Abstract

Surfactants are ubiquitous molecules, holding a $33 billion market (2014) due to a broad range of products from cosmetics, cleaning reagents to foods or pharmaceuticals. Surfactants exhibit activity at interfaces between aqueous and non-aqueous phases by reducing surface tension, wetting, foaming, and emulsification. However, currently these molecules are mainly chemically synthesized, relying on petrochemical sources and are thus governed by petroleum prices, negative environmental impact and the limited resources of fossil fuels.

Biosurfactants are a green alternative as they are surfactants produced by microorganisms, enabling companies to meet current demands for industrial sustainability and address the increasing awareness of consumers. They can potentially be produced from cheap renewable feedstocks, are robust enough to be used in industrial processes and yet biodegradable and eco-friendly. In particular protein or peptide biosurfactants additionally offer vast design opportunities and variability in their properties due to the flexibility in amino acid side chain characteristics. This is why they are particularly used where biocompatibility or additional functionality is required.

However problems associated with yield, cost of production and purification, and the tailoring of the molecules for specific applications due to a lack of fundamental understanding, make these molecules currently not economically competitive with cheaper conventional chemical surfactants.

DAMP4, designed by Anton Middelberg, is a member of a family of four-helix bundle biosurfactant proteins, and addresses key issues frequently related to biosurfactants, as it can be expressed at remarkably high yield in Escherichia coli and can be purified by a simple and cost-effective process due to its extraordinarily high thermostability. Further DAMP4 is a foaming surfactant, and foaming behaviour can be controlled by the pH. It is highly surface active due to a conformational change upon interfacial adsorption. This study aimed to create a more fundamental understanding of the underlying sequential and structural features supporting these characteristics enabling specific design of new highly functional four-helix bundle biosurfactants for particular applications while keeping the production costs low enough for industrial use. This was achieved by applying all-atom molecular dynamics simulation to establish the link between the biosurfactants’ sequence, three-dimensional structure and characteristics such as overexpression yield, purification and interfacial behavior.

In order to keep the cost low, high expression yields are required for newly designed four-helix bundle biosurfactant proteins. This is a sequence-dependent feature, and could prior to this thesis only be determined by experiments. To save time and expenses to identify high expressing proteins in the future when new designs will be created, a test tool was developed based on the combination of the behavior of these structures in a molecular dynamics simulation and a statistical classifier.
This prediction model was based on data from our four-helix bundle library, and was successful in differentiating highly expressing four-helix bundles from those that were expected to be problematic upon overexpression.

To further be economically competitive new designs are required to maintain a high level of stability, as for the cost-effective purification process high temperatures (>90 °C) in combination with a chaotropic salt induce bacterial cell lysis and denaturation of all bacterial contaminants, while the four-helix bundle DAMP4 remains stable in solution and can be recovered with solid-liquid separation. With different strategically-designed \textit{in silico} DAMP4 variants we identified the tightly packed hydrophobic core as the most important structural feature for its high thermal stability. In addition the simulation showed integration of up to three hydrophilic residues into the hydrophobic core is possible without major loss of stability, maintaining the ability to be purified with the above process. This makes them adaptable for a range of applications. Despite the high thermal stability of DAMP4, experiments and foaming behavior indicate a conformational rearrangement of DAMP4 from stable four-helix bundle in bulk, to a chain of four single helices at the interface. It was unknown, how the rearrangement occurred, and what structural and environmental features were triggering it. The simulations showed an unfolding of the bundle with the opening of helices 1 and 4 which are parallel to the interface, with the final conformation as the four helices being in parallel, hydrophobic residues facing the air phase, hydrophilic ones facing the bulk water. In the simulations this behavior was triggered by a change in environment, namely the pH shifting from neutral to acidic or basic conditions, a correlated change in protonation state leading to an access of positive or negative charges, together with general fluctuations of the flexible molecule near the interface and the right orientation with helices 1 and 4 facing upwards.

By applying molecular dynamics simulations on this family of four-helix bundles, including purposefully designed \textit{in silico} variants, this thesis gave new insights into the sequence-structure-function relationships of protein biosurfactants. It contributed to the knowledge of how to design these molecules in order to overcome cost-barriers while maintaining high functionality as well as create other desired properties and interactions.
Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the Copyright Act 1968.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.
Publications during candidature

Peer-reviewed papers


Oral presentations


Poster presentations


**Publications included in this thesis**


- Incorporated as Chapter 3

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Statement of contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrea Schaller (Candidate)</td>
<td>Designed simulations (50%)</td>
</tr>
<tr>
<td></td>
<td>Performed simulations (100%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (80%)</td>
</tr>
<tr>
<td></td>
<td>Wrote paper (100%)</td>
</tr>
<tr>
<td>Natalie K. Connors</td>
<td>Designed simulations (20%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (45%)</td>
</tr>
<tr>
<td>Stefan. A. Oelmeier</td>
<td>Designed simulations (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (5%)</td>
</tr>
<tr>
<td>Juergen Hubbuch</td>
<td>Edited paper (5%)</td>
</tr>
<tr>
<td>Anton P. J. Middelberg</td>
<td>Designed simulations (20%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (45%)</td>
</tr>
</tbody>
</table>


- Incorporated as Chapter 4
<table>
<thead>
<tr>
<th>Contributor</th>
<th>Statement of contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrea Schaller (Candidate)</td>
<td>Designed simulations (50%)</td>
</tr>
<tr>
<td></td>
<td>Performed simulations (100%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (50%)</td>
</tr>
<tr>
<td></td>
<td>Wrote paper (100%)</td>
</tr>
<tr>
<td>Natalie K. Connors</td>
<td>Designed simulations (20%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (30%)</td>
</tr>
<tr>
<td>Mirjana Dimitrijev Dwyer</td>
<td>Designed experiments (50%)</td>
</tr>
<tr>
<td></td>
<td>Performed experiments (100%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (30%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (30%)</td>
</tr>
<tr>
<td>Stefan. A. Oelmeier</td>
<td>Designed simulations (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (5%)</td>
</tr>
<tr>
<td>Juergen Hubbuch</td>
<td>Edited paper (5%)</td>
</tr>
<tr>
<td>Anton P. J. Middelberg</td>
<td>Designed simulations (20%)</td>
</tr>
<tr>
<td></td>
<td>Designed experiments (50%)</td>
</tr>
<tr>
<td></td>
<td>Analysed data (10%)</td>
</tr>
<tr>
<td></td>
<td>Edited paper (30%)</td>
</tr>
</tbody>
</table>
Contributions by others to the thesis

Prof. Anton Middelberg: Conception and design of this project, thesis design, and critical review.
Prof. Juergen Hubbuch: Prof. Design of this project, critical review.
Dr. Natalie Connors: Simulation design, critical review.
Prof. Debra Bernhardt: Design of Chapter 5, critical review.
Dr. Stefano Bernardi: PMF simulations in Chapter 5, design and setup.
Dr. Stefan Oelmeier: Simulation design, critical review.
Dr. Mirjana Dimitrijev Dwyer: Helped perform LCMS measurements in Chapter 3.
Alice Yu: Helped perform expression experiments in Chapter 3.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.
Acknowledgements

I am very grateful to Prof. Hubbuch and Prof. Middelberg for giving me the great opportunity to work on this project. I would like to thank them for their continuous support, help and guidance throughout, as well as all their ideas and mentoring during my PhD. I would like to further thank Prof. Bernhardt for joining my advisory team during the last year of my thesis and who contributed greatly with her computational and simulation knowledge, and encouraged with her kind words. I am grateful for the day-to-day supervision by Dr. Natalie Connors, who guided me from my first days at CBE and who I learned from a lot.

I would like to thank all members of CBE, CTCMS and the Nielsen Group for their daily support, help, advice, and companionship during the last three and a half years here at AIBN. Many thanks to my family, who are always there for me and support me in everything I do.

I am grateful for financial support by an UQI Tuition Fee Scholarship and an AIBN RHD Living Scholarship funded by an ARC Discovery Grant awarded to Professors Middelberg and Hubbuch. Further I would like to gratefully acknowledge the allocation of computational time on their supercomputers by the Research Computing Centre (RCC) at the University of Queensland and the National Computational Infrastructure (NCI).
Keywords
biosurfactant, four-helix bundle, molecular dynamics simulation, expression, stability, interface.

Australian and New Zealand Standard Research Classifications (ANZSRC)
ANZSRC code: 030402, Biomolecular Modelling and Design, 50%
ANZSRC code: 090499, Chemical Engineering not elsewhere classified, 30%
ANZSRC code: 100306, Industrial Molecular Engineering of Nucleic Acids and Proteins, 20%

Fields of Research (FoR) Classification
FoR code: 0306, Physical Chemistry (incl. Structural), 50%
FoR code: 0904, Chemical Engineering, 30%
FoR code: 1003, Industrial Biotechnology, 20%
# Table of contents

Abstract ................................................................................................................................. ii
Declaration by author ........................................................................................................ iv
Publications during candidature ........................................................................................ v
Publications included in this thesis ....................................................................................... vi
Contributions by others to the thesis .................................................................................... viii
Statement of parts of the thesis submitted to qualify for the award of another degree ....... viii
Acknowledgements ............................................................................................................... ix
Keywords ............................................................................................................................... x
Australian and New Zealand Standard Research Classifications (ANZSRC) ................. x
Fields of Research (FoR) Classification .............................................................................. x
Table of contents ................................................................................................................ xi
List of figures and tables ...................................................................................................... xiv
List of abbreviations used in the thesis ................................................................................ xviii

## Chapter 1 Introduction ........................................................................................................ 1
  1.1 Research Objectives .................................................................................................... 4
  1.2 Thesis structure ......................................................................................................... 5
  1.3 References .................................................................................................................. 7

## Chapter 2 Literature review .............................................................................................. 8
  2.1 Surfactants .................................................................................................................. 8
    2.1.1 Chemical surfactants ............................................................................................ 8
    2.1.2 Biosurfactants ...................................................................................................... 12
    2.1.3 Recombinant protein expression in Escherichia coli ......................................... 19
    2.1.4 Protein adsorption to interfaces: Foam formation and stability ......................... 20
  2.2 Molecular dynamics simulation .................................................................................. 23
    2.2.1 History and overview .......................................................................................... 23
    2.2.2 Introduction to MD methodology ......................................................................... 25
    2.2.3 Structure modelling .............................................................................................. 27
    2.2.4 Method ................................................................................................................ 31
    2.2.5 High temperature MD simulation to assess unfolding and stability .................... 33
    2.2.6 MD simulation at interfaces ............................................................................... 34
    2.2.7 Potential of mean force (PMF) ............................................................................ 35
Chapter 5: Adsorption behaviour of a four-helix bundle biosurfactant to an air/water interface

5.1 Introduction

5.2 Material and methods

5.2.1 Molecular dynamics simulation

5.2.2 Unfolding simulations of DAMP4 in the presence of an air/water interface with ions

5.2.3 Potential of mean force (PMF) calculation

5.2.4 Calculation of water distribution around DAMP4

5.2.5 Calculation of dipole moments and quantification of orientation of water molecules

5.3 Results and discussion

5.3.1 Investigation of the water structure at the interface and around DAMP4 for unfolding

5.3.2 Other factors influencing DAMP4 adsorption to the interface: Drivers, mechanism of unfolding and conformation of DAMP4 at the interface

5.4 Conclusion

5.5 References

Chapter 6: Discussion

6.1 Key research findings

6.2 Outlook

6.3 Conclusion

6.4 References

Appendix A: Supplementary Information for Chapter 5
List of figures and tables

List of figures

Figure 2-1  Structures of a surfactant and typical micelle configurations ................................................. 8
Figure 2-2  Schematic structures of different surfactants. ............................................................................. 9
Figure 2-3  Structural comparison of a) classical chemical surfactants and b) peptide or c) protein surfactants. ........................................................................................................................................... 10
Figure 2-4  Possible arrangements for surfactants hydrophobic (filled circle) and hydrophilic (empty circle) moieties in surfactants ......................................................................................... 12
Figure 2-5  Schematic depiction of the “Bake-to-break and precipitate” process used for DAMP4 purification. ........................................................................................................................................... 15
Figure 2-6  Foaming behaviour of DAMP4 under different pH conditions. ................................................... 16
Figure 2-7  Concentration-dependent interfacial tension reduction of DAMP1 and DAMP4. ......................... 17
Figure 2-8  Suggested possible pathways for DAMP4’s adsorption to an air/water interface. ................. 18
Figure 2-9  Schematic representation of the structure of a foam .................................................................. 21
Figure 2-10  Stages of the adsorption process of proteins to an air/water interface. ................................... 21
Figure 2-11  Time and length scale of various biophysical techniques. ......................................................... 26
Figure 2-12  Schematic illustration of theoretical model prediction methods a) homology modelling and b) fold recognition ............................................................................................................ 28
Figure 2-13  Schematic representation of periodic boundary conditions. ......................................................... 33
Figure 3-1  SDS-PAGE gel showing NOVEX® Sharp Pre-stain molecular weight ladder (lane 1), DAMP4 prior to induction of recombinant protein expression (lane 2) and when harvested (lane 3), as an example of soluble expression. Lanes 4 & 5 are for sequence SP17 and lanes 6 & 7 for SP18 at induction and at harvest, respectively. Both sequences were not expressed under these conditions. The image was produced using Bio-Rad Gel-Doc™ imaging system ........................................................................................................ 48
Figure 3-2  The 3-dimensional structures of DAMP4 a) front view and c) top view and ASProt b) front view and d) top view, derived from homology modelling ............................................. 49
Figure 3-3  Rmsd raw data in production simulation of single structures showing destabilization and partial unfolding ........................................................................................................................................ 54
Figure 3-4  Flow chart of model building process .......................................................................................... 55
Figure 3-5  Model descriptor Energyelectrostatic for investigated sequences ................................................. 56
Figure 3-6  The 3-dimensional structure of SP12 .......................................................................................... 57
Figure 3-7  Progression of simulation variable Energytotal for DAMP4 ......................................................... 58
Figure 3-8  The 3-dimensional structure of EDP-11 ...................................................................................... 60
Figure 3-9  SDS page gels for a) EDP-11 before induction (lane 2 (T) & 3 (S)) and 2 h (lane 4 (T) & 5 (S)), 5 h (lane 6 (T) & 7 (S)) and 8 h post induction (lane 8 (T) & 9 (S)) with standard ladder (lane 1). b) QDP-11 before induction (lane 2 (T) & 3 (S)) and 2 h (lane 4 (T) & 5 (S)), 5 h (lane 6 (T) and lane 7 (S)) and 8 h post induction (lane 8 (T) & 9 (S)) with standard ladder (lane 1). T stands for total protein, and S for soluble protein..............61

Figure 3-10 Simulation data as used for prediction model for variants AS1-AS5 averaged over three runs...........................................................................................................................................67

Figure 3-11 Comparison of data for the prediction descriptors a) Energy, b) Energy\_bond, and c) Energy\_electrostatic for 1 ns (o) and 25 ns (x)..............................................................................................68

Figure 3-12 SDS-PAGE gels showing expression results of AS variants..........................................................69

Figure 3-13 LCMS results for heat purified DAMP4 a) HPLC, b) MS, and heat purified AS5 c) HPLC, d) MS. .................................................................................................................71

Figure 3-14 SDS-PAGE gel of AS5 and DAMP4. ........................................................................................................72

Figure 3-15 SDS PAGE gel of AS5 after heat process. .....................................................................................73

Figure 4-1  Fold structure of the biosurfactant protein DAMP4. ..................................................................81

Figure 4-2  Differential scanning calorimetry showing a melting point above 120 °C. ....................84

Figure 4-3  Circular dichroism spectra of DAMP4 at a) 20 °C and b) 90 °C........................................84

Figure 4-4  Stability of DAMP4 at different concentrations of GdmCl at 20 °C. .............................85

Figure 4-5  Simulation results of DAMP4 stability variants. ........................................................................87

Figure 4-6  Simulation results of DAMP4 and hydrophobic core variants. .............................................90

Figure 4-7  Selected structures of hydrophobic variants after 100 ns simulation..................................91

Figure 4-8  a) Top view, designated C\_\alpha distances between the four helices at the bottom b) and top c) of four-helix bundles and d) difference in solvent accessible surface area compared to DAMP4 for selected hydrophobic variants.................................................................92

Figure 4-9  Multiple sequence alignment[58] for monomer DAMP1 and homology modelling template 3EDC of E. coli’s lactose operon repressor. .................................................................95

Figure 4-10 Ramachandran plot for DAMP4 backbone atoms.................................................................96

Figure 5-1  Structural characteristics of DAMP4. .....................................................................................103

Figure 5-2  Distribution function of water molecules around DAMP4 as defined by equation (28). .................................................................................................................................107

Figure 5-3  Visualization of location of cuts of DAMP4 along the z-axis in 5 Å slices to identify hydration sites. ....................................................................................................................108

Figure 5-4  Cross sections of 3D plot of location of water molecules around DAMP4 in 10 ns simulation with the protein fixed at its initial location in 5 Å slices from a) z=-20 to h) z=+20 as visualized in Figure 5-3. .................................................................109
Figure 5-5  Dipole moments of all waters in 1 Å bins along the z-coordinate normalized by the number of molecules..........................................................................................................................110

Figure 5-6  Angle of the dipole moment vector of each water molecule and the xy-plane..............111

Figure 5-7  Magnitude of average dipole moment vectors.........................................................................................................................111

Figure 5-8  Starting configuration for DAMP4 unfolding simulations in the presence of an air/water interface................................................................................................................................................115

Figure 5-9  Suggested unfolding pathway of DAMP4 into “W-conformation” when an interface is approached. ..............................................................................................................................................116

Figure 5-10  Averaged hydrophobic solvent accessible surface areas of DAMP4 in the presence of an air/water interface......................................................................................................................117

Figure 5-11  Top (Pro 3-Leu 94), bottom (Glu 23-Ser 76) and middle His-His (His 13-His 85) distances for DAMP4 unfolding simulations at different ion- and pH conditions to follow the pathway of opening..................................................................................................................118

Figure 5-12  DAMP4 unfolding pathway at an air/water interface as observed in MD simulation. 120

Figure 5-13  Secondary structure contents of DAMP4 during the unfolding simulation are maintained....................................................................................................................................................121

Figure 5-14  Potential of mean force relative to the initial distance or free energy and mean force for unfolding of DAMP4 in varying interfacial conditions.............................................................................122

Figure 5-15  Structure of DAMP4 at pH 4.8.................................................................................................................................124

Figure 5-16  Summary of the trigger and pathway of DAMP4 unfolding at an air/water interface. 126

Figure 6-1  Flow chart of the aim of this thesis. ................................................................................................................................134
List of tables

Table 3-1  Four-helix bundle sequences[39] used to build the model (Training Set) and independent sequences used to test predictions (Test Set). Solubility in overexpression was determined experimentally.................................................................47

Table 3-2  Prediction results into insoluble (I) and soluble (S) from existing prediction tools of our set of four-helix bundle sequences.................................................................53

Table 3-3  Overview of the descriptors of selected prediction models that showed the best prediction performance. E stands for energy.................................................................55

Table 3-4  Details of the final prediction model (Model 1) with the three descriptors Energy\textsubscript{total}, Energy\textsubscript{bond} and Energy\textsubscript{electrostatic}. With the scale factors normalized support vectors $x_i$, to the support vectors corresponding weights $\alpha_i$ and bias $b$ to use in the decision function (Equation 21)..................................................................................................................59

Table 3-5  Model input data for sequences EDP-11 and QDP-11 as derived from MD simulation. Brackets indicate different starting structures.................................................................61

Table 3-6  DAMP4 variants created to validate prediction of expression outcomes using MD. 64

Table 4-1  Sequence design for DAMP4 \textit{in silico} stability variants for MD simulation. In these variants one of the suggested stability mechanisms (hydrophobic core, salt bridges, water shell) is eliminated or weakened. Red marks site of mutations in comparison to DAMP4.................................................................86

Table 4-2  Sequence design for hydrophobic variants with 1 to 8 Leu to hydrophilic Asp mutations in the hydrophobic core. Red marks site of mutations in comparison to DAMP4. 89

Table 4-3  Relative variance, variance and mean rmsd $C_\alpha$ as numerical descriptors for rmsd $C_\alpha$ plots depicted in Figure 4-5 and Figure 4-6. .................................................................96

Table 5-1  Overview of different starting ion locations for DAMP4 unfolding simulations for the three basic systems. Each of these systems was simulated three times with different starting conformation.................................................................102
# List of abbreviations used in the thesis

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIBN</td>
<td>Australian Institute for Bioengineering and Nanotechnology</td>
</tr>
<tr>
<td>BPTI</td>
<td>Bovine pancreatic trypsin inhibitor</td>
</tr>
<tr>
<td>CASP</td>
<td>Critical assessment of structure prediction</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>FTIR</td>
<td>Fourier transform infrared spectroscopy</td>
</tr>
<tr>
<td>GRAVY</td>
<td>Grand average of hydropathy</td>
</tr>
<tr>
<td>GSH</td>
<td>Glutathione</td>
</tr>
<tr>
<td>HPLC</td>
<td>High-performance liquid chromatography</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl-β-D-thiogalactopyranosid</td>
</tr>
<tr>
<td>kDa</td>
<td>Kilodalton</td>
</tr>
<tr>
<td>LC/MS</td>
<td>Liquid chromatography and mass spectrometry</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>PEF</td>
<td>Protein Expression Facility</td>
</tr>
<tr>
<td>pI</td>
<td>Isoelectric point</td>
</tr>
<tr>
<td>PMF</td>
<td>Potential of mean force</td>
</tr>
<tr>
<td>rmsd</td>
<td>Root mean square deviation</td>
</tr>
<tr>
<td>rpm</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulphate-polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>VdW</td>
<td>Van der Waals</td>
</tr>
<tr>
<td>VLP</td>
<td>Virus-like particle</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

Computational chemistry is an area of growing interest, as this discipline applies computational methods to chemical problems and thus opens new possibilities where classical experimental methods fail or experimental results cannot be understood in detail[1].

