residue solvation, backbone torsion angles, sidechain rotamer energies and an average unfolded state reference energy (Leaver-Fay *et al.*, 2013).

There are many recent methods developed in Rosetta, each tackling different scientific problems. Developers have also focused on the increase of the accessibility and usability of Rosetta, and they have developed user-friendly interfaces, such as Foldit Standalone (Kleffner *et al.*, 2017). The interface helps the designer improve the designed sequence by the assessment of the substituted amino acids, and in the case of the new structure's energy is considered higher than the original, even after considering the temperature input probabilities, it visually prompts localised sterical clashes that require change for achieving a higher structural stability.

The reported limitations based on the score function include the incomplete estimation of entropy (it is a fast sampling method at the cost of high accuracy), results-based Monte Carlo sampling of fragments from crystal structures, and the use of the implicit solvation model (it contributes to the method's speed but hinders the effect of the explicit ions, water, and lipid models) (Leman *et al.*, 2020).

2.1.9.3. Considering the membrane: Implicit membrane simulations

Studies have shown that design methods cannot rely only on evaluating the score functions and introducing accuracy through increasing the system's complexity in the study would provide further comprehension of the design process.

In this sense, RosettaMP, which has its roots in the Rosetta Membrane protein structure prediction tool developed by Barth and colleagues, implements core modules for sampling and scoring proteins in an implicit membrane (Alford et al., 2015). Implicit membranes are models that respect the dielectric features and the hydrophobic depth of a natural membrane.

The Hippo software developed by Ulmschneider and co-workers in the early 2000s is also a computational method designed with a specific focus on the membrane proteins, which focuses on the insertion, folding and orientation of protein helices in implicit membrane environments.

Although both RosettaMP and Hippo impose an implicit membrane model for the sequence sampling depending on the insertion and orientation in the membrane, some noteworthy differences need to be mentioned. In terms of ease of use and accessibility, the Hippo software seems to be more user-friendly for specific tasks. Furthermore, Hippo software relies on the crystallographic database in terms of general orientation and insertion (membrane protein focused). In contrast, RosettaMP is still based on sequence fragment libraries, which would make the Hippo approach more applicable for the sampling of purely *de novo*-designed sequences for insertion and orientation.

2.1.9.4. Considering the membrane complexity-Molecular dynamics simulations

Implicit membranes can be considered reliable as intermediary steps for sampling *de novo* sequences, considering the general features of membranes. However, these membrane models fail to reliably model the membrane protein interiors and to intuitively consider the interaction complexity levels imposed by all-atom models.

In response to these limitations, the molecular dynamics simulations allow the direct analysis of protein dynamics and conformational changes by integration of Newton's motion equations described by the MD force fields (Leman *et al.*, 2020). Whilst the implicit membranes can offer approximate barriers of the hydrophobic environment and lack the molecular components, the all-atom explicit membrane models incorporate with accuracy the atom and physical diverse terms characteristic to the phospholipid bilayer, ions and water molecules.

A few examples of molecular dynamics packages widely used in studies for protein dynamics are CHARMM (Brooks *et al.*, 2009), AMBER (J. Wang *et al.*, 2004) and GROMACS (Van Der Spoel *et al.*, 2005). These methods have proven to be useful in model (structural) refinement and determination of complex molecular interactions and functional mechanisms.

The MD simulation packages involve potential energy calculations and usage of force fields to model the physical interactions within a system accurately. Although the force fields are imposed on all the molecules involved in the system (protein, lipids, water, ions), the choice of force fields that are going to be used in the study depends on the desired final answer. For example, if an MD-based study aims to analyse the stability of a certain *de novo*-designed protein, the applied forced fields would be chosen in such a way that they do not interfere with the reference key energetic points that define the stability of the chosen structure in a given environment. A specific example would be the AMBERff98 force field, which was proven not to influence the secondary structure of proteins. GROMACS is known for its fast and efficient algorithms for the calculation of non-bonded interactions, which represent the most complex term throughout the simulations. Furthermore, it involves the particle mesh Ewald (PME) technique for long-range electrostatics. In terms of the computational cost required, it is quite a versatile package, allowing the simulations even on a few CPU cores, but can be transferrable easily on GPU or a combination of both and depending on the chosen distribution, the time required for the completion of the simulation can vary. It is also a versatile method when referring to the inputs, making accessible the modification of the pressure and system-applied pressure modes (isotropic, semi-isotropic, anisotropic) and temperature. Considering that the MD simulations need to be relevant for the eventual subsequent experiments, the

temperature and pressure need to be kept constant. Thus, there are widely used the Parrinello-Rahman and Berendsen thermostats and barostats. The system equilibration steps are essential prior to the final production run, due to the fact that they bring the system to a more realistic, stable state in the given physical conditions. Before the final production run, GROMACS also involves energy minimisation steps, which contribute to the searching for a stable starting structure and the reduction of the potential unrealistic conformations. For this, it is often used the steepest descent method for the relief of the out-of-range steric clashes by iteratively moving the component atoms in the direction of a lower local minimum. Additionally, during the system preparation, periodic boundary conditions can be applied for the study of bulk properties and dynamic behaviours. The PBC can be seen as infinite lattice simulations (or mirror images of the simulation box) used in order to reduce the surface effects of the finite size of the simulation box (the intervention of further pressure-related forces). The resulting calculations are transposed into a so-called trajectory file, which comprises the totality of time-dependent conformational changes.

Overall, GROMACS is considered a reliable computational method, which applies the principles of classical mechanics defined by computational algorithms, and which is able to predict a protein's dynamic behaviour and interactions over a given time range.

Overall, if taken separately, the chances for a single computational method to accurately elucidate scientific questions, such as the governing forces driving protein folding, especially in the case of a *de novo* protein, are low. Thus, it is of interest to elaborate additional computational frameworks that would efficiently combine sampling techniques (fast and less accurate and also slow with increased accuracy) and allow the sampling of sequences that are not library-based.

In this chapter, we proposed an alternative multi-step computational framework for the *de novo* design of membrane protein structures based on a single, repetitive sequence. The framework consisted of the use of fast and slow techniques, balancing the speed and accuracy of the study, as presented in the Figure 2.1.

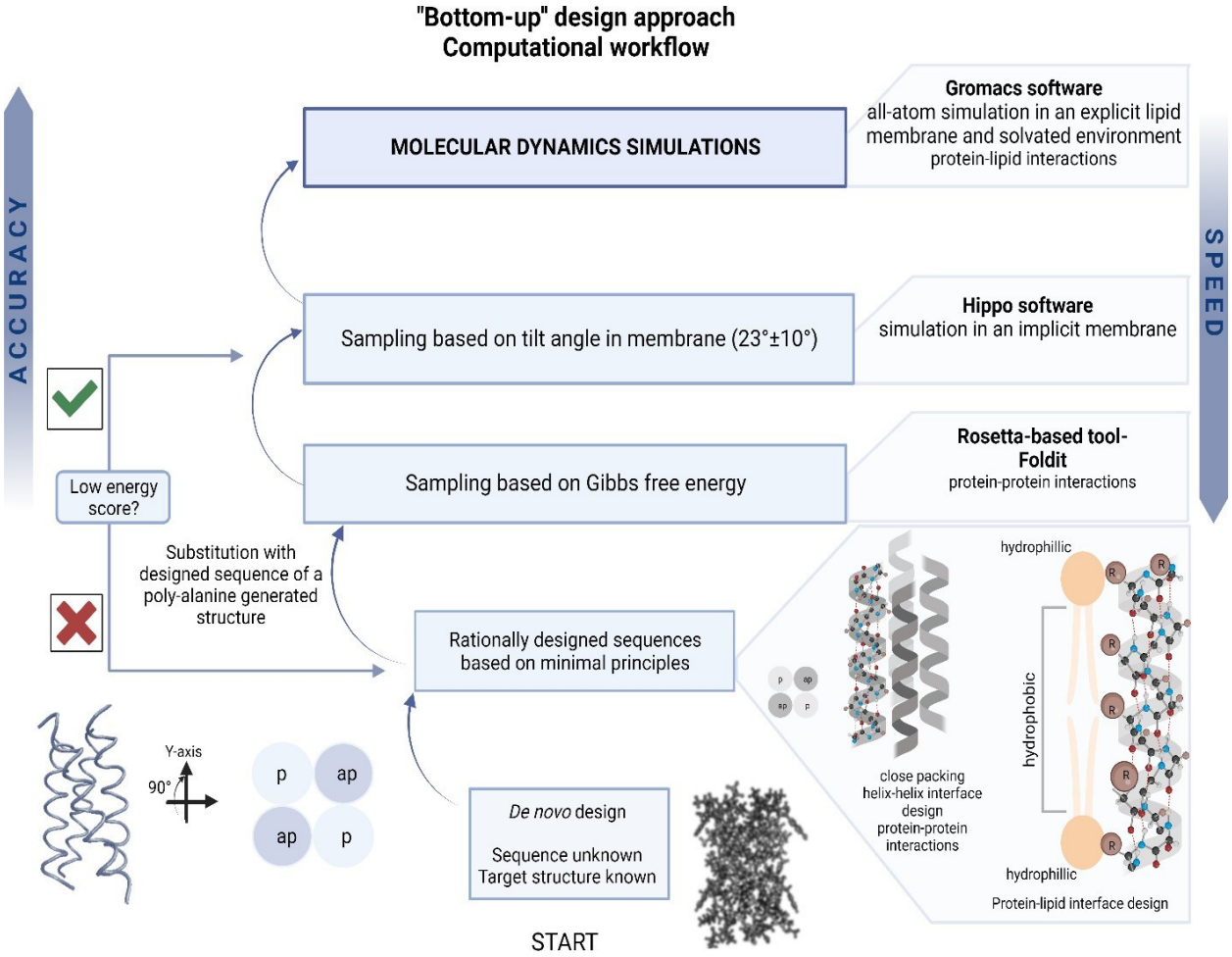


Figure 2.1. Computational sequential framework for sampling the rationally designed sequences. The workflow represents a bottom-up approach. It started with the establishment of an antiparallel tetrameric bundle of transmembrane helices as the target structure. This was followed by the rational design of sequences, taking into consideration the protein-protein and the protein-lipid interfaces. The sequences substituted a poly-alanine software-generated backbone, and the structures based on the designed sequences were sampled based on their free Gibbs energy. The target was the identification of a sequence which forms a stable tetrameric structure (exhibiting a low free Gibbs energy). The sequence was taken further for simulations for the tilt angle in an implicit membrane for detecting the correct positioning of the different types of amino acids within the sequence with respect to the hydrophobicity and depth of the membrane. The sequence (structure), which respected all the imposed criteria, was subjected to molecular dynamics simulations in an explicit environment containing all-atom simulated lipids, water molecules and ions for the detection of the protein-lipid interactions. (Created with BioRender)

Multiple sequences based on a minimal complexity were designed and aimed for the stabilisation of an antiparallel transmembrane four-helix bundle. This architecture was chosen, to facilitate the architecture of a membrane protein channel that needs to be designed, this being the original scope of the project, and a tetramer would be considered as being the first simplistic version of such design,

compared to a dimer, or a trimer. Moreover, a tetramer would have enabled us to design and study paired interhelical, or helical-membrane interface, as future work, and the antiparallel feature would have given an insight of specific amino acids motifs that would stabilise, or destabilise the protein-protein association. This would be a valuable point for future work on the study of orthogonal sequences.The single and repetitive sequence, which proved to stabilise the bundle, was AVGKGGVFVGFGGVGVGFYGV, named CC1. The sequence was sampled for stability implementation of the sequence in a Crick parametrised template, and evaluated using a Rosetta-based score function user-friendly interface, Foldit. Furthermore, the design was assessed for correct orientation in an implicit membrane using Hippo software. The accuracy of the study for the design's dynamic behaviour in explicit environments was made possible with the aid of the MD simulations package GROMACS. The design was simulated in pure POPC, POPE and POPG bilayers for 100 ns. Essential protein- lipid interactions were observed, in addition to the predicted protein-protein interactions during the design stage.

2.2. Materials

The sequences were sampled using Hippo (https://www.biowerkzeug.com/) and Foldit Standalone (https://els2.comotion.uw.edu/product/foldit-standalone). Templates of antiparallel structures were generated using the online source CCCP (Coiled-coil Crick Parametrization): (https://grigoryanlab.org/cccp). The molecular dynamics simulations were performed using Gromacs 4.5.6 (https://www.gromacs.org/), and the explicit membranes were obtained using CHARMM-GUI. The structures and simulations were visualised using the Swiss PDB Viewer (https://spdbv.unil.ch/), USCF Chimera tool (https://www.cgl.ucsf.edu/chimera/) and VMD (https://www.ks.uiuc.edu/Research/vmd/). The MemProtMD database for the amino acids distribution in different regions of the membrane lipid bilayer is available at the following source: http://memprotmd.bioch.ox.ac.uk/stats/.

2.3. Methods

The design framework of the *de novo* transmembrane bundle of helices, namely CC1, was based on a sequential multi-step approach, combining different types of software for the iterative sampling of the rationally designed sequences. Initially, there were designed sequences to provide stable structures, as reflected by the free Gibbs energy, and to respect the amino acids distribution across the membrane. This step provided insight into the understanding of design rules. Using this knowledge, there were designed sequences which imposed symmetry, the final goal was to design a repetitive sequence, able to

stabilise and contribute to the self-association of an antiparallel tetrameric bundle. The rational design of sequences was followed by computational sampling and molecular dynamics simulations.

2.3.1. Rational design of sequences

Concisely, several sequences were rationally designed for the target design of an antiparallel tetrameric bundle based on minimal sequence complexity, designing the inter-helical and helix-membrane interfaces, following seven design rules:

a) Design of a sequence consisting of 21 amino acids; the length of the sequence was chosen to be sufficient to span the width of a membrane lipid bilayer;

b) Choice of small amino acids for close packing of the design;

c) Choice of hydrophobic amino acids for the transmembrane region of the sequence to match the hydrophobic feature of the transmembrane region of the lipid bilayer;

d) Choice of polar amino acids to match the hydrophilic region of the lipid bilayer;

e) Choice of the amino acids based on their propensity in different membrane regions, as reported in the database for the membrane proteins embedded in lipid bilayers (MemProtMD);

f) Choice and positioning of 'anchoring' amino acids;

g) Reduction of the sequence complexity and re-positioning of amino acids in the sequences, if necessary, to achieve a minimal complexity;

2.3.2. Simulations of the protein-protein interactions of the tetrameric antiparallel structures using the Foldit Standalone software

A poly-alanine structure template was generated using the Coiled-coil Crick Parametrization online tool available from the Grigoryan lab (https://grigoryanlab.org/cccp). The template was chosen to be a bundle of four-21 amino acids-long antiparallel helical chains. The structure was generated according to the following parameter values, as they are calculated at the generation step in CCCP: superhelical radius R_0= 7.360; helical radius R_1=2.260; superhelical frequency ω_0=-2.447; helical frequency ω_1= 102.857; α=-12.010; starting helical phase Φ_1=[-9, -9, -9, -9]; chain superhelical offset $\Delta\Phi_0$=[90; 180; 270], chain axial offset $Z_{aa'}$=[0, 0, 0]. The bundle consisted only of the helices and was not connected by any loops. The structure was generated from the online tool as a file in the PDB format. The PDB file was used as input in the Foldit Standalone software. In this program, each of the poly-alanine backbone sequence

amino acids was mutated to the desired rationally designed sequence and energy scores pre- and post energy minimisation were recorded.

As positive control, the reported REAMP sequence was chosen as a tetrameric bundle of helices. As a negative control, a poly-leucine tetrameric bundle was chosen.

2.3.3. Simulations for the evaluation of tilt angles in an implicit membrane using the Hippo software

. The sequences were sampled for tilt using different implicit membrane widths in the range of 24-36 Å by inserting the sequence of a single helix of the input text file of the software and running the simulation.

2.3.4. Molecular dynamics (MD) simulations for stability in explicit environments using Gromacs software

All-atom lipid membranes were generated as PDB files. The composition of the lipid membranes was symmetrical; both lipid leaflets of each membrane were composed of the same type of phospholipid. The lipid membranes were composed of either of the following phospholipids: POPC (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine,) POPE (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine), and POPG (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol). The Gromacs lipid molecules were translated to the specific AMBER98 force field and used to initially simulate the area per lipid and lipid ordering parameters (John Simms, personal communication). The membrane PDB file and the design (CC1, REAMP, poly-leucine) PDB file obtained from Foldit were loaded as individual layers in the visualisation software of structures- Swiss PDB Viewer. The designed protein was oriented and centred manually in the lipid bilayer, using the Swiss PDB Viewer. The membrane-centred design model represented the input PDB file for the MD simulations. As a first step for the MD, the coordinates and topology were defined. The PDB file containing the protein centred in the lipid bilayer was converted into a Gromacs-compatible file GRO, and the corresponding force field was chosen (gromos52a6). The dimension of the simulation box was defined as 64x64x64 Å on the x, y and z-axes. The pressure of the system was chosen as constant at 1 atmosphere, using the Parrinello-Rahman barostat (1 atm) and the system's compressibility was set as semi-isotropic. The temperature for the simulation was maintained at 300 degrees on the Kelvin (K) scale, using the Berendsen temperature coupling system. The system was solvated with water molecules (spc216 MD model for water molecules; 50:1 water: lipid number of molecules), and sodium (Na^+) and chloride (Cl^-) ions were introduced through a solution of 0.15 M sodium chloride (NaCl). Once the system

was solvated and ions were added, there were performed a set of equilibration steps before the final production run. The equilibration steps included equilibration of pressure (NPT-particles number, pressure and temperature constant) and volume (NVT-particle number, volume and temperature constant). Positional restraints of 1000kJ/mol were applied to different parts of the system as part of the equilibration steps. The tetramer was treated as a continuous chain, all domains were simulated at once and used as the input. The positional restraints were applied using different atoms and released in stages. Initially, they were used to restrain all heavy atoms within the tetramer for 500 ps, followed by restraints relaxation. The following restraining step was applied to the main chain atoms for 500 ps, and the cycle was repeated for the backbone atoms. The final positional restraints were applied for the alpha carbon (C_α) atoms before the final production run. The final MD simulation was performed for 100 ns in 50000 steps with a steep descent integrator.

2.3.5. Visualisation of the resulting trajectories of the MD simulations in VMD (Visual Molecular Dynamics) software

The corresponding GRO and XTC (compressed trajectory) files of each Gromacs simulation were loaded into the software. In most of the cases for the lipid, water and ions molecules it was chosen the "VDW" graphical representation (appeared as spheres). The visual clarity was improved, for the same molecules, by choosing the "CPK" representation (spheres of smaller size compared to VDW connected with lines), or the "Lines" representation. To highlight the protein in the simulation box, in most cases, it was chosen the graphical representation "NewCartoon" (which appeared as ribbons). To improve the visual clarity in some of the cases it was chosen the "Cartoon" representation (appeared as tubes). At different points, specific amino acids were selected, and the structures were illustrated using the "Bonds" graphical representation. Differences between the trajectory frames were visually inspected, and the out-of-ordinary events were captured in snapshots.

2.3.6. Visualisation of Ramachandran plots in VMD

The corresponding structure file (PDB file) represented the input file for the calculations of the dihedral angles and the GRO and XTC files for structure comparison. The Ramachandran plots were generated from the analysis tool.

2.4. Results

2.4.1. Rational design- the choice of the amino acids

Based on the amino acids propensities in different regions of the membrane, reported by the statistics available on MemProtMD (Figure 2.2), there were chosen sets of amino acids for the rational design of the *de novo* sequence. In the limit between the hydrophilic and hydrophobic regions, there are found aromatic amino acids tryptophan and tyrosine, positively charged amino acids arginine and lysine and negatively charged aspartate or glutamate. The hydrophobic amino acids valine, alanine, methionine, cysteine, leucine, isoleucine, glycine and phenylalanine are found in the transmembrane region. More specifically, the distribution of alanine, valine, glycine and phenylalanine is more extended in the membrane, reaching the hydrophilic-hydrophobic limit, an aspect which was also considered in the design process. This wide distribution is not considered for leucine and isoleucine, the propensity of those being towards the hydrophobic core of the bilayer, as expressed by the database, limiting the use of these amino acids for the design of other regions of the helix.

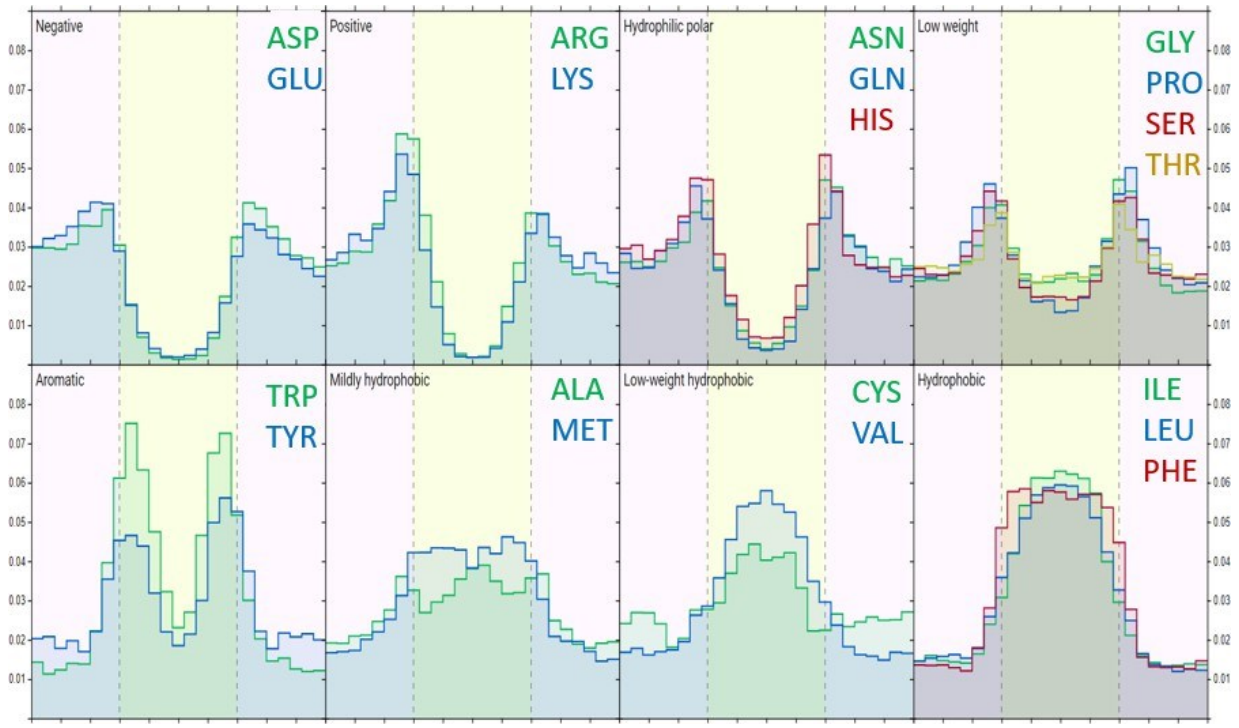


Figure 2.2. The distribution and abundance of amino acids in different regions of the membrane. The regions coloured in light pink represent the hydrophilic regions, the dotted lines mark the placement of the phosphate groups of the annular lipids, and the light yellow fragment represents the transmembrane region. (Adapted from http://memprotmd.bioch.ox.ac.uk/)

In Figure 2.3, presented below there are listed the types of amino acids chosen in the design process within this project. In panel A, there are highlighted the regions in the helix that interact with the limit between the hydrophilic lipid head groups and hydrophobic parts of the lipid bilayer (represented as region 1) and the transmembrane region comprising the hydrophobic lipid tails. The designed set of sequences was based on variable combinations between the following amino acids, as presented in panel B: for the hydrophilic-hydrophobic limit, there were tested: the positively charged lysine and arginine, the aromatic tyrosine (polar) and tryptophan (non-polar), the negatively charged aspartic acid, and the polar histidine. Alanine, valine, glycine, leucine and phenylalanine were chosen for the design of the transmembrane section of the helix.

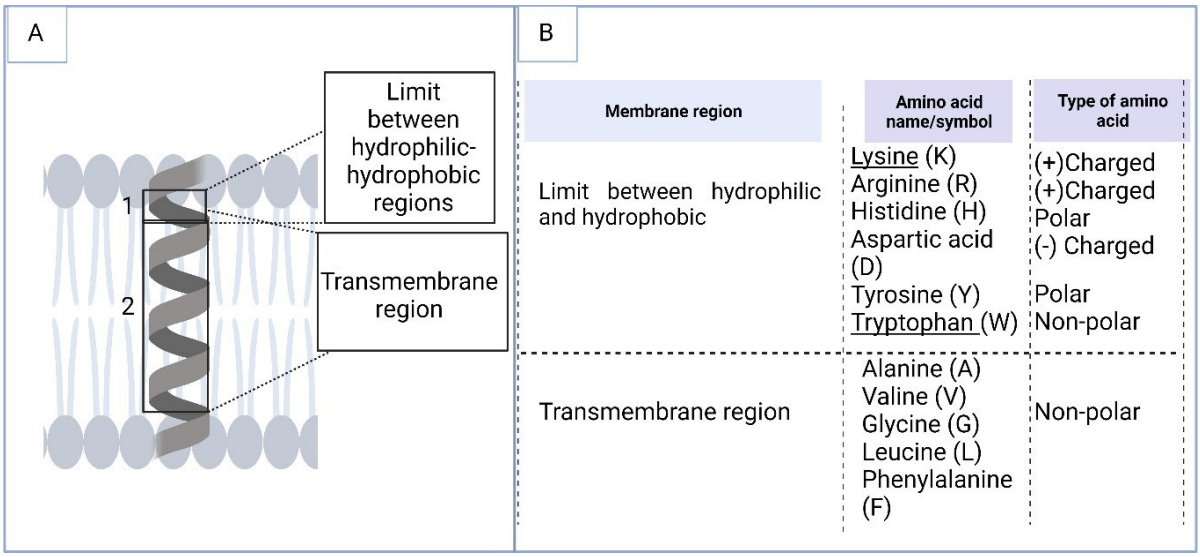


Figure 2.3. The amino acids were chosen for the design of sequences. A. Representation of a protein helix embedded in a lipid bilayer. (1) The region in the helix which interacts with the limit between the hydrophilic and the hydrophobic regions of the lipids bilayer. (2) The transmembrane region of the helix, which interacts with the lipid tails. B. The amino acids chosen for the design of different helix regions are charged and aromatic (polar and non-polar) amino acids for the region close to the head groups and non-polar, hydrophobic amino acids for the transmembrane region. (Created with BioRender)

Initially, there have been designed sequences which imposed either homo-dimeric symmetry between the helices or homo-tetrameric, the latter being the end goal. When the mutations were completed, it was noted the computed energy score, which reflected the prediction of the structural stability. The obtained structure was subjected to free energy minimisation. At this step, there were indicated by the program, any local mismatched interactions that would have needed improvement. In the case of fixing the prompted suggestions in the structure, the displayed energy score decreased accordingly, depending on the energy of each modified bond or sterical impediment. From this analysis,

it was selected the structure, based on the repetitive sequence, for which it was detected a small difference between the energy scores before and after the minimisation steps. The selected sequence was considered to stabilise the helical bundle. The final structure was exported as a PDB file. The homo-dimeric sequences were designed to accommodate the rationale of amino acid choice depending on the propensity and abundance in the membrane. There were tested several sequences, as shown in Tables 2.1 and 2.2 and there have been detected favourable changes in the energy score results when certain amino acids were introduced in the sequence (e.g. alanine, as can be seen in Table 2.1). The homo-dimeric structures were tested as a proof-of-concept for the selection of different amino acids for the desired design. In these structures, it could also be observed that the presence of phenylalanine in the sequence contributed to further stability, as shown by the difference between the energy scores before and after the energy minimisation.

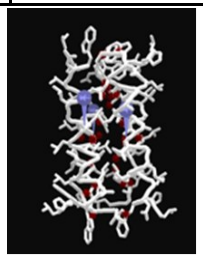
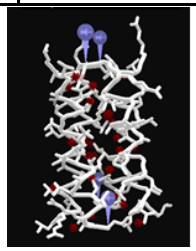
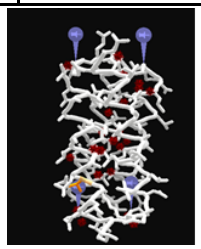
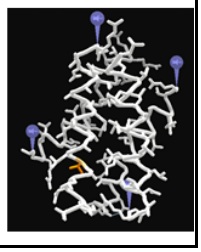
Table 2.1. Sequences of homo-dimeric structures (test) and corresponding Rosetta energy scores, as determined in the Foldit program. The first two columns correspond to the sequences of the symmetrical transmembrane helices (TM) 1/3, respectively, in column two, the symmetrical helices 2/4. The underlined sequence fragments in the first column indicate the sequence fragment that is distinct between the two pairs of symmetrical helices. The following columns show the energy score values of the structures before the energy minimisation ($E_{initial}$) and after the energy minimisation (E_{final}), as displayed in the Foldit program.

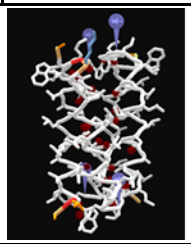
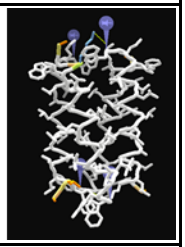
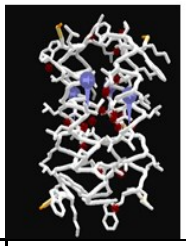
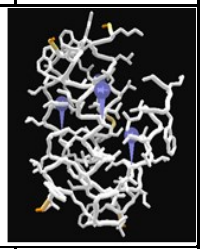
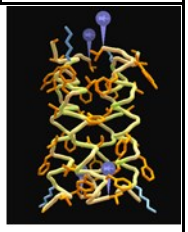
Sequences		**Energy score**	
TM 1/3	**TM 2/4**	**$E_{initial}$**	**E_{final}**
AAAKAAAFAAFAAAAAAFYAA	**AAAKFAAFAAAAAAFAAAYAA**	**-175.61**	**-175.67**
AVGKGAVFVGFGAVAVGFYGV	**AVGKFGVFVAGAGVFVAVYGV**	**-72.092**	**-72.230**
AAAKAAAFAAFAAAGAAFYAA	**AAAKFAAFAAAGAAFAAAYAA**	**-169.181**	**-169.275**
AAAKAAAFAAFAAAGAAFWAA	**AAAKFAAFAAAGAAFAAAWAA**	**-167.76**	**-167.87**
AVGKGAVFVGFGGVGVGFWGV	**AVGKGGVFVGFGGVGVGFWGV**	**-55.776**	**-81.27**
RVGWGAVFVGFGGVGVGFWGV	**RVGWGGVFVGFGGVGVGFWGV**	**-42.961**	**-79.89**

The design approach in this study was based on inverse folding. The target amino acid sequence was desired to stabilise the chosen theoretical structure and minimise variability. In this sense, amino acid

sequences were designed and tested for the stabilisation of homo-tetrameric structures, as shown by the calculated energy scores. It was observed that, as expected, introducing small amino acids (such as glycine) in the structure promoted the close-packing and stabilisation of the tetramer.

Table 2.2. Sequences of homo-tetrameric structures and corresponding Rosetta energy scores, as determined in the Foldit program. The first column shows the sequence of the four antiparallel transmembrane helices (TM 1-4). The second column shows the structures and the energy score values of the structures before the energy minimisation step ($E_{initial}$) and after the energy minimisation (E_{final}).

Sequences	**Energy score**	
TM 1-4	**Einitial**	**Efinal**
LHWDLALLALALLALALDWHL	**1525.96**	**190.76**
		
LRVDLALASLALASLALDVRL	**468.59**	**-152.48**
		
LRVDLALALLALALLALDVRL	**762.89**	**-177.16**
		

AWKLLALALLLLALALILLKW	**1049.16**	**-209.24**
		
AKWLLAALALLLLALALALKA	**821.69**	**-191.59**
		
AVGKGGVFVGFGGVGVGFYGV	**-51.036**	**-77.37**
		

This was a direct approach to the design. The sequences were chosen based on theoretical information and less on structures or designs that had previously been tested or sampled, and rather to test and get an insight into the visual aspect and the impact of the placement of the amino acids in different parts of a helix and interactions that can take occur. The sequence space was not thoroughly explored, as is the case when sequences are generated with specific software based on an algorithm and other techniques which can sample thousands of amino acid combinations and could diminish the risks which could be encountered in downstream applications.

2.4.1.1. Evaluation of Foldit as a computational tool for protein design

Once the amino acid composition and desired positions were set onto the generated structure

template, the Foldit method, recognised as a tool for rational design (Pan and Kortemme, 2021), aided the screening of the potential sequences for the stabilisation of a tetrameric bundle of antiparallel helices, through the calculation of the potential energy and comparison between the initial states and minimised states. Although at this stage, multiple sequences were tested, and although some of them might have shown better (lower) values of the computed potential energy, the visual inspection during the minimisation showed several local interacting residue rearrangements. The CC1 tetramer with the repetitive sequence displayed preferred low energy score differentials in completion to the compatibility of sequence to structure stabilisation. The introduction of the small residues, such as glycine, promoted tighter helical packing in the homotetramer. Furthermore, when the template alanine residues were mutated to phenylalanine residues, there were observed favourable contributions. Free energy minimisation resulted in a reduced Rosetta score (-82, as prompted after the energy minimisation), consistent with increased modelled stability. These results indicated the successful design of intra-helical interfaces. The initial promising results obtained from this method encouraged the proceeding of the study with a single, repetitive CC1 sequence (AVGKGGVFVGFGGVGVGFYGV) in further experiments. The need for further computational sequence sampling is based on the fact that Foldit, as a fast computational method, the calculated energy results for the packing and stability are performed on the system (the protein structure) perceived in a vacuum, without considering external factors (e.g. lateral pressure in the membrane).

Upon identifying a low-energy sequence with Foldit, Hippo was used to evaluate tilting in an implicit membrane due to its specialisation for orienting transmembrane protein structures.

2.4.2. Ramachandran plot of the design

In order to further predict the secondary structure of the designed tetramer, a Ramachandran plot was generated using the Visual Molecular Dynamics (VMD) software. As it can be seen in Figure 2.4 in panel A, there are illustrated the four quadrants of the Ramachandran plot and the secondary structures described by the dihedral angle values ranges of different quadrants. (I.- left-handed α-helix, II.- β-sheet (parallel/antiparallel), III. right-handed α-helix). The blue regions describe the quadrant region with the most suitable values corresponding to the allocated secondary structure. Panel B represents the Ramachandran plot of the CC1 tetramer. The constituent amino acids are represented by yellow dots. For the CC1 design, it can be seen that most of the constituent amino acids are located in the third quadrant (III), predicting a right-handed α-helix secondary structure.

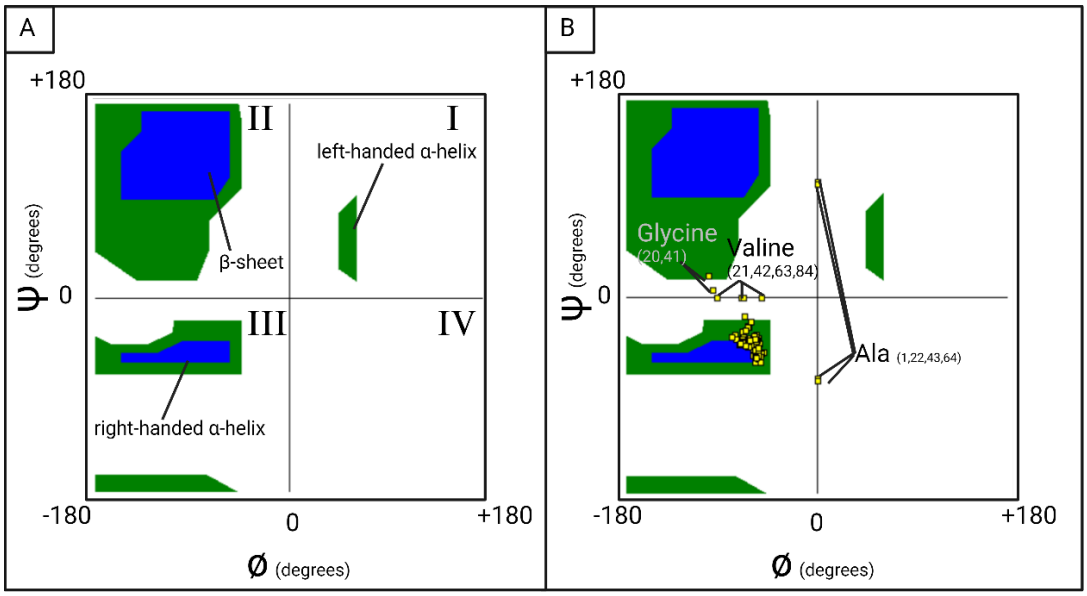


Figure 2.4. Ramachandran plots. A. General Ramachandran plot including the secondary structures found in different quadrants: I.- left-handed α-helix, II.-β-sheet, III- right-handed α-helix. On the X axis, there are found the phi (Φ) dihedral angles with values in the range of -180° and +180°. On the Y axis, there are the psi (Ψ) dihedral angles B. Ramachandran plot obtained for the CC1 tetramer structure not embedded in any lipid bilayer. The amino acids are illustrated as yellow dots. The flanking alanine (1, 22, 43, 64), valine (21, 42, 63, 84) and certain glycine (20, 41) residues had different phi/psi angle values and were indicated by lines. (Created with BioRender)

As highlighted in panel B, few amino acids from the structure were observed that were not found in the third quadrant. On the axis of 0 degrees, the amino acids alanine and valine were found, which delimit the sequence of each helix. Also, it appeared in the Ramachandran plot that two of the glycine residues situated towards the ends of the first and second helix tended to be part of the second quadrant, the dihedral angles describing a secondary structure of β-sheet.

In order to test the protein-membrane interactions, the simulations were performed using two different software: the Hippo software, which gives an insight into the possible orientation of the design in an implicit membrane; however, it uses a simplified model of a membrane, and Gromacs software, which introduces accuracy in the study, being an all-atom simulation, the design is simulated in a natural-like environment, and it contains an explicit membrane with defined lipid molecules, the system is solvated using water molecules and ions, and the temperature and pressure can also be controlled.

2.4.3. Simulations for the evaluation of tilt angles in an implicit membrane using the Hippo software

Hippo software was used to assess the tilt exhibited by an alpha-helix in an implicit membrane. This was a fast method to test the correct positioning of the amino acids in rapport with the hydrophobic

and hydrophilic regions of the membrane. The helix was generated in the same software, using as input the designed sequence selected from the preceding sampling step, comprising the simulations in Foldit. The tilt angle is described by the position of the membrane-embedded helix along the Z axis at the end of the simulation (Z') against the reference position of the Z axis (Z), as represented in Figure 2.5 A. Figure 2.5 B illustrates the tilt measurement in Hippo software. The membrane model is simplified and is not an all-atom model, also called an 'implicit membrane', with the same properties and functions as an all-atom membrane. The implicit membrane is delimited by two plans, and the distance between those plans represents the depth of the membrane, measured in angstroms. It can be modified and sampled in the simulations. The tilt measured in Hippo is influenced by the hydrophobic mismatch between the sequence and the membrane and the depth of the membrane. Considering this, it was assessed the tilt of the selected sequence (CC1) in membranes with depths between 24-36 angstroms (Å). In Figure 2.5, panel C are visual representations from the output PDB files, of the variation of the tilt angle of the helix based on the chosen sequence in implicit membranes of 24 respectively 36 angstroms.

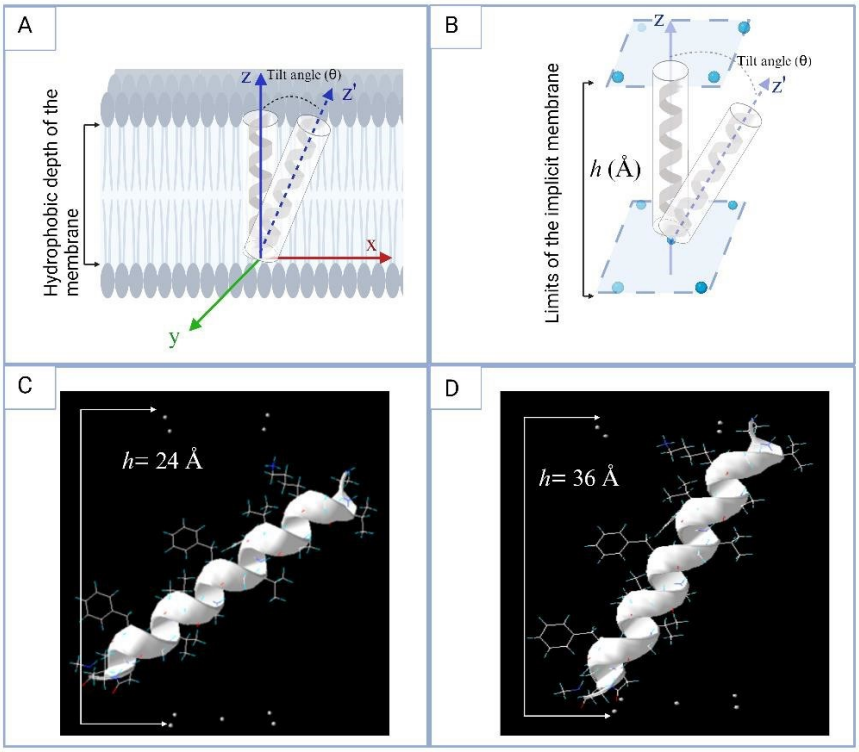


Figure 2.5. Measuring the tilt angle of helices-Simulations in Hippo software. A. Example of a helix embedded in a lipid bilayer. The tilt angle is the angle formed by the helix along the Z axis (blue arrow) from the reference position (Z) until the final position (Z', blue dotted arrow). The hydrophobic depth of a lipid bilayer is described by the transmembrane region or the lipid tails region. B. Example of measuring the tilt angle in a Hippo simplified membrane model (implicit membrane). The tilt of the helix is influenced by the width (h) of the membrane and the hydrophobic mismatch. C-D. Representations of the designed sequence simulated in Hippo in implicit membranes

of width 24 and 36 angstroms. (Created with BioRender)

Table 2.3 comprises the values of the measured tilt angle in implicit membranes of various widths between 24 and 36 angstroms.

Table 2.3. Tilt angle with respect to the implicit membrane width. The first column comprises the values measured in angstroms for each tested implicit membrane. In the second column, there are found the tilt angles of the helix are based on the selected, designed sequence.

Implicit membrane thickness (h, Å)	Tilt angle (°)
24	53.34
26	49.01
28	44.47
30	39.87
32	35.38
34	31.75
36	28.73

There is a direct correlation between the membrane width and the tilt angle corresponding to the designed sequence; more specifically, the increase in the width of the membrane corresponds to a decrease in the tilt angle. The increment of the membrane width with 2 Å corresponds to a decrease in the tilt angle with 4-5°. In the last instance (h=36), the decrease of the tilt is 3°, indicating a corresponding tilt angle of 28.73°. In this case, the positioning of the helix is influenced by two factors. The hydrophobic mismatch plays a role in the modification of the tilt angle. The difference between the length of the helix and the membrane width impacts the orientation of the helix, exhibiting a large tilt angle (53°) at small membrane widths (24 Å), which is attenuated at a membrane width above 32 Å when the tilt decreases below 35°. Another impacting factor on the tilt angle is the presence of the lysine and the tyrosine, each present towards the extremities of the helix and their interactions with the polar region of a membrane drive the orientation of the helix, depending on the membrane width (distance between the polar region).

2.4.3.1. Evaluation of Hippo as a computational method

Implicit membrane modelling with Hippo evaluated helix bundle tilting within the membrane. In

addition to the minimal sequence complexity, the applied design rationale included the size and the amphiphilic feature of the membrane's lipid bilayer. Implicit membrane modelling with Hippo evaluated helix bundle tilting within the membrane. It is important to mention that despite the fact that membranes have a complex composition and architecture, the efforts for the protein design are primarily purposed for elucidation of the impact of the lipid bilayer on the structure (lipid-mediated conformational changes), stability (lipid-induced protein-protein association), and also, on the function of specific proteins. To gradually introduce the accuracy of the study, the complexity of the environment was also gradually introduced through the use of different membrane models. The gradual introduction of accuracy in the study allowed the validation of the rest of the principles applied to the design but gradually relaxed the restraining factors. In other words, it was avoided the application of a direct all-atom molecular dynamics simulation, due to the complexity of external factors, which could influence the overall conformation and increase the number of variables in a potential correlation sequence-structure-function.

For the design of membrane proteins, the consideration of the tilt angle in the membrane was an important aspect, and part of the rational design was to restrain the tilt angle, due to literature experimental data, which reported a specific tilt of approximately 23±10°. The screening of sequences in different membrane dimensions, depending on the orientation in a more simplified membrane model (intrinsic membrane), was a fast and validated method enabled with the aid of the Hippo software. These experiments using this type of software had proven to be robust. It was previously shown that the results obtained from the algorithm were in the expected ranges, as experimentally obtained from NMR structural measurements (Ulmschneider *et al.*, 2007). In the present study, there could be detected relationships between membrane dimensions and tilting with a further extent to the amino acids distribution in the membrane, more specifically, a mismatch between the protein-membrane dimensions (protein sequence length insufficient to span the membrane) or inadequate hydrophobic/polar (or charged) residue distribution would determine a tilt change towards the membrane normal on the X axis in an XYZ representation. The tilt angles of the CC1 sequence fell within the expected distribution of membrane proteins. The Hippo testing in this study involved the sampling of a single, individual helix due to the repetitive and symmetrical feature of the CC1 sequence, and the results of the helix tilting within the implicit membrane from one test could be assigned to the rest of the individual helices. The repetitive and symmetrical sequence between the transmembrane domains of other potential designs could enable the speed of the sampling. However, it has to be highlighted the limitations associated with the test, which were also encountered within the present study, namely, the sampling of the tilt within the membrane of the entire bundle (or tetramer), which would be rather relevant in the case, for example, of a dimer of

dimers the input file needing further modifications for this application. The limitation associated with the input required by Hippo is that it uses a single chain, recognised from the starting peptide amino group to the carboxyl end, making it difficult a clear separation between the contribution of each TM chain on an overall tilt. Hence, the single helix tilt angle would be important in the success of the design of a repetitive structure or for the assignment of individual tilt contributions, in case helices are designed separately, which are aimed to be part of a more complex structure. Also, the membrane 'snorkelling' effect of the lysine residues was further confirmed in experiments; in the case of lysine presence towards each of the sequence ends, it promoted the decrease in tilt against the membrane normal on the Z axis. This further emphasised on the importance of hydrophobic mismatch as the molecular driving forces dictating the membrane's depth fitting.

2.4.4. Molecular dynamics (MD) simulations for stability in explicit environments using Gromacs software

The simulations in explicit environments provide the possibility of determining the potential existence of interactions between the protein and the surrounding environment (lipids, solvent, ions), which can influence any structural conformations and stability of the protein. There were performed simulations in Gromacs to explore the dynamics of the three tetrameric bundles: CC1, REAMP and poly-leucine in symmetric POPC, POPE and POPG lipid bilayers, solvated with water molecules and supplemented with sodium (Na^+) and chloride (Cl^-) ions. The time-dependent dynamic evolution of the system (the bundles in the explicit environments) was comprised in the corresponding trajectory files, which indicated changes in the root-mean-squared-deviation at the protein's backbone level, but also localised changes when individual snapshots of the adopted conformations were visually inspected.

Notably, each molecular dynamics simulation was performed only once (n=1), due to computational performance limitations encountered. The computational results presented in the following subsections rather highlight preliminary observations of the interactions that may occur; however, they cannot be classed as definite, as multiple runs must be tested, in order to achieve computational convergence and express finite conclusions on the molecular events which can be expected in the given parametrised conditions.

2.4.4.1. Helical bundle stability

The stability of the three tetramers in different explicit environments was observed by comparison

of the corresponding root-mean-square deviation (RMSD) values. The RMSD plots for the C_α, presenting the differences between the adopted conformations of the backbone throughout the time of the simulations, were obtained from each corresponding trajectory file. Root-mean squared deviation (RMSD) trajectories over the simulation duration demonstrate that CC1 exhibits the greatest stability across all tested lipid bilayers (Figure 2.6). In the case of CC1, it can be seen in Figure 2.6A that the root-mean-square deviation plots have shown that the differences in the distances between the Cα in the adopted structural conformations in the trajectory towards the minimum are between 0.1-0.25 nm. On the other hand, when REAMP was simulated in POPC, POPE and POPG bilayers, it was observed a different behaviour. As can be seen in the RMSD plot presented in Figure 2.6B, when REAMP was embedded into POPE, the bilayer imposed resistance, decreasing the degree of freedom in terms of adopting different structural conformations towards the minimum energetic state. In contrast, the conformational dynamics are enhanced in the POPC and POPG bilayers. Furthermore, poly-leucine's initial conformation is drastically altered; the distances between the alpha-carbon atoms increased up to 0.8 nm, which was the maximum recorded value for the root-mean-square deviation, thus affecting the stability of the peptide when simulated in a POPC bilayer. The degree of freedom in terms of conformational changes was observed as decreasing when poly-leucine was simulated in POPE, respectively POPG, the peptide exhibiting dynamics with a similar trajectory.

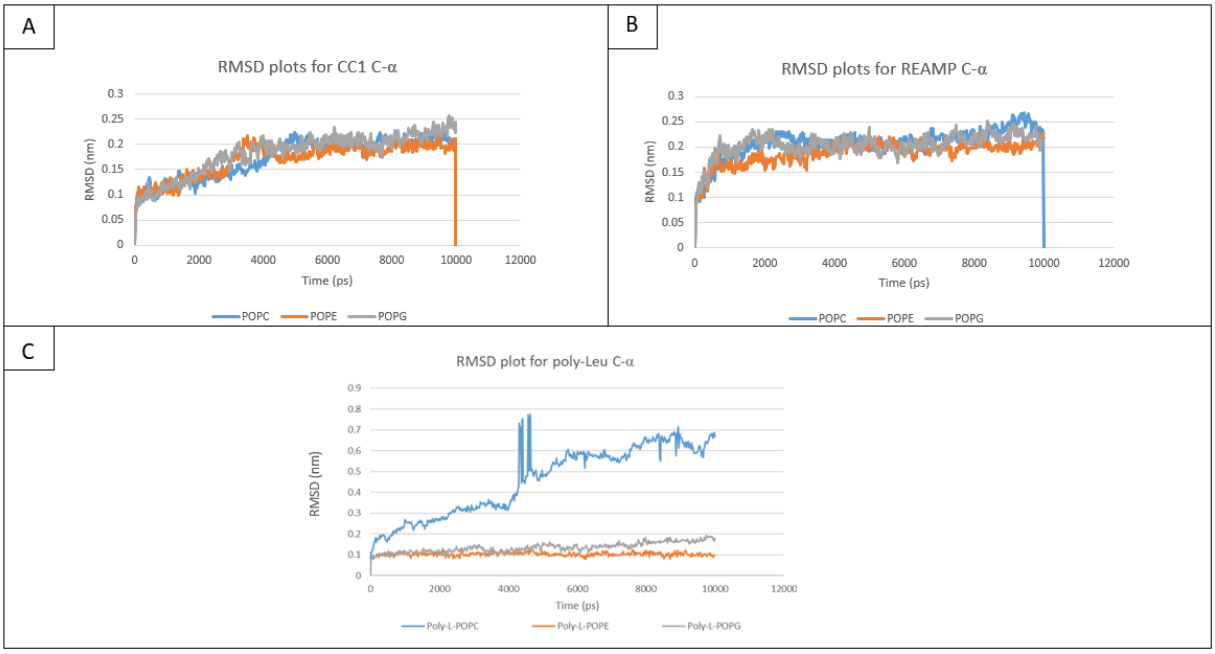


Figure 2.6(A-C). Root-mean-square deviation plots of CC1, REAMP and poly-L bundles in different lipid bilayers. On the X-axis is represented the time of the simulation measured in picoseconds (ps), the duration of the simulation

was 10000 ps. On the Y-axis, there are represented the structural differences measured between the C_α in the adopted conformations throughout the simulation. A. RMSD plot of CC1 C_α in POPC (blue line), POPE (orange line), and POPG (grey line) bilayers. The measured structural differences are in the range between 0.05-0.25 nm. B. RMSD plot of REAMP Cα in POPC (blue line), POPE (orange line), and POPG (grey line) bilayers. The measured structural differences are in the range between 0.05-0.25 nm. C. RMSD plot of poly-Leu C_α in POPC (blue line), POPE (orange line), and POPG (grey line) bilayers. The measured structural differences are in the range between 0.15-0.8 nm for simulation in POPC and between 0.1-0.2 nm in POPE and POPG.

Figure 2.7 A-C illustrates the comparisons between the RMSD plots of the three peptides when simulated in the same type of bilayer.

As can be seen in Figure 2.7A, corresponding to the RMSD plots obtained from simulations in the POPC bilayer, the poly-Leu exhibits the largest structural differences, while CC1 and REAMP present smaller deviations, the most stable being CC1.

The RMSD plot obtained from the simulations in the POPE bilayer shows that the most stable component appears to be the poly-leucine; however, the deviations of CC1 and REAMP are not concerning. The maximum value of the recorded deviation compared to the initial conformation is less than 0.3nm, as represented in Figure 2.7B.

The POPG bilayer appears to allow motion in equal measure for CC1, REAMP, and poly-leucine, as can be seen by comparison in Figure 2.7C. CC1 and REAMP have similar trajectories with structural deviations less than 0.25nm, and the structures stabilise after 4000ps. Whether poly-leucine, despite lower initial structural differences, the overall dynamics are inclined towards larger deviations.

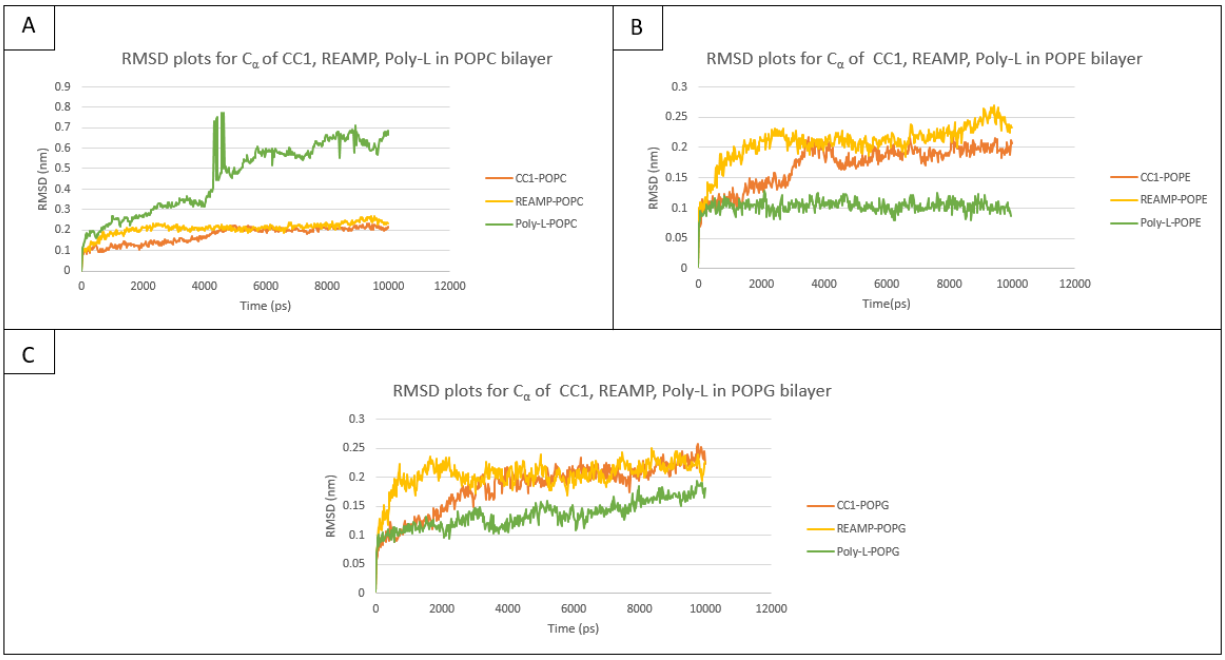


Figure 2.7 (A-C). Comparison plots of the root-mean-square deviations of CC1, REAMP and poly-Leu bundles in POPC,

POPE and POPG bilayers. On the X-axis is represented the time of the simulation measured in picoseconds (ps), the duration of the simulation was 10000 ps. On the Y-axis, there are represented the structural differences measured between the Cα in the adopted conformations throughout the simulation. A. Comparison between the RMSD plots of CC1 (orange line), REAMP (yellow line) and poly-Leu (green line) Cα in POPC bilayer. The measured structural differences are in the range between 0.1-0.8 nm for poly-Leu and between 0.1-0.3 for CC1 and REAMP. B. Comparison between the RMSD plots of CC1 (orange line), REAMP (yellow line) and poly-Leu (green line) Cα in. The measured structural differences are for all peptides in the range between 0.1-0.25 nm. C. Comparison between the RMSD plots of CC1 (orange line), REAMP (yellow line), and poly-Leu (green line) Cα in POPG bilayers. The measured structural differences measured for all peptides are in the range between 0.1-0.25 nm.

In addition to the structural differences observed for the alpha-carbons of each peptide, inspections of the snapshots of each of the trajectories indicated further structural differences and interactions with the surrounding lipid or solvent molecules, important in the determination of essential protein-lipid interactions as impact factors in the protein folding. In the following subsections there are presented the localised structural modifications, as observed in the Visual Molecular Dynamics (VMD) software.

2.4.4.2. Sequence-Structure relationships- Orientation of the helices of the tetramers

In the following subsections there are presented some of the identified out-of-ordinary protein structural conformational changes and protein-lipid interactions. Each figure, corresponding to different identified events, also contains a snapshot of the specific orientation of the tetramer (illustrated as ribbons) in the figure for better clarity. The colour scheme of the helices of the tetramer is from red to blue, and the helical termini (N- and C-) are illustrated in the colours with higher intensity. The helical end region coloured in intense red represents the N-terminus of the tetramer, and the intense blue represents the C-terminus. At appropriate stages, specific amino acid structures were highlighted in yellow. In Figure 2.8 there are exemplified different orientations of a tetramer (e.g. CC1 tetramer).

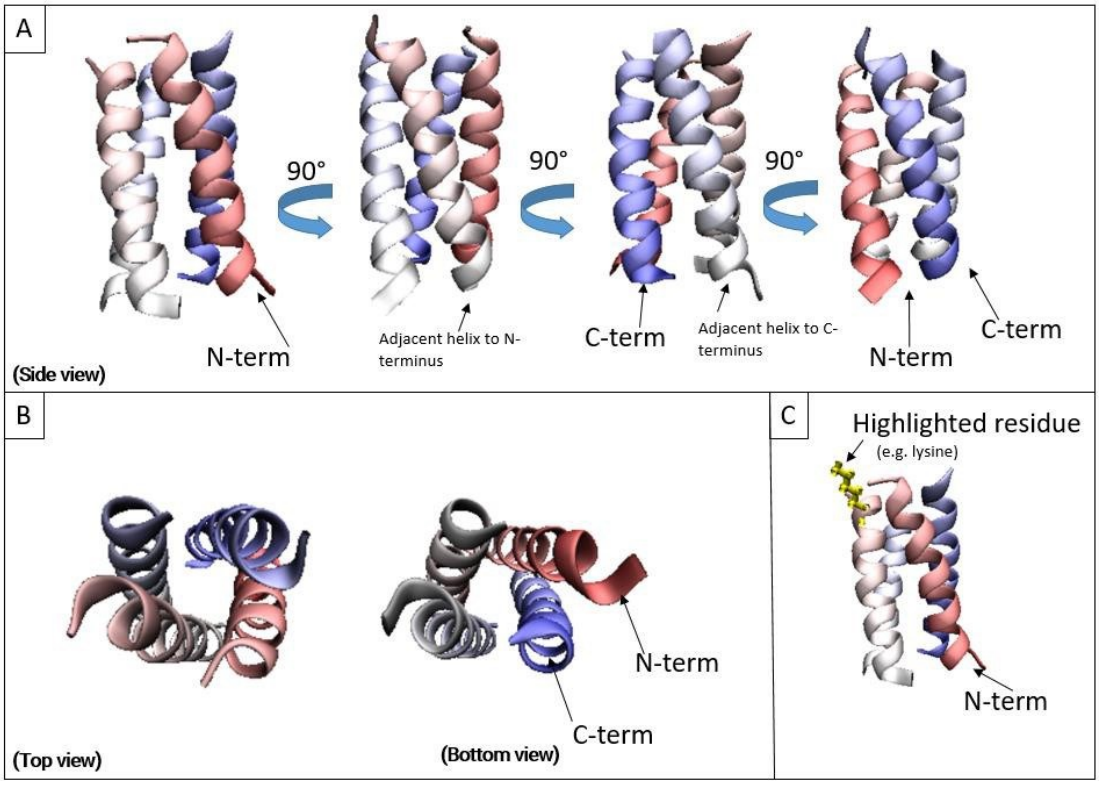


Figure 2.8. Different orientations of the tetramer. A. Side view of the tetramer. Different orientations of the tetramer. Rotations by 90 degrees allow the illustration of all helices. The N-terminus (red) and C-terminus (blue) and corresponding adjacent helices are indicated by arrows. B. Top and bottom view of the tetramer. The N- (red) and C- (blue) termini are indicated by arrows. C. Side view of the tetramer with the highlighted lysine residue illustrated in yellow and N-terminus (red) indicated by an arrow. (Created with Biorender)

2.4.4.3. CC1 embedded in POPC bilayer

Figure 2.9 (A-C) illustrates the structure of the CC1 tetramer (represented as ribbons) embedded in the solvated POPG bilayer (lipid and solvent molecules were removed for visual clarity), along with the orientation of the tetramer (from N—to C—termini) at different time points (A—starting point, B—midpoint, C—ending point) during a 100 ns simulation.

From the inspection of the corresponding trajectory, it could be observed the helical distortion of the transmembrane helix 1 and the series of conformational changes can be seen in Figure 2.9. This phenomenon could be explained by the presence of the lysine in the structure and by its known snorkelling behaviour in the lipid head group region, which determines the change in the structure of the helix, becoming elongated. When the lysine residue was highlighted in the visual representation of the simulation, it could be seen that the lysine is involved in multiple structural rearrangements. The bonds can be seen as stretching, and the amino acid adopts different conformations in order to favour the position and interactions of the NH_3^+ group with the solvent or with the lipid head groups. This reorientation of lysine can be seen between panels A-C. Around the midpoint of the simulation it can be seen the movement of the lysine is almost 45 degrees upwards, reaching an almost vertical position. In the snapshots corresponding to the second half of the simulation, it could be observed clearly the snorkelling behaviour of this lysine residue could be when bonds are stretched, and the Lys is found in a vertical position against the membrane. However, the conformation is not completely stabilised; it continues to move either by 90 or 180 degrees, the bonds suffering further torsion or bending. This reorientation across the trajectory demonstrates the predicted snorkelling behaviour of the lysine residue.

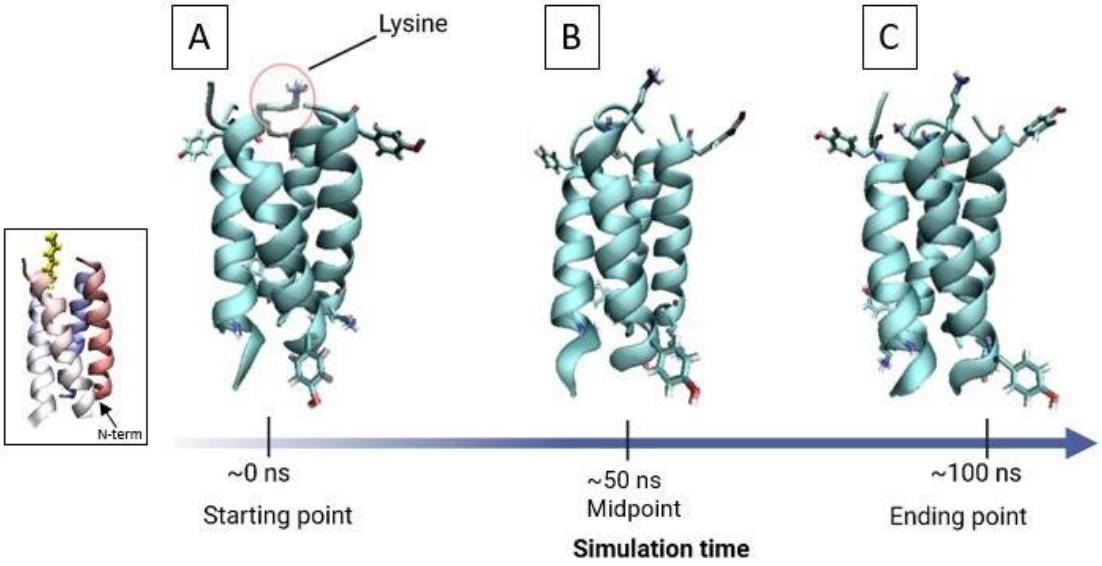


Figure 2.9. Snorkelling behaviour of lysine in the CC1 tetramer structure embedded in POPC. Snapshots from the simulation of CC1 in the POPC bilayer were obtained from the visualisation of the trajectory using VMD software. CC1 transmembrane helices are represented as ribbons (turquoise), with highlighted lysine and tyrosine residues, which are placed towards the limits of the helices. The snapshots of the structure were taken at different time points in the simulation: start point (A), midpoint (B) and endpoint (C). Highlighted in the orange circle (panel A) is the lysine residue, which corresponds to the helix region, which undergoes major structural changes and presents a distinct snorkelling behaviour into the polar region.