Molecular dynamics (MD) simulations are a widely used approach applied for the calculation of the behaviour of particles as a function of time based on physical principles. This method has a key advantage in that it allows the user to see dynamic movement of biomolecules in atomic detail in varying environmental conditions, which is largely not possible experimentally. These simulations provide the user with exact control of specific contributions, and by altering them their influence on particular properties in synergy with experiments. Developments in the accuracy of simulation methods, together with the advances in computational power have recently enabled the studies of increasingly larger systems, extended simulation times approaching the range of milliseconds, as well as the simulation of much more complex systems such as membranes, entire cells or viruses, and MD studies to complement experimental data are abundant[2]. The importance and rising recognition and acceptance of these methods has been acknowledged by the 2013 award of the Nobel Prize in Chemistry to Martin Karplus, Arieh Warshel, and Michael Levitt, three pioneers in the field of computational biophysics and structural biology. Many researchers saw this as the long-awaited confirmation of the remarkable value of computational methods for understanding biomolecules[3]. Experiments are the standard method used to characterise biomolecular systems, showing their behaviour, reactions and/or function. However experimental results cannot always be fully understood or explain the mechanisms and phenomena observed in the relevant spatial and temporal scales, and this is when the “computational microscope” comes into play[4].

In this thesis we used the combination of experimental methods and molecular dynamics simulation to characterise a family of four-helix bundle biosurfactant proteins. Biosurfactants are surface active agents produced by microorganisms, and a sustainable alternative to commonly chemically synthesized surfactants which come from either petrochemical or oleochemical sources[5]. Petroleum derived products are associated with the emission of greenhouse gases, which are linked to air pollution and global warming. Further negative environmental impact can be seen by the destruction of land and ocean area in the oil exploration process, resulting in the loss of vegetation and natural habitats of animals; the same outcome can be caused by oil spills or leaks in the process of oil production[6]. In addition to environmental issues there is a problem with the limited
resources of fossil fuels, of particular concern given the constantly growing population and demand[7]. Biosurfactants address the increasing demand for more industrial sustainability. As surfactants comprise an estimated market of $33 billion (2014)[8], including applications in two of Australia’s leading industries, food and mining, a green alternative is highly desirable and required in the long term. This large market is due to their versatility with a broad range of predominantly everyday products which require surfactants such as cosmetics, pharmaceuticals, personal care products and washing agents. Biosurfactants have the advantages of being biodegradable, ecofriendly and able to be produced from cheap, renewable feedstocks which can include waste materials. Additionally they are also robust enough to be used in industrial applications as they can generally withstand high temperatures and extreme pHs. This is why it is on the green agenda of many companies to start replacing some or all of the petrochemically derived surfactants in their products with sustainable biosurfactants[5].

However this is not always an easy task, mostly due to the higher costs associated with biosurfactants compared to cheap petroleum-derived surfactants, which makes them currently not economically viable in the mainly low-cost industrial sectors that surfactants are used in. Additional problems which currently impede a more widespread use of biosurfactants in commercial products and applications are related to yield, cost for production and/or purification, and a general lack of fundamental understanding which includes the design and tailoring of the molecules for particular applications[5].

The family of four-helix bundle biosurfactant proteins, and in particular one member therein, DAMP4[9], which was designed to address the problems frequently associated with biosurfactants, offers a unique model system to study the sequence-structure-function relationships of this group of biosurfactants. These molecules are not naturally occurring, but were rather designed purposefully and their study can give meaningful insights into the design strategies for biosurfactants. Experimental results are readily accessible and DAMP4’s behaviour upon expression, purification and functionality have been studied. This PhD aims at elucidating some of the biosurfactant’s underlying features and mechanisms with molecular dynamics simulations, with particular focus on how and why DAMP4’s design addresses the issues with

a. expression yield
b. economical purification
c. fundamental understanding with focus on its adsorption behaviour to an air/water interface
with the use of DAMP4 and other variants of the four-helix bundle biosurfactant library. Each of these points will be discussed further in the next few sections.

\textit{a. Expression yield}

DAMP4 can be expressed at very high levels of 15 mg L\textsuperscript{-1} OD\textsubscript{600}\textsuperscript{-1} in \textit{Escherichia coli} (\textit{E. coli}) in a minimal culture media. This is one of the highest peptide expression levels achieved to date\cite{10}. However a small modification of the sequence, such as a shorter linker in the turn regions, leads to incorrectly folded and insoluble product\cite{9}. Frequently, designed sequences are either undesirably expressed at low yield or as inclusion bodies, or no yield is observed at all if the folding of the target molecule is incorrect and the product is proteolytically degraded by bacterial host proteases. Unfortunately, expression outcome is not yet predictable and can only be determined after time-consuming and expensive experiments. Our library of four-helix bundles contains a whole range of sequences with different experimental expression outcomes, and despite a very high sequence similarity between the different variants, expression can never be guaranteed. As expression of a particular protein depends on its primary amino acid sequence\cite{11}, and can be related to the stability of the correctly folded protein, molecular dynamics simulation offers the opportunity to investigate previously used sequences, with varied expression outcome, on their stability \textit{in silico} in order to draw conclusion on the expression outcome of future designed sequences \textit{in vivo}. This is a novel and, to date, unique approach to a well-known and studied problem, as so far the majority of the accessible prediction tools are based on sequence statistics, rather than the dynamic stability of the 3-dimensional structure\cite{12}. The correct indication of a promising expression outcome of a biosurfactant protein can help overcome initial cost in the development process of newly designed biosurfactant candidates, as well as contribute to a general understanding of underlying reasons for why bacterial protein expression in other cases is doomed to failure.

\textit{b. Economical purification}

DAMP4 can be purified in a simple, fast, and cheap two-step process, in which 0.5 M Na\textsubscript{2}SO\textsubscript{4} is added to the culture media and the entire cell broth heated up to 90 °C for 30 minutes. This causes cell lysis and denaturation of all impurities, while DAMP4 remains soluble and stable in solution. It can be recovered with a solid-liquid separation step, such as filtration or centrifugation, as industrial grade DAMP4 with a yield of 84\%\cite{13}. Such process design is feasible because DAMP4 can withstand extreme temperatures, and maintains its stable four-helix bundle fold even above the boiling point of water. But what makes DAMP4 so stable? Different structural characteristics well known to support stability in proteins, namely salt bridges, the hydrophobic core, and the hydration
shell, might provide the reason, and molecular dynamics simulation used in combination with homology modelling allows the simple sequence modification of DAMP4 in order to eliminate these stabilizing mechanisms. The stability observed in the MD simulation and the mechanism contributing to that stability can be directly inferred as a contributor to the extraordinary thermal stability observed experimentally. This knowledge can then be used in new biosurfactants, which can be designed to be stable enough to be purified with the aforementioned process, but yet maintain their functionality. This would enable the use of biosurfactants even in low-cost industrial sectors where the use of expensive chromatographical steps is not economically viable.

c. *Fundamental understanding of DAMP4’s adsorption behaviour to an air/water interface*

This section addresses the lack of fundamental understanding, with particular focus on the surface active properties and interfacial adsorption. Surfactants consist of a hydrophilic and a hydrophobic part, and because of this structure they are active at interfaces between aqueous and non-aqueous phases. To fully exhibit its surface active behaviour, DAMP4 needs to undergo conformational rearrangements, as all its hydrophobic residues are buried inside the core in aqueous solution while its surface is hydrophilic. Neutron reflectometry experiments confirm this hypothesis of changing tertiary structure, as the results show an interfacial film thickness of DAMP4 adsorbed to an air/water interface of 10-14 Å. This corresponds to the thickness of a single $\alpha$-helix, and further indicates an unfolding of the four-helix bundle despite its high thermal stability in bulk[14]. However experiments cannot give insights into how the transition between the four-helix bundle in bulk to an extended chain of helices occurs, what its final structure at the interface is, or why the changes even arise when DAMP4 is so stable in bulk. In contrast molecular dynamics simulation can show the structures at the interfaces, and for processes that are sufficiently fast it can also examine processes occurring at the air/water interface in atomic detail and help answer the above questions. This will give insights into sequence features related to the functionality of the molecules as well as elucidate fundamental mechanisms and drivers for adsorption processes of proteins to hydrophilic/hydrophobic interfaces by simulating these environments or understand mechanisms of foam control.

1.1 Research Objectives

In summary the aim of this PhD project is to understand the fundamentals of the relations of sequence to heterologous expression outcome, stability and adsorption behaviour of this family of four-helix bundle biosurfactant proteins and to extend the newly gained knowledge to generate
design recommendations for new and improved sequence designs, which ultimately aims to incorporate biosurfactants in more industrial products and processes.

This PhD addresses the following main objectives by applying all-atom molecular dynamics simulation:

(i) To establish a MD-based method for the prediction of experimental expression outcome of designed four-helix bundles solely on the basis of their primary sequences to ensure expression yield for new sequence designs.

(ii) To identify the contributions of the hydrophobic core, the water shell and ion-pair interactions to the stability of DAMP4 and to apply this knowledge in new designs to ensure functionality, stability and ease of manufacture.

(iii) To investigate the interfacial adsorption behaviour of DAMP4 in a simulation to identify the course and mechanism of adsorption to an air/water interface and to confirm an α-helical conformation at the interface. This includes the understanding of the influences of sequence, pH and ions on adsorption and the ability to draw conclusions on the driving mechanism beneath the observations.

This research uniquely links the molecular descriptors (structure, sequence) to both product and process behaviours and aims at understanding their connections, ultimately leading to new and improved soft matter foam materials. Furthermore it adds to the knowledge in the fields of engineering protein sequences, protein stability as well as interfacial adsorption and physicochemical characteristics of interfaces at a fundamental level which is relevant to a range of industrial sectors, including two of Australia’s leading industries, food and mining.

1.2 Thesis structure

This thesis has in addition to the introduction, five chapters including a literature review, three research chapters addressing the objectives above, two of which were published as original research articles (Chapters 3 and 4), and the last chapter containing the discussion, conclusions and future work.

In Chapter 2 a literature review is presented on the key matters considered in this thesis. This includes the current state of knowledge relevant to this research, as well as the necessary background information on DAMP4 and the family of four-helix bundle biosurfactants.
In Chapter 3 a statistical model to predict the expression outcome of biosurfactant four-helix bundle sequences which uses the output of high temperature MD simulations is presented. The first part of this chapter describes how the model was created with different stability-related parameters from MD simulations of 15 four-helix bundles by statistical modelling with a support vector machine algorithm and how the models’ functionality was successfully challenged with two new sequences. The second part contains the experimental validation of the model with the experimental expression of five newly designed (specifically for this purpose) four-helix bundle sequences in *E. coli* and the comparison of experimental outcomes and model predictions.

Chapter 4 explores the stability of the four-helix bundle biosurfactant DAMP4 by the experimental characterisation of its stability, as well as the design of *in silico* DAMP4 variants which do not possess stabilizing features in their sequences. Their stability in thermal unfolding MD simulations is compared and the importance of each feature discussed, revealing the hydrophobic core as the most essential stabilizing characteristic. Design recommendations for future four-helix bundle biosurfactant proteins are given which comply with stability and thus cost-effective purification, adsorption behaviour and functionality.

In Chapter 5 the adsorption behaviour of DAMP4 is elucidated in all-atom MD simulations of a water slab with two air/water interfaces. The importance of the environmental conditions including pH and ions for successful unfolding at the interface is shown as well as the pathway of conformational changes and the final structure of DAMP4 at the interface elucidated. The MD results suggest a combination of factors playing a role in the initiation of the conformation rearrangement at the interface. This chapter gives an insight into the fundamentals of interfacial adsorption of biomolecules, playing an important role for their functionality.

Chapter 6 summarizes the key findings of this thesis and how they address the aforementioned research objectives and problems associated with biosurfactants, and concludes with suggestions for future research.
1.3 References

This chapter will outline the relevant theoretical background and current knowledge, as well as the key literature related to this research.

There are two main parts, the first focuses on surface active agents (section 2.1), for which this research is seeking a more sustainable alternative. The second part describes the main method used in this thesis, which is molecular dynamics simulation (section 2.2), including methodology, applications, and limitations.

### 2.1 Surfactants

#### 2.1.1 Chemical surfactants

Surface active agents (‘surfactants’) are molecules which exhibit activity at interfaces between aqueous and non-aqueous phases, for example in the form of reduction of surface tension, wetting, foaming, detergency, dispersion, or emulsification[1, 2]. Because of this variety of functions they are among the most versatile products of the chemical industry, with applications in pharmaceuticals, cosmetics, personal care and hygiene products, cleaning agents, laundry powder, or food, as well as the use in drilling muds to prospect for petroleum or as flotation agents in ore beneficiation.

![Figure 2-1 Structures of a surfactant and typical micelle configurations.](image)

- **a)** Structure of a classical surfactant with hydrophilic headgroup and hydrophobic tail; Examples of micelle shapes, **b)** spherical micelle, **c)** lamellar configuration. The micelle shape depends on the surfactant concentration[2].
In recent years the areas of application of surfactants have continuously expanded from the classical applications to high-technologies such as electronic printing, magnetic recording, biotechnology or microelectronics[3]. Surfactants are amphiphilic molecules, which means they are organic compounds with at least one lyophilic (‘solvent-loving’) and one lyophobic (‘solvent-fearing’) moiety. In case that water is the solvent these groups are referred to as hydrophilic and hydrophobic, respectively. The hydrophilic part can be neutral or carry a charge, either positive, negative, or both. These surfactants are called non-ionic, cationic, anionic, and amphoteric, respectively. In addition they can have different structures, including one hydrophilic head and two hydrophobic tails, two hydrophilic parts separated by one hydrophobic part, two classical surfactant structures separated by a spacer molecule (gemini surfactant), or comprise a covalent linkage of several hydrophobic and hydrophilic parts (polymeric surfactants) as depicted in Figure 2-2.

Figure 2-2 Schematic structures of different surfactants.
Possible structural arrangements of hydrophilic (purple circle) and hydrophobic (black chain) parts with example surfactants. a) Classical surfactant, b) single headgroup, double hydrophobic tail, c) double headgroup, single tail, d) polymeric, and e) gemini[2].
In contrast to these surfactants peptide or protein surfactants vary from the classical hydrophilic headgroup – hydrophobic tail structure in a rather facial hydrophilicity or hydrophobicity through the facial separation of respective amino acid side chains along the peptide.

**Figure 2-3 Structural comparison of a) classical chemical surfactants and b) peptide or c) protein surfactants.**

This is depicted in Figure 2-3 with the example of helical peptides or proteins, which have the classical heptad sequence of amino acids where every 1\(^{st}\) and 4\(^{th}\) side chain is hydrophobic, resulting in the formation of an \(\alpha\)-helix[4] and the facial separation of hydrophilic and hydrophobic residues and thus surface active characteristics. Middelberg and Dexter[5] used this fact in combination with known interactions and forces in peptides to create new soft matter materials which were responding in their function to an external stimulus. Lately entire proteins can be artificially designed in order to act as surfactants with desired properties due to the flexibility in side chain characteristics[6].

Structure and charges influence the surfactants’ surface properties as well as their field of application. A classical surfactant with a hydrophilic headgroup and a hydrophobic tail can be seen in Figure 2-1 a). The hydrophilic part makes the surfactant soluble in polar solvents, and the hydrophobic part in non-polar solvents and oil. Because of this structure the molecules will aggregate above a minimum bulk concentration (critical micelle concentration) forming so called micelles of different shapes (Figure 2-1b) and c)), and in addition adsorb to hydrophilic/hydrophobic interfaces. Both processes aim at minimizing the exposure of hydrophobic residues to the hydrophilic solvent.

Thus surfactants alter the properties of aqueous media in the bulk, where hydrophobic solutes can be dissolved or emulsified within the hydrophobic inside of the micelles, as well as at the hydrophobic/hydrophilic interface. At the interface surfactants reduce the surface (air/liquid) or interfacial (liquid/liquid, liquid/solid, solid/solid) excess free energy through adsorption.
In the case of two immiscible liquids the interfacial tension \( \gamma_{AB} \) between the liquids A and B is defined as

\[
\gamma_{AB} = \gamma_A + \gamma_B - 2\psi_{AB}
\]

with \( \gamma_A \) and \( \gamma_B \) being the surface tension of liquids A and B, and \( \psi_{AB} \) the interaction energy per unit area between the two.

The measured surface tension corresponds to the free energy per unit area of the boundary between water and air because the uniform interaction network of the molecules is broken at the surface and the molecules mainly interact with the subsurface layer. Alternatively it is quantified as the force normal to the interface per area. Because of this force the liquid tries to minimise its surface area, which is the reason why for example bubbles in liquid are spherical[2]. Surface tension is measured as the minimum work \( W_{\text{min}} \) that needs to be done to increase the surface area and the work is defined as the product of the surface tension \( \gamma \), and the increased area:

\[
W_{\text{min}} = \gamma \Delta \text{interfacial area}
\]

As surfactants adsorb at the surface they significantly change the work required to expand the interface. Surfactants play a major role in systems, where the ratio of surface area to volume is large, which is the case in foams, emulsions or dispersions[3]. In this process, the surfactants replace interfacial molecules and disrupt cohesive interactions between polar and non-polar molecules. The hydrophobic part of the surfactant starts to interact with the non-polar phase and the hydrophilic head with the polar. This lowers the surface tension as the new interactions are stronger than the ones between polar and non-polar molecules before. The higher the surfactant concentration at the interface, the higher the reduction in surface tension[2].

The use of the oldest surfactant, soap, goes back to the times of the ancient Egyptians, and synthetic surfactants have been produced since the first half of the 20\textsuperscript{th} century. Most of the production today relies on approximately even volumes of petrochemical (crude oil) and oleochemical (e.g. coconut oil, tallow, palm oil) sources, whereby fatty alcohols are the most common feedstocks for a whole range of non-ionic and anionic surfactants. Other common raw materials for surfactant production are fats, fatty acids or methyl fatty esters derived from oil. The largest field of application are detergents and cleaning agents (for household as well as industrial applications, toiletries and personal care products), which makes up for around 60% of the surfactant use. The other 40% are applications in the field of agrochemical and industrial usages[2]. The highest surfactant consumer is the Asia-Pacific region with a share of 36% of the overall market, which has grown immensely.
over the last few decades. The estimated global surfactant market in 2014 was 33.2 billion US$ with the potential to increase to 40.4 billion US$ by 2022[7], underlining the need for a more sustainable alternative in the future.

2.1.2 Biosurfactants

Biosurfactants are surfactant molecules produced by microorganisms, including bacteria, fungi, and yeast, from cheap, renewable feedstocks[1]. They have the same structure as chemically synthesized surfactants, with a hydrophobic moiety, which is generally an alkyl chain or hydrophobic amino acids, and a hydrophilic moiety like a lipid head, a sequence of hydrophilic amino acids, sugar ring, short peptide and DNA strand, allowing for a wide range in physical and biological characteristics by variations in these moieties[8].

They have become an attractive alternative for chemical surfactants for many companies in the last few years due to the current trend for industrial sustainability. Biosurfactants are readily biodegradable, which complies with the legislations for surfactant biodegradability[2]. This is not only important for environmental considerations, but also for the customers’ acceptance, who have increasing concerns about biodegradability and toxicity of chemical products due to increasing awareness[3]. In addition to also being non-toxic and ecofriendly, biosurfactants are stable within different environments, including relatively high temperatures or high salt concentrations, which is important for their robustness in possibly harsh industrial process conditions. They are classified as either low molecular weight molecules, such as glycolipids, lipopeptides and flavolipids, or high molecular weight molecules which include polysaccharides, proteins, lipopolysaccharides, and lipoproteins.

![Figure 2-4 Possible arrangements for surfactants hydrophobic (filled circle) and hydrophilic (empty circle) moieties in surfactants.](image)

a) Classical helical heptad sequence, b) block copolymer arrangement.

Glycolipid biosurfactants are particularly promising as replacements for some of the widely used detergents such as linear alkylbenzene sulfonates (LASs). This includes sophorolipids and mannosylerythritol lipids which are produced by yeasts of genera *Candida* and *Pseudozyma,*
respectively, and rhamnolipids produced bacterially by *Pseudomonas aeruginosa*. The problem with these molecules, however, is that they are generally expressed as a mixture of different congeners[1]. In general, properties such as solubility, surface tension reduction, CMC and function as foaming, wetting, or emulsification determine the suitability of a surfactant for a particular application[9]. The problem with the expression of a mixture of congeners is that they vary in structure and thus behave differently in product formulations. Separation through extensive downstream processing is in general not economical for the rather low-cost surfactant products. To improve this, the genetics of the microorganism that produces the surfactant can be modified in order to get a more uniform product, and it drives the research for ‘designer biosurfactants’, such as DAMP4, which are not a mixture of different molecules and can be designed for example for a particular function or behaviour[1]. These designed helical peptides or proteins with their difference in hydrophilic and hydrophobic moiety arrangement (see Figure 2-4a)) break the “block copolymer” paradigm (Figure 2-4b)) which is limited by the CMC and have the key advantage of not having a CMC and thus resulting in better bulk behaviour compared to the classical molecules while being highly active at the interface.

Designed protein surfactants are very diverse as the 20 naturally occurring amino acids, which can be charged, polar or nonpolar, offer a vast array of possible combinations of these components into peptides or proteins, maintaining the amphiphilic character while allowing the fine tuning of physicochemical properties. Many natural or designed peptides and proteins are not obviously amphiphilic, but reveal this property upon structural rearrangements and formation of secondary structures like α-helix or β-sheet in solution and at interfaces. This is a well-known phenomenon for peptides with a reoccurring sequence of different hydrophilic and hydrophobic residues and aids in the formation of a hydrophilic plane by polar and charged groups, and a hydrophobic one by hydrophobic side chains[8, 10].

Major issues with biosurfactants are currently still related to production and cost. Large scale production of the molecules requires the consideration of different factors. Safety and health are one of them in case of a production in bacteria, some of which may be classified as pathogens, or genetically modified yeast. It might require the companies to take special measures, but no real problems have been reported so far. Another point for commercial production of biosurfactants is the yield required to be economical and the development of high producing strains to improve yields is desirable, as well as the ability for the microorganisms to grow from cheap substrates, ideally waste material. This can considerably reduce the production cost. Another cost-related factor is the need for downstream processing, in particular if the molecules are produced as different congeners or products are expressed intracellularly, showing the need for new process technologies
or innovative processes to remove impurities, such as the “bake-to-break and precipitate” process applied for DAMP4[11], or genetic modifications of the producing strain. A non-cost related issue is the lack of fundamental understanding, and how to design the molecules for a particular application[1].

For these reasons, biosurfactants are not well established in industry yet, and only few commercially available biosurfactants exist, among them surfactin, sophorolipids, and rhamnolipids. Two American companies, AGAE Technologies and Jeneil Biotech, are for example producing rhamnolipids on the commercial scale with yields of around 12-20 g/l. Sophorolipids are already produced by several companies including Saraya Co. Ltd. in Japan, Soliance in France, or MG Intobio in Korea. Yields are in the range of 30-100 g/l. These molecules are mainly used in soaps and skin care products[1, 12].

DAMP4 is a designed stimuli-responsive four-helix bundle biosurfactant protein with promising, unique characteristics and it is a member of a new family of designed four-helix bundle protein surfactants, referred to as “Felix Biosurfactants”.

The 98 amino acid long protein DAMP4[6] was designed by Anton Middelberg based on the AM1 peptide (MKQLADS-LHQLARQ-VSRLEHA), a 21-residue stimuli-responsive surfactant derived from the lac repressor protein Lac21 which exists as a monomer/tetramer mixture in solution[13]. The AM1 peptide, which varies from Lac21 in the insertion of two metal-binding histidines, showed the ability to reversibly stabilize foams and emulsions by building metal-ion complexes with other AM1 molecules at hydrophobic/hydrophilic interfaces. This process could be reversed by the disruption of the histidine-metal binding through acidification to change the protonation state of histidines, or by the addition of a chelating agent. It presented one of the first peptides designed for active stimuli-responsive foam control. AM1 was predicted to form α-helices at the interface and tetramers in solution[5]. For economic reasons, as the synthetic manufacture of AM1 was too cost-intensive for industrial use, the peptide was advanced by joining four AM1 to build a stable and soluble four-helix bundle protein, named DAMP4 (MD-[PS-MKQLADS-LHQLARQ-VSRLEHA-D]4). DAMP4 consists of four identical helices, connected with an aspartic acid (D) - proline (P) – serine (S) linker. The linker has the purpose of connecting the four helices and was designed purposefully to promote repulsion between neighbouring loops with the single negatively charged D; P has the lowest helix propensity and will thus break the helicity; and S provides enough flexibility and rotational freedom for a stable fold. DAMP4 has a molecular weight of 11.1 kDa and a theoretical isoelectric point, pI, of 6.7. DAMP4 is a member of a family of designed four-helix bundle biosurfactant proteins, which all follow a basic hydrophilic-hydrophobic pattern motif: MD-[PS-XYYYYYY-YYYYYY-XYYYYY-D]4, with X and Y representing a hydrophobic and
hydrophilic amino acid, respectively. This allows the free design of desired characteristics. However the freedom of sequence design comes along with problems associated with expression yield, as under given experimental conditions the expression outcome in *E. coli* depends solely on the amino acid sequence[14] and despite the existence of different statistical prediction tools based on the primary structure[14-17] a successful heterologous expression is still a trial and error process. This issue will be addressed and discussed in Chapter 3 as a high expression yield in combination with a cheap purification strategy can reduce the production cost significantly. Felix biosurfactants, as demonstrated on DAMP4, can be purified in a cost-effective two step “bake-to-break and precipitate” (BBP) process, which is schematically demonstrated in Figure 2-5. It consists of a thermal treatment in combination with a kosmotropic salt, followed by solid-liquid separation[18]. DAMP4 is expressed intracellularly at a level of 15 mg/(l OD600) in simple culture media in *E. coli*.

![Figure 2-5 Schematic depiction of the “Bake-to-break and precipitate” process used for DAMP4 purification.](image)

In detail, following expression 0.5 M Na₂SO₄ is added to the media. In the next step, the entire culture broth is heated up to 90 ºC for 30 minutes. The high temperature causes cell lysis and release of DAMP4 molecules into the media, as well as denaturation of bacterial contaminants, while DAMP4 remains stable in solution, unaffected by the temperature due to its high thermal stability. Therein the kosmotropic salt seems to salt-in and further stabilize DAMP4, whereas the contaminants are salted-out. Subsequently soluble DAMP4 is separated from insoluble cell debris and denatured impurities with two microfiltration steps with 1.2 μm and 0.2 μm pore size, resulting in ‘industrial-grade’ DAMP4, a molecule pure enough to be used in industrial applications. Recovery of expressed material was around 84% after filtration. Further purification of DAMP4 is possible with concentration of DAMP4 through an ultrafiltration membrane with 10 kDa molecular weight cut-off and the retentate fraction diluted to a Na₂SO₄ concentration of 0.125 M, causing
precipitation of DAMP4 due to the lack of ‘salting-in’. Filtration at 0.2 μm captures the final highly purified product with recovery of 41%[11]. The application of this intensified process relies on the stability of the protein, which retains the majority of its native structure even at temperatures > 90 °C. This thermal stability is above the average of proteins and is in the region of that of proteins from thermophilic microorganisms[19]. DAMP4 offers this process advantage due to its melting point of around 115 °C as determined with differential scanning calorimetry (DSC), but it is yet to be understood why DAMP4 has this extremely high melting point. This knowledge gap will be addressed in Chapter 4, which contains the experimental characterisation of DAMP4 stability as a premise for the heat process, and investigation of the underlying mechanisms with MD simulation[20].

Figure 2-6 Foaming behaviour of DAMP4 under different pH conditions.
Foaming of 0.3 mg mL⁻¹ DAMP4 in 25 mM HEPES buffer, 200 μM EDTA, 10 mM NaCl. a) Stable foam at pH 8.5 dissipates by the addition of 14 μl of 1 M HCl lowering the pH to 7.5, b) control experiment, foam propensity unchanged by addition of 14 μl 1 M NaCl, c) dilution from pH 8.5 to pH 7.1 with milli-Q water causes foam collapse[6].
Experimental foaming tests revealed that DAMP4 is surface active at oil-water interfaces forming stable emulsions[21], or at air/water interfaces, where it exhibits excellent foaming behaviour at pH 8.5 (Figure 2-6). At this pH and in the presence of ethylenediaminetetraacetic acid (EDTA) substantial and stable foam is formed and the interfacial tension decreased to 49 mN m\(^{-1}\). This foam can be dissipated by a pH decrease of as little as one unit to pH 7.5, either by the addition of a small amount of HCl (Figure 2-1a)), or by dilution with milli-Q water (Figure 2-1b)). This gives the biosurfactant promise as a new stimuli-responsive foam control application. DAMP4’s pH responsiveness derives from its amino acid composition, and the foam stabilization mechanism is based on electrostatic forces. It was designed with an overall neutral net charge at neutral pH, with even numbers of positive and negative charges on each helix. Without net charge, an electrostatically stabilized foam cannot be formed despite an approximately similar surface excess of molecules under both pH conditions (pH 7.5 and pH 8.5).

![Figure 2-7 Concentration-dependent interfacial tension reduction of DAMP1 and DAMP4.](image)

Interfacial tension reduction for different concentrations of a) DAMP1, and b) DAMP4 and corresponding schematic free energy diagram for the adsorption process to an air/water interface. Figures adapted from [22].

It has been hypothesized[6] that at pH 8.5, further away from DAMP4’s isoelectric point of 6.7, the aminium side chain of lysine possibly deprotonates at the interface, which can have an influence on
pK<sub>2</sub> values in its proximity, resulting in an overall net charge of -1 on each helix and the charge stabilizes the thin film electrostatically. This behaviour is reversible and pH changes can actively switch the film-state on and off[6]. DAMP4 as well as its monomer DAMP1 (PSMKQLADS-LHQLARQ-VSRLEHAD) adsorb to an air/water interface and lower the interfacial tension concentration-dependent (Figure 2-7). DAMP1 is unstructured in bulk but expected to adopt an α-helical conformation upon adsorption. This is because of the strategic placement of hydrophilic and hydrophobic residues in the peptide’s design, which drives interfacial helical formation. Its adsorption kinetics are fast and correspond to that of a diffusion-controlled process as seen in Figure 2-7a). DAMP4 has a very stable four-helix bundle fold in bulk, and the adsorption rates are rather slow compared to DAMP1 (minutes vs. seconds for DAMP1) when studied at very low concentrations. This is even slower than expected for a simply diffusion controlled process and indicates an additional barrier for adsorption which does not exist for DAMP1. In addition, neutron reflectometry results show that the thickness of a DAMP4 film adsorbed to the air/water interface is the same for AM1 and DAMP1: 10-14 Å thick. This corresponds to the thickness of a single helix. These experimental results indicate DAMP4 is undergoing conformational changes in its tertiary structure at or before adsorption to the interface[22].

![Figure 2-8 Suggested possible pathways for DAMP4's adsorption to an air/water interface.](image)

a) Diffusion to, followed by unfolding at the interface; b) partial unfolding upon approximation to the interface to fully unfolded chain of helices when adsorbed.

DAMP4 likely unfolds from the stable four-helix bundle in bulk to a chain of four-single helices at the interface, revealing its surface active characteristics and inducing the high energy barrier as pictured in Figure 2-7b). However it is not clear what triggers these changes at the interface, or what transition and interfacial arrangement look like. The interfacial behaviour of molecules however plays an important role for product formulations, but the link between interfacial properties, ideally directly derived from the molecular structure, and functional behaviour is not
well established yet[22]. This issue will be addressed in Chapter 5, where DAMP4’s molecular structure is linked to its adsorption behaviour and the results will suggest a mechanism for the processes proceeding at the air/water interface.

It has been shown that the addition of a small peptide like DAMP1[22] or a small amount of chemical surfactant as for example sodium dodecyl sulphate (SDS)[23] can enhance the biosurfactant’s performance and reduce the amount of protein needed for foam stabilisation without the loss of switchability. This is because of the synergistic interactions between DAMP4 and these molecules, which significantly speeds up the interfacial kinetics. The cost reduction in the formulation through chemical surfactants or peptides, in combination with the cost-efficient purification strategy, which eliminates the need for expensive chromatographical purification, are promising steps towards the broad application of designed biosurfactants like DAMP4 in industry[23].

2.1.3 Recombinant protein expression in Escherichia coli

*Escherichia coli* is one of the most successful, and thus popular host systems for heterologous protein expression[24, 25]. This is for several reasons. For example its genetics as well as transcription, translation, and protein folding processes in *E. coli* are well understood. In addition it requires only inexpensive substrates, and grows very rapidly with a doubling time of about 20 minutes[26] and to high cell density[25].

In the ideal case the gene, which encodes the amino acid sequence of the desired protein, is cloned into an expression vector, which is then transformed into the host strain of choice. This can be done with electroporation or heat shock methods. Overproduction of the target protein is induced, for example with isopropyl β-D-1-thiogalactopyranoside (IPTG) which triggers the transcription of the lac operon, and the molecules can be recovered and purified[27]. In practice however there are many problems which can be associated with recombinant protein expression. Frequent issues include expression as aggregates called inclusion bodies, poor growth of the bacteria, no biological activity, or expression at low to zero concentration[25, 27]. Continuous research has led to different approaches to overcome these challenges. One method to increase solubility, detect the molecule upon expression and purification, and easily purify it from bacterial contaminants is to add a fusion tag (peptide or polypeptide) to the amino acid sequence. Frequently used peptide tags are poly-Arg-, FLAG-, poly-His-, or S-tags, for which antibodies for easy recognition and affinity chromatography detection are possible. Popular fusion proteins which in addition increase solubility are maltose-binding protein, N-utilization substance Protein A, or ubiquitin. Tags can be removed with enzymatic or chemical cleavage. In case of toxicity to the host the secretion of the protein into the media by using particular signal peptides is possible[27]. Overproduction of proteins into the *E. coli*
cytoplasm often causes misfolding and formation of inclusion bodies. Inclusion bodies can facilitate the purification process; however there is a need to denature and refold the proteins in vitro, which is not always successful and can result in low yields of biologically active target molecules. To prevent this it is common practice to decrease the cultivation temperature. Alternatively chaperones, which help in the folding process, or foldases, which accelerate rate-limiting folding steps, can be co-expressed and potentially prevent the formation of inclusion bodies[25]. Despite these efforts it is still recommended to try multiple conditions and combinations of constructs, expression vectors, and expression strains, as the experimental process is still trial-and-error and time-consuming[27].

A different approach to the same problem, namely soluble expression of a target protein sequence or not, is to apply statistical methods on the amino acid sequence a priori to any experimental trial. Numerous different theoretical prediction tools based on recombinant protein libraries exist[15-17, 28, 29] (for a detailed review see Chapter 3), which can help to find more promising expression candidates. However they are still limited and not always suited for all classes of proteins. For example for our set of four-helix bundle sequences the prediction success was low as demonstrated in section 3.1. To address this problem and improve the prediction power, and ultimately experimental expression success, for the family of biosurfactant proteins used in this thesis, we created a new prediction model. In addition to simply analysing the primary sequence of the proteins, their 3-dimensional structure and behaviour is determined using a molecular dynamics simulation. Combining these methodologies is to this date a unique new approach. The model is tailored in particular to four-helix bundles and its parameter and results can be found in Chapter 3.

2.1.4 Protein adsorption to interfaces: Foam formation and stability

Protein foams are used in a number of different applications, most commonly they are found in the food industry e.g. in beer, whipped creams or mousse s, or personal hygiene products such as soap. Frequently an understanding of the mechanisms underlying foam formation is not only important to create a stable foam, but also to efficiently control, and if required destroy, foams as they can be unwanted byproducts for example in fermentations[30]. Foams are networks of gas bubbles within thin films of a continuous water phase (Figure 2-9), also called lamella. It is the properties of the air/water interfaces which are the most important factor to determine if a foam is stable or not. Three main mechanisms are responsible for foam instability with a liquid thin film, namely drainage, coalescence, and disproportionation. In unstable foam, drainage of the liquid in the thin film occurs, which causes the gas bubbles to approach each other and ultimately coalesce, leading to a loss of structure, stability and texture of the foam. Another mechanism which causes a foam to break is by diffusion of gas between different particles due to different Laplace pressures, and thus gas concentrations.
Figure 2-9 Schematic representation of the structure of a foam.
Foam consists of gas bubbles in a liquid phase. Figure adapted from [30].

The extent of this effect, called disproportionation, depends on the elasticity of the thin film, but leads ultimately to a coarsening of the bubbles, more rapid drainage, and foam collapse[31]. Proteins stabilize foams by adsorption to the interfaces due to their amphiphilic structure with hydrophilic and hydrophobic residue side chains. Frequently, this involves a conformational change of the protein.

Figure 2-10 Stages of the adsorption process of proteins to an air/water interface.
The adsorption of proteins to an hydrophobic/hydrophilic interface includes a) the transport of the molecule to the interface by diffusion, b) the penetration or adsorption process, and c) the structural rearrangement at the interface[30].

To be an effective stabilizer the protein needs to adsorb quickly, as foam formation is only a short temporary state if no stabilization occurs and the gas bubbles collapse rapidly. Protein adsorption involves several phases as depicted in Figure 2-10. In the first phase the molecule moves to the interface by diffusion, convection, or both. In a simply diffusion-limited
system, small molecular weight and thus high diffusion coefficients, are important for foamability. Depending on the surface concentration, the surface pressure changes due to the lowering in surface tension by the adsorption of the protein. However a diffusion limited process is only valid at the initial stages of surface adsorption, and later the influence of already adsorbed molecules becomes stronger, imposing a new barrier for adsorption. The more hydrophobic the molecules are, the more rapid the increase in $\pi$, and the higher the barrier for desorption from the interface[30]. At the interface the adsorbed proteins interact with each other through electrostatic or hydrophobic forces, building hydrogen or covalent bonds, and like this build a stabilizing viscoelastic network which helps to prevent the coalescence of the bubbles[30]. However this network is not tight enough to avoid, for example, gas diffusion completely[31]. As will be discussed in Chapter 5, for some molecules including DAMP4, conformational rearrangements post adsorption, also called surface denaturation, can occur, promoting the protein-protein interactions. This is another factor influencing the interfacial properties of the molecules, and it can vary from no structural changes, to radical rearrangements. This is a unique property limited to protein surfactants, and thus needs to be considered in surfactant designs due to their influence on adsorption kinetics. Foam stability is influenced by the protonation state of the proteins, which varies depending on the pH. At pHs far away from the pI the molecules have a higher net charge, which is not favourable for the stability of foams, whereas at a pH close to the pI a foam is more stable[32]. This is in contrast to what is observed for DAMP4, where good foaming is observed at pH 8.5 but no stable foam at pH 7.5, which is closer to the isoelectric point of 6.7, showing the need for a more fundamental understanding of this molecule[6]. At the pI the number of molecules adsorbed to the interface is highest, as is the surface viscosity. It has been shown that an increased bulk viscosity of the protein solution helps to prevent drainage, as the more liquid in the thin film the less likely the gas bubbles are to coalesce. Coalescence additionally depends on the charge state of the interface, as similarly charged interfaces cause electrostatic repulsion. Covalent bonds can facilitate the foam stability further, as well as strong hydrophobic interactions between the molecules. The exposure of hydrophobic residues to the air phase is one of the main driving forces for adsorption and the conformational changes of the molecules at the interface. This results in the formation of a continuous, cohesive film. This viscoelastic film is able to withstand thin film stretches by the deformation of proteins within limits, and dampen fluctuations in the thickness of the film, which is referred to as the viscoelastic mechanism of stabilization and is unique to protein surfactants[30].