The behaviour of the surrounding lipids was also analysed. In Figure 2.10 there are represented the snapshots of the embedded CC1 tetramer embedded in the solvated POPC bilayer (lipids and solvent molecules represented as spheres) at different time points of the simulation (A- start, B- middle, C- end). Highlighted in purple circles are the lipid molecules POPC 78 and 91, which were found to interact closely with the region of interest throughout the entire simulation time. At the initial stage of the simulation, when the observed helical distortion can be distinguished, the NH_3^+ group interacted with an oxygen atom contained by the POPC 91 molecule. However, it can be detected a distinct structural modification of the helix twist. This is most likely due to the steric clashes imposed by the packing of a POPC molecule (POPC 91), which intercalates the acyl chain diagonally. This packing is also influenced by further steric clashes imposed by the orientation of the aromatic ring of the phenylalanine of the adjacent helix, oriented perpendicular to the lipids plan. After the simulation midpoint, the abnormal conformation of the helix disappears, and the correlation with the orientation of the acyl chain of POPC 91 can be confirmed; at this stage, the lipid tails change from a 'kinked' conformation to a closely packed structure, this change creating a larger void between the protein and lipids, allowing the re-formation of the helix. The re-formation of the helix is also allowed by the re-arrangement of the acyl chain of POPC 78, with which the

NH_3^+ group of the lysine interacts with the oxygen atom in the glycerol lipid backbone. The tails of POPC 78 pack closely in the midpoint frame (panel B), creating reduced lipid density and eliminating the void steric clashes imposed on the phenylalanine residue, allowing the adoption of a favourable conformation; hence, the helix to re-form. When the lipid tails of POPC 91 and POPC 78 were not packed closely, they re-adopt a 'kinked' conformation, leading to the investigation of other possible interactions which could influence the modifications in the helical structure. Comparing the structures observed around the midpoint of the simulation (not all figures included), there could be identified two phenylalanine residues within the two adjacent helices, the aromatic rings placed in proximity and with parallel orientation against each other, the possible formation of π-π stacking type of interactions could contribute to the stabilisation of the helix structure and re-formation. However, when the phenylalanine rings were placed perpendicular to each other, the interaction was disrupted, and the helix structure was affected. Towards the end of the simulation, it could be seen that a distinct packing of the POPC 78 and 91 and the lysine residue interacted with the choline group of the head group of POPC 91, the phosphate group of an adjacent lipid molecule.

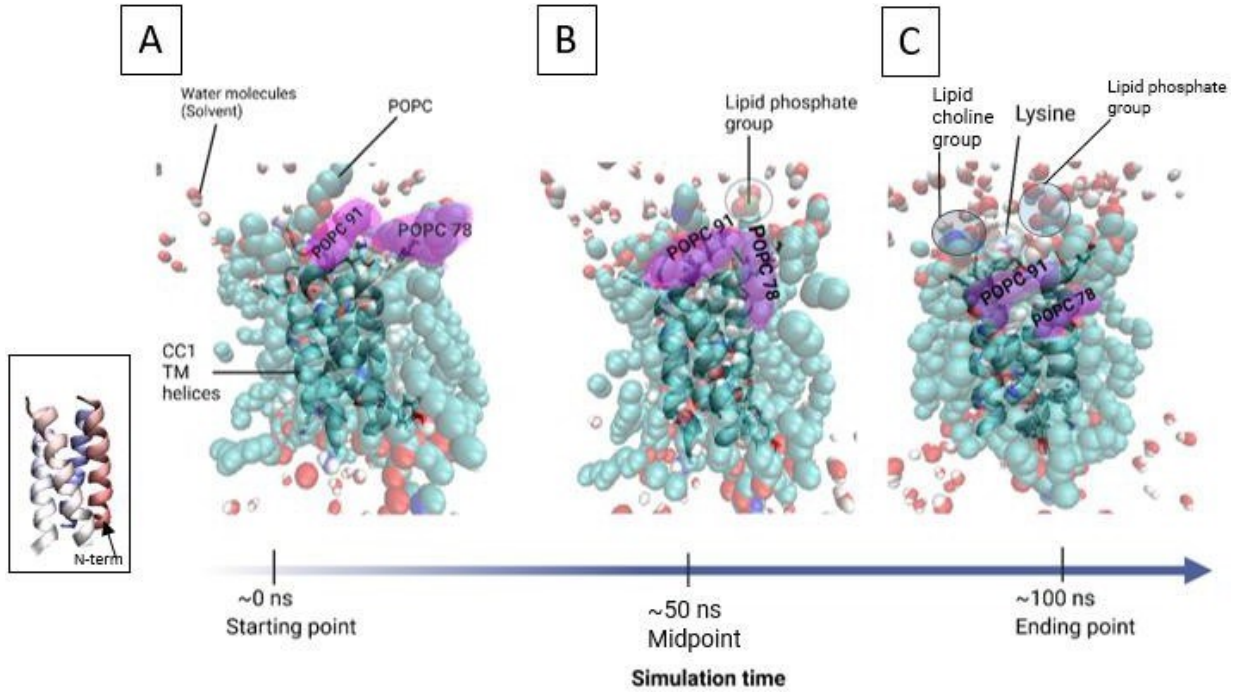


Figure 2.10. Protein-lipid interactions were observed for CC1 embedded in the POPC bilayer. Snapshots are taken from the visualisation of the trajectory file using VMD software. The CC1 helical structure is illustrated as ribbons. There are presented only the lipids within a distance of 5 Å from the structure of the protein. The tails of the POPC molecules are illustrated as turquoise spheres, and the phosphate atom in the lipid phosphate group is yellow. In the surrounding environment are present also the solvent (water) molecules. The three snapshots were taken from different time points of the simulation (around the start, midpoint and end). The lipid molecules which were observed to change position throughout the simulation are highlighted in purple balloons (POPC 91 and POPC 78).

2.4.4.4. CC1 embedded in POPE bilayer

At the beginning of the simulation, it could be observed the conformational changes of a tyrosine residue, as illustrated in Figure 2.11 B. As observed from the visual analysis of the trajectory, the simulation started with the tyrosine having the aromatic ring positioned parallel to the plan of the lipids and the hydroxyl group embedded towards the hydrophobic core of the bilayer. The adopted conformational changes oscillated between this orientation and a tendency of perpendicular positioning of the aromatic ring on the plan of the lipids. The adopted conformational changes are a result of the tendency of hydrogen-bonding type of interaction between the hydroxyl group of the tyrosine residue and the solvent and lipids head groups. It could also be observed that the level of helical distortion could be assigned to the high amount of water molecules interacting with the protein. When the interaction of the tyrosine residue with the solvent was stopped throughout the conformational changes, the helical structure was reformed. In Figure 2.10 A there is represented the CC1 tetramer (illustrated as ribbons) and the 5 Å surrounding area of the solvated POPE bilayer (represented using the VDW representation) in which the tetramer was embedded. It can be seen in Figure 2.11 A that there could be considered a correlation between the helical distortion and the lipid density. When the structure of the helical region was distorted, the lipid density in that region was decreasing. In Figure 2.11 A are labelled the POPE molecules which could be identified to interact with the helical region of interest. The POPE molecules which tend to interact with the highlighted tyrosine are the POPE 95 and POPE 115. The decrease in the lipid density between these two lipid molecules could be correlated with the conformational changes of the tyrosine and the possibility of the orientation of the hydroxyl group towards the oxygen atoms in the lipid head groups. In addition to POPE 95 and POPE 115, there could be observed consistent interactions with POPE 131 and POPE 135. It was shown a tendency of the POPE 131 lipid molecule to support the stabilisation of the structure of the helical region, the upper orientation of the tyrosine and contributed to the deterministic lipid density mentioned above.

The inspected individual trajectory frames, as illustrated in Figure 2.11, indicated a distinct arrangement of the detected interacting lipid molecule. It could be observed a dislocation of the POPE 95 molecule, hence the interaction disruption with the tyrosine residue.

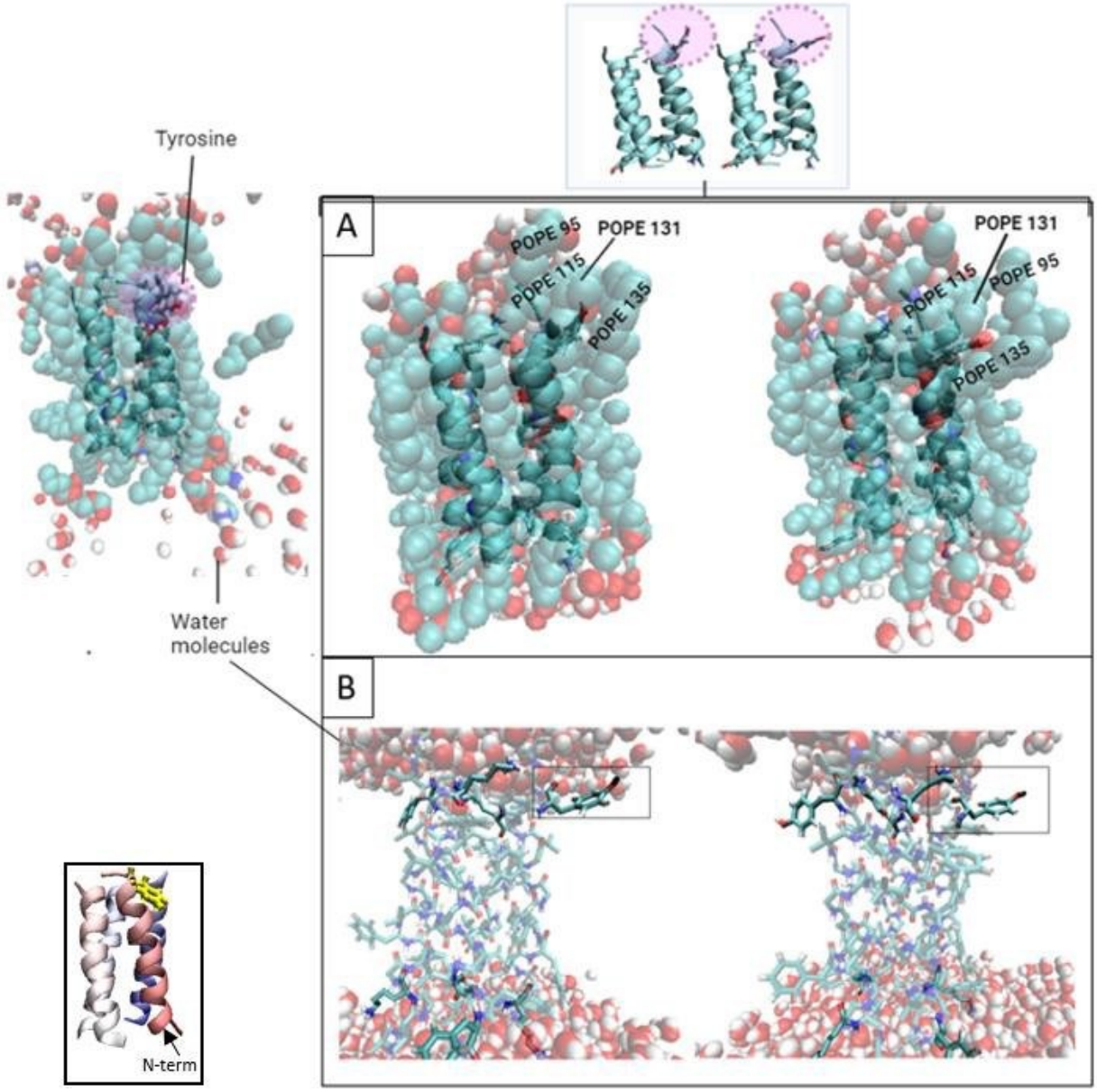


Figure 2.11. CC1 embedded in POPE: Conformational changes and interactions of tyrosine with the environment. Snapshots were obtained from the visualisation of the trajectory using VMD software. The snapshots correspond to the initial frames after the starting point of the simulation. (Left) is presented by the peptide and the surrounding environment present at a distance of 5 Å from the tetramer. The tyrosine which undergoes conformational changes is highlighted in a purple balloon. (Right) Snapshots of the tetramer highlighting the conformational changes of the tyrosine. A. Snapshots of the tetramer with the tyrosine in different conformations and surrounded by the explicit environment (lipid tails illustrated as turquoise spheres) and water molecules (illustrated as red and white spheres). The transmembrane helices are illustrated as ribbons (turquoise). There are highlighted the lipid molecules which were identified to change position or contribute to protein-lipid interactions along with the protein conformational changes: POPE 95, POPE 115, POPE 131 and POPE 135. B. Snapshots of the tetramer with the tyrosine in different conformations and surrounded by the solvent. The transmembrane helices are represented as a stick (bonded) model, with the tyrosine interacting with the solvent and changing conformations highlighted in borders.

As can be seen in Figure 2.12, there are some bulk water molecules which trespass the lipid bilayer, possibly due to the decrease in the lipid density in that region. Although not shown, it could be seen that the attraction forces between the highlighted tyrosine determined the orientation of the helices and the lipid density (the lipid molecules are not shown in Figure 2.12 for better visual clarity), and so, the number of water molecules able to migrate towards the lipid bilayer. The water molecules, which could enter the core of the bilayer, would be able to interact with the oxygen atoms within the protein's backbone through hydrogen bonds.

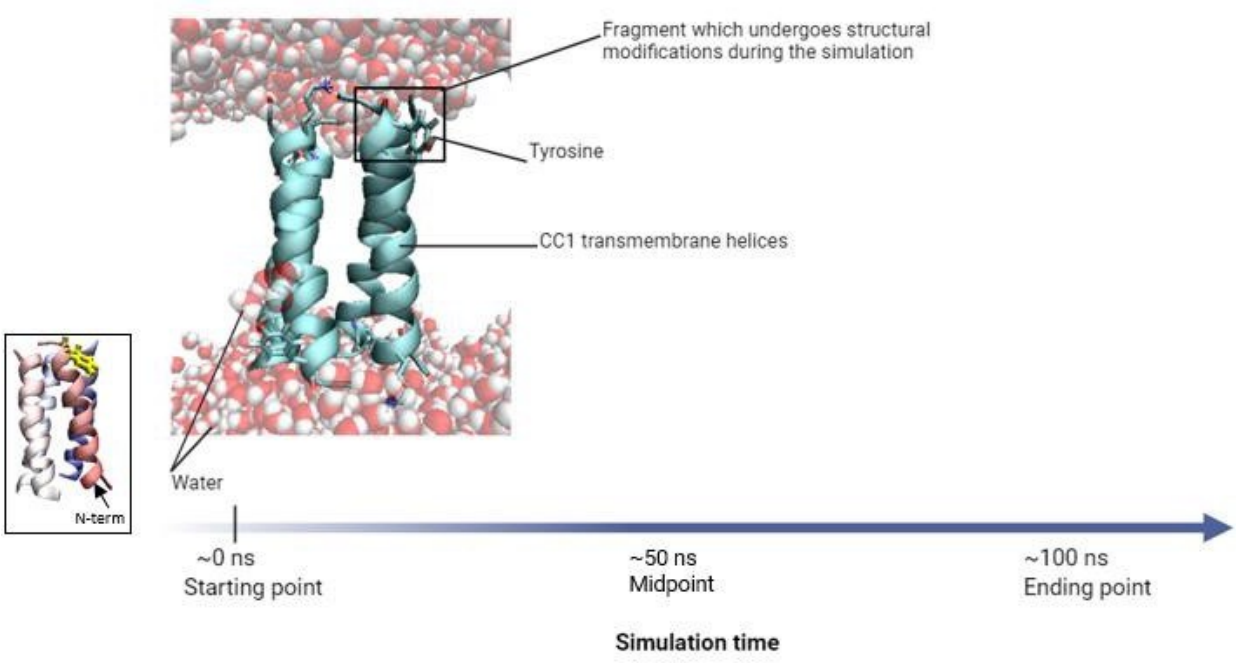


Figure 2.12. Snapshot of CC1 embedded in POPE bilayer highlighting the interactions between CC1 and solvent. It is highlighted the region in the helix which undergoes structural modifications during the simulation. In this region, it is also highlighted one of the tyrosine residues. One of the CC1 transmembrane helices (turquoise ribbons) interacts with solvent (water) molecules illustrated as white and red spheres.

In Figure 2.13 (A-C) there are shown the modifications of the specific protein-lipid interactions, as well as selected lipid dislocations or re-arrangements at different time points observed during the 100 ns simulation. It could also be seen that the tilt of the tetramer changed during the simulation. Towards the endpoint of the 100 ns simulation, as shown in Figure 2.13 C, the tetramer is almost perpendicular to the plan of the lipids, having a low average tilt between the tetramer helices. This orientation of the tetramer facilitated the interactions between the highlighted tyrosine residue and the oxygen atoms of the glycerol backbone of the lipid head groups. During the tilt modification, around the midpoint of the simulation, it

could be seen that the structure of the tetramer helices suffered a compression to some degree in the area placed in the centre of the bilayer, as represented in Figure 2.13 B, resulting in the kinking of the helices.

Regarding the protein-lipid interactions, there could be seen that some interactions with the helical region of interest were disrupted around the midpoint of the 100 ns simulation. Due to the change in the lipid density and arrangement, the interactions with POPE 131 and 135, as detected in the initial stage (Figure 2.13 A), could not be detected in an area of 5 Å within the protein in the middle range of the trajectory frames, however, it can be seen to be recovered with POPE 131 towards the completion of 100 ns (Figure 2.13 C). As a comparison between panels A and C, the interaction between the tyrosine and the POPE 131 and 135 molecules was stronger in the beginning (Figure 2.13 A) due to the arrangement of the lipids, especially of POPE 135. Also, as can be seen between panels A and C, the presence of the lipid molecule POPE 135 in the proximity of the tyrosine residue influenced the structural modifications of the helix (Figure 2.13 A), respectively influenced helical reformation when dislocated (Figure 2.13 B and C).

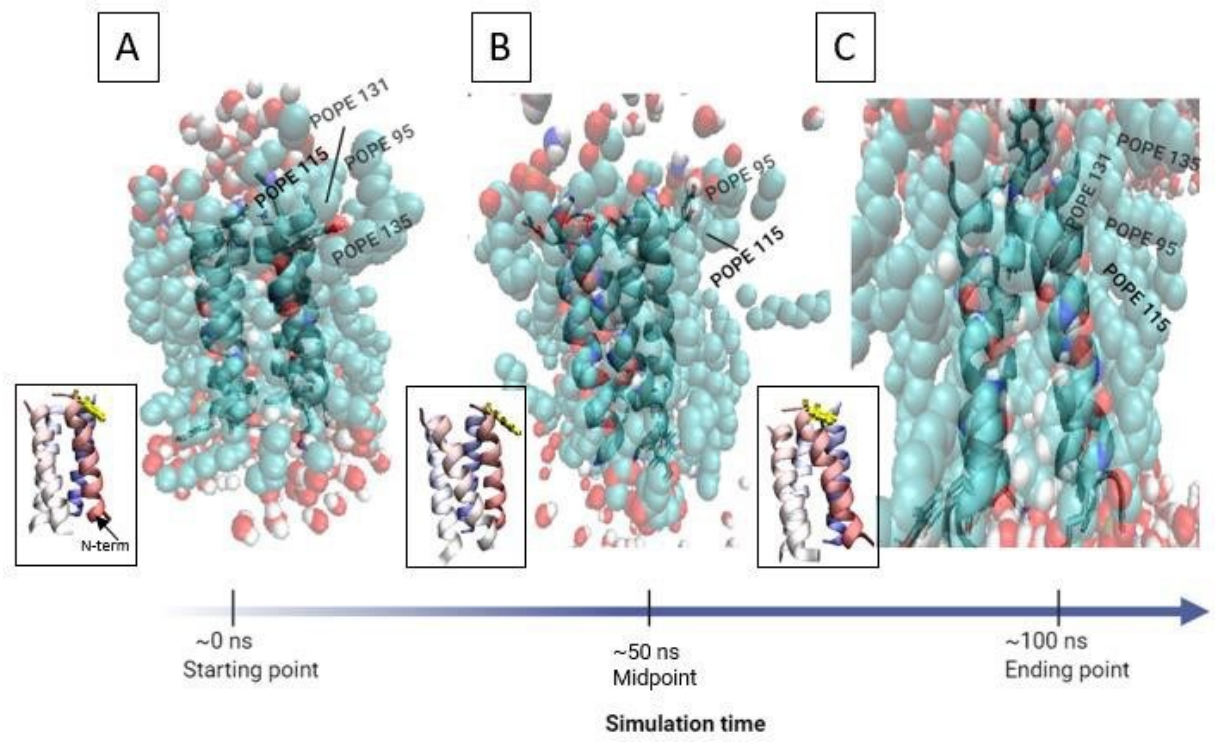


Figure 2.13. Protein-lipid interactions were observed for CC1 embedded in the POPE bilayer. Snapshots were obtained from the visualisation of the trajectory using VMD software and taken from different time points of the simulation (from left to right: starting point, midpoint and end). There are illustrated the tetramer (helical structure represented as turquoise ribbons) embedded in the explicit environment (selected shell thickness of 5 Å distance from the tetramer), which consisted of solvent (water molecules illustrated as white and red spheres) and lipid molecules (lipid tails pictured as turquoise spheres in the van der Waals representations). At each time point, the lipid molecules which interacted with the helical region which changed conformations, were labelled

correspondingly. At the starting point, the interacting lipid molecules were POPE 95, POPE 115, POPE 131 and POPE 135. Around the midpoint of the simulation, the identified interacting lipids were POPE 95 and POPE 115. Towards the end of the simulation, the highlighted lipids were POPE 95, POPE 115, POPE 131 and POPE 135.

In Figure 2.14 there are highlighted the changes in the tilt of the CC1 tetramer between the starting point of the simulation (Figure 2.14 A) and the ending point (Figure 2.14 B). In each of the panels at the top there is highlighted the CC1 tetramer (illustrated as tubes) embedded in the solvated POPE bilayer (VDW representation). At the bottom of the panels there is represented only the CC1 tetramer, the lipid and solvent environments were removed for visual clarity. In order to assess the differences in the tilt, there were measured distances between the alpha-carbon atoms (C_α) of the amino acids placed at the ends of the helices - alanine and valine. To distinguish the placement of the amino acids within the antiparallel bundle, the helical ends were labelled as positions A and B. The corresponding values were compared between the start and the end of the simulation. The helices were labelled from 1 to 4, starting from the N-terminus containing helix. The graphical representation could indicate that the helices 1 and 4 had large differences in the tilt angles.

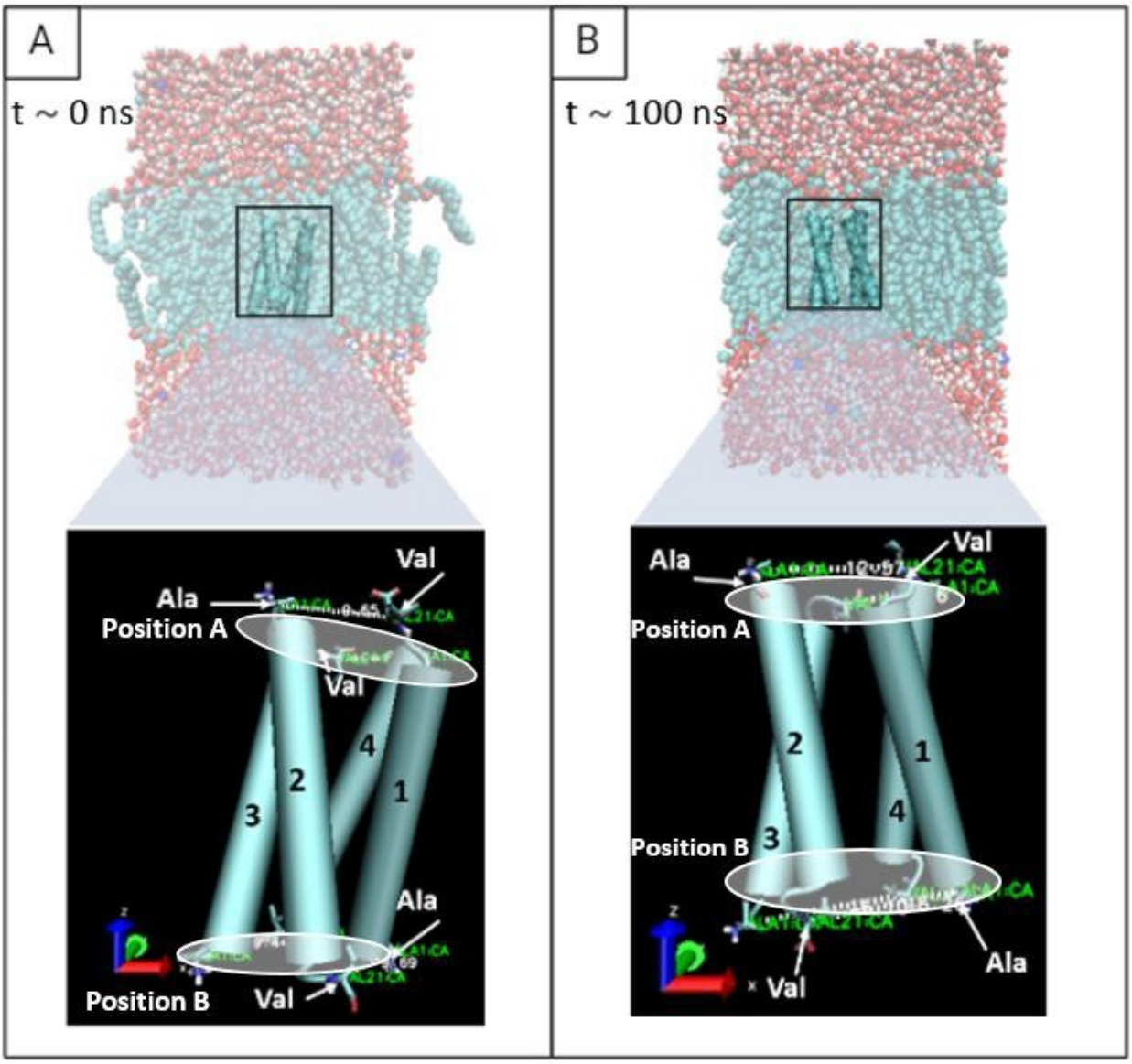


Figure 2.14. Helix structural repositioning at different time points of CC1 tetramer embedded in POPE bilayer. A, B (Top) Snapshots at different time points (A. t~0 ns/ starting point; B. t~100 ns/ ending point) of the simulation of the CC1 tetramer embedded in the POPE bilayer (side view). Graphical representation of the simulation box comprising solvent molecules (red and white spheres), lipid molecules (turquoise spheres) and embedded CC1 tetramer (turquoise bars). The CC1 tetramer is highlighted in the black border. (Bottom) Detailed representations of the changes in the positions of the transmembrane helices. The helices are labelled 1-4, and the limiting amino acids of each helix are alanine and valine.

The distances between the helices measured in angstroms (Å) are listed in Table 2.4. In addition to the visual information, the values of the measured distances provided insight into a potential pore-gating mechanism formed by the helical ends. The comparison between the distance values of positions A and B at the start and the end of the simulation indicated that the distances at the position A of the helices are within a range of 1-3 Å. In the case of position B, the distances measured at the start and at

the end showed differences within the range of 5-12 Å. The tilt angle measurements indicated that helix 3 exhibited most orientation modifications and showed a tilt increase by approximately 12 degrees over the 100 ns trajectory.

Table 2.4. Measured distances between C_α of the limiting amino acids of the CC1 helices embedded in the POPE bilayer. Comparison between different helices pairs (column 1) and helix regions (position A and position B) at the start and the end of the simulation.

Helices pairs	**Measured distances between C_α of the limiting amino acids of the helices (Å)**			
	Simulation start		**Simulation end**	
	Helix region		Helix region	
	Position A	Position B	Position A	Position B
1-2	9.65	9.69	12.57	16.26
3-4	8.12	9.44	8.26	16.10
1-3	10.53	9.48	7.32	20.64
2-4	15.69	8.84	16.57	20.34

2.4.4.5. CC1 embedded in POPG bilayer

A change in the tilt of the tetramer was also observed in the case of CC1 embedded in the POPG bilayer. Figure 2.15 (A-B) includes at the top the CC1 tetramer highlighted in a border (tube representation) and embedded in the solvated POPG lipid environment (VDW representation). At the bottom of the figure, there are simplified representations; the lipid and solvent molecules were not included in order to enhance the clarity of the graphical representation. The change in the tilt was determined by measurements between the alpha-carbon (C_α) atoms of the helix-delimiting amino acids alanine and valine. The limits of the antiparallel helices were considered as positions A and B, as highlighted in Figure 2.15 A-B. The simplified graphical representations in Figure 2.15 A and B showed that the helices which would exhibit the largest tilt would the helices 2 and 3, as labelled in the figure.

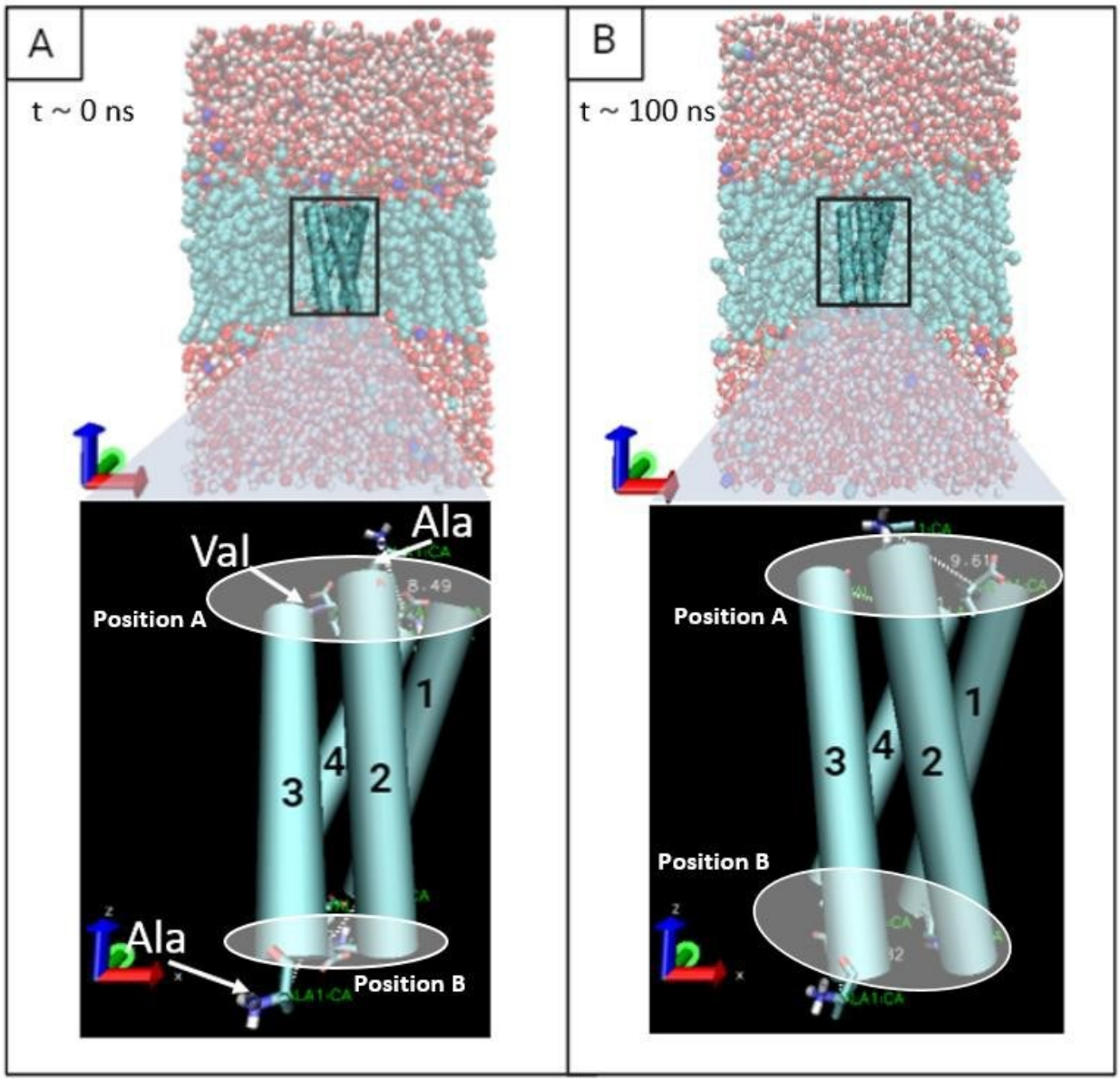


Figure 2.15. Helix structural repositioning at different time points of CC1 tetramer embedded in POPG bilayer. A, B (Top) Snapshots at different time points (A. t~0 ns/ starting point; B. t~100 ns/ending point) of the simulation of CC1 tetramer embedded in POPG bilayer (side view). Graphical representation of the simulation box comprising solvent molecules (red and white spheres, as VDW representation), lipid molecules (turquoise spheres, as VDW representation) and embedded CC1 tetramer (turquoise bars). The CC1 tetramer is highlighted in the black border. (Bottom) Detailed representations of the changes in the positions of the transmembrane helices. The helices are labelled 1-4, and the limiting amino acids of each helix are alanine and valine.

The distances measured in angstroms (Å) between the alpha-carbon atoms of the limiting helical pairs in the positions A and B at the start and at the end points of the simulation are included in Table 2.5. As a comparison with the case of CC1 embedded in POPE, in this case, the degree of the potential opening of the "pore" formed by the amino acids placed at the extremities of the helices shows an increase for

both positions; however, it indicated a similar trend. Between the start and the end of the simulation, the differences in the recorded distances between the amino acids in position A varied in a range within 1-7 Å. The distance differences for position B varied within a range of 2-12 Å. Similar to POPE bilayer observations, helix 3 exhibited a tilt increase of approximately 5 Å, and helix 4 exhibited a tilt of 4 Å, over the course of 100 ns simulation. In this case, also similar to CC1 embedded in POPE, position B confers wider possible pore opening throughout the simulation compared to position A.

Table 2.5. Measured distances between C_α of the limiting amino acids of the CC1 helices embedded in the POPG bilayer. Comparison between different helices pairs (column 1) and helix regions (position A and position B) at the start and at the end of the simulation.

Helix pairs	**Measured distances between C_α of the limiting amino acids of the helices (Å)**			
	Simulation start		**Simulation end**	
	Helix region		Helix region	
	Position A	Position B	Position A	Position B
1-2	9.61	9.79	8.49	16.23
3-4	9.91	10.82	11.17	22.50
1-3	11.90	20.53	6.25	22.50
2-4	11.97	7.14	17.80	10.48

In addition to the side view representations of the tilt change illustrated in Figure 2.15, a bottom view of the POPG-embedded CC1 tetramer at the start and at the end of the simulation is represented in Figure 2.16 A-B. This figure enabled the determination of the specific helix for which the tilt angle with respect to the lipid bilayer changes the most compared to the others. In this case, it could be seen that, in fact, the N-terminus containing helix (helix 1 as labelled in Figure 2.15) has the largest tilt angle at the end of the simulation, which could not be visible only from the side view representation in Figure 2.15. The identification of the N-terminus helix can be done with the aid of the additional label imposed on the figure, containing the orientation of the tetramer and the standard colour scheme red-blue corresponding to the N- and C-termini. The change in the tilt angles of the helices 2 and 3 are also visible in the bottom view representation, as identified previously in Figure 2.15.

In panel B, which corresponds to a snapshot taken at the end of the simulation, the N-terminus helical region was distorted.

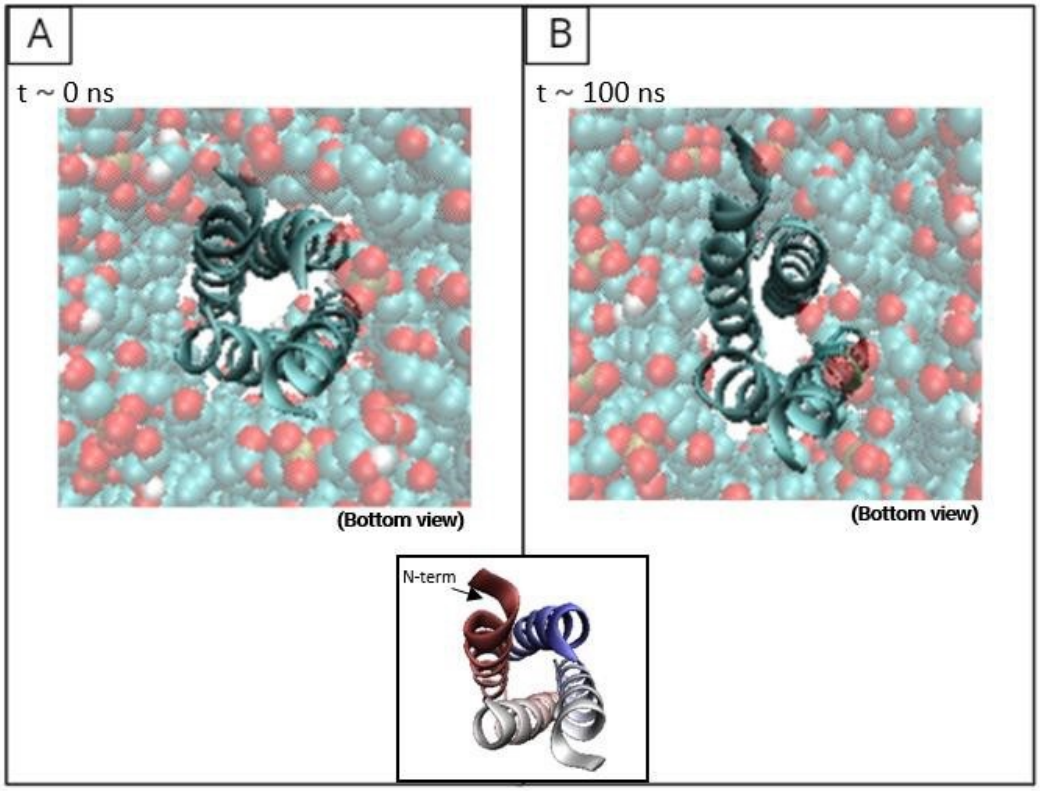


Figure 2.16. Helical repositioning and structural modifications in the CC1 tetramer structure embedded in the POPG bilayer. A, B. Snapshots (bottom view) of the simulation of CC1 embedded in POPG bilayer from different time points (A. t~0 ns/starting point; B. t~100 ns/ending point). A, B. Graphical representations of the bottom lipid monolayer and the CC1 tetramer (turquoise ribbons) in the centre of the lipid box (spheres/ VDW graphical representation). A. Initial position of the helices. The helices are labelled from 1 to 4, as found in the CC1 tetramer. B. The final position of the helices. The arrow indicates the first helix, for which occur the most structural modifications.

As it could be seen in Figure 2.1, the N-terminus helical region was distorted during the simulation, however not only towards the end of the simulation, as seen in the previous figure. The helical region was distorted since the initial stage, as shown in Figure 2.17. The snapshot presented in Figure 2.17 has the CC1 tetramer illustrated as ribbons. However, the surrounding lipid and solvent environment is illustrated as a simplified version of VDW (CPK) for increasing the visual clarity. The area containing the distorted N-terminus helical region is highlighted in grey. In this figure, it can be observed the interaction (possibly electrostatic) between the N-terminus nitrogen atom of the alanine residue and the oxygen atoms of the phosphate groups contained by the POPG 29 and POPG 39.

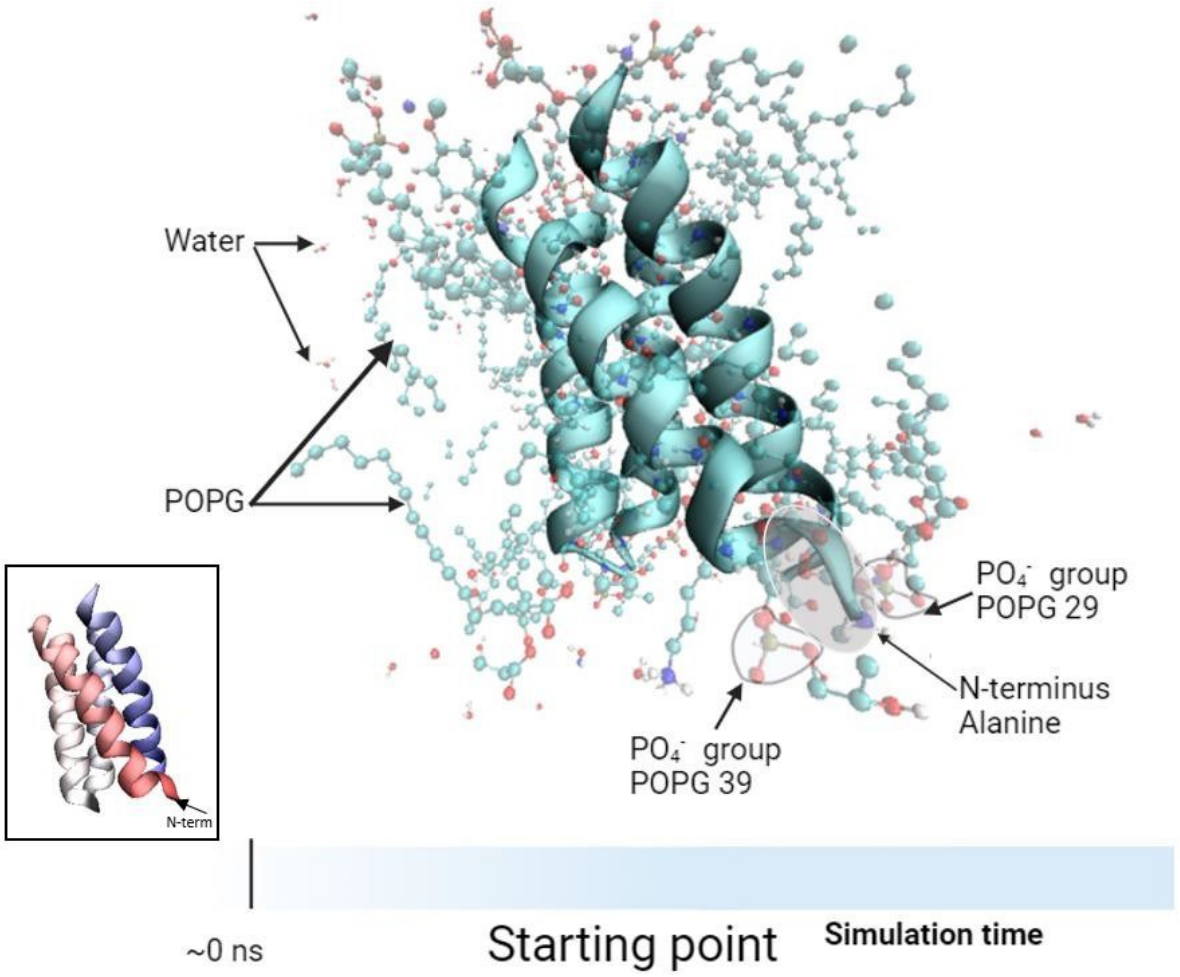


Figure 2.17. Protein-lipid interactions were observed for the N-terminus of the CC1 tetramer embedded in the POPG bilayer at the beginning of the simulation. Snapshot of the simulation of CC1 tetramer (turquoise ribbons) embedded in POPG (spheres, CPK representation in VMD software) bilayer (side view). The selected lipid shell thickness is within a distance of 5 Å from the peptide bundle. The N-terminus of alanine interacts with the phosphate groups of the POPG molecules 29 and 39. Arrows indicate surrounding POPG and solvent (water) molecules.

Lysine snorkelling behaviours were exhibited, with flexible side chain conformational sampling enabling hydrogen bonding with lipid phosphate moieties. The corresponding snapshots are included in the Figure 2.18 (A-C). In panel A there is illustrated the side view of the CC1 tetramer (ribbons) embedded in the POPG bilayer (simplified representation as lines for visual clarity). It can be seen that the lipid density towards the bottom of the tetramer is decreased, and the helical ends towards the bottom are dispersed. The lipid density towards the top of the helix is higher compared to the opposite end. The arrangement of the lipids around the centre of the bilayer was shown to not be in the closest proximity of the N-terminus-containing helix. It could be considered an influencing factor in the change of the tilt angle. Also, it could be observed that the top of the tetramer (Figure 2.18 B) is stabilised by the interaction between

the nitrogen atom of the lysine, the hydroxyl group of the tyrosine or the nitrogen atoms within the backbone and five lipid phosphate groups. The number of identified interactions decreased with the decrease of the lipid density at the bottom of the tetramer (Figure 2.18 C), and the tetramer tended to interact rather with the oxygen atoms in the glycerol lipid head group. The exact tetramer orientations are included as labels highlighted on the borders.

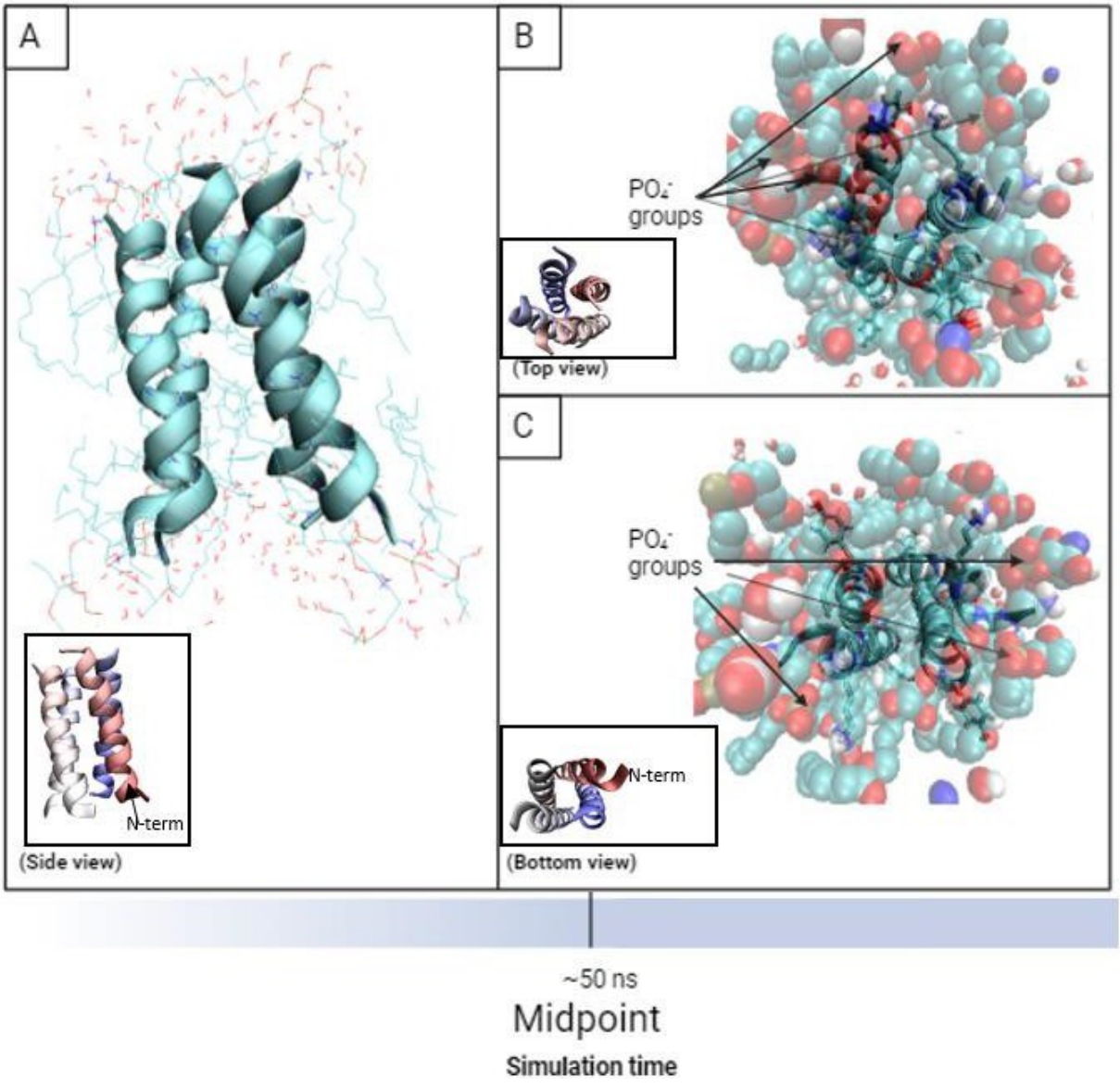


Figure 2.18. Protein-lipid interactions were observed for the CC1 tetramer embedded in the POPG bilayer around the midpoint of the simulation. A, B, C. Snapshots of the CC1 tetramer (turquoise ribbons) embedded in the solvated POPG bilayer (in A illustrated as lines and in B and C illustrated as spheres/VDW representations, selected lipid shell thickness was 5 Å distance from the tetramer). A. (Side view) Higher lipid density in the upper lipid leaflet of the POPG bilayer compared to the lower leaflet. B. (Top view). The upper limits of all four helices interacted with five

lipid phosphate groups (indicated with arrows). C. (Bottom view). The lower limits of only two helices interacted with three phosphate groups (indicated by the arrows).

2.4.4.6. REAMP embedded in POPC bilayer

Figure 2.19 (A-C) represents the REAMP tetramer (ribbons) when embedded in the solvated POPC bilayer (the lipid and water molecules were not shown for visual clarity) at different time points (A- star point, B- midpoint, C- ending point) in the 100 ns simulation. In the grey circle it is highlighted the tryptophan residue contained by the helix adjacent to the N-terminus helix. The orientation of the tetramer in the snapshots is shown in the reduced-size figure in the bottom left corner and highlighted in the border. The label also contains the tryptophan residue highlighted in yellow. From the comparison of the snapshots, it could be observed that the helical region containing the tryptophan residue is distorted throughout the simulation (Figure 2.19 B and C); the tryptophan adopted a series of different conformations, which influenced the degree of helical distortion.

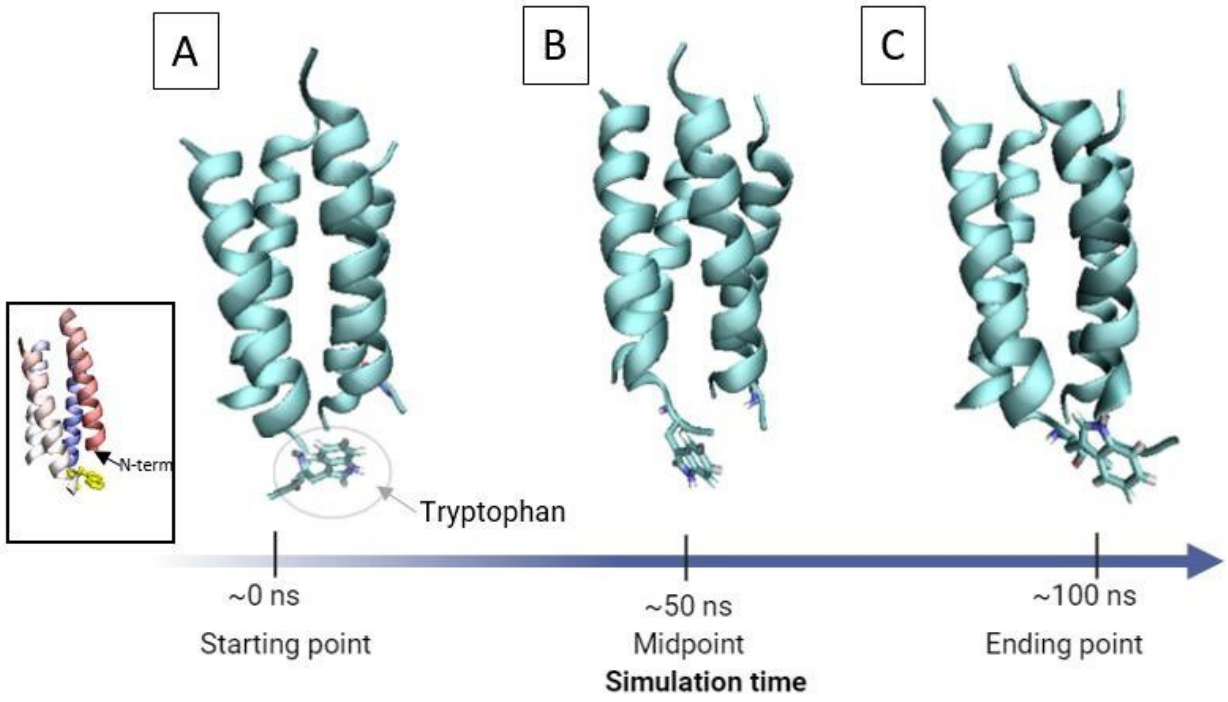


Figure 2.19. Tryptophan-containing helix structural rearrangement of REAMP embedded in POPC bilayer. Snapshots taken from the simulation of REAMP embedded in POPC bilayer at different time points (from left to right: starting point, midpoint and ending point) illustrating the conformational changes of tryptophan (highlighted in grey circle) and helix regional distortion.

The observed helical distortion was influenced by the protein-lipid interaction which was occurring in that specific helical region.

In Figure 2.20 (A-C) there are presented the different time points snapshots, including the graphical representation (VDW representation) of the solvated POPC lipid bilayer. The solvated bilayer represented in the figure is the area at a distance within 5 Å from the tetramer. It could be observed that the POPC molecule, which interacts with the tryptophan residue across the entire duration of the simulation (100 ns), was POPC 76. The lipid molecule interacts with the amino acid with the phosphate group. The helical distortion it could be correlated with the POPC 76- TRP interaction, but also with the adopted conformation of the TRP. In the case of the interaction of the phosphate group of POPC 76 with the hydrogen atoms of the aromatic ring, the hydrogen bonding did not cause the helical distortion (Figure 2.20 A). This was caused in the case of the parallel orientation of the aromatic ring to the phosphate group, generation π (pi)-cation interactions and driving the helix structural modifications.

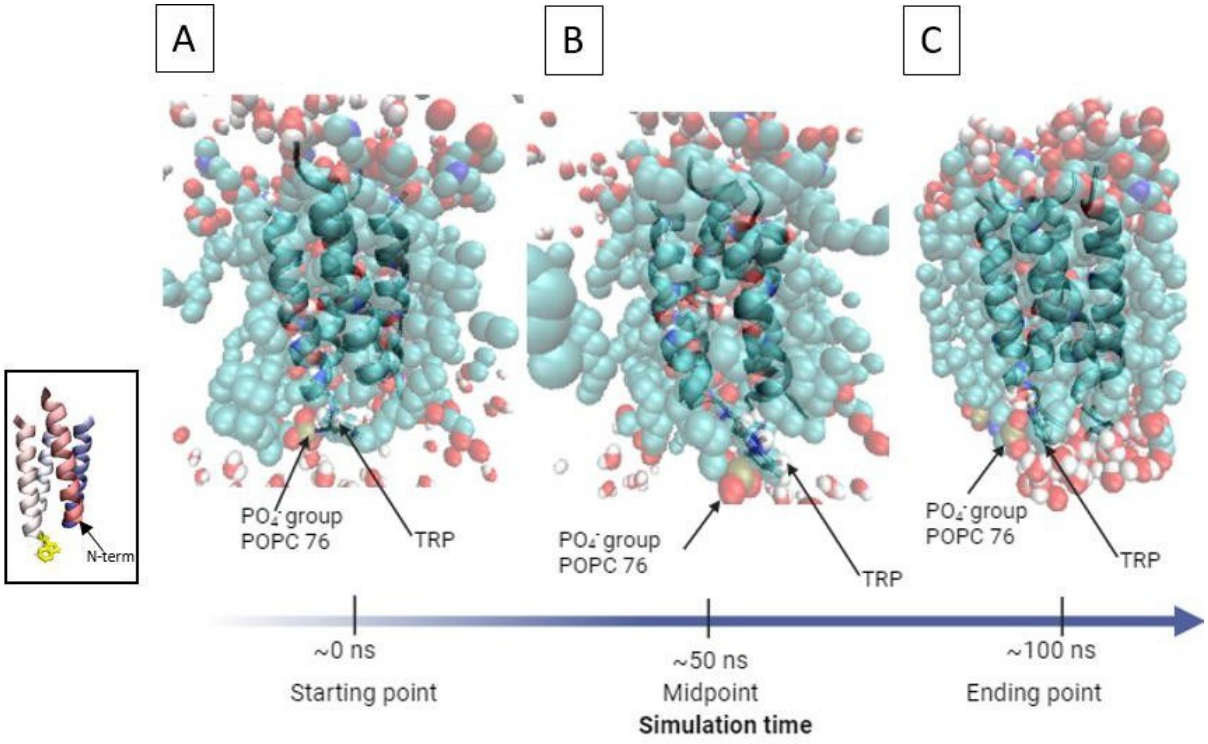


Figure 2.20. Protein-lipid interactions were observed for REAMP embedded in the POPC bilayer. Snapshots from the simulation (side view) taken at different time points (from left to right: starting point, midpoint, ending point) illustrating the REAMP tetramer (turquoise ribbons) embedded in the solvated lipid bilayer (spheres/VDW representation, selected lipid shell thickness is within 5 Å from the protein) and, the specific interactions of the tryptophan with the phosphate group of the POPC 76 lipid molecule, corresponding with regional helical distortion.

2.4.4.7. REAMP embedded in POPE bilayer

In Figure 2.21 (A-C) there are shown different snapshots of the REAMP when embedded in the POPE bilayer (lipid molecules are not shown for visual clarity) at different time points of the simulation (A- start, B- midpoint, C- end) and the tetramer orientation highlighted in the border. There could not be detected any regional helix structural modifications. This phenomenon was also confirmed by the root-mean-square deviation plots, which showed an increased stability of the REAMP in the POPE bilayer. In the present figure in can be detected a small change in the tilt angle of the helix containing the N-terminus. In panel C, it can be also be detected a kink in the helix containing the C-terminus.

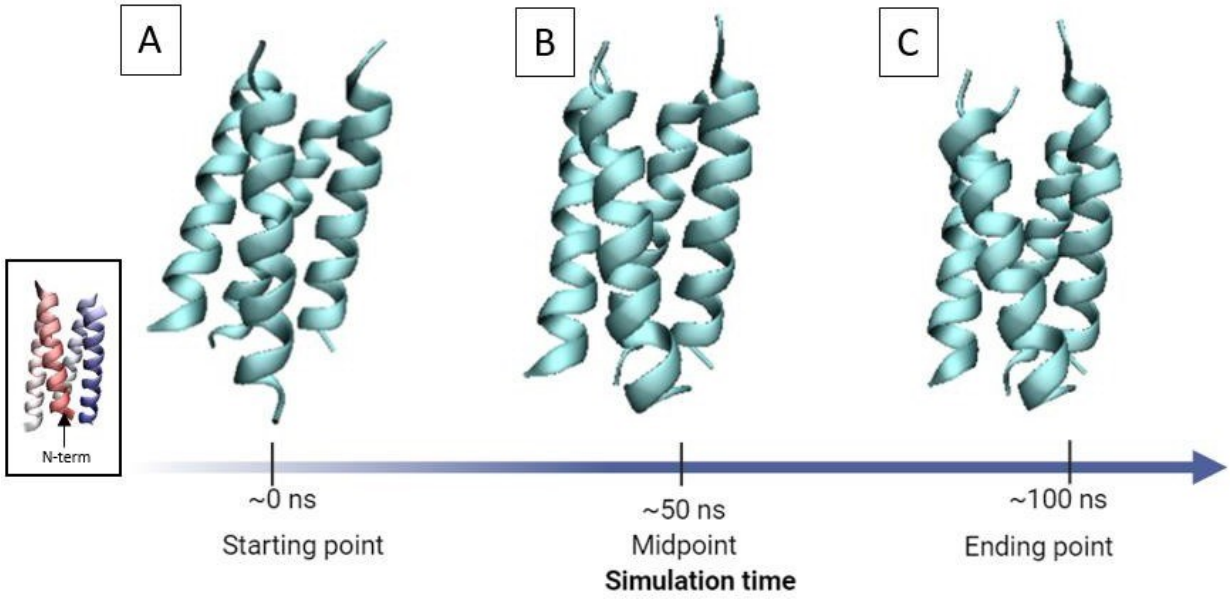


Figure 2.21. Structural arrangements of REAMP tetramer embedded in POPE bilayer. Snapshots (side views) taken from the visualisation of the trajectory of the simulation (VMD software) at different time points of the simulation (from left to right: starting point, midpoint and endpoint). The structure of REAMP did not suffer any major structural modifications.

Figure 2.22 (A-C) contains the same type of snapshots as in the figure presented above. The selected frames from the middle range of the trajectory present in panel B showed a kink in the helix containing the C-terminus. This showed that the kink is present in the structure of the helix for almost half of the duration of the simulation (50 ns). In this figure, there is also represented the solvated lipid bilayer. As can be observed in panel B, the lipid density increased around the middle section of the simulation. Still, also, there was also a higher order of freedom for the lipid molecules compared to the other time points.

It can be seen that different protein-lipid interactions can be distinguished throughout the simulation.

From the comparison of the frames in panels A, B, and C, the tetramer begins by interacting through the helical ends with the solvent molecules. As the simulation progressed, the helical ends interacted with glycerol oxygen atoms and the phosphate groups of the head groups of the POPE molecules. The number of these interactions increased and was strengthened towards the end of the simulation (panel C).

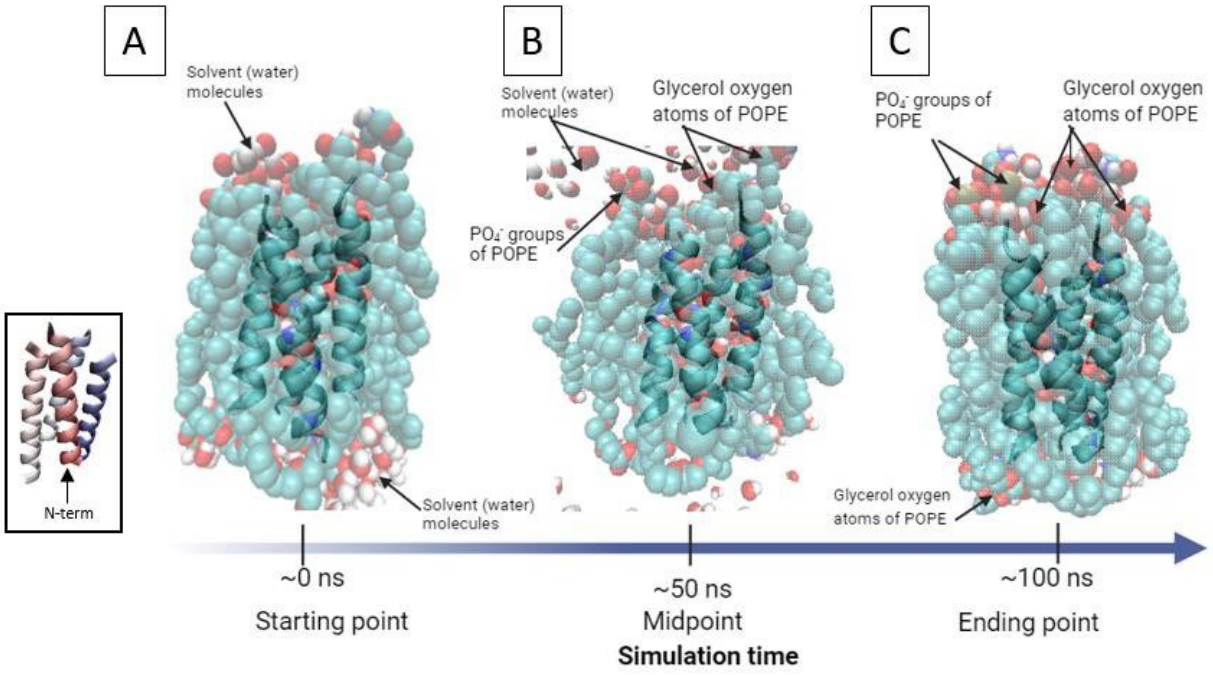


Figure 2.22. Protein-lipid interactions were observed for REAMP embedded in the POPE bilayer. Snapshots (side views) taken from the visualisation of the trajectory of the simulation at different time points (from left to right: starting point, midpoint and ending point). REAMP tetramer is illustrated as turquoise ribbons embedded in the solvated POPE bilayer, illustrated as spheres. The lipid selected lipid shell thickness is within 5 Å distance from the tetramer. Around the starting point, the limits of the helices interact with the solvent molecules (indicated by the arrows in the first snapshot). Towards the midpoint of the simulation (second snapshot), it could be observed the interaction of the helices limits with the glycerol-oxygen atoms of the lipid head groups (as indicated by the arrows). Towards the end of the simulation the structure could establish more interactions with both glycerol-oxygen atoms but also with phosphate groups (indicated by the arrows). It was also observed an increase in lipid density.

2.4.4.8. REAMP embedded in POPG bilayer

Figure 2.23 (A-B) comprises the snapshots of the REAMP tetramer embedded in the POPG bilayer from the initial stage of the simulation, along with the overall orientation of the tetramer. Panel A shows the REAMP tetramer (represented as ribbons) embedded in the solvated POPG lipid bilayer (selected area within 5 Å from the tetramer). In panel B it is represented the same snapshot, however the lipid molecules were removed for a better visual clarity. Highlighted in a circle, there is the distorted helical region of the N-terminus-containing helix.