Proteins and peptides vary in their structure from that of classical surfactants, which generally consist of a hydrophilic headgroup and a hydrophobic tail, in their rather facial amphiphilicity. Because of the large number of possible side chain characteristics – and combinations, biomolecules offer the advantage to specifically engineer particular interactions at the interface.
Depending on these interactions the intermolecular network between adsorbed molecules is designed to be either strong, or weak. A special feature is to develop these characteristics within a given environmental condition which can be changed. This means, the particular function, in this case the foaming, is only evoked by a particular trigger, such as the pH, and these molecules are called stimuli-responsive[33].

In summary, the mechanism which defines the ability of proteins to stabilize foams is a complex mixture of forces, governed by the intrinsic properties of the protein, including size, surface hydrophobicity, and flexibility, as well as environmental factors such as pH, temperature, or viscosity. However, a well-understanding of these underlying mechanisms enables the design of new and improved soft materials with special features, such as switchability, depending on the environmental conditions.

### 2.2 Molecular dynamics simulation

#### 2.2.1 History and overview

The application of computer simulation to biological problems goes back to the 1950s, when Alder and Wainwright pioneered in the method using hard two-dimensional disks. By then, the simulation was restricted to a few hundred molecules, limited by computer memory capacity and speed[34, 35]. In 1960, Gibson et al.[36] applied molecular dynamics (MD) to investigate the damaging of a metal copper crystal lattice system of 500 atoms by high-energy radiation, which was by then a complex system. Aneesur Rahman can be considered as the father of the field, he made many contributions to methodology and promoted the use of MD simulation in the scientific community[37]. In 1964 he simulated liquid argon interacting with a Lennard-Jones potential by solvation of the classical equation of motion and characterised a number of its properties[38], and in the 1970s liquid water[39] and superionic conductor α-Agl[37, 40]. The first small protein simulated was bovine pancreatic trypsin inhibitor (BPTI) in 1977[41], and in the 1980s simulations of biomolecules of increasing size were getting feasible and more and more popular[37].

MD simulations can be used to address a number of different biological problems including conformation and structural rearrangements of proteins, membrane transport, protein folding, and ligand binding.

MD is useful to investigate conformational changes as biomolecules are not rigid but flexible, and constantly switch between different conformational states, often related to their function and regulation. MD can help to reveal the pathway between the different states, to identify new conformations, to determine equilibrium states or distributions between the different conformations,
and elucidate the influence of mutations or ligands on these[42]. Shaw et al.[43] used extremely long all-atom MD simulations to characterise the folding of a WW domain (FiP35) and the folded-state transition of BVTI. They found that FiP35 showed several folding-unfolding events when they simulated it for 100 μs at 395 K. Thereby the folding followed the same pathway of disordered and unfolded to native structure via formation of the tip of the first hairpin, followed by the entire first hairpin, then formation of the second hairpin and the consolidation of the hydrophobic core. Unfolding proceeded along the reversed pathway. By simulation of the individual hairpins and analysis of transition states they explained this behaviour with the different intrinsic thermodynamic stabilities. To investigate the native state dynamics of BVTI, 1 ms simulation at 300 K was performed to reproduce and analyse the kinetics and found that it transitioned reversibly between a number of structurally distinct long-lived states[43].

In membrane transport, MD simulation is particularly useful due to its ability to capture the movement and transport of single atoms at very fine spatial and temporal steps. Transport efficiency and substrate selectivity often depend critically on these detailed spatial configurations and the subtle movement through the membrane. Questions that have arisen regarding substrate permeation rate, substrate selectivity or transport regulation with different stimuli could be answered using MD[42]. Gurtovenko and Vattulainen[44] explored the poorly understood ion leakage mechanism through lipid membranes using MD and found that in response to an electric field Na+-ion transport occurs through water pore formation in the membrane. The selectivity of potassium channels of K+-ions over Na+-ions was investigated by Noskov and co-workers[45]. With their simulations they found that it was not the precise fit of the K+-ions into the pore that is responsible, but the intrinsic physical properties, like electrostatic interactions, result in coordination of the ions at the binding sites and causes ionic selectivity.

Another field of application is protein folding, where MD can help to identify the native structure of a protein, or elucidate folding or unfolding pathways, directly showing the atomic details of these events. This detailed view further allows the determination of heterogeneity of folding pathways and their rate-limiting steps, nature of misfolded states and other complex features of the folding process[42]. Snow et al. investigated the folding of the small, 20-residue Trp cage protein and concluded that folding using MD does not only generate folding trajectories, but as well quantitative prediction power for folding rate and structure[46]. With simulations of larger molecules and longer timescale becoming achievable, Duan et al. showed in a 200 ns simulation the beginning of the folding process of the villin headpiece subdomain, a 36-residue large peptide[47].

Ligand binding is another area of active research, in particular in medical research to identify possible drug target binding sites. These ligands usually change the functionality of the protein by either blocking the active site directly or by leading to an altered conformation. Advances in the
accessible time scales allow the modelling of these binding events, without any prior knowledge of the binding sites. Data such as binding site, binding conformation, binding affinity or pathway can be derived from these simulations[42]. Buch et al. investigated the binding of serine protease β-trypsin inhibitor benzamidine to trypsin in 187 MD simulations which represented the diffusion and specific binding event. Quantitative reconstruction with Markov models lead to the identification of binding via two metastable intermediate states of the complex. They concluded that one of the transition states is the rate-limiting step and suggested, an alteration of the residues involved in the transition state might alter the binding kinetics of benzamidine[48]. Another example for the importance of the inclusion of atomic motions in the binding process of ligands is the mollusc acetylcholine binding protein (AChBP)[49]. It was suggested that the flexible unbound protein might be selectively stabilized by the binding agonists and antagonists, all of which conformations might be druggable and pharmacologically relevant[49, 50].

This was only a short overview of the possibilities given by MD simulations in the biological field, but there is a large array of applications for MD in other areas, which are beyond the scope of this thesis. But to conclude, due to the increasing success of simulations, peaking in the award of the Nobel Prize in Chemistry to Martin Karplus, Arieh Warshel, and Michael Levitt, this methodology is becoming more and more attractive to theorists as well as experimentalists, and its usage and acceptance among researchers is growing continuously.

2.2.2 Introduction to MD methodology

Molecular dynamics (MD) simulation has emerged as an important and powerful tool to investigate biomolecular systems in the last few decades. This is due to advances in the accuracy of force fields (potential energy functions), as the underlying model systems, and simulation speed associated with enhanced hardware and algorithms leading to simulation times closer to timescales in the range of real-time biological events. Biological molecules are not rigid, but they are highly dynamic, and MD is a technique which can capture and describe these dynamics, whereas other experimental methods like nuclear magnetic resonance (NMR) or X-ray diffraction only capture static structures[42]. Further, molecular dynamics simulation is able to describe dynamic events occurring on the atomic level of proteins, which is so far not accessible to experimental techniques as illustrated in Figure 2-11. A molecular dynamics simulation solves Newton’s law of motion, which is the force that is acting on each atom of the system, at each time step considering the influence of its neighbouring atoms[51]. It is the calculation of the molecule’s behaviour over time based on physical theory[52].
Figure 2-11 Time and length scale of various biophysical techniques.

The different colours represent the resolution of the various techniques; bold indicates the ability to obtain single-molecule data. NMR can provide information on broad timescales, but is limited in its ability to motion on certain intermediate timescales (indicated by shading and dashed lines). AFM=Atomic force microscopy; EM=Electron microscopy; FRET=Förster resonance energy transfer; NMR=Nuclear magnetic resonance [42].

The potential energy function $E$, which is used to calculate the force acting on each atom, needs to be selected to be realistic and computationally efficient. A commonly used form is:

$$E(R) = \frac{1}{2} \sum_{\text{bonds}} K_b (b - b_0)^2 + \frac{1}{2} \sum_{\text{bond angles}} K_\theta (\theta - \theta_0)^2 + \frac{1}{2} \sum_{\text{torsional}} K_\rho ([1 + \cos(n\rho - \delta)]) + \sum_{\text{nb pairs}} \left( \frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr} \right)$$

(3)
In equation (3), the first three sums describe interactions between atoms which are directly bonded, and the last sum describes non-bonded intramolecular and intermolecular interactions, including electrostatic and van der Waals interactions as a function of their position $R=(r_1, r_2, r_3, ..., r_N)$. The first sum describes the deviation from the equilibrium bond length $b_0$ with bond force constant $K_b$ and instantaneous bond length $b$ by a Hooke’s law harmonic potential. The second sum includes energy due to perturbation of bond angles with equilibrium angle $\theta_0$, force constant $K_\theta$ and angle $\theta$. The third sum accounts for rotations around bonds and is a periodic potential with a cosine function for torsion angle $\rho$, multiplicity $n$ and phase difference $\delta$. The last term summed over all atoms includes energies due to non-bonded van der Waals interactions, described via a Lennard Jones potential, and electrostatic interactions, described by a Coulomb potential, between two atoms with interatomic distance $r$ and charges $q_1$ and $q_2$ and dielectric constant $D$. Parameters $A$ and $B$ are specific for the different types of atoms[37, 53]. Together with the functional form, the parameters for this equation define the force field. The parameters are determined empirically based on experiments, first-principles physics and parameter fitting to quantum mechanical computations. Their usefulness depends on the underlying model systems and a particular force field needs to be chosen depending on the application[42, 53]. In this thesis the force field Amber03 was used, a point-charge all-atom force field specifically designed for the simulation of proteins[54].

2.2.3 Structure modelling

The starting structure of a simulation can be obtained experimentally by X-ray crystallography and NMR structure determination, or by theoretical model building. As experimental structure determination with X-ray crystallography or NMR spectroscopy is still a rather difficult, time-consuming and costly process[55, 56], in this thesis we use theoretical modeling techniques to create and validate a robust structure of DAMP4. Different methods for theoretical structure determination exist. Among them comparative or homology modelling is a method which has been shown most promising and useful in predicting 3D structures, getting closer to a level comparable to experimental methods[57-59]. Thereby the accuracy of homology derived models varies from low resolution models, with just an overall correct fold, to more accurate models which are comparable to medium resolution structures determined by crystallography or NMR spectroscopy[58]. Homology modelling is based on the two principles that the tertiary structure of a protein is uniquely determined by its amino acid sequence, and that similar sequences very likely adopt similar folds[60].

The modelling process is divided into the following seven steps[60]

1. Template identification and first alignment
In the first step the target sequence is aligned with all sequences in the protein data bank (PDB) to find possible matches with structurally related proteins. Sequence identity can be checked with programs like BLAST[61] and FASTA[62]. The steadily growing number of structures in the PDB increases the successful identification of appropriate template structures by using residue exchange matrices, where residues are aligned depending on identity or property similarity. After the identification of the best matches, a multiple sequence alignment helps in the second step to correct the alignment.

Figure 2-12 Schematic illustration of theoretical model prediction methods a) homology modelling and b) fold recognition.

Homology modelling is based on the assumption that similar sequences adopt similar structures and thus searches the protein data bank for sequences matches and uses them as templates for the 3D model of the protein of interest. Fold recognition is based on the assumption that there is a limited amount of folds adopted by proteins and thus tries all the different folds for the target sequences and judges their quality based on energetic functions.
In the backbone generation, only model backbone atoms are placed according to the template if the residues differ, or backbone atoms including side chains if they are the same. Multiple templates can be used for different regions of the protein if certain parts of the templates align better with particular areas. Structure error prediction tools (for example the PDBREPORT database[63]) can be used to identify errors on experimentally determined pdb structures to avoid these parts in the model. Loop conformations (step 4) are more difficult to predict. Loops can be modelled with two different approaches, knowledge based by copying the loops from pdb templates, or energy based. Side chains are modelled with position-specific rotamer libraries derived from high resolution X-ray data, facilitated by the backbone conformation, which favours certain rotamer arrangements. Their quality is judged by a variety of energy functions. Optimization of the model is performed by an iteration of rotamer predictions and energy minimization steps, or by running a short molecular dynamics simulation of the model. The last step is a model validation. This can either be done by calculating force-field based energies, which checks the local geometry such as bond lengths and angles, but does not include entropic effects. The other possibility is to directly use normality indices, which includes the comparison of for example bond lengths, angles, torsions, the burying of hydrophobic residues or radial distribution functions to real structures[60].

Another approach for theoretical modelling involves de novo methods, in which the three dimensional structure is determined solely based on the primary sequence. These methods are frequently applied when no solved homologous structure is available. The two most important factors for a successful de novo structure prediction are a) an efficient search through all possible conformations, and b) the free energy function used to assess the quality of the conformations for the given sequence[57]. Fold recognition is a widely used de novo method and based on the assumption that there is a limited amount of possible protein folds. With the PDB now containing almost all possible structural folds, fold recognition has become a key factor in protein structure determination. Fold recognition methods can be divided into only sequence based- and sequence and structure based. Latter assess the compatibility of the target sequence with each known structure with a score function, which takes structural characteristics such as accessible surface area, contact area of solvent with polar atoms, atom distances or backbone torsion angles into account[64]. Among different fold recognition web servers are SPARKS-X[65] (http://sparks-lab.org/yueyang/server/SPARKS-X/), ROBETTA[66] (http://www.robeta.org/submit.jsp) or Pmodeller[67] (http://pcons.net/index.php).

All obtained models that are obtained are incorrect, however they can still be useful[60]. Therefore it is essential that every theoretical model building process is followed by a critical valuation of the
model quality. Different tools[63, 68-70] exist to check the quality of the three dimensional structure, in particular when there is low sequence similarity to the template structure or de novo methods were used.

One popular method for the quality evaluation, which was also applied for the DAMP4 structure (section 4.8), is a Ramachandran plot. This method is based on the observation that proteins show preferences for conformations with particular combinations of values for angles based on energetic considerations. The torsion angles $\rho$ and $\psi$ of residue $i$ define these conformations, whereby $\rho$ corresponds to the torsion $C_{i-1}-N_i-C\alpha_i-C_i$, and $\psi$ to $N_i-C\alpha_i-C_i-N_{i+1}$[71]. The distribution of $\rho$ and $\psi$ values is called the Ramachandran diagram and plots $\rho$ versus $\psi$ for each residue. In the plot the areas are divided into most favoured, allowed, generously allowed, and disallowed region[68]. A deviation of the model angles from the preferred areas indicates problems in the modelled structure.

It can be considered as one of the most simple and sensitive methods to reveal major errors in protein models because $\rho$ and $\psi$ angles are usually not constrained during refinement[72].

Despite the increasing number of template structures for homology modelling in the protein data bank and increasing accuracy of software and force fields there are a number of limitations and errors in homology modelling which one needs to be aware of when applying theoretical structure modelling. In general these errors can be divided into five categories[58]:

1. Errors in side chains
   Side chain conformations are not necessarily conserved and might lead to crucial errors if the side chain plays an important role for the molecule’s function, such as in protein binding.

2. Distortions and shifts in correctly aligned regions
   Due to sequence divergence a main chain might change even though the overall fold remains the same. This implies that even though the sequence alignment was correct, there are local errors in this region. Sometimes this is also related to the method and environment in which structure was determined. These errors can be minimized by using a number of different template structures.

3. Errors in regions in which no template is available
   These regions are particularly problematic if they are more than nine residues long and the magnitude of the error depends on the quality of the surrounding regions where a template is available.

4. Errors due to misalignment
   Misalignments happen frequently when the sequence identity with the template is less than 30%. These errors can be decreased by using multiple sequence alignments and by iteratively modifying alignments with identified errors in the model.

5. Incorrect templates
   Errors in model structures will increase with the errors in the underlying template structures.
Further improvements in recognition of problematic structure-sequence similarities and alignments, modelling of loops and side chains, as well as the resolution and quality of template structures are needed in combination with good methods to detect these errors[58]. However the wide success and use of these models is an indicator that it is still possible to create models at a sufficient accuracy to create useful results for a wide range of applications.

In this thesis the molecular graphics, modelling and simulation package YASARA structure was used. The force fields of this software were particularly optimized for homology model refinement[73] and performed extraordinary well in the CASP (Critical Assessment of Structure Prediction)[74].

As there are no homologous sequences for the entire DAMP4 sequence in the protein data bank, a combination of both methods, homology modelling (based on the homology modelling experiment in YASARA structure) and fold recognition (SPARKS-X) was used, as techniques from the CASP pointed out the importance of refinement of structures derived from fold recognition methods[65]. High model quality was verified with YASARA’s internal structure validation tool[63], see section 4.4.2, and the external ProSA (protein structure analysis) program[69, 70]. In addition, a Ramachandran plot[71] to confirm the geometrical accuracy around the Cα-atoms was created.

2.2.4 Method

Once the potential energy function is chosen and a high quality starting conformation determined, either experimentally or theoretically, a dynamic simulation can be begun. Starting velocities are derived from a Maxwellian distribution according to the temperature. Newton’s law of motion

\[ F_i = m_i a_i = \frac{-\partial E(r_1, r_2, ..., r_N)}{\partial r_i} \]  

is used to compute the acceleration \( a_i \) of atom \( i \) with atomic mass \( m_i \) and the force \( F_i \) which is calculated from the differentiation of equation (3) with respect to the current atom position and its surrounding atoms[53]. Newton’s equation of motion can be integrated with different algorithms, however the simplest, and usually best, is the so-called Verlet algorithm[75, 76]. The new atom position \( r_i \) at new time \( t + \Delta t \) is calculated with a Taylor expansion of the coordinate \( r \) forward and backward in time:

\[ r_i(t + \Delta t) = r_i(t) + v_i \Delta t + \frac{1}{2} a_i(t) \Delta t^2 + \frac{\Delta t^3}{3!} \dddot{r}_i + O(\Delta t^4) \]  

and

\[ r_i(t - \Delta t) = r_i(t) - v_i \Delta t + \frac{1}{2} a_i(t) \Delta t^2 - \frac{\Delta t^3}{3!} \dddot{r}_i + O(\Delta t^4). \]

Summed up this results in
\[ r_i(t + \Delta t) + r_i(t - \Delta t) = 2r_i(t) + a_i(t)(\Delta t)^2 + O(\Delta t^4) \]  

(7)

or

\[ r_i(t + \Delta t) \approx 2r_i(t) - r_i(t - \Delta t) + a_i(t)\Delta t^2. \]  

(8)

Since Newton’s law of motion is integrated, the acceleration corresponds to

\[ a_i(t) = -\frac{1}{m} F_i(r_1(t), r_2(t), ..., r_N(t)) \]  

(9)

whereby the force \( F_i \) is a function of the position \( r_i(t) \) and derived from the potential energy function[75].

Continued integration of these equations for all atoms of the system (including the protein and the surrounding solvent), at each time step enables the calculation of the behaviour of the system in the course of time[53].

In our simulations the isobaric-isothermal ensemble was used. It is also called the constant-NPT ensemble, as the number of atoms \( N \), pressure \( p \) as well as the temperature \( T \) are kept constant while the total energy \( E \) and volume \( V \) fluctuate around the thermal equilibrium. It is widely used as most real experiments are carried out under these conditions[75]. The temperature was kept constant with a Berendsen thermostat through weak coupling of the system to an external bath of the target temperature[77]. Thereby the atom velocities are rescaled to adjust the force and thus selected temperature \( T_0 \).

\[ m_i \dot{v}_i = F_i + m_i \gamma \left( \frac{T_0}{T} - 1 \right) v_i \]  

(10)

with the damping constant \( \gamma \), which determines the strength of the coupling, being defined as

\[ \gamma = 1 + \frac{\Delta t}{2\tau_t} \left( \frac{T_0}{T} - 1 \right). \]  

(11)

Equation (10) shows the modified equation of motion, in which the atom velocities \( v_i \) are scaled proportionally to \( \gamma v_i \) with time constant \( \tau_t \)[77].

YASARA does not use the classical Berendsen thermostat which uses the strongly fluctuating instantaneous temperature, but rather the time average temperature in order to calculate the scaling factor under minimal influence on the system[78].

To simulate the system realistically it is important that the system of interest is surrounded by an infinite number of bulk molecules, which is usually achieved through the use of periodic boundary conditions. This means, the simulation cell containing the N particles is considered as one of an infinite periodic lattice of equal cells (see Figure 2-13). In this case any particle in the simulation box interacts with all the other particles, independent if they are in the same simulation cell or any other. In practice however it is common to apply a cutoff distance for intermolecular interactions in order to save computational cost[75].
Figure 2-13 Schematic representation of periodic boundary conditions. Figure adapted from [75].

2.2.5 High temperature MD simulation to assess unfolding and stability

Protein unfolding and folding events have been studied for several decades[79], and it is known that the thermal stability of a protein can be used as a measure for its general stability in bulk[80]. Protein folding plays a role for many biological processes, including human diseases, protein degradation and translocation, as well as structure determination and prediction[81]. A challenge which is still associated with the assessment of folding or unfolding events with molecular dynamics simulation is the time scale required for the events to occur, which is frequently not accessible in the simulation due to the computational requirements.

Computational demand can be lowered for example by the use of an implicit solvent or a ‘united atom’ approach, however all-atom MD simulations are the most realistic, and explicit treatment is particularly important to reveal unfolding mechanism and pathway[82]. Alternatively the simulation can be performed for a rather long time, for example Kundu and Roy simulated the antifreeze protein AFPIII for 10 ns to reveal its unfolding pathway at different temperatures from 277 K up to 473 K[83]. Frequently, in addition to the large differences in experimental vs. computationally available time scale for unfolding that exist, there are high energy barriers associated with major structural rearrangements such as a thermal denaturation. This is why these simulations are often run under elevated temperature conditions, and it is common to use temperatures as high as 498 K, or 225 °C, in order to induce the unfolding of proteins in MD. This temperature may seem rather drastic, but Daggett and co-workers have investigated the influence of the increased temperature on the unfolding pathway of chymotrypsin inhibitor 2 (CI2) with the hypothesis that the increased
temperature only affects the rate of unfolding by helping to overcome the activation barrier, but not the energy landscape. In a total of 0.344 μs of simulation time at seven different temperatures, below, at, and above the experimental melting point of Cl2, they confirmed the expected outcomes, which is the protein being stable and remain folded, unfolding and refolding, and unfolding, respectively. The unfolding pathway consistently followed a particular order, and the same transition states were observed. In the end they concluded the high temperatures change the time scale of events and helps to activate the unfolding process, and is therefore a suitable and useful method to investigate unfolding. Thereby the energy landscape and consequently the unfolding pathway are not altered, only times and order of the breaking of particular intramolecular contacts varies[79]. Numerous other studies have applied this method to speed up unfolding processes successfully [83-88]. In this thesis we applied high temperature MD simulations at recommended 498 K[79] to compare the stability of DAMP4 to the stability of different DAMP4 variants in Chapter 3 and Chapter 4.

2.2.6 MD simulation at interfaces

Interfacial behaviour of proteins has been studied and characterised widely with experimental techniques[89-93]. However, information on protein structural changes in the course of surface adsorption is still limited. Experimental conclusions on unfolding of globular proteins at interfaces even led to contradictory results[93]. Simulation studies of protein behaviour and adsorption in the presence of gas/liquid[94, 95], liquid/liquid, and solid/liquid[96] become more and more popular to study the mechanisms in atomic detail excluded in experiments. Wijman and Dickinson[97] investigated the competitive displacement of a protein monolayer at a liquid interface by more affine particles such as surfactant molecules using Brownian dynamics simulation. In their simulation, the protein layer was bonded to form a connected network as a gel-like structure with certain holes for displacement molecules. At the beginning of the simulation, the first displacement molecules entered these holes in the network, and then the displacement started from these patches until the entire protein layer was displaced from the interface. Their study was limited by the representation of the film and adsorber particles as uniform spheres, however the authors concluded the validity and transferability of their work to more complex systems and did not expect the pathway or findings to change[97]. Euston et al.[95] studied the beer foam stabilization mechanism of hop derived iso-α-acid on barley lipid transfer protein (LTP) with molecular dynamics simulation in bulk and in the presence of a vacuum/water interface. The bulk simulations showed the preferential binding sites of hop acid to the protein by hydrophobic interactions. In the interfacial simulations the hop acids rapidly adsorbed to the interface and offered a preferential binding site of LTP compared to the bare vacuum/water interface. Their results confirmed the previously suggested
stabilization mechanism of cross-linkage of LTP at the interface by the hop acid and additionally suggested increased adsorption leading to a higher LTP concentration, further stabilizing the foam bubbles in beer[95].

Engin and Sayar[94] (2012) investigated the secondary structure, adsorption behaviour and monolayer assembly of a 24-residue amphiphilic peptide with molecular dynamics simulation. The simulations of a single molecule in bulk and in the presence of an air/water interface showed that in both cases the peptide adopts a β-hairpin structure, but in the latter adsorption and secondary structure formation occurred simultaneously. In silico mutants revealed the hydrophobic residues are necessary for secondary structure formation and adsorption, whereas the turn region is only required for perfect β-hairpin structure. They calculated the free energy of adsorption of this peptide and the potential of mean force between two molecules in the parallel and antiparallel arrangement, whereas the latter was preferential due to electrostatic and van der Waals interactions between aromatic residues. The optimal packing density was validated with the simulation of a 32-molecules peptide monolayer.

Penna et al.[96] investigated the adsorption mechanism of two peptides, SD152 and A3, to a water/solid interface with over 240 MD simulations. They suggested a general adsorption mechanism comprising the three phases of biased diffusion towards the interface, anchoring and lockdown of the peptides. They identified preferential amino acids for these events and concluded the relevance of their results for protein molecules[96]. However the simulation of biomolecules at interfaces is still limited to small proteins and peptides, and large conformational changes such as the unfolding of an entire four-helix bundle protein at an interface is computationally very demanding. Prévost and Ortmans (2001) [98] aimed to investigate the structural changes of Apolipoprotein E (Apo E) at water/lipid interfaces with molecular dynamics simulation. This molecule exists as elongated bundles of four α-helices in aqueous systems, but reveals an unfolding similar to DAMP4 upon lipid association. However, due to computational limitations of this study their production simulations were restricted to 1.5 ns of Apo E in the presence of an interface (a water/CCl₄ model system) and in bulk, and only the influence of the interface on the structure without any unfolding or adsorption happening was analysed. They found that no change in helical content, but an increased backbone hydration with a higher number of H-bonds to water in proximity to the interface occurred. The interface favoured hydration of the protein as well as more water molecules penetrating the core.

2.2.7 Potential of mean force (PMF)

The relative probability of finding a molecule in a particular conformation is determined by the free energy difference between the conformational states[99]. In Chapter 5, the free energy difference
between the folded four-helix bundle structure and an unfolded conformation with all four helices in parallel was calculated using the potential of mean force (PMF), $W(r)$, which corresponds to the free energy profile of the system. The two states of interest are linked through a particular pathway, the reaction coordinate $r$. Simulations are carried out with the molecules constrained to points along $r$. The free energy profile can be obtained by integration of the constraint force along the pathway. The free energy difference between the constrained initial and final points can then be obtained [99].

The PMF relative to a reference state $W(r^*)$ is determined by

$$W(r) = W(r^*) - k_B T \ln \frac{\langle \rho(r) \rangle}{\langle \rho(r^*) \rangle}$$

(12)

where

$$\langle \rho(r) \rangle = \frac{\int e^{-\beta H(\Gamma)} \delta(r-r_1) d\Gamma}{\int e^{-\beta H(\Gamma)} d\Gamma}$$

(13)

is the probability that the molecule is at position $r_1$ on the reaction coordinate. $\Gamma$ represents the positions and velocities of all the molecules in the system, $H(\Gamma)$ is the energy when the system has those positions and velocities, $\delta(r-r_1)$ is the Dirac delta function that is zero unless the configuration of the molecule is at $r_1$ on the reaction coordinate, and $\beta = 1/k_B T$ with Boltzmann constant $k_B$ and temperature $T$ [100]. We take $r^*$ to be the initial configuration and set $W(r^*) = 0$.

In most cases the potential of mean force cannot be determined directly from an unconstrained MD simulation, as the system space cannot be sampled evenly and there might be several local energy barriers within the transition process along the reaction coordinate. These areas would result in a low sample density, whereas other low energy states would result in high sampling. To overcome this problem, methods such as umbrella sampling [101] or thermodynamic integration [102] can be applied. Therefore, the system is constrained with a force (“spring”) at different selected configurations along the reaction coordinate, enforcing equilibrium simulations with the collection of averaged constrained forces, and ensuring a complete coverage of the sampling space.

It can be shown that if the constraint is very strong, then

$$W(r) - W(r^*) = G(r) - G(r^*)$$

(14)

where $G$ is the free energy of the system with a constraint applied. Typically harmonic potentials are applied to constrain the system, so this corresponds to a case with high force constants. The total energy of the system then becomes

$$H'(\Gamma, r_0) = H(\Gamma) + g(r, r_0)$$

(15)

where $g(r, r_0)$ is the constraint potential, and $r_0$ is the fixed position on the reaction coordinate to which the system is constrained.

The derivative of the free energy $G$ with respect to $r_0$ is calculated as the ensemble:
\[
\frac{\delta G(r_0)}{\delta r_0} = \langle \frac{\delta H(\Gamma, r_0)}{\delta r_0} \rangle_r = \langle \frac{\delta g(r, r_0)}{\delta r_0} \rangle_{r_0}. \tag{16}
\]

\(...\)_{r_0} is the ensemble average calculated with the system constrained to oscillate about \(r=r_0\).

In our study we constrain the distance between selected sites, \(r_0\), so a constraining potential is added to the Hamiltonian \(H\) with the constraining force constant \(k\)

\[
H(\Gamma, r_0) = H(\Gamma) + \frac{1}{2} k [r - r_0]^2 \tag{17}
\]

so that the difference in free energy along the reaction coordinate from \(r_a\) to \(r_b\) results in

\[
\Delta G = \int_{r_a}^{r_b} \langle \frac{\delta g(r, r_0)}{\delta r_0} \rangle_{r_0} d r_0
= \int_{r_a}^{r_b} \langle k(r - r_0) \rangle_{r_0} d r_0
= - \int_{r_a}^{r_b} \langle F(r, r_0) \rangle_{r_0} d r_0. \tag{18}
\]

Consequently \(\Delta G\) corresponds to minus the integral of the mean force acting due to the constraint and the addition of the constraining force enables an even sampling of the mean force at particular chosen values of the reaction coordinate[102].

### 2.2.8 Limitations

Despite the widespread use and success, molecular dynamics has its limitations. For the simulations, two main challenges exist, including the accuracy of the current force fields, and the simulation time, which is presently limited to approximately one microsecond due to the high computational demand[49].

One has to be aware that force field parameters such as partial charges, van der Waals interactions or values for bonds and angles, are determined empirically from either experimental data (spectroscopy) or QM calculations. These values are approximate and thus not exact and accurate, so simulation results are subject to these errors[103]. The force fields are only approximations of the quantum mechanical reality and require further refinement to accurately reproduce the behaviour of macromolecules in certain cases. In particular, when quantum effects are important, the force fields are poorly suited for these systems. This problem can be addressed by adding quantum mechanical components to the force fields for certain parts of the system, which are used in addition to the molecular-dynamics parameters for the larger system as for example applied by Hong et al.[104]. Further the development of new and improved force fields has been a focus in recent years, with more extensive validation against experimental data[42]. Hereby it is important to
choose a force field which is adequate for the desired application in order to reproduce the behaviour of the molecules under the conditions the simulation is run realistically[105].

An effect which is currently not included in most of the simulations is electron polarization. All charges assigned to atoms are fixed and usually atom centred, whereas in reality electron clouds are flexible and constantly shift between the different atoms bonded depending on their environment. Effort has been made towards the development of polarizable force fields[106-108] but so far with limited success and application, but might improve the accuracy of simulations in the future[49, 109].

Due to the computational demand, simulation time might be limited, which can result in insufficient conformational sampling. In this case results might not be representative and the accessibility of particular states might be limited through high energy barriers. It is a general problem that the amenable time scale of a simulation is usually a magnitude shorter than the time scale in which the majority of real biological events take place, which is in the region of microseconds to milliseconds.

Two ways this issue could be addressed is by lowering the energy barrier for certain events, e.g. by applying extremely high temperature for protein unfolding[79]; or with improved hardware specialized to speed up molecular simulations to extend their time scale, as for example the special-purpose parallel supercomputer Anton. This machine allows classical MD simulations to be on the millisecond-scale and with increasing simulation time the chances for a rare event to be sampled increase. This new time scale of simulation time enables researchers to investigate a whole range of biochemical phenomena on the atomic level including protein-protein interactions, molecular structural changes due to their function, and protein folding[110].

The equations of motion for interacting atoms are integrated by an MD program through its time integration algorithm. It is based on finite difference methods and integration is performed in distinct time steps $\Delta t$. The iteration results in approximates which are associated with a certain error such as through truncation of the Taylor expansion, rounding off errors in computation, or a specified cutoff for non-bonded interactions. This can be minimized by choosing a small timestep $\Delta t$, however this is always a tradeoff between accuracy of the calculation and simulation speed or time[105].

In addition, a good starting model needs to be available for a successful simulation, and the quality of a theoretical model requires thorough verification before use, as demonstrated in section 2.2.3. Only a high quality starting model can give reliable and high quality results in a simulation. Despite these existing shortcomings, the number and success of molecular dynamics simulations as of this day have proven it an important tool in modern structural biology, and the current work and solutions addressing these problems at the moment and in the future make great progress in overcoming these limitations[49].
2.3 References

52 D. C. Rapaport, in The Art of Molecular Dynamics Simulation; Cambridge University Press, **2004**.
57 A. Fiser, R. Sánchez, F. Melo, A. Salis, in Computational Biochemistry and Biophysics; Marcel Dekker, inc., New York, **2001**.
103 S. Qamar, in *DNA Translocation through a Molecular Nanopore: A Molecular Dynamics Study*; ProQuest, Ann Arbor, **2009**.
105 A. Gunaratne, in *A Penalty Function Method for Constrained Molecular Dynamics*; ProQuest, Ames, **2006**.
Chapter 3    Predicting recombinant expression

3.1 Predicting recombinant expression experiments using molecular dynamics simulation

This section consists of the journal article published as:

3.1.1 Abstract

Soluble expression of de novo-designed proteins in Escherichia coli (E. coli) remains empirical. For given experimental conditions expression success is determined in part by protein primary sequence. This has been previously explored with varying success using a variety of statistical solubility prediction tools though without taking fold stability into account. In the present study, the three-dimensional structure of proteins in molecular dynamics (MD) simulations is used to predict expression as a new approach with a set of four-helix bundles. Stability-related parameters for ten structures were determined in a thermal unfolding MD simulation and used to build statistical models with a support vector machine (SVM) classifier. The most accurate models were identified by their performance on five independent four-helix bundle sequences. The final model provided accurate classification prediction for this test set and was successfully applied in a model challenge with two newly designed sequences.

The combination of simulation-derived parameters and an SVM classifier has potential to predict recombinant expression outcome for this set of four-helix bundles. With further development, this approach of utilizing higher-dimensional protein structural information to predict expression may have potential to advance recombinant biotechnology through modern computational and statistical science.

3.1.2 Introduction

Solubility and stability of a protein upon heterologous expression under a given set of conditions in the most widely characterised and used host system Escherichia coli (E. coli) are determined in part by the protein's amino acid sequence[1-3]. Depending on the sequence and the growth conditions, overexpression frequently does not lead to the desired stable, soluble and active protein, but rather to the formation of insoluble aggregates called inclusion bodies, or proteolytic degradation of the
product, meaning no expression is observed[4-6]. Misfolding, low folding rates, and structural instability can, alone or in combination, lead to aggregation or degradation, thus causing low or no levels of the desired protein to accumulate. Different computational tools have been used to address these problems. Statistical methods were applied to predict proteolytic degradation[4] or formation of inclusion bodies[4, 7]. Molecular dynamics and related simulation methods have been used to elucidate mechanisms and origin of protein folding, misfolding and aggregation[8-10]. Bioinformatic analysis can assist to predict protease cleavage sites and stop non-specific cleavage of proteins[11]. But despite these efforts and existing design recommendations to improve expression yield for de novo sequences, this outcome is often not preventable[4, 6]. Experimental methods to circumvent this problem, including reduction of cultivation temperature, coexpression of chaperones or foldases, secretion of the target protein into the periplasm or media, and expression of the target sequence as fusion proteins, are applied empirically and routinely[3, 5, 12]. The success of these methods is nevertheless not guaranteed and the entire process of obtaining a soluble protein from a known sequence relies on slow and costly trial and error[5, 13].

Researchers have been addressing this problem computationally since Wilkinson and Harrison developed the first prediction model for solubility upon overexpression in E. coli in 1991[7]. They predicted the formation of inclusion bodies using statistical analysis. Six parameters related to folding and solubility in 81 proteins were examined with discriminant analysis. In descending order, charge average, turn forming residue fraction, cysteine and proline fraction, hydrophilicity and total number of residues were identified to be important[7]. Later it was discovered that only the first two were necessary to distinguish between soluble and insoluble expression. The new model was applied to predict solubility of the originally-insoluble protein human interleukin-3 upon fusion to the native E. coli proteins NusA, GrpE and bacterioferritin[14]. Idicula-Thomas and Balaji (2004) studied the relationship between the primary structure of a protein and its solubility on overexpression in E. coli for the identification of the most important sequence-dependent features causing this behaviour. They systematically investigated these features in their 170 data sets corresponding to solubility or inclusion body formation and used them to build a final mathematical model with discriminant analysis and solubility index computation. From a test set of 40 protein sequences 62.5% were predicted correctly, with their solubility index formula containing the variables of thermostability, in vivo half-life, Asn, Thr and Tyr fraction and the tripeptide composition[15]. Idicula-Thomas et al. further modified their model using a support vector machine-based algorithm and the primary sequences of 192 proteins in the form of residue and dipeptide composition, together with six physicochemical properties. They reached a prediction accuracy of 72% and identified the aliphatic index, several residue or mono- or di-peptide contents like Glu, His-His, Arg-Gly, Arg and Gly, the instability index, and the net charge of the protein, to
be among the most important factors influencing solubility. One of the main advantages of their models is that the influence of point mutations on solubility during expression under the same conditions can be predicted easily[16]. Smialowski et al. (2007) presented an approach called PROSO, which was later modified to PROSOII[1], where a combination of a support vector machine and a Naïve Bayes classifier led to an improved capacity to distinguish between soluble and insoluble expression outcome, with overall prediction accuracy of 72%. Validation was conducted with experimental measurements of 31 variants of two different proteins. Among features selected as most important for classification were frequencies of eight amino acids (Arg, Asp, Cys, Glu, Gly, Leu, Met, Ser), relative fraction of negatively charged residues and three frequencies of dipeptides (Glu-Gly, His-Met and Lys-Gly)[17]. Chan et al. were the first to take the entire cloning and expression region for their calculations by means of a support vector machine algorithm into account, as some combinations of vectors and target proteins lead to no expression in *E. coli*[4]. Whereas previous studies assumed that the protein was expressed, they created different models including 617 features that distinguished between three different classes: no expression, soluble expression and expression but formation of inclusion bodies. They demonstrated that their best model performed better than models in previous studies[4, 7, 17] with a test accuracy of 83%. Huang et al. (2012) developed an optimised scoring card method based on dipeptide composition to calculate the solubility probability of target proteins. This approach provided a model that emphasised the high solubility tendency of thermophilic α-helical proteins and had a performance comparable to Chan et al. It concluded a high contribution to solubility of the dipeptides Leu-Ala, Ile-Pro and Met-Cys and of the single amino acids Ala, Glu, Asp and Lys[18].