The visual inspection of the snapshot illustrated in panel A indicated the presence of specific lipid molecules which interacted with the distorted helical region, namely POPG 125, 140 and 132. The POPG 130 and 132 contribute with the glycerol oxygen (POPG 130 and POPG 125), and the phosphate group (POPG 132) to the interactions with the serine residue present in the distorted region.

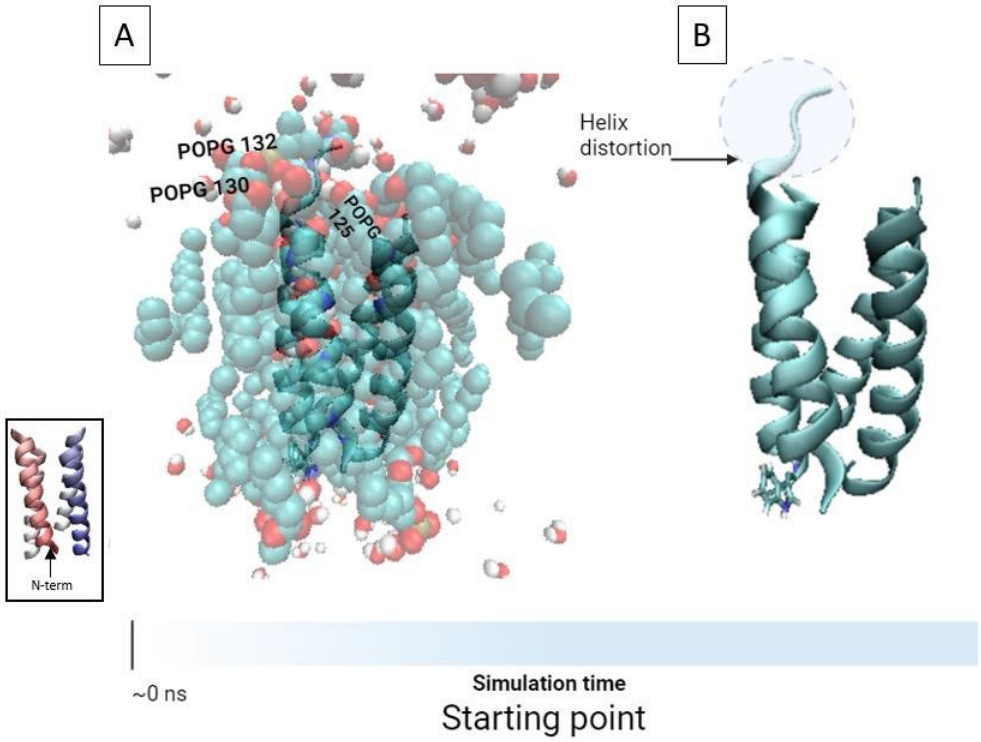


Figure 2.23. Structural modifications were observed for the REAMP tetramer embedded in the POPG bilayer. Snapshots (side views) from the visualisation of the trajectory of the simulation close to the starting point. (Left to right) The first snapshot represents the graphical representation of the REAMP tetramer as turquoise ribbons. The region which undergoes structural modification (helical distortion) is highlighted in the dashed circle. The second snapshot comprises the REAMP tetramer (illustrated as turquoise ribbons) embedded in the solvated POPG lipid environment (selected lipid shell thickness was 5 Å distance from the tetramer). There are labelled the POPG molecules observed to interact with the distorted helical region: POPG 125, POPG 130 and POPG 132.

As the time of the simulation progressed, the protein-lipids interactions modified from the presence of different lipid molecules in the proximity of the helix. Figure 2.24 (A-C) shows snapshots of the embedded REAMP in the POPG bilayer from the second half of the simulation (panels A and B- around the midpoint and panel C towards the end of the simulation). In panel A, it is shown the POPG 141 lipid molecule that interacts through the phosphate group with the distorted helical region, in addition to the POPG 132. In panel B, the highlighted region in the circle refers to an observed kink in the helix, containing the C-terminus, and the lipid molecules observed to interact with this region were POPG 107 and POPG 131. Towards the end of the simulation, as shown in panel C, the POPG 107 molecules were dislocated and

interacted with the initial distorted helical region through the glycerol-oxygen atoms in the head group.

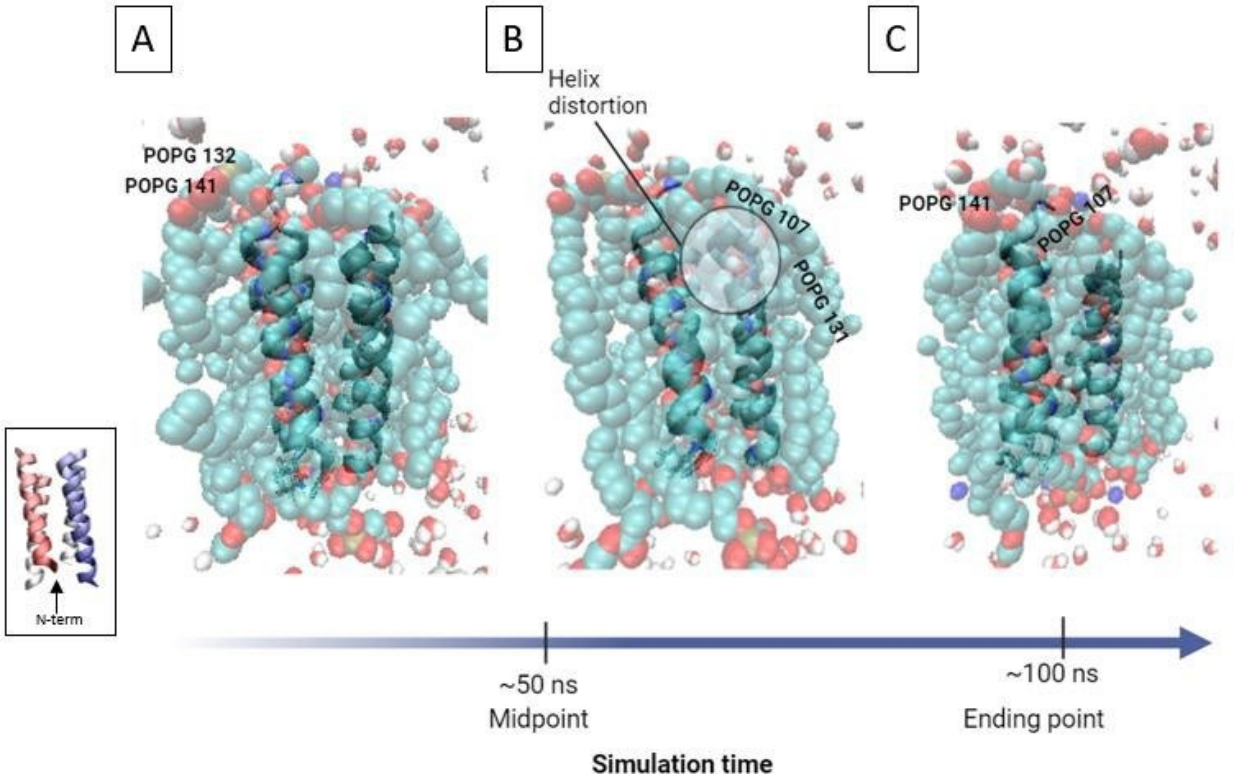


Figure 2.24. Protein-lipid interactions were observed for REAMP embedded in the POPG bilayer. Snapshots (side views) taken from the visualisation of the trajectory of the simulation at different time points (from left to right: two snapshots around the midpoint and endpoint). REAMP is illustrated as turquoise ribbons, and the surrounding environment (solvated POPG bilayer within 5 Å distance from the tetramer) is illustrated as spheres. The initial distorted helical region could be observed to interact with the POPG 132 and POPG 141 lipids. In the second snapshot it could be observed an additional helical distortion at the adjacent helix (highlighted in the circle), which interacts with the POPG 107 and POPG 131. Towards the end of the simulation, the POPG 107 shifts position towards the left, and the distorted helical region interacts with POPG 107 and POPG 141.

2.4.4.9. Poly-leucine embedded in POPC bilayer

In this case, the simulation, as shown by the RMSD plot in Figure 2.6, indicated a high instability of the tetramer in the bilayer, and the visual representation of the simulation's trajectory indicated that the tetramer was unfolding shortly after the initial stage (figure not shown).

2.4.4.10. Poly-leucine embedded in the POPE bilayer

Figure 2.25 (A-C) comprises the poly-leucine tetramer (ribbons) embedded in the solvated POPE bilayer (VDW representation) at different time points in the simulation (A- start, B-middle, C-end), along with the representation of the overall tetramer orientation highlighted in the border. It can be seen that throughout the simulation, the tilt angle of the tetramer changed, and towards the end of 100 ns (panel

C), the tetramer was close to being perpendicular to the plan of the lipids. In panel B, it can be observed that, the structure of the helices tends to suffer deformation (compression) around the middle point of the simulation. This could possibly be due to the hydrophobic mismatch between the bundle and the lipid head groups.

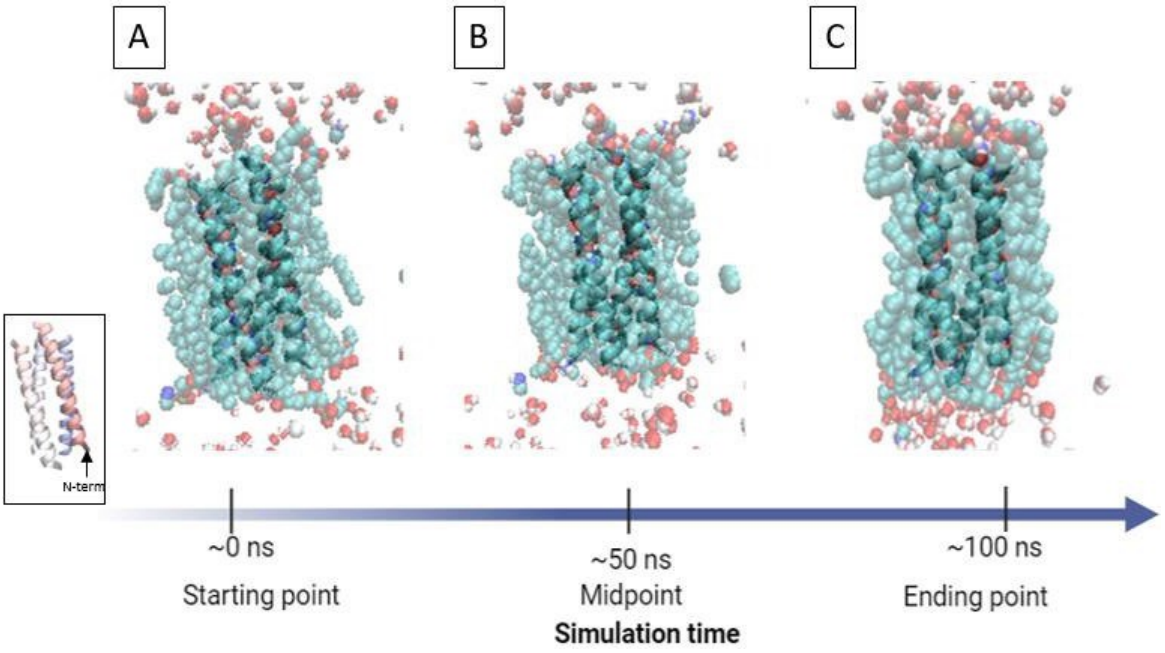
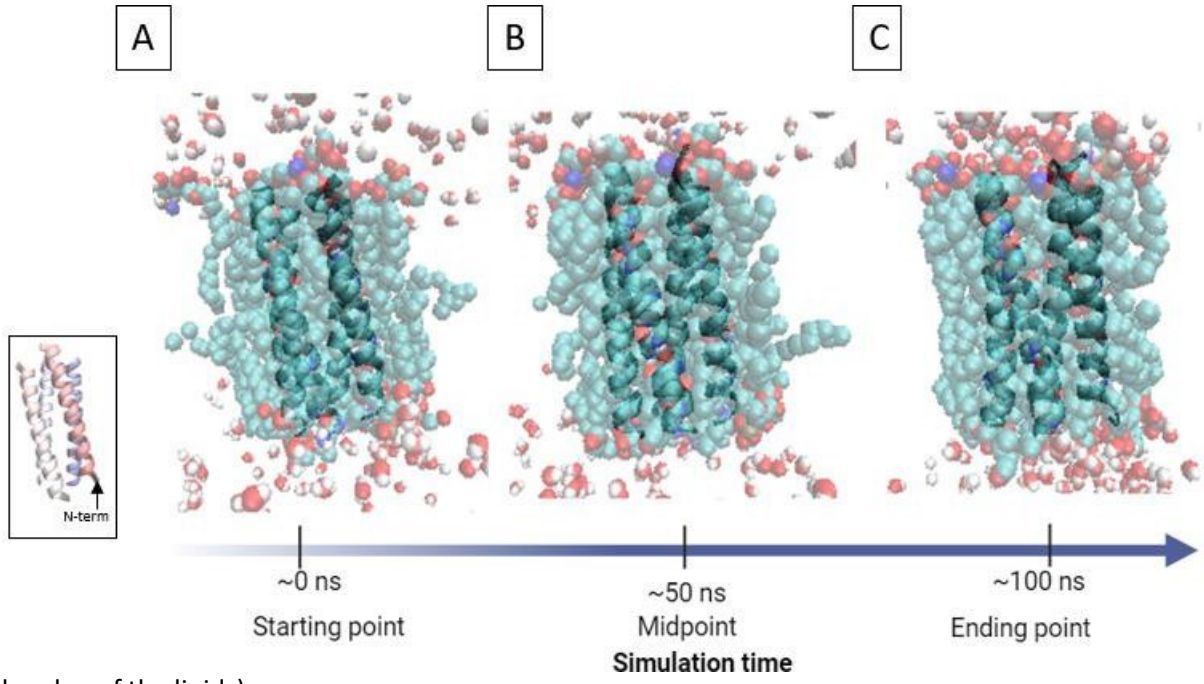


Figure 2.25. Structural arrangement and modifications of the Poly-Leu tetramer embedded in the POPE bilayer. Snapshots are taken from the visualisation of the trajectory of the simulation at different time points (from left to right: starting point, midpoint, ending point). The Poly-Leu tetramer is illustrated as turquoise ribbons embedded in the surrounding solvated POPE lipid environment (illustrated as spheres). From the starting point until the ending point, it can be observed in the shift in the tilt of the bundle (at the ending point, the helices of the tetramer are almost perpendicular to the plan of the POPE bilayer). Around the midpoint of the simulation, it could also be detected a deformation in the structure of the helices (compression).

2.4.4.11. Poly-leucine embedded in POPG bilayer

Similar to the case above, in Figure 2.26 (A-C) there is presented the poly-leucine tetramer embedded in the POPG bilayer at different time points (A- starting point, B- midpoint, C-end point). In this case it can also be seen the change in the tilt of the tetramer. The helix compression seen around the midpoint of the simulation is not as pronounced as in the case of embedding in POPE, possibly due to the reduction in the lipid density around that midpoint. Towards the end of the simulation (panel C) it could be seen an increase in the lipid density and a change in the tilt angle (tetramer almost perpendicular to



the plan of the lipids).

Figure 2.26. Structural arrangement and modifications of the Poly-Leu tetramer embedded in the POPG bilayer. Snapshots are taken from the visualisation of the trajectory of the simulation at different time points (from left to right: starting point, midpoint, ending point). The Poly-Leu tetramer is illustrated as turquoise ribbons embedded in the surrounding solvated POPG lipid environment (illustrated as spheres). From the starting point until the ending point, it can be observed the shift in the tilt of the bundle (at the ending point, the helices of the tetramer are almost perpendicular to the plan of the POPG bilayer). Around the endpoint of the simulation, it could also be detected an increase in the lipid density and the lipids grade of the order.

2.4.4.12. Summary of the MD simulations and their significance

In summary, the chosen computational approach appeared to be reliable, especially due to the consistent results obtained for the REAMP and poly-Leu tetrameric bundles. As expected from the reported experimental values, REAMP has shown a preferential tendency for the POPE bilayers. Due to the validation of expression experiments performed in *E. coli* cell membranes (having in composition an increased composition in POPE phospholipids), and the relative predicted stability of REAMP in POPE bilayers from the performed molecular dynamics simulations experiments, the sequential computational framework established for this study has proven to be an additional, reliable method. The observations for the poly-leucine tetrameric bundle further completed the results. The errors seen in the 100 ns-molecular dynamics simulations performed in the POPC bilayer have further confirmed the impossibility of self-association of the transmembrane helices, as communicated in previous studies.

The molecular dynamics simulation results observed for the CC1 tetrameric bundle indicated that the sequence would not necessarily be lipid-dependent for the helical association (compared to REAMP's preference for POPE) and the structure would be stable across POPC, POPE, and POPG bilayers, as indicated by the computed RMSD values. The even distribution of the computed stability across the tested bilayers would have been expected, due to the novelty of the sequence (which did not follow any natural existing sequence consensus, which would have determined to a certain extent the preference for the lipid bilayers depending on the natural expression host patterns).

2.4.4.12.1. Mechanistic implications in the assessment of design's stability

The CC1 bundle exhibited reduced structural deviations, as shown by the RMSD results from the trajectories, in comparison with the controls, especially in the POPE bilayer.

The qualitative comparison of the root-mean-square deviation (RMSD) values of the CC1 design to previously reported bundles indicated specific preferential protein-lipid interactions.

For example, the analysed RMSD plots of the CC1 design showed a variation in the structural deformations in the range between 0.1- 0.25 nm, with an enhanced structural stabilisation observed especially in the POPE bilayer, observed between 50- 80 ns throughout the 100 ns molecular dynamics simulation. Furthermore, after a 40 ns time point in the molecular dynamics simulation, the CC1 structure showed equal stabilisation across all tested bilayer compositions. CC1 structure exhibited RMSD values of approximately 0.15 nm in the POPC simulation and showed high stability compared to poly-leucine bundles, which exhibited distortions up to 0.8 nm in the POPC bilayer. The RMSD values of the CC1 design could be compared to the ones observed for the REAMP structure, for which preferential interactions with POPE were expected from the reported expression studies. REAMP structure showed deviations in the range between 0.1- 0.25 nm across the entire timescale of the simulation, with an enhanced stabilisation in the POPG and POPC between the 40-60 ns simulation timescale range. The results obtained for the poly-leucine bundle were expected, as the association of the helices are reported not to be favoured by the intrinsic structural properties, as observed from the high obtained RMSD values when simulated in POPC, whilst POPE and POPG could contribute to the lateral pressure forces to the stabilisation of the structure.

The stability was also expected in the all-atom simulations, as suggested by the lack of local rearrangements during the Foldit energy minimisation. These results would imply a positive impact in the case of the expression of the tetramer in organisms with membranes rich in POPE, such as *E. coli* (Rowlett *et al.*, 2017). The *E. coli* inner membrane mimetic POPE preferentially stabilised CC1, guiding production

vector choices matching this phospholipid composition dominance. The simulations in the POPC bilayers indicated that the intercalation of this type of phospholipid in a bilayer would induce certain local helical distortions in some moments. Important to be noted is that the RMSD values obtained for the stability of the backbone are indicators of the trend of the protein's behaviour throughout the trajectory, rather than specific time-points dependent events. The occurring local distortions could be explained by an interplay between the amino acid composition of the protein, combined with the packing trend and lateral pressure exhibited by the lipids. The glycine residue content could influence the distortions in CC1, these being the residues observed to suffer the most modifications in the phi/psi angles in Ramachandran plots. These residues contributed to the local flexibility of the helix in the conditions of increased lateral pressure and strong electrostatic interactions. POPC molecules have a cylindrical shape and tend to form planar bilayers, thus in the case of occurring helical distortions, these are due to the electrostatic properties of the PC head group. Although it is not a head group capable of acting as a hydrogen bond donor (Corin and Bowie, 2020), the strongly electrostatic head group, as it was previously described (Cantor, 1999), could interact with the glycine residues placed towards the polar interface of the helix and with the aromatic ring of the tyrosine residues around the same region. For the simulations in POPE and POPG there could be observed subtle compressions of the helices, as other types of local distortions. These conformational changes would be in line with the already described structural modifications of other membrane proteins in response to the curvature stress imposed by the surrounding lipid bilayers, especially by the negative curvature imposed by the POPE bilayer (Corin and Bowie, 2020). As a general observation, the lipid head groups represented major interaction sites across sequences and bilayer compositions. They were involved in interactions either through the phosphate groups contained in PG, choline groups (PC), or the oxygen atoms contained within the glycerol backbone of PE. The strong specific interactions through the oxygen atoms of the glycerol backbone of the PE head group re-emphasized its previously observed behaviour as a hydrogen bond donor (Lessen *et al.*, 2022). Local distortions in the structure of the CC1 were seen as an influence of the PG head groups, as a result of the hydrogen bonding and van der Waals interactions between the hydrogen bond donor (backbone atoms and N-terminal alanine amino group) and the acceptor, represented by lipids which contained the PG head group (Murzyn, Róg and Pasenkiewicz-Gierula, 2005).

In addition to the general influence of the lipid bilayer composition on the local structural modifications, other essential conformational changes observed, also highlighted the residue-specific molecular driving forces for this phenomenon. For example, in the case of the CC1 structure, the flexibility and positive charge of the lysine residues ensured the proper accommodation within the lipid bilayer and

complemented the membrane polarity transitions. The tyrosine residues' hydroxyl groups established hydrogen bonding with the adjacent phospholipid head groups and water molecules. The tryptophan within the REAMP structure adopted different conformations and modulated local helical distortion and π-cation interactions between the aromatic ring and the lipid head group (especially influenced by the PC). Furthermore, the exposed peptide-lipid complexes exhibit experimentally consistent enthalpic favourability, circumventing distortions seen in other non-natural analogues.

Nevertheless, in the absence of empirical stability data, computational measures alone restrict the strength of conclusions on improved stability.

2.4.4.12.2. Conformational changes could indicate a potential functionality of the CC1 design

The increased stability of the CC1 design expressed by the RMSD values is generally designated for describing the stability of the backbone between the different conformational changes. The observations of the conformational changes of the entire design detected the tendency of reorientation of the tetramer and a helical tilting, which would correspond to a pore-like opening-closing behaviour. This was observed especially in the pure POPE and the POPG bilayers, with approximately 12 Å "pore" expansion towards the intracellular side of the membrane (or the side containing the N- and the C-termini). The pore opening towards the intracellular side of the membrane is also found in naturally occurring channels, such as TRPV1, which was also reported to form homotetrameric structures around the pore (Cao *et al.*, 2013). Furthermore, it was shown by Liao and co-workers that the helices that surround the pore have aromatic residues such as tyrosine and phenylalanine, which form a hydrophobic domain, in contrast with the charged one observed in voltage-gated ion channels (Liao *et al.*, 2013). This finding further supports the potential of the CC1 structure for a channel architecture. The tilt of the helices was variable depending on the lipid bilayer. In the POPE bilayer, helices 1 and 4 displayed the highest tilt differences, whereas, in the POPG, helices 2 and 3 had the most significant changes in tilt. In the POPC bilayer it could not be detected such changes in the helical tilt in the membrane. The computational results show again the impact on the structure and function of membrane proteins of the curvature stress and lateral pressure imposed by the lipid bilayers. Their role in stabilising conformational active/inactive states has been shown previously in solid-state NMR, X-ray and cryo-EM studies (Jodaitis, van Oene and Martens, 2021). The results observed for the tilt changes of CC1, depending on the lipid environment, could be associated with the behaviour of mechanosensitive channels, which are known to open in response to membrane tension. The exact determinant factor for the activation mechanism of this type of channel is

still not completely solved, oscillating between different causal hypotheses, including hydrophobic mismatch, changes in the membrane curvature or anisotropic forces (Flegler, Rasmussen and Böttcher, 2021). The results obtained from the present experiments could be associated with the latter two hypotheses. In the case of CC1, it was also observed from the visual inspection of the simulations that the tilt changes coincided with a reduced lipid density; in this case, it could be considered the influence of a potential dipole moment between the lipid groups and the presence of the N- and the C- termini of the structure on the same side of the lipid membrane. Considering that previously, in different oligomers, it was found that the dipole moment is parallel to the direction of the insertion (Sarkar and Chattopadhyay, 2022), and it is caused primarily by the anisotropic (also known as non-random) distribution of the lipids, could be the contributing force to the conformational changes observed for CC1. It has to be re-emphasised the fact that the computed MD simulations were performed in pure POPE and POPG bilayers, and the results regarding the behaviour, tilting and "channel-like" behaviour would most probably differ, when simulations would be performed in a more realistic environment, such as bilayers comprising a mix between PE/PG. This aspect would be highly desirable for future experiments, to enable a higher accuracy of the MD results, hence a higher relevance for the potential *in-vitro* applications.

2.5. Discussion

2.5.1. Design of membrane proteins based on first principles

The great feature of protein design is that it can be bi-directional, starting from certain general structures and identifying key structural elements for determining the structure-activity relationship or, starting from function, creating new structures.

The present chapter comprised an alternative framework to the *de novo* design of a transmembrane helical bundle. The design pathway used in this project concatenated principles of the rational design of membrane proteins with the benefits of speed and accuracy of the multi-step computational-aided sequence sampling techniques. The strategy proposed in this study aims to contribute as a comprehensive method for efficiently designing transmembrane helical bundles, especially with antiparallel orientation and based on a single repetitive sequence.

As derived from the fundamental key concepts of *de novo* design, this study attempts to contribute to the expansion of the protein repertoire (Pan and Kortemme, 2021) and to emphasise the importance of reducing the possible degrees of freedom during the design process through inverse folding or a bottom-up approach (Yang *et al.*, 2021).

The membrane protein design methodology integrated both physics-based and knowledge-based techniques through a multi-step approach. A Crick parameterisation method was first used to construct an ideal 4-helix antiparallel bundle. The repetitive CC1 sequence was rationally designed and then screened for stability against this backbone template using Foldit. The resulting structures were evaluated for tilt angles in implicit membranes with Hippo before explicit solvent molecular dynamics stability simulations using GROMACS.

2.5.2. Comparison to prior work antiparallel bundles and starting considerations for the design process

The starting point of the present bottom-up approach was to design a novel *ab initio* amino acid sequence, which was able to stabilise a tetrameric bundle of antiparallel alpha-helices. Unlike the previous antiparallel dimer-of-dimers designs (Grayson, Anderson and Jlr, 2018), this work achieved a stable homogeneous helical arrangement. Though antiparallel packing arrangements have been shown for some water-soluble sequences (Smith *et al.,* 2015), membrane proteins strongly favour parallel orientations (Roberts *et al.*, 2012). By overcoming this natural bias, the current methodology expands the design boundaries. The sequence determinants differentiating these impact association preferences merit investigation. Unlike most of the current contributions to the *de novo* (computational) design, which focus on either predicting and generating new protein structures (Huang, Boyken and Baker, 2016), on the modifications of the existing naturally occurring amino acid protein sequences through mutagenesis (Cobb, Sun and Zhao, 2013), or, on the identification of sequence patterns in different protein classes and use them as design templates, this study aimed to design novel unique tectons, able to self-assemble into the desired structure and potentially into higher order structures.

The design and choice of the amino acid sequence followed a multi-step workflow. Starting from the immense number of possible amino acid sequence combinations and a large number of possible backbone adopted conformations of a single combination (DeGrado and Korendovych, 2020), a sequence based on minimal complexity, comprising a few types of amino acids was desirable and in completion, its repetitive feature in the overall structure.

The application of rational design has significantly contributed to the reduction of these possible numbers. An important aspect of the chosen design pathway is represented by the possibility of correlation of the impact of specific amino acids or amino acid motifs on the formation and stability of the desired structure.

Despite the fact that the widely used computer-based techniques, such as Rosetta, developed by David Baker and co-workers, are under constant development (Ovchinnikov *et al.*, 2018), and structural predictions are substantially more achievable nowadays, compared to previous times, the Monte Carlo sampling algorithms for conformations do not provide a simple and clear set of information about the impact or contribution to adopting a certain conformation of specific amino acids. This limitation resides in the algorithm itself, the sampling being randomised and based on selections between three to nine amino acids from the libraries of existing protein structures. Although the proven accuracy of this method in structure prediction was valued within multiple assessments (CASP) (Rohl *et al.*, 2004), and this method was used by Baker and co-workers to engineer repeating motifs to self-assemble into superhelical structures (Xu *et al.*, 2020), the algorithm is not based entirely based on *de novo* design first principles.

From a thermodynamic point of view, reducing sequence complexity would also significantly reduce the system's entropy ("chaos"), reduce the degrees of freedom, and thus, should achieve a more stable structure, taking into consideration the energy landscape theory of protein folding (Cheung, Chavez and Onuchic, 2004). In other words, the rotamer entropy would significantly influence the stability of a considered protein folding state, more specifically, acting as an opposite force to sustain a more favourable folding of the protein. In addition to the rotamer entropy, studies have also shown that rotamer strain energy also conflicts with protein folding (Penel and Doig, 2001), and could occur when sequence-predominant amino-acid side-chains are not found in their lowest energy rotamer conformations, resulting in shifts of the dihedral angles; hence, overall protein conformations. Although it has been shown that for nascent proteins, ribosomes can adjust the rotamer entropy (Streit *et al.*, 2024), this hypothesis must have been circumvented from the design step. Considering the previous statements, the computationally designed and probed protein must have satisfied from early stages a low entropy (low rotamer conformational changes, support the lowest calculated energy score and, thus, provide a computationally stable structure). If all these conditions were satisfied in the early stages of the design process and probed in the initial computational refining steps, it would have aided the study of the impact of the membrane environment on the overall packing, assembly and folding, which was one of the main objectives of the present study.

2.5.3. Evaluation of the Applied Computational Methods

Currently, the great advances in computational biology and the development of algorithms and software offer a wide variety of potential methods for designing, structure folding prediction, or

sequence-motif identification (Koehler Leman, Ulmschneider, and Gray, 2015; Chen *et al*., 2020) for assigning or repurposing the activity. Each of these methods addresses specific scientific questions.

In this study, the multi-step computational framework extended the current approaches for *de novo* design. This approach combined physics-based modelling with knowledge-based techniques. One of the strong points of the chosen framework is the directed, rational iterative sequence-structure sampling, which improved the sampling efficiency. Another strong point was represented by the enabling of sequence customisation (in Foldit) for stability compatibility against previously fixed templates (poly-alanine Crick parametrised template). Traditionally, the antiparallel homo-tetrameric helical bundle has proven to be a relevant motif for the channel architecture. In addition, the first part of the framework comprised computationally inexpensive methods, which enabled rapid sequence optimisation. The introduction of the simplified, implicit membrane models through Hippo allowed the screening for the positioning of the design within the transmembrane domain. It detected relationships between the dimensions of the membrane and the tilting. The introduction of all-atom explicit lipid bilayers through Gromacs added environmental complexity and revealed specific protein-lipid interactions. The control comparisons revealed sequence stability determinants; the stability observed for REAMP was consistent with the expected lipid preferences (POPE is the major phospholipid in the *E. coli* membrane, a host in which it was expressed (Lalaurie *et al.*, 2018)) and the instability of the poly-leucine re-emphasised the suboptimal hydrophobic forces. The limitation of the molecular dynamics simulations was the timescale of the simulation production. The dynamics perspective provided by the 100 ns simulations provided a limited capacity to conclude on the long-term stability of the design. Although longer timescales are now achievable through the use of supercomputers, access to such equipment was not possible within the present study. In summary, the multi-step approach efficiently navigated the initial sequence optimisation prior to the analysis of protein dynamics. Each method contributed with individual strengths, improving the reliability and each of the methods was proven to be robust against other experimental studies.

2.5.4. Impact of the chosen computational framework

As a case-specific impact, the use of the chosen multi-step computational framework emphasised the fundamental understanding of sequence determinants for membrane protein folding and assembly. It elucidated the contributions of the specific small amino acid motifs to the helix association and the key roles of the charged and aromatic residues in the stabilisation and positioning of transmembrane structures. Comprehensive assessments of insertion stability have challenged membrane protein designs,

a gap these cost-efficient multi-stage simulations helped overcome. It also enabled a logical and direct connection between the molecular dynamics observations and the thermodynamic principles, and the results obtained favoured this way the 'inverse folding' design principles applied. The entropic and the enthalpic thermodynamic factors were observed to dictate the occurrence of the conformational events within the protein-lipid interfaces. The comparison between the results obtained from the computational methods used in this work gives an insight into the specific environmental factors which influenced the stability. This was observed from the sequential computational experiments: from a theoretical prediction of the CC1 design stability in Rosetta Foldit to the identification of specific lipid type-dependent lipid-protein interactions, which marked localised stability disruptions in the all-atom Gromacs MD simulations. In this study, we tested and established a broadly applicable and accessible computational workflow. This current framework could be easily transferable to other membrane protein targets. By revealing key sequence drivers of buried helical packing, this investigation enables better *de novo* control in engineering custom membrane protein activities. Despite the fact that it could be seen as a less automated (or machine learning) driven computational framework, it is not computationally expensive. The inputs used in Foldit and as predecessors for the MD software-compatible files (GRO files) are easily accessible PDB structure files, which can be either designed and customised as proposed, using a specific template (as obtained from the grigoryanlab.org, as referenced in the Methods section), or could be found in the repository of the open access Protein Data Bank if naturally occurring (https://rscb.org). Not only that the workflow address the *de novo* design first principles for protein design, but it harnesses the benefits of the combined coarse-grained modelling, bioinformatics and all-atom simulations, providing valuable insights on the relationship between the sequence and the environment on the stability of the secondary structure. It is not less valid that artificial intelligence-aided computational studies for protein design indicate improved accuracy and targeted optimisation (Mallik *et al.*, 2023). Thus, the combined used scripts and variables could contribute to the composing of the training datasets of an AI algorithm for future optimisations, especially for the simulations of the CC1 design in multiple combinations of lipid bilayers to provide a broader perspective on the stability and protein-lipid interactions. Regarding the versatility of the proposed workflow, as it employs transferrable tools like Foldit and GROMACS, it could be easily adapted to other α-helical architectures like 7-TM GPCRs.

The choice of the constituent methods within the framework enabled the advancement of the state-of-the-art in *de novo* membrane protein design. The results obtained using this approach matched or exceeded the reported stability of previously designed systems. Furthermore, the demonstrated replicability facilitates any future optimisation.

2.5.5. The potential of the CC1 sequence

The results obtained from the computational sampling and simulations support the fact that the CC1 could be a stable structure if embedded in a lipid bilayer. Furthermore, the observed 'open-close '-like behaviour in POPE and POPG indicates that even in the incipient design stage, it could act as a channel/transporter.

Protein design could target a certain known function and to reverse engineer a possible structure to achieve the specific function, or new functions could be assigned: protein logic gates (Chen *et al.*, 2020), protein switches (for example, pH-dependent (Boyken *et al.*, 2019), ion-dependent (Wei *et al.*, 2020)), self-assembling biomaterials (Shen *et al.*, 2018), therapeutic properties (antimicrobial (Huong T. Kratochvil *et al.*, 2021), antiviral (Chevalier, D. Silva, *et al.*, 2017)).

There are also cases when the design end result has additional functions or behaviour compared to the initial design plan. Such an example is based on the helical peptides designed by Schafmeister and co-workers (Schafmeister *et al.*, 1997), which have proven to self-associate into a homotetrameric bundle rather than solubilise membrane proteins, as initially intended (DeGrado and Korendovych, 2020).

The design of the CC1 sequence was not based on any naturally occurring protein sequence or structure, and it is a novel sequence. Although it was not part of the design plan or rationale, the sequence has some similar key features common to some antimicrobial peptides. The 'AVGK' sequence motif is present in the CC1 sequence and is also present in a few small antimicrobial peptides found in amphibians, namely Brevinin-2DYe and Dybowskin-5. These peptides also have an overall high content of glycine and valine. This could be observed in a reported sequence alignment for the simulation-guided rational design of a small-pore forming antimicrobial peptide (sequences found in the Supplemental Information: (Chen *et al.*, 2019)). Moreover, glycine-rich peptides were discovered by a group of scientists, which have been shown to have antimicrobial properties and to be highly expressed endogenously in mud-crab haemocytes and reported to be a preferred future therapeutic direction for crawfish' disease battle in favour of traditional antibiotics (Xie *et al.*, 2020). These being said, it would be of high-interest for future studies to investigate such therapeutic properties of the CC1 sequence.

2.5.6. Future directions

2.5.6.1. Coarse-grained MD simulations

To be able to certainly associate the sequence with the structural stability and adoption of specific conformational changes in different bilayers, it is essential to observe the behaviour in MD simulations of

longer timescales. This would enable a broader perspective on the occurring structural events. Also, longer simulation timescales would be more relevant to the natural biophysical phenomena. Replicate simulations are needed to sample stochastic effects better and quantify the magnitude of trajectory uncertainties through metrics like standard deviation between runs. Furthermore, achieving computational convergence would be desirable. Similar to the statistical significance for *in vitro* experiments, it is necessary to replicate the computational simulations in order to determine numerically any calculation errors and, once determined, to evaluate how fast the iterative process achieves the final solution.

2.5.6.2. Test further sequences and engineer the pore

In order to achieve the final design goal of this study (design of membrane channels), it would be necessary to test additional sequence candidates and to engineer the interhelical interfaces of the helical bundle which are forming the pore. Additional designs would still need to satisfy the observed dominant interactions with the lipid phosphate groups over the lipid glycerol moieties, which highlight the significance of the electrostatic forces over the hydrogen bonding for the driving of helical association of the design. In order to design the selectivity filter and conducting pore, there would be some key aspects to consider. For example, special attention would be directed to the distribution and possible space arrangement of the backbone carbonyl oxygen atoms able to coordinate dehydrated ions. The charged amino acids are also determinants for the selectivity capacity of the channel: the negative charges of the sidechains of aspartate or glutamate would contribute to selectively filtering the cations. Similarly, the positively charged sidechains of arginine or lysine would select anions. In structural studies, there were observed that hydrogen bonding and cation-π interactions are occurring within the selectivity pore. To be able to interact with the solvated ions, or depending on the electronegativity of the specific ion, the hydroxyl groups of threonine, serine, or tyrosine act as hydrogen bond donors. The cation-π interactions would be satisfied by the presence of rigid aromatic rings of either tryptophan, phenylalanine, or histidine. At the current stage, cation-π pinning of lysine amino groups with aromatic phenylalanine rings likely stabilises observed helical packing motifs, with optimal sub-nm sidechain separations arising at interaction-favoured spots. Mutations regarding this aspect and their impact on the tilt in the membrane and the stability and packing of the overall structure could be sampled using the computational approach, including Foldit and Hippo, enabling a fast analysis. Considering the conductivity pore, the control of the size is essential. In order to study this, molecular dynamics simulations would be essential for detailed and dynamic size studies. The geometry and size of the pores could be assessed using the VolMap tool

available in the Visual Molecular Dynamics (VMD) software, which was used in this study. This would be a fast and easy method, that does not require a high computational cost, which could provide quick estimates of the pore size. The Gromacs computational approach used in this study to assess the behaviour of the current design within an all-atom lipid membrane model used could further be completed by umbrella sampling and the potential of mean force calculations to estimate the narrow regions within the pore and the size needed for a given molecule to pass. Hydrophobic amino acids such as alanine, leucine and valine are known to contribute to the control of a certain size cut-off for the permeability of hydrated ions. The distribution of these amino acids could be determined depending on the size of the known hydration shell (dependent also on the ionic radius: e.g. $Na^{+}_{(aq)} > K^{+}_{(aq)}$) of the specific ion/molecule desired to be conducted through the pore. Furthermore, introducing rings of alternating glutamate and lysine residues within the helical barrel interior could enable engineering pH-gated pores via salt-bridge formations. Similar to the selectivity filter, the composing positive charged amino acids would conduct anions and negatively charged amino acids would transit cations. As part of future sequence optimisation, these considerations provide size constraints, charge attractions and polar interactions, all required to mimic the behaviour of natural channels and confer desired specificity. Considering the current CC1 sequence, the propensity of the hydrophobic amino acids and the presence of rigid aromatic amino acids would not represent design points to be altered. It would be, however, of interest to increase to number of charged and polar amino acids depending on the type of ion desired to be transported.

This aspect would lead to the sampling of multiple mutations within the sequence. For this, the quantification of the mutational sensitivity using *in silico* alanine scanning would be an ideal approach, compared to randomised mutations. This would also enable the possibility of further association of the specific introduction/deletion of amino acids with the observed stability and conformational changes. This set of information would further contribute to the establishment of a library of *de novo* designed tectons, able to self-assemble and to be possibly purposed as self-assembling materials or nanoreactors.

Part of future design strategy would also be the incorporation of loop engineering. Although it was not the focus of this chapter, the designed transmembrane helices will be connected through short loops rich in glycine and serine (see Chapter 3); however, they are reported not to have any specific function. In order to contribute to the design of the function, the engineering of loops to be attached to the transmembrane helices and append functional features such as ligand binding would be an interesting study point.

2.5.6.3. Validate computational designs

Nevertheless, the variety and complexity of the computational experiments would be immense, however, it is necessary also to test the designs in a relevant system. These would be either cell-based systems to test the biocompatibility of the design and further assess the reliability of the applied design principles or simplified liposome-based systems. This would further enable the assessment of insertion and stability of the design in a lipid bilayer. The artificial probing system would also enable the testing of ion conductivity across designed membrane pores (through electrophysiology experiments). Also, would be of high demand the validation of the computationally predicted structure of *de novo* design CC1 via solution NMR or crystallisation trials. These experiments would be trivial as the next steps in the study in order to benchmark the computational predictions before pursuing functional screenings.

The bioavailability and possibility of assimilation in an *E. coli* expression system and artificial systems will be discussed in Chapter 3 and Chapter 4 of the present study.

3. – Expression and purification of *de novo* designs

3.1. Introduction

The biocompatibility of the newly designed, non-natural sequence represented the objective presented in this chapter, along with obtaining insights into the capacity of insertion of this design in the natural membrane. It is known that some naturally occurring membrane protein insertions can be frequently inefficient or toxic, leading to the production of not properly folded MPs (Le Bon *et al.*, 2018). The novel nature of *de novo* designs can enhance the toxicity perceived by the cell and implement certain defence mechanisms within the host.

Among the main design considerations for the CC1 design, it was chosen an antiparallel arrangement of the helices, which was also found in some of the naturally occurring ion channels (this represents the end goal of this design). This design was chosen due to multiple considerations, apart from the generic structural architecture desired for a membrane channel, such as the structural initial scaffold that, according to the computational studies performed on it, it satisfied the desired constrained conformational state, it was based on a minimal sequence and comprised in its sequence, amino acids placed at strategical positions, to prove their importance in structural biology (membrane-anchoring properties of lysine, size-to-close packing properties of small amino acids such as alanine, valine and glycine, satisfaction of hydrophobic-hydrophilic interactions exhibited by amino acids which can interact with an amphipathic environment, as lipid polar head groups regions, such as tyrosine, and further sterical and embedded hydrophobic interactions satisfied by phenylalanine residues). The antiparallel tetrameric arrangement of the CC1 helices is known to be a type of arrangement that provides stability and helps achieve the function (Joh *et al.*, 2017), and generally, in membrane proteins, the antiparallel orientations (66%) are preferred over the parallel ones (Eilers *et al.*, 2002). The sequence was designed following first principles for protein close packing (Grayson, Anderson and Jlr, 2018) for the specific antiparallel orientation and, furthermore, for the tetra-helical association. Also, the aim was to design a novel sequence without any natural homology, although certain domains or motifs in membrane proteins are associated with antiparallel preference and dimerization tendencies, such as the TMD4 of the EmrE (Julius *et al.*, 2017). Given that the stability of the model was already established in the molecular dynamics simulations (presented in Chapter 2), it was of interest to test the possibility of expression and insertion in *E. coli* membranes.

Escherichia coli represents the most used expression host system due to its ease of manipulation, cost-efficiency and fast cell growth (Hattab *et al.*, 2015), achieving protein expression in 2-3 days starting

from plasmid transformation. It also represented the host target for the expression of some designed soluble miniproteins or protein binders (Haven *et al.*, 2022), reflecting the versatility of this cell host. The architecture of *E. coli* membranes and cell membranes, in general, includes phospholipids as major structural constituents. They are amphipathic molecules consisting of a hydrophilic head and a hydrophobic tail. The head comprises the link between a glycerol scaffold and a polar group, which is the phosphate group. It is negatively charged and hydrophilic and can be linked to other polar molecules, which can be either choline, serine, ethanolamine, glycerol, inositol, etc. The hydrophobic tail is represented by fatty acids, which can be saturated or unsaturated, and of different lengths, which are bound at stereospecific *sn* positions of the glycerol backbone. The membrane of *E. coli* has been shown to contain 75% phosphatidylethanolamine (PE), 20% phosphatidylglycerol (PG) and 5% cardiolipin. Cardiolipin is a type of diphosphatidylglycerol lipid and has two moieties of phosphate, hence displaying two negative charges per head group and four distinct acyl chains as hydrophobic parts. The packing of the lipids is influenced by the geometry determined by the difference in size between the polar group and the fatty acids and by the type of fatty acids composition; the saturated acyl chains are linear, whereas the unsaturated ones present a kink. Different types of packing determine the curvature of a natural-mimicking membrane. Packing can also be influenced by the temperature in rapport with the phase transition temperature - at temperatures below the phase transition, an ordered packing is, whilst, at temperatures close to or above the phase transition, the orientation of the tails is more disordered. Depending on the lipid shape and packing, the interaction of the protein with the membrane can change, as shown for amphipathic peptides to change from being tilted in a membrane with negative curvature to being bound to the surface with positive curvature (Strandberg *et al.*, 2012).

If there is an expression of a toxic protein, this can be sent to inclusion bodies and there can occur certain modifications in the membrane, as a stress response, such as lipids rearrangements.

The level of toxicity of the MP expression can be diminished by either manipulating the DNA constructs (choice of promoters, tags), choice of more tolerant strains or by controlling the membrane fluidity through the control of thermodynamic and osmotic parameters of the cell growth conditions (temperature and growth medium).

In this study, there were available the pET 28b (+) plasmids (and pSTBlue-1 as alternatives) containing the widely used T7 promoter, the lac IPTG inducible operon, histidine tag and the sequences of mistic and the superfolder YFP, in which we cloned the synthetic gene which encodes the designed protein. The purpose of introducing the mistic sequence was to minimise the implication of the translocon

machinery and to help with the insertion of the design in the membrane, as it was reported to localise GPCRs in the membrane (Ananda Chowdhury, Rentian Feng, Qin Tong, Yuxun Zhang, 2014). The superfolder YFP was introduced as an internal chaperone, and the split sequence, through the bimolecular fluorescence complementarity (BiFC) to help gain information regarding the folding, fluorescence occurring only in the case of association of the split N-terminus and C-terminus parts (Ottmann *et al.*, 2009). For later applications, it was chosen the extraction of the proteins from the natural membranes using dodecyl-β-D maltoside (DDM) as a detergent.

Detergents are historically the most used agents for membrane protein extraction and stabilisation (Stetsenko and Guskov, 2017). They are amphipathic molecules, presenting a polar or charged hydrophilic head group and a hydrophobic tail. Having these structural features is capable of disrupting the membrane by insertion of the tails in the membrane and extracting at higher concentrations. They are known to form spontaneously pseudo-spherical assemblies, called micelles. The spontaneous event occurs as soon as the appropriate concentration of detergent is achieved, known as critical micelle concentration (CMC). There are different classes of detergents used for solubilising membranes, such as non-ionic (e.g. dodecyl maltoside, decyl maltoside, octyl glucoside), ionic (e.g. sodium dodecyl-sulfate), or zwitterionic (e.g. lauryl dimethylamine oxide, FOS-Choline 12) (Kalipatnapu and Chattopadhyay, 2005). Non-ionic detergents are generally mild and non-denaturing. However, they can give low solubilisation yields. When the micelle is formed, the protein is included in a detergent-protein complex. In this complex, lipids can be partially retained or sometimes completely removed, modifying this way the lateral pressure exhibited on the protein, thus destabilising the structure in some cases (Van Den Brink-Van Der Laan *et al.*, 2004).

The focus of this part of the study was determining the importance of cell strain choice and controlling the growth conditions, hence membrane fluidity on the expression of the design, and to determine whether the chosen sequence needs further improvements for successful insertion.

3.2. Materials

All chemicals were purchased from Melford (UK) and Sigma-Aldrich (UK). PEX-A128-CC1/REAMP/poly-Leu DNA plasmids were synthesised and purchased from Eurofins (UK). PET-28a (+)-REAMP/(S) sfYFP was synthesised and purchased from Twist Bioscience (U.S.). XL-2 Blue *E. coli* cells were purchased from Agilent Technologies (UK). The plasmid DNA isolation and purification kits were purchased from Thermo Fisher (UK), Sigma-Aldrich (UK), and Qiagen (UK). TAE buffer 50X was purchased from ThermoFisher, UK. DNA gel extraction and clean-up kits were purchased from Qiagen (UK), New England

Biolabs (UK) and Fisher Scientific (UK). Enzymes and enzyme buffers were purchased from New England Biolabs (UK). Antibodies were purchased from Sigma-Aldrich (UK). Polypropylene chromatography columns were purchased from Bio-Rad Laboratories (UK). Adsorbent polystyrene bio-beads SM2 were purchased from Bio-Rad Laboratories. The lipids were purchased from Avanti Polar Lipids Inc. (US). All chemicals which were purchased and not listed were of analytical grade.

3.3. Methods

3.3.1. Molecular biology - DNA templates

This section presents all molecular biology approaches tested, corresponding to the isolation from the purchased DNA plasmids of the CC1, REAMP, and poly-Leu inserts and molecular cloning of these into different customised DNA plasmids existent in the laboratory.

3.3.1.1. DNA plasmids

This subsection describes the general design of the laboratory-existent DNA plasmids, the design of the synthesised CC1, REAMP and poly-leu genes, and the molecular cloning strategies.

The DNA plasmids for bacterial expression existent in the laboratory were pET-28b (+) and pST Blue-1. They had inserted between *XhoI* and *XbaI* restriction sites (Figure 3.1. A and B) three different sequence cassettes: empty (E), split (S) sfYFP, and non-split (N) sfYFP, of 545 bp, 1295 bp, respectively 1301 bp. The diagrams of the DNA sequences are represented below in Figure 3.2 (A-C).

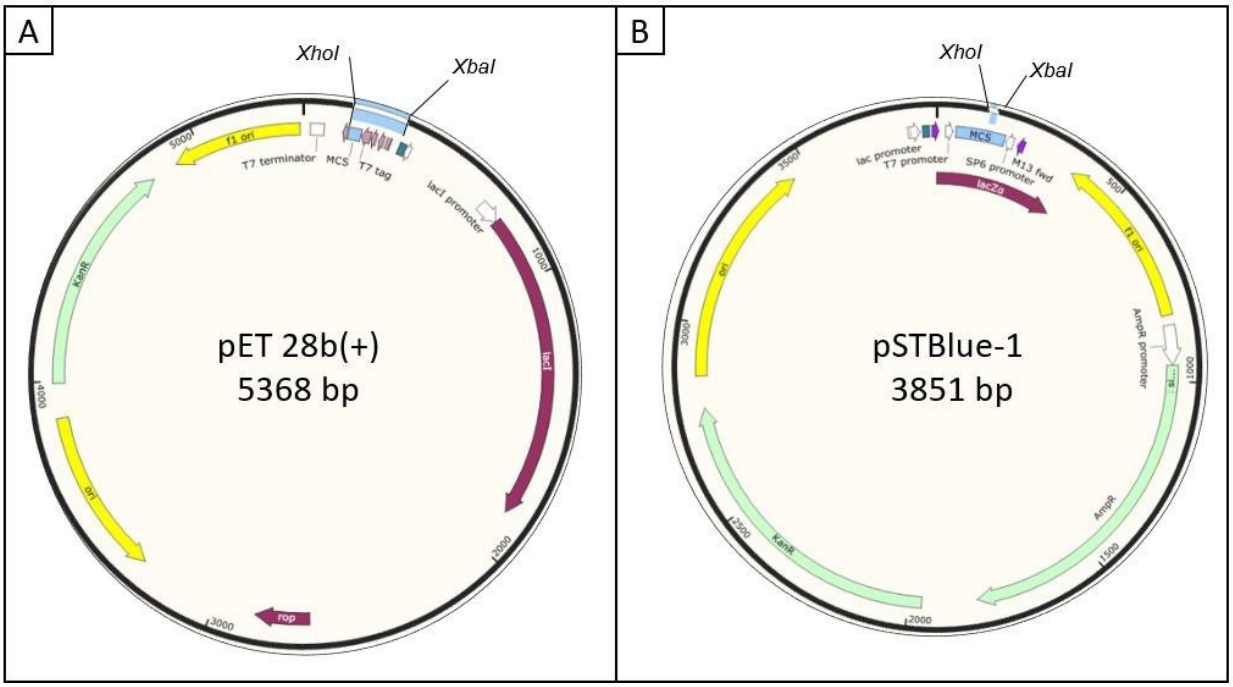


Figure 3.1. Maps of pET28b (+) and pSTBlue-1 plasmids. A. pET-28b (+) (5368 bp) plasmid map, with key features: *lac*

operator, kanamycin resistance gene, T7 promoter and terminator. B. pSTBlue-1 (3851 bp) DNA plasmid map with key features: kanamycin and ampicillin resistance genes, T7 promoter and terminator, and *lac* operator. Both plasmids have highlighted the region between *XhoI* and *XbaI* restriction sites, where DNA sequence cassettes (E), (S) sfYFP and (N) sfYFP were introduced.

The (E) cassette (Figure 3.2, A) presented as key elements: the leader sequence for pET-28 b(+), the mistic sequence, linkers, restriction sites, highlighting *NdeI* and *NheI*, factor Xa cleavage sites, and a C-terminus histidine (10 residues) tag.

The (S) sfYFP cassette (Figure 3.2, B) presented the same features as (E), with the addition of split sequences of superfolder YFP at the N- and C-termini.

The (N) sfYFP cassette (Figure 3.2, C) had the same features but with the undivided sequence of the sfYFP at the C-terminus.

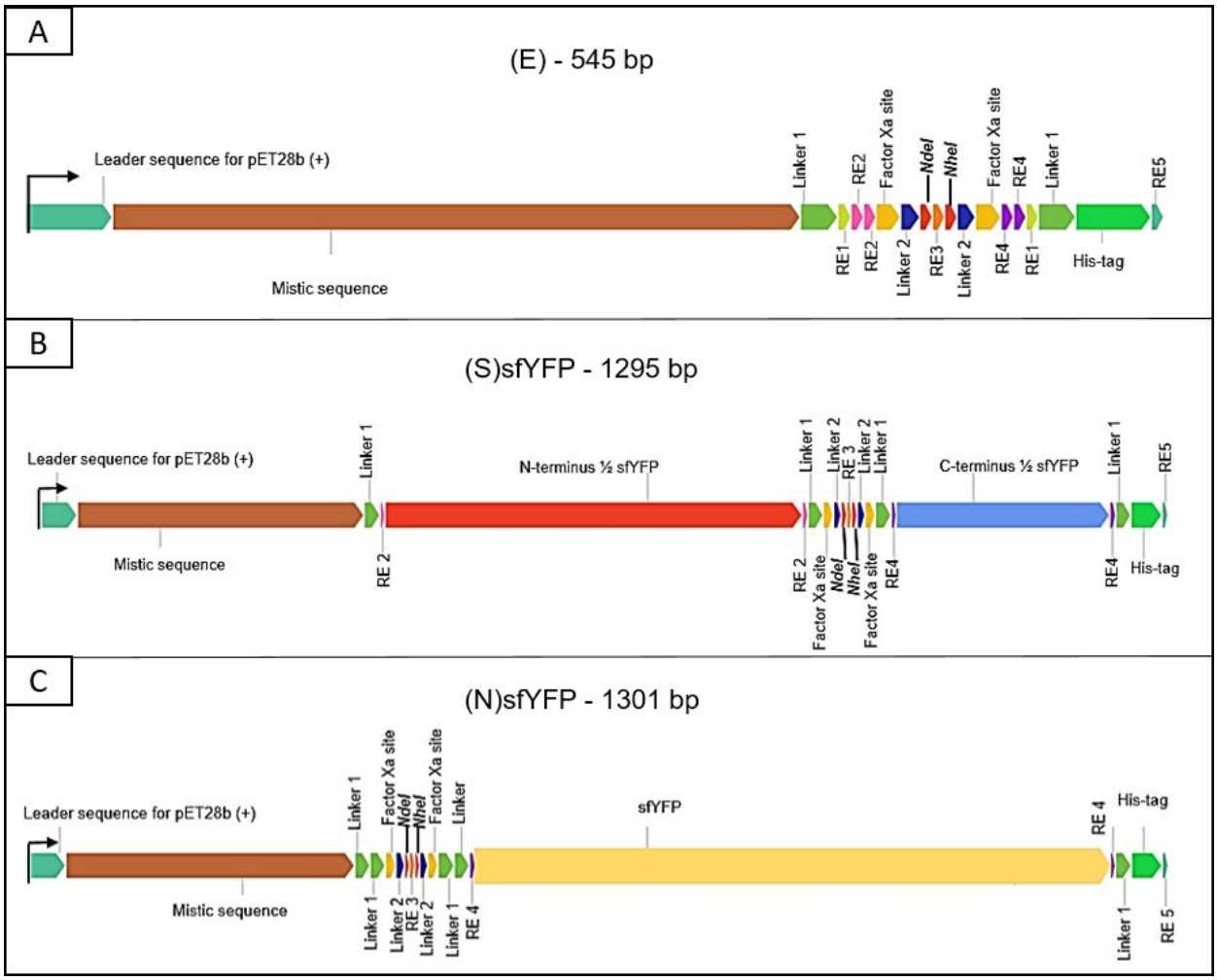


Figure 3.2. Diagrams of the customised (E), (S) and (N)-pET 28b(+) plasmids. A. Diagram of (E) DNA sequence cassette. The length of the fragment is 545 bp and comprises the leader sequence for pET 28b(+), mistic sequence, linkers (Linker 1: GGSGGS and Linker 2: GGS), restriction enzymes (in bold: *NdeI* and *NheI;* RE1- *KpnI*; RE2-*EcoRI*; RE3-*ScaI*; RE4- *HindIII*; RE5-*XhoI)*, factor Xa cleavage sites, C-terminus his-tag. The DNA segments of the (E) cassette are

common among all three types of cassettes. B. Diagram of (S) sfYFP DNA sequence cassette. The length of the fragment is 1295 bp and includes, apart from (E) components, a split-sequence of sfYFP: halves of the sfYFP sequence at the N- and C-termini. C. Diagram of (N) sf YFP customised DNA sequence. The length of the fragment is 1302 bp and contains the common features and the entire sfYFP sequence at the C-terminus.

As a general molecular cloning strategy, it was intended to ligate the CC1, REAMP and poly-Leu DNA sequences into the plasmids mentioned above between the *NdeI* and *NheI* restriction sites.

The CC1, REAMP and poly-Leu genes were synthesised and provided by Eurofins Genomics (UK) in commercial pEX-A128 DNA plasmids and inserted between *NdeI* and *NheI* restriction sites. Figure 3.3. represents a general map of commercially available pEX-A128 DNA plasmid highlighting the position of the synthesised genes, and in Table 3.1. there are represented the DNA sequences of the CC1, REAMP and poly-Leu genes and the corresponding translated (amino acids) sequences.

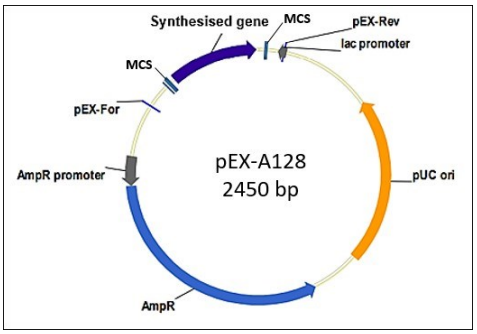


Figure 3.3. General map of pEX-A128 DNA plasmid (2450 bp) with the synthesised gene, illustrated as the dark blue arrow, inserted in the multiple cloning site between the *NdeI* and *NheI* restriction sites. (adapted from plasmid map provided by Eurofins Genomics, UK). The plasmid has an ampicillin resistance gene.

Table 3.1. DNA and amino acid sequences of CC1, REAMP and poly-Leu. The first column presents the gene names, the corresponding DNA sequences are shown in the second column, the third column mentions the corresponding size of the sequences in base pairs (bp), and the fourth column contains the translated amino acid sequences of the corresponding genes. The amino acids in the transmembrane domains of CC1 and REAMP are underlined and the non-underlined amino acids constitute the linkers placed between the helices.

Gene	DNA sequence	Size	Amino acid sequence
CC1	GGCCCGGCGGTGGGCAAAGGCGGCGTGTTTGTGGGCTTT GGCGGCGTGGGCGTGGGCTTTTATGGCGTGGGCGGCTCT GGCTCTGGCGGCGGCTCTGGCCCGGCGGTGGGCAAAGGC GGCGTGTTTGTGGGCTTTGGCGGCGTGGGCGTGGGCTTTT ATGGCGTGGGCGGCTCTGGCTCTGGCGGCGGCTCTGGCCC GGCGGTGGGCAAAGGCGGCGTGTTTGTGGGCTTTGGCGG CGTGGGCGTGGGCTTTTATGGCGTGGGCGGCTCTGGCTCT GGCGGCGGCTCTGGCCCGGCGGTGGGCAAAGGCGGCGTG TTTGTGGGCTTTGGCGGCGTGGGCGTGGGCTTTTATGGCG TGGGCGGCTCTGGCTCTGGCGGCGGCTCT	396 bp	GPAVGKGGVFVGFGGVGVGF YGVGGSGSGGGSGPAVGKGG VFVGFGGVGVGFYGVGGSGS GGGSGPAVGKGGVFVGFGGV GVGFYGVGGSGSGGGSGPAV GKGGVFVGFGGVGVGFYGVG GSGSGGGS

REAMP	GTGCTGCTGCTGCTGTCTGGTCTGGGTCTGCTGCTGCTGTC TCTGCTGGGTCTGCTGCTGCTGTCTTCTGGTGAAGAAGGTT CTTCTCTGCTGCTGCTGTCTGGTCTGGGTCTGCTGCTGCTG TCTCTGCTGGGTCTGCTGCTGTGGGGTTCTGGTAAAAAAG GTTCTTCTCTGCTGCTGCTGTCTGGTCTGGGTCTGCTGCTG CTGTCTCTGCTGGGTCTGCTGCTGCTGTCTTCTGGTGAAGA AGGTTCTCTGCTGCTGCTGTCTGGTCTGGGTCTGCTGCTGC TGTCTCTGCTGGGTCTGCTGCTGCTGTCTGGTAAA	336 bp	VLLLLSGLGLLLLSLLGLLLLSSGE EGSSLLLLSGLGLLLLSLLGLLLW GSGKKGSSLLLLSGLGLLLLSLL GLLLLSSGEEGSLLLLSGLGLLLL SLLGLLLLSGK
Poly-Leu	CTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT GCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGC TGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG CTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT GCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGC TGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG CTGCTG	333 bp	LLLLLLLLLLLLLLLLLLLLLLLLL LLLLLLLLLLLLLLLLLLLLLLLLL LLLLLLLLLLLLLLLLLLLLLLLLL LLLLLLLLL

The complete pET 28a (+)-REAMP(S) sfYFP was synthesised by Twist Biosciences (U.S.), including the backbone. The (S) sfYFP cassette containing REAMP sequence between *NdeI* and *NheI* sites was synthesised into pET28a (+) DNA plasmid between *XhoI* and *XbaI* sites. The map of the pET28a (+) DNA plasmid (5369 bp) is represented in Figure 3.4.A, with a highlighted region between *XhoI* and *XbaI*. In Figure 3.4. B is represented the diagram of the REAMP-(S)sfYFP DNA cassette, with the REAMP sequence being introduced between the *NdeI* and *NheI* restriction sites.

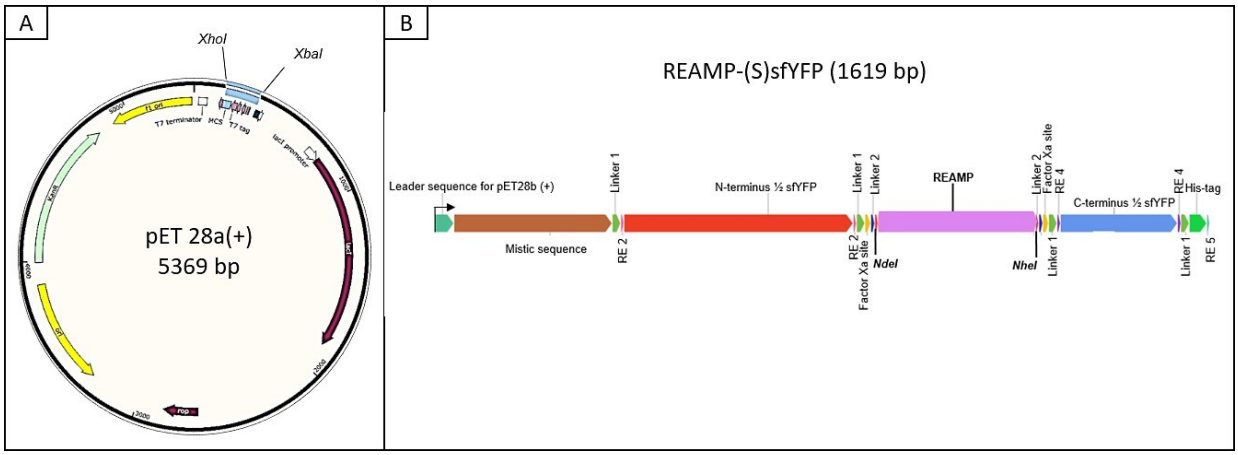


Figure 3.4. Diagram of cloning strategy of the REAMP-(S) sfYFP cassette into pET28a (+) DNA plasmid. A. Map of pET28a (+) DNA plasmid (5369 bp), with the highlighted region between *XhoI* and *XbaI*, where the REAMP-(S) sfYFP was inserted. B. Diagram of the REAMP-(S) sfYFP DNA sequence cassette (size of the fragment is 1619 bp). The DNA sequence components of the cassette were identical to the ones contained by (S) sfYFP, with the additional insertion of the REAMP sequence between the *NdeI* and *NheI* restriction sites.

3.3.2. Optimisation of the molecular cloning workflow

The molecular cloning was a challenging process for this project. For the DNA gel extraction step, increasing the DNA recovery yield represented a great challenge, and equally, improving the efficiencies of the ligation and transformation steps. Multiple optimisation methods and techniques (Figure 3.5.) were employed at each stage and are elaborated in the following subsections.

The diagram illustrated below, in Figure 3.5., includes the general molecular cloning workflow followed within this project and the troubleshooting methods. Concisely, optimisation of the DNA gel extraction method consisted of testing commercially available kits from different manufacturers and alternative traditional methods. The latter approaches involved low-temperature incubations, high-speed centrifugation steps, filtration steps through different types of glass wool, or ethanol DNA precipitation. The optimisation of the ligation step included testing two separate ligases (T7 and T4 ligases), incubation times and temperatures, and comparing the ligation efficiencies of the reactions when the plasmid DNA was pre-dephosphorylated and the reactions when the plasmid DNA was not subject to any pre-treatment. The transformation efficiency was compared between three different *E. coli* strains (XL-1 Blue, XL-2 Blue and BW25113). The methods observed to improve the yield or efficiency were taken in further experiments.

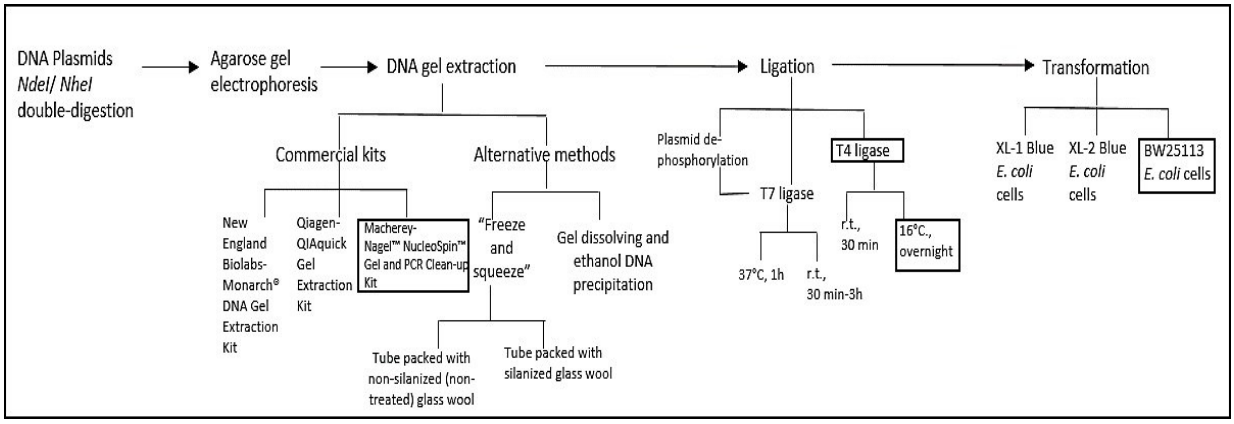


Figure 3.5. Molecular cloning workflow starting from *NdeI*/*NheI* DNA plasmids double-digestions. The branched information represents the troubleshooting of specific workflow steps. The methods which were taken for further application are highlighted in borders. DNA gel extraction different commercially available kits tested: Monarch DNA gel extraction kit (New England Biolabs, UK), QIAquick gel extraction kit (Qiagen, UK), Nucleospin gel and PCR Clean-up kit (Macherey-Nagel, Fisher Scientific, UK). The alternative methods tested: "freeze and squeeze" with the tube packed with silanised and non-silanized glass wool and gel dissolving completed by ethanol DNA purification. For the ligation: T7 and T4 DNA ligases at different incubation temperatures and times and before using T7 ligase: gel extracted plasmid DNA dephosphorylating pre-treatment, along with no treatment. Transformation: different *E. coli*

competent cell strains tested: XL-1 Blue, XL-2 Blue and BW25113.

The plasmids DNA were amplified in competent *E. coli* cells (purchased plasmids pEX-A128-CC1/REAMP/Poly-Leu and existent (E)/(S) sfYFP/ (N) sfYFP versions of pET28b (+) and pSTBlue-1), isolated and purified by maxiprep, as described in following subsections. Prior to bacterial transformation, the lyophilised pEX-A128 plasmids were rehydrated with 10 µl molecular biology-grade water.

3.3.2.1. Bacterial transformation

Stocks of 100 µl of competent *E. coli* cells (e.g. XL-1 Blue, XL-2 Blue, BW25113 strains) stored at -80°C were thawed on ice. Each stock of cells was transformed using 1 µl plasmid DNA. The cells were transformed using the heat shock method. The microcentrifuge tube containing the cells was incubated on ice for 30 minutes and then at 42°C for 45 seconds using a heat block. The tube was placed on ice for another 2 minutes. The tube was supplemented with 900 µl of sterile Luria broth (LB) (Melford, UK) and incubated at 37°C for 1 hour at 180 rpm in an orbital shaking incubator. Sterile LB agar was melted in a laboratory microwave oven until completely dissolved and incubated at room temperature for approximately 30 minutes. Under aseptic conditions, the LB agar was supplemented with the appropriate antibiotics (final concentrations: 100 µg/ml ampicillin for ampicillin-resistance genes and 50 µg/ml kanamycin for kanamycin-resistance genes were used throughout experiments) and 20 ml were transferred into sterile polycarbonate Petri dishes. The cells were harvested at 3220 x g for 5 minutes, and 950 µl of the supernatant was removed. The cells were resuspended, and the inoculum was spread on dry LB agar plates supplemented with antibiotics. The inoculated plates were incubated overnight at 37°C in a static incubator.

3.3.2.2. Plasmid isolation and purification – Maxiprep

Ten millilitres of sterile LB broth were transferred into individual sterile universal tubes and supplemented with appropriate antibiotics. One bacterial colony from each Luria broth (LB) agar plate was selected and used to inoculate the media-containing tubes. The bacteria were incubated for 6 hours at 37°C under continuous agitation at 180 rpm in an orbital shaking incubator. (There were also tested day cultures starting from freshly transformed cell stocks; the obtained results were similar). The cultures were transferred into 1L glass flasks containing 200 ml of LB medium supplemented with appropriate antibiotics (final concentrations of 100 µg/ml ampicillin for pEX-A128 plasmids and 50 µg/ml kanamycin

for pET-28b(+), pET-28a(+) and pSTBlue-1 plasmids) and incubated overnight under the same conditions. The next day, the cells were harvested at 3220 x g for 30 minutes, and the plasmid isolation and purification were performed following the manufacturer's instructions (Qiagen, UK). Steps 10 and 11 in the manufacturer's instructions were performed in a different manner. The 30 ml high-speed polycarbonate Oak Ridge tubes (Nalgene, Thermo Scientific, UK) were initially tested for the > 15,000 x g centrifugation of the precipitated DNA, however, the detection and retention of the DNA pellet were challenging. For this reason, the DNA was eluted and precipitated in a 50 ml polypropylene conical tube and centrifuged at 3220 x g for 90 minutes at 4°C for step 10 and for 20 minutes for step 11. The purified DNA was quantified using a NanoDrop spectrophotometer, and absorbance ratios measured at 260/280 nm were observed and recorded. The DNA samples were stored at -20°C for further use.

3.3.2.3. DNA plasmid restriction enzyme digestion

Purified DNA plasmids, restriction enzymes, and buffers were thawed on ice. The DNA plasmids were double-digested simultaneously with the *NdeI* and *NheI* (1 µl of each, corresponding to 20 enzymatic units) restriction enzymes in 1X NEBuffer r.2.1 (5 µl) (New England Biolabs, UK) and molecular biology-grade water for a total reaction volume of 50 µl. The reaction was incubated at 37°C for 1 hour. The restriction digestion reaction was stopped after 1 h by adding 1X purple SDS loading dye (8 µl) (New England Biolabs, UK). The digested DNA fragments were separated depending on their size by agarose gel electrophoresis.

3.3.2.4. Agarose Gel Electrophoresis

1 % agarose gel was prepared by mixing molecular biology-grade agarose powder (Melford, UK) with 1X TAE running buffer (ThermoFisher, UK). The agarose was melted by heating the mixture in a laboratory microwave oven and mixed regularly until completely dissolved. The mixture was incubated at room temperature for approximately 15 minutes until the appropriate temperature was reached (about 50°C/mixture became visibly more viscous when stirred). The gel solution was mixed with ethidium bromide (0.003%), poured into the gel mould and incubated at room temperature until it was set.

The DNA samples mixed with 1X purple loading dye (NEB) and the DNA Fast ladder (New England Biolabs, UK) were loaded into the wells of a 1% agarose-TAE gel stained with 0.003% ethidium bromide. Gel electrophoresis was performed at 100 V (50 mA, variable) for 1 hour in 1X TAE running buffer.

3.3.2.5. DNA gel extraction

Multiple approaches were employed at this step to optimise the DNA recovery yield. In this subsection, there are described the multiple tested methods (as presented in Figure 3.5.): a) use of commercially available kits, or b) alternative methods: "Freeze and squeeze", gel dissolving-ethanol DNA precipitation).

The stained DNA bands were visualised using a UV transilluminator and excised using a sterile scalpel.

a) Commercial DNA gel extraction kits:

The DNA bands were isolated and purified following the manufacturer's instructions from where the kits were purchased (New England Biolabs, UK/ Qiagen, UK/Fisher Scientific, UK). There were tested the elution suggestions in the manufacturer's instructions (using the elution buffer provided in the kit and using molecular biology-grade water).

b) Alternative methods:

- "Freeze and squeeze" gel extraction

The excised gel bands were incubated in a sterile 1.5 ml microcentrifuge tubes at -20°C for 1 hour. Frozen gel bands were placed into a 0.5 ml microcentrifuge tube bottom-pierced with a 23-gauge sterile needle and packed at the bottom with a) silanised or b) non-silanized glass wool using ethanol-wiped forceps. The 0.5 ml microcentrifuge tube was placed in a 1.5 ml tube and centrifuged at 13,000 x g for 5 minutes at room temperature. Corresponding recovered DNA was quantified using a NanoDrop spectrophotometer, and absorbance ratios at 260/260 nm were observed.

- Gel dissolving and ethanol DNA precipitation

The excised gel band was dissolved by adding the chaotropic salts-containing buffer from commercial kits (e.g. QG buffer from QIAquick DNA gel extraction kit, Qiagen, UK) to the pre-weighed gel band in a volume ratio of 3:1 (100 mg gel ~100 µl) and incubated at 50°C for 10 minutes, with regular mixing. The solution was loaded onto a DNA gel purification column from available kits (e.g. QIAquick spin column from Qiaquick DNA gel extraction kit) and centrifuged at 13,000 x g for 1 minute at room temperature. The flow-through was transferred into a sterile microcentrifuge tube and mixed with 0.7 volumes of isopropanol. The solution was centrifuged at 13,000 x g for 30 minutes. The supernatant was

discarded, and the DNA pellet was washed twice with 500 µl of 75% ethanol and centrifuged at 13,000 x g for 5 minutes. The DNA pellet was resuspended in 50 µl of molecular biology-grade water and quantified as mentioned above.

3.3.2.6. DNA Ligation

This molecular cloning step required optimisation for maximising the efficiency. The T7 and T4 ligases and different incubation times and temperatures were tested, as detailed below in this subsection.

The purified fragments extracted from the agarose gel represented the DNA template fragments, respectively the plasmid DNA (e.g. pET28b (+)/pSTBlue-1 (E)/(S) sfYFP/ (N) sfYFP) and the insert DNA (e.g. CC1/REAMP/Poly-Leu). The ligation reaction was performed in microcentrifuge tubes in different ratios of plasmid: insert DNA, using the following formula:

Required mass insert (ng) = desired insert/plasmid molar ratio x mass of plasmid (ng) x ratio of insert to plasmid length

3.3.2.6.1. Ligation reaction with T7 ligase

No plasmid treatment: The purified plasmid DNA was mixed with the purified inserts DNA in the following calculated plasmid: insert DNA ratios: 1:1, 1:3, 1:5, 1:7, 3:1, and 5:1. In the mixtures there were added 1 µl (3000 enzymatic units; volume of 1 µl was chosen for experiments convenience) of T7 ligase, 1X T7 ligase buffer (10 µl) (New England Biolabs, UK), and molecular biology-grade water until a total reaction volume of 20 µl. To optimise the ligation efficiency, the reaction was incubated at either a) 37°C for 2 hours or b) room temperature for 30 min to 3 hours.

Plasmid DNA 5'-end dephosphorylating pre-treatment: The purified plasmid DNA was mixed with 5 µl (5 enzymatic units) of shrimp alkaline phosphatase (rSAP), 1X rCutSmart Buffer (2 µl) (New England Biolabs, UK) and incubated at 37°C for 30 minutes. Due to the low yield of plasmid DNA recovery, it was used entire 20 µl volume of eluted DNA from the gel extraction steps. The enzyme was deactivated at 65°C for 5 minutes. The ligation reaction conditions mentioned above were tested.

3.3.2.6.2. Ligation reaction with T4 ligase

The ligation reaction contained purified plasmid and DNA inserts, in the ratios mentioned above and 1 µl (2000 enzymatic units, volume was chosen for experiments convenience) of T4 ligase, 1X T4 ligase buffer (2 µl) (New England Biolabs, UK), and molecular biology-grade water until a total reaction volume of 20 µl. To optimise the ligation efficiency, the reaction was incubated at either a) room temperature for

30 minutes or b) 16°C overnight.

3.3.2.7. Bacterial colony PCR

The entire volume of the ligation reaction was used to transform competent *E. coli* cells (e.g. BW25113). The transformed cells were plated on Luria broth (LB) agar plates supplemented with appropriate antibiotics (e.g. final concentration of kanamycin: 50 µg/ml). The colony templates for the PCR reactions were obtained by selecting an individual bacterial colony from the plate and resuspended by trituration in 10 µl of molecular biology-grade water. Two sets of negative controls were used: a colony of cells transformed with the undigested plasmid DNA (pET28b (+)-(S) sfYFP) and a colony of cells transformed with *NdeI*/*NheI* double-digested pET28b (+)-(S) sfYFP. The PCR reaction was performed using T7 primers.

The T7 primers used had the sequences presented in Table 3.2.

Table 3.2. Sequences of primers used for bacterial colony PCR. The primers' names and orientation are presented in the first column, and the oligonucleotide sequences (starting from 5') are included in the second column.

Primer orientation	**Oligonucleotide sequence**
Forward (F) T7 primer	5' GTG GCA GCA GCC AAC TCA GC
Reverse (R) T7 primer	5' TAA TAC GAC TCA CTA TAG GG

The bacterial colony was mixed with the PCR mixture: 1X Phusion buffer (New England Biolabs, UK), 200 µM dNTPs (New England Biolabs, UK), 0.5 µM of primers, 1 unit Phusion polymerase (New England Biolabs, UK) and incubated in a thermal cycler, following 30 cycles of denaturation (98°C, 10 s), annealing (55°C, 30 s), and extension (72°C, 60 s). Following the thermal cycles, the reaction was incubated in the thermal cycler at 72°C for 120 s and held at 4°C.

3.3.2.8. Plasmid isolation and purification – Miniprep

One colony of transformed BW25113 *E. coli* competent cells was used to inoculate 10 ml of sterile Luria broth (LB) medium supplemented with kanamycin (50 µg/ml final concentration) and incubated overnight at 37°C under constant agitation at 180 rpm in an orbital shaking incubator.

The cells from the overnight culture were harvested at 12,000 x g, at 4°C for 1 minute, and the plasmid DNA was isolated and purified according to the manufacturer's instructions (Sigma-Aldrich, UK,

or ThermoFisher, UK). The different kits were used based upon availability.

3.3.3. Optimisation routes for *de novo* protein expression, purification and characterisation

The expression, purification and characterisation processes of the *de novo* proteins associated with this project were challenging and impeded due to the nature of the proteins and the independent behaviour from the proteins available in the literature. These processes involved long optimisation routes (as presented in Figure 3.6.), and several techniques were employed to maximise the results from the cell growth rate, to the concentration of purified expressed protein. Multiple steps were optimised in the workflow, such as the cell culture (cell strains and culture conditions), protein inducer titration, cell lysis, buffer components concentrations for purification and antibody dilutions for detection. The following subsections present the detailed optimisation steps for the *de novo* protein expression in bacterial *E.coli* cells, nickel nitrilotriacetic acid immobilised metal affinity chromatography (Ni^{2+}-NTA IMAC) purification and detection by sodium dodecyl-sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting. Figure 3.6. shows a diagram comprising the optimisation steps of the workflow.

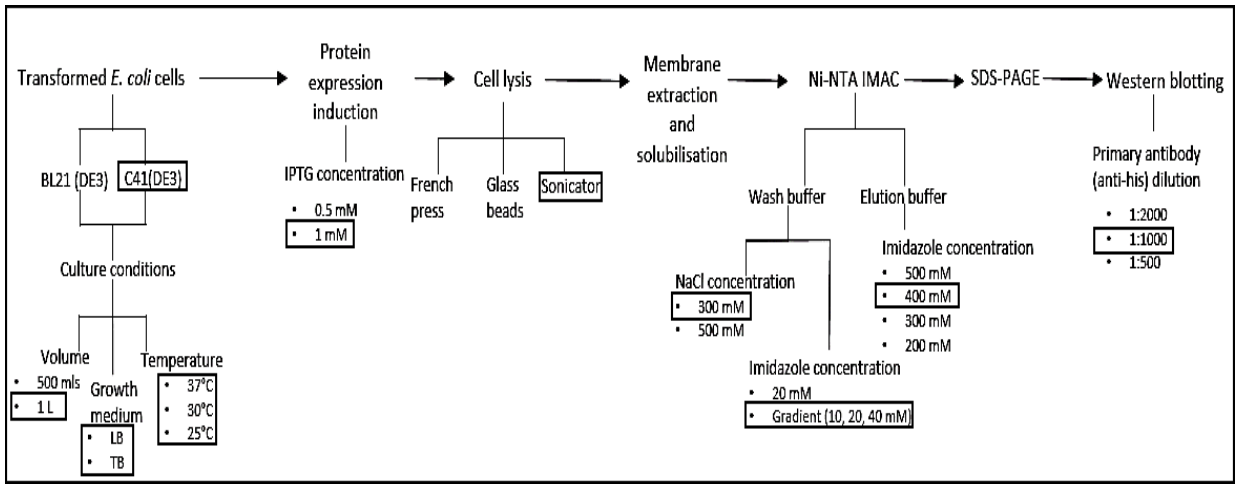


Figure 3.6. Workflow of cell culture, protein expression, purification and characterisation. The branched information represents the optimisation strategies. Highlighted in borders are the methods taken for further experiments. For cell culture: *E. coli* strains for protein expression: BL21 (DE) and C41 (DE3); culture volume (500 ml/1L), culture growth medium (LB/TB), incubation temperatures (37°C, 30°C, and 25°C), IPTG concentration (0.5 mM/1 mM); cell lysis methods (French press/glass beads/sonicator); buffer compositions for Ni^{2+}-NTA IMAC (wash buffer- 300 mM/500 mM NaCl; 20 mM or gradient of 10, 20 and 40 mM imidazole; elution buffer- 200/300/400/500 mM imidazole); western blotting: anti-his antibody dilutions (1:2000/1:1000/1:500).

3.3.3.1. Bacterial E. coli cell culture and protein expression

50 µl of competent *E. coli* cells for expression (e.g. BL21 (DE3) and C41 (DE3)) were transformed

with 1 µg of purified DNA (the ligated product) and grown overnight at 37°C on kanamycin-selective Luria broth (LB) agar plates. Colonies obtained were used as inoculants of 200 ml overnight cultures in LB medium supplemented with kanamycin (final concentration 50 µg/ml) and incubated at 37°C, 180 rpm in an orbital shaker. The next day, flasks with: a) 500 ml or b) 1L of sterile growth medium: Luria broth (LB), and Terrific broth (TB) containing antibiotics were inoculated with overnight culture (1:10 culture dilution) and grown at 37°C under constant agitation until the $O.D._{600}$ was approximately 0.6. The protein expression was induced with: a) 0.5 mM isopropylthio-β-galactoside (IPTG) or b) 1 mM IPTG, and the cultures were grown for 3 hours at different temperatures (37°C, 30°C, and 25°C) under constant agitation (180 rpm in an orbital shaking incubator).

3.3.3.2. Bacterial E. coli cell lysis, membrane extraction and solubilisation

Cells were harvested by centrifugation at 3220 x g, 4°C for 30 minutes, using a benchtop centrifuge. The cell pellets were resuspended until a homogeneous solution was observed in 10 ml of ice-cold lysis buffer (50 mM sodium phosphate buffer, 300 mM sodium chloride, 15% (v/v) glycerol, pH=7.5 and containing 1 tablet/10 ml buffer of cOmplete protease-inhibitor cocktail EDTA-free tablets (Merck, UK, or Roche, Switzerland). The different protease inhibitor mixtures were used depending on availability.

The cells were lysed using different approaches:

3.3.3.2.1. Using a French press

The resuspended cells were loaded into the cylinder body of the French press (Thermo Fisher, UK). The cylinder body and piston were pre-chilled in a laboratory refrigerator at 4°C for 2 hours. Cells were disrupted at a pressure of 1500 psi. The flow-through was collected in a polypropylene conical tube and incubated on ice. The lysate was recirculated through the press seven times.

3.3.3.2.2. Using glass beads

Glass beads (436-600 µm, acid washed, Sigma-Aldrich, UK) were mixed with the resuspended cells (0.5 g glass beads/ 10 ml cell resuspension) and vortexed for 30 minutes at 4°C. The beads were separated from the lysate by centrifugation at 3220 x g at 4°C for 5 minutes.

3.3.3.2.3. Sonication

The resuspended cells were disrupted by sonication, using a 50 Sonic Dismembrator (Fisher Scientific, UK), at 75% amplitude, in cycles of 30 seconds of pulses followed by 30 seconds of rests for a total sonication time of 10 minutes. During the sonication cycles, the sample was incubated on ice.

The cell lysate was separated from the cell debris by centrifugation at 10,000 x g. The bacterial membranes were separated from the soluble fraction by centrifuging the supernatant at 100,000 x g for

1 h, at 4°C. The bacterial membranes were resuspended in 5 ml of n-dodecyl-β-D-maltoside (DDM)-solubilisation buffer (lysis buffer supplemented with 2.5% (w/v) DDM) and homogenised using a pestle and a glass homogeniser. The homogenised solution was transferred into a sterile 100 ml glass flat-bottom conical flask and incubated at 4°C for 1 h under constant mixing, using a magnetic stirrer.
The insoluble fraction was separated from the DDM-solubilised bacterial membrane components by centrifugation at 100,000 x g for 1 h, at 4°C.

3.3.3.3. Ni^{2+}-NTA IMAC purification

The soluble fraction was mixed with 5 ml of buffer pre-equilibrated nickel-nitrilotriacetic acid (Ni^{2+}-NTA) resin (ThermoFisher, UK) and incubated overnight at 4°C under constant mixing. A filter pre-packed 10 ml polypropylene column (Bio-Rad Laboratories, UK) was washed by passing through 10 ml of 100% ethanol and 10 ml of sterile distilled water. The mixture was loaded onto the polypropylene column and eluted under gravity flow at room temperature. The flow-through was re-circulated for 4 times. The nickel resin was washed with 4 column volumes (CV) of ice-cold washing buffer.

Multiple washing buffer compositions were tested with different sodium chloride concentrations: 300 mM, or 500 mM and different imidazole concentrations: 20 mM, or gradients of 10, 20, and 40 mM. The suitable washing buffer composition was 50 mM sodium phosphate buffer, 300 mM sodium chloride, 5% v/v glycerol, 0.25% w/v DDM, imidazole 10, 20, 40mM gradient, pH= 7.5. The protein was eluted with 6 ml ice-cold elution buffer in fractions of 0.5 ml volume. The elution buffer composition was optimised by testing different imidazole concentrations: 200, 300, 400, and 500 mM imidazole. The elution buffer composition used for further experiments was: 50 mM sodium phosphate buffer, 300 mM sodium chloride, 400 mM imidazole, 0.1% DDM, pH=7.5. Eluted protein fractions were stored at 4°C until further use.

3.3.3.4. Sodium dodecyl-sulphate polyacrylamide gel electrophoresis

Separating 12% polyacrylamide gels and 4% stacking gels (1.0 mm thickness) were prepared (see Table 3.3) and hand-casted.

Table 3.3. Recipes for 12% separating and 4% stacking polyacrylamide gels. The quantities presented are sufficient for the preparation of 4 separating gels (12%) and 4 stacking gels (4%) of 1.0 mm thickness.

12% Separating gel (25 ml)		**4% Stacking gel** (12.5 ml)	
Distilled water	8.8 ml	Distilled water	6.3 ml

1.5 M Tris- HCl pH=8.8	9.4 ml	0.5 M Tris- HCl pH=6.8	4.2 ml
40 % acrylamide solution	6 ml	40 % acrylamide solution	1 ml
10% SDS	200 µl	10% SDS	125 µl
10% APS	200 µl	10% APS	1 ml
TEMED	10 µl	TEMED	5 µl

Protein samples (from the crude extract and the purified fractions) in volumes of 20 µl were mixed with 4 µl of 6X Laemmli reducing sample buffer (LSB) (Alfa Aesar, Thermo Scientific, UK) and loaded into the polyacrylamide gels, along with 5 µl of protein molecular weight marker (PageRuler Prestained Protein Ladder, cat. no. 26616, ThermoFisher, UK), as standard. Gel electrophoresis was performed at room temperature at constant voltage (100 V, variable Amp) for approximately 1.5 hours in 1X SDS running buffer (22.87 mM Tris base, 191.82 mM glycine, 0.1% SDS) until the dye reached the end of the gel.
For protein staining, after the electrophoresis run, the acrylamide gels were incubated overnight in Coomassie Blue stain (Instant Blue, Abcam, UK). The stained protein bands were identified by visual inspection.

3.3.3.5. Western blotting

SDS-PAGE was performed on the proteins of interest, as described in the previous subsection. A PVDF membrane was incubated for activation at room temperature in 100% methanol for 5 minutes under constant agitation. The pre-activated membrane, acrylamide electrophoresed gel, foam sponges, 1.5-mm filter papers, and transfer cassette were pre-soaked in 1X transfer buffer (25 mM Tris, 192 mM glycine, 20% (v/v) methanol), and the ‘’transfer sandwich’’ was assembled. The electrotransfer was performed on ice in 1X transfer buffer under constant voltage (100 V) for 1 h.

The membrane was isolated from the transfer sandwich and incubated in a clean recipient containing 20 ml of blocking solution (20 mM Tris, 150 mM sodium chloride, 0.01% (v/v) Tween-20, 5% (w/v) bovine serum albumin) for 1 hour, under constant mixing using a platform rocker. The blocking solution was removed, and the membrane was incubated under agitation for 5 minutes in TBS-T buffer (20 mM Tris, 150 mM sodium chloride, 0.01% (v/v) Tween-20). The TBS-T buffer was replaced with 10 ml of primary anti-his antibody solution: a) 1:2000, b) 1:1000, or c) 1:500 dilution in blocking solution and incubated at room temperature for 1 h under constant mixing. The primary antibody was removed, and the membrane was washed three times for 5 minutes with TBS-T buffer. The membrane was incubated

with the secondary horseradish peroxidase (HRP) conjugate antibody solution (1:5000 concentration in blocking solution) at room temperature for 1 hour under constant mixing. The membrane was washed five times for 5 minutes with TBS-T buffer. The membrane was incubated for 5 minutes at room temperature in enhanced chemiluminescence (ECL) substrate (SuperSignal West Pico PLUS Chemiluminescent substrate, ThermoFisher, UK), and the protein bands were visualised using the automated default gel imaging settings of the G:Box Chemi XX6 system (Syngene, U.K.)

3.3.3.6. Densitometry quantitative assay- Bovine serum albumin (BSA) assay

BSA protein standards were prepared in compositions of 0.125 µg, 0.25 µg, 0.75 µg, 1 µg and 1.25 µg. The protein standard samples (20 µl) and the protein of interest samples were mixed with 4 µl of 6X LSB and loaded into the wells of the polyacrylamide gel, along with the prestained protein molecular weight marker. SDS-PAGE, protein staining, and protein visualisation were performed in the same manner as previously described. The density of the stained bands reflected by the distribution of the peak areas was measured using ImageJ software. The background of the images was subtracted using the ImageJ software. The peak areas were used to generate the BSA standard curve and to calculate the concentration of the purified proteins of interest.

3.3.4. Reconstitution into proteoliposomes

3.3.4.1. Lipid film preparation

1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) and 1-Palmitoyl-2-oleoyl-sn-glycero-3-[phospho-rac-(1-glycerol)] (POPG) phospholipids (Avanti Polar Lipids Inc., U.S.) in a ratio of 1:1 were dissolved in 2 ml of chloroform, dried using a rotary evaporator under vacuum (474 mmHg), and resuspended above the phase transition temperature in 50 mM sodium phosphate buffer, pH= 7.4 to achieve a final lipid concentration of 10 mg/ml. The lipid suspension was extruded 21 times through an Avanti mini-extruder with 400 nm polycarbonate membranes to generate unilamellar liposomes.

3.3.4.2. Liposome detergent saturation

The liposomes were mixed with DDM in a 1:1 molar ratio (for POPC: POPG 10 mg/ml 1:1, it was added 4.9 mg DDM). The solution was incubated at room temperature for 3 hours under constant mixing.

The lipid-detergent mixture was bath-sonicated for 10 minutes at room temperature. Bio-beads SM-2 were washed three times with methanol, three times with distilled water, and once with liposome buffer (50 mM sodium phosphate buffer pH=7.4). The lipid-detergent mixture was mixed with the DDM-purified protein of interest in a 60:40 protein-to-lipid ratio. The protein-lipid-detergent mixture was added to the pre-equilibrated bio-beads (10:1 (w/w) wet beads: detergent) and incubated overnight above the lipid phase transition temperature under constant mixing. The proteoliposomes were recovered by centrifugation at 100,000 x g for 3 hours at 4°C and resuspended in 100 µl of liposome buffer.

3.3.4.3. Sucrose density gradient flotation

The proteoliposomes were separated from empty liposomes using sucrose density gradient centrifugation.

Sucrose solutions of different concentrations (60% and 20%) were prepared in 50 mM sodium phosphate buffer, pH=7.4. In a sterile 2 ml microcentrifuge tube, 400 µl of 60% sucrose solution was mixed with the recovered liposomes to reach a final concentration of 40% sucrose, then 900 µl of 20% sucrose solution was gently added to avoid phase mixing and supplemented with 100µl of liposome buffer (50 mM sodium phosphate buffer, pH=7.4). The tube containing the 3 phases was centrifuged at 180,000 x g for 3 hours.

3.4. Results

3.4.1. Molecular biology

The pEX-A128-CC1 and pET28b (+)-(S) sfYFP plasmids were double digested using *NdeI* and *NheI* restriction enzymes. Electrophoresis of the digestion of the pEX-A128-CC1 plasmid on a 1% agarose gel resulted in a DNA band of larger size (2450 bp) corresponding to the plasmid backbone and a smaller size band corresponding to the CC1 inserted fragment (396 bp). The electroseparation of the digested pET28 b (+)-(S) sfYFP resulted in two fragment bands: a band corresponding to the large fragment at approximately 6468 bp visible on the gel, and a short fragment (13 bp), which migrated fast through the gel, thus, not present in the gel. Agarose gel electrophoresis of the two digestions is illustrated below in Figure 3.7.

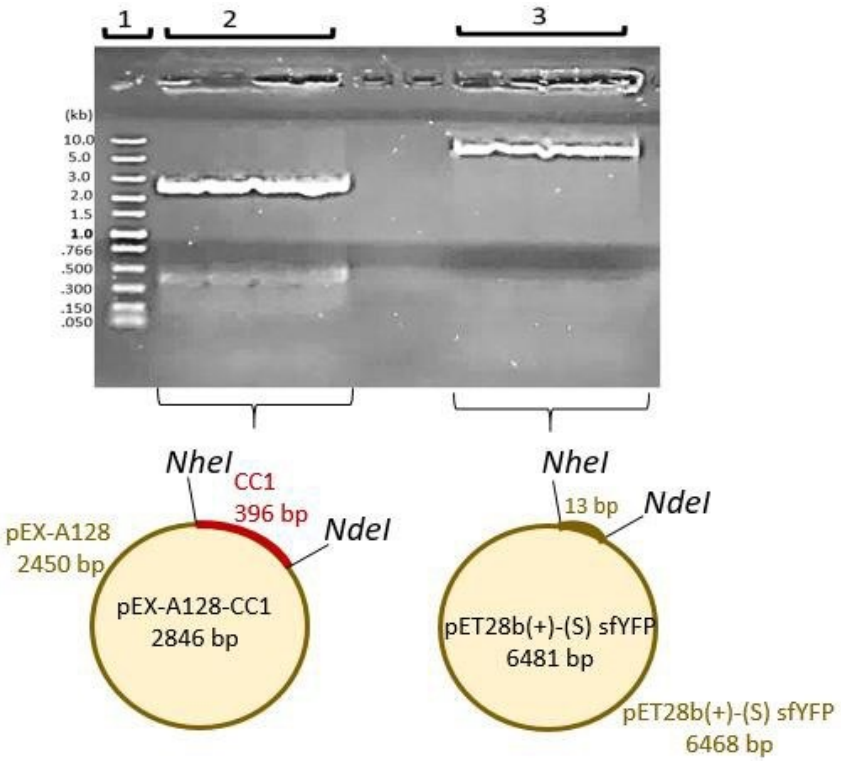


Figure 3.7. Electrophoresed agarose gel of *Nde*I-*Nhe*I- double-digested pEX-A128-CC1. 1% Agarose gel-TAE buffering system (100 V, 1h). Lane 1. Fast ladder (NEB). Lane 2. *NdeI/NheI* double-digested pEX-A128-CC1 plasmid: pEX-A128 backbone fragment (size of fragment – 2450 bp), CC1 insert fragment (size of fragment – 396 bp). Lane 3. *NdeI/NheI* double-digested pET28b (+)-(S) sfYFP plasmid fragment (size of fragment - 6468 bp).

The DNA bands corresponding to the CC1 insert fragment, and the pET28 b (+)-(S) sfYFP plasmid fragment were isolated, and the DNA was extracted from the gel and purified, as described in the Methods section. The two DNA fragments were ligated, and BW25113 *E. coli* competent cells were transformed. Ligation of the fragments was successful in the case of ratio plasmid: insert of 1:5.

Colony PCR was performed using T7 promoter primers to further confirm the ligation of the CC1 insert into the pET28b (+)-(S) sfYFP plasmid. Agarose gel electrophoresis of the colony PCR products illustrated in Figure 3.8 showed a size increase in the PCR product of the DNA cassette from the ligation reaction compared to the PCR reaction for the unligated plasmid. The bands present on the gel derived from the PCR reaction highlight the amplification of the plasmid region between the T7 promoter and terminator, containing the (S) sfYFP cassette, approximately 1412 bp (lane 3). The difference in size suggested the potential incorporation of the CC1 insert in the (S) sfYFP cassette, shown by the size of the band of approximately 1808 bp (lane 2). A control PCR reaction (PCR mix and molecular biology grade water as template) was performed to ensure no impurities were present in the PCR mix.

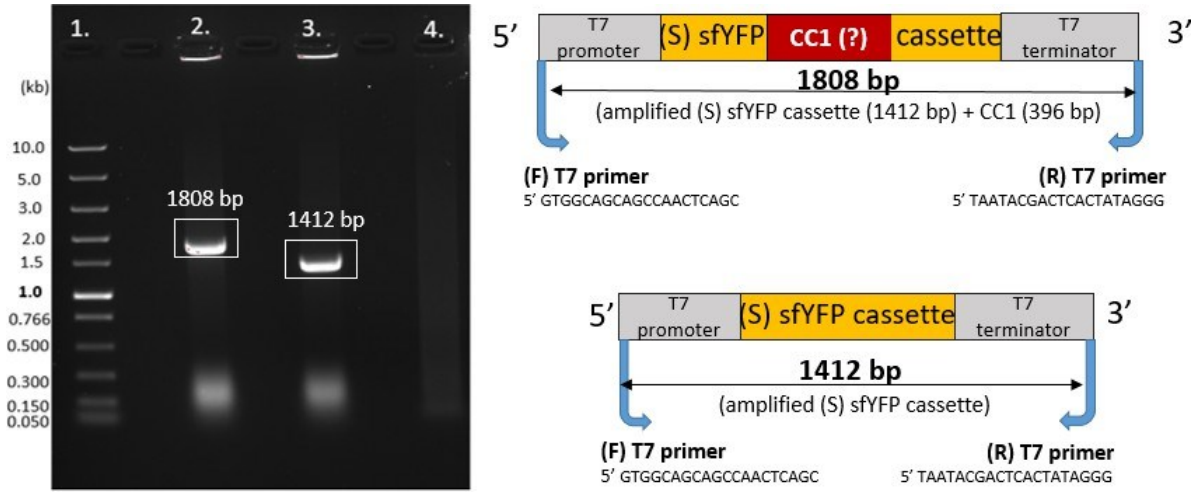


Figure 3.8. Electrophoresed agarose gel of colony PCR reactions. (Left) 1% Agarose gel-TAE buffering system (100 V, 1 h). Lane 1. Fast ladder DNA marker. Lane 2. The colony PCR reaction of cells transformed with the ligated product CC1-(S) sfYFP (1808 bp). Lane 3. The colony PCR reaction of cells transformed with undigested (S) sfYFP (1412 bp). Lane 4. Control PCR reaction with molecular biology-grade water. (Right) Schematic representation of the amplified regions situated between the recognition and binding sites of the T7 primers.

The ligation product was further analysed by agarose gel electrophoresis to confirm the association between the band size and the insertion of the CC1 sequence in the plasmid by comparing it with different restriction digests. In the lane containing the double-digested ligated product, as shown in Figure 3.9, it could be seen a single faint band at the corresponding fragment size for CC1 (396 bp).

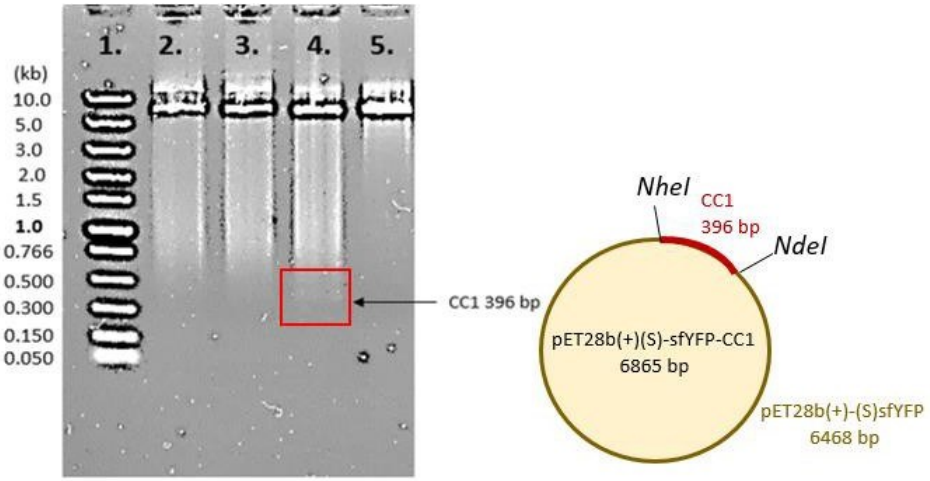


Figure 3.9. Electrophoresed agarose gel of digested pET28b (+)-CC1-(S) sfYFP plasmid. (Left) 1% Agarose gel-TAE Buffering system (100 V, 1 h). Lane 1. Fast Ladder. Lane 2. *NdeI*-digested pET28b (+)-CC1-(S) sfYFP plasmid. Lane 3. *NheI*-digested pET28b (+)-CC1-(S) sfYFP plasmid. Lane 4. *NdeI/NheI* double-digested pET28b (+)-CC1-(S) sfYFP plasmid, presenting two bands: high molecular weight corresponding to the plasmid backbone of approximately 6468 bp; the lower molecular weight band corresponding to the size of the CC1 insert (396 bp), as indicated by the

arrow. Lane 5. Undigested pET28b (+)-CC1-(S) sfYFP plasmid. (Right) Schematic representation of the CC1 gene insertion.

The ligated pET28b (+)-CC1-(S) sfYFP plasmid and pET28a (+)-REAMP-(S) sfYFP plasmid were purified by miniprep, as described. NanoDrop quantification of the purified plasmids indicated yields of 636.3 ng/µL and 389.9 ng/µL, respectively. The absorption ratios measured at 260/280 nm were 1.88 and 1.86, respectively, indicating pure DNA. The purified plasmids were subsequently used to transform *E. coli* cells for protein expression.

3.4.2. Protein expression, purification, quantification and reconstitution into proteoliposomes

3.4.2.1. Protein expression in E. coli cells

C41 (DE3) *E. coli* competent cells were transformed with purified plasmids DNA. Cultures were seeded and grown under different conditions, and protein expression was induced with IPTG, as specified in the Methods section. Cell growth was monitored pre- and post-IPTG induction (N=3) to observe any effect of the protein expression on cell growth rate (i.e. possible toxicity of the protein on the cells) and, if adverse effects were observed, to assess whether changing the growth conditions could counteract them. Based on the different culture conditions, the differences in the cell culture optical densities of cells transformed with pET28b (+)-CC1-(S) sfYFP can be seen in Figure 3.10. The optical density measurements for the cultures of cells expressing CC1-(S) sfYFP showed that cells grown in the TB medium and incubated at 30°C exhibited the most extended exponential growth and highest absorbance values. Lower and comparable values of absorbance and a similar growth trend had the cell cultures grown in LB, incubated at 30°C and in TB, and incubated at 25°C. Cultures grown in LB and incubated at 25°C had a slow growth rate compared to 30°C and 37°C, as shown by the absorbance values. Cultures grown in either media (LB or TB) and incubated at 37°C showed contrasting results compared to the other conditions. The absorbance measurements indicated that the cells grown in Luria broth (LB) at 37°C manifested a drastic decline in terms of cell growth, approximately immediately after IPTG induction. However, when grown in TB at the same temperature, the cells continued to grow for 160 minutes post-IPTG induction and then the observed growth pattern suffered a change and the measured value optical density decreased. The growth curve has shown for cells grown in either LB or TB medium, at 30°C, that the cells had an extended growth phase until the harvesting time point. Cell growth in LB or TB at 25°C also indicated an extended

exponential phase (about 120 minutes post-IPTG induction, followed by a decline in the value of O.D.), with longer generation times for cells grown in TB until the harvesting point. As a general observation, TB as a growth medium prolonged cell growth. The temperature of 30°C exhibited a positive effect on cell viability in both growth media. In Figure 3.10 panels A-C, it is presented the entire growth trend, from the culture seeding (t_0) until the harvesting time (t_{final}). In Figure 3.10 D there are selected and highlighted the same, normalised results obtained from the growth curves presented in the previous panels but only selected in the growth range between the IPTG induction until the harvesting time points, to enable a clearer image of the influence of the growth conditions on the cell growth duration, in the context of an ongoing, potentially considered toxic, membrane protein expression.

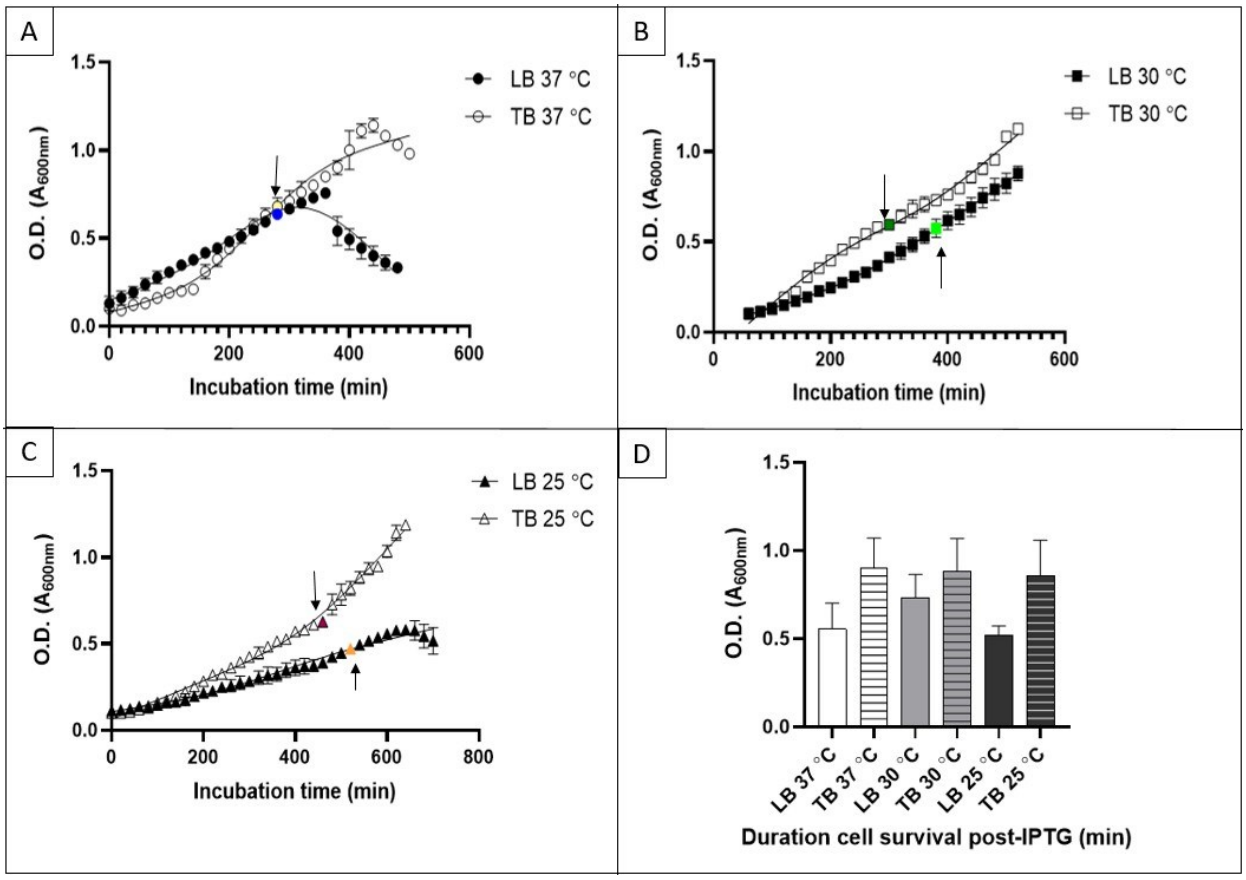


Figure 3.10. (A-C) Graphs presenting growth curves of C41 (DE3) *E. coli* cells transformed with pET28b (+)-CC1-(S) sfYFP grown in different conditions (t_0= culture seeding; t_{final}= harvesting, arrows indicate the start of IPTG induction). The measurements were performed in triplicates for 3 independent biological replicates (N=3), analysis performed using GraphPad Prism 8 software. Plotted values represent the mean absorbance values, and the error bars represent the standard errors (SEM). (X-axis: incubation time expressed in minutes; Y-axis: optical density (O.D.) expressed in measurements of absorbance at 600 nm (A_{600nm}). Allocated symbols A. **(filled circle) LB, 37°C** (SEM=0.02173;) **(unfilled circle) TB, 37°C** SEM=0.40041) B. **(filled triangle) LB, 30°C** (SEM= 0.0306); **(unfilled triangle) TB, 30°C** (SEM= 0.03731); C. **(filled square) LB, 25°C** (SEM= 0.01506); **(unfilled square) TB, 25°C** (SEM= 0.03195). (D) Duration of cell survivals after initiating the IPTG induction, depending on culture conditions. Allocated bars: **(white) LB, 37°C**; **(white/horizontal lines) TB, 37°C; (grey) LB, 30°C; (grey/horizontal lines) TB, 30°C; (black) LB, 25°C;**

(black/horizontal lines) TB, 25°C.

Similarly to Figure 3.10, the growth curves of cells transformed with pET28a (+)-REAMP-(S) sfYFP are illustrated in Figure 3.11. The measurements of the optical densities of the cultures of cells with the pET28a (+)- REAMP-(S) sfYFP plasmid incorporated also reflected an effect of the conditions on cell survival. Overall, the cell cultures' growth rates increased compared to the cells expressing CC1-(S) sfYFP. The cells cultured at 37°C in LB medium stopped to grow approximately 60 minutes post-IPTG induction. At the same temperature but grown in a TB medium, the cells exhibited a higher tolerance, prolonging the stationary growth phase until approximately 160 minutes post-IPTG. For the rest of the conditions (LB, TB, at 30°C and 25°C), the cells' linear growth phase extended until the harvesting point. In addition, cells have shown a higher tolerance and longer growth time during the protein expression when cultured in TB medium. Similarly, lower incubation temperatures impacted positively the cell growth during the protein expression, and 30°C has shown to be optimum in either growth media.

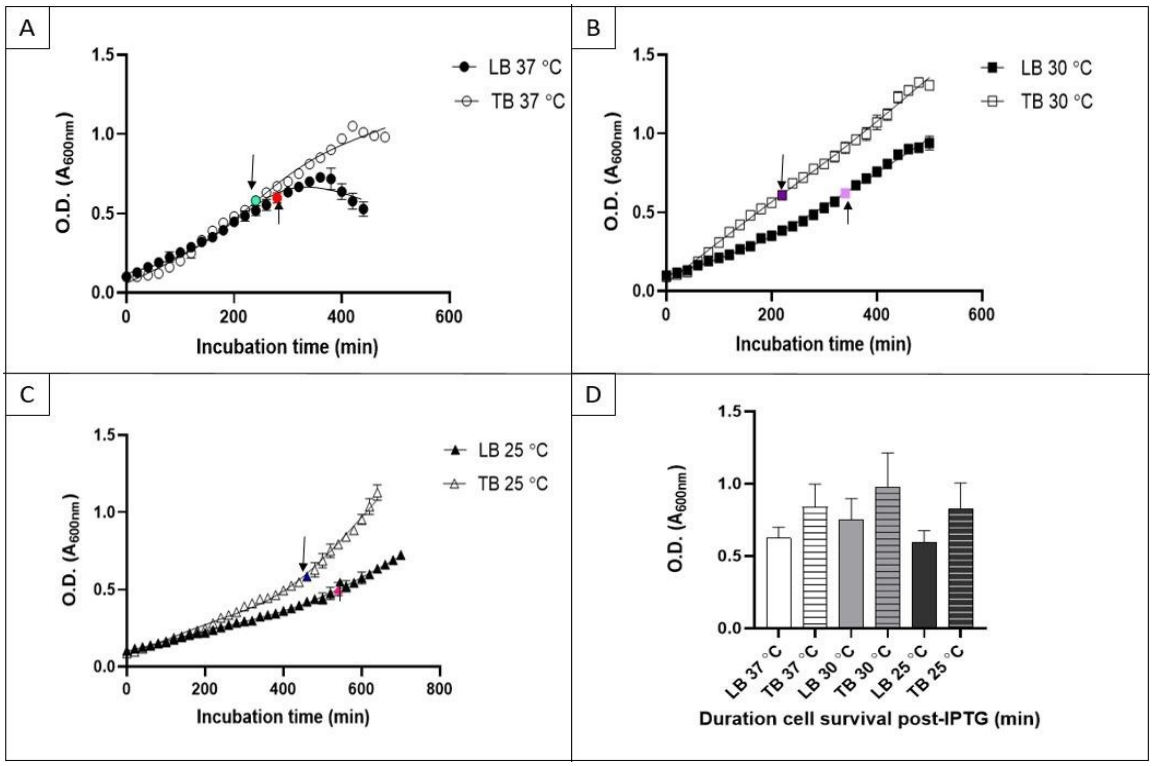


Figure 3.11. (A-C) Graph presenting growth curves of C41 (DE3) *E. coli* cells transformed with pET28a (+)-REAMP-(S) sfYFP grown in different conditions (t_0= culture seeding; t_{final}= harvesting, arrows indicate the start of IPTG induction). The measurements were performed in triplicates for 3 independent biological replicates (N=3), analysis performed using GraphPad Prism 8 software. Plotted values represent the mean absorbance values, and the error bars represent the standard errors (SEM). (X-axis: incubation time expressed in minutes; Y-axis: optical density (O.D.) expressed in measurements of absorbance at 600 nm (A_{600nm}). Allocated symbols: **(filled circle) LB, 37°C** (SEM= 0.02414); **(unfilled circle) TB, 37°C** (SEM= 0.03691); **(filled triangle) LB, 30°C** (SEM= 0.03056); **(unfilled triangle) TB, 30°C** (SEM= 0.04396); **(filled square) LB, 25°C** (SEM= 0.01699); **(unfilled square) TB, 25°C** (SEM= 0.02889). (D)

Duration of cell survivals after initiating the IPTG induction, depending on culture conditions. Allocated bars: **(white) LB, 37°C**; **(white/horizontal lines) TB, 37°C; (grey) LB, 30°C; (grey/horizontal lines) TB, 30°C; (black) LB, 25°C; (black/horizontal lines) TB, 25°C**.

3.4.2.2. Protein Ni^{2+}-NTA IMAC purification

The expressed proteins were isolated from the bacterial membranes, solubilised using DDM, and purified by IMAC chromatography, as described in the Methods section. Below are listed the stained acrylamide gels containing the crude extract samples and samples from the washing steps on the left side and the elution samples from the Ni^{2+}-NTA chromatography on the right side, obtained from the cultures grown in different media at different temperatures. Figures 3.12 and 3.13 correspond to the crude extracts and purified fractions containing CC1-(S) sfYFP. In Figure 3.12, panels A-C represent the gels corresponding to the samples obtained from cells expressing CC1-(S) sfYFP cultured in LB medium. Panel A presents gels corresponding to samples obtained from cells grown in LB at 25°C, panel B- samples from cells grown at 30°C and panel C- cells grown at 37°C. It can be observed in all gels containing the crude lysate that the protein is expressed (protein band around 56.6 kDa) when compared with the lane containing uninduced cells; expression was also confirmed by mass spectrometry (shown in Figure 3.20 A). It can be observed in the representative gels images, protein bands of high intensity at the corresponding size in lane 4, corresponding to the supernatant. It has to be noted that, during the membrane preparation, the supernatant obtained after the 100,000 x g centrifugation step was not clear, possibly indicating small-size remainder cell debris in suspension. This aspect could be associated with the lysis method; in this case, it was sonication. When the French press lysis method was tested, the obtained supernatant was visibly clear, the size of the cell debris pellet was smaller, and the bands observed in this fraction were of considerably lower intensity. Another indicator that the cell growth conditions impacted the expression is represented by the protein bands that appeared in the crude membrane fractions (lane 5). A good separation was observed in panel A of bands of high intensity, suggesting the presence of high amounts of the protein of interest. This hypothesis was confirmed by the yield results for pure protein. Although intense protein bands were observed for the crude membrane in panel C, the purified fractions contained low protein amounts. Presumably, this was because, at this incubation temperature, most of the over-expressed protein led to the potential aggregation or formation of insoluble inclusion bodies. This was observed after analysing by SDS-PAGE the pellet obtained from the final centrifugation step (post-DDM solubilisation). This pellet was not observed in the rest of the conditions. There were observed protein bands in the Ni^{2+}-NTA resin (lane 6) and low-intensity bands in the flow-through (excepting panel C) in lane 7, suggesting successful binding and no cleavage of the his-tag during the preparation. For panel

C, high-intensity protein bands in lane 7 can be explained by the possible aggregation. If the results obtained in panels A and C were consistent with the hypotheses formulated from the cell growth curves, they were not consistent in the case of LB, 30°C. The gels obtained from these conditions indicated overall low expressed protein recovery in the intended fractions. Interestingly, in all cases, a high amount of protein of interest was co-eluted in the washing step with 10 mM imidazole and not with 20 mM or 40 mM.

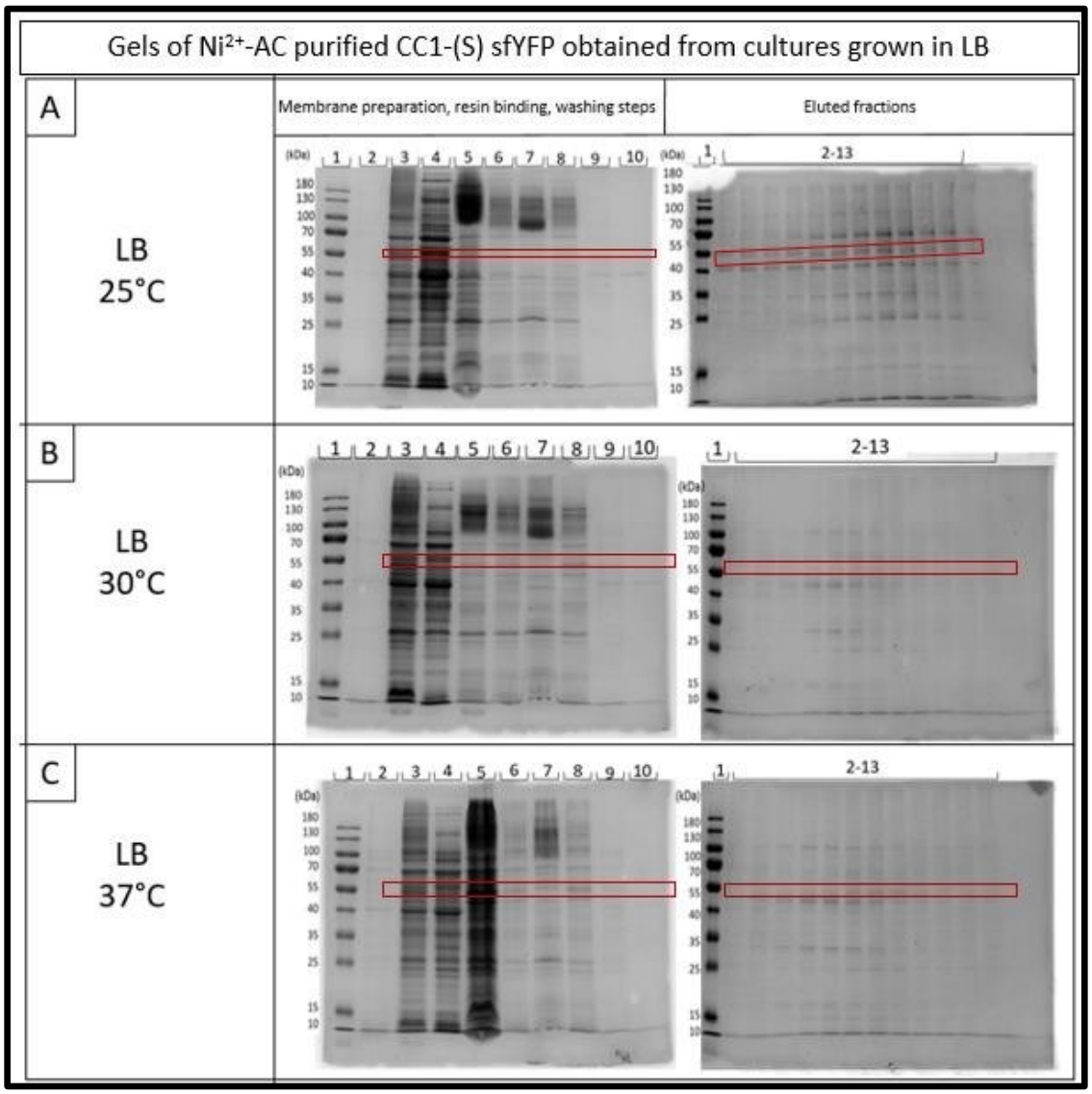


Figure 3.12. (A-C) Electrophoresed polyacrylamide gels corresponding to CC1-(S) sfYFP protein samples obtained from cells grown in LB medium. Culture incubation temperatures: 25°C (panel A); 30°C (panel B); 37°C (panel C). Highlighted in red borders are the protein bands at the corresponding theoretical molecular weight (approximately 56.6 kDa). Each set of gels was loaded as follows: (Left) Lane 1. Prestained protein molecular weight marker; Lane 2. Uninduced whole cells; Lane 3. Lysed induced cells; Lane 4. Supernatant; Lane 5. Membrane; Lane 6. Ni^{2+}-NTA resin

incubated with eluted DDM-purified CC1-(S) sfYFP; Lane 7. Flow-through; Lane 8. Wash with 10 mM imidazole; Lane 9. Wash 20 mM with imidazole; Lane 10. Wash with 40 mM imidazole. (Right) Lane 1. Prestained protein molecular weight marker; Lane 2-15: Eluted fractions.

In Figure 3.13, panels A-C are represented the gels corresponding to the samples obtained from cells expressing CC1-(S)sfYFP cultured in TB medium. Panels A-C illustrate gels with fractions obtained from lysates of cells incubated at different temperature post-IPTG induction: panel A- cells incubated at 25°C, panel B- 30°C, and panel C- 37°C. Again, the presence of a band at the expected molecular weight corresponding to the protein of interest in the induced cell lysate fraction can be observed when compared with the uninduced cells.

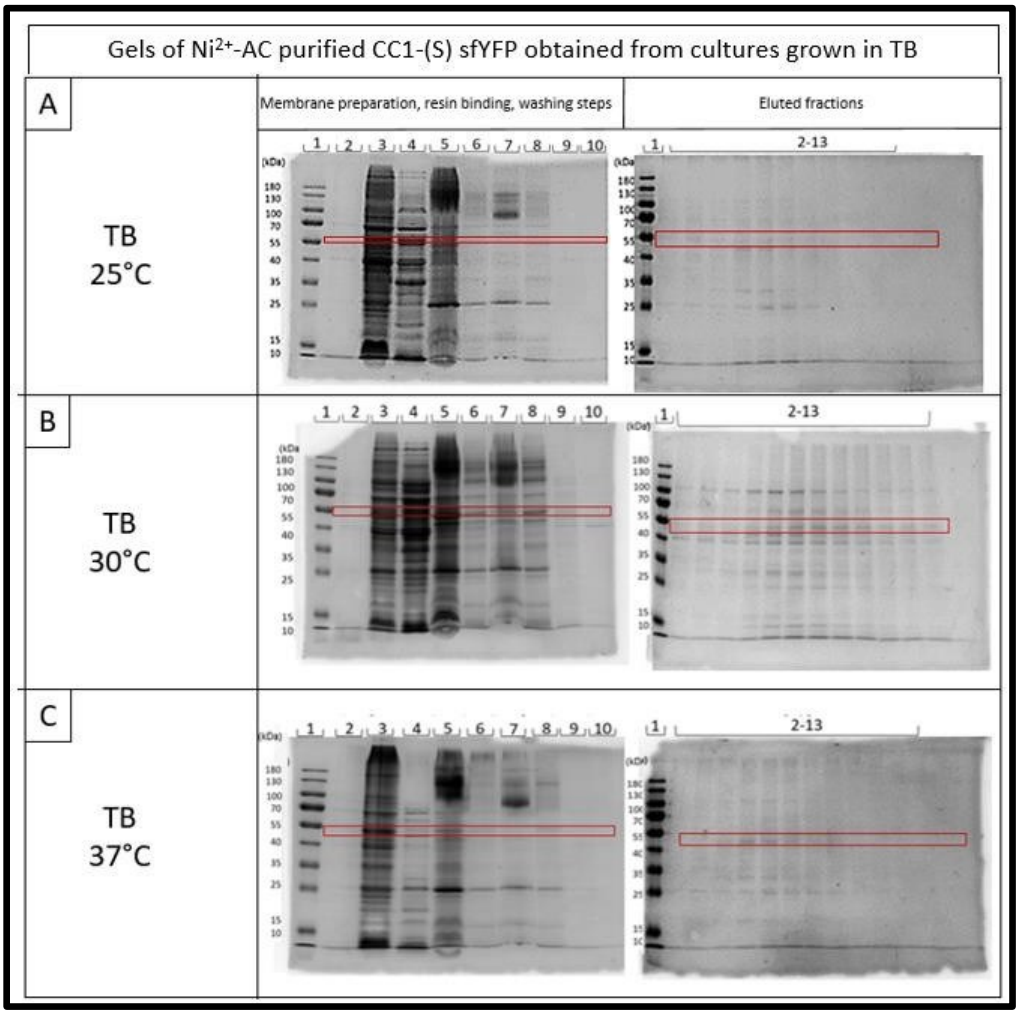
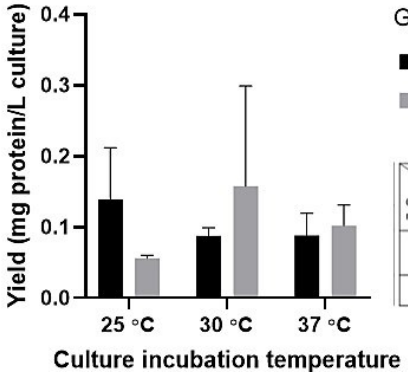


Figure 3.13. (A-C) Electrophoresed polyacrylamide gels corresponding to CC1-(S) sfYFP protein samples obtained from cells grown in TB medium. Culture incubation temperatures: 25°C (panel A); 30°C (panel B); 37°C (panel C). Highlighted in red borders are the protein bands at the corresponding theoretical molecular weight (approximately 56.6 kDa). Each set of gels was loaded as follows: (Left) Lane 1. Prestained protein molecular weight marker; Lane 2. Uninduced whole cells; Lane 3. Lysed induced cells; Lane 4. Supernatant; Lane 5. Membrane; Lane 6. Ni^{2+}-NTA resin incubated with eluted DDM-purified CC1-(S) sfYFP; Lane 7. Flow-through; Lane 8. Wash with 10 mM imidazole; Lane 9. Wash 20 mM with imidazole; Lane 10. Wash with 40 mM imidazole. (Right) Lane 1. Prestained protein molecular weight marker; Lane 2-15: Eluted fractions.