All these studies directly link the primary structure of the proteins, including some physicochemical properties that can be related therefrom, to expression. A more precise assertion might be possible, if sequence and expression could be linked via the 3-dimensional (3-D) structure and the behaviour of the molecule. Molecular dynamics (MD) simulation is a suitable tool for this inter-linkage as it has developed as a powerful device in the last decades and can provide molecular and atomistic insights into mechanisms, kinetics and chemical processes. MD can give a deeper understanding of the fundamental principles and be an important complement to experimental results. With increasing advances in computing and software, the improving accuracy and description of the underlying physical systems and the use of supercomputer power, the simulation of larger and more realistic systems is becoming feasible[19]. MD simulations have been used for a large range of applications recently[20-28]. Despite this variety, studies directly linking MD simulations and heterologous overexpression are rare and limited to specific case studies. Tang et al. ascribed the altered activity of 14 Galactose-1-phosphate uridylyltransferase variants, which leads to a potentially lethal disease in humans, to an altered stability and solubility. This theory was
consolidated by successfully relating \textit{in vitro} expression results to structural changes on the molecular base in a combination of molecular modelling analysis and molecular dynamics simulations\cite{29}. More frequently expression behaviour for specific cases is explained using MD, for example the role of the steric chaperone Lif26 from Acinetobacter sp. XMZ-26 upon expression of the lipase Lip26\cite{30} or the different activity after expression of mutant R344D of human cathepsin A to the wild type and other R344X mutants\cite{31}.

To the best of our knowledge MD simulation has not been applied to predict expression behaviour of proteins from the amino acid sequence so far. In the present study we performed all-atom MD simulations to predict expression as a case study of de novo-designed four-helix bundles in \textit{E. coli}. Four-helix bundles have been studied extensively\cite{6, 32-34} as they offer a small and simple structural motif but have, however, a well-defined fold and hydrophobic core exposing polar side chains to the solvent. They consist of one defined secondary structure element and are easily designed by following a binary patterning of polar and non-polar amino acids\cite{6, 32}. Using this technique Kamtekar et al. systematically created 48 four-helix bundle sequences, but only 60\% of them folded into a stable and soluble compact globular structure leading to recombinant expression in \textit{E. coli}\cite{6}. The substitution of a single amino acid can cause a destabilisation of the native structure of natural and de novo-designed proteins that can give rise to proteolytic degradation\cite{6}, which makes it desirable to develop an a priori prediction tool.

The advantage of a MD simulation based prediction tool is that sequence-dependent features or physicochemical properties that turned out to be important in previous studies are included automatically by means of the protein's behaviour in the simulation. In this study thermal unfolding molecular dynamics simulations were performed, as successful expression of a certain amino acid sequence is suggested to be related to the stability of the native structure of a protein\cite{6, 29, 34}. The thermal stability can be a key determinant to the protein's sensitivity for proteolytic degradation\cite{6}. Homology models of ten new four-helix bundle sequences designed as variants of a recently-reported thermostable four-helix bundle\cite{35, 36}, for which \textit{in-vitro} expression behaviour was tested, were created and simulated at 498 K for 25 ns. Fifteen stability-related features were recorded. The strongest descriptors to distinguish between the two groups (soluble expression, no expression) were identified and used to build a statistical model with a support vector machine (SVM) algorithm. Previous studies could show an improvement in prediction power by addition of simulation-derived descriptors to a statistical approach for other biological problems\cite{37, 38}. Five new sequences from a test set were modelled and simulated under the same conditions, and were successfully classified into the correct group with the final model. A conclusive model challenge with two independent sequences was conducted.
3.1.3 Methods

Four-helix bundle design
DAMP4, MD(PS-MKQLADS-LHQLARQ-VSRLEHA-D)_4[35], and anionic AS_{Prot} MD(PS-ANSVAES-LANLAES-VSELVSNA-D)_4, are two de novo-designed four-helix bundle sequences[39]. They follow a seven residue α-helical pattern, where every 1st and 4th residue is hydrophobic. The four identical helices are connected via a DPS-linker sequence which ensures helix termination, promotes electrostatic repulsion between adjacent loops and provides sufficient flexibility for folding into a stable four-helix bundle[35, 40].

Table 3-1 Four-helix bundle sequences[39] used to build the model (Training Set) and independent sequences used to test predictions (Test Set). Solubility in overexpression was determined experimentally.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Theoretical isoelectric point[41]</th>
<th>Solubility in over-expression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAMP4</td>
<td>MD-[PS-MKQLADS-LHQLARQ-VSRLEHA-D]_4</td>
<td>6.7</td>
<td>yes</td>
</tr>
<tr>
<td>DAMP4ΔS</td>
<td>MD-[P- MKQLADS-LHQLARQ-VSRLEHA-D]_4</td>
<td>6.7</td>
<td>no</td>
</tr>
<tr>
<td>AS_{Prot}</td>
<td>MD-[PS-ANSVAES-LANLAES-VSELVSNA-D]_4</td>
<td>2.98</td>
<td>yes</td>
</tr>
<tr>
<td>SP1</td>
<td>MD-[PS-AKSVAES-LHSLARS-VSRLVEHA-D]_4</td>
<td>6.7</td>
<td>yes</td>
</tr>
<tr>
<td>SP2</td>
<td>MD-[PS-AKSVAKS-LHSLARS-VSRLVEHA-D]_4</td>
<td>11.6</td>
<td>no</td>
</tr>
<tr>
<td>SP3</td>
<td>MD-[PS-AHSAVES-LHSLARS-VSRLVEHA-D]_4</td>
<td>6.14</td>
<td>yes</td>
</tr>
<tr>
<td>SP4</td>
<td>MD-[PS-AHSAHSA-LHSLARS-VSRLVEHA-D]_4</td>
<td>6.98</td>
<td>no</td>
</tr>
<tr>
<td>SP5</td>
<td>MD-[PS-AHSAHSA-LHSLARS-VSRLVHHA-D]_4</td>
<td>11.52</td>
<td>no</td>
</tr>
<tr>
<td>SP6</td>
<td>MD-[PS-AHSAHSA-LHSLARS-VSRLVSHA-D]_4</td>
<td>11.55</td>
<td>no</td>
</tr>
<tr>
<td>SP7</td>
<td>MD-[PS-AHSAHSA-LHSLAES-VSELVSNA-D]_4</td>
<td>4.67</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP12</td>
<td>MD-[PS-AESVAES-LAELAES-VSELVSEA-D]_4</td>
<td>2.83</td>
<td>no</td>
</tr>
<tr>
<td>SP13</td>
<td>MD-[PS-AQSAQAS-LQLAQAS-VSQLVSAQ-D]_4</td>
<td>3.24</td>
<td>yes</td>
</tr>
<tr>
<td>SP17</td>
<td>MD-[PS-ANSAKS-LANLAES-VSKLVSNA-D]_4</td>
<td>10.00</td>
<td>no</td>
</tr>
<tr>
<td>SP18</td>
<td>MD-[PS-ANSAKS-LANLAES-VSKLVSNA-D]_4</td>
<td>12.00</td>
<td>no</td>
</tr>
<tr>
<td>SP19</td>
<td>MD-[PS-ANSAVS-LANLAES-VSKLVSNA-D]_4</td>
<td>6.12</td>
<td>yes</td>
</tr>
<tr>
<td>EDP-11 [42]</td>
<td>- - [ PG-IAELEAE-LSAVEAE-LEAI LAEL-D]_4</td>
<td>2.81</td>
<td>no</td>
</tr>
<tr>
<td>QDP-11</td>
<td>MD-[PS-IQLQAQL-LSAVQAQL-LSAVQAQL-D]_4</td>
<td>3.24</td>
<td>yes</td>
</tr>
</tbody>
</table>

Additional sequences reported in this study are variants of these two proteins, which vary in sequence length, net charge and charge distribution from the original molecules[39]. All sequences are provided in Table 3-1.

Recombinant expression in *E. coli*
Cloning and expression experiments were conducted by the Protein Expression Facility (PEF) at The University of Queensland.
To express the sequences, a codon-optimized synthetic gene was cloned into the pET-48b(+) vector. Chemically-competent *E. coli* BL21(DE3) were transformed with the vector via heat shock treatment and used to express the sequences. Bacterial cells were streaked on LB plates (Amresco LB agar, Miller formulation, tissue culture grade, Solon, OH, USA) containing kanamycin sulphate (15 μg ml\(^{-1}\), Gibco, Invitrogen, SKU# 11815) and a single colony was selected to inoculate 100 ml of LB medium (Amresco LB broth, Miller formulation, tissue culture grade, Solon, OH, USA) containing kanamycin sulphate (15 μg ml\(^{-1}\)) in shake-flask cultures. Cultures were incubated at 37 °C on an orbital shaker until OD\(_{600}\) reached approximately 0.5. Each culture was induced with 1 mM Isopropyl-β-D-thiogalactopyranosid (IPTG) and further incubated at 26 °C overnight until harvest at approximately 20 h post induction. Expression was assessed by SDS-polyacrylamide gel electrophoresis (PAGE). Samples were taken immediately prior to induction and at harvest and were lysed in BugBuster (Novagen, Madison, WI, USA). SDS-PAGE (NuPAGE Bis-Tris 4-12% gels (Invitrogen, Carlsbad, CA, USA)) was conducted using standard procedures and staining to assay for a new band at the expected size calculated with ProtParam (http://web.expasy.org/protparam/).

Figure 3-1 SDS-PAGE gel showing NOVEX® Sharp Pre-stain molecular weight ladder (lane 1), DAMP4 prior to induction of recombinant protein expression (lane 2) and when harvested (lane 3), as an example of soluble expression. Lanes 4 & 5 are for sequence SP17 and lanes 6 & 7 for SP18 at induction and at harvest, respectively. Both sequences were not expressed under these conditions. The image was produced using Bio-Rad Gel-Doc™ imaging system.
Figure 3-1 shows a SDS-PAGE gel of a solubly-expressed sequence (DAMP4) having a molecular weight of ~11 kDa which resulted in high yield and two sequences (SP17, SP18) having an expected molecular weight of ~10 kDa that resulted in no expression, as an example of the type of experimental data obtained to assess experimental overexpression.

**Homology modelling**

All homology modelling experiments and simulations were conducted in the MD software package YASARA structure version 12.6.28 (YASARA Biosciences GmbH, Vienna, Austria)[43, 44]. Molecular graphics were created with YASARA (www.yasara.org) and POVRay (www.povray.org). As no homologous sequences using PSI-BLAST[45] for DAMP4 were found in the protein data bank (PDB) a single helix was modelled first. The helix model was based on the aligned sequence of template 3EDC[46]. Three copies were created and the four helices were connected manually. The final model was optimized with another modelling step to enhance side-chain rotamers, energy minimized and refined using the YASARA2 force field[47] with its optimized cutoff for Van der Waals interactions of 7.86 Å. For modelling the ASProt sequence the online fold recognition server SPARKS-X[48] was used, as homology modelling also failed due to lack of sequence similarity. The four-helix bundle fold, which was the one with the highest alignment score and hence the most likely, was based on the designed helical protein part of template 1Y4C[49].

![Figure 3-2 The 3-dimensional structures of DAMP4 a) front view and c) top view and ASProt b) front view and d) top view, derived from homology modelling.](image)

As model quality according to YASARA's structure validation tool[50] was not optimal this was followed by a homology modelling experiment providing the fold recognition model as a template
to optimize side-chain rotamers. An energy minimization and a refinement simulation were performed.

To assure the quality of the models displayed in Figure 3-2, two different tools were used. The in-YASARA implemented structure validation tool revealed overall Z-scores of 2.732 for DAMP4 and 0.284 for ASProt. This quality check compares the target structure to a set of gold-standard references, made comparable by normalized knowledge-based potentials. It describes how many standard deviations a structure is away from the average, therefore the DAMP4 and ASProt models are considered optimal [50]. The ProSA (protein structure analysis) program [51, 52] produced similar results. With ProSA Z-scores of -6.69 and -7.59 for DAMP4 and ASProt, respectively, both lie within the range of Z-scores typically found for proteins in the PDB. The energies were well below zero indicating no problematic or defective parts. As sequences DAMP4ΔS, SP1-SP7, SP17-SP19, EDP-11 and QDP-11 (Table 3-1) are variants of DAMP4 and ASProt, they were built in a homology modelling experiment providing these two structures as templates.

**MD simulation**

All simulations were carried out using the Amber03 force field [53] and explicit water molecules. The distance for van der Waals interactions was set to a medium range cutoff of 7.86 Å for consistency with the homology modelling, and long-range electrostatics were calculated using the Particle Mesh Ewald (PME) [54] algorithm. Periodic boundary conditions and a Berendsen Thermostat [55] for temperature control were applied. The standard time step of 2.5 fs and 1.25 fs for inter- and intramolecular forces, respectively, was used as the optimal compromise between simulation speed and accuracy of the integration. For analysis purposes simulation snapshots were saved every 2 ps. All simulations were performed in triplicate. A cubic simulation box was built with a distance of 35 Å around all protein atoms. The box was filled with TIP3P water molecules in a neutralization experiment [56], which adds water molecules and counter ions according to the set density and equilibrates them in a short MD simulation of the solvent. An equilibration simulation was run for 100 ps at 298 K, saving snapshots at 60 ps, 80 ps and 100 ps as starting conformations for the production simulation (25 ns at 498 K). Using temperatures significantly above the melting point is a common practice in unfolding MD simulations as it speeds up the unfolding process without altering its pathway [57, 58]. The study by Daggett and co-workers [57] concluded 498 K as a suitable temperature for protein unfolding simulations at a minimum of computational cost. In the solubility prediction model building process the three starting structures were considered separately as single runs, whereas in the following model application the three runs were considered together leading to a final prediction. If the results of these runs were inconsistent the result of the majority was taken. During the simulation, 15 stability-related parameters of the protein were recorded:
total-, bond-, angle-, dihedral-, planarity-, van der Waals (vdW)- and electrostatic-energy, radius of gyration, root mean square deviation (rmsd), solvent accessible surface area, secondary structure percentages (α-helix, turn and coil), intramolecular H-bonds and H-bonds between the molecule and water.

**Data processing**

Analysis and processing of simulation results were performed in MATLAB R2012a (Mathworks®, Natick, USA). A distinction was made between parameters that were normalized to the starting point of the simulation prior to analysis such as solvent accessible surface area, helix-, turn- and coil-content, intramolecular H-bonds and radius, and parameters where the raw data such as energies and rmsd values were used. Data were fitted with a smoothing function by performing a local regression using a weighted linear least squares and a 1st degree polynomial model with span 80%. As only the endpoints of the fitting spline at simulation time 25 ns are relevant to distinguish between the behaviour of the two groups (expression, no expression), these points were used to build the final prediction model.

**Support vector machine classification**

In this study a support vector machine classifier (SVM) was applied as this tool does not require the variables have a normal distribution. The SVM approach has been successfully applied to a variety of biological problems[4, 59-63]. This algorithm is based on a class of hyperplanes

\[
\langle w, x \rangle + b = 0
\]  

with normal vector \( w \in R^N \) and bias \( b \in R \). \( \langle \cdot \rangle \) defines the inner product and \( x_i \in R^N \) are the data points, the vectors which comprise the \( N \) descriptors and belong to one of the two classes \( y_i = \mp 1 \) (soluble expression or not). The optimal hyperplane is the one that separates these classes so that the distance between the points closest to the hyperplane and the plane is maximized. Like this, points which are not as close to the training data are nevertheless on the right side of the hyperplane and thus classified correctly. The closest points are called the support vectors.

To determine the optimal hyperplane, \( w \) and \( b \) are scaled so that

\[
|\langle w, x_i \rangle + b| = 0
\]  

and thus

\[
y_i\langle w, x_i \rangle + b \geq 1
\]

is true.

As the margin perpendicular to the optimal hyperplane equals \( 2/\|w\| \), it is maximized, when \( \|w\| \) is minimized.
and therefore the optimal hyperplane can be determined by the solvation of the quadratic optimization problem with a Lagrange function which results in \( w \)

\[
w = \sum_i \alpha_i y_i x_i
\]  

(22)

with the Langrangian multipliers \( \alpha_i \geq 0 \) for the support vectors. This allows the calculation of \( b \) and new data points \( x_{new} \in \mathbb{R}^N \) can be classified using the decision function

\[
f(x_{new}) = \text{sign}((w, x_{new}) + b) = \text{sign} \left( \sum_i y_i \alpha_i \langle x_i, x_{new} \rangle + b \right)
\]  

(23)

as this equation determines the direction of the point subject to the hyperplane and thus the affiliation to one of the groups.

In the case of non-linear separable data, the fact that this function only depends on the inner product, is used to lift the separation problem into a higher dimensional feature space \( F \)

\[
\phi: \mathbb{R}^N \rightarrow F
\]  

(24)

and with an efficiently evaluated kernel-function \( k \), the same linear algorithm is solved. In this study a quadratic kernel

\[
k(x_i, y_i) = (1 + \langle x_i, y_i \rangle)^2
\]  

(25)

with the corresponding decision function

\[
f(x_{new}) = \text{sign} \left( \sum_i y_i \alpha_i k(x_i, x_{new}) + b \right)
\]  

(26)

was used[64, 65].

3.1.4 Results and discussion

Sequence analysis

Sequence features leading to no or insoluble expression are still poorly understood, mainly due to the complexity of the balance of underlying biochemical effects and processes which determine expression outcome[66]. Thus reasons behind failure of soluble expression can only be suggested. For example DAMP4 variant DAMP4ΔS is lacking the Ser in the linking turn region designed to connect the four helices (Table 3-1). This could cause a lack of flexibility that results in an insoluble and incorrectly folded protein which is likely degraded by internal bacterial proteases. SP1 is another variant, designed to be thermally stabilized by removal of Met and Gln, and addition of an extra hydrophobic residue in the final heptad to adjust for helix tilt which results in high yield. Compared to SP1, two negatively charged Glu on each helix of SP2 were replaced with positively
charged Lys. This causes an unbalanced charge structure with five positively charged residues vs. only one negatively charged residue on each helix. As suggested previously[17, 66-69] with a decreasing number of negative charges related to a high isoelectric point[66, 70] of 11.6, this protein is less likely to express, which was confirmed experimentally. Furthermore the stabilizing salt bridges between Lys and Glu, which can help to facilitate folding, are broken. In comparison to SP1, only a single amino acid on each helix of SP3 was altered by replacing a positively charged Lys with His which resulted in an accumulation of soluble protein and no change in expression outcome. From SP3 to SP4, however, a negatively charged Glu was replaced by another His resulting in good expression but an insoluble product. This can either be caused by the lower number of negatively charged residues resulting in less electrostatic repulsion and thus aggregation, or the high number of His residues may enable metal-mediated histidine linkage. For SP5 another negative charge was removed and replaced with a His leading to proteolytic degradation of eventually expressed protein, possible difficulties in expression are as well indicated by a high PI (11.52). The PI of SP6 is also in this range, which corresponds to an expected outcome of no expression. SP7 is, compared to SP5 and SP6, stabilized by negative charges. As Arg residues correlate with decreased expression[66], the replacement of two Arg residues with Glu on each helix promotes soluble expression.