In this case, TB as a growth medium and incubation at 30°C as the growth condition is correlated to the highest expressed protein recovery in the intended fractions, compared to the rest of the temperatures. At 37°C (panel C) and 25°C (panel A), similar behaviours as for LB at 37°C could be assigned (aggregation) by observing the reduced resin binding and overall low yield.

Protein bands of higher intensity were observed, hence a higher purified protein yield corresponding to the eluted DDM-purified protein obtained from the cultures, as presented in Figure 3.14 in LB at 25°C (0.139 mg/L culture), TB at 30°C (0.157 mg/L culture) and TB at 37°C (0.102 mg/L culture). Low yields were obtained for cells cultured in LB at 37°C (0.088 mg/L culture), LB at 30°C (0.087 mg/L culture), and in TB at 25°C (0.056 mg/L culture).

Yield (mg protein /L culture) of CC1-(S) sfYFP depending on culture conditions



Growth medium \ Temperature	25°C	30°C	37°C
	Yield mean ± SD (mg protein/ L culture)		
LB	0.139 ± 0.07302	0.087 ± 0.1128	0.088 ± 0.03086
TB	0.056 ± 0.00418	0.157 ± 0.14142	0.102 ± 0.03086

Figure 3.14. Graph presenting the yields of purified CC1-(S) sfYFP, depending on culture conditions. Yields were obtained based on calculations using densitometry data. Densitometry data was obtained from BSA quantitative assays, performed in duplicate for independent biological replicates (N=2), analysis performed using GraphPad Prism 8 software. Error bars represent the standard deviations between calculated yields. Highlighted in red are the highest yield values obtained between the different temperatures (X-axis: Culture incubation temperature: 25°C, 30°C and 37°C; Y-axis: yield, expressed in mg protein/L culture; Bars: Growth medium: (light grey): Terrific broth (TB), (black): Luria broth (LB). **LB, 25°C** (Yield mean= 0.13904; SD= 0.07302); **TB, 25°C** (Yield mean= 0.05615; SD= 0.004181); **LB, 30°C** (Yield mean= 0.08776; SD= 0.112); **TB, 30°C** (Yield mean= 0.15768; SD= 0.14142); **LB, 37°C** (Yield mean= 0.08856; SD= 0.03086).

The presence of CC1-(S) sfYFP was further confirmed by the presence of bands at the corresponding molecular weight (approximately 56.6 kDa) on the anti-his blots. After yield measurements, for Western blots, the fractions obtained from the culture grown in TB at 30°C presented the highest protein yield. The blots correspond to the crude extract (Figure 3.15, A) and the elutions (Figure 3.15, B).

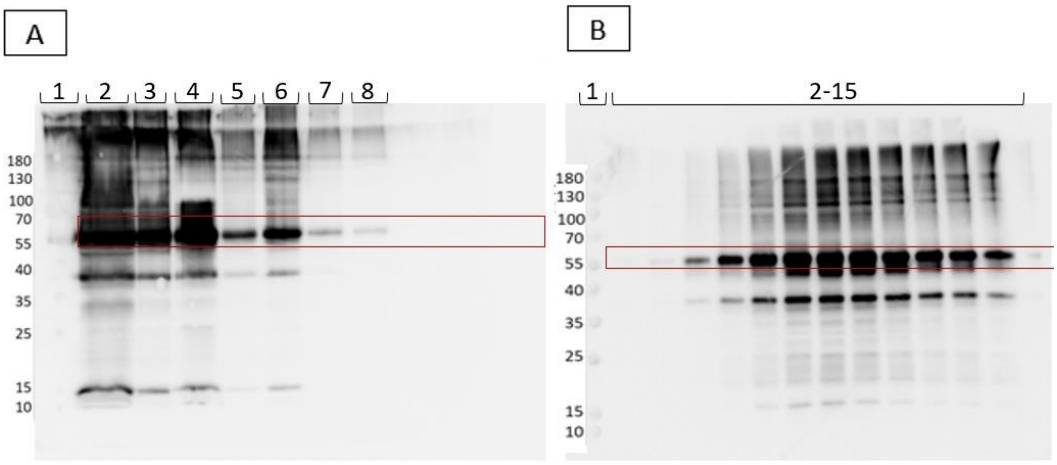


Figure 3.15. Western blot analysis of his-tagged CC1-(S) sfYFP-containing fractions obtained from cells cultured in TB at 30°C. Primary anti-his antibody dilution 1:1000, secondary HRP conjugated antibody dilution 1:5000 . Highlighted in red borders are the protein bands at the corresponding theoretical molecular weight (approximately 56.6 kDa) A. Lane 1. Prestained protein molecular weight marker; Lane 2. Lysed cells; Lane 3. Supernatant, Lane 4. Membrane; Lane 5. Ni^{2+}-NTA resin incubated with eluted DDM-purified CC1-(S) sfYFP; Lane 6. Wash 10mM imidazole; Lane 7. Wash 20mM imidazole; Lane 8. Wash 40mM imidazole. B. Lane 1. Prestained protein molecular weight marker; Lane 2-15. Eluted fractions.

In Figures 3.16 and 3.17, there are represented in a similar manner the acrylamide gels corresponding to the protein samples obtained from cells expressing REAMP-(S) sfYFP. In Figure 3.16, panels A-C represent the gels corresponding to the samples obtained from cells expressing REAMP-(S) sfYFP cultured in LB medium. Panel A presents gels corresponding to samples obtained from cells grown in LB at 25°C, panel B- samples from cells grown at 30°C and panel C- cells grown at 37°C. It can be observed in all fractions a protein band at the corresponding molecular weight of REAMP-(S) sfYFP (56.7 kDa) also identified by mass spectrometry analysis (shown in Figure 3.20 B).

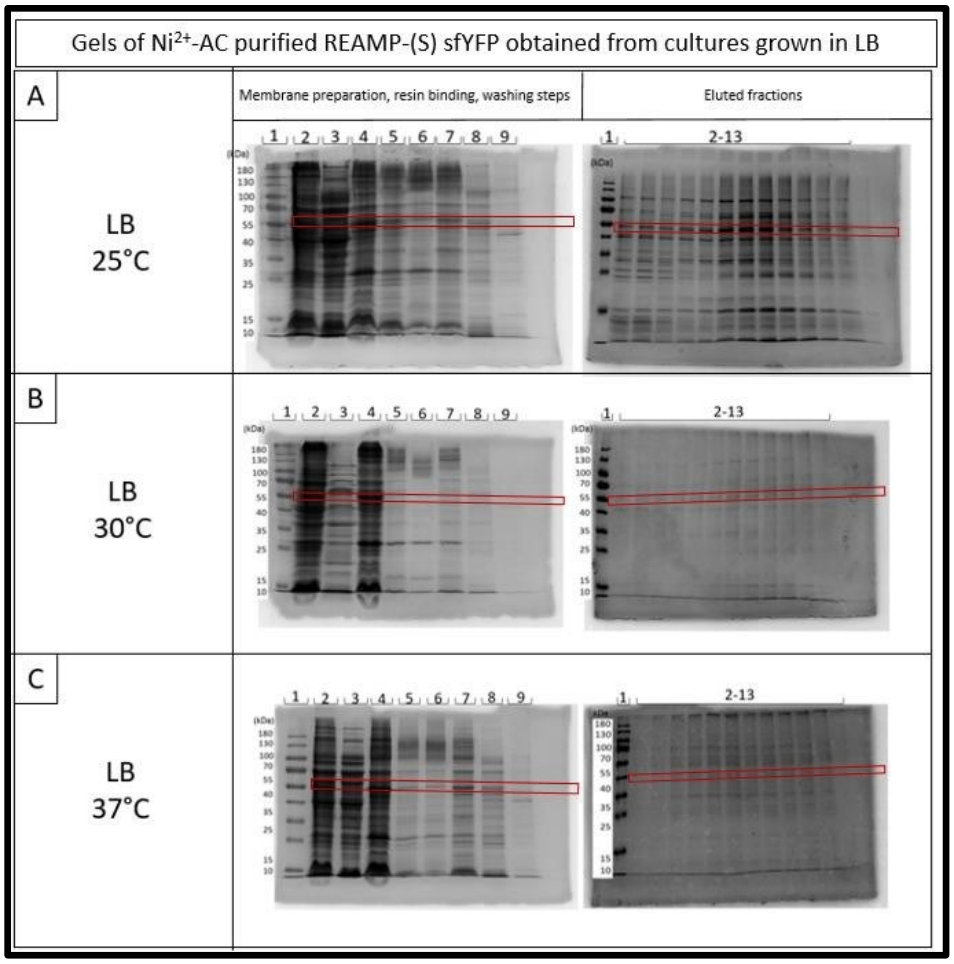


Figure 3.16. (A-C) Electrophoresed polyacrylamide gels corresponding to REAMP-(S) sfYFP protein samples obtained from cells grown in LB medium. Culture incubation temperatures: 25°C (panel A); 30°C (panel B); 37°C (panel C). Highlighted in red borders are the protein bands at the corresponding theoretical molecular weight (approximately 56.6 kDa). Each set of gels was loaded as follows: (Left) Lane 1. Prestained protein molecular weight marker; Lane 2. Lysed induced cells; Lane 3. Supernatant; Lane 4. Membrane; Lane 5. Ni^{2+}-NTA resin incubated with eluted DDM-purified REAMP-(S) sfYFP; Lane 6. Flow-through; Lane 7. Wash with 10 mM imidazole; Lane 8. Wash with 20 mM imidazole; Lane 9. Wash with 40 mM imidazole. (Right) Lane 1. Prestained protein molecular weight marker; Lane 2-15: Eluted fractions.

For the expression of REAMP-(S) sfYFP, LB, 25°C and 37°C growth conditions can be correlated to high yields of purified expressed protein from the intended fractions, and lower for LB, 30°C.

In Figure 3.17, panels A-C there are represented the gels corresponding to the samples obtained from cells expressing REAMP-(S)sfYFP cultured in TB medium. Panels A-C illustrate gels with fractions obtained from lysates of cells incubated at different temperature post-IPTG induction: panel A- cells incubated at 25°C, panel B- 30°C, and panel C- 37°C. In the case of TB, similar to the case of expression of CC1-(S) sfYFP, the highest purification yield from intended fractions were obtained from cultures incubated at 30°C.

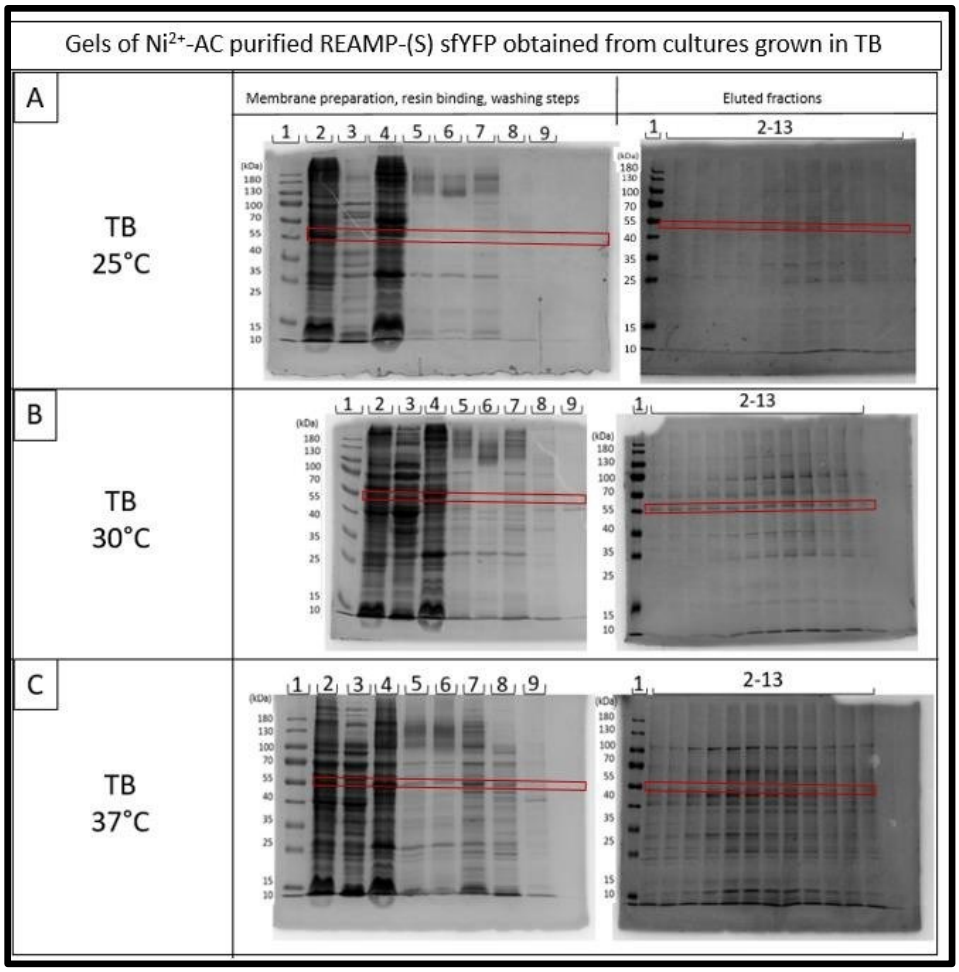
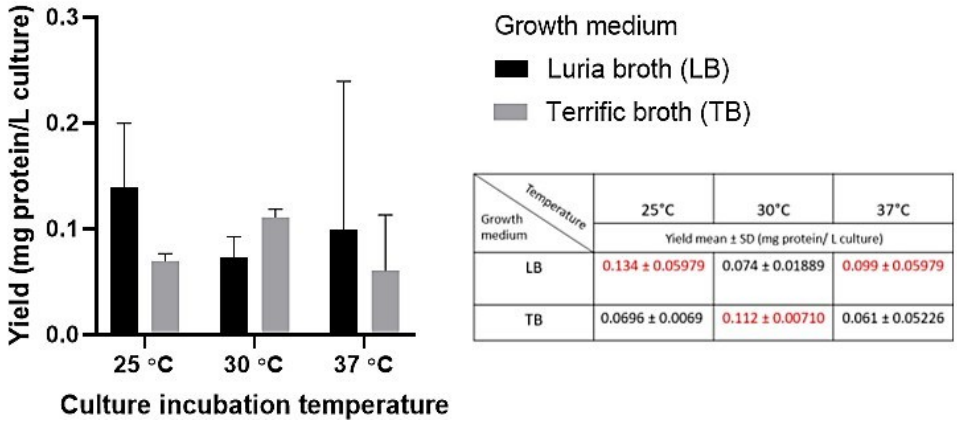


Figure 3.17. (A-C) Electrophoresed polyacrylamide gels corresponding to REAMP-(S) sfYFP protein samples obtained from cells grown in TB medium. Culture incubation temperatures: 25°C (panel A); 30°C (panel B); 37°C (panel C). Highlighted in red borders are the protein bands at the corresponding theoretical molecular weight (approximately 56.6 kDa). Each set of gels was loaded as follows: (Left) Lane 1. Prestained protein molecular weight marker; Lane 2. Lysed induced cells; Lane 3. Supernatant; Lane 4. Membrane; Lane 5. Ni^{2+}-NTA resin incubated with eluted DDM-purified REAMP-(S) sfYFP; Lane 6. Flow-through; Lane 7. Wash with 10 mM imidazole; Lane 8. Wash 20 mM with imidazole; Lane 9. Wash with 40 mM imidazole. (Right) Lane 1. Prestained protein molecular weight marker; Lane 2-15: Eluted fractions.

Protein bands of higher intensity were observed, hence a higher purified protein yield corresponding to the eluted DDM-purified protein obtained from the cultures, as presented in Figure 3.18 in LB at 37°C (0.0995 mg/L culture) and at 25°C (0.13997 mg/L culture) and in TB at 30°C (0.11121 mg/ L culture). Lower yields were observed for the cultures grown in LB at 30°C (0.0736 mg/L culture) and in TB at 37°C (0.06073 mg/L culture) and at 25°C (0.06961 mg/L culture).



Growth medium \ Temperature	25°C	30°C	37°C
	Yield mean ± SD (mg protein/ L culture)		
LB	0.134 ± 0.05979	0.074 ± 0.01889	0.099 ± 0.05979
TB	0.0696 ± 0.0069	0.112 ± 0.00710	0.061 ± 0.05226

Figure 3.18. Graph presenting the yields of purified REAMP-(S) sfYFP, depending on culture conditions. Yields were obtained based on calculations using densitometry data. Densitometry data was obtained from BSA quantitative assays, performed in duplicate for independent biological replicates (N=2), analysis performed using GraphPad Prism 8 software. Error bars represent the standard deviations between calculated yields. (X-axis: Culture incubation temperature: 25°C, 30°C and 37°C; Y-axis: yield, expressed in mg protein/L culture; Bars: Growth medium: (light grey): Terrific broth (TB), (black): Luria broth (LB). **LB, 37°C** (Yield mean= 0.09953; SD= 0.14001); **LB, 30°C** (Yield mean= 0.07362; SD= 0.01889); **LB, 25°C** (Yield mean= 0.13997; SD= 0.05979); **TB, 37°C** (Yield mean= 0.06073; SD= 0.05226); **TB, 30°C** (Yield mean= 0.11121; SD= 0.00710); **TB, 25°C** (Yield mean= 0.06961; SD= 0.00690).

The protein of interest was further detected in the selected eluted samples by observing bands at the corresponding molecular weight on a western blot, as shown in Figure 3.19.

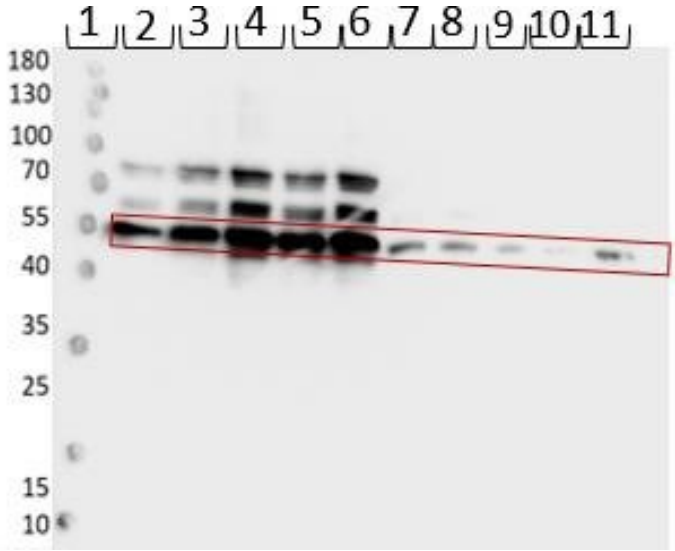


Figure 3.19. Western blot analysis of his-tagged REAMP-(S) sfYFP. Primary anti-is antibody dilution 1:1000, secondary HRP conjugated antibody dilution 1:5000. Highlighted in the red border are the bands at the corresponding molecular weight of REAMP-(S) sfYFP (56.7 kDa). Lane 1. Prestained protein molecular weight marker; Lane 2. Elution 6, LB 37°C; Lane 3. Elution 7 LB 37°C; Lane 4. Elution 8 LB 37°C; Lane 5. Elution 6 TB 37°C; Lane 6. Elution 7 TB 37°C; Lane 8. Elution 6 TB 30°C; Lane 9. Elution 7 TB 30°C, Lane 10. Elution 8 TB 30°C; Lane 11. Elution 6 TB 25°C.

3.4.2.3. Protein identification by mass spectroscopy

Stained protein gel bands at the corresponding theoretical molecular weight were isolated (as highlighted in Figure 3.20 A, B) and sent for analysis at the proteomics mass spectrometry facility at the University of St. Andrews, Scotland. The trypsin digestion of the extracted protein samples resulted in a mixture of peptides. The analysis of the peptide sequences assigned a 28% sequence coverage for CC1-(S) sfYFP (Figure 3.20 A) and 40% sequence coverage for REAMP-(S) sfYFP (Figure 3.20 B). The sequences coverage are marked in red.

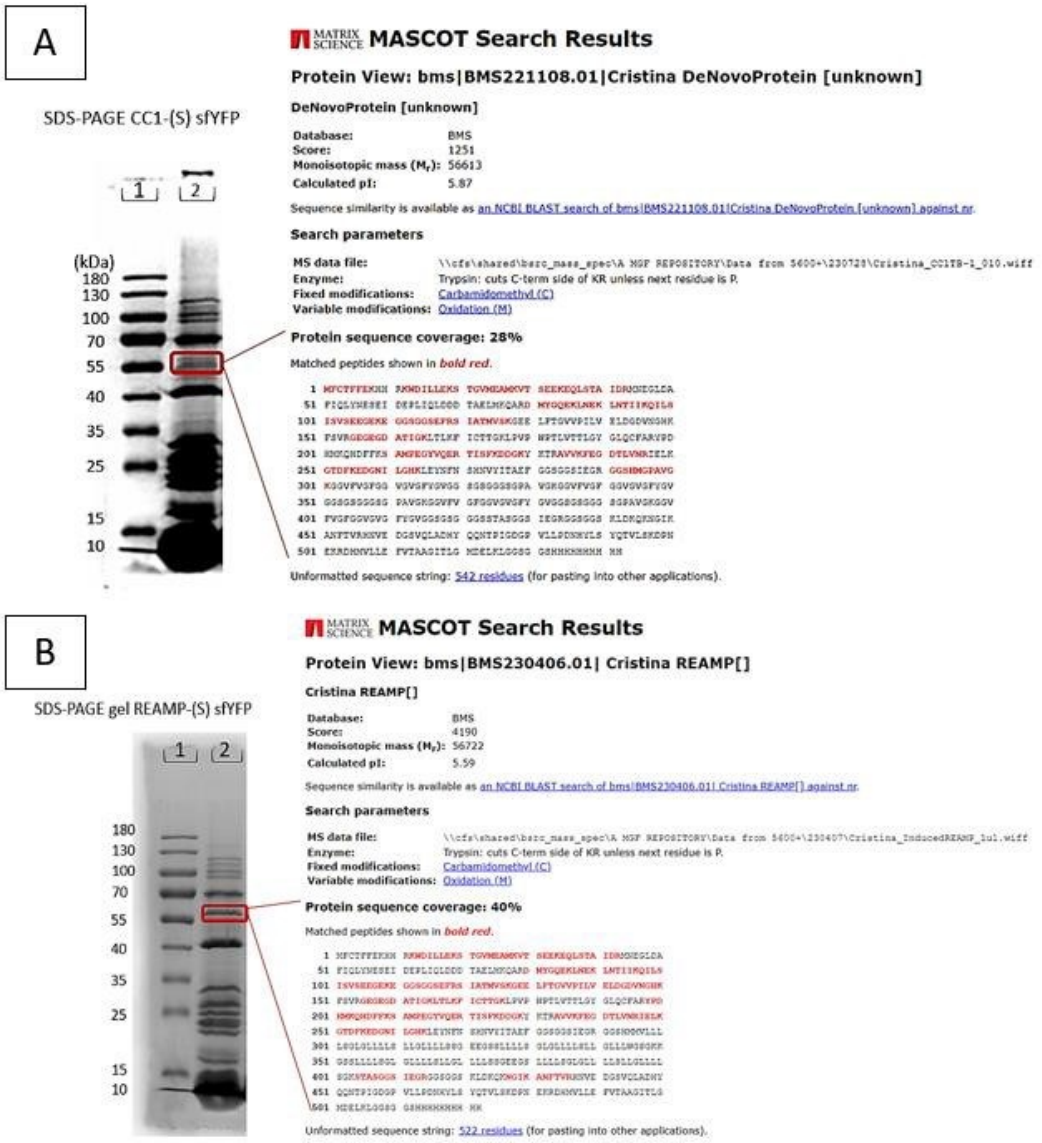


Figure 3.20. Mass spectrometry coverage data for expressed CC1-(S)sfYFP and REAMP-(S)sfYFP. A. Gel and MS data for expressed CC1-(S)sfYFP. (gel) Lane 1. Prestained protein marker. Lane 2. Lysed C41 (DE3) cells expressing CC1-(S)

sfYFP. B. Gel and MS data for expressed REAMP-(S)sfYFP. (gel) Lane 1. Prestained protein marker. Lane 2. Lysed C41 (DE3) cells expressing REAMP-(S) sfYFP. Highlighted in the red borders are the gel bands which were isolated and sent for analysis. (Right) Results of the digestion of the band and sequence coverage (28% for CC1-(S)sfYFP and 40% for REAMP-(S)sfYFP).

In addition to the CC1-(S)sfYFP protein coverage data, the MASCOT search results indicated the presence of another protein (ATP synthase) situated around 55 kDa in the analysed gel band. It was observed a difference in coverage corresponding to the expression of the b-subunit of the ATP synthase between the BL21 (DE3) and C41 (DE3) strain. Increased expression of the ATP synthase corresponded to the protein mixture extracted from the BL21 (DE3) variant (49%), as seen in Figure 3.21 A) and a decreased expression in C41 (DE3), with a protein coverage of 15% (Figure 3.21 B).

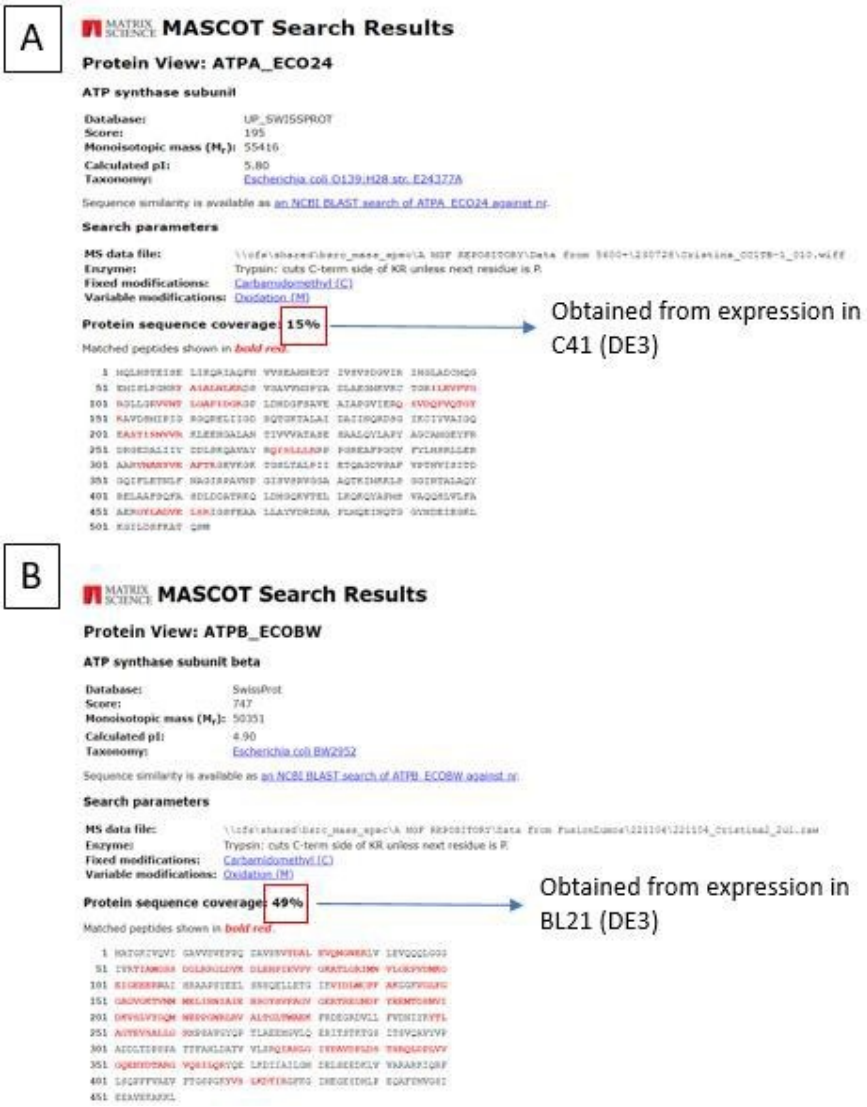


Figure 3.21. Mass spectrometry coverage for the ATP synthase contained in the protein band around 55 kDa. A. Protein sequence coverage obtained from expression in C41(DE3) *E. coli* cells (15%) and B. Protein sequence

coverage obtained from expression in BL21 (DE3) cells

3.4.2.4. Reconstitution into proteoliposomes

The DDM-purified CC1-(S) sfYFP protein was reconstituted in POPC: POPG liposomes, as mentioned in the Methods section and incorporation were detected by sucrose gradient flotation and western blotting (Figure 3.22). A band at the corresponding molecular weight (approximately 56.6 kDa) was observed in the higher percentage sucrose gradient fraction.

Figure 3.22. Anti-his western blot analysis of proteoliposomes. Lane 1. Prestained protein molecular weight; Lane 2. Fraction containing empty liposomes; Lane 3. Fraction containing the proteoliposomes consisting of CC1-(S) sfYFP incorporated in POPC: POPG liposomes.

3.5. Discussion

The study of membrane proteins is particularly challenging due to their topology in a complex lipid environment (Lalaurie *et al.*, 2018). Furthermore, the sequence novelty of the *de novo* proteins enhances the complexity and diverts the studies from the patterns reported in the literature corresponding to the naturally occurring proteins. For this reason, each step in the optimisation of the extraction, solubilisation and purification represents a bottleneck and requires an individually tailored experiment depending on the subjected protein. It is well known from the plethora of studies on membrane proteins that the over-expression of recombinant membrane proteins can disturb the homeostasis of the host cell, perceived as toxic. It was shown that the host cells, under stress conditions, can undergo certain mechanisms of adaptation, which can focus on the restructuring of the membrane composition or the induction of the formation of inclusion bodies. Nevertheless, the expression of *de novo* membrane proteins would represent a disturbing factor in the normal metabolism of the chosen expression host. In the context of

the expansion of the membrane protein *de novo* design, it would be valuable to add experimental results to the computational predictions and to be able to assign certain sequences, consensus, or motifs to certain cell responses. This would have a great impact on the membrane engineering field for different purposes (e.g. modification of membranes for the production of high-valuable chemicals for industrial applications). Furthermore, this would serve as a contribution to the understanding of the membrane folding mechanisms through the detection of the specific essential interactions between the *de novo* protein and the annular and non-annular lipids and to probe experimentally the integrity of the time-saving computational studies, increasing this way the accuracy of the research.

This chapter comprised the optimisation of the molecular cloning of the synthesised CC1 and REAMP genes, along with the optimisation of their expression in *E. coli* bacterial cells, extraction and purification.

3.5.1. Molecular cloning optimisation

Referring to molecular cloning optimisation, this part of the project was very challenging, especially due to the low yield of the DNA obtained from the gel extraction step, which impeded the remainder of the cloning workflow. Although the commercial gel extraction kit (Nucleospin®, Macherey-Nagel, Thermo Fisher, UK) has shown to be the most efficient in solving the issue, the experiments involving it had to be delayed due to the temporary unavailability at that time and delayed product shipping, caused by the COVID-19 pandemic. For this reason, many other methods and techniques had to be employed. It was observed from the results obtained from the other commercial kits, also based on the silica technology, that the prevalent DNA loss was within the first flow-through, as shown by NanoDrop quantification of each flow-through, suggesting that the DNA was not binding efficiently to the silica membrane of the provided spin columns. The following rationale was to substitute the silica-based columns. Some of the employed methods, which substituted the silica-based spin columns, such as the "freeze and squeeze" method, seemed to give better results compared to the other tested commercial kits. However, the results were not consistent across multiple tests and were also more time-consuming, involving longer centrifugation steps and incubation times. The Nucleospin® columns were successful, most likely due to the difference in the binding capacity and efficiency of the silica-based membrane and not on the material itself. Their binding capacity is 25 µg of DNA fragments with sizes between 50 bp-20 kbp, which is greater compared to QiAquick® having a binding capacity of 10 µg of fragment sizes between 70 bp-10 kbp or Monarch ® of 5 µg.

The conditions of the ligation reaction also had to be improved by not only testing the conditions recommended by the manufacturer but also findings from the literature. Up to date, there have been developed alternatives for traditional cloning, such as Gibson assembly, Golden gate assembly, NEBuilder HiFi assembly, combinatorial cloning, etc., which are used for diverse synthetic biology projects; however, within this part of the project, we opted for the versatility and cost-advantage characteristic of the traditional cloning. For future perspectives on the creation of a library of *de novo* sequences derived from CC1, the use of the other types of DNA assembly would be desirable. For the transformation of ligated plasmid products, there were initially tested the commonly used strains for routine cloning XL-1 Blue competent *E. coli* cells ($1x10^8$ transformants/ µg) and XL-2 ultracompetent *E. coli* cells ($5x10^9$ transformants/ µg); however, without success, although the designed genes sent for synthesis (Eurofins Genomics, UK), were codon-optimised for the *E. coli* K-12 cells beforehand. It was then tested the *E. coli* BW25113 strain, a genetically modified (*recA*$^+$, *hsdR*$^-$) K-12 derivative, which was described previously to have a transformation efficiency 100-fold greater compared to XL-Blue or DH5α strains (Yang *et al.*, 2022) and further proven by the successful transformations within this project.

3.5.2. Optimisation of expression-initial considerations

The expression of the *de novo* designs also highlighted the importance of using modified host cell strains for successful production, and we also noted some possible implications of *de novo* protein expression on the membrane lipid restructuring.

It is important to emphasise that membrane protein design is a challenging journey. Despite the recent advances in computational design and predictions, there are certain imperative research questions, such as biocompatibility and function, to be answered jointly with the experimental results. To date, the focus of *de novo* protein design has been mainly on the soluble proteins (DeGrado and Korendovych, 2020) with different folds. In this case, the design follows the arrangement of the hydrophilic amino acids on the protein surface; the folding is driven by the hydrophobic effect and the non-covalent interactions that occur. For membrane proteins, the complexity is enhanced by their topology in a complex environment mainly represented by lipids. The folding of the membrane proteins requires the insertion of the membrane. For this, the protein's sequence has to be complementary with the different regions of the membrane- hydrophilic amino acids as anchors in the hydrophilic border of the membrane and hydrophobic amino acids interacting with the fatty acids chains. There is also progress in redesigning naturally occurring proteins for different functions: to bind metals and small molecules or to act as

catalysts in different reactions. However, the design, especially of membrane proteins, even from minimal principles, remains an open research question. An empiric example to emphasize the challenge involved in the expression and folding of computationally re-designed proteins, would be the small percentage of expressed, purified and folded designed out of the total designable proteins to achieve a specific fold, as presented by Pan and colleagues in their work (Pan *et al.*, 2020).

3.5.2.1. Expression of de novo antiparallel homotetramers

CC1 was designed as a bundle of tetrameric antiparallel transmembrane helices joined by short, flexible linkers (Van Rosmalen, Krom and Merkx, 2017), with the focus on the design of inter-helical interfaces and the protein-lipid interfaces, trying to orchestrate the stability and the possibility of oligomerisation of helices based on the same sequence. Due to the novelty of the sequence, the literature data for the expression of *de novo* transmembrane proteins in *E. coli* cells was limited.

3.5.2.2. Solid-phase peptide synthesis

Furthermore, multiple studies on the synthesis of *de novo* proteins rather focused on the solid-phase peptide synthesis, (using either Fmoc or Boc strategies), as is the case of the Zn^{2+} -Rocker (2MUZ PDB entry), an antiparallel homo-tetrameric transmembrane helices bundle, designed by the DeGrado research group (Joh *et al.*, 2014). Other *de novo* proteins homo-4-mer with similar symmetries (dihedral D2/cyclic C2), such as PDB:1SNE (Ali *et al.*, 2004) or the porphyrin binding PRIME protein (Korendovych *et al.*, 2010) or homo-2-mer bundles such as PDB:1JM0 (Di Costanzo *et al.*, 2001), PDB:1G6U (Ogihara *et al.*, 2001) have been reported to be obtained via solid-phase synthesis.

3.5.2.3. Biocompatible de novo antiparallel transmembrane bundles

Recent studies have elaborated the design of biocompatible antiparallel transmembrane helices bundles, and functional for the electron transfer (Lalaurie *et al.*, 2018), REAMP, the sequence which we used as a comparison term, for design stability and biocompatibility in *E. coli* cells of the CC1 sequence in this study. In the initial studies, REAMP was expressed in C43 (DE3) *E. coli* cells and had a full variant of the superfolder GFP tag attached at the C-terminus. In this study however, we had modified constructs, including the mistic sequence and the split sequences of the superfolder YFP (sfYFP) on either side of the designed protein gene. We tried the expression in different expression *E. coli* strains, such as the BL21 (DE3) and the C41 (DE3) strains.

3.5.2.4. Choice of expression host. The robustness of E.coli cells for protein expression varies between strains

The results of this study highlight once again the importance of the selection of strains capable of expressing MP that are associated with lower yield (Rosano and Ceccarelli, 2014). This is a case-specific optimisation step, that depends on multiple factors, such as the availability of choices at the time of the study, or on the features comprised in the DNA construct that encodes the target protein. In this case, *E.coli* was the chosen expression system due to its cost-efficiency and its capacity to produce the protein of interest in a timely manner (pure protein obtained after 3 days from the cell transformation). In this study it was compared the expression of the newly designed protein in the *E. coli* strains BL21(DE3) and C41(DE3), representing the strains available at that point. The C41 (DE3) strain could express the proteins, as shown by the mass spectrometry data, in comparison with the commonly used BL21 (DE3). There were multiple assumptions made for the reason which stands behind this result, which are detailed in the following subsections. Although experiments did not include this aspect of the study, it was reported in the literature that for the BL21 (DE3) and C41 (DE3) strains were found essential differences in the lac repressor gene of the lac operon as well as in the lacUV5 promoter of the DE3 region, by which the toxic effect on the host caused by overexpression of membrane proteins could be diminished in the C41 strain. A mutation in lacI was demonstrated to be crucial for conferring tolerance to membrane protein overexpression.

C41(DE3) and C43(DE3), the latter being reported in some expression studies of *de novo* proteins, are widely used for the production of a variety of membrane proteins and toxic proteins. It was observed that these strains suffered less from Sec translocon saturation, however, this is not a general rule. It was also reported that changes in the lacUV5 promoter of these strains are responsible for improving membrane protein overexpression (Kwon *et al.*, 2015). The characteristics of different promoters can be harnessed to efficiently express the protein after the metabolic adaptation of the cell. The T7 promoter-based plasmids could integrate, in this sense, promoters which are known for low leaky expression, such as arabinose (Cronan, 2006) or rhamnose (Giacalone *et al.*, 2006). Interactions between the operon-repressor have shown to improve the expression of toxic membrane proteins, as studies on the expression in *E. coli* BL21 (DE3) cells of a series of calcium channels and viral ion channels have indicated. The expression improved by 5-20 fold as a result of this type of interaction (Narayanan, Ridilla and Yernool, 2011). Regarding the basal expression, or plasmid “leakiness”, this could not be assigned for our constructs containing only the T7 promoter because of the absence of protein bands of considerable intensity on the SDS-PAGE gels for the uninduced C41 (DE3) cells, however, the use of other promoters in future studies would be beneficial

to aid in the mechanisms of the cell.

3.5.2.5. Expression results for CC1 in E. coli cells emphasize on the need of sequence optimisation

It could be seen from the CC1-(S)sfYFP expression levels in *E. coli* cells were low. This could be observed from both the intensity of the protein bands observed on the Coommassie-stained, and from the mass-spectrometry coverage results. Apart from the need of choice of an optimal cell strain, to sustain a toxic *de novo* protein, results indicate the need for future sequence optimisation. This can be concluded especially from the comparison between the bands observed in the electrophorised acrylamide gels, especially when the soluble fraction was compared to the membrane fraction. It could be seen a much darker band around the expected molecular weight corresponding to the CC1-(S)sfYFP, in the soluble fraction. This would lead to the conclusion that the partition between the aqueous (soluble) and oil (membrane), determined by the CC1 sequence, could be inclined towards the aqueous phase. Nonetheless, the size of the superfolder YFP tag could contribute to this phenomenon as well, thus is would be essential for further proof-of-concept the comparison between the expression levels of the sfYFP-tagged version versus the non-tagged version of the protein, considering that the hydrophobicity scale (Monné, Hermansson and Von Heijne, 1999) of amino acids would place the current sequence of CC1 on a favourable territory. More importantly, before the molecular cloning optimisation, it would have been essential to simulate any major conformational changes of the entire protein model (with the GGSGSGGGS linkers and the YFP tag). This would have enabled a prediction of the predominant forces that occur, hence would have expected a specific partition. The GGSGSGGGS linkers were added between the transmembrane regions as short, non-functional linkers, as reported (Van Rosmalen, Krom and Merkx, 2017).

Also, another reason behind the low cell expression, that is related to the sequence, would be the occurrence of unfavourable events at the mRNA level, such as hairpin structures that can hinder the ribosome binding or progression (De Smit and Van Duin, 1990), or errors in the sequence recognition pathways. This would require for further optimisation of the codons, while maintaining the encoded protein sequence and to reduce in the same time the formation of the hairpin structure. Softwares like mFold (Reeder *et al.*, 2006), SparseRNAfold (Gray, Will and Jabbari, 2024) or NUPACK (Fallmann *et al.*, 2017) could be used to predict artefacts of the mRNA. Another approach would be to introduce spacer regions, such additional nucleotides between the ribosome binding site and the start codon, as seen to influence protein production in *Bacillus subtillis* (Volkenborn *et al.*, 2020). On the other hand, errors in sequence recognition could be avoided by overexpressing factors that aid in the translation initiation or elongation, or to optimise the sequence around the start codon (such as Shine-Dalgarno in prokaryotes).

The formation of inclusion bodies is another aspect that could be associated especially with the expression in the BL21 (DE3) strain. Although the data was not presented, experimental observations were seen after the final centrifugation step of the DDM-solubilised membrane, obtaining a white-opaque pellet, most probably represented by inclusion bodies (Singh *et al.*, 2015). This further sustains that the membrane of the BL21(DE3) could not assimilate the protein, causing protein aggregation, hence formation of inclusion bodies.

3.5.3. Potential mechanisms of adaptation of the cell to expression of *de novo* proteins

3.5.3.1. Expression levels of ATP synthase as indicator of cell stress

A comparison between the mass spectrometry results obtained for the expression of CC1 in BL21 (DE3) and C41 (DE3) led to further hypotheses related to the biocompatibility of the sequence and triggered adaptation mechanisms. Notably, this was observed from preliminary data, from only a single mass-spectrometry comparative analysis and multiple factors needs to be considered, such as equipment-related, or technical errors. Along with the expression of CC1, in the mass spectrometry results it has also been detected the expression of the beta-subunit of the ATP synthase, with a coverage of 50% in the case of BL21 (DE3), decreasing to 15% in C41 (DE3). Inversely proportional it has been detected the sequence coverage for the expression of CC1. As identified in previous studies, the mistic sequence and the co-expression of the b-subunit of the ATP synthase with the protein of interest are considered important factors for targeting the protein of interest in the bacterial membrane (Zoonens and Miroux, 2010). Furthermore, ATP synthase has been associated with the proliferation of the inner membrane of C41 (DE3) and C43 (DE3) and modification of the lipid composition, with increased levels of cardiolipin in the membranes expressing the b-subunit (Arechaga *et al.*, 2000). Although this could lead to a possible explanation of the results obtained, replicates would be needed to form a sustained conclusion.

3.5.3.2. Membrane restructuring as stress response

It has been presented from early studies that the cell increases its tolerance for the overexpression/ expression of toxic proteins through the biosynthesis of lipids (Raetz, 1978). Interestingly, regarding the regulation of the synthesis of fatty acids and phospholipids, there have been identified in *E. coli,* the glycerol auxotroph genes *plsB* and *gpsA* and for which it has been reported that the glycerol starvation leads to the inhibition of the lipids biosynthesis (Raetz, 1978). In this case, regarding the cell growth conditions, this could be considered one of the reasons why, predominantly, the growth and

expression results in this project were favourable, corresponding to the use of terrific broth as a growth medium. Terrific broth is supplemented before sterilisation with glycerol as an energy source for growth and, consequently, as a contributor to the cell's mechanism of adaptation to protein expression through lipid biosynthesis and membrane restructuring. In contrast, the results observed for Luria broth could be explained following this rationale. Lately, there has been an increasing interest in the study of membrane lipid modifications, especially since they have been observed in different pathologies (Pichler and Emmerstorfer-Augustin, 2018). One of the reported modifications is the increased level of cardiolipin in the stationary growth phase in *E. coli*, hence stabilising the membrane (Hiraoka, Matsuzaki and Shibuya, 1993) due to the presence of four acyl chains and special anchoring behaviour in the membrane. In addition to the terrific broth glycerol content, the corresponding growth curves have shown a prolonged stationary/slow growth post the IPTG induction, especially in the cases of incubation at 25°C and 30°C; in the meantime, possibly the level of cardiolipin increased and stabilised the membrane, increasing the chances of accommodation of the *de novo* designs in this specific strain.

3.5.4. Impact of the temperature on the *de novo* protein expression

The results in this study (presented in Chapter 2) from the MD simulations (indicated that the sequence of CC1 would be compatible and would confer stability of the design across bilayers of different lipid compositions (POPC, POPG, POPE). In contrast, the sequence of REAMP tends to have an increased affinity towards the POPE. The simulations were performed at 300° K (27°C). The computational results, in this case, were mirrored in the experimental results observed for the cultures incubated at 25°C.

The results that have shown that the lower incubation temperatures were correlated to higher expression yields for both the CC1-(S)-sfYFP and for the REAMP-(S)sfYFP. This phenomenon could be explained by a cascade of events which are determined by the modification of the membrane as a response to the stress conditions. It is well known that *Escherichia coli* cells are robust cells, one of the reasons that are being extensively used, and experiments on different strains have shown that they can tolerate growth at incubation temperatures of up to 42°C, with the optimum incubation temperature of 37°C. The membrane, having in composition lipids, the fluidity can be altered through temperature modifications and higher temperatures would determine a more fluid membrane, depending on the lipid composition. Moreover, it is known that the cells can adopt certain mechanisms of adaptation as a response to changes in the membrane fluidity, in order to maintain the homeoviscous environment. It is also to be noted that not only temperature can represent a stress factor for membrane modifications, but

also the osmotic stress, or pH changes can cause changes in the membrane composition, protein expression and fluidity (Sawant, Singh and Appukuttan, 2022). Although for a better understanding of the mechanism, all these factors should be monitored, from the results of the presents study, we could correlate the expression with lower incubation temperatures (25°C and 30°C), with also a correlation with the cell growth rate in the different cell growth media used (Luria Broth, correlated to better expressed protein recovery from intended fractions for incubation at 25°C, Terrific broth enabled better expression yields for incubation at 30° of both proteins). The medium chemical composition could contribute to the corresponding osmotic stress for the membrane modifications, LB increasing the osmotic stress through the salt content, whilst TB could diminish it. In the case of LB, its formula also contains salt (NaCl) in the concentration of 10 g/L (or 1%). Salt provides essential electrolytes, and it is usually used in concentrations between 5 and 10 g/L for the growth of transformed bacteria with different plasmids, the lower concentration being used for growing bacteria transformed with plasmids with resistance to salt-sensitive antibiotics. TB, on the other hand, lacks salt in its formula. However, it is more prosperous than LB, containing potassium phosphates, which have the role of lowering the chances of cell death caused by changes in pH, and glycerol as an extra source of energy apart from the peptone. In addition to the observations of the present study, TB was reported in other studies as a growth medium which could enhance the cellular stress response through its nutritional support, which could explain certain phenomena in the present study, although without experiments performed in this sense in order to prove this hypothesis. It was previously reported that the amino acids in TB could be used by the cell to overexpress the stress-response proteins (Kram and Finkel, 2015) , like chaperones such as DnaK, or GroEL, which could act in the reduction of protein damage during stressful growth conditions such as nutrient depletion, or drastic pH shifts. Nevertheless, the inclusion of glycerol in the chemical composition of TB, as an extra carbon/ energy source, contributes to the ATP (re)generation during osmotic stress (a scenario which could be sustained in the context of the non-tuned expression of a membrane channel that passes the membrane in an uncontrolled manner) or nutrition depletion. Although in the experimental design step, it might seem trivial, as a case-specific impact analysis, overall, TB, in comparison to LB, ensured a higher resistance to cell's stress response to "toxic" membrane protein expression through its buffering properties, nutrient enrichment and sustained carbon source.

At the same time, it is recommended for bacteria transformed with low copy plasmids and known to lead to fast growth and high yield of bacteria (Lessard, 2013). The plasmids, in which the genes encoding *de novo* sequences for CC1-(S) sfYFP and REAMP-(S) sfYFP were cloned, also contain the kanamycin resistance gene. We observed that TB represented the growth medium for the cells expressing

these *de novo* proteins, generally resulting in faster growth and longer growth time frames. This can be explained also by the higher amount of nutrients, but also by the lack of salt in composition. It was previously shown that kanamycin-resistant bacterial growth can be affected even by low salt concentrations of 1% or lower (Park, Choi et al. 2021), as is the case of LB.

Another consideration for the impact of temperature would be related to the oxygen consumption. Inductively, higher temperatures would correspond to stimulated cell metabolism and thus, an increase in oxygen consumption from the cell growth chamber. As we could see in this study from the growth experiments, cells indeed transitioned the growth phases faster when incubated at 37°C; this temperature, in these cases, perhaps promoted a boost of the metabolism of the transformed cells encoding the CC1-(S)sfYFP and REAMP-(S)sfYFP genes. Early studies on the membrane lipid modification as a stress response indicated that *E. coli* cells stimulate the formation of a less fluid (more rigid and impermeable) membrane as a response to a low dissolved oxygen supply (Arneborg, Salskov-Iversen and Mathiasen, 1993). Due to the fact that in our experiments, the cultures were grown in flasks, with only periodic, limited aeration (aeration was introduced at the time points corresponding to the cell density measurements), and the literature studies were performed in bioreactors with controlled, constant supply of oxygen resource, this mechanism could be plausible in the present case. The molecular mechanism of this event is determined by an increase in the synthesis of cyclopropane fatty acids, which reduce the degree of unsaturation and keep the content of saturated lipids constant, all influenced by the consumption of oxygen. It is known that the synthesis of these fatty acids acts as a faster substitution mechanism for decreasing the membrane fluidity, in favour of the time-consuming replacement of unsaturated fatty acids by saturated fatty acids. From a thermodynamic point of view, this can be seen as an energy (Gibbs-free energy) - conserving substitution mechanism.

Independent of direct correlation to oxygen consumption, Marr and Ingraham (Marr and Ingraham, 1962) have shown in their work that low incubation temperatures encourage the production of unsaturated fatty acids (to counteract the determined decreased fluidity) through the activity of the FabF enzyme, and high temperatures encourage the production of saturated fatty acids (analogue to unsaturated fatty acids) through the activity of the FabB enzyme.

Other quantitative studies performed on *E. coli* modifications as a result of temperature modifications looked at how temperature impacted the overall stoichiometry between the carbon, nitrogen and phosphorus biomass content, and this increased with the decrease in incubation temperature (Cotner, Makino and Biddanda, 2006). Their study could divide the proportion between the allocation of the high content of P for ribosomal assembly at lower temperatures. These observations

further serve as support statements for the expression of CC1 at lower temperatures.

Another possible molecular mechanism proven in previous studies and referring to the protein expression sustained at lower temperatures could be explained by the ribosomal content. It is known that cells growing at lower temperatures contain a higher content of ribosomal subunits. The high content of ribosomal subunits is synthesized by the cells which are found in low growth rate, also called by scientists 'starved' cells. The reason for the synthesis of these ribosomes is a response of the cell to the growth condition, known as 'waiting for better days' (Farewell and Neidhardt, 1998). The high level of ribosomes would be useful to rapidly respond to an increase in nutritional richness, for protein synthesis and to overcome any blockage in protein initiation. In this study, the growth curves obtained from the different temperatures indicate the growth rates were slower at 25°C and 30°C until the induction point, especially in the case of 25°C. After the induction point, the growth at 25°C presented a steeper ascension, compared to a more linear growth observed pre- and post-induction for incubation at 30°C. This could be explained by the fact that the 25°C temperature point is closer to the experimentally-determined one at which it was determined the highest content of polysomal subunits (Farewell and Neidhardt, 1998).

3.5.4.1. Membrane fluidity sustains protein stability and interactions

The relationship between temperature and membrane fluidity can be extended also between temperature and kinetic energy. A decrease in temperature would be correlated to a lower membrane fluidity (higher rigidity) which would correspond to a decrease in the kinetic energy, or reduced mobility of the phospholipids. This promotes the establishment of interactions between the hydrophobic side chains of the amino acids within the transmembrane region of the protein's helices and the hydrophobic tails of the lipids. The temperature decrease determines both the increase of lipid ordering and the decrease the molecular vibration, rotation, and torsion, restricting the movement of the side chains, hence promoting the close protein-lipid interactions. Despite the fact that aromatic amino acids are involved in the stabilisation of the design through the interaction with the lipid acyl chains, this can disturbed by increasing the temperature of the system. CC1 design comprises phenylalanine residues in the transmembrane region, which are known to interact well with the hydrophobic core of the membrane and to promote membrane permeability by the formation of pore-like complexes with the lipids (Adamczewski and Tsoukanova, 2018) and also to have a preference for the type of lipid packing for its intercalation. However, when the incubation temperature increased to 37°C, the interaction between the phenyl ring and the acyl chains was destabilised, hence the decreased capacity of maintaining the design

in the membrane. Taking into consideration the plausible mechanisms presented previously, if the membrane lipid composition modifies and the content of unsaturated fatty acids increases, the number of π-interactions between the phenylalanine aromatic ring and the double bonds of the lipid tails would be enhanced. Moreover, through this type of interaction, the structure of the protein would be stabilised and possibly would further support an 'open-closed'-like mechanism of function of the CC1, already determined in the MD simulations (described in Chapter 2), due to the possibility of cis-trans isomerization of the double bonds engaged in the interactions with the phenylalanine residues.

Furthermore, alteration of the incubation temperature impacts the membrane fluidity through the modification of the phase transition of the membrane-component phospholipids, and this could alter the spatial arrangement and interactions of the phospholipids. The phase transition temperatures (from gel-ordered state to liquid crystalline) for the phospholipids of interest in this study are POPG T_m = -2 °C, POPE T_m =25°C and CL T_m =47-62.2 °C (Avanti Polar Lipids, no date).

The phase transition temperature can increase with the increase of the fatty acids chain length and headgroup size and decrease with the increase of the number of unsaturated fatty acids. Rappolt and co-workers have shown in their studies on SOPE artificial models (monounsaturated stearoyl-oleyl-PE) that the phase transition from lamellar to hexagonal H-II was determined at 30.9°C (Rappolt *et al.*, 2008). Moreover, the distance between the headgroups increased as a function of decreasing the temperature in the lamellar phases. These observations would support analogously the preference for expression and membrane insertion at lower temperatures of CC1, along with the high content of PE in *E. coli*. The membrane architecture determined in H-II phases observed at a temperature close to the expression temperature obtained in our experiments could favour the spontaneous protein insertion, due to the reduction of lateral pressure in the regions between the membrane rods and the lamellar bilayer. Thus, the incubation temperatures below the phase transition temperatures could sustain the preference for insertion in an ordered bilayer of the novel sequence and represent the optimal temperature for this event.

If we consider the hypothesis of increased cardiolipin synthesis in the stationary growth phases applicable in this study as well, and the lipid phase transition temperature of cardiolipin for the cell cultures incubation temperatures of 37°C and 30°C, the T_m can further explain the highlighted the importance of cardiolipin in the cell growth, membrane stabilisation and yields of the expressed designs, however to clearly establish the existence of a trend, it has to be increasing the number of experiments (currently N=2). Furthermore, 25°C expression was favoured for REAMP due to its predicted affinity for POPE, the percentage of PE in *E. coli* membrane (75%) and compatibility with the phase transition

temperature of this lipid. An increased level of cardiolipin in the membrane possibly modifies the temperature of the overall membrane, maintaining it rather in a generally ordered gel-like architecture with fully extended and tightly packed acyl chains. This would enable better interaction between the sidechains of the transmembrane residues and the hydrophobic region of the membrane, lipid architecture which would be disrupted and oriented to more random distribution, hence weaker protein-lipid interaction with increased temperature. Interestingly, the general growth trend of the C41 (DE3) cells expressing REAMP showed a better tolerance of the cells for the protein. In addition to the POPE preference shown in the MD simulations, there might also be a high affinity of REAMP for cardiolipin, due to the presence of arginine residues, thus acting as an annular lipid. High affinity towards cardiolipin was previously shown for the hemoprotein cytochrome c (Trusova *et al.*, 2010). The REAMP sequence did not have any homology to the natural existent proteins, however it was shown to have resembling functionality to the b-type cytochromes (Lalaurie *et al.*, 2018). This remark leads to the conclusion that optimisation of *de novo* sequences can not only influence the folding but also the function, as a result of the amino acids sidechains-lipid affinity and interactions.

3.5.5. Membrane lipid composition and impact on the expression considerations

The observed changes in the membrane lipid composition depending on the growth phase and further possible contribution to protein expression and recovery, lead to the idea of determining any protein motifs able to be selective to specific lipids and to contribute to the expression stabilisation, membrane insertion and perhaps, oligomerisation depending on the lipid environment.

The CC1 had in its amino acid composition tyrosine at the interfacial head group region and stabilised the structure through π-interactions, the 'aromatic belt' at this region being observed in cytochrome c oxidase. The stabilisation in the membrane was again, promoted at lower temperatures, when the percentage of saturated content increased, and this is in line with previous work performed on LacY, PheP, or GalP, expressing the importance of the presence of saturated chain PE, which supported the correct folding and membrane topologies in the membrane (Corin and Bowie, 2020). Previous structural studies and current design approaches also indicate that the importance of the lipid-binding protein motifs has rather a functional significance, as it was determined for cardiolipin. For this lipid it was described to have a very specific binding pattern: KKY, RKY, HRN, or more generally XYZ, where X,Y are positively charged residues and Z polar residue (Hunte, 2005). Considering this aspect, this statement further supports the possible affinity of REAMP towards the cardiolipin, due to the EExS and KKxS (where x is glycine) motifs contained in the linkers, resembling to the XYZ cardiolipin binding motif. Furthermore,

functional studies on the homotetrameric KcsA channel have shown that cardiolipin binding to the arginine and lysine (Iwamoto and Oiki, 2013) residues found at the intracellular surface (Inada et al., 2022) of the M0 helix of the tetramer. Although currently, CC1 does not contain such a motif and only has a locally limited distribution of lysine and tyrosine residues, these observation points can contribute to the design strategy for the functionality of the new protein.

In this manner, it would contribute to either the improvement of the sequence, increase the biocompatibility, either detect possible interactions between the specific protein and the lipids (annular), and to be able to assign certain sequence motifs to different lipids affinities similar to the cholesterol-binding consensus detected in GPCRs ($W/Y\text{-}(X)_{1\text{-}3}\text{-}I/V/L\text{-}(X)_{1\text{-}7}\text{-}K/R$), or in the peripheral-type benzodiazepine receptor ($L/V\text{-}(X)_{1\text{-}5}\text{-}Y\text{-}(X)_{1\text{-}5}\text{-}R/K$) (Contreras *et al.*, 2011). Current approaches for the *de novo* design of helical transmembrane assemblies mainly rely on computational sampling and molecular dynamics simulations (Niitsu and Sugita, 2023).

3.5.6. Mass spectrometry for detection the expression of *de novo* protein

The insertion and localisation in the membrane of the expressed proteins can be detected by western blotting of the membrane. In this study, the common histidine tag was only used as a detection factor. To enhance the accuracy, further epitopes can be assigned to the sequence to help detection (Lalaurie *et al.*, 2018). It was also confirmed by the mass spectrometry of the band at the correct size in the membrane fraction. Classical mass spectrometry, like any technique, has its limitations as well. In this project, trypsin was used trypsin for the cleavage of peptide fragments. However, trypsin is known to recognise and cut at the sites where arginine and lysine residues are present. Referring strictly to the sequence, this can explain the increased protein coverage for REAMP. In the case of the CC1 sequence, it does not contain arginine residues, and the propensity of lysines is reduced, and this could introduce a possible artefact in the interpretation of detected peptide fragments. An alternative to overcome this challenge would be the native mass spectrometry approach. Yet, in this case, it has to be attributed special attention to the use of detergents which are compatible with this technique (Behnke and Urner, 2023), DDM not being part of them. Another alternative would be to increase the number of charged residues in the sequence in the corresponding positions assigned for the interaction with the hydrophilic region of the membrane.

Despite the minimal sequence complexity of both CC1 and REAMP, the novelty of the sequences imposes difficulties in cellular integration. Thus, regarding the optimisation of expression of the *de novo*

transmembrane designs, it would be desirable to compile all factors which are known to diminish the toxic effect exhibited by the expression, such as choosing an appropriate strain and promoters and modulating the growth conditions, depending on the tendencies observed from the molecular dynamics simulations.

3.5.7. Optimisation of solubilisation and purification

In this study, the bacterial membranes were solubilised using DDM as a non-ionic detergent. As solubilisation method, we chose the widely used DDM detergent, due to its efficiency and widely studied applications and for the desired latter applications (reconstituting the *de novo* protein in proteoliposomes and further study of the embedding in droplet-interface artificial bilayers). DDM was chosen a solubilising agent for the preliminary, proof-of-concept studies, remaining for the future to test other solubilising agents such as amphipols or different polymers (SMA, DIBMA, AASTY etc.).Liposomes are one of the most used models to mimic the natural lipid bilayer and to reduce the complexity of the lipid environment in order to gain sufficient information possibly related to folding and function. It has been reported that the presence of lipids and detergent introduces high anisotropy in the system (Harris and Booth, 2012). In relationship with the molecular dynamics simulations, it would be of interest to simulate the proteins in an anisotropic environment, in addition to the semi-isotropic system, which was used as part of the simulations for this project. Increasing the accuracy of the computational study like this would facilitate further determination of (in)dependence of the protein design sequence from the lack of curvature stress, as encountered in micelles, which would give further insights into the folding characteristics. The project did not advance to the point of folding determinations or structural characterisation; however, in the future, it would be desirable to use DM as a detergent, this being associated with better micelle formation, as it was reported initially for the study of REAMP. In general, a screening of detergents would imply a high financial cost for the experiments, but considering the reported features of the polyoxyethylene (8/9) dodecyl ether, its low CMC (0.003-0.005%), lower than the ones of DDM and DM and formation of large micelles (Stetsenko and Guskov, 2017). Furthermore, it would be of interest to test the stability in other types of micelles, as it was reported for the Zn^{2+} -Rocker that the rigidity of the antiparallel helices decreased in the DPC micelles, compared to the lipid bilayer (Joh *et al.*, 2017), when envisaging the functionality of the CC1 design. The purification, solely based on the nickel affinity of the histidine tag, has proven to be not sufficient; on the SDS-PAGE gels were visible multiple bands, being either other co-eluted histidine-containing membrane proteins, either different oligomer states or protein-detergent complexes. The fractionation of the complementary size exclusion chromatography purification step was impeded by

the presence of the detergent in the buffers, making impossible the correct fractionation and analysis due to the specific equipment limitations. The concentration of detergent of 0.1% in the elution buffer was slightly above the DDM critical micelle concentration of approximately 0.086%; thus, altering the concentration of detergent was not a valid option. As a result, it would be of interest in the future to alter the DNA constructs by adding different affinity tags, such as the streptavidin tag (Maertens *et al.*, 2015).

Also, part of the future studies would be testing the folding of the proteins by circular dichroism both in DDM and other micelles like DPC (Korendovych *et al.*, 2010) and studying the self-association properties through analytical ultracentrifugation (Correia, 2010).

3.5.8. Additional future perspectives for *de novo* protein expression

As future perspectives, it would be essential to test wider temperature ranges for both experimental and MD simulations and, for the latter, to include cardiolipin in the bilayers of the simulations, as well as other lipid species, which are characteristic to other host cells, such as mammalian, or yeast.

3.5.8.1. Culture growth in controlled environments

Additional to simulations, as an experimental future perspective, it would be of interest to culture the cells encoding the CC1-(S)sfYFP and REAMP-(S)sfYFP in more controlled environments, such as bioreactors, using both LB and TB, repeating the expressing tests at 25°C, 30°C and 37°C, and also introducing other incubation temperatures such as 17°C and 42°C. This would enable the control over the consumption of oxygen and also the supplementation of the growth media with necessary nutrients. Given the fact that the observations on the membrane modifications depending on oxygen consumption, completed by the depletion, or distribution of resources to different cellular subunits, it would be of interest to determine if there would be any changes for the already observed expression patterns and also to have a broader perspective on the temperature tolerance. The addition of the low incubation temperature would serve to determine the validity of the ribosomal content hypothesis, as well as the faster membrane destabilisation at 42°C. Achieving these incubation temperatures would also represent a challenging point, and would be dependent on the equipment limitations, especially for achieving temperatures below room temperature. Achieving 25°C in this study represented a challenge while using the orbital shaking incubators, when not placed in a temperature-controlled room, especially during the seasons with increased temperatures.

Auto-sampling for the determination of the cell culture optical density would also be a factor

which would positively impact the control of the growth environment and also diminish technical and human error.

3.5.8.2. Lipidomics analysis

Due to the existence of many supporting studies referring to membrane lipid composition modification as a function of temperature influence, it would be desirable to use high-resolution mass spectrometry to enable precise lipid profiles. This would be done at various stages of the growth in different temperatures. This would enable to track and quantify any changes in the cardiolipin content and to, have a better correlation between the computational and experimental observed lipid preference of the specific CC1 sequence and further determine if motifs should be designed to support any needed lipid-interactions. This technique enables accurate determination, however it involves high costs. There are other cost-effective methods, however, not at the expense of accuracy. Determining the lipid profiles could also be done by extracting the lipids from the obtained membranes at different time points throughout the growth and analysed using thin-layer chromatography (Deranieh, Joshi and Greenberg, 2013). This method, however, proved to have certain limitations, due to the similar, or even identical retention factor of different phospholipids, even when the migrated lipid spots were revealed using different, lipid headgroup-classes selective staining dyes, such as molybdenum blue (dye which stains PG), or ninhydrin (stains PE and PS). Another drawback related to the lowered sensitivity of the method is the oxidation of the unsaturated lipids, due to the long exposure of the silica plate to the atmospheric oxygen (Fuchs *et al.*, 2011). In the case of mixtures of fatty acids taking part from the same class, the results referring to the impact of membrane restructuring on expression would be inconclusive.

Additionally, if the determination of the lipid profile in the case of both designs would follow the imposed hypotheses, this would represent a further step in the biocompatible *de novo* design and understanding of insertion and folding, and testing whether the presence of these lipids is necessary for the insertion, folding, or if the designed sequence is independent in this sense, but requires the lipids for the function, as it was shown in the case of LacY (Harris and Booth, 2012). As a comparison between the potential lipids requirements by the *de novo* designs, although the CC1 is situated in the developing design phase (comprises only the design of the inter-helical interfaces and the helix-lipid interface), REAMP would require POPE and cardiolipin for high stability and function, whilst CC1 would not necessarily require the presence of a specific lipid for stability. However, the results seen from the reconstitution in liposomes with POPC: POPG bilayers, would further suggest the possible preference of CC1 to insert in bilayers presenting negative charges in the lipids head group region. Determining the lipid affinity, it was

shown that this can contribute to the association between the sequence and the capability of insertion in the membrane depending on packing and curvature (Huong T Kratochvil *et al.*, 2021).

4. Cell-free expression of *de novo* designs and reconstitution into droplet interface bilayers

4.1. Introduction

4.1.1. Cell-free protein expression

Cell-free protein expression has emerged as a powerful tool for the production of membrane proteins, offering several advantages over traditional cell-based expression systems. This technique involves the use of cell lysates or purified components from cells to carry out transcription and translation processes *in vitro*, allowing for the rapid and efficient synthesis of proteins without the need for living cells. This approach is especially useful to synthetic biology as it has the potential to overcome the challenges associated with expressing membrane proteins in living cells, such as toxicity, low yields, and difficulties in purification (Sachse *et al.*, 2014).

One of the primary advantages of cell-free protein expression is its ability to produce proteins that are toxic to living cells e.g. channels. Membrane proteins, in particular, often require specific lipid environments and can cause cell death when overexpressed *in vivo* (Bernhard and Tozawa, 2013). Cell-free systems circumvent this issue by providing a controlled environment for protein synthesis, where the necessary components for transcription and translation are supplied in the reaction mixture. This approach allows for the expression of membrane proteins that might otherwise be impossible to produce in living cells, enabling their study and characterisation (Sachse *et al.*, 2014). Furthermore, cell free approaches have the ability to be used as the basis for artificial cells, as well as immense potential in synthetic biology. As any technique, it presents advantages and disadvantages, when compared to the common cell-based protein expression. In Table 4.1 there are highlighted the main advantages identified for the cell-free expression (Reckel *et al.*, 2010), as well as the downfalls that could interfere with certain project purposes, in comparison to cell-based expression (Pedro, Queiroz and Passarinha, 2019).

Table 4.1. Advantages and disadvantages of cell-free membrane protein expression and cell-based expression

	CELL-FREE EXPRESSION	**CELL-BASED EXPRESSION**
Toxicity	Reduced	Known to be associated with MP expression in cells
Production time	Fast, can be completed in a few hours, depending on the kit used	Long, can take from few days to weeks, depending on the cell line that is used
Cost	Moderate, depending on the amount of expression reactions	Cheaper, but can vary depending on the cell type

	needed (suitable for small-scale production)	used. For example, expression in *E. coli*, it is cheaper than CFE, however, expression in mammalian cells could be price comparable to CFE
Protein synthesis	Generally good yield, very little contamination in CFE kit components; Cell-extract-based CFE could contain nonspecific enzymes such nucleases or proteases that affect the synthesis	Could be affected by the presence of proteases
Protein membrane insertion	CFE allows flexibility in the direct manipulation of the environment (allows addition of membrane mimics, or detergents) and the control over the insertion by manipulation of the lipid composition for downstream reconstitution	Cell-systems can impose difficulties to the membrane insertion, multiple variables involved
Folding	The system allows flexibility over the direct control of the factors that can affect protein folding (pH, redox state etc.)	Can lead to protein aggregation and loss of function

Furthermore, cell-free protein expression systems can be easily adapted to incorporate membrane proteins into liposomes or other lipid bilayer systems. By supplying lipids or pre-formed liposomes to the reaction mixture, membrane proteins can be directly expressed and inserted into the lipid bilayer, providing a native-like environment for their study (Sachse *et al.*, 2014). This approach has been successfully applied to a wide range of membrane proteins, including G protein-coupled receptors, ion channels, and transporters, enabling their functional characterisation and the investigation of their interactions with ligands or other proteins (Khambhati *et al.*, 2019).

Cell-free expression can be performed using either cell-free extracts or commercially available kits. Latest studies have shown a specific interest, especially in the use of kits, enhancing the speed of the study and avoiding the complexity associated with the process of the cell-based extract. As all experiments are unique, they require constant and thorough optimisation to find suitable expression conditions. The commercial kits contain protein synthesis organelles extracted from different systems, such as *E. coli* cells, tobacco cell lysate, wheat germ, or HeLa cells. In Table 4.2 it is presented a comparison between different commercially available cell-free expression kits, with a special focus on the ones corresponding to *E. coli*

expression systems.

Table 4.2. Comparison between some of the commercially available cell-free expression kits

Kit name (Manufacturer)	Expression system	Yield (mg/ml)	Protein synthesis time (hours)	Price (GBP)* *reported in 2024	Compatibility with downstream processes
NEBExpress Cell-Free E. coli Protein synthesis system (New England Biolabs)	*E. coli*	0.5	3	150	Compatible with SDS-PAGE and Western blot, offers high expression levels, ability to produce high molecular weight proteins, scalability and cost-effective for high-throughput expression
PURExpress In Vitro Protein Synthesis Kit (New England Biolabs)		0.1	3	296	Contains purified components required for *E. coli* transcription and translation, compatible with studies for expression and folding; For MP expression needs supplementation with lipids and buffer optimisation
Magic Cell-Free PURExpress Protein Expression kit (Creative biolabs)		0.75	3	-	Confirmation of open reading frames, examinations of the effects of mutations on ORFs, identify active domains and functional residues; Compatible with HTS/directed evolution, synthetic biology, studies on protein folding, activity and protein-protein interactions
Expressway		0.8	2		Suitable for HTS

(Invitrogen)					experiments and expressing genes toxic to in vivo systems. For MP expression needs supplementation with MembraneMax reagent that ensures proper folding and helps with the nickel-affinity chromatography

4.1.2. Droplet-interface bilayers (DIBs)

Despite the advancements in cell-free protein expression and the use of liposomes and nanodiscs for membrane protein incorporation, there remains a significant knowledge gap in understanding the folding, stability, and function of *de novo*-designed membrane proteins. These proteins, which are engineered from scratch to have specific structural and functional properties, often pose unique challenges in terms of expression, purification, and characterisation. Traditional cell-based expression systems may not be suitable for producing these proteins, and the lack of a natural evolutionary context makes it difficult to predict their behaviour in artificial membrane environments. Moreover, the complex interplay between the protein and the lipid bilayer, as well as the influence of bilayer properties on protein folding and function, is not yet fully understood for *de novo*-designed membrane proteins.

This chapter aims to bridge this knowledge gap by investigating the cell-free expression of *de novo*-designed membrane proteins CC1 and REAMP and their reconstitution into droplet interface bilayers (DIBs). By employing a cell-free expression system, we eliminate the limitations associated with cell-based expression and enable the rapid and efficient synthesis of these proteins in a controlled environment. The use of liposomes with various lipid compositions allows us to systematically investigate the influence of bilayer properties on protein folding, stability, and insertion efficiency. Furthermore, by reconstituting the cell-free expressed proteins into DIBs, we create a platform for studying their function and interactions in a biomimetic membrane environment. This approach enables us to gain new insights into the behaviour of *de novo*-designed membrane proteins and their relationship with the lipid bilayer, ultimately advancing our understanding of membrane protein engineering and synthetic biology.

Droplet interface bilayers (DIBs) have emerged as a powerful platform for studying membrane proteins in a controlled and biomimetic environment. The approach uses the formation of a lipid bilayer at the interface between two aqueous droplets submerged in an oil phase, providing a unique system for the reconstitution and characterisation of membrane proteins (Bayley *et al.*, 2008). The ability to create stable bilayers with well-defined compositions and the potential for high-throughput screening make DIBs an attractive tool for investigating the structure, function, and interactions of membrane proteins (Leptihn *et al.*, 2013).

In Figure 4.1, the general concept of droplet interface bilayer formation is represented. Monolayer lipid-coated aqueous droplets, when placed in an oil environment (typically n-hexane, undecane, or silicone oil), have the aspect of inverted micelles, with the hydrophobic tails oriented towards the oil phase. The droplets can be formed by manual or robotic micro pipetting, resulting in droplets of size greater than 100μm. Smaller droplets are usually obtained using microfluidic devices, usually the T-junction type. In this case, the droplets are obtained by breaking an aqueous flow by a perpendicular oil flow. In DIB systems, the formed droplets can be pushed into place using a micropipette, to allow the DIB formation. Other methods involve the use of electrodes linked to micromanipulators, which can then bring the droplets into contact and form the bilayer. The latter method is more expensive and complex; however, it allows a better study of the bilayer formation and surface contact area (Villar and Bayley, 2013).

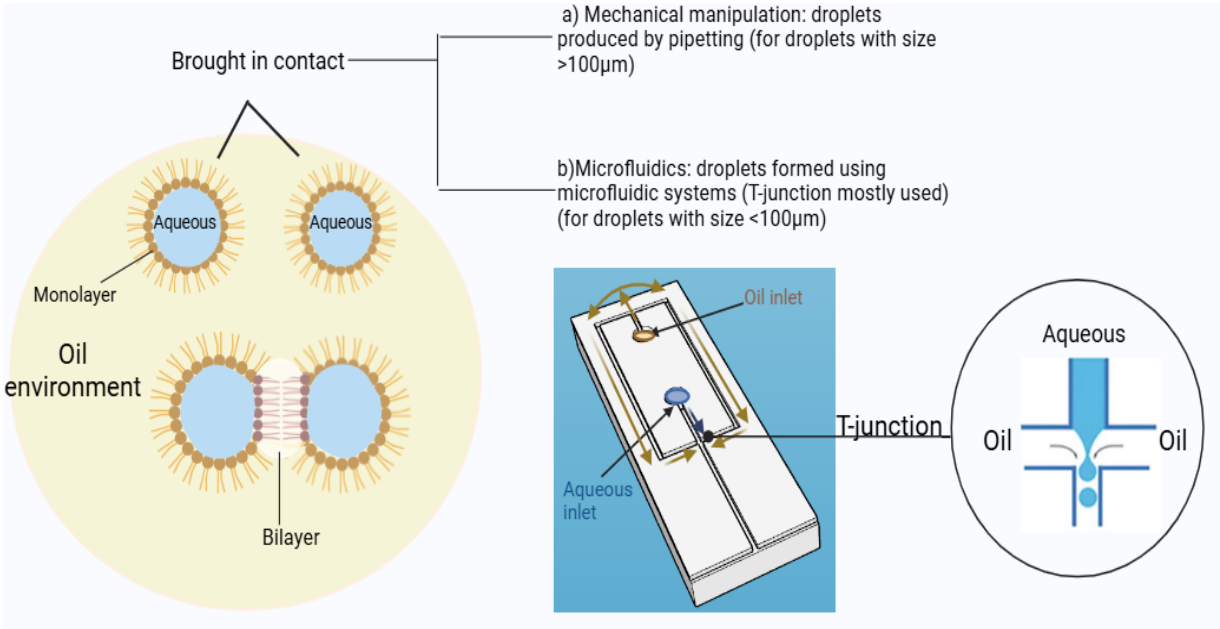


Figure 4.1. General droplet interface bilayer formation. Two monolayer lipid-coated aqueous droplets can be obtained from pipetting, resulting in droplets of sizes greater than 100µm, or small-sized droplets from the use of

microfluidic devices. A T-junction microfluidic device is usually used to obtain droplets. Within the junction, an aqueous flow is broken by an oil flow, each of varying speeds. (Created with BioRender)

One of the key advantages of using DIBs for studying membrane proteins is the ability to form stable bilayers with precisely controlled lipid compositions. By carefully selecting the lipids used to form the monolayers surrounding the aqueous droplets, researchers can create bilayers that mimic the composition of specific cell membranes, or investigate the effects of lipid composition on protein function (Venkatesan *et al.*, 2015). This level of control over the bilayer environment is particularly important for membrane proteins, as the lipid composition can significantly influence their stability, folding, and activity (Sezgin and Schwille, 2012). Moreover, the DIB platform allows for the formation of asymmetric bilayers, where each leaflet of the bilayer has a different lipid composition, enabling the study of protein-lipid interactions in more physiologically relevant contexts (Hwang *et al.*, 2008).

Another significant advantage of DIBs is the potential for high-throughput screening of membrane proteins. The small volume of the aqueous droplets (typically in the range of microliters) and the ability to generate arrays of droplets using 3D printed devices enable the rapid and parallel analysis of multiple protein variants or conditions (Venkatesan *et al.*, 2015). This high-throughput capability is particularly valuable for directed evolution experiments, where large libraries of protein mutants can be screened for desired properties, such as improved stability, altered substrate specificity, or enhanced activity (Freeman *et al.*, 2015). Additionally, the DIB platform is compatible with a wide range of analytical techniques, such as electrical measurements, fluorescence imaging, and mass spectrometry, allowing for the comprehensive characterisation of membrane protein function and interactions (Czekalska *et al.*, 2015).

The DIB approach has been successfully applied to study a diverse range of membrane proteins, including ion channels, transporters, and receptors. For example, the mechanosensitive channel of large conductance (MscL) has been reconstituted into DIBs, allowing for the investigation of its gating mechanism and response to membrane tension (Trantidou *et al.*, 2017). Similarly, the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel has been studied using DIBs, providing insights into its gating and conductance properties (Czekalska *et al.*, 2015). The ability to reconstitute and study these proteins in a well-defined and controlled environment has greatly advanced our understanding of their structure-function relationships and potential as drug targets.

Additionally, the use of DIBs as a reconstitution platform also represents a novel method for the study of the folding process and function of *de novo*-designed proteins.

In Figure 4.2 there are illustrated some of the studied uses of droplet interface bilayers and how various compartmentalisation can be achieved from DIB networks. DIBs can be used to screen different

compounds against membrane proteins. For example, Holden and co-workers (Holden, Needham and Bayley, 2007) screened a series of cyclodextrins against αHL. DIBs were used as high-throughput screening method using a chip format to study different blocking agents against the Kcv potassium channel, by Syeda and colleagues (Syeda *et al.*, 2008). As presented in panel A, the chip brings into contact series of droplets, one that contains the protein that needs to be studied, and the other, the analyte of which influence on the protein's mechanism needs to be assessed. When the droplets are brought into contact and the bilayers is formed, the membrane protein inserts and the transport mechanism can be analysed from electrical recording. Transport activity of reconstituted proteins can be assessed in DIBs, as it was the case of the MscL channel (Haylock *et al.*, 2020). Also, membrane proteins can be used to purposefully functionalise DIBs in different architectures, to obtain functional nanobiodevices. In panels B and D there are highlighted examples of such devices. Bacteriorhodopsin was used as a light sensor in a DIB network formed of five droplets (Holden, Needham and Bayley, 2007). Three of them contained the BR and were linked to a common electrode and to another droplet network in which the DIB contained the αHL heptamer. The electrical recordings have shown changes in the current, when the light source was turned on and off. The electric signal was determined by the amount of protons transported by the BR across the membrane, dependent on the amount of photons of light that was absorbed. Optical nano-device was also obtained in studies of Friddin et. al, which showed that DIBs represent an easy platform to manipulate the bilayer asymmetry across the bilayer and the influence of the lipid composition on the preference of membrane protein insertion and activity, as proven in studies on the calcium transport across αHL (Friddin *et al.*, 2016).

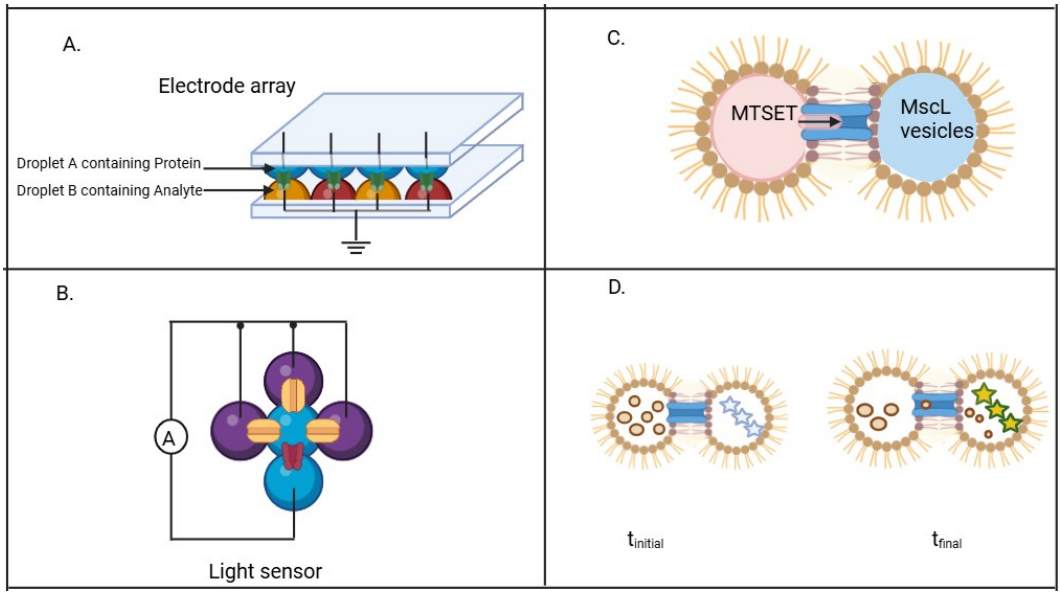


Figure 4.2. Different applications of droplet interface bilayers. A. Screening chip for membrane proteins, containing

on side droplets containing the membrane proteins and on the other side different analytes. When droplets are put in contact, the bilayer is formed, membrane protein inserts in the bilayer and the activity can be assessed through electrical recording. B. Droplet networks can be used as light sensors through electrical recording. C. Assessment of MscL transport activity in DIBs. D. DIBs as optical networks. Abstractisation of the calcium ions transport across αHL nanopore inserted into droplet interface bilayer into the adjacent droplet that contained EDTA and FLuo-4 dye, obtaining fluorescence, confirming the protein's activity in the bilayer.(Created with BioRender)

4.1.2.1. Bilayer stability

The stability of droplet interface bilayers (DIBs) formed with liposomes containing cell-free expressed proteins is a critical factor in the study of membrane protein function and interactions using this innovative platform. DIBs are a versatile and powerful tool for investigating membrane proteins in a controlled and biomimetic environment, offering several advantages over traditional reconstitution methods, such as the ability to form stable bilayers with well-defined compositions and the potential for high-throughput screening (Bayley *et al.*, 2008). However, the successful application of DIBs in membrane protein studies relies heavily on the formation of stable bilayers that can withstand the rigours of experimental manipulation and maintain the structural and functional integrity of the incorporated proteins (Leptihn *et al*., 2013).

The stability of DIBs is influenced by a complex interplay of factors, including the composition and physicochemical properties of the lipid bilayer, the nature and concentration of the incorporated proteins, and the experimental conditions used for bilayer formation and characterisation (Tsuji et al., 2013). One of the key determinants of DIB stability is the composition of the lipid bilayer, which can significantly impact the bilayer's thickness, fluidity, and mechanical properties (Butt et al., 2010). The choice of lipid species and their relative proportions can be tailored to mimic the native environment of the incorporated proteins or to optimise the bilayer's stability and compatibility with the experimental setup (Maglia *et al.*, 2009). For example, the incorporation of lipids with longer and more saturated acyl chains, such as 1,2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC), has been shown to enhance the stability of DIBs by increasing the bilayer's thickness and reducing its permeability to ions and small molecules (Bayley *et al.*, 2008). Similarly, the inclusion of cholesterol or other sterols can modulate the bilayer's fluidity and mechanical properties, improving its resistance to rupture and facilitating the incorporation of membrane proteins (Zhang *et al.*, 2019). The presence of charged lipids, such as phosphatidylserine (PS) or phosphatidylglycerol (PG), can also influence the bilayer's electrostatic properties and interactions with the incorporated proteins, affecting the stability and functionality of the resulting DIBs (Syama *et al.*, 2022). The nature and concentration of the incorporated membrane proteins can also have a significant impact on the stability of DIBs. The size, hydrophobicity, and oligomeric state of the proteins can influence their

insertion efficiency and compatibility with the lipid bilayer, affecting the overall stability and integrity of the DIBs (Tsuji *et al.*, 2013). High protein concentrations or the incorporation of large, multimeric proteins can lead to bilayer destabilisation or rupture, while low protein concentrations may result in insufficient functional activity or difficulty in detecting protein-mediated events (Maglia *et al.*, 2009). Therefore, optimising the protein concentration and reconstitution conditions is crucial for achieving stable and functionally active DIBs.

The experimental conditions used for DIB formation and characterisation, such as the composition of the aqueous solutions, the applied voltage, and the temperature, can also significantly influence the stability of the bilayers (Tsuji *et al.*, 2013). The ionic strength and pH of the aqueous solutions can affect the electrostatic interactions between the lipid head groups and the incorporated proteins, modulating the bilayer's stability and permeability (Fenz and Sengupta, 2012). The application of voltage across the DIBs, which is often used to study the electrical properties of incorporated ion channels or transporters, can also impact the bilayer's stability by inducing electrostriction or electroporation effects (Sengel and Wallace, 2016). Temperature variations can alter the bilayer's fluidity and phase behaviour, affecting the stability and functionality of the incorporated proteins (Zhang *et al.*, 2019).

To assess the stability of DIBs and optimise the experimental conditions for membrane protein studies, a combination of optical imaging, electrical measurements, and spectroscopic techniques can be employed (Bayley *et al.*, 2008). Optical imaging methods, such as fluorescence microscopy or interferometry, can provide real-time information on the formation, size, and stability of DIBs, allowing for the monitoring of bilayer integrity and the detection of protein incorporation (Zhang *et al.*, 2019). Electrical measurements, such as capacitance and conductance recordings, can offer insights into the bilayer's thickness, permeability, and the functional activity of incorporated ion channels or transporters (Maglia *et al.*, 2009). Spectroscopic techniques, such as fluorescence resonance energy transfer (FRET) or surface-enhanced Raman spectroscopy (SERS), can be used to probe the structural and conformational changes of the incorporated proteins, providing valuable information on their stability and functionality within the DIBs (Fenz and Sengupta, 2012).

4.1.3. Designed proteins- model systems for understanding membrane protein structure and function

Designed proteins, such as REAMP (Lalaurie *et al.*, 2018), have emerged as valuable tools for studying the structure, function, and dynamics of membrane proteins in a controlled and systematic

manner, as well as generating designs with specific functions. These engineered proteins are created through the rational design of amino acid sequences, allowing researchers to introduce specific structural features, functional properties, or labelling sites that facilitate their study and application (Chung *et al.*, 2012). By carefully selecting the desired characteristics and incorporating them into the protein sequence, researchers can create customised membrane proteins that are optimised for specific experimental or technological purposes (Joh *et al.*, 2014).

One of the primary motivations behind the use of designed proteins is the ability to overcome the limitations associated with studying native membrane proteins. Many membrane proteins are difficult to express, purify, and manipulate due to their hydrophobic nature and complex structural organisation (Joh *et al.*, 2014). By designing proteins with simplified structures or enhanced stability, researchers can circumvent these challenges and obtain high-quality samples for structural and functional characterisation (Perez-Aguilar and Saven, 2012). For example, the REAMP protein (Lalaurie *et al.*, 2018) includes a redesigned extracellular loop that enhances its expression and stability, making it more amenable to structural and biophysical studies (Barth and Senes, 2016).

Another key advantage of using designed proteins is the ability to introduce specific functional properties or molecular recognition capabilities. By carefully selecting the amino acid sequences and incorporating functional motifs, researchers can create membrane proteins with tailored activities, such as ion channels with altered selectivity or gating properties or receptors with enhanced affinity for specific ligands (Schmidpeter *et al.*, 2023). This level of control over the protein's function enables the development of novel biosensors, drug-delivery systems, or therapeutic agents that exploit the unique properties of membrane proteins (Sowlati-Hashjin, Gandhi and Garton, 2022).

The use of designed proteins as model systems for studying membrane protein behaviour has also gained significant attention in recent years. By creating simplified and well-characterised proteins that mimic the essential features of native membrane proteins, researchers can investigate fundamental aspects of protein-lipid interactions, membrane protein folding and stability, or the mechanisms underlying protein function (Mahendran *et al.*, 2017). These model systems provide a controlled and reproducible platform for testing hypotheses and developing new experimental approaches, ultimately advancing our understanding of the complex world of membrane proteins (Perez-Aguilar and Saven, 2012).

4.1.4. 3D printing and microfluidics

The advent of 3D printing and microfluidic techniques has revolutionised the field of droplet

interface bilayer (DIB) research, enabling the creation of customised reservoirs and devices for high-throughput screening and automation of membrane protein studies (Czekalska *et al.*, 2015). These innovative approaches have greatly expanded the potential applications of DIBs, allowing for the rapid and parallel analysis of multiple protein variants or experimental conditions, as well as providing better control over bilayer composition and stability (Theberge *et al.*, 2010). By leveraging the precision and flexibility of 3D printing and microfluidics, researchers can now design and fabricate tailored platforms for DIB formation, overcoming the limitations of traditional manual methods and paving the way for more efficient and reliable membrane protein characterisation (Nguyen *et al.*, 2016).

3D printing, also known as additive manufacturing, has emerged as a powerful tool for creating customised reservoirs and devices for DIB formation (Zhang, 2019). This technology allows for the rapid prototyping and production of complex three-dimensional structures with high resolution and accuracy using a variety of materials such as polymers, resins, or hydrogels (Su, Wang and McAlpine, 2023). By designing and printing reservoirs with specific geometries, dimensions, and surface properties, researchers can optimise the conditions for DIB formation and stability, while also incorporating features that facilitate the integration of electrodes, sensors, or fluidic interfaces (Nguyen *et al.*, 2016).

For example, 3D-printed reservoirs with precisely controlled volumes and shapes can be used to generate arrays of uniform DIBs, enabling the high-throughput screening of membrane proteins or the investigation of bilayer properties under different experimental conditions (Czekalska *et al.*, 2015). The incorporation of micro-channels or perfusion systems into the 3D-printed devices can allow for the continuous exchange of aqueous solutions, facilitating the study of protein-ligand interactions or the modulation of bilayer composition (Su, Wang and McAlpine, 2023). Moreover, the use of transparent or conductive materials in the 3D printing process can enable the integration of optical or electrical detection methods, such as fluorescence microscopy or electrophysiological recordings, providing real-time information on protein function and bilayer stability (Zhang, 2019).

Microfluidic techniques have also proven to be invaluable for the automation and high-throughput analysis of DIBs (Theberge *et al.*, 2010). Microfluidic devices, which consist of micron-scale channels and chambers fabricated using soft lithography or other microfabrication methods, can precisely control the generation, manipulation, and analysis of droplets, allowing for the formation of large arrays of DIBs with well-defined compositions and sizes (Nguyen *et al.*, 2016). By automating the droplet generation and coalescence processes, microfluidic platforms can significantly increase the throughput and reproducibility of DIB experiments while also reducing the sample volumes and reagent costs (Czekalska *et al.*, 2015).

One of the key advantages of microfluidic devices for DIB research is their ability to create highly

controlled and dynamic environments for membrane protein studies (Su, Wang and McAlpine, 2023). For instance, microfluidic platforms can be designed to generate concentration gradients or to sequentially expose DIBs to different chemical or physical stimuli, enabling the investigation of protein-ligand interactions, conformational changes, or bilayer remodelling processes (Theberge *et al.*, 2010). The integration of valves, pumps, and other active components into the microfluidic devices can further enhance the control over the experimental conditions, allowing for the precise modulation of flow rates, pressures, or temperatures (Nguyen *et al.*, 2016).

Moreover, the combination of 3D printing and microfluidics can lead to the development of even more sophisticated and versatile platforms for DIB research (Theberge *et al.*, 2010). By 3D printing the microfluidic devices themselves or by incorporating 3D-printed elements into the microfluidic systems, researchers can create highly customised and integrated platforms that combine the benefits of both technologies (Nguyen *et al.*, 2016). For example, 3D-printed microfluidic devices with embedded electrodes or sensors can enable the simultaneous electrical and optical characterisation of membrane proteins while also providing precise control over the bilayer composition and stability (Czekalska *et al.*, 2015). The application of 3D printing and microfluidic techniques in DIB research has already led to significant advances in the study of membrane proteins, including the development of high-throughput screening assays for ion channels and transporters, the investigation of protein-lipid interactions and phase behaviour, and the creation of biomimetic membranes for drug discovery and biosensing applications (Zhang, 2019). As these technologies continue to evolve and mature, it is expected that they will play an increasingly important role in unlocking the full potential of DIBs as a powerful and versatile platform for membrane protein research (Nguyen *et al.*, 2016).

4.1.5. Impact of material surface modifications on DIB stability

The development of alternative materials and fabrication methods for creating reservoirs and devices for droplet interface bilayer (DIB) formation has been a key focus in advancing this powerful platform for membrane protein studies (Nguyen *et al.*, 2016). While traditional DIB experiments have relied on the use of glass or plastic substrates, such as coverslips or Petri dishes, these materials often present limitations in terms of their surface properties, optical transparency, or compatibility with high-throughput screening and automation (Tsuji *et al.*, 2013). To overcome these challenges, researchers have explored the use of novel materials, such as polydimethylsiloxane (PDMS) and silicone elastomers, as well as innovative fabrication techniques, including soft lithography and replica moulding, to create customised

and optimised platforms for DIB formation and characterisation (Czekalska *et al.*, 2015).

PDMS, a silicone-based organic polymer, has emerged as a particularly attractive material for DIB research, due to its unique combination of properties, including its optical transparency, gas permeability, flexibility, and biocompatibility (Sia and Whitesides, 2003). PDMS can be easily moulded into complex three-dimensional structures with high fidelity and reproducibility using soft lithography techniques, such as micro-moulding or replica moulding (Duffy *et al.*, 1998). By designing and fabricating PDMS-based reservoirs and devices with specific geometries, dimensions, and surface chemistries, researchers can create highly optimised platforms for DIB formation and manipulation, while also enabling the integration of various analytical techniques, such as optical imaging or electrophysiological recordings (Bayley *et al.*, 2008).

For example, PDMS microfluidic devices with precisely controlled channel dimensions and flow profiles can be used to generate large arrays of uniform DIBs, allowing for the high-throughput screening of membrane proteins or the investigation of bilayer properties under different experimental conditions (Czekalska *et al.*, 2015). The incorporation of pneumatic valves or pumps into the PDMS devices can further enhance the control over the droplet generation and manipulation processes, enabling the automated and programmable formation of DIBs with well-defined compositions and sizes (Tsuji *et al.*, 2013). Moreover, the gas permeability of PDMS can be exploited to create hypoxic or anaerobic environments for studying the effects of oxygen tension on membrane protein function and stability (Sei *et al.*, 2014).

The surface chemistry of PDMS and silicone elastomers can be readily modified using various chemical or physical treatments, such as plasma oxidation, silanization, or polymer coating, to optimise their properties for DIB formation and stabilisation (Trantidou *et al.*, 2017). These surface modifications can be used to control the wettability, charge, or adhesion of the materials, promoting the stable and reproducible formation of DIBs with desired compositions and properties (Nguyen *et al.*, 2016). For example, the hydrophobicity of PDMS can be increased by silanization with alkyl-silanes, enhancing the stability of the lipid monolayers and preventing the adhesion of proteins or other biomolecules to the surface (Punnamaraju, You and Steckl, 2012).

In addition to PDMS and silicone elastomers, other materials, such as hydrogels or thermoplastics, have also been explored for DIB research (Leptihn *et al.*, 2013). Hydrogels, which are cross-linked networks of hydrophilic polymers, can be used to create biocompatible and transparent substrates for DIB formation, while also allowing for the encapsulation of cells or other biological components (Trantidou *et al.*, 2017). Thermoplastics, such as polycarbonate or poly(methyl methacrylate), can be moulded or 3D printed into complex structures with high precision and reproducibility, providing an alternative to PDMS

for the fabrication of microfluidic devices or high-throughput screening platforms (Czekalska *et al.*, 2015). The development of novel materials and fabrication methods for DIB research has greatly expanded the possibilities for studying membrane proteins in a controlled and biomimetic environment. By enabling the creation of customised and optimised platforms for DIB formation and characterisation, these approaches have significantly enhanced the throughput, reproducibility, and versatility of DIB experiments. As researchers continue to explore and refine these innovative materials and techniques, they will undoubtedly contribute to a deeper understanding of membrane protein function and interactions, ultimately facilitating the development of new therapeutic strategies and biotechnological applications.

4.1.6. Objectives of the study comprised in this chapter

The primary objectives of this chapter are two-fold. First, we aim to investigate the cell-free expression of the *de novo*-designed membrane proteins CC1-(S)sfYFP and REAMP-(S)sfYFP using a commercially available cell-free protein synthesis kit (PURExpress) and evaluate the influence of different liposome compositions on protein expression, folding, and insertion efficiency. Second, we seek to reconstitute the cell-free expressed proteins into droplet interface bilayers (DIBs) and assess the stability and functionality of these proteins in a biomimetic membrane environment. By achieving these objectives, we hope to gain insights into the behaviour of this case-specific *de novo*-designed membrane proteins and establish a platform for studying their structure-function relationships in a controlled and systematically variable lipid environment. To achieve these objectives, we employed a multidisciplinary approach combining cell-free protein synthesis, liposome preparation, and droplet interface bilayer formation. The CC1 and REAMP proteins were expressed using the PURExpress cell-free protein synthesis kit in the presence of liposomes with various lipid compositions, including POPC, POPC: DPhPC, POPC: POPE, POPC: POPG, and POPC: POPS. The expressed proteins were then purified and characterised using fluorescence measurements, SDS-PAGE, and Western blotting to assess their expression levels, folding, and insertion efficiency. To study the proteins in a biomimetic membrane environment, we reconstituted the cell-free expressed proteins into DIBs using a novel 3D-printed device and assessed the stability and functionality of the proteins using fluorescence microscopy and electrophysiological measurements. Furthermore, we investigated the influence of material surface modifications on DIB stability and optimised the experimental conditions to enhance the reliability and reproducibility of our results.

4.2. Materials and equipment

4.2.1. Lipid film preparation

The lipids were purchased from Avanti Polar Lipids Inc., U.S. Analytical grade chloroform solution was purchased from Fisher Scientific, UK. Glass round-bottom 5-ml flasks were purchased from Fisher Scientific. The rotary evaporator used was a Buchi R-100, coupled to a Buchi V-100 vacuum pump attached to a Buchi I-100 interface (Buchi Labortechnik, Germany).

4.2.2. Liposome preparation

There were used 0.4 µm and 1 µm polycarbonate Whatman membrane filters (Sigma-Aldrich, UK). The filter supports (10 mm) were purchased from Avanti Polar Lipids Inc., U.S..

4.2.3. Cell-free protein expression

UK.
The cell-free expression reaction kits (PURExpress kit) were purchased from New England Biolabs,

4.2.4. Droplet interface bilayer formation

The undecane and silicone oil AR-20 were purchased from Sigma-Aldrich UK. The polycarbonate Petri dishes were purchased from Corning, UK. Sterile 1.5 ml Eppendorf tubes and 1µl pipette tips were purchased from Starlabs, UK.

4.2.5. 3D Resin moulds

The model of the used 3D printer was Elegoo Mars 3. The 3D printing UV-LED curing resin used was Fast Clear resin, Siraya Tech. The rest of the chemicals were of analytical grade and purchased from Sigma-Aldrich, UK and Fisher Scientific, UK. The uncured resin washing and UV-light curing machine model was Anycubic Wash & Cure Plus. The used three-dimensional computer-aided design (CAD) software wasSketchUp Make 2017. The software used for refining was Autodesk Netfabb. The computer-aided manufacturing (CAM) software was Chitubox.

4.2.6. Silicone negative moulds

A silicone rubber mixture (GP-3481-F-RTV Silicone Rubber) from Polycraft was used for silicone

negative moulds.

4.2.7. Polydimethylsiloxane (PDMS) moulds

The polydimethylsiloxane (PDMS Sylgard 184, Dow Corning) was purchased from Fisher Scientific, UK. The PDMS curing oven was a laboratory oven with natural convection (Memmert UN55).

4.2.8. Material coating

The coating chemicals dopamine hydrochloride, Tris-HCl, bovine serum albumin (BSA), and polyethylene glycol 20000 (PEG20K) were purchased from Sigma-Aldrich, UK and Melford, UK. The styrene-maleic acid 2000 polymer was manufactured by Cray Valley U.S. and hydrolysed in-house by lab members. The enamel paint was purchased from PlastiKote, UK. The acrylate-based (cyano-acrylate) adhesive formulation was available on the market as Gorilla Glue transparent super glue gel.

Visual inspection of the droplet-interface bilayer formation and stability

The recording and visual inspection for the DIB formation and stability were performed using a USB Microscope LED Camera (1000x 1920x1080P), which was compatible with the OTG View software (compatible with the Android operating system).

All the chemicals which were purchased and not added to the list were of analytical grade.

4.3. Methods

4.3.1. Lipid film preparation

The lipids were weighed and mixed in desired concentrations and molar rations for the lipid stock solutions: POPC stock concentration 40 mg/ml; POPC: DPhPC (1:1) 25 mg/ml; POPC: POPE (10:1) 10 mg/ml; POPC: POPG (10:1) 10 mg/ml; POPC: POPS (10:1) 10 mg/ml. The lipid mixtures were transferred in individual 5 ml round-bottom glass flasks, solubilised in 1 ml of chloroform and dried under vacuum using the rotary evaporator with the vacuum set at 484 mmHg, until no traces of liquid chloroform were observed in the flask. The dried lipid films were resuspended in 1 ml sodium phosphate buffer pH=7.4 and vortexed until a homogeneous mixture was obtained. The lipid solutions were extruded and used immediately.

4.3.2. Liposome preparation

The resuspended lipid mixtures were extruded using an Avanti mini-extruder coupled to two- 1 ml Hamilton syringes. The polycarbonate membranes and filter supports were incubated for approximately 5 minutes in a 50 mM sodium phosphate pH=7.4 solution prior to assembling the mini-extruder. The same solution was passed a few times through the 1 ml syringes (Hamilton, U.S.) prior to coupling to the extruder. The resuspended lipid solutions were passed through the extruder 21 times.

4.3.3. Cell-free protein expression

Cell-free protein expression reactions for CC1-(S)sfYFP and REAMP-(S)sfYFP were prepared using the PURExpress Protein Synthesis kit (New England Biolabs, UK), as indicated in the manufacturer's instructions (New England Biolabs, 2021b) and summarised in Table 4.3. The DNA templates were represented by the plasmid DNA obtained from miniprep, as presented in Chapter 3: pET28b(+)-CC1-(S)sfYFP and pET28b(+)-REAMP-(S)sfYFP. The DNA concentrations were: for pET28b(+)-CC1-(S)sfYFP it was 636.3 ng/µl ($A_{260/280}$=1.88), and respectively 389.9 ng/µl ($A_{260/280}$=1.86). For technical clarity, a volume of 2 µl was considered for the pET-28b(+)-CC1-(S)sfYFP and 3 µl for pET-28b(+)-REAMP-(S)sfYFP as the template DNA. The cell-free reactions were incubated for 3h at 37°C in a static incubator and mixed with liposome solutions until a final lipids concentration of 10 mg/ml (for POPC: POPC (1:1); POPC: POPE (10:1); POPC: POPG (10:1); POPC: POPS (10:1)), or 25 mg/ml (for POPC: DPhPC (1:1)). The reactions were then centrifuged at 100,000 g for 1 h (TLA 55 rotor, Beckman, U.S.).

One of the controls was represented by the DHFR, and the reaction was performed as indicated in the manufacturer's instructions. Another control was represented by the cell-free reaction mixed only with liposome solution, lacking the DNA template.

Experiments were performed with the liposome solution added at the start of the reaction, respectively, at the end of the reaction incubation.

Table 4.3. Cell-free protein expression reactions for incorporation of expressed proteins into liposomes. In the first column there are presented the mixture components and in the second column there are presented the volumes. The volume of the liposome solution is expressed as "x" due to the fact that the volumes varied depending on the initial liposome stock solutions.

Cell-free expression mixture component	**Volume (µl)**
PURExpress Solution A	10
PURExpress Solution B	7.4
Water	5.5
DNA template: pET-28b(+)-CC1-(S)sfYFP or pET-28b(+)-REAMP-(S)sfYFP or DHFR (+)control	 2 or 3 or 2
Liposome stock solution	X

4.3.4. Sucrose gradient liposome flotation

Sucrose solutions of different concentrations (60% and 20%) were prepared in 50 mM sodium phosphate buffer, pH=7.4. In a sterile 2 ml microcentrifuge tube, 400 µl of 60% sucrose solution was mixed with the cell-free expression reaction containing liposomes to reach a final concentration of 40% sucrose, then 900 µl of 20% sucrose solution was gently added to avoid phase mixing and supplemented with 100µl of liposome buffer (50 mM sodium phosphate buffer, pH=7.4). The tube containing the 3 phases was centrifuged at 180,000 x g for 3 hours. The solution was divided into phases and isolated separately for further analysis.

4.3.5. Fluorescence measurements

The isolated phases from sucrose gradient flotation experiments were loaded along with controls

into a polystyrene OptiPlate. The fluorescence values of each sample were measured using 485 nm $_{excitation}$/510nm $_{emission}$ filters (Mithras LB940, Berthold Technologies). The fluorescence data corresponding to the fractions, including the expressed proteins, were normalised against the controls.

4.3.6. SDS-PAGE and Western blotting

The detection of protein cell-free expression was determined by SDS-PAGE and Western blotting. Twenty microliters of the liposome-cell-free expression reaction fractions were used for the SDS-PAGE and Western blotting. The primary antibody (mouse anti-6x his antibody, Sigma-Aldrich, UK) concentration was 1:500, diluted in blocking solution (5% BSA in TBS-T), and the secondary antibody (anti-mouse HRP conjugated antibody, Merck, UK) concentration was 1:2000, diluted in blocking solution. The rest of the steps of the SDS-PAGE and Western blotting were performed as described in the methods section of Chapter 3.

4.3.7. Droplet-interface bilayers (DIB) formation

In a polycarbonate Petri dish, there were mixed silicone oil AR-20 and undecane in a 1:1 volume ratio. In the mixture, there were carefully pipetted 1µl droplets of the liposome-containing cell-free protein expression solutions fractions, in which it was identified the highest fluorescence and liposome-only solutions. The droplets were allowed to equilibrate in the silicone oil-undecane bath for approximately 5 minutes at room temperature. The cell-free reaction containing and liposome-only containing droplets were mechanically put into contact using two 1 µl pipette tips. The DIB formation workflow is illustrated schematically in Figure 4.3.

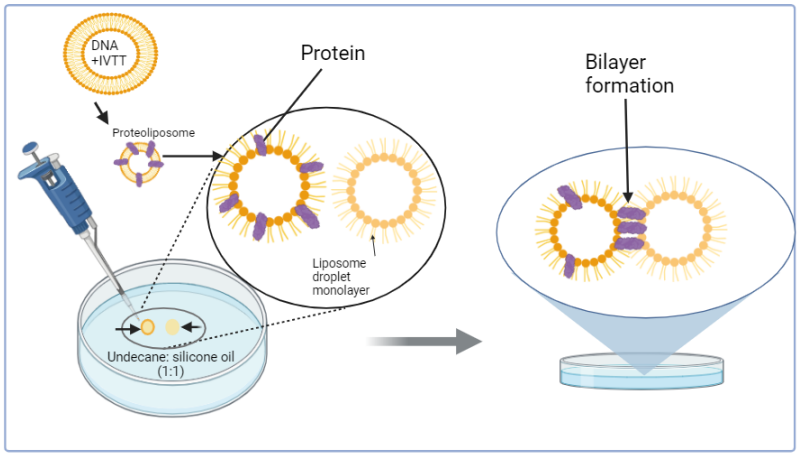


Figure 4.3. Schematic representation of theoretical DIB formation. (Left) In a dish containing a mixture of undecane and silicone oil in a ratio of 1:1, pipetted droplets of different composition are present: one droplet containing

liposomes (yellow circles), cell-free expression reaction components, and the expressed protein (purple icon), and another droplet containing a liposome-only solution. (Right) When brought in contact, the inverted lipid monolayers form a bilayer interface (droplet interface bilayer). (Created with BioRender).

4.3.8. Visual inspection of droplet-interface bilayers formation and stability

Droplet contact, bilayer formation, and bilayer stability were visually inspected using a microscope LED camera and recorded in real time using the OTG View software. The key events (bilayer formation, droplets coalescence/repulsion) were isolated in snapshots of the corresponding recordings.

4.3.9. 3D printing of vessels purposed for DIB formation

The 3D designs were obtained using three-dimensional computer-aided design (3D CAD) software (SketchUp Make 2017). The geometries and boundary edges were checked and repaired to reduce the design artefacts in the Autodesk Netfabb software. The repaired designs were imported into computer-aided manufacturing (CAM) software (Chitubox), compatible with the 3D resin printer (Elegoo Mars 3, Elegoo). Supports for the design were added automatically in Chitubox. In Chitubox, it was chosen the characteristic printing parameters (resin printing density, printing angle) for the type of used resin (Fast Clear resin, Siraya Tech). Prior to the printing start, there were performed essential inspection of the printer (level of resin in the printer tank, cleaning of building plate and LCD mirror screen from any remaining pieces of cured resin). Upon printing completion, the resin model was removed from the building plate and detached from the supports, cleaned under constant mixing in an isopropanol bath, to remove all residues of uncured resin. The resin-printed design was cured under 360° ultraviolet light (405 nm wavelength) exposure for 15 minutes (Anycubic Wash and Cure Plus Machine, Anycubic).

4.3.10. Preparation of silicone and PDMS moulds

4.3.10.1. Silicone negative mould

A silicone (GP-3481-F RTV Silicone Rubber, Polycraft) mixture was prepared according to the manufacturer's instructions in a ratio of 10:1 silicone base to silicone red fast catalyst. The mixture was degassed in a vacuum chamber for 10 minutes and poured onto a 3D-printed design. The silicone was cured at room temperature (T°=~20°C) for 16 hours.

4.3.10.2. Polydimethylsiloxane (PDMS) mould

The PDMS (Sylgard 184 elastomer, Dow Corning) mixture was prepared following the manufacturer's instructions in a ratio 10:1 elastomer base: elastomer curing agent and mixed vigorously until a homogeneous opaque mixture was obtained. The mixture was placed in a vacuum chamber for 10

minutes to remove any air bubbles. The degassed PDMS mixture was poured on a negative mould (resin/silicone moulds were pre-coated with the elastomer curing agent). The PDMS was cured in a laboratory oven with natural convection (Memmert UN55) at 65°C for 3 hours. After curing the PDMS mould was cleaned for impurities with isopropanol.

4.3.11. Materials surface chemistry modification

The resin and PDMS moulds' surfaces needed improved hydrophobicity. Hence, multiple coating substances for surface chemistry modification were tested.

4.3.11.1. PDMS coatings

Individual PDMS moulds were incubated overnight (approximately 16 hours) at room temperature in Bijou polycarbonate sterile 5 ml tubes (Starlabs, UK), containing dopamine hydrochloride (Sigma-Aldrich, UK) 2 mg/ml in 100 mM Tris-HCl (Melford, UK), pH=8.0 solution; bovine serum albumin (Melford, UK) 2 mg/ml in distilled water solution for 3 hours separately. Another PDMS mould was also coated with a thin, uniform layer of enamel paint using an airbrush spray gun attached to an air compressor (FD-18-2 mini air compressor, Fengda).

4.3.11.2. Resin coatings

Multiple coatings were tested on resin 3D-printed rectangular chips (2 x 2 x 1 cm). For this, there were prepared multiple solutions: dopamine hydrochloride (Sigma-Aldrich, UK) 2 mg/ml in 100mM Tris-HCl (Melford, UK) pH=8.0; bovine serum albumin (BSA, Melford, UK) 2mg/ml in distilled water; 2% hydrolysed styrene-maleic acid 2000 polymer dissolved in 50mM sodium phosphate buffer (Melford, UK) pH=8.0; 2% polyethylene glycol 20000 (PEG20K) (Sigma-Aldrich, UK) in hot (T°=˜100 °C) distilled water. Two copies of the printed chips were incubated with each solution for 16 hours at room temperature, under constant mixing, on a rocking shaker (SSL3 gyro-Rocker, Stuart). Two copies of the resin chips were coated with a thin layer of acrylate-based adhesive and let to air dry for approximately 16 hours (overnight). The coated chips were separated into populations to be transferred separately into individual silicone oil and water baths and incubated for 24 hours at room temperature under constant mixing. The excess of water and silicone oil was removed from the coated resin surface. As controls there were also incubated two copies of the uncoated chips into individual silicone oil and distilled water baths.

4.3.12. Water droplet test for surface chemistry determination

For all the coated chips it was performed the water droplet test was to classify the type of resulting surface after chemical modification. A 1µl water droplet was placed on the coated surfaces. Using ImageJ software->Contact Angle plug-in, it was measured the contact angle between the tangent to the water droplet and resin surface. For contact angles smaller than 90°, the surface was considered hydrophilic; for angles greater than 90°, the surface was considered hydrophobic after incubation in both types of phases (silicone oil and water). Mixed surfaces were considered if there were obtained both smaller and greater than 90° contact angle values after incubation in different phases. In Figure 4.4 it is exemplified the determination of surface chemistry by water droplet contact angle in Image J.

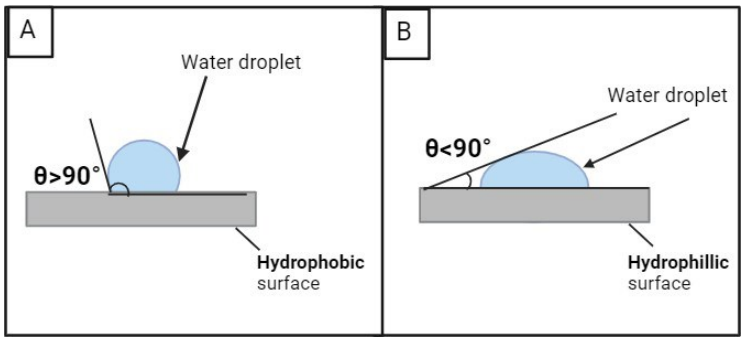


Figure 4.4. Examples of water droplet tests. A. Example of a hydrophobic surface. The contact angle (θ) formed by a tangent to the water droplet and the surface is greater than 90°, due to the shape adopted by the water droplet on the surface (more round) B. Example of a hydrophilic surface. The contact angle between the water droplet tangent and the surface is smaller than 90° due to the shape adopted by the water droplet (more flat, droplet contact surface is larger with a hydrophilic surface). (Created with BioRender)

4.4. Results

4.4.1. Fluorescence measurements

The cell-free expression reactions were mixed with liposomes of various compositions, as described in the Methods section. There were performed comparison cell-free expression reactions with liposomes added at the start and at the end of the incubation time of the cell-free expression reactions. The fluorescence was measured for the fraction containing the proteoliposomes after the sucrose gradient centrifugation at 180,000 g for 3 h. Data was compared between the time points of the addition of the corresponding liposomes. The liposomes that were used had the following compositions: POPC: POPC (1:1); POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPG (10:1); POPC:POPS (10:1).

In Figure 4.5, it is represented the graph containing the normalised fluorescence intensities data

obtained from excitation at 485 nm and emission at 510 nm (on the Y-axis), corresponding to the cell-free expressed sfYFP tag of the CC1-(S) in reactions mixed with the liposomes at the start (S), respectively, at the end (E) of the reactions. The data was normalised against the fluorescence measured for samples containing the components of the IVTT reaction mix and each liposome composition that was used in the experiments. On the X- axis, there is the plotted the grouped average data (N=3), depending on the type of liposomes added to the reaction. The fluorescence intensities measured for the reactions mixed with the liposomes at the start of the reactions did not exceed 20000 a.u. Depending on the liposomes composition used, the fluorescence intensities of the cell-free expressed CC1-(S)sfYFP decreased as follows: POPC: POPC (1:1); POPC: POPS (10:1); POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPG (10:1). Overall, the highest fluorescence intensities were observed for the reactions mixed with the liposomes at the end of the cell-free expression reactions. Among the liposome compositions, CC1-(S)sfYFP exhibited the highest fluorescence intensity when the reaction was mixed at the end of the reaction with POPC: POPE (10:1) liposomes (approximately 73369 a.u.), and decreasing for the rest of the liposomes compositions used as following: POPC: POPG (10:1); POPC: POPC (1:1); POPC: POPS (10:1); POPC: DPhPC (1:1).

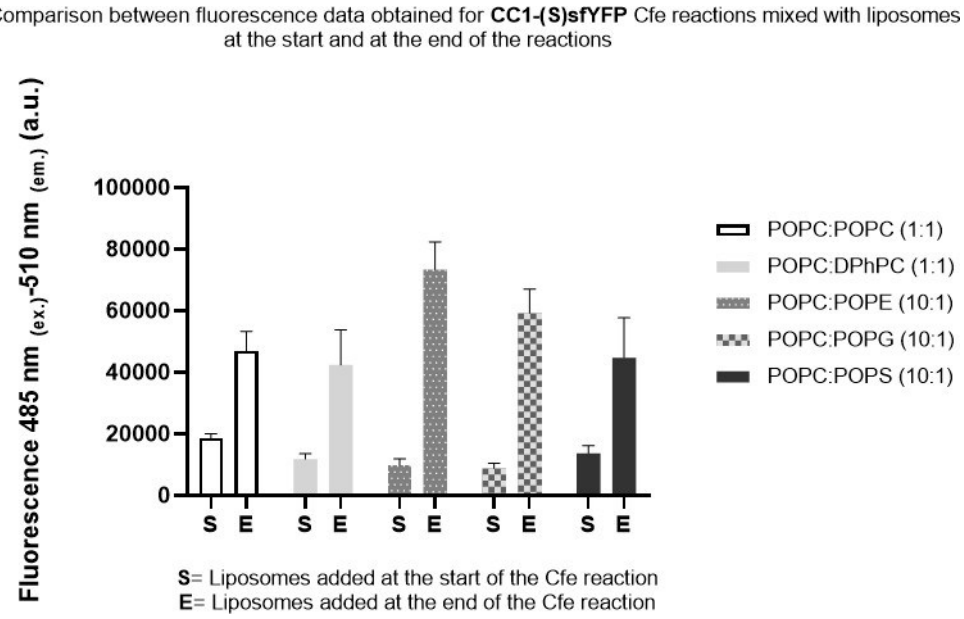


Figure 4.5. Normalised fluorescence data was obtained for CC1-(S)sfYFP from cell-free expression (CFE) reactions and incorporated in liposomes with various compositions after sucrose gradient (bottom fraction) in triplicates, analysed using GraphPad Prism 8 software. The error bars represent the standard deviations between the values across group replicates. The bar graph representation order (left to right) corresponds to fluorescence obtained for cell-free expressed CC1-(S)sfYFP in: **POPC: POPC** (1:1) (Start: Mean= 18588.3; SD=1479.368; N=3; **End**: Mean=46992; SD=6275.52; N=3); **POPC: DPhPC** (1:1) (Start: Mean=11788; SD=1788.96; N=3; **End**: Mean=42336.1; SD=11442.5; N=3); **POPC: POPE** (10:1) (Start: Mean=9764.31; SD= 2144.23; N=3; **End**: Mean=73369.6; SD=8987.01; N=3); **POPC:**

POPG (10:1) (Start: Mean=8789.23; SD=1657.23; N=3; **End**: Mean=59225.9; SD=7855.51; N=3); **POPC: POPS** (10:1) (Start: Mean=13689; SD=2398.23; N=3); **End**: Mean=44723.2; SD=13020.7; N=3).

Figure 4.6 comprises a graph in which there are represented the corresponding normalised fluorescence intensities (Y-axis) observed for the cell-free expressed REAMP-(S)sfYFP in reactions mixed with the liposomes of various compositions at the start and at the end of the reactions. On the X axis there is the grouped average normalised data depending on the type of liposome composition used for each reaction. The liposome compositions used were identical to the ones presented for CC1-(S)sfYFP. Similar to the graph observed in the previous figure, the fluorescence intensities obtained for the cell-free expressed REAMP-(S)sfYFP in reactions mixed with the liposomes at the start (S) of the reactions were lower than the ones having liposomes added at the end (E) of the reactions. The (S) fluorescence intensities observed were not exceeding in this case 25000 a.u. and were decreasing depending on the liposome compositions added to the mixture, as following: POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPE (10:1); POPC: DPhPC (1:1); POPC:POPC (1:1). For the (E) fluorescence intensities, the decreasing order depending of the liposomes composition mixed at the end of the reactions is the following: POPC: DPhPC (1:1); POPC: POPE (1:1); POPC: POPS (10:1); POPC: POPC (1:1); POPC: POPG (10:1). Comparing to the high differences between absolute values observed for the average fluorescence intensities of the cell-free expressed CC1-(S)sfYFP, the differences between the (S) and (E) of cell-free expressed REAMP-(S)sfYFP, when POPC: POPG (10:1), respectively POPC: POPS (10:1) were used, were of approximately 15000 a.u.

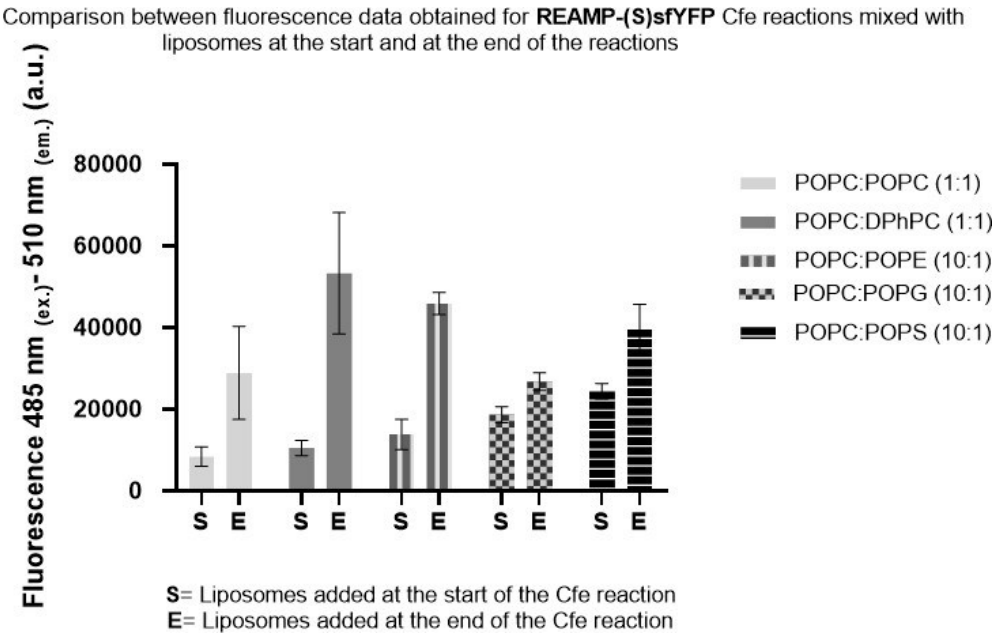


Figure 4.6. Normalised fluorescence data was obtained for REAMP-(S)sfYFP from cell-free expression (CFE) reactions and incorporated in liposomes with various compositions after sucrose gradient (bottom fraction). The error bars represent the standard deviations between values across group replicates, analysis performed using GraphPad Prism 8 software. The bar graph representation order (left to right) corresponds to fluorescence obtained for cell-free expressed REAMP-(S)sfYFP in **POPC: POPC** (1:1) (Start: Mean=8389.9971; SD=2346.85; N=2; **End**: Mean= 28889.49; SD=11384.28733; N=2); **POPC: DPhPC** (1:1) (Start: Mean=10506.6744; SD=1867.98; N=3; **End**: Mean=53288.56006; SD=14879.70296; N=3); **POPC: POPE** (10:1) (Start: Mean=13826.28979; SD= 3699.3547; N=3; **End**: Mean=45857.54757; SD=26969.43258; N=3); **POPC: POPG** (10:1) (Start: Mean= 18697.76173; SD=1967.35; N=2; **End**: Mean= 26770.35192; SD=2117.734736; N=2); **POPC: POPS** (10:1) (Start: Mean=24391.18079; SD=1899.657; N=2; **End**: Mean=39547.05092; SD=6165.980531; N=2).

In Figure 4.7 it is illustrated the graph comprising comparative data of the normalised grouped average fluorescence intensities (on the Y-axis) obtained for the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP in the reactions mixed with liposomes of various compositions at the end of the reactions (on the X-axis). It could be observed that cell-free expressed CC1-(S)sfYFP exhibited higher fluorescence intensities, compared to REAMP-(S)sfYFP, when obtained in reactions mixed at the end of the incubation time with most liposome compositions, excepting POPC: DPhPC (1:1), when REAMP-(S)sfYFP was more fluorescent than CC1-(S)sfYFP. It could be observed for the fluorescence intensities recorder for both cell-free expressed proteins and mixed with POPC: DPhPC (1:1) liposomes that the proteins are found in an antithesis relationship. More specifically, for this liposome composition, it was observed that CC1-(S)sfYFP exhibits the lowest fluorescence intensity and REAMP-(S)sfYFP exhibits the highest intensity from their groups.

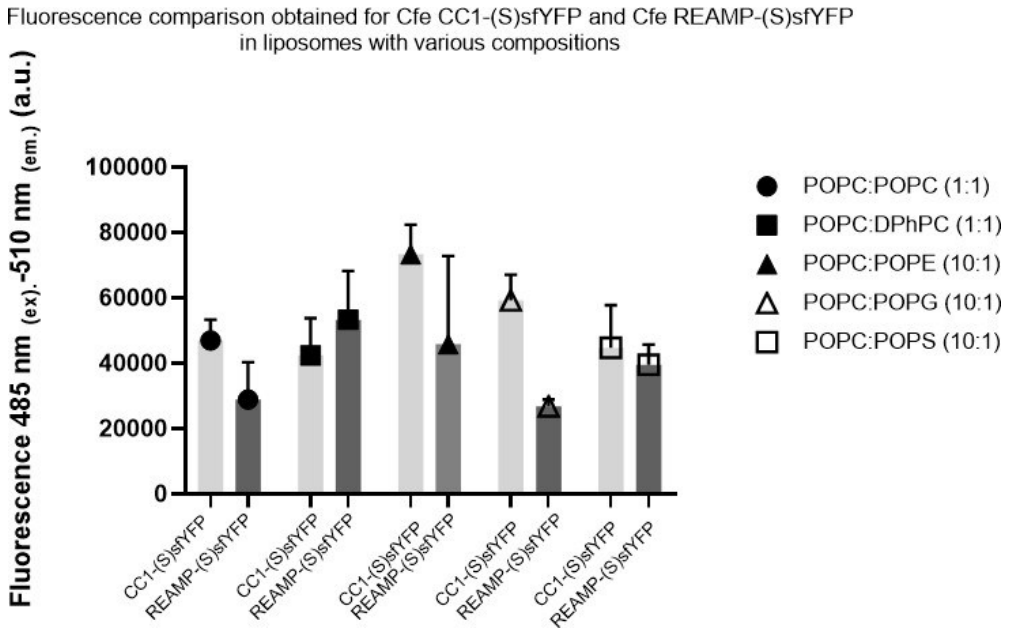


Figure 4.7. Comparison between the normalised fluorescence data obtained for the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP (represented as dark grey bars) incorporated in liposomes of various compositions. The normalised fluorescence intensity values corresponding to CC1-(S)sfYFP are represented as light grey bars and for REAMP-(S)sfYFP as dark grey bars. The error bars represent the standard deviations for the measurements across group replicates, analysis performed using GraphPad Prism 8 software. The fluorescence data is grouped in series of two bars (CC1-(S)sfYFP and REAMP-(S)sfYFP), depending upon the liposome composition in which the proteins were reconstituted and are represented as following: POPC: POPC (1:1) (black filled circle); POPC: DPhPC (1:1) (black filled square); POPC: POPE (10:1) (black filled triangle); POPC: POPG (10:1) (white triangle); POPC: POPS (10:1) (white square). Across all data, CC1-(S)sfYFP tended to exhibit higher fluorescence compared to REAMP-(S)sfYFP, with the highest values observed when incorporated into POPC: POPE (10:1) liposomes.