Table 3-2 Prediction results into insoluble (I) and soluble (S) from existing prediction tools of our set of four-helix bundle sequences.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Solubility upon expression</th>
<th>PROSO[17][a] Prediction (probability)</th>
<th>PROSOSII[1][a] Prediction (solubility score)</th>
<th>SOLpro[2][b] Prediction (probability)</th>
<th>Harrison et al.[7, 14][c] Prediction (probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAMP4</td>
<td>yes</td>
<td>I (0.574)</td>
<td>I (0.164)</td>
<td>S (0.807)</td>
<td>S (59.2%)</td>
</tr>
<tr>
<td>DAMP4ΔS</td>
<td>no</td>
<td>I (0.571)</td>
<td>I (0.049)</td>
<td>S (0.741)</td>
<td>S (71.9%)</td>
</tr>
<tr>
<td>ASProt</td>
<td>yes</td>
<td>I (0.572)</td>
<td>I (0.338)</td>
<td>S (0.940)</td>
<td>S (81.5%)</td>
</tr>
<tr>
<td>SP1</td>
<td>yes</td>
<td>I (0.580)</td>
<td>I (0.281)</td>
<td>S (0.953)</td>
<td>I (79.4%)</td>
</tr>
<tr>
<td>SP2</td>
<td>no</td>
<td>I (0.585)</td>
<td>S (0.715)</td>
<td>S (0.934)</td>
<td>S (71.7%)</td>
</tr>
<tr>
<td>SP3</td>
<td>yes</td>
<td>I (0.576)</td>
<td>I (0.589)</td>
<td>S (0.856)</td>
<td>I (54.4%)</td>
</tr>
<tr>
<td>SP4</td>
<td>no</td>
<td>I (0.576)</td>
<td>S (0.815)</td>
<td>S (0.814)</td>
<td>I (79.4%)</td>
</tr>
<tr>
<td>SP5</td>
<td>no</td>
<td>I (0.579)</td>
<td>S (0.843)</td>
<td>S (0.865)</td>
<td>I (93.9%)</td>
</tr>
<tr>
<td>SP6</td>
<td>no</td>
<td>I (0.579)</td>
<td>S (0.947)</td>
<td>S (0.822)</td>
<td>I (88.7%)</td>
</tr>
<tr>
<td>SP7</td>
<td>yes</td>
<td>I (0.567)</td>
<td>S (0.883)</td>
<td>S (0.791)</td>
<td>S (95.2%)</td>
</tr>
<tr>
<td>SP12</td>
<td>no</td>
<td>I (0.572)</td>
<td>I (0.465)</td>
<td>S (0.978)</td>
<td>S (97.0%)</td>
</tr>
<tr>
<td>SP13</td>
<td>yes</td>
<td>I (0.572)</td>
<td>I (0.304)</td>
<td>S (0.978)</td>
<td>I (54.4%)</td>
</tr>
<tr>
<td>SP17</td>
<td>no</td>
<td>I (0.577)</td>
<td>S (0.695)</td>
<td>S (0.958)</td>
<td>I (97.5%)</td>
</tr>
<tr>
<td>SP18</td>
<td>no</td>
<td>I (0.582)</td>
<td>I (0.065)</td>
<td>S (0.956)</td>
<td>I (97.5%)</td>
</tr>
<tr>
<td>SP19</td>
<td>yes</td>
<td>I (0.575)</td>
<td>S (0.738)</td>
<td>S (0.792)</td>
<td>I (88.9%)</td>
</tr>
</tbody>
</table>

Additionally, four positively charged Lys were replaced with Glu, which favours expression as positively charged amino acids may reduce translation efficiency in *E. coli* possibly due to the attraction to negatively charged rRNA[66]. In a preliminary study, we tested existing tools for the prediction of solubility upon expression available online with our set of four-helix bundle sequences (Table 3-2). We found that these tools were either not able to distinguish between the different groups within these sequences and predicted all sequences as soluble (SOLpro[2]) or insoluble (PROSO[17]), or were able to identify differences, but their performances were with 33% (PROSOII[1]) and 53% (Harrison et al.[14]) correct prediction, respectively, not suitable for our set of four-helix bundles, urging for a new approach. It has to be considered that these prediction tools are limited by their assumption that the protein is overexpressed in *E. coli* and do not consider the case of no accumulation of soluble protein due to proteolytic degradation. Further, the scope of these models was not solely inclusive of four-helix bundles, and it is likely a very low number of these structures are included in the test- and training-sets causing this rather random allocation of the bundle sequences.

**Model development**

A support vector machine classifier was applied to create prediction models comprising different combinations of variables from MD simulations. As can be seen from the rmsd values for all sequences (Figure 3-3) the molecules are destabilized within the applied simulation time frame, but do not unfold completely. Nevertheless differences in stability between the different sequences can be observed.

![Rmsd raw data in production simulation of single structures showing destabilization and partial unfolding.](image)

**Figure 3-3** Rmsd raw data in production simulation of single structures showing destabilization and partial unfolding.
The model building process starting from the simulation data collection to the final models in Table 3-3 can be found in Figure 3-4. A set of ten sequences (DAMP4, DAMP4ΔS, SP1, SP2, SP3, SP4, SP5, SP6, SP7, ASProt), each with a triple determination, was used to build the models (training set) and to predict a set of five sequences (SP12, SP13, SP17, SP18, SP19) that was not included in the model-building process (test set).

![Figure 3-4 Flow chart of model building process.](image)

All possible models with data from the training set with two to five descriptors (including total-, bond-, angle-, dihedral-, planarity-, vDW- and electrostatic-energy, radius of gyration, rmsd, solvent accessible surface area, α-helical, turn and coil percentage, intramolecular H-bonds and H-bonds between the molecule and water) were looped as e.g. combinations of two of these descriptors, combinations of three descriptors, etc. The performance of each of these models was judged by means of a 5-fold cross-validation to identify the most important descriptors.

### Table 3-3 Overview of the descriptors of selected prediction models that showed the best prediction performance. E stands for energy.

<table>
<thead>
<tr>
<th>Model</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E_{total}</td>
</tr>
<tr>
<td>2</td>
<td>E_{planarity}</td>
</tr>
<tr>
<td>3</td>
<td>E_{bond}</td>
</tr>
<tr>
<td>4</td>
<td>E_{total}</td>
</tr>
<tr>
<td>5</td>
<td>E_{total}</td>
</tr>
<tr>
<td>6</td>
<td>E_{bond}</td>
</tr>
<tr>
<td>7</td>
<td>E_{total}</td>
</tr>
<tr>
<td>8</td>
<td>E_{dihedral}</td>
</tr>
</tbody>
</table>
A 5-fold cross validation divides the training set into five subsets. Four of these are used to build a model of the selected descriptors, and the remaining part is used for validation. This is repeated five times, whereas each of the subsets will be left out once, resulting in a mean percentage accuracy over the five sets. This method is used to judge the general prediction performance of statistical models. In the next selection step, taking only models with 100% mean correct prediction in the cross-validation into account, the models were further selected through their performance on the test set, resulting in one model with three descriptors, three models with four descriptors, and eight models with five descriptors which all predicted the test set correctly.

A final test with arbitrarily changed test- and training-set composition reduced the number of possible prediction models to one model with three descriptors, four models with four descriptors and three models with five descriptors (Table 3-3).

**Support vector machine results**

The applied SVM algorithm was able to identify models that best distinguished between proteins that could be expressed soluble and stable in *E. coli*, and proteins that could not. The consideration of a single descriptor did not lead to a clear result as expression behaviour is complex and subject to interference by a variety of different effects. As an example, charges can be mentioned.

![Figure 3-5 Model descriptor Energy\_electrostatic for investigated sequences.](image-url)

Three markers for each sequence correspond to the three different starting structures. The red dotted lines represent a notional 'expression slot', used to guide visual interpretation.
In the majority of the prediction tools available in literature a higher number of, especially negative, charges has been identified as one of the most important factors promoting solubility during overexpression[7, 15, 16, 67-69]. Due to electrostatic repulsion between the polypeptides, less or slower aggregation leading to higher soluble expression yield has been observed. It has also been shown that fusion of highly acidic peptides can solubilize proteins during recombinant expression in *E. coli*[71].

A simulation parameter that can be directly linked to charge is $\text{Energy}_{\text{electrostatic}}$. In the MD simulations, it can be observed that expression most likely occurs within a certain range of electrostatic energies, as displayed in Figure 3-5. If the negative charge is too large, and especially when charges are clustered, intramolecular electrostatic repulsion can destabilize the native protein conformation, hampering overexpression[71]. The electrostatic interactions might on the one hand be favourable for the molecule's secondary structure, causing the highly negative electrostatic energy, however their influence on the stable four-helix bundle tertiary fold can be unfavourable bringing about misfolding and degradation. This effect could be observed for expression of sequence SP12, where adjacent negative charges within one turn of the helix lead to intramolecular repulsion (Figure 3-6).

![Figure 3-6 The 3-dimensional structure of SP12.](image)

Negative charges are highlighted in yellow. Negative charges in close proximity within one turn of the helix lead to strong repulsive forces, which destabilize the structure.

These sequences reveal a large negative electrostatic energy which is not favourable for expression. Expression therefore predominantly occurs within a certain electrostatic energy 'slot', yet there are exceptions like DAMP4 and SP17 where other effects must dominate and define the expression outcome. Thus, $\text{Energy}_{\text{electrostatic}}$ is not a sufficient predictor alone. To account for interfering effects and to improve predictive power, two different statistical classifiers, discriminant analysis (data not
shown) and a support vector machine, were applied. The latter performed best. Not all of the simulation parameters appear in the final model to ensure its robustness and to prevent overfitting due to the limited number of training sets. Considering models with 2-5 descriptors lead to a total of 4928 models, which were reduced to the most accurate ones. The resulting eight models, listed in Table 3-3, all showed an accuracy of 100% in predicting the test set. These models mainly contain energy-descriptors, but as well some normalized secondary structure percentages such as helix- or coil-fraction. The most frequent variable is Energy\textsubscript{electrostatic}, which is in good agreement with sequence-based expression prediction tools available in literature.

**Final model selection**

Although all models in Table 3-3 revealed the same optimal prediction accuracy there are a few reasons why Model 1 should be selected for future predictions. Firstly, it is simple; if a model with a smaller number of variables can reach the same prediction power as a more complex system, it should always be preferred. It reduces the risk of overfitting, which is according to Occam's Razor (the principle of parsimony), the use of models or procedures that include more complex systems or more terms than necessary for prediction. This means, descriptors are used without any beneficial value and, even worse, more descriptors can lead to poorer performance[72]. Secondly, Model 1 has less variation in the prediction results. Usually it is expected that prediction accuracy increases when using an additional descriptor. But as an example, compared to Model 1, Model 5 contains supplemental $\alpha$-helical content as a descriptor (Table 3-3).

![Figure 3-7 Progression of simulation variable Energy\textsubscript{total} for DAMP4.](image)

For Model 1, all predictions for the three different starting structures are consistent within each protein; the prediction for SP17 with Model 5 for one of the starting structures is incorrect, but still lead to the overall correct prediction of 'no expression' as two out of three structures infer. This is
due to the fact that the variation in \( \alpha \)-helical content within the three starting structures for each protein is with a variation coefficient of 6.48\% clearly larger than for energy values with 0.12\%, 0.16\% and 1.78\% for Energy\textsubscript{electrostatic}, Energy\textsubscript{bond} and Energy\textsubscript{total}, respectively. An inconsistent result may indicate that a sequence is at a threshold between expression and no expression, e.g. resulting in poor solubility. As can be seen in Figure 3-1, SP17 and SP18 both result in poor expression compared to DAMP4 and are thus considered to be grouped to "no expression". It can be observed that the band at 10 kDa is slightly darker for SP17 than for SP18. However, where to set the threshold in the experimental data between soluble expression and no expression is open to discussion. Thirdly, Model 1 should be chosen to minimize simulation time. Compared to other variables, energies do not continuously increase nor decrease during the simulation. Rather, they reach an equilibrium state at which they oscillate around a certain value according to small fluctuations in temperature (see Figure 3-7) describing an intrinsic energetic state of the protein. This equilibrium is reached after approximately 1 ns, which means, future simulation time can be set much shorter. This approach drastically reduces computational time, while at the same time, ensures equivalent prediction quality. The final model contains the three descriptors Energy\textsubscript{total}, Energy\textsubscript{bond} and Energy\textsubscript{electrostatic} and details from the SVM algorithm for predictions using this model can be found in Table 3-4.

Table 3-4  Details of the final prediction model (Model 1) with the three descriptors Energy\textsubscript{total}, Energy\textsubscript{bond} and Energy\textsubscript{electrostatic}. With the scale factors normalized support vectors \( x_i \), to the support vectors corresponding weights \( \alpha_i \) and bias \( b \) to use in the decision function (Equation 21).

<table>
<thead>
<tr>
<th>alpha</th>
<th>Support vectors</th>
<th>Support vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.06441</td>
<td>-1.42405</td>
<td>0.99804</td>
</tr>
<tr>
<td>0.12121</td>
<td>-0.48305</td>
<td>-1.72367</td>
</tr>
<tr>
<td>-0.40812</td>
<td>-1.46944</td>
<td>0.40042</td>
</tr>
<tr>
<td>1.07143</td>
<td>-0.25043</td>
<td>0.55418</td>
</tr>
<tr>
<td>0.86296</td>
<td>-0.26918</td>
<td>0.48625</td>
</tr>
<tr>
<td>1.07143</td>
<td>-0.24614</td>
<td>0.50335</td>
</tr>
<tr>
<td>-0.93750</td>
<td>0.27028</td>
<td>0.64541</td>
</tr>
<tr>
<td>-0.93750</td>
<td>0.31335</td>
<td>0.66126</td>
</tr>
<tr>
<td>-0.93750</td>
<td>0.29479</td>
<td>0.65699</td>
</tr>
<tr>
<td>0.13433</td>
<td>0.87692</td>
<td>-0.80934</td>
</tr>
<tr>
<td>-0.1051</td>
<td>1.48919</td>
<td>-0.45865</td>
</tr>
</tbody>
</table>

Scale data

| Shift            | 3211.07174 | -3177.37552 | 35560.74587 |
| Scale factor     | 3.8442e-04 | 7.0728e-03  | 3.3131e-04  |
| Bias             |            |            | 1.29766     |

Energy\textsubscript{total} is a measure for the internal potential energy of the molecule and can be seen as an approximation for enthalpy\cite{73}. Energy\textsubscript{bond} describes interactions between directly covalently
bonded atoms and comprises, for example, unfavourable effects due to unnatural bond lengths or distortions of the local geometry. Energies due to electrostatic interactions between charged residues are included in the Energy\textsubscript{electrostatic} descriptor, which confirms the importance of charge for the estimation of expression outcome observed in previous studies.

**Model challenge**

Two sequences that were neither included in the model building nor selection process were used to test the performance of the newly developed final model. The designed sequences, EDP-11[39, 42] and QDP-11[39], were used to trial the potential of the model to distinguish between the two groups 'soluble expression' and 'no soluble expression'. The first sequence, QDP-11 was according to our experience expected to express, as it contains a lot of neutral, hydrophilic Gln residues, which was determined by Bertone et al. to favour expression[69]. In the second sequence, EDP-11, all Gln residues were replaced with negatively charged Glu together with one additional replacement of Ser with Glu on each helix. This leads to a total number of 32 negatively charged residues, which should strongly promote repulsive forces within the molecule and thus lead to a strong destabilization and no expression yield.

![Figure 3-8](image)

*Figure 3-8 The 3-dimensional structure of EDP-11.*

Negatively charged residues are highlighted in yellow.

As can be seen from Figure 3-8 the repulsive interactions are not only due to repulsion within one turn of the helix, but there are as well interhelical repulsive forces promoting the destabilization of the entire molecule; the sequence is therefore not expected to accumulate upon overexpression and due to the lack of a flexible linker. Experimental expression results confirmed these hypotheses as depicted in Figure 3-9. EDP-11 did not show yield accumulation upon overexpression of the target
protein (Figure 3-9a) lane 9) when a negative staining with zinc/imidazole is applied following the Coomassie blue staining[74], whereas for QDP-11, a clear white band at approximately 8 kDa can be identified (Figure 3-9b) lane 9).

Figure 3-9 SDS page gels for a) EDP-11 before induction (lane 2 (T) & 3 (S)) and 2 h (lane 4 (T) & 5 (S)), 5 h (lane 6 (T) & 7 (S)) and 8 h post induction (lane 8 (T) & 9 (S)) with standard ladder (lane 1). b) QDP-11 before induction (lane 2 (T) & 3 (S)) and 2 h (lane 4 (T) & 5 (S)), 5 h (lane 6 (T) and lane 7 (S)) and 8 h post induction (lane 8 (T) & 9 (S)) with standard ladder (lane 1). T stands for total protein, and S for soluble protein.

This is below the theoretically calculated molecular weight of 10.51 KDa and 10.33 kDa for QDP-11 and EDP-11, respectively, possibly because the molecules do not unfold entirely or due to their negative charges disrupting interaction with negatively charged SDS[75]. The sequences were then modelled and simulated and the data prepared as described in methodology.

Table 3-5 Model input data for sequences EDP-11 and QDP-11 as derived from MD simulation. Brackets indicate different starting structures.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Energy\text{total}</th>
<th>Energy\text{bond}</th>
<th>Energy\text{electrostatic}</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDP-11 (1)</td>
<td>$-7.4863\times10^3$</td>
<td>$3.0284\times10^3$</td>
<td>$-4.2970\times10^4$</td>
</tr>
<tr>
<td>EDP-11 (2)</td>
<td>$-7.4795\times10^3$</td>
<td>$3.0317\times10^3$</td>
<td>$-4.2941\times10^4$</td>
</tr>
<tr>
<td>EDP-11 (3)</td>
<td>$-7.4382\times10^3$</td>
<td>$3.0346\times10^3$</td>
<td>$-4.2954\times10^4$</td>
</tr>
<tr>
<td>QDP-11 (1)</td>
<td>$-4.8677\times10^3$</td>
<td>$3.2836\times10^3$</td>
<td>$-3.5021\times10^4$</td>
</tr>
<tr>
<td>QDP-11 (2)</td>
<td>$-4.8467\times10^3$</td>
<td>$3.2657\times10^3$</td>
<td>$-3.5006\times10^4$</td>
</tr>
<tr>
<td>QDP-11 (3)</td>
<td>$-4.8032\times10^3$</td>
<td>$3.2811\times10^3$</td>
<td>$-3.5016\times10^4$</td>
</tr>
</tbody>
</table>

The resulting variables (Table 3-5) were used as input for our new model with the model parameters of Model 1 (Table 3-4) and the theoretical prediction with the simulation data matched the experimental finding, as QDP-11 was predicted to express soluble and EDP-11 was predicted not to. This result shows that the simulation, and therefore the presented model, is sensitive to small
alterations in the amino acid sequence (for similarity between the four-helix bundle primary structures see Table 3-1), which confirms molecular dynamics simulation as a suitable and promising tool to capture these differences and refer them to an altered stability and expression outcome.

3.1.5 Conclusion

This study shows that simulation-derived data in combination with a support vector machine algorithm can be useful to predict recombinant expression of four-helix bundles in *E. coli*. The 3-dimensional structure of the proteins can add important information and leads, in this case study, to a high prediction accuracy. To the best of our knowledge this approach has not been previously used for protein expression prediction models. It has to be noted that the resulting models presented here are expected to be only valid when the same conditions, experimentally and in the simulation, are applied and the availability of the 3-dimensional structure, or at least a similar sequence and structure to use as homology modelling template, is required for the prediction. Due to the confined availability of experimental data in this case study the usefulness of this new approach was tested with a limited number of four-helix bundles. Nevertheless the methodology of using molecular dynamics simulation to predict expression outcome shows promise, and is, to the best of our knowledge, completely novel, and may have high potential and wide application. The models have been successfully tested on four-helix bundles thus far; however the effective inclusion of other groups of proteins may be confirmed in a future study.

3.1.6 Acknowledgments

This work was kindly funded by an Australian Research Council Discovery Grant (DP120103683). The authors gratefully acknowledge the Protein Expression Facility at the University of Queensland, specifically Emelyn Tan and Brad Ryan, for conducting cloning and expression experiments as described in the manuscript. A.S. was financially supported by an UQI Tuition Fee Scholarship and an AIBN RHD Living Scholarship; A.P.J.M. thanks the Queensland Government award of the 2010 Smart State Premier's Fellowship. The authors further gratefully acknowledge the HPC support from the University of Queensland. Sequence EDP-11 peptide was designed by Dr. Annette Dexter from The University of Queensland and patented with international publication number WO 2011/116411 A1. A.P.J.M. designed the sequence variants and the expression testing experiments which were conducted by the Protein Expression Facility. A.S. conducted and designed the simulations, analysed the data, and wrote the manuscript. N.K.C., S.A.O., A.P.J.M., and J.H. contributed to research design and analysis, and manuscript editing and review. All authors have approved this article.
3.2 Validation of predictions of expression experiments using molecular dynamics simulation

3.2.1 Introduction

Thirty-five years ago the only way to obtain a protein was from its natural sources[76]. A lot has changed since then and modern recombinant protein expression techniques are able to produce desired proteins in different host systems like bacteria, plants, mammalian, yeast, fungal and insect cells[77, 78]. Overexpression of recombinant proteins is a widely used technique to gain proteins used for pharmaceuticals, cosmetics, food and research. However, their use can be limited by the expression of large enough amounts of correctly folded, active and soluble product[78]. Not all proteins can be expressed at high solubility, and frequently overexpression results in low or no yields of soluble protein[66]. Despite a lot of effort to solve this problem experimentally[14, 79-82] or to predict the outcome computationally[4, 7, 16-18], an accumulation of the product cannot be guaranteed and new and improved methods are required. In the previous section we reported a completely novel approach to this problem by applying molecular dynamics (MD) simulation in combination with a statistical classifier[83]. This differentiates from existing approaches for the prediction of overexpression outcome, which so far exclusively relied on the primary sequence and physicochemical properties derived therefrom, by taking the three-dimensional (3D) structure into account. This approach directly links the stability and fold behaviour of the molecules to the outcome of recombinant expression experiments.