4.4.2. Liposome insertion of the cell-free expressed proteins

It was also calculated, based on the fluorescence intensities and theoretical distribution in a sucrose gradient, the liposome insertion efficiencies of the presumably folded states of the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP. The top fraction of the sucrose gradient was considered to contain the empty liposomes, the middle fraction to contain the proteoliposomes and the bottom fraction to contain the aggregated protein.

Figure 4.8 comprises the comparison between the liposome insertion efficiencies of the presumably folded states of the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP in liposomes of various compositions, calculated based on the fluorescence data obtained for different fractions of the sucrose gradient.

Overall, they could be assigned higher insertion efficiencies for REAMP-(S)sfYFP in most of the liposome compositions, except POPC: DPhPC (1:1). The liposome compositions which could be correlated

to increasing liposome insertion efficiency for the cell-free expressed REAMP-(S)sfYFP were the following: POPC: DPhPC (1:1); POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPC (1:1); POPC: POPE (10:1). The liposome compositions which could be correlated to increasing values of the calculated liposome efficiency of the presumably folded state of the cell-free expressed CC1-(S)sfYFP were: POPC: POPE (10:1); POPC: POPG (10:1); POPC: POPC (10:1); POPC: POPS (10:1); POPC: DPhPC (1:1).

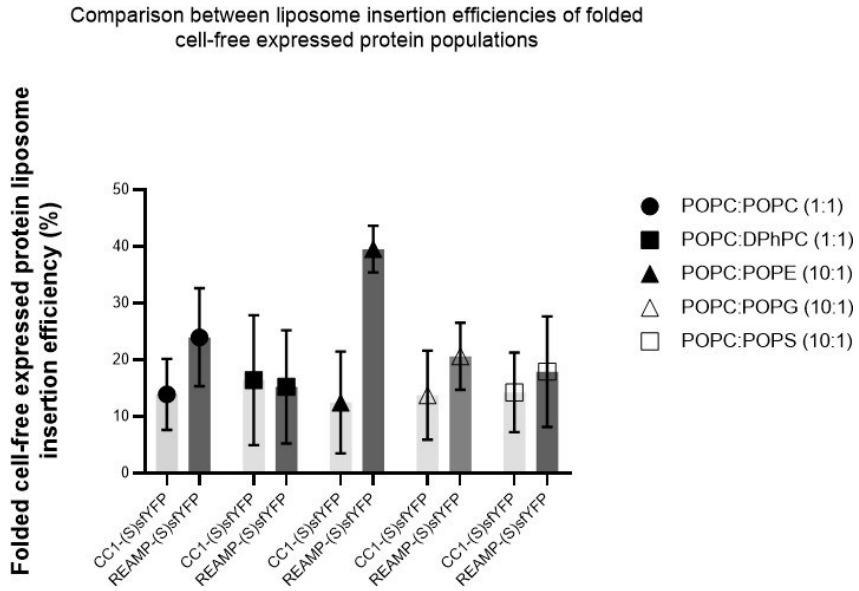


Figure 4.8. Liposome insertion efficiencies of the folded states of the cell-free expressed proteins CC1-(S)sfYFP and REAMP-(S)sfYFP. The data corresponding to CC1-(S)sfYFP is illustrated as light grey bars and for REAMP-(S)sfYFP as dark grey bars, analysis performed using GraphPad Prism 8 software. The **increasing** insertion efficiencies depending on the liposome compositions of the presumably folded states of the **CC1-(S)sfYFP** and calculated based on the fluorescence data obtained for the sucrose gradient fractions are the following: **POPC: POPE (10:1)** (Mean = 12.4575; SD= 8.9872; N=3); **POPC: POPG (10:1)** (Mean= 13.7397; SD=7.855; N=3); **POPC: POPC (1:1)** (Mean= 13.8830; SD=6.2750; N=3); **POPC: POPS (10:1)** (Mean= 14.2368; SD=7.0281; N=3); **POPC: DPhPC (1:1)** (Mean= 16.3877; SD= 11.4486; N=3). The **increasing** insertion efficiency values obtained for the presumably folded state of the cell-free expressed **REAMP-(S)sfYFP**, depending on the liposome compositions, are the following: **POPC: DPhPC (1:1)** (Mean= 15.211; SD=9.9662; N=3); **POPC: POPS (10:1)** (Mean= 17.9013; SD= 9.7454; N=3); **POPC: POPG (10:1)** (Mean= 20.6063; SD= 5.9328; N=3); **POPC: POPC (1:1)** (Mean= 23.9540; SD= 8.6515; N=3); **POPC: POPE (10:1)** (Mean= 39.5084; SD= 4.117; N=3).

4.4.3. Influence of the liposome compositions on the fluorescence and insertion of the CFE proteins

From the comparison between the data obtained for the fluorescence intensities exhibited by the cell-free expressed CC1-(S)sfYFP and its potential folded state to insert into liposomes of various compositions, it could be extracted a direct relationship between the order of the liposomes compositions

and their antithetic effect on the fluorescence intensities and the calculated liposome insertion efficiencies, as presented in the diagram from Figure 4.7.

In the Figure 4.9 there are presented the directions for the increase of the fluorescence intensities (opposite from the order of compositions) and the liposome insertion efficiencies (same as the order of compositions) of the cell-free expressed CC1-(S)sfYFP, based on the liposomes compositions presented in the following order: POPC: POPE (10:1); POPC: POPG (10:1); POPC: POPC (1:1); POPC: POPS (10:1); POPC: DPhPC (1:1). The data used for this diagram corresponded to the data presented for the cell-free expression reactions mixed with the liposomes at the end of the reaction's incubation period.

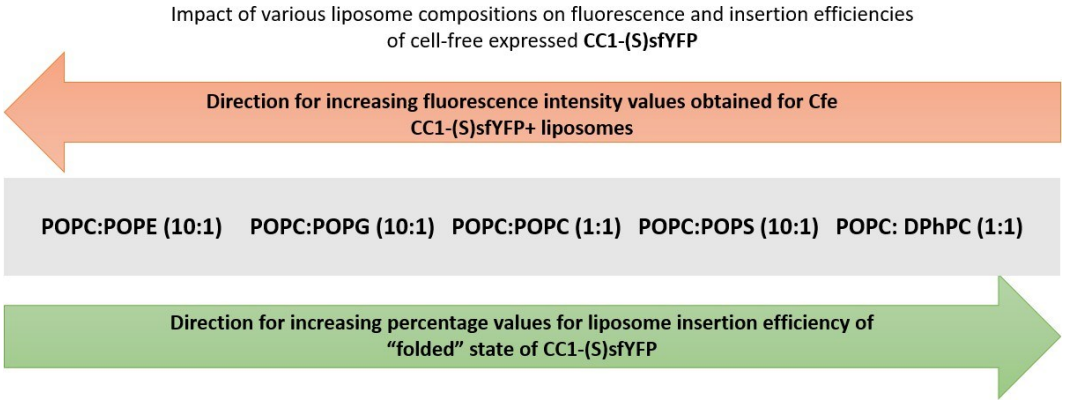


Figure 4.9. Direct relationship established between the liposome compositions and their impact on the fluorescence intensities and insertion efficiencies of the folded states of the cell-free expressed CC1-(S)sfYFP. The order of liposomes lipid compositions presented in the grey region (POPC: POPE (10:1); POPC: POPG (10:1); POPC: POPC (1:1); POPC: POPS (10:1); POPC: DPhPC (1:1)), corresponded to the order of increasing values obtained for the liposomes insertion efficiencies of the folded state of the cell-free expressed protein CC1-(S)sfYFP (increasing direction indicated by the green arrow). The liposome compositions presented in the order as in the grey box corresponded to the decreasing order observed for the fluorescence intensities measured for the cell-free expressed proteins with liposomes added at the end of the reaction (the direction for increasing fluorescence as function of liposome composition order is indicated by the orange arrow). For the liposome compositions order as presented in the grey box, it could be assigned an exact inverse proportional relationship for the fluorescence intensities and insertion efficiencies.

The data corresponding to the variability of the fluorescence intensity exhibited by the cell-free expressed REAMP-(S)sfYFP and the insertion efficiency is illustrated in the diagram presented in Figure 4.10. In comparison to the diagram comprising the analogue data corresponding to CC1-(S)sfYFP, for the cell-free expressed REAMP-(S)sfYFP, the increase of fluorescence intensity and the liposome insertion efficiency of the potential folded state, were influenced by two separate liposome compositions orders. One liposome composition was found to have an antithetic effect (POPC: DPhPC (1:1) was correlated to the highest fluorescence intensity and the lowest liposome insertion efficiency). The rest of the liposome

compositions that were correlated to the increase in the fluorescence intensity of the REAMP-(S)sfYFP were ordered, as following: POPC: POPC (1:1); POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPE (10:1); POPC: DPhPC (1:1). The order correlated to the increase of the liposome insertion efficiency was POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPE (10:1).

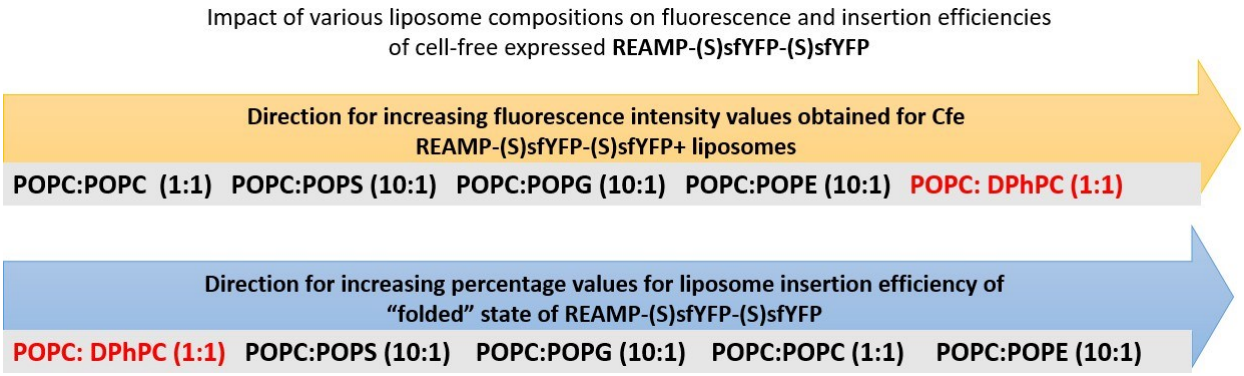


Figure 4.10. The impact of various liposome compositions on the fluorescence intensities of the cell-free expressed proteins with liposomes added at the end of the reactions and the insertion efficiencies of the folded states of the REAMP-(S)sfYFP. (Yellow arrow and grey box): The liposomes compositions order which impacted the increase in fluorescence intensities is POPC: POPC (1:1); POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPE (10:1); POPC: DPhPC. (Blue arrow and grey box): The liposomes compositions order which impacted the increase in liposome insertion efficiencies of the folded states of the cell-free expressed REAMP-(S)sfYFP intensities is: POPC: DPhPC (1:1); POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPC (1:1); POPC: POPE (1:1). Highlighted in red is the liposome lipids composition POPC: DPhPC (1:1), for which it could be assigned an exact inverse proportional effect on the fluorescence (highest values) and liposome folded protein insertion efficiency (lowest) for REAMP-(S)sfYFP.

Regarding the correlations established between the liposome compositions and the liposome insertion efficiencies of the folded states of the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP, a further correlation between the bilayer's properties and protein's insertion tendency could be established. The same data as for the previous diagrams was used, namely the data obtained for the cell-free expressed proteins from reactions with the liposomes added at the end of the reaction, eliminating the data corresponding for the POPC: POPC (1:1) and POPC: DPhPC (1:1) liposomes.

The comparison of the data is illustrated in the diagram presented in Figure 4.11. It could be observed a single liposome composition order correlated to opposed tendencies for liposome insertion efficiency of the cell-free expressed proteins. The order POPC: POPS (10:1); POPC: POPG (10:1); POPC: POPE (10:1) corresponded to a tendency for the increase of liposome insertion efficiency of the REAMP-(S)sfYFP and a decrease for the same of the CC1-(S)sfYFP. The presented order of the lipids would influence the gradual increase of the bilayer negative curvature.

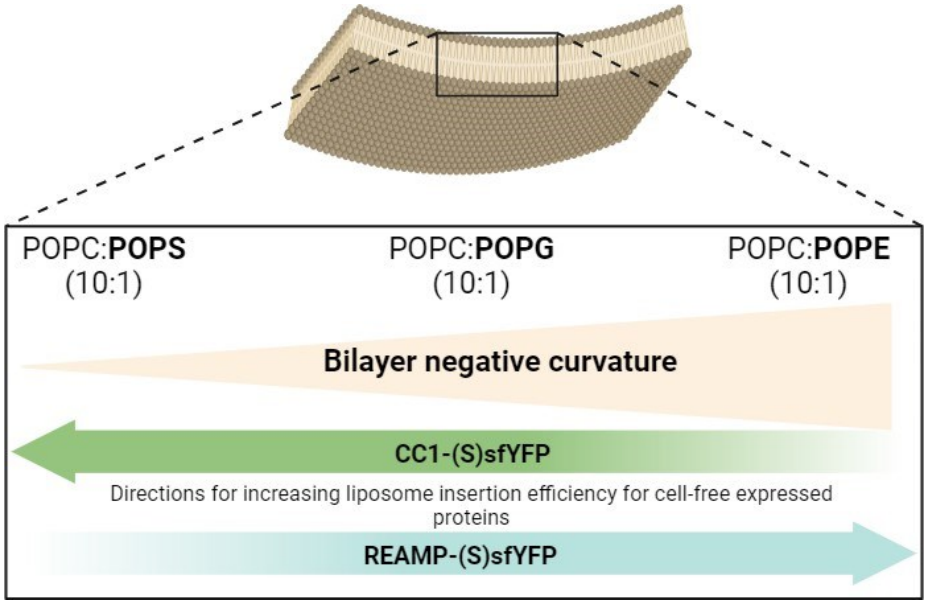


Figure 4.11. Relationship between the bilayer negative curvature and the insertion of folded states of the cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP. The liposome compositions were selected from the orders presented above for their impact on the liposome insertion efficiencies of CC1-(S)sfYFP and REAMP-(S)sfYFP, depending on the effect on the bilayer curvature. Direct relationships were observed for the order of liposome compositions: POPC: POPS (10:1), POPC: POPG (10:1), POPC: POPE (10:1), which corresponded to the increase in bilayer curvature. For this order, the liposome insertion efficiency for CC1-(S)sfYFP decreased, whilst for REAMP-(S)sfYFP increased.

4.4.4. Cell-free expression (CFE) of the CC1-(S)sfYFP and REAMP-(S)sfYFP

For the cell-free protein expression reactions it was used the commercial *in vitro* transcription and translation (IVTT) kit PURExpress (New England Biolabs). The cell-free protein expression reaction consisted of a mixture of a specified ratio between the kit components, the template DNA, nuclease-free water and liposome-buffered solution. Multiple liposome compositions were tested, such as POPC: POPC (1:1); POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPG (10:1), and POPC: POPS (10:1). It was also tested the influence of the liposomes on the transcription and translation efficiency of the kit components, by comparisons of the exhibited fluorescence of the translated proteins in reactions mixed with liposomes

at the start of the incubation time with the results obtained from reactions mixed with liposomes at the end of the reaction incubation time.

Control CFE reactions were performed in order to assess the viability of the kit components, to determine the background protein expression and the impact of liposomes on the reaction's components.

Figure 4.12 represents the electrophoresed polyacrylamide gel containing the separated protein bands obtained from the control cell-free protein expression reactions represented by the CFE of dihydrofolate reductase (DHFR control plasmid was provided in the PURExpress kit), and the CFE components mixed with a liposome solution. It could be observed a similar banding pattern between the two reaction (containing the DNA encoding the positive control protein, in this case DHFR, and the one containing only the cell-free expression reaction components and liposomes). As presented in the supporting product literature materials published by the manufacturer, the bands that migrate below 16 kDa represent the majority of the ribosomal proteins and all the other bands, except the protein of interest (i.e. DHFR at approximately 18kDa) represent other endogenous proteins present in the kit, as expressed in the product manual. Hence, in both cases it could be considered that the protein synthesis reaction was successfully completed.

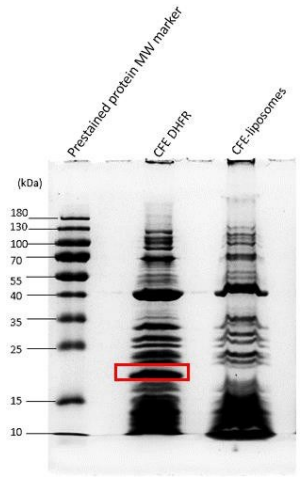


Figure 4.12. Electrophoresed polyacrylamide gels containing the cell-free protein expression control reactions. Lane Prestained protein ladder (Pierce 16616 prestained protein marker). Lane 2. Cell-free expressed DHFR (˜18 kDa, highlighted in red border). Lane 3. Cell-free expression reaction solutions mixed with liposome solution.

From the visualisation of the electrophoresed gel, it could be observed that the kit was successful in the synthesis of the control DHFR protein. It could be observed on the electrophoresed gel an intense protein band at the corresponding molecular weight of the DHFR protein (˜18 kDa, highlighted in the red border). The protein bands obtained for the CFE reaction containing liposomes only and no template DNA

were comparable to the literature results.

4.4.5. Cell-free expression of CC1-(S)sfYFP and REAMP-(S)sfYFP influenced by the addition timepoints of the liposomes

For the REAMP-(S)sfYFP CFE reactions, liposomes of different compositions were added at the start and at the end of the reaction's incubation time and there were observed differences for the CFE expression of REAMP-(S)sfYFP, as suggested by the intensity of the bands at the corresponding molecular weight, detected on an anti-his Western blot.

Figure 4.11 comprises anti-his blots of the electrophoresed samples containing REAMP-(S)sfYFP CFE reactions mixed with liposomes of different compositions (POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPS (10:1) at the start (S) of the CFE reaction incubation time and at the end (E) of the incubation (panel A). In Figure 4.13, panel B it is illustrated an anti-his blot comprising samples of the REAMP-(S)sfYFP CFE reactions mixed with liposomes of various compositions (POPC: POPC (1:1); POPC: POPG (10:1); POPC: POPE (10:1) and POPC: DPhPC (1:1)), at the end of the CFE reactions only. In both panels, there were observed protein bands at the corresponding molecular weight to the REAMP-(S) sfYFP (˜56.7 kDa) and were highlighted in red borders.

For the samples corresponding to the CFE of REAMP-(S)sfYFP from reactions mixed with the liposomes at the start of the incubation time, no bands or very low-intensity bands at the corresponding MW could be detected. Protein bands of various intensities, situated at the corresponding MW or REAMP-(S)sfYFP could be observed for the samples containing the CFE reactions mixed with liposomes at the end of the reactions. As it could be seen in Figure 4.13 panel B, bands with higher intensities situated at approximately 56.7 kDa could be detected in the samples containing the CFE mixed with POPC: POPC (1:1) and POPC: POPG (10:1) liposomes, compared to the other liposome compositions.

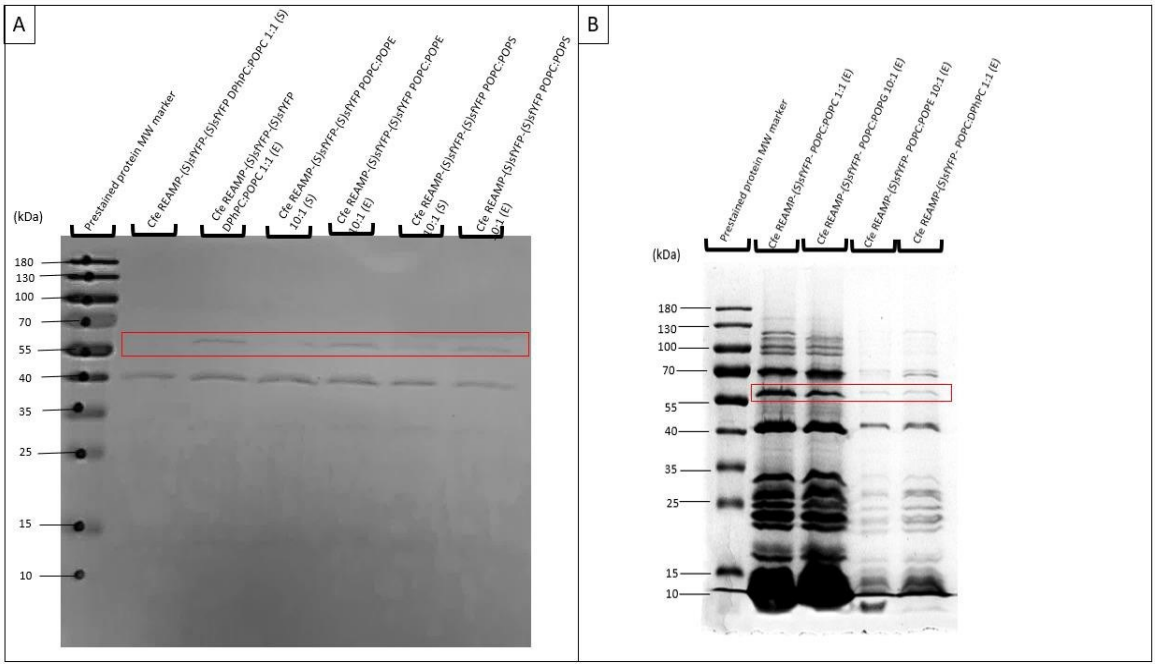


Figure 4.13. Anti-his western blots of cell-free expressed REAMP-(S)sfYFP and liposomes of various compositions. A. Comparison between expression in PURExpress cell-free system of REAMP-(S)sfYFP when liposomes of various compositions were added at the start and at the end of the reaction: Lane 1. Prestained protein MW marker. Lanes 2-3 CFE REAMP-(S)sfYFP with DPhPC: POPC (1:1) liposomes (Lane 2-liposomes added at the start of reaction; Lane 3-liposomes added at the end of the reaction); Lanes 4-5 CFE REAMP-(S)sfYFP with POPC: POPE (10:1) liposomes (Lane 4-liposomes added at the start of the reaction; Lane 5- liposomes added at the end of the reaction); Lanes 6-7 CFE REAMP-(S)sfYFP with POPC: POPS (10:1) liposomes (Lane 6- liposomes added at the start of the reaction; Lane 7-liposomes added at the end of the reaction). It could be observed that the intensities of the bands of the cell-free expressed REAMP-(S)sfYFP (~56.7 kDa) were corresponding to the reactions to which the liposomes were added at the end of the cell-free reactions. B. Comparison between expression in PURExpress cell-free system of REAMP-(S)sfYFP when liposomes of various compositions were added at the end of the reactions. In addition to panel A, panel B contains in Lane 2- CFE of REAMP-(S)sfYFP with POPC:POPC (1:1) liposomes added at the end of the reaction, in Lane 3. CFE of REAMP-(S)sfYFP with POPC: POPG (10:1) liposomes added at the end of the reactions. Lanes 4 and 5 represent replicates of the CFE reactions with POPC:POPE (10:1), respectively POPC: DPhPC (10:1) liposomes added at the end of the reactions. The bands of higher intensities for the CFE REAMP-(S)sfYFP (˜56.7 kDa) were obtained for reactions mixed with POPC: POPC (1:1) and POPC: POPG (10:1) liposomes. In panels A and B there are highlighted in red border the areas of interest that corresponds to approximately 56.7 kDa (the molecular weight of REAMP-(S)sfYFP).

Similarly, for the CC1-(S)sfYFP CFE reactions, liposomes of different compositions were added at the start and at the end of the reaction's incubation time, and there were observed similar differences for the CFE expression of CC1-(S)sfYFP.

Figure 4.14 comprises anti-his blot, stained with Ponceau stain, containing the samples corresponding to the CFE reaction for CC1-(S)sfYFP mixed with liposomes of different compositions (POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPS (10:1) at the start (S) of the CFE reaction incubation time and at the end (E) of the incubation (panel A).

In Figure 4.14, panel B it is illustrated an anti-his blot comprising samples of the CC1-(S)sfYFP CFE reactions mixed with liposomes of various compositions (POPC: POPC (1:1); POPC: POPE (10:1); POPC: POPS (10:1); POPC: POPG (10:1); POPC: DPhPC (1:1)), at the end of the CFE reactions only. In both panels, there were observed protein bands at the corresponding molecular weight to the REAMP-(S) sfYFP (˜56.7 kDa) and were highlighted in red borders.

For the samples corresponding to the CFE of CC1-(S)sfYFP from reactions mixed with the liposomes at the start of the incubation time, no bands or very low-intensity bands at the corresponding MW could be detected. Protein bands of similar intensities, situated at the corresponding MW or CC1-(S)sfYFP could be observed for the samples containing the CFE reactions mixed with liposomes at the end of the reactions.

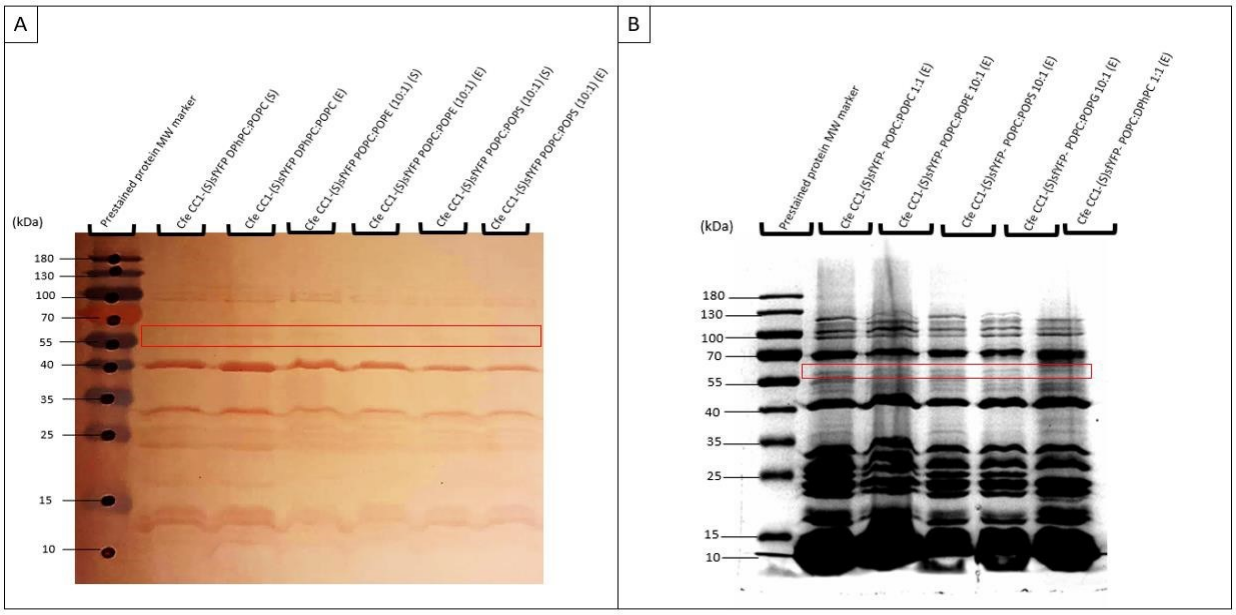


Figure 4.14. Anti-his western blots of cell-free expressed CC1-(S)sfYFP and liposomes of various compositions. A. Blot stained with Ponceau stain. Comparison between expression in PURExpress cell-free system of CC1-(S)sfYFP when liposomes of various compositions were added at the start and at the end of the reaction: Lane 1. Prestained protein MW marker. Lanes 2-3 CFE CC1-(S)sfYFP with POPC: DPhPC(1:1) liposomes (Lane 2-liposomes added at the start of reaction; Lane 3- liposomes added at the end of the reaction); Lanes 4-5 CFE CC1-(S)sfYFP with POPC: POPE (10:1) liposomes (Lane 4-liposomes added at the start of the reaction; Lane 5- liposomes added at the end of the reaction); Lanes 6-7 CFE CC1-(S)sfYFP with POPC: POPS (10:1) liposomes (Lane 6- liposomes added at the start of the reaction; Lane 7- liposomes added at the end of the reaction). It could be observed that the intensities of the bands of the cell-free expressed CC1-(S)sfYFP (˜56.6 kDa) corresponded to the reactions to which the liposomes were added at the end of the cell-free reactions. B. Comparison between expression in PURExpress cell-free system of CC1-(S)sfYFP when liposomes of various compositions were added at the end of the reactions. Lane 2- CFE of CC1-(S)sfYFP with POPC: POPC (1:1) liposomes added at the end of the reaction, in Lane 3 CFE of CC1-(S)sfYFP with POPC: POPE (10:1) liposomes added at the end of the reactions. Lane 4. CFE of CC1-(S)sfYFP with POPC: POPS (10:1) liposomes added at the end of the reactions. Lane 5. CFE of CC1-(S)sfYFP with POPC: POPG (10:1) liposomes added at the end of the reactions. Lane 6. CFE of CC1-(S)sfYFP with POPC: DPhPC (1:1) liposomes added at the end of the reactions. In panel

B, the bands corresponding to the cell-free expressed CC1-(S)sfYFP (the regions close to the CC1-(S)sfYFP ˜56.6 kDa were highlighted in red borders in panels A and B) were of similar intensities, despite the liposome compositions mixed with the CFE reactions.

4.4.6. Cell-free expressed protein identification by mass spectrometry

It was observed from the blots that the cell-free protein expression reactions led to histidine-containing protein bands situated at various molecular weights, and multiple protein bands around 55 kDa were migrating at close distances; hence, mass-spectrometry analysis was additional to the Western blotting. The mass-spectrometry analysis was performed by a team of scientists from the mass-spectrometry facility of the University of St. Andrews.

In Figure 4.15, panels A and B, on the left side, there are illustrated the corresponding images of the electrophoresed gels comprising the CFE reactions for CC1-(S)sfYFP (A) and REAMP-(S)sfYFP (B). The CFE reactions were not mixed with liposome solutions. In both panels on the right there are presented screenshots of the MASCOT results for protein sequence coverage. In both panels, the highlighted in red borders protein bands at the corresponding MW for CC1-(S)sfYFP (˜56.6 kDa) and REAMP-(S)sfYFP (˜56.7 kDa) were isolated and sent for analysis.

The mass spectrometry analysis resulted in protein sequence coverage of 14% for CC1-(S)sfYFP and 35% for REAMP-(S)sfYFP.

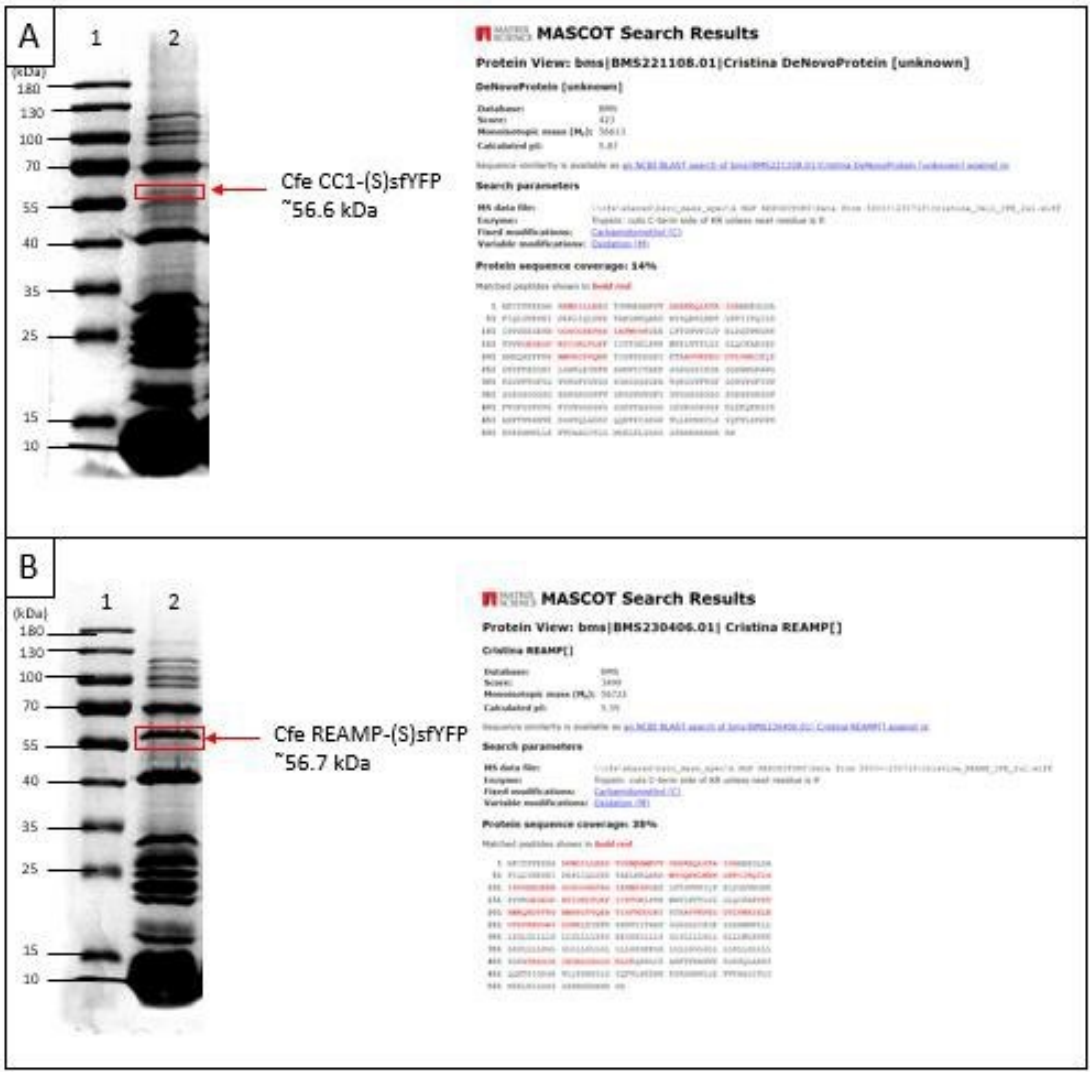


Figure 4.15. Mass-spectrometry results for identification of cell-free expressed CC1-(S)sfYFP and REAMP-(S)sfYFP. A. (Left) Electrophoresed acrylamide gel (12%). Lane 1. Prestained protein molecular weight marker. Lane 2. Cell-free expression reaction (PURExpress) corresponding to CC1-(S)sfYFP. The protein band of approximately 56.6 kDa, corresponding to the molecular weight of CC1-(S)sfYFP, is highlighted in a red border and indicated by an arrow. This band was isolated and sent for mass-spectrometry analysis. (Right) Results for mass-spectrometry identification. The CC1-(S)sfYFP sequence coverage was of 14%. B. (Left). Electrophoresed acrylamide gel (12%). Lane 1. Prestained protein molecular weight marker. Lane 2. Cell-free expression reaction (PURExpress) corresponding to REAMP-(S)sfYFP. The protein band of approximately 56.7 kDa, corresponding to the molecular weight of REAMP-(S)sfYFP, is highlighted in a red border and indicated by an arrow. This band was isolated and sent for mass-spectrometry analysis. (Right) Results for mass-spectrometry identification. The REAMP-(S)sfYFP sequence coverage was of 35%.

4.4.7. Droplet interface bilayer assembly

The reactions containing the cell-free expressed CC1-(S)sfYFP mixed with liposomes of various compositions, along with liposome solutions, were used in droplet interface bilayer assay, in order to

determine what liposome composition formed a stable bilayer, and whether the CC1-(S)sfYFP-containing lipid droplets could form stable bilayers for the insertion and functionality testing of the CC1-(S)sfYFP.

In Figure 4.16 is represented the arrangement of the types of droplets used in order to assess the possibility of the formation of stable bilayers.

All droplets had a volume of 1µl and were pipetted into a Petri dish containing a mixture of undecane: silicone oil (1:1) and incubated for approximately 5 minutes at room temperature for equilibration.

The first group of droplets (group 1, indicated by the arrow) was represented by two liposomes-only droplets (L). The liposome solution pipetted for this group had the same composition and concentration, as used for the corresponding protein cell-free expression reaction. The second group (2) was represented by two droplets of different compositions. The (L) droplet represented a liposomes-only droplet, and the (X) droplet contained the solution of the cell-free expressed CC1-(S)sfYFP obtained from the reactions with the liposomes (of the same composition as used for the (L) droplets), added at the end of the reaction, The last group (3) comprised the droplets containing two X droplets (containing the solutions of the cell-free expressed CC1-(S)sfYFP from the reactions mixed with liposomes).

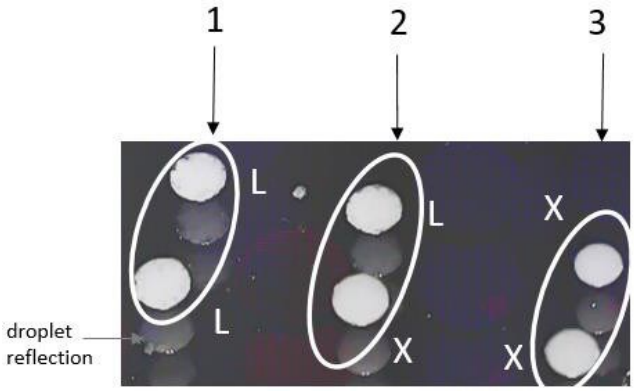


Figure 4.16. Droplets group arrangement for droplet interface bilayer formation assessment. Group 1 consisted of two droplets immersed in undecane: silicone oil (1:1) bath of liposomes of the same composition as used for the corresponding cell-free protein expression reaction (L). Group 2 consisted of two droplets: one of liposomes only (L) and one of the cell-free protein expression reaction mixed with the same type of liposome, added at the end of the reaction (X). Group 3 consisted of two droplets of the cell-free protein reaction mixed with the liposomes of the same composition as in the other groups. The droplets images of lower intensity found in the background are represented by the LED light reflected images.

4.4.8. Droplet interface bilayer formation and stability depending on lipid composition

The droplets were mechanically brought into contact using two sterile pipette tips (10 µl), and the formation and stability were recorded using an LED microscope camera connected to the OTG View software.

The liposomes compositions tested for the bilayer formation and stability were POPC: POPC (1:1); POPC: DPhPC (1:1); POPC: POPE (10:1); POPC: POPG (10:1); POPC: POPS (10:1).

In Figure 4.17 and 4.18 there are presented the DIB formation and stabilities of different types of droplets, grouped as presented in Figure 4.16, depending on the type of lipid composition that was used for the liposomes used for the droplets. The (L) droplet was represented by a droplet containing only liposomes of specified composition and the (X) droplet contained the cell-free expressed CC1-(S)sfYFP in liposomes of specific lipid composition, containing IVTT reactions. Figure 4.17 illustrates snapshots of the droplet groups, containing in their composition lipids that are known to give planar bilayers. In panel A is presented the DIB formation between droplets having in composition POPC: POPC (1:1), and in panel B, the series of events corresponding to droplets having in composition POPC: DPhPC (1:1).

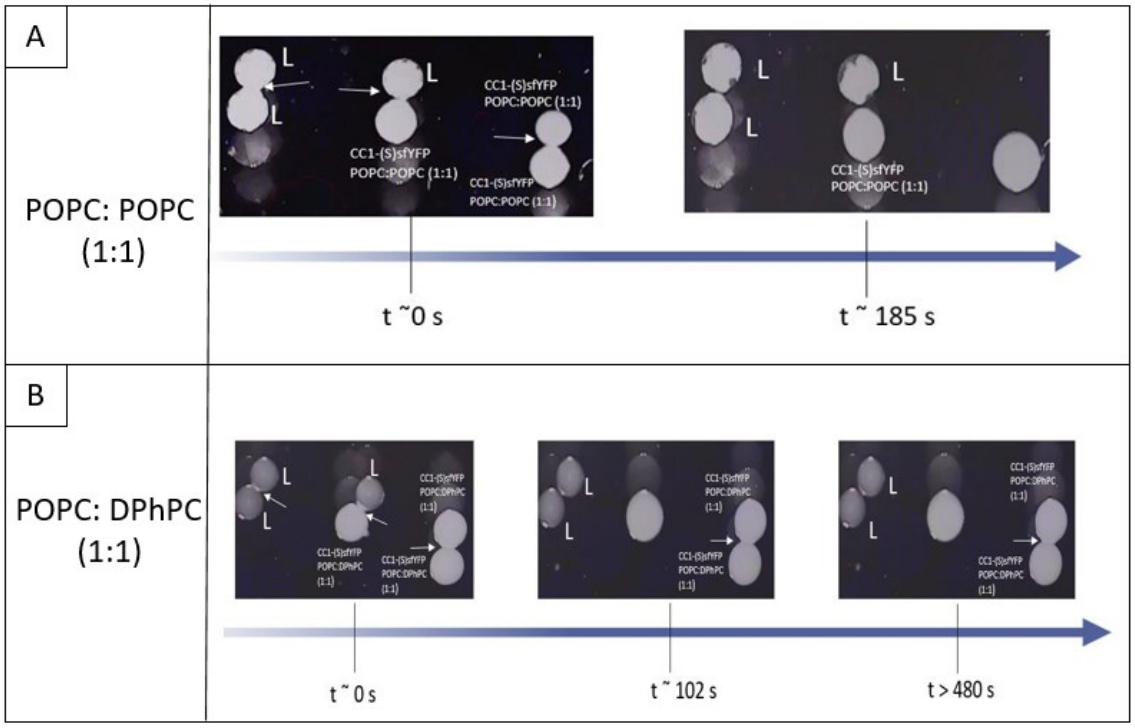


Figure 4.17. DIB formation assessment for planar bilayers. Panel A. DIB formation assessment for POPC: POPC (1:1) bilayer. At starting point (t~0), each group formed the bilayer. L is represented by the POPC:POPC (1:1) liposomes only droplets and the X is represented by the cell-free reaction for the expression of CC1-(S)sfYFP, mixed at the end

of the reaction with POPC:POPC (1:1) liposomes. At˜ 185 s, the first two groups of droplets separate (the droplets interface is destroyed in the case of the group with liposomes-only droplets and the one with mixed droplets). The third group of droplets after ˜185 s coalesced.
Panel B. DIB formation assessment for POPC: DPhPC (1:1) bilayer. At t˜0 each group of droplets formed a DIB. L was represented by the POPC: DPhPC (1:1) liposomes-only droplets and the X was represented by the cell-free reaction for the expression of CC1-(S)sfYFP, mixed at the end of the reaction with POPC: DPhPC (1:1) liposomes. At t˜102 s, the liposomes-only droplets (first group) separated and the mixed droplets (second group) coalesced, whilst the DIB formed between the cell-free expression reaction-only droplets (third group) remained intact.. Even after 480 s the bilayer formed by the cell-free expression reactions-only droplets (third group) remained intact.

In panel A, it could be observed that after approximately 185 s from the droplets contact, the second droplet group ((L-L) and (L)-CC1-(S)sfYFP)) separated and the last group coalesced.

In panel B, it could be observed an increased stability of the DIB, when DPhPC lipid was introduced in the liposomes composition. Modifications in the stability were observed after approximately 102 s, when the first droplet group separated and the second group coalesced. The DIB formed by the droplets containing the cell-free protein expression mixtures was observed to remain stable even after 480 s from the formation starting point.

In Figure 4.18 is presented a comparison between the DIB formation and stability of the droplet groups that have in composition lipids that induce bilayer curvature. In panel A is presented the effect on DIB formation and stability of POPC: POPS (10:1) lipid composition; in panel B, the effect of POPC: POPG (10:1) and in panel C, the influence of POPC: POPE (10:1). In each snapshot the droplet group order is the same as in the general example.

In panel A, it could be seen that, overall, the DIB stability was affected compared to the same events presented in Figure 4.17. After approximately 28 s, the third group separated, and after approximately 83 s, the second group coalesced.

Panel B illustrates a similar impact of the lipid composition on the DIB formation and stability. In this case, the droplets containing POPC: POPG (10:1), present in the first group separated after approximately 32 s, whereas the second and third group of droplets, respectively the groups that have at least one droplet containing the cell-free expressed protein, coalesced after 70 s.

The trend is respected in panel C, for droplets having in composition POPC: POPE (10:1) liposomes. In this case, the DIB stability was slightly increased, the first and second droplet groups separated after approximately 39 s, whilst the DIB formed between the droplets of the third group remained stable until approximately 92 s from the formation, and then coalesced.

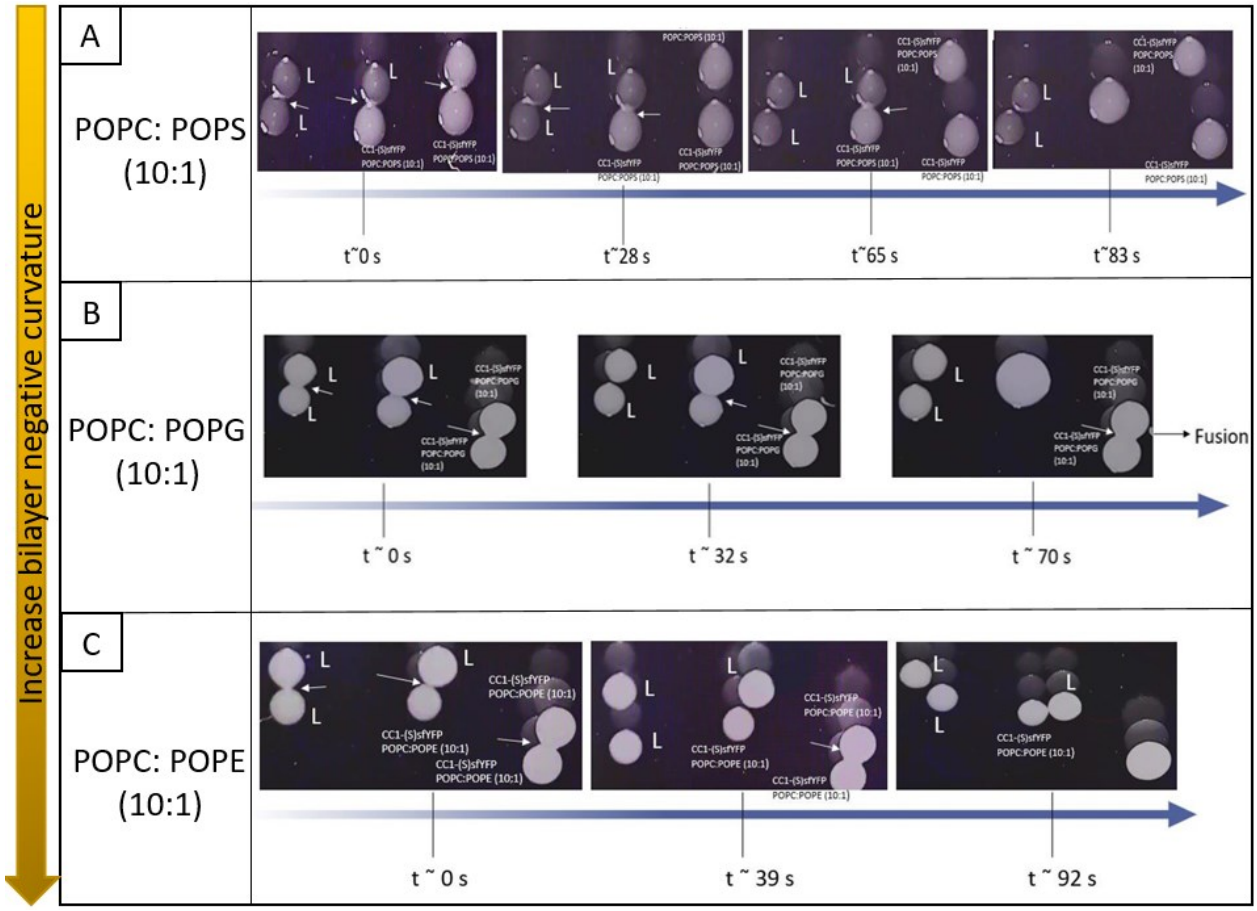


Figure 4.18. DIB formation and stability of negatively curved bilayers. Panel A. DIB formation assessment for POPC: POPS (10:1) bilayer. At t˜0 each group of droplets formed a DIB. L was represented by the POPC: POPS (10:1) liposomes-only droplets and the X was represented by the cell-free reaction for the expression of CC1-(S)sfYFP, mixed at the end of the reaction with POPC: POPS (10:1) liposomes. At t˜28 s, the cell-free reaction-only (third group) droplets separated. At t˜65 s, in addition to the third group of droplets, also the liposomes-only (first group) droplets separated. At t˜83 s, the mixed (second group) droplets coalesced. Panel B. DIB formation assessment for POPC: POPG (10:1) bilayer. At t˜0, each group of droplets formed a DIB. L was represented by the POPC: POPG (10:1) liposomes-only droplets, and the X was represented by the cell-free reaction for the expression of CC1-(S)sfYFP, mixed at the end of the reaction with POPC: POPG (10:1) liposomes. At t˜32 s, the liposomes-only droplets (first group) separated, whilst the bilayers formed by the other two groups of droplets remained intact. At t˜70 s, the mixed droplets (second group) coalesced, whilst the bilayer remained intact for 1-2 seconds more, then fused (snapshot not shown due to error in recording). Panel C. DIB formation assessment for POPC:POPE (10:1) bilayer. At t˜0 each group of droplets formed DIB. L was represented by the POPC:POPE (10:1) liposomes-only droplets and the X was represented by the cell-free reaction for the expression of CC1-(S)sfYFP, mixed at the end of the reaction with POPC:POPE (10:1) liposomes. At t˜39 s, the liposomes-only (second group) and the mixed (third group) droplets separated. The DIB formed between the cell-free reaction droplets remained intact. At t˜92 s the cell-free reaction droplets coalesced.

From the analysis of Figures 4.17 and Figures 4.18 it can be seen an impact on the DIB formation and stability, depending on the lipid compositions used for the liposomes comprised by the droplets. Planar bilayer forming lipid compositions such as POPC: POPC (1:1) and POPC: DPhPC (1:1) enabled the

formation of more stable bilayers, compared to the bilayers with induced curvature formed of POPC: POPS (10:1), POPC: POPG (10:1) and POPC: POPE (10:1). The POPC: DPhPC (1:1) was identified to be the most suitable lipid composition for stable DIB formation. Moreover, in most of the cases the third droplet group (droplets containing the cell-free expressed protein), formed a more stable bilayer, compared to the rest of the combinations.

4.4.9. Development of alternative prototype models for reservoirs for DIB formation

From the previous experiments, it was observed that the high volume of the oil-solvent bath in the Petri dish (approximately 10 ml) oppressed the mechanical manipulation of the droplets and DIB formation.

New models of containers (or reservoirs) for a silicone oil-undecane bath of a smaller volume (microliters scale) were designed. The designed reservoirs were 3D printed using UV-light curing clear resin. Reservoirs of the same designs and made from polydimethylsiloxane (PDMS) were also tested.

Figure 4.19 presents the 3D computer-aided design (using Sketch Up) chip containing eight reservoirs for the DIB formation.

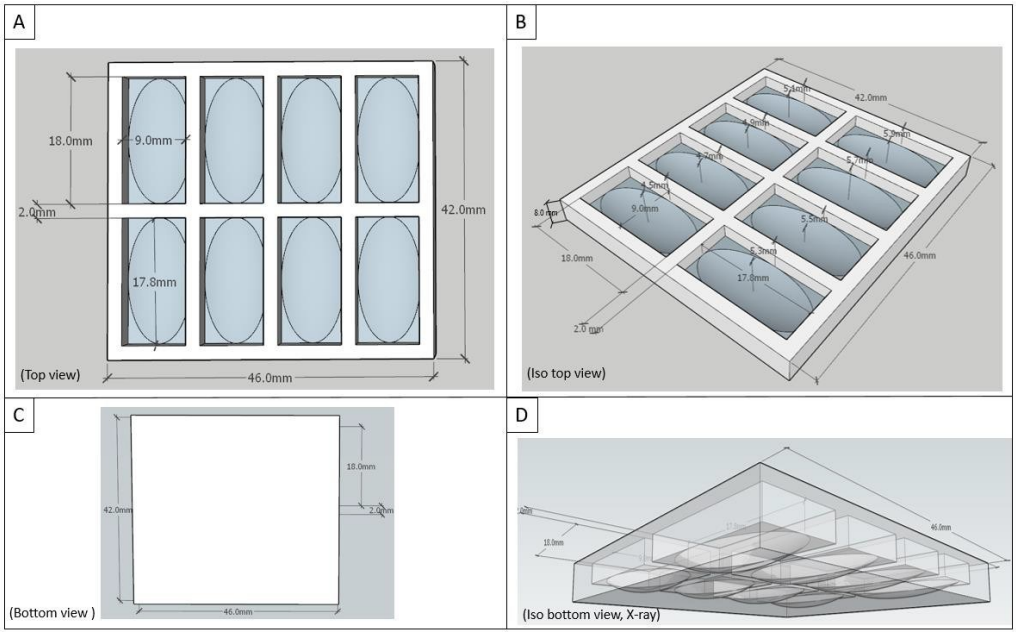


Figure 4.19 3D chip design for DIB formation. A. Top view of the chip containing 8 oval-shaped wells contained in rectangular-shaped cassettes. The total length of the chip was 46.0mm; the total width was 42.0 mm and the total height was 8.0 mm. The length of a cassette was 18.0 mm, the height was 2.0 mm, and the length of an oval-shaped

well was 17.8 mm. The width of a well was 9.0 mm. B. Iso view, top perspective. The depths of the oval shaped well were 4.5 mm; 4.7 mm; 4.9 mm; 5.1 mm; 5.3 mm; 5.5 mm; 5.7 mm and 5.9 mm. C. Bottom view of the chip with flat bottom surface. D. Iso view, bottom and X-ray perspective. The oval-shaped wells had rounded surfaces at the bottom, and the chip had a flat surface at the bottom.

Panel A represents the top view of the design comprising the oval-shaped wells contained in rectangular cassettes (length of 18.0 mm). The length of the entire chip was 46.0 mm, and the width was 42.0 mm. Panel B represents the isometric view of the chip, in which the depths are highlighted. The oval-shaped wells had a length of approximately 17.8 mm, a width of approximately 9.0 mm (panel A) and a thickness of 8.0 mm (panel B). The series of wells had gradient depths between 4.5 -5.9 mm, with an increment of 0.2 mm (panel B). The wells were separated by walls of 2.0 mm width (panel B). The wells had rounded surfaces at the bottom, as indicated by the X-ray view presented in panel C, and the entire chip assembly had a flat surface at the bottom (panels B and C).

The DIB formation in the resin and PDMS chips wells was tested using droplets containing liposomes-only solutions of various compositions and concentrations: of POPC: POPC (1:1); POPC: POPE (10:1); POPC: POPS (10:1) and 25mg/ml; 2.5 mg/ml; 250µg/ml of POPC: DPhPC (1:1); 10mg/ml; 1mg/ml; 100µg/ml POPC: POPG (10:1).

In both cases, the droplets were not stable when pipetted in the wells of any of the chips containing 100 µl of undecane: silicone oil mix. The droplets burst before the mechanical manipulation (hydrophilic materials). These results were consistent across all experiments with the various liposome composition droplets.

4.4.10. Material coatings

The results observed for the uncoated resin and PDMS wells led to the conclusion of the necessity of testing different coatings of the materials for the modification of surface chemistry.

Different resin and PDMS moulds were incubated in dopamine solution 2mg/ml dissolved in 100mM Tris-HCl pH=8.0, bovine serum albumin solution 2mg/ml or coated with a uniform layer of enamel paint. Successful DIB formation could be seen only for the PDMS mould coated with the dopamine solution, and for the resin mould coated with enamel paint, as presented in Table 4.2. The POPC: POPC (1:1), 10mg/ml lipid droplets represented the only set which formed DIBs in the resin reservoirs. The resin mould was coated with enamel paint and the DIB formed was stable for approximately 180 s. The lifetime of the lipid droplet assemblies with the following composition: POPC: POPC; POPC: POPE; POPC: POPS in the same concentrations described previously, did not exceed, on average more than 30 s, resulting in the coalescence of droplets in both the resin and PDMS moulds, hence considered as not formed (-).

For the POPC: DPhPC (1:1), 25mg/ml lipid droplets, the lifetime of the formed DIB was approximately 600 seconds before coalescence when the PDMS mould coated with the dopamine solution was used. Similarly, for the same lipid droplet composition, at a concentration of 2.5mg/ml, the lifetime was approximately 480 seconds in the same environment (PDMS coated with dopamine solution), as presented in Table 4.4

Table 4.4. The influence of material coatings on DIB formation. In the first columns are presented the droplets of different liposomes composition and concentrations (POPC: POPC; POPC: POPE; POPC: POPS; -10mg/ml-100µg/ml; POPC: DPhPC 25mg/ml-250µg/ml). Resin and PDMS coated with dopamine (D) solution, enamel paint (EP) and BSA (B). In the last columns are presented the results for the DIB formation (DIB formed: (+); DIB not formed: (-)) and lifetime measured in seconds from the formation starting point. Highlighted in grey boxes are the corresponding liposome concentrations that formed DIBs in the corresponding wells with coated surface.

Droplets		Material coatings		DIB formation (+/-) Coating (#) for DIB formation DIB lifetime (s)	
		Resin	PDMS	Resin	PDMS
Lipid composition	Lipid concentration				
POPC: POPC (1:1)	10mg/ml	Dopamine 2mg/ml (D) Enamel paint (EP) BSA 2mg/ml (B)		+ (EP) 180 s	-
	1 mg/ml			-	-
	100 µg/ml			-	-
POPC: POPE (10:1)	10 mg/ml			-	-
	1 mg/ml			-	-
	100 µg/ml			-	-
POPC: POPS (10:1)	10 mg/ml			-	-
	1 mg/ml			-	-
	100 µg/ml			-	-
POPC: DPhPC (1:1)	25 mg/ml			-	+ (D) 600 s
	2.5 mg/ml			-	+ (D) 480 s
	250 µg/ml			-	-
POPC: POPG (10:1)	10 mg/ml			-	-
	1 mg/ml			-	-
	100 µg/ml			-	-

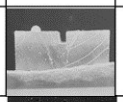
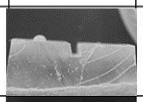
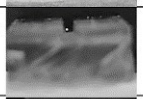
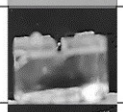
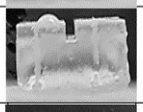
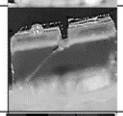
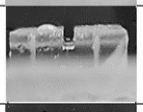
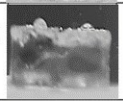
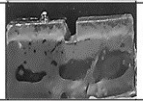
4.4.11. Further post-processing surface chemistry modification

Based on the majority of negative results obtained for the tested coatings and the deformations observed for the PDMS moulds, it was intended the use of resin moulds for future experiments. Further

coatings were tested for the modification of the surface chemistry of the wells of the moulds. Substitute rectangular resin chips (2 cm x 2 cm) with flat surfaces were 3D printed, cured using UV light and used to test different coatings. The coating substances were: dopamine hydrochloride 2mg/ml solution dissolved in 100 mM Tris-HCl, pH= 8.0; bovine serum albumin solution 2mg/ml dissolved in sterile RNA-free water; polyethylene glycol (PEG) 20K 2% solution; hydrolysed styrene-maleic acid (SMA) 2000 2% solution, and acrylate-derived adhesive. Table 4.5 presents the measurements of the contact angles of the water droplets for each coated resin chip incubated in separate phases.

It was observed from the droplet test angle measurements that the uncoated chip surface and the BSA (2mg/ml) and PEG20K (2%) coatings provided hydrophilic surfaces, whilst dopamine (2mg/ml) and SMA2000 (2%) provided surfaces with mixed chemistry (acted as both hydrophilic and hydrophobic when incubated in oil or water). The coating, which provided a hydrophobic resin surface with an angle between the surface and the tangent to a water droplet greater than 90° after the incubation in both silicone oil (θ=122.66°) and water (θ=133.26°), was the acrylate-derived adhesive.

Table 4.5. Surface chemistry modification with coatings. The first column comprises the types of coating applied. The second column contains pictures of the water droplet pipetted on the corresponding coated resin surface which was incubated in oil. The third column contains the corresponding calculated contact angle. The fourth column contains pictures of the water droplet pipetted on the corresponding surface, which was incubated in water. The fifth column contains the contact angles of the water droplets from column 4. The sixth column describes the type of surface chemistry formed by each coating.

Type of material	OIL	Angle	WATER	Angle	Type of surface
Uncoated resin chip		62.35°		52.49°	Hydrophilic
Dopamine 2mg/ml		106.85°		0°	Mixed
BSA 2mg/ml		37.65°		37.51°	Hydrophilic
PEG20K 2%		88.26°		61.54°	Hydrophilic
SMA 2000 2%		105.95°		~0°	Mixed
Acrylate-derived adhesive		121.66°		133.26°	Hydrophobic

The chip containing oval-shaped wells was coated with a thin layer of the acrylate-derived adhesive, dried and tested for DIB formation compatibility. The same compositions of liposome solutions were used for the droplets immersed in the 100 µl of the undecane: silicone oil mixture contained in the wells.

The droplets containing POPC: DPhPC (1:1) liposomes solution formed a DIB in the chip well coated with the acrylate-derived adhesive. The DIB was stable for approximately 144 seconds, after which the droplets coalesced. The DIB formation and modification in stability are presented in the recording snapshots comprised in the Figure 4.20. In panel A is represented the formation of the DIB and in panel B is indicated the coalescence of the droplets after approximately 144 seconds.

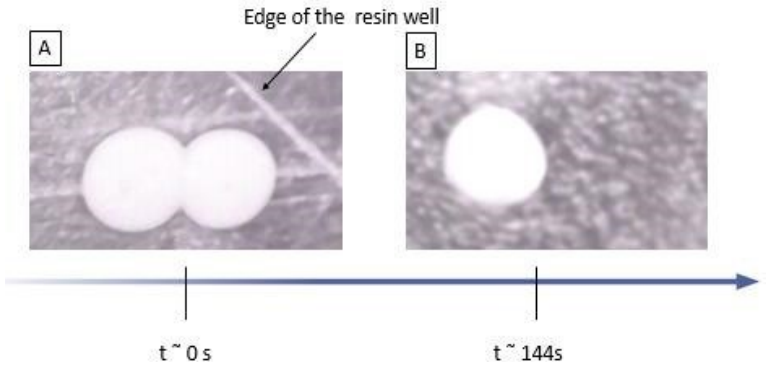


Figure 4.20. Assessment of DIB formation and stability formed by POPC: DPhPC (1:1) liposomes containing droplets in resin well coated with acrylate-derived adhesive. A. Starting point indicating the formation of the DIB. Panel B. After approximately 144 seconds, when the droplets coalesced.

4.5. Discussion

4.5.1. Cell-free expression of membrane proteins

The use of the cell-free expression approach also referred to as *in vitro* transcription and translation (IVTT) of membrane proteins, has been increasing, especially in the past decade. The choice reasoning of this expression approach is based on one of its great advantages- the ability to express toxic or difficult-to-express proteins in natural hosts. As numerous reports have shown, the expression of membrane proteins in common cell hosts (bacterial, yeast, mammalian) can be challenging. Nevertheless, the expression of *de novo* membrane proteins would increase the difficulty, as could be seen in Chapter 2. Thus, the use of cell-free-based expression for *de novo*-designed membrane proteins represents a promising alternative to overcome the challenges encountered in traditional cell-based expression.

The cell-free expression principle is based on the use of a cell lysate, which confers the minimum

necessary cell components for transcription and translation, eliminating the rest of the cell components that inhibit the assimilation of the synthesised protein. The cell lysate can be obtained from *E. coli*, yeast cells, and even rabbit reticulocytes. Due to the fact that obtaining the lysate components using traditional separation methods can be time-consuming and yields can vary depending upon experimental conditions, alternatives became more popular.

Such alternatives are commercially ready-to use available, such as cell-free protein synthesis kits (for example, PURExpress), which are composed of recombinantly expressed and purified transcription/translation components. In order to replicate the natural environment in which membrane proteins rely (lipid bilayer), these biomolecules need to be supplemented to the CFPS reaction. Often, they are supplied in the forms of liposomes, or nanodiscs, as artificial reconstitution systems for membrane proteins, in which the newly synthesised protein can insert and fold.

4.5.2. Bimolecular fluorescence complementarity as an additional contributor to membrane protein folding

Chaperones are essential proteins that play crucial roles, assisting protein folding and preventing protein aggregation. Based on their mechanisms of activation through conformational changes, chaperones can be classified as adenosine triphosphate (ATP)- dependent and independent (Mas *et al.*, 2020). Chaperones, along with the translocation machinery-corresponding molecules, are present among the rest of the CFPS kit components (Harris *et al.*, 2022). Due to the fact that the solutions of a CFPS kit would have a limited amount of energy (ATP) supply, the segments of the (S)sfYFP might further contribute to the assessment of the overall folding through the bimolecular (fluorescence) complementarity, acting as an “internal chaperone”. The use of own chaperone-like molecules, in addition to the host chaperones, was also observed in some viruses (Xiao, Wong and Luo, 2010). This hypothesis led us to test further fluorescence-related experiments for the determination of a potential relationship between the observed fluorescence intensity values and insights correlated to the folding state of the synthesised protein.

The synthetic nature of *de novo*-designed proteins, in order to be successfully expressed without any impediments, would need to circumvent the translocon machinery. Thus, the addition of tags represented by strongly folding proteins may assist and drive the folding of synthetically derived proteins. GFP is known to be a pseudo-irreversible folded protein. In addition, other fluorescent proteins have been engineered to aid in the folding studies, such as the superfolder version of the yellow fluorescent protein (sfYFP) (Romei and Boxer, 2019).

Bimolecular fluorescence complementarity (BiFC) involves the reconstitution of a fluorescent protein when two non-fluorescent fragments are brought together by the interaction of attached proteins (Kodama and Hu, 2012). BiFC is a valuable tool for studying protein-protein interactions and protein folding, as demonstrated by the recovery of fluorescence when split GFP fragments are attached to interacting peptides (Ghosh *et al.*, 2000). In this study, we have exploited this approach to drive the folding of the *de novo*-designed peptide. The strong folding behaviour of the split YFP acts as an internal chaperone and monitors membrane protein folding. The high fluorescence values observed compared to the background indicate successful expression and folding of the proteins. The antiparallel orientation of the CC1 helices, driven by the complementary side-chain interactions, likely contributes to the proper folding of the sfYFP fragments, as evidenced by the fluorescence intensities.

Due to the fact that there could be observed high fluorescence values compared to the background (CPFS components, liposomes), it indicated the successful expression of the proteins, and indicated the folding of the molecule. The side chains of the transmembrane helices of CC1 are highly complementary in an antiparallel orientation, as it was designed (see Chapter 2), and the fluorescence intensities exhibited (fluorescent molecule folded) would suggest the appropriate orientation of the helices. For future studies, the choice of appropriate control DNA plasmids would be desirable, with the sfYFP segments to be attached to non-associating proteins (such as the poly-leucine helices, as mentioned in Chapter 2 and Chapter 3). This would enable us to distinguish between the association driving forces (either the P-P interactions need to be satisfied within the structure of the attached molecule to fold the YFP, or solely the segments found in proximity can fold, which drives the overall folding).

4.5.3. Influence of the addition time points of the liposomes on the expression efficiency

The difference in fluorescence intensities could be attributed to several factors. First, the early addition of liposomes might lead to the sequestration of essential CFE reaction components, such as amino acids, nucleotides, or enzymes, thereby reducing the available resources for protein synthesis. Second, liposome-ribosome interactions could hinder translation by interfering with mRNA engagement, resulting in delayed or inhibited protein synthesis. Importantly, the macrocycle formed by the *de novo* designed peptide and the superfolder YFP may allow the subsequent folding of the bundles in to the lipid bilayers despite their absence near the start of the expression. In the experiments, the lower fluorescence intensities observed when liposomes were added at the start of the reaction support these hypotheses. The liposomes likely sequestered essential reaction components and interfered with protein synthesis,

leading to reduced expression of CC1-(S)sfYFP and REAMP-(S)sfYFP.

Considering the mechanism of the bimolecular fluorescence complementarity (BiFC), following the folding of the sfYFP segments, a low fluorescence signal but still higher than the background was recorded for the reactions with liposomes added at the start. The sfYFP protein-protein interactions are not completely satisfied. This could be the possible reason why there are differences between the fluorescence intensity values observed when liposomes were added at the start versus at the end of the CFE reaction. Also, the differences between the fluorescence across the lipid compositions used could be explained by the lipid structural characteristics (space between the head group, between the tails, curvature formed) and the impact on the folding or self-association of the proteins TMs.

Also, the fluorescence intensities obtained for the proteins across the different lipid groups could be due to the additional competitive interactions with the existing ions and the CFE components; if they create a shield, they diminish the protein-lipids interactions, enhancing the possibility of protein-protein interaction, association and fluorescence of the YFP, hence the differences observed for the results of the fluorescence values measured for the reactions having the liposomes added at the start and at the end of the reactions.

Whilst the fluorescence of the YFP can be used as a guide for the study of the folding mechanism and protein-protein interactions, there is also the possibility of the existence of false positive results.

One possible explanation for the differences observed in the fluorescence intensities could be related to some inhibitory factors of the autocatalytic formation of the chromophore in YFP. The autocatalytic formation of the chromophore involves the folding of the protein chromophore moieties (formed of Ser, Tyr and Gly), cyclization of the amino acids for the formation of the imidazolinone ring, oxidation (for the formation of the characteristic double bonds conjugated system), system conjugation extension by Y203 and maturation (Stepanenko *et al.*, 2013). These essential steps for the chromophore formation could be affected by certain conditions in a cell-free expression context, such as: oxygen limited availability (especially in dense and poorly mixed reactions), chances in the pH due to the formation of certain by-products during the protein synthesis, ionic concentrations and presence of reducing agents which inhibit the oxidation step. In this case, although it is a weak reducing agent, the presence of chloride ions, as a result from the pH adjustment, in the sodium phosphate buffer in which the liposomes were resuspended, could be considered to interfere.

It could be more plausible the hypothesis that there is a compatibility between the cell-free reaction critical components and liposomes, a phenomenon which inhibits the protein synthesis reactions.

For example, some liposomes could sequester essential biomolecules from the cell-free reaction

(such as amino acids, nucleotides, cofactors, or enzymes). The sequestration of these biomolecules would decrease the concentration of the biomolecules for maximum protein expression efficiency. Furthermore, the liposomes might bind to the ribosomes included in the solutions of the CFE reaction, and this interaction would interfere with the ability of ribosomes to engage with the mRNA and to delay or inhibit the protein translation, therefore inhibiting protein synthesis. Moreover, the liposomes added at the start of the reaction altered the pH and the ionic concentration required for the successful completion of the IVTT reaction. Although not specified clearly the type of ions not to be added, we could conclude that the presence of Na^+ from the sodium phosphate buffer in which the liposomes were resuspended coincided with the requirement for the lack of sodium, as specified for the potential inhibitors such as NaCl, glycerol, EDTA and Mg^{2+} or K^+ salts (New England Biolabs, 2021b).

Currently, cell-free reaction components are made available in commercial reaction kits, overcoming the challenges associated with the extraction of essential protein synthesis components from cell lysates.

However, the efficiency of the kits can vary depending on the end-applicability and the templates used (New England Biolabs, 2021b). Depending on the composition of the templates, protein synthesis can be suppressed or even not initiated.

To this date, groups of researchers focused on the most frequent challenges encountered in the cell-free synthesis of *de novo* proteins. They developed alternative kits, modifying the compositions of the solutions and making *de novo* cell-free protein synthesis more accessible.

For example, Matsuura and colleagues used a PURE cell-free expression kit (PURExfrex 1.0) for which they modified the concentrations of some components. They obtained significantly higher expression for the GFP protein in POPC giant unilamellar vesicles (Uyeda *et al.*, 2022).

4.5.4. Fluorescence intensity variation depending on liposome composition

The CC1-(S)sfYFP and REAMP-(S)sfYFP were expressed in liposomes, and sucrose gradient flotation was used to remove excess empty liposomes and unfolded or aggregated protein. Looking at the data and excluding the planar bilayers (POPC-POPC and POPC-DPhPC), the increase of fluorescence was enhanced by the increase in the bilayer negative curvature as follows: PS<PG<PE for both CC1 and REAMP. In the case of DPhPC, the data showed an antithetic impact on the fluorescence of CC1 and REAMP (CC1 exhibiting the lowest fluorescence, REAMP exhibiting the highest). This could be possible due to the competitive protein-lipid interaction forces; the sidechains of the CC1 TMs, and probably the ones of phenylalanine,

exhibited van der Waals and π-cation interactions with the methyl groups of the branched tails of DPhPC. These interactions possibly stabilised the TMs within the bilayer, however did not necessarily promote the closest association of the TMs.

Although the interactions through hydrogen bonds are prevalent in soluble proteins, due to the prevalence and specific disposition of hydrophilic amino acids (such as Ser, Thr), it has also been shown in the early studies on membrane proteins, such as glycophorin A, that carbon backbone hydrogen atoms can form hydrogen bonds (Smith *et al.*, 2002).

REAMP, contains in the structure of the TMs serine residues that could be hydrogen bond donors and interact with the hydrogen atoms of the methyl groups of the DPhPC tails. It is known that in most of the cases, the hydrogen bonds are typically stronger interactions than the van der Waals dispersion, π-π, or π-cation interactions. REAMP also contains leucine residues, which could contribute to the side chains atom groups for van der Waals interactions with the DPhPC lipid tails. Considering the sum of interactions that the protein transmembrane helices structures could exhibit, stronger interactions through the side chains would exhibit REAMP due to the hydrogen bonding, whilst electrostatic, van der Waals and π-π would be predominant for CC1.

These interactions could promote better the self-association of the TMs of CC1 in the POPC: DPhPC bilayer, leading to lower fluorescence values obtained in this case, compared to REAMP.

The computational simulations showed that the structure of the REAMP is flexible (see Chapter 2). The interactions exhibited by the residues placed in the transmembrane regions, especially the DPhPC lipid molecules, combined with the architecture of DPhPC tails would lead to a "relaxed" conformation of REAMP, with a weak association between the transmembrane helices and lipid tails in a low packing density state.

Although the established dependence can vary upon experimental conditions, it is known that the amount of dispersed light can increase the fluorescence values. Furthermore, it is known that in the case of planar bilayers (as is the case for POPC: DPhPC), the lipid tails contribute to the increase of dispersed light. Considering the type of interactions of REAMP and disposition of the DPhPC tails, the fluidity of the bilayer and the exposed lipid tails led to the increase in the amount of light dispersed, thus to the fluorescence intensity in the case of CFE REAMP in POPC: DPhPC liposomes.

Furthermore, it is known that the dispersed light effect is exhibited by both the lipid head groups and the lipid tails; however, for the bilayers under curvature stress, the dispersion effect of the head groups is more pronounced. Studies have shown that bilayers under negative curvature impact more the dispersion of the light in comparison with bilayers under positive curvature.

The fluorescence data obtained in the present experiments support the literature statements. If the data observed for the planar bilayer is not considered, the fluorescence intensities observed in the experiments involving either CFE CC1 or CFE REAMP increased with the increase of the negative bilayer curvature, as influenced by the lipid composition (PS< PG< PE).

4.5.5. Cell-free protein expression efficiency is influenced by multiple factors

The expression levels of CC1-(S)sfYFP and REAMP-(S)sfYFP in our cell-free system were influenced by several factors, as evidenced by the Western blot analysis. One of the key factors affecting expression levels was the liposome composition. For REAMP-(S)sfYFP, the protein band intensities varied depending on the lipid composition, with the highest intensities observed in POPC: POPC (1:1) and POPC: POPG (10:1) liposomes. This observation aligns with the preference of REAMP for POPE-containing bilayers, as demonstrated in our molecular dynamics simulations (Chapter 2) and *E. coli* expression experiments (Chapter 3). In contrast, the expression levels of CC1-(S)sfYFP appeared to be less influenced by the liposome composition, with only slight enhancements in the presence of POPC: POPC (1:1), POPC: DPhPC (1:1), and POPC: POPE (10:1) liposomes. Another critical factor influencing expression efficiency was the CFPS kit components. The amino acid sequence of the expressed protein and the codon usage can affect the availability of essential components, such as deoxyribonucleoside triphosphates (dNTPs). For CC1-(S)sfYFP, supplementing the kit with additional dGTP, dTTP, and dCTP might improve expression levels by ensuring an adequate supply of the necessary building blocks for the predominant amino acids in the CC1 sequence. Other factors, such as the presence of chaperones, the stability of the mRNA, and the overall energy supply in the CFPS reaction, can also impact protein expression levels. While these factors were not directly investigated in our study, they represent potential areas for future optimisation and exploration

The workflow based on the fluorescence observations was meant to represent a screening method for insertion, folding, stability, and further functionality testing. Hence, it has to be determined if the expression and insertion levels of the folded states were influenced depending on the lipid environment.

Overall, it could be concluded that the expression of both proteins was profoundly inhibited by the liposome addition at the start of the reaction. Furthermore, the expression levels of the CC1-(S)sfYFP were decreased in comparison with the REAMP-(S)sfYFP.

Interestingly, the expression levels of the proteins varied depending on the liposome

compositions added to the CFE reactions.

The expression levels of the CC1-(S)sfYFP, as observed from the visual inspection of the protein bands, were considered to be similar and did not indicate a direct influence from the liposome composition. The intensities of the protein bands were enhanced, however, not considerably, in the presence of POPC: POPC (1:1), POPC: DPhPC (1:1), and POPC: POPE (10:1) liposomes.

The variation of the expression levels of the cell-free expressed REAMP-(S)sfYFP could be attributed to the direct influence of the liposome compositions. The protein band intensities decreased depending on the following liposome compositions: POPC: POPC (1:1)> POPC: POPG (10:1)> POPC: DPhPC (1:1)> POPC: POPE (10:1)> POPC: POPS (10:1). The low expression level observed in the case of PE containing liposomes could be due to the low concentration of POPE in the liposomes (10 POPC: 1 POPE molar ratio).

One of the experimental challenges regarding the determination of protein expression levels relies on the encountered challenges associated with protein quantification. The determination of the CFE protein expression was performed from the visual inspection of the protein bands observed at the corresponding molecular weights, and obtained from Western blotting. The protein quantification post-cell-free expression reaction was attempted using commercial kits (Pierce BCA protein assay kit, Thermo Fisher, UK). The results were inconclusive even after performing the troubleshooting methods indicated by the manufacturer (related to the absorbance errors introduced by the presence of the lipid molecules in the reaction). Also, performing dilution of the reactions would still not be a clear indicator regarding the amount of protein expressed due to the variation of the protein liposome insertion, etc. The traditional densitometry assays against bovine-serum albumin standards would still not represent a reliable test, especially due to the high background noise.

Hence, the implementation of more sensitive methods for assessing cell-free synthesised protein yield would be part of future work.

For example, in other studies, the yield of the synthesised protein was measured using TCA to precipitate labelled protein after the synthesis in the presence of ^{35}S methionine (Blackholly *et al.*, 2022). This approach was not possible in this study due to cost and experimental design accessibility; however, it would be desirable for future determinations. However, this method would imply the CC1 sequence alteration. Instead, based on the presence of tyrosine molecules in the CC1 sequence and the wide range of available radiolabelled molecules, ^{123} I-3-iodo-α-methyl-L-tyrosine could be used, following a similar strategy (Jager *et al.*, 2001) .

Another essential aspect to consider in the expression level comparative analysis between the

CC1-(S)sfYFP and REAMP-(S)sfYFP is related to the CFPS kit components and the protein sequence.

For future experiments envisioning the increase of the cell-free based expression levels of CC1-(S)sfYFP, it would be essential to consider the supplementation of the kit solutions with deoxyribonucleoside triphosphates (dNTPs). It is possible that due to the protein amino acid sequence and the potential codon usage, the concentrations of the dNTPs supplied in the kit solutions were not sufficient for the synthesis of the amino acids composing the CC1 sequence. It would be of future interest to consider the supplementation of the concentrations of dGTP (deoxyguanosine triphosphate), dTTP (deoxythymidine triphosphate) and dCTP (deoxycytidine triphosphate). These deoxyribonucleosides contain the corresponding nitrogenous bases guanine (dGTP), thymine (dTTP) and cytosine (dCTP). These bases encode the predominant amino acids in the CC1 sequence, more specifically the glycine residues (9 residues per TM), being encoded favourably by the GGC codon (2.02%), followed by the valine residues (6 residues per TM), encoded by GTG (2.71%), and phenylalanine residues (3 residues per TM), encoded by TTC (1.81%), as described in the codon usage table (Subramanian *et al.*, 2022).

Moreover, regarding the CFPS inhibition by the addition of liposomes at the start of the reaction, it would be desirable to increase the concentration of other kit components in order to overcome the sequestration of the kit biomolecules by the liposomes.

4.5.6. Incorporation of the cell-free expressed protein folded states influenced by the liposome composition

Research studies on cell-free expressed membrane proteins show they can fold better in certain types of curved bilayers. The lipid composition of the cell membrane affects this. Its properties like head group charge, chain saturation, chain length, and phase behaviour can influence protein folding. Different lipids like DOPC, DOPG, and DOPE, as they are predominantly used in experiments, proved to have varying effects. They impact protein insertion into the membrane, folding rate, transmembrane helix topology, and structural stability (Harris *et al.*, 2022; Shinoda *et al.*, 2016). Non-bilayer-forming lipids like DOPE can be added to mixed lipid bilayers to alter membrane properties, too. The presence of bound-lipid components and binding-partner proteins in the lipid bilayer environment optimises the folding of membrane proteins. This leads to functional and structurally intact proteins in the end.

Furthermore, studies focused on the determination of the exact lipid composition impact on the insertion and folding of the TMs of membrane proteins. It was concluded that the addition of bilayers with head group charge (DOPG) and increased lateral chain pressure (by the addition of DOPE) led to

more successful co-translation insertion and folding of various proteins (Harris *et al.*, 2022) such as rhomboid protease GlpG (Harris *et al.*, 2017), the GPCR β-1 adrenergic receptor (Rues, Dötsch and Bernhard, 2016), or the pentameric mechanosensitive channel (MscL) (Roos *et al.*, 2013; Harris *et al.*, 2018).

It is known that the lipid substrates can be provided in the cell-free protein synthesis reaction for the insertion and folding, often as either liposomes or nanodiscs (Harris *et al.*, 2022). The cell-free protein synthesis kits can vary in composition, as previously mentioned. Thus, a comparative analysis for the insertion and folding of the IVTT proteins obtained using the PURExpress protein synthesis kit, with lipids supplied in the form of liposomes, comprised the studies on XylE and LacY secondary transporters GlpG rhomboid protease, Connexin-43 small molecule channel, and DsbB disulphide bond enzyme.

In the cases of LacY and XylE, it was observed that both proteins preferred DOPG and DOPE as bilayers for insertion and folding (Harris, Pellowe and Booth, 2020). The same lipid composition that exhibits high lateral chain pressure and contains charge (DOPG and DOPE) was preferred by GlpG (Harris *et al.*, 2017). For the PURExpress IVTT connexin-43 channel, the reported experimental results indicate a preference for DOPC, and the insertion was observed to decrease when DPPC or DOPG were also introduced in the reaction (Moritani *et al.*, 2010). Neutral head group and reduced lateral chain pressure, as exhibited by DMPC was reported to favour the insertion of DsbB (Harris *et al.*, 2017) . Recent studies on the outer membrane proteins OmpA and BamA have shown that the excess of negative charge in the outer leaflets of asymmetric liposomes, impeded the insertion and folding and were favoured when the charge was found in the inner leaflet (Machin *et al.*, 2023).

The experiments based on the sucrose gradient separation determined the theoretical classification of the protein folding states, depending on the fluorescence of the sucrose gradient fractions. The top fraction was considered to contain the empty liposomes (due to the basal fluorescence intensity values), and the bottom fraction was considered to contain the aggregated protein, along with other CFPS kit components. The middle fraction of the sucrose gradient was considered to contain the cell-free expressed protein incorporated in the corresponding liposomes of various composition.

The liposome incorporation of REAMP-(S)sfYFP tended to increase with the increase of bilayer negative curvature, with the highest values obtained for POPC: POPE (10:1), as it was expected from previous experiments.

When CC1-(S)sfYFP was inserted into a liposome, an opposition relationship was observed with the fluorescence intensity results in the tested liposome compositions. The incorporation of CC1-(S)sfYFP showed a tendency to increase with the decrease of the negative curvature of the bilayer, with the highest

values observed for the POPC: DPhPC (1:1) liposomes. These results further show that the CC1 structure is stabilised by moderate interactions with the DPhPC lipid tails, and the insertion of the transmembrane helices of the CC1 is favoured in bilayers with reduced curvature. These results could be resembled to the study on connexin-43, for which it was also indicated that the increase in the lateral chain pressure and introduction of charges would be inhibitory for the protein insertion.

The increase of bilayer curvature, hence the modifications in the bilayer lateral pressure, would require additional energetic contribution for the insertion of the structure in curved bilayers. As the PURExpress cell-free protein synthesis kit might contain chaperones in the composition, and although the split version of the YFP should act as an additional folding contributor, they might still not be enough to fold the structure of the CC1-(S)sfYFP in highly curved bilayers, in the context of artificial expression systems.

In order to test this hypothesis, it would be of future interest to perform the experiments using as a DNA template the plasmid containing either the non-split version of the YFP or the non-tagged version (as described in Chapter 3), to attribute any folding contribution of the split version of the YFP.

Furthermore, other cell-free expression studies reported that supplementing the kit components with exogenous chaperones enhanced the folding of the newly synthesised proteins and prevented aggregation. This approach, however, would serve as an additional step in the comparative analysis of the influence on the overall folding mechanism of the YFP molecule.

Nevertheless, protein secondary structure determination experiments (such as circular dichroism) on the fractions containing the protein (middle and bottom) would further contribute to the establishment of a clear, valid relationship between the fluorescence measurements and the folding states of the cell-free expressed proteins, and consider it as a screening method in case of the use of other liposome compositions.

There are also reported other valuable techniques for the study of the secondary and tertiary structures during the co-translational folding. One of these techniques is represented by the surface-enhanced infrared spectroscopy (SEIRAS), and which has been proven successful for the co-translation folding studies of the bacteriorhodopsin (bR) (Baumann *et al.*, 2016), GlpG and DsbB (Harris *et al.*, 2017) in DMPC nanodiscs (Harris *et al.*, 2022).

4.5.7. Droplet interface bilayers stabilised by the insertion of the folded CFE protein structure

The results observed for the droplet interface bilayer assembly led to further insights regarding the interconnected protein-lipid stabilisation and interactions. The composition of the liposomes

influenced the formation of the droplet interface bilayers and the lifetime of the DIBs from droplets containing only CFE proteins and the corresponding lipids.

The stability (lifetime) of the formed DIB decreased depending on the following liposome composition order: POPC: DPhPC (1:1)> POPC: POPC (1:1)> POPC: POPE (10:1)> POPC: POPG (10:1)> POPC: POPS (10:1). Although the lifetime of the DIB formed by droplets containing POPC: DPhPC (1:1) was considered to be case-specific extended (approximately 480 s), the duration for the stability is relatively short, compared to reported DIB lifetimes (180 min) of other lipid composition (DOPC, DOPE and DOPG) (Findlay, Harris and Booth, 2021).

These results further support the observed tendency for the stabilisation of the folded state of CC1 by planar bilayers.

The DPhPC lipid tails adopt a special conformation when hydrophobic solvents surround them (Venkatesan *et al.*, 2018). The packing orientation changes from a relatively close packing of the tails to a disordered packing and the creation of a larger void in the lipid tails bilayer region, which would facilitate a spontaneous insertion of a transmembrane helix.

The artificial bilayer models containing DPhPC were reported to be efficient in multiple studies of membrane protein channels for testing functionality. This type of lipid was reported to reduce the exhibited lateral pressure profile, hence reducing the bias, which could impact the function of the channels.

The orientation of the DPhPC tails in solvent (in this case, undecane) would, in theory, permit the interdigitalisation of the tails, which are found in the proximate distance and the formation of a stable bilayer from monolayers with such architecture. However, this theory was contradicted to a certain extent by the results observed for the POPC: POPC bilayers in the cases of DIBs being formed by droplets containing empty liposomes only.

Interestingly, the presence of the CC1-(S)sfYFP in both droplets considerably extended the lifetime of the formed DIB.

This phenomenon could be explained by the introduction of further forces (van der Waals, π-π stacking, electrostatic, hydrogen bonds) by the presence of the CC1 structure in the bilayer. Considering the dissociation constants and the partitioning coefficients established by the pioneers of the study of membrane proteins insertion and folding, based on the amino acid composition of CC1, the K_d from water to “oil” environment (Wolfenden, 2007) would describe a spontaneous and more rapid distribution than for vice-versa. Indeed, the high glycine content in the CC1 sequence influences the theoretical partitioning values calculated for w/o and o/w (w= water; o= oil) to be empirically close, which would further explain

the tendency of preference of CC1 for planar lipid bilayers over bilayers under curvature stress for protein insertion.

In the case of POPC: DPhPC, the contact surface area is larger compared to the curved bilayers, creating a large bilayer which confers CC1 the possibility of insertion and establishment of a network of interactions with the surrounding lipid tails. In this way, the stability of the DIB and the association of the transmembrane helices of CC1 were influenced reciprocally.

The results observed for the stability of the DIBs formed by droplets containing POPC: POPS, POPC: POPG, or POPC: POPE liposomes could be explained by the decreased size of the contact surface between the monolayers, with the increase of bilayer curvature.

Furthermore, it is likely that the results obtained for DIBs having the liposome as mentioned above compositions are reversed compared to the results for the insertion efficiency of the folded cell-free expressed proteins due to the reverted biomechanics in the presence of the solvent. In further detail, the observed insertion was determined for liposomes found in an aqueous solution. In this case, the lipids influence the increase in the negative curvature of the bilayer as follows: PS< PG< PE. When the lipid tails were exposed to the hydrophobic solvent, the packing of the lipids changed, and the theoretical void volumes (Pichot, Watson and Norton, 2013) encountered in the head group regions and the lipid tail regions would change inverse proportionally, the properties of the lipids acting on the bilayer curvature reversely (theoretically influencing the increase of the negative bilayer curvature as follows: PE<PG<PS).

4.5.8. Material chemistry modification

Three-dimensional (3D) printing represents an emerging technology in engineering that has played an important role in biology and biotechnology for the past four decades. It enables the development of various tools, from high-throughput screening devices, molecular biology assays, and biochips for disease diagnosis to the maintenance of synthetic biological organisms for the study of evolutionary biology (Murphy and Atala, 2014).

This technology involves designing devices of small sizes with channels varying from 10 µm-100 µm. However, they can have larger features (1mm), and manipulating small volumes (fl, pl, nl, µl) of liquids (Niculescu *et al.*, 2021) is an advantage (cost-effective).

To overcome the limitations of PDMS and resin materials in forming stable droplet interface bilayers (DIBs), we investigated various surface coatings to modify the material chemistry and improve compatibility with the lipid-oil system. Coatings such as bovine serum albumin (BSA), polyethylene glycol (PEG20K), and styrene-maleic acid (SMA2000) did not significantly improve the surface chemistry for DIB

formation. BSA-coated surfaces exhibited non-uniformity, likely due to protein aggregation, while PEG20K and SMA2000 coatings failed to provide the desired hydrophobicity for stable bilayer formation. Dopamine coating of PDMS surfaces showed promise in stabilising DIBs, particularly for POPC: DPhPC (1:1) liposomes. The polydopamine layer likely interacted with the organic phase and PDMS sidechains, contributing to the formation of relatively stable bilayers. However, the inconsistency in results across different lipid compositions suggests that further optimisation of the coating process is necessary to ensure uniform surface modification. The most successful coating in our study was the acrylate-derived adhesive, which significantly improved the hydrophobicity of the resin surface. The water droplet test confirmed a uniform hydrophobic surface, with contact angles exceeding 90° after incubation in both water and silicone oil. This coating effectively eliminated the competing interaction forces between the resin surface and the lipid droplets, allowing for stable DIB formation. The acrylate-derived adhesive's success can be attributed to its chemical structure, fast curing time, and ease of application, making it a promising candidate for future microfluidic experiments involving DIB formation.

PDMS devices are widely used for biomedical purposes. The elastomer is reported to have excellent optical, electrical, and mechanical properties, and most importantly, it is a biocompatible material. In the experiments presented in this chapter, Sylgard 184 PDMS elastomer was used. It has further favourable properties; it is gas-permeable only and impermeable to liquids (Lamberti, Marasso and Cocuzza, 2014).

However, it was observed in the presented experiments that the surface of the PDMS deformed as a result of short incubation in the undecane-silicone oil phase. This observation suggested the swelling of the material, as it was previously shown; the PDMS has a preference for the absorption of non-polar solvents as undecane. In the literature it is described another disadvantage of PDMS, represented by the absorption of small hydrophobic molecules, as it was observed for the lipid droplets.

To overcome PDMS's limitations, chemical modification of the material's surface was tried, testing different coatings (enamel paint (Hagemann *et al.*, 2024), BSA (Azizipour *et al.*, 2022), and dopamine hydrochloride (Dabaghi *et al.*, 2021) solutions). The purpose of the coatings was to form a uniform surface compatible with the solvent phase, avoiding PDMS swelling and lipid droplet absorption.

The hydrophobic nature of PDMS, associated with the surface exposed methyl groups, leads to its lipophilic behaviour, resulting in the absorption of the lipid droplets. In the literature, it is also correlated to poor wettability and poor cell adhesion. In previous studies, it was shown that polydopamine (PD) coating represents an adaptor. The advantages of the PD coating include simplicity, the PD coating does not involve chemical synthesis, and the polymerisation is in-situ. The rationale for the current study was

not relying on improving the cell adhesion but introducing an adaptive layer to counteract the natural PDMS lipophilicity; however, the obtained results showed an inconsistent effect, and most of the lipid droplets were still absorbed by the PDMS. Interestingly, only the lipid droplets composed of POPC: DPhPC in the concentrations of 25 mg/ml and 2.5 mg/ml successfully formed a droplet-interface bilayer, which was stable for approximately 480 s, to which possibly the self-assembling characteristic of DPhPC lipids contributed.

The droplet formation and stability are influenced by a series of competing forces which can increase the entropy of the system. The electrostatic forces, van der Waals interactions, hydrogen bonds, as well as the overall lipid packing are influenced by the lipid head group structure. The structure of the lipid head groups further influences the packing of the lipid alkyl chains, which further affects the curvature of the bilayer. These forces count for the formation and stability of an individual droplet. The forces encountered for the packing of the lipid tails influence the formation and stability of the bilayer formed between droplets. In completion, gravitational force represents another competing force which influences the lipid droplet formation, contact and bilayer stability. It is known that the size of the phosphatidylethanolamine (PE) head group is smaller than the size of the phosphocholine (PC) head group, which leads to the dense packing of the lipid alkyl chains. The PE head group also has the ability to form intermolecular hydrogen bonds with the carbonyl oxygen from adjacent PE groups. These structural features of the PE head group would suggest the formation of a stable lipid monolayer, which would further stabilise the droplets.

The choline (PC) head groups are known to form weak hydrogen bonds or charge pairs with the phosphate groups or carbonyl groups from adjacent PC groups. In the context of a lipid droplet used in these experiments for the formation of DIBs, the head groups interact with each other. They are also exposed in individual droplets to the buffer solution (sodium phosphate buffer, pH 7.5), which is contained by the droplets. In contrast, the lipid tails are exposed towards the solvent phase (undecane: silicone oil). When POPC: POPE (10:1 ratio) lipid droplets were used, it was observed instability of bilayer and droplet fusion. Nonetheless the dense chain packing influenced by the PE head group and the bilayer stability, the presence of PE head groups in these experiments destabilised the system. In the case of single droplets, additional hydrogen bonds could occur between the PE head groups and the hydroxyl groups in the buffer, electrostatic (charge) interactions between the positive charge of PE and negative charge of phosphate groups on the buffer and between the choline (PC) head group and the phosphate groups of the buffer. All these forces combined within a single droplet already would create surface tension of a single lipid monolayer from inside-out and at the contact of two droplets, the result would be the fusion. It is highly

possible for the intrinsic forces exhibited by the individual droplets to be sufficient for the destabilisation of the system. When these forces were supplemented by the surface tension exhibited by the solvent phase, the result was droplet fusion.

It was further observed that no adaptive layer could compensate for the negative influence of these interactions. Hence, no coating had an impact on the DIB formation regardless of the substrates (resin or PDMS). When POPC: POPS lipid droplets were used, it could be also observed droplet fusion and no impact from the coating. Unlike PE, phosphatidylserine (PS) is a larger head group compared to PC, and it is negatively charged. It presents high electrostatic potential and can be involved in intermolecular hydrogen bond interactions. When POPC: POPC (1:1) lipid droplets were used, enamel paint as an adaptive layer had a positive impact on the bilayer formation and stability with resin as substrate. The absence of the PE head group diminished the number of intrinsic interactions which could occur; only remaining the weak hydrogen bonds exhibited between the adjacent PC head groups. In this case, the surface tension exhibited by the organic solvent phase and the lipid droplet absorption by the resin were alleviated by the enamel paint coating. The extrinsic competing forces were diminished by the π-π (pi-pi) stacking interactions between the phthalic acid residues in enamel paint and the aromatic rings in the resin's polyurethane composition (as specified in the material safety data sheet by the manufacturer Siraya Tech. U.S.) and silicone oil.

Enamel paint could exhibit interactions as well between the glycerol hydroxyl groups within its structure and the cyano group within the resin's polyacrylate chains. This coating was not successful; however, with PDMS as substrate, the hydrophobic forces exhibited by the PDMS could not be counteracted by the coating. Possibly, the coating could not be uniform due to coating-substrate chemical mismatches. Thus, the lipid droplets were absorbed by the PDMS material. In contrast to the positive influence of the coating on counteracting the extrinsic forces and stabilisation of the lipid bilayer, it was difficult to achieve the formation of the bilayer due to the neutral membrane curvature exhibited by the PC head group geometry (planar bilayers).

Relatively stable DIBs could be obtained when POPC: DPhPC (1:1) lipid droplets were used, with a positive effect from the dopamine coating of the PDMS mould. The introduction of DPhPC lipids improved the association of the lipid monolayers and the stability of the lipid bilayer. This can be explained mainly by the chemical structure of the lipid tails, which drive the overall packing, also having the choline as a head group, as POPC. DPhPC lipid tails being are composed of saturated alkyl chains, enhancing the stability of the monolayers, and branched with methyl groups.

The sterical mismatch introduced by the methylated acyl chains in a POPC: DPhPC monolayer

would result in the disorder in the chain configuration, the system re-arranging towards equilibration of disorders and leading to a better association between the monolayers zipper-like and once formed, stable bilayers. This lipid composition was not successful when resin moulds were used, possibly due to the high entropy resulting from the competing forces mentioned above. However, polydopamine-coated PDMS attenuated these forces, the polydopamine interacting with the organic phase and PDMS' sidechains, contributing to the formation of relatively stable DIBs between POPC: DPhPC droplets. Although polydopamine coating could represent a good adaptive layer, in the rest of the experiments, the results were not consistent with the ones obtained for POPC: DPhPC. This could be further explained by the non-uniform coating and the alteration of the physicochemical properties which occur due to the incorporation of Tris and unoxidised dopamine molecules. This can be further avoided by the replacement of the Tris buffer with phosphate buffer; however, the consequent deposition of PD nano aggregates (Hemmatpour *et al.*, 2023) could also lead to non-uniform coating.

Based on a similar principle related to droplet stabilisation, bovine serum albumin (BSA) coating was chosen. It was shown in previous studies that following BSA coating, the hydrophobic or (lipophilic as given in this case) characteristic of PDMS is lost, as desired in the present experiments. It could represent a better adaptive layer due to the ability to diminish the competing forces exhibited by the substrates (resin or PDMS), represented by its molecular size. However, it was observed that the obtained surface was still non-uniform, possibly due to BSA aggregates, as suggested in previous studies. The ATR-FT-IR studies on BSA-coated PDMS surfaces have shown large peaks occupied by amide moieties (Cámara-Torres *et al.*, 2021), present even after the washing of the surface, suggesting irreversible coating , compared to the PD coating. Experimental results have shown that BSA did not represent a good coating method for the desired applicability. Styrene-maleic acid and PEG20K solutions also did not represent good material coatings in order to successfully modify the surface chemistry for improvement of hydrophobicity.

The material coating using the cyano-acrylate-derived adhesive successfully modified the resin surface and has proven to provide a uniform hydrophobic surface, as suggested by the water droplet test results, after the incubation of the chip in either phase (water and silicone oil). The cyano-acrylate material coating thus allowed the effective bond coupling with the exposed resin material groups, thus eliminating the competing interaction forces with the droplets. Apart from the chemical structure advantages, it provided further experimental design advantages (fast curing times, ease of use and cost-effectiveness). The successful modification of surface chemistry using dopamine and acrylate-derived adhesive coatings opens up new possibilities for studying membrane proteins in DIBs. Future experiments could involve

incorporating these coatings in 3D-printed microfluidic chips to enable automated control of droplet formation, contact, and bilayer assembly, paving the way for more efficient and high-throughput studies of membrane protein function and interactions. The future experiments dedicated for the continuation of this chapter would be the inclusion of the adaptive coating in 3D printed microfluidic chips, for the automated control of droplet formation, contact and droplet-interface bilayer assembly.

4.5.9. Optimisation of experimental conditions

To improve the efficiency and reliability of our cell-free expression (CFE) experiments and droplet interface bilayer (DIB) formation, we identified several key areas for optimisation. One of the primary optimisations involves the choice of buffer for liposome resuspension. Our results suggest that the high ionic strength of the sodium phosphate buffer may have contributed to the sequestration of essential CFE reaction components and the inhibition of protein synthesis when liposomes were added at the start of the reaction. To mitigate this issue, we propose using a buffer with lower ionic strength, such as Tris or HEPES, which have been successfully employed in previous CFE studies (Findlay, Harris and Booth, 2021). However, careful analysis of the buffering system would need to be performed as not to preclude further biophysical experiments such as circular dichroism.

Another critical optimisation is the use of orbital shaking incubators instead of static incubators for CFE reactions. Proper mixing is essential for ensuring an even distribution of oxygen, which is crucial for the formation of the YFP chromophore. Orbital shaking would enhance the mixing of the reaction components, potentially improving the efficiency of protein synthesis and the resulting fluorescence intensities. Several other factors could be explored to further improve our experimental setup. For example, the inclusion of appropriate DNA and protein controls would provide a more comprehensive understanding of the relationships between fluorescence, expression, and folding of the CFE *de novo* proteins. Moreover, the use of automated 3D-printed microfluidic devices for DIB assembly and channel functional assays could enhance the reproducibility and throughput of our experiments. Continual optimisation of our experimental conditions is crucial for obtaining reliable and reproducible results in the study of *de novo* membrane proteins. By prioritising the most critical optimisations, such as buffer selection and mixing conditions, and exploring additional areas for improvement, we can enhance the efficiency and robustness of our CFE and DIB experiments. This, in turn, will enable more comprehensive studies of membrane protein folding, function, and interactions, ultimately advancing our understanding of these complex biological systems.

Moreover, it was observed a high sensitivity in rapport to the subtle changes in the environmental conditions (atmospheric temperature, pressure, aerodynamic flow patterns) of the biomechanics of the droplets used for the droplet-interface bilayer assembly. For this, the utilisation of the positive results obtained regarding the lipid concentrations and the surface chemistry modification, would be of interest to be used in experiments using automated 3D-printed microfluidic devices for the DIB assembly and channel functional assays. Future experiments regarding CC1 functionality would also include other types of liposome compositions, and based on the initial observed preference for planar bilayers, lipids with neutral head group charge and low lateral chain pressure, such as DMPC would be desirable.

4.5.10. Future impact

Cell-free expression of membrane proteins allows their production outside living cells. This enables researchers to study membrane protein properties in a controlled environment. Reconstitution in droplet-interface bilayers provides a platform to mimic biological processes in artificial cells. Consequently, researchers can study membrane protein functions and interactions.

Cell-free expression means no cells are involved when making membrane- proteins. This makes it easy for scientists to examine membrane protein qualities. They can carefully control the environment. Reconstituting membrane proteins in droplet-interface bilayers creates artificial cell-like structures. These mimic real cells, so researchers observe how membrane proteins work and interact. Droplet-interface-bilayers with reconstituted membrane proteins provide a robust, stable system. This system overcomes challenges when studying membrane proteins outside living cells. Sophisticated *in vitro* studies are possible using this setup. The incorporation of membrane transporters into droplet-interface bilayers allows for the investigation of specific, uphill transport reactions, providing insights into the transport mechanisms and functions of these proteins (Allen-Benton, Findlay and Booth, 2019).

Overall, the application of cell-free expression and reconstitution of membrane proteins in droplet-interface bilayers has significant implications for studying membrane protein properties, creating artificial membranes for biological studies, enhancing *in vitro* research capabilities, and facilitating functional studies on membrane transporters. As any technique, the studies involving droplet-interface bilayers have their own pitfalls, such as the limited scalability and the challenges introduced by the manipulation of small volumes of reactants (Baxani *et al.*, 2022). Despite this, it serves as a promising platform for drug screening, the development of biosensors, and optical tweezers (using lipids with photosensitive structures, such as azo-lipids) formed by droplet networks, which would further enable the

study of potential protein orthogonality.

5. Conclusions

Deciphering the few-decades-old enigmas of the relationship sequence-structure-function and of the folding mechanism of membrane proteins is a very challenging process. It does not only resume the understanding of the molecular chemistry of the structures and the associated forces involved but also the understanding of the impact of the surrounding environment, which, for membrane proteins, is trivial.

In this sense, advances in the field of synthetic biology and *de novo* protein design come in the aid of the existing methods, and aim to contribute with providing essential answers referring to the sequence-structure-function relationship problem. Advances in the implementation of computational techniques in these research fields, promoted the speed of research. However, we still find ourselves in the scientific era where computational predictions or computer-only-based studies, although they give unforeseen results, can still give errors or incomplete outputs, especially due to insufficient real-life-based training information. For this reason, the need of interdisciplinary studies and combining computational techniques with laboratory-based experiments are required.

De novo protein design aims to contribute to the design of small building blocks, which can self-associate into known or new structures, stabilising existing scaffolds. Thus, it enables the possibility of creating a library of small building blocks and their influence on structure formation and stability.

5.1. Design approach and computational framework

In this study, we aimed to bring the *de novo* membrane protein design to the limit to a certain extent. A completely new amino acid sequence was designed based on minimal and rational principles and statistical studies. This sequence appeared to stabilise an antiparallel tetrameric bundle in explicit membranes, as shown in molecular dynamics simulations. In this study it is also proposed an additional computational framework for design and sequence validation.

The theoretical computational results were assessed against controls (REAMP and poly-leucine bundles).

The CC1 sequence was designed using a multi-step approach. It did not include the use of coiled coils or GxxxG motifs. It was a reversed approach, starting from a given architecture and choosing a sequence able to stabilise an antiparallel tetramer, systematically. The rational choice of amino acids was based on the established statistical distribution across the membrane and the specific characteristics of other amino acids, reported in studies throughout time (lysine snorkelling effect, the aromatic belt, positive-inside rule). Using a parametrised poly-alanine structural template obtained from the CCCP online tool, the alanine residues were substituted with desired amino acids, based on real-time visual and energy score modifications in FoldIt software, until a sequence (repetitive across the antiparallel helices) that was able to provide a stable

structure was obtained.

The rationale for the choice of amino acids, completed by the impact exerted on the stability, was tested for impact on the structure in the context of an implicit membrane using the Hippo software. The sequence was sampled for tilt in a desired range, characteristic to membrane proteins, a test which was successfully passed by the obtained sequence. These computational tools are user-friendly and easily accessible and fast, being able to perform multiple iterations of the design strategy. It is also true that there is a plethora of structural predicting tools, having AlphaFold as representative and widely used for protein design and structure prediction. Most of the current computational-aided protein design are based on the use of algorithm-based calculations, Monte Carlo calculations etc., that sample thousands of sequence fragments and provide a new sequence with a predicted structure. The advantage of these methods is the speed of the study and the accuracy demonstrated for various applications, however, there are still some challenges related to the design of membrane-embedded proteins. These methods are fast for the fast sampling thousands of amino acid combinations and motifs.

Although the approach used in this study does not thoroughly sample the sequence space, or it could take a long time, it has the advantage of providing a clear and visual conclusion on the presence of each certain amino acid in different places of a TM helix and on the protein-protein interactions that could be involved.
The accuracy was introduced by the molecular dynamics simulations of the tetramer in the GROMACS software, analysing the structure stability of the tetramer, and possible protein-environment interactions with an explicit all-atom membrane model. The simulations in pure POPC, POPE and POPG bilayers have shown once again from the analysis of the root-mean-square deviations that the overall structure was stable across all membranes tested. It has further confirmed the lysine residues snorkelling effects and the influence of the aromatic belt in the transmembrane region of the helices.

The structural changes observed in POPE and POPG, such as an “open-close” mechanism, could represent a point for future investigation, starting with simulations in native-like membranes, having POPE and POPG in a more complex lipid environment.
For this part of the study, for future experiments, it would be interesting to combine the design tools that sample more thoroughly the sequence space, making additional substitutions in the sequence and to design the interfaces of the helices that would form the channel’s pore, this being the main objective of this study. It would be also desirable to complete the MD simulations, not only with simulations of natural-like lipid compositions, but also to simulate the exact design that would be used in the experiments (with links between the helices and tags).

5.2. Biocompatibility

The novel CC1 structure was also expressible in *E. coli* cells, which opens further perspectives for the assessment of expression in other host cells and to possibly re-design against other lipid compositions. The expression of the CC1-(S)sfYFP design re-emphasised on the case-specific experimental optimisation need. The results have shown that the C41 (DE3) strain accommodated better the protein, whilst when expressed in BL21(DE3), inclusion bodies were obtained. Also, it was shown that the growth conditions could be related to a better expression of the protein, terrific broth and lower incubation temperatures (25°C), increased the yield. The successful expression in bacterial cells, adds to the relevance of the CC1 design as a biocompatible *de novo* design, this representing a niche in the realm of *de novo* membrane protein design. Further comparative experiments need to be done with the YFP-tagged and the non-tagged CC1 constructs, to be able to assess the partition between the soluble fraction and the membrane fraction of the CC1 design. Nonetheless, circular-dichroism experiments would be highly desirable for this study.

5.3. Cell-free expression

The sequence was also expressible in a cell-free environment, enabling future study of its potential channel function, which was pre-implied by the conformational changes in the MD simulations, in the pre-established and tested artificial cell-like environments, such as DIBs. There were tested multiple lipid compositions for a possible correlation to the insertion efficiency of the protein in the bilayer, and it was seen a correlation between the negatively-curved bilayer-forming lipid compositions and the CC1-(S)sfYFP design, increasing from the negatively-curved toward planar bilayers. The most stable bilayer, which was also correlated to the best insertion efficiency, was for the POPC: DPhPC (1:1) lipid composition. For future experiments it would be desirable to test multiple lipid compositions for the effect on the insertion, as well as possibly finding other lipid compositions to form more stable bilayers rather than POPC: DPhPC, such as DMPC, DPPC bilayers, for example.

6. References

Arinaminpathy, Y. *et al.* (2009) 'Computational analysis of membrane proteins: the largest class of drug targets', *Drug Discovery Today*, 14(23–24), pp. 1130–1135. Available at: https://doi.org/10.1016/j.drudis.2009.08.006.

Baeza-Delgado, C. *et al.* (2016) 'Biological insertion of computationally designed short transmembrane segments', *Scientific Reports*, 6(March), pp. 1–9. Available at: https://doi.org/10.1038/srep23397.

Baker, J.A. *et al.* (2017) 'Charged residues next to transmembrane regions revisited: "Positive-inside rule" is complemented by the "negative inside depletion/outside enrichment rule"', *BMC Biology*, 15(1), pp. 1–29. Available at: https://doi.org/10.1186/s12915-017-0404-4.

Bonander, N. *et al.* (2005) 'Design of improved membrane protein production experiments: Quantitation of the host response', *Protein Science*, 14(7), pp. 1729–1740. Available at: https://doi.org/10.1110/ps.051435705.

Booth, P.J. (2012) 'A successful change of circumstance: A transition state for membrane protein folding', *Current Opinion in Structural Biology*, 22(4), pp. 469–475. Available at: https://doi.org/10.1016/j.sbi.2012.03.008.

Booth, P.J. and Clarke, J. (2010) 'Membrane protein folding makes the transition', *Proceedings of the National Academy of Sciences of the United States of America*, 107(9), pp. 3947–3948. Available at: https://doi.org/10.1073/pnas.0914478107.

Caillon, L. *et al.* (2020) 'Triacylglycerols sequester monotopic membrane proteins to lipid droplets', *Nature Communications*, 11(1). Available at: https://doi.org/10.1038/s41467-020-17585-8.

Chen, Y. *et al.* (2003) 'DnaK and DnaJ facilitated the folding process and reduced inclusion body formation of magnesium transporter CorA overexpressed in Escherichia coli', *Protein Expression and Purification*, 32(2), pp. 221–231. Available at: https://doi.org/10.1016/S1046-5928(03)00233-X.

Cheung, M.S., Chavez, L.L. and Onuchic, J.N. (2004) 'The energy landscape for protein folding and possible connections to function', *Polymer*, 45(2), pp. 547–555. Available at: https://doi.org/10.1016/j.polymer.2003.10.082.

Chorev, D.S. and Robinson, C. V. (2020) 'The importance of the membrane for biophysical measurements', *Nature Chemical Biology*, 16(12), pp. 1285–1292. Available at: https://doi.org/10.1038/s41589-020-0574-1.

Corin, K. and Bowie, J.U. (2022) 'How physical forces drive the process of helical membrane protein folding', *EMBO reports*, 23(3), pp. 1–20. Available at: https://doi.org/10.15252/embr.202153025.

Curnow, P. and Booth, P.J. (2009) 'The transition state for integral membrane protein folding', *Proceedings of the National Academy of Sciences of the United States of America*, 106(3), pp. 773–778. Available at:

https://doi.org/10.1073/pnas.0806953106.
Cymer, F., Von Heijne, G. and White, S.H. (2015) 'Mechanisms of integral membrane protein insertion and folding', *Journal of Molecular Biology*, 427(5), pp. 999–1022. Available at: https://doi.org/10.1016/j.jmb.2014.09.014.
Daley, D.O. *et al.* (2005) 'Biochemistry: Global topology analysis of the Escherichia coli inner membrane proteome', *Science*, 308(5726), pp. 1321–1323. Available at: https://doi.org/10.1126/science.1109730.
DeGrado, W.F., Gratkowski, H. and Lear, J.D. (2003) 'How do helix–helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles', *Protein Science*, 12(4), pp. 647–665. Available at: https://doi.org/10.1110/ps.0236503.
Dowhan, W. and Bogdanov, M. (2009) 'Lipid-dependent membrane protein topogenesis', *Annual Review of Biochemistry*, 78, pp. 515–540. Available at: https://doi.org/10.1146/annurev.biochem.77.060806.091251.
Elfaramawy, M.A. *et al.* (2018) 'Quantitative analysis of cell-free synthesized membrane proteins at the stabilized droplet interface bilayer', *Chemical Communications*, 54(86), pp. 12226–12229. Available at: https://doi.org/10.1039/c8cc06804f.
Engelman, D.M. *et al.* (2003) 'Membrane protein folding: Beyond the two stage model', *FEBS Letters*, 555(1), pp. 122–125. Available at: https://doi.org/10.1016/S0014-5793(03)01106-2.
Fallmann, J. *et al.* (2017) 'Recent advances in RNA folding', *Journal of Biotechnology*, 261, pp. 97–104. Available at: https://doi.org/10.1016/j.jbiotec.2017.07.007.
Freigassner, M., Pichler, H. and Glieder, A. (2009) 'Tuning microbial hosts for membrane protein production', *Microbial Cell Factories*, 8(December 2009). Available at: https://doi.org/10.1186/1475-2859-8-69.
Friddin, M.S. *et al.* (2016) 'Optically assembled droplet interface bilayer (OptiDIB) networks from cell-sized microdroplets', *Soft Matter*, 12(37), pp. 7731–7734. Available at: https://doi.org/10.1039/c6sm01357k.
Fujita, S., Kawamura, I. and Kawano, R. (2023) 'Cell-Free Expression of De Novo Designed Peptides That Form β-Barrel Nanopores', *ACS Nano*, 17(4), pp. 3358–3367. Available at: https://doi.org/10.1021/acsnano.2c07970.
Ghosh, I. *et al.* (2000) 'Antiparallel Leucine Zipper-Directed Protein Reassembly : Application to the Green Fluorescent Protein Department of Molecular Biophysics and Biochemistry The dissection and subsequent reassembly of a protein from peptidic fragments provides an avenue for', (11), pp. 5658–5659.
Gray, M., Will, S. and Jabbari, H. (2024) 'SparseRNAfolD: optimized sparse RNA pseudoknot-free folding with dangle consideration', *Algorithms for Molecular Biology*, 19(1), pp. 1–16. Available at: https://doi.org/10.1186/s13015-024-00256-4.
Grayson, K.J., Anderson, J.L.R. and Jlr, A. (2018) 'Designed for life : biocompatible de novo designed proteins

and components'.

Haylock, S. *et al.* (2020) 'Membrane protein mediated bilayer communication in networks of droplet interface bilayers', *Communications Chemistry*, 3(1), pp. 1–8. Available at: https://doi.org/10.1038/s42004-020-0322-1.

Von Heijne, G. (2006) 'Membrane-protein topology', *Nature Reviews Molecular Cell Biology*, 7(12), pp. 909–918. Available at: https://doi.org/10.1038/nrm2063.

Hessa, T. *et al.* (2005) 'Recognition of transmembrane helices by the endoplasmic reticulum translocon', *Nature*, 433(January), pp. 377–381.

Holden, M.A., Needham, D. and Bayley, H. (2007) 'Functional bionetworks from nanoliter water droplets', *Journal of the American Chemical Society*, 129(27), pp. 8650–8655. Available at: https://doi.org/10.1021/ja072292a.

Horne, J.E., Brockwell, D.J. and Radford, S.E. (2020) 'Role of the lipid bilayer in outer membrane protein folding in Gram-negative bacteria', *Journal of Biological Chemistry*, 295(30), pp. 10340–10367. Available at: https://doi.org/10.1074/jbc.REV120.011473.

Hylton, R.K. and Swulius, M.T. (2021) 'Challenges and triumphs in cryo-electron tomography', *iScience*, 24(9), p. 102959. Available at: https://doi.org/10.1016/j.isci.2021.102959.

Joh, N.H. *et al.* (2014) 'De novo design of a transmembrane zn2+-transporting four-helix bundle', *Science*, 346(6216), pp. 1520–1524. Available at: https://doi.org/10.1126/science.1261172.

Joh, N.H. *et al.* (2017) 'Design of self-assembling transmembrane helical bundles to elucidate principles required for membrane protein folding and ion transport', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1726). Available at: https://doi.org/10.1098/rstb.2016.0214.

Kawamala, B.K. and Abrol, R. (2022) 'Three-stage model of helical membrane protein folding: Role of membrane-water interface as the intermediate stage vestibule for TM helices during their in membrano assembly', *Biochemical and Biophysical Research Communications*, 624, pp. 1–7. Available at: https://doi.org/10.1016/j.bbrc.2022.07.006.

Kleinschmidt, J.H. (2003) 'Membrane protein folding on the example of outer membrane protein A of Escherichia coli', *Cellular and Molecular Life Sciences*, 60(8), pp. 1547–1558. Available at: https://doi.org/10.1007/s00018-003-3170-0.

Kram, K.E. and Finkel, S.E. (2015) 'Rich medium composition affects Escherichia coli survival, glycation, and mutation frequency during long-term batch culture', *Applied and Environmental Microbiology*, 81(13), pp. 4442–4450. Available at: https://doi.org/10.1128/AEM.00722-15.

Leach, M.D. and Cowen, L.E. (2014) 'Membrane fluidity and temperature sensing are coupled via circuitry

comprised of Ole1, Rsp5, and Hsf1 in Candida albicans', *Eukaryotic Cell*, 13(8), pp. 1077–1084. Available at: https://doi.org/10.1128/EC.00138-14.

Lu, P. *et al.* (2018) 'Accurate computational design of multipass transmembrane proteins', *Science*, 359(6379), pp. 1042–1046. Available at: https://doi.org/10.1126/science.aaq1739.

Machin, J.M. *et al.* (2023) 'Protein–lipid charge interactions control the folding of outer membrane proteins into asymmetric membranes', *Nature Chemistry*, 15(12), pp. 1754–1764. Available at: https://doi.org/10.1038/s41557-023-01319-6.

Mbaye, M.N. *et al.* (2019) 'A comprehensive computational study of amino acid interactions in membrane proteins', *Scientific Reports*, 9(1), pp. 1–14. Available at: https://doi.org/10.1038/s41598-019-48541-2.

Meyer, E.A., Castellano, R.K. and Diederich, F. (2003) *Interactions with aromatic rings in chemical and biological recognition*, *Angewandte Chemie - International Edition*. Available at: https://doi.org/10.1002/anie.200390319.

Monné, M., Hermansson, M. and Von Heijne, G. (1999) 'A turn propensity scale for transmembrane helices', *Journal of Molecular Biology*, 288(1), pp. 141–145. Available at: https://doi.org/10.1006/jmbi.1999.2657.

Mravic, M. *et al.* (2019) 'Packing of apolar side chains enables accurate design of highly stable membrane proteins', *Science*, 363(6434), pp. 1418–1423. Available at: https://doi.org/10.1126/science.aav7541.

Mueller, L.K. *et al.* (2020) 'Challenges and Perspectives in Chemical Synthesis of Highly Hydrophobic Peptides', *Frontiers in Bioengineering and Biotechnology*, 8(March), pp. 1–17. Available at: https://doi.org/10.3389/fbioe.2020.00162.

Narayanan, A., Ridilla, M. and Yernool, D.A. (2011) 'Restrained expression, a method to overproduce toxic membrane proteins by exploiting operator-repressor interactions', *Protein Science*, 20(1), pp. 51–61. Available at: https://doi.org/10.1002/pro.535.

Newberry, R.W. *et al.* (2013) 'n →π* Interactions of Amides and Thioamides: Implications for Protein Stability', *Journal of the American Chemical Society*, 135(21), pp. 7843–7846. Available at: https://doi.org/10.1021/ja4033583.

Pan, X. and Vachet, R.W. (2022) 'Membrane Protein Structures and Interactions From Covalent Labeling Coupled With Mass Spectrometry', *Mass Spectrometry Reviews*, 41(1), pp. 51–69. Available at: https://doi.org/10.1002/mas.21667.

Pandi, A. *et al.* (2023) 'Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides', *Nature Communications*, 14(1), pp. 1–14. Available at: https://doi.org/10.1038/s41467-023-42434-9.

Pedro, A.Q., Queiroz, J.A. and Passarinha, L.A. (2019) 'Smoothing membrane protein structure

determination by initial upstream stage improvements', *Applied Microbiology and Biotechnology*, 103(14), pp. 5483–5500. Available at: https://doi.org/10.1007/s00253-019-09873-1.

Penel, S. and Doig, A.J. (2001) 'Rotamer strain energy in protein helices - Quantification of a major force opposing protein folding', *Journal of Molecular Biology*, 305(4), pp. 961–968. Available at: https://doi.org/10.1006/jmbi.2000.4339.

Pleshakova, T.O. *et al.* (2018) 'Atomic force microscopy for protein detection and their physicochemical characterization', *International Journal of Molecular Sciences*, 19(4). Available at: https://doi.org/10.3390/ijms19041142.

Pope, A.L. *et al.* (2020) 'A Conserved Proline Hinge Mediates Helix Dynamics and Activation of Rhodopsin', *Structure*, 28(9), pp. 1004-1013.e4. Available at: https://doi.org/10.1016/j.str.2020.05.004.

Popot, J.L. and Engelman, D.M. (1990) 'Membrane Protein Folding and Oligomerization: The Two-Stage Model', *Biochemistry*, 29(17), pp. 4031–4037. Available at: https://doi.org/10.1021/bi00469a001.

Reckel, S. *et al.* (2010) 'Strategies for the cell-free expression of membrane proteins.', *Methods in molecular biology (Clifton, N.J.)*, 607, pp. 187–212. Available at: https://doi.org/10.1007/978-1-60327-331-2_16.

Reeder, J. *et al.* (2006) 'Beyond Mfold: Recent advances in RNA bioinformatics', *Journal of Biotechnology*, 124(1), pp. 41–55. Available at: https://doi.org/10.1016/j.jbiotec.2006.01.034.

Rees, D.C. and Engelman, D.M. (2017) 'Forces involved membrane in the assembly proteins', *The FASEB Journal*, 6(15), pp. 3397–3402.

Van Rosmalen, M., Krom, M. and Merkx, M. (2017) 'Tuning the Flexibility of Glycine-Serine Linkers to Allow Rational Design of Multidomain Proteins', *Biochemistry*, 56(50), pp. 6565–6574. Available at: https://doi.org/10.1021/acs.biochem.7b00902.

Sachse, R. *et al.* (2014) 'Membrane protein synthesis in cell-free systems: From bio-mimetic systems to bio-membranes', *FEBS Letters*, 588(17), pp. 2774–2781. Available at: https://doi.org/10.1016/j.febslet.2014.06.007.

Schafmeister, C.E. *et al.* (1997) 'A designed four helix bundle protein with native-like structure', *Nature Structural Biology*, 4(12), pp. 1039–1046. Available at: https://doi.org/10.1038/nsb1297-1039.

Schmitt, C. *et al.* (2016) 'Compartmentalization and transport in synthetic vesicles', *Frontiers in Bioengineering and Biotechnology*, 4(FEB), pp. 1–12. Available at: https://doi.org/10.3389/fbioe.2016.00019.

Senes, A. (2011) 'Computational design of membrane proteins', *Current Opinion in Structural Biology*, 21(4), pp. 460–466. Available at: https://doi.org/10.1016/j.sbi.2011.06.004.

Sezgin, E. and Schwille, P. (2012) 'Model membrane platforms to study protein-membrane interactions', *Molecular Membrane Biology*, 29(5), pp. 144–154. Available at:

https://doi.org/10.3109/09687688.2012.700490.
Shi, L. *et al.* (2002) 'β2 adrenergic receptor activation: Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch', *Journal of Biological Chemistry*, 277(43), pp. 40989–40996. Available at: https://doi.org/10.1074/jbc.M206801200.
Silva, D.A. *et al.* (2019) 'De novo design of potent and selective mimics of IL-2 and IL-15', *Nature*, 565(7738), pp. 186–191. Available at: https://doi.org/10.1038/s41586-018-0830-7.
Singh, A. *et al.* (2015) 'Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process', *Microbial Cell Factories*, 14(1), pp. 1–10. Available at: https://doi.org/10.1186/s12934-015-0222-8.
Skach, W.R. (2009) 'Cellular mechanisms of membrane protein folding', *Nature Structural and Molecular Biology*, 16(6), pp. 606–612. Available at: https://doi.org/10.1038/nsmb.1600.
Slovic, A.M. *et al.* (2004) 'Computational design of water-soluble analogues of the potassium channel KcsA', *Proceedings of the National Academy of Sciences of the United States of America*, 101(7), pp. 1828–1833. Available at: https://doi.org/10.1073/pnas.0306417101.
De Smit, M.H. and Van Duin, J. (1990) 'Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis', *Proceedings of the National Academy of Sciences of the United States of America*, 87(19), pp. 7668–7672. Available at: https://doi.org/10.1073/pnas.87.19.7668.
Sparks, K.A. *et al.* (2014) 'Comparisons of interfacial phe, tyr, and trp residues as determinants of orientation and dynamics for GWALP transmembrane peptides', *Biochemistry*, 53(22), pp. 3637–3645. Available at: https://doi.org/10.1021/bi500439x.
Streit, J.O. *et al.* (2024) 'The ribosome lowers the entropic penalty of protein folding', *Nature*, 633(8028), pp. 232–239. Available at: https://doi.org/10.1038/s41586-024-07784-4.
Syeda, R. *et al.* (2008) 'Screening blockers against a potassium channel with a droplet interface bilayer array', *Journal of the American Chemical Society*, 130(46), pp. 15543–15548. Available at: https://doi.org/10.1021/ja804968g.
Ulmschneider, M.B. and Sansom, M.S.P. (2001) 'Amino acid distributions in integral membrane protein structures', *Biochimica et Biophysica Acta - Biomembranes*, 1512(1), pp. 1–14. Available at: https://doi.org/10.1016/S0005-2736(01)00299-1.
Villar, G. and Bayley, H. (2013) 'Functional Droplet Interface Bilayers', in *Encyclopedia of Biophysics*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 861–868. Available at: https://doi.org/10.1007/978-3-642-16712-6_567.
Volkenborn, K. *et al.* (2020) 'The length of ribosomal binding site spacer sequence controls the production yield for intracellular and secreted proteins by Bacillus subtilis', *Microbial Cell Factories*, 19(1), pp. 1–12.

Available at: https://doi.org/10.1186/s12934-020-01404-2.
Vorobieva, A.A. *et al.* (2021) 'De novo design of transmembrane β barrels', *Science*, 371(6531), pp. 1–25. Available at: https://doi.org/10.1126/science.abc8182.
Wimley, W.C. (2003) 'The versatile β-barrel membrane protein', *Current Opinion in Structural Biology*, 13(4), pp. 404–411. Available at: https://doi.org/10.1016/S0959-440X(03)00099-X.
Winogradoff, D., John, S. and Aksimentiev, A. (2020) 'Protein unfolding by SDS: the microscopic mechanisms and the properties of the SDS-protein assembly', *Nanoscale*, 12(9), pp. 5422–5434. Available at: https://doi.org/10.1039/C9NR09135A.