In this study the new model was validated with five de novo sequences. These sequences were specifically designed to result either in high yield, an accumulation of soluble protein, or not by applying sequence-based features derived from literature which promote or impair solubility upon overexpression. The sequences were cloned and the vectors transformed into *E. coli*, and expression outcome determined via SDS-PAGE. The results were compared to computational predictions assessed with molecular dynamics simulation based on the methodologies of section 3.1. Therefore the amino acid sequences were homology modelled and simulated at 498 K to determine energy values needed for the support vector machine classifier for the prediction. The model was successful to predict if the four-helix bundles would be expressed at high yield, as the molecules which were classified as ‘no expression’ revealed distinctly lower yield compared to the one which were predicted to express. Further the aim of this study was to confirm that a truncated simulation time of 1 ns is enough to obtain the same prediction accuracy as the long simulations of 25 ns from our previous work, but saves computational time and resources, and leads to faster results in future predictions. This suggestion was endorsed with the present results.
3.2.2 Methods

Sequence design
Based on the amino acid sequence of DAMP4 [MD-(PS-MKQLADS-LHQLARQ-VSRLEHA-D)₄] but modified in order to yield in an accumulation of soluble protein when expressed in *E. coli* or not, five new sequences were created. Design rules for sequence-dependent features determined empirically and previously described in literature that either promote or impair overexpression in *E. coli* were followed. Two variants were created to yield in expression, AS1 and AS2. For AS1 one additional negative charge, which is the most frequently described sequential factor promoting expression[7, 15, 16, 67-69, 84], on each of the helices was introduced. For AS2 serine content (GRAVY index -0.8)[85] in the sequence was decreased[7, 17, 67], replacing it with asparagine (GRAVY index -3.5)[85] and thereby increasing hydrophilicity[7, 84, 86]. All these sequence changes were previously described to potentially lead to high expression yield in *E. coli*.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence Features Used</th>
<th>Theoretical Isoelectric Point[87]</th>
<th>Sequence Features Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants created to express</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS1</td>
<td>MD-[PS-MKQLADS-LHQLERQ-VSRLEHA-D]₄</td>
<td>5.85 Negative charge</td>
<td></td>
</tr>
<tr>
<td>AS2</td>
<td>MD-[PS-MKQLADN-LHQLARQ-VNRLEHA-D]₄</td>
<td>6.70 Serine content, hydrophilicity</td>
<td></td>
</tr>
<tr>
<td>Variants created not to express</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS3</td>
<td>MD-[PS-MKQLASS-LHQLARQ-VSRLSHA-D]₄</td>
<td>11.55 Negative charge, serine content</td>
<td></td>
</tr>
<tr>
<td>AS4</td>
<td>MD-[PS-MEQLADS-LEQLAEQ-VERLEHA-D]₄</td>
<td>3.91 Negative charge</td>
<td></td>
</tr>
</tbody>
</table>

Three sequences, AS3, AS4 and AS5, were modified to lead to no accumulation of soluble protein. For AS3, the previously described features were applied reversed: two negative charges on each helix were replaced with serine residues. AS4 contains four additional negative charges on each helix, strongly promoting a destabilization by intramolecular repulsive forces. For AS5, negative charges were replaced with positively charged lysine[66].

Recombinant overexpression in *E. coli*

The genes encoding the sequences from Table 3-6 were synthesized (Geneart AG, Regensburg, Germany) and cloned into the pET-48b(+) expression vector by the Protein Expression Facility (PEF) at the University of Queensland. The plasmids were transformed into BL21(DE3) competent *E. coli* cells via heat shock treatment (42 °C for 45 s). The bacteria were streaked on LB agar plates (10 g L⁻¹ tryptone, 5 g L⁻¹ yeast extract, 10 g L⁻¹ NaCl, 15 g L⁻¹ agar) containing kanamycin sulphate.
(50 μg ml⁻¹, Gibco, Invitrogen, Carlsbad, CA, USA) and incubated at 37 °C for 16-17 h. For the overnight culture a single colony was selected to inoculate 5 ml of 2YT medium (16 g L⁻¹ tryptone, 10 g L⁻¹ yeast extract, 5 g L⁻¹ NaCl) containing kanamycin sulphate (50 μg ml⁻¹). The cultures were incubated at 30 °C overnight on an orbital shaker with 180 rpm. 100 ml of 2YT medium (50 μg ml⁻¹ kanamycine sulphate) were inoculated with the overnight culture to set the starting OD₆₀₀ to 0.001. The main cultures were incubated at 37 °C and 180 rpm until OD₆₀₀ reached 0.5. Overexpression of the target protein was induced with 100 μl 1 M Isopropyl-β-D-thiogalactopyranosid (IPTG) and the cultures were incubated at 180 rpm for approximately 20 h at two expression temperatures (26 °C and 37 °C) until harvest. Cultures with expression temperature 26 °C were cooled down with tap water prior to induction. To assess expression outcome by means of SDS-polyacrylamide gel electrophoresis (PAGE), samples were taken prior to induction and at harvest. ODs of the samples were matched by concentration of the sample prior to induction. For SDS-PAGE (NuPAGE Bis-Tris 4-12% gels (Invitrogen, Carlsbad, CA, USA)) the samples were denatured with heat treatment (5 min, 95 °C) and standard procedures and staining were conducted.

**Material collection and preparation for molecular weight analysis (AS5)**

To analyse the molecular weight of AS5 a main culture of 400 ml was prepared following procedures described above. For harvest the culture was centrifuged (Beckman Avanti® J-20XP (Beckman Coulter, Brea, CA, USA), 5000 rpm, 4 °C, 20 min; Beckman Allegra® X-15R (Beckman Coulter, Brea, CA, USA), 4500 rpm, 4 °C, 20 min) to remove the supernatant. The pellet was resuspended in 40 ml lysis buffer (11.7 g L⁻¹ NaCl, 4.9 g L⁻¹ tris-base, 0.372 g L⁻¹ EDTA, 5% (v/v) glycerol, 0.771 g L⁻¹ DTT, 3.073 g L⁻¹ GSH, pH 8) for a 10 times concentration. The samples were divided into 2 x 20 ml and one half was sonicated for 3 x 45 s (30 W) with a Branson Sonifier 450 mounted with a micro tip (Branson Ultrasonics, Danbury, CT, USA) followed by a centrifugation step (18000 rpm, 4 °C, 20 min) to remove insoluble contaminants. The supernatant was collected for further analysis. The other half of the samples was purified with the DAMP4 heat process[88] by addition of Na₂SO₄ to a final concentration of 0.3 M and heating of the samples to 90 °C for 45 min. Centrifugation (15000 rpm, 4 °C, 5 min) was used to remove precipitated contaminants. The same procedures were conducted for DAMP4 as a control and for comparison to the original molecule.

**Liquid chromatography and mass spectrometry (LC/MS)**

To analyse the mass of AS5 a Waters Quattro Micro API quadrupole mass spectrometer (Waters Corporation, Milford, MA, USA) in positive-ion mode was used. A Kinetex C18 column (Phenomenex, Torrance, CA, USA) was applied with running buffer A (0.01% TFA v/v in Milli-Q
water) and elution with a gradient from 30% to 65% v/v buffer B (90% v/v acetonitrile, 0.01% TFA) with a flow rate of 0.6 mL min\(^{-1}\). The supernatant from the sonicated and heat process samples was injected with a flow rate of 10 \(\mu\)L min\(^{-1}\) and control and data analysis was conducted with the Waters MassLynx software.

**Molecular dynamics simulation**

All simulations were conducted and data analysed as described before[83]. To confirm the applicability of a shorter simulation time for the suggested Model 1 (section 3.1.4), one set of simulations was conducted for a truncated simulation time of 1 ns while maintaining all other simulation parameters.

3.2.3 **Results and discussion**

**Computational prediction results**

The five new variants of DAMP4 (Table 3-6) were either designed to express (AS1, AS2) or not to express (AS3, AS4, AS5). Therefore the DAMP4 sequence was modified to match findings described in literature that define sequence dependent features that either promote or impair soluble expression of target proteins in *E. coli*. Our previously published model[83] as described in section 3.1, which applies a combination of MD simulation and a SVM classifier, was used to predict the expression outcome of these sequences computationally. This model uses three energy descriptors derived from a 25 ns thermal unfolding MD simulation to predict if a certain sequence likely results in an accumulation of soluble protein when expressed in *E. coli* experimentally or not. The simulation data for each of these variants averaged over three runs can be found in Figure 3-10, however each of these energies considered by themselves are not enough to suggest the final expression outcome which is why we used the statistical classifier. AS1 and AS2 were designed to express due to higher stability. This outcome is reflected in the energy plots, where the total energy is compared to the other variants lower, indicating a more stable structure as this can be seen as an approximation for enthalpy. These two variants are as well within the previously defined electrostatic energy slot within which expression is likely (Figure 3-11c)). Compared to that the sequences designed not to express, AS3 and AS5 are outside the range. Their total energy is as well higher than that of AS1 or AS2, indicating a general lower stability in the simulation. This can be attributed in part to the design principles applied to lower expression yield, as well as in the associated loss of salt bridges between lysine and aspartic acid in the first heptad of each helix. AS4 has a very low electrostatic energy, due to the high density of negative charges, designed to be too high to promote expression and rather lead to internal destabilization. These data points combined were in the following used to predict the final experimental expression outcome with the prediction
model. For all the AS-sequences (Table 3-6) the MD model predictions matched the purposes the sequences were created for, which is an accumulation of soluble protein and thus high yield for AS1 and AS2 or not for AS3-AS5, in all cases.

![Figure 3-10 Simulation data as used for prediction model for variants AS1-AS5 averaged over three runs.](image)

This result shows that the MD prediction model is able to distinguish between small alterations in the sequences, and it is able to refer them to a potentially altered expression outcome. Further these changes match findings of previously published studies, as these were the base for the new variants and the same trends as in literature can be seen.

**Confirmation of a shorter simulation time of 1 ns**

As proposed previously (section 3.1.4) when predictions are made with a model containing only energy descriptors, the simulation time can be set much shorter than in our initial simulations, which does not only lead to a faster available result, but as well saves computational resources (cpu*h and data which has to be analysed and stored). This is due to the energy conservation of the system, as after the molecule reaches equilibrium, it only reveals small fluctuations around a certain energy level due to fluctuation in the current system temperature. As this equilibrium state is
reached after approximately 1 ns, we suggested a truncation of simulation length to this time period.
Conducting the simulation with the same conditions and parameter, but for only 1 ns lead for all AS-variants, as expected, to the same prediction results.

Figure 3-11 Comparison of data for the prediction descriptors a) Energy, b) Energy\textsubscript{bond}, and c) Energy\textsubscript{electrostatic} for 1 ns (o) and 25 ns (x).

Red, blue and green correspond to starting structures 1, 2, and 3, respectively.

Figure 3-11 shows that the variations in energy values for 1 ns and 25 ns are small and data from the shorter simulations can be used for future predictions without the loss of prediction accuracy. This is verified with the mean variation coefficients (ratio of standard deviation to mean value), which are with 1.02%, 0.33%, and 0.2% for Energy, Energy\textsubscript{bond}, and Energy\textsubscript{electrostatic}, respectively, extremely low.

**Recombinant expression in *E. coli***

Expression experiments of five new constructs, all variants of DAMP4[35, 89], were conducted and expression outcome determined via SDS-PAGE to confirm if the results predicted with the MD tool (see section Computational prediction results) were correct.
Figure 3-12 shows the expression results of AS1-AS5 and DAMP4 at 26 °C and 37 °C. The gels reveal no big differences between the two expression temperatures. Both sequences designed to express, AS1 and AS2, were expressed at high yield (lanes 2 & 4, respectively), even higher than the original molecule DAMP4 (lane 12). AS3 was expressed, but at a lower yield compared to DAMP4. AS4 was not expressed at all but likely degraded by internal bacterial proteases as no band is visible at the expected molecule size of 11 kDa[87].

Lane 10 shows the expression of a molecule, but at a higher molecular weight (MW) than expected for AS5. The band is located at 15 kDa which is more than 1/3 higher than the calculated MW of 11.2 kDa. The yield of this protein is approximately half of the yield of DAMP4, which, as expected, resulted in a high accumulation of soluble protein (lane 12). AS1 resulted in a very high expression yield. Compared to DAMP4, on each helix Ala on the third heptad was replaced by negatively charged Glu. This mutation introduces additional negative charges, which has been suggested to help the structures in folding due to electrostatic repulsion between the different molecules reducing the number of misfolded, and thus degraded, proteins and the risk for aggregation[71]. By mutating two non-polar amino acids on the surface of different sides of S1 Dihydrofolate reductase (DHFR) to negatively charged amino acids, Dale et al. reached a vastly improved solubility, linked to a 264-fold improvement in activity, when overexpressed in E. coli[90]. Furthermore by increasing the hydrophilicity of the molecule, as Ala has a GRAVY index of 1.8 and Glu of -3.5[85] and this site is located on the protein’s surface, we hypothesized...
that this will increase the solubility of the protein. This could be confirmed with an even higher expression yield than DAMP4, as can be seen in the results in Figure 3-12. For AS2 the same effect could be observed. On each helix two Ser residues (GRAVY index -0.8) were mutated to Asp (-3.5) designed to increase solubility by an increase in hydrophilicity and by decreasing the number of Ser, which both were previously[7, 67] shown by statistical analysis of sequences, which are expressed soluble or as inclusion bodies, to be negatively related to solubility. Compared to DAMP4 the expression yield could be increased by these features (Figure 3-12). AS3 was designed not to express by applying the principles described before. However, a yield approximately half of the yield of DAMP4 can be seen on the SDS-PAGE gel. The reduction in negative charges likely caused an increase in misfolded and aggregated product, but did not lead to the expected outcome of no accumulation of soluble protein. This is probably because overexpression outcome is determined by a mixture of different sequence-dependent and -independent features and the influence of a certain mutation can vary accordingly[15]. AS4 has seven negative charges on each helix. Despite the findings that suggest negative charges promote soluble expression, if the sequence contains too many negative charges, and in particular when the charges are clustered, it can lead to no accumulation of soluble protein[71, 83]. In this case, the negative charges cause repulsion within the helix and within different helices of the same molecule, and thus causing large repulsive forces in folding. As expected this variant lead to no expression yield. For AS5 SDS-PAGE showed the expression of a molecule but at a higher molecular weight than calculated. This might be due to the fact that AS5 is positively charged. It was designed with six positive charges to promote repulsion and, as it was suggested by Price et al., positive charges can cause problems by interactions with negatively charged rRNA[91]. Highly basic molecules can have a lower charge to mass ration due to the high number of positive charges, resulting in a higher apparent molecular weight on an SDS-PAGE gel[92]. However it has to be confirmed that this molecule truly is the target protein. The expression yield of this molecule is notably lower than for DAMP4.

Our MD prediction tool was able to predict AS1, AS2 and AS4 correctly. For AS3 and AS5 it predicted no accumulation of soluble protein, however the results suggest that these sequences are expressed to a certain extend; however their yield is notably lower than the yield for DAMP4 and the SDS-PAGE gel indicates problems when these molecules are overexpressed.

**Identification of AS5**

To identify if the band on the SDS-PAGE gel corresponds to the target protein AS5, liquid chromatography and mass spectrometry (LC/MS) measurements were conducted to determine the molecular weight of the overexpressed molecule. LC/MS is the combination of high-performance liquid chromatography (HPLC), to separate the sample in its different components, and mass
spectrometry (MS) to identify these by their molecular masses[93]. A RP C18 column was used for separation and the samples prepared as described in methods. In case the protein did not withstand the denaturing conditions in the heat process, a sonicated sample was run in parallel.

As can be seen in Figure 3-13, for AS5 no molecular weight could be determined with MS as HPLC was not able to separate the sample and no distinct peaks corresponding to AS5 can be seen. For both, DAMP4 and AS5, peaks at 1.8 min and 31 min can be observed, which therefore correspond to bacterial contaminants and native bacterial proteins and not to the target protein AS5. The first peak likely caused by contaminants, which get washed through the column and do not hydrophobically interact with the resin, and the latter by small peptide contaminants. DAMP4 can be identified in the HPLC peak between 19.5 min and 23 min in the corresponding MS spectrum (see Figure 3-13a)). However, this peak is widely distributed. A possible explanation might be that ions included in the lysis buffer (Cl, Mg, Triton) interact with and bind to DAMP4 and thus cause this broadening of the DAMP4 peak. Possible reasons why AS5 is not showing up in the chromatogram might be that it was denatured in the DAMP4 heat process and thus removed with the precipitated contaminants in solution in the centrifugation step, it did not interact with the resin and got washed trough, or electrostatic interactions were covering hydrophobic interactions in the
column and thus elution of the molecules was not successful. As the contaminant peaks for DAMP4 and AS5 were comparable in height, a washing through is rather unlikely. To confirm the presence of the target protein after heat and sonication treatment a SDS-PAGE gel was run.

Figure 3-14 SDS-PAGE gel of AS5 and DAMP4.

AS5 before induction (lane 1), at harvest (lane 2), after sonication (lane 3) and after the heat process (lane 4), and DAMP4 before induction (lane 5), at harvest (lane 6), after sonication (lane 7) and after the heat process (lane 8); Lane 9 shows the Novex® Sharp pre-stain standard ladder.

The gel shows that cell lysis with sonication leads to a very high number of contaminants in the sample used for LC/MS (lane 3) compared to the heat process (lane 4); however this should not affect MS results as the components should be resolved via HPLC. More likely the target protein interacts electrostatically with the column material and as electrostatic interactions are stronger than hydrophobic interactions, the protein cannot be separated with an RP-HPLC column. Further it can be seen from the gel that AS5 is not denatured during the heat process. The band of AS5 is still visible after the heating (lane 4); however approximately half of the protein is lost in this step. This result indicates that it might be the target protein as it can only withstand the heat process with a stable, four-helix bundle fold and in the case of a host cell protein it would have been denatured and separated in the centrifugation step. However it cannot be excluded that the target protein has a tag due to a failed stop codon which results in this increased molecular weight and another dark band at approximately 9 kDa is also visible after the heat process. An analysis of the cloned DNA-sequence showed that in case of one failed stop codon the molecule would result in a molecular weight of 14.136 kDa, in case of another failure to 14.796 kDa. Another expression at 37°C purified by the DAMP4 heat process with 0.5 M Na₂SO₄ after resuspension in equilibrium buffer (50 mM NaCl, 25 mM Na₂HPO₄, Figure 3-15, lane 1) revealed the previously observed band
at 9 kDa more clearly. This band has approximately the same intensity as the band above expected to be the target but is below the calculated molecular weight for AS5.

![Figure 3-15 SDS PAGE gel of AS5 after heat process.](image)

 Supernatant after centrifugation (lane 1) and precipitate (lane 2), and the standard ladder (lane 3).

The band at 14.5 kDa is as well visible in the precipitate (lane 2). This might indicate that a stop codon failed and produced a version of AS5 which has a small tag of 28 amino acids. This variant might still have some kind of four-helix bundle fold which enables it to withstand the heat process, but is not as stable as the original variant, thus causing this large band in the precipitate. The 9 kDa band could be the correctly expressed variant as this band does not appear in the precipitate, indicating a stable and correct four-helix bundle fold. This would mean that the stop codon failed in some cases and in others not.

### 3.2.4 Conclusion

In conclusion, this study showed that a truncated simulation time of only 1 ns is enough to reach the same prediction accuracy as the long simulation of 25 ns, saving time and resources. Further, our previously reported MD prediction tool for recombinant expression in *E. coli* was able to predict, if sequences would be expressed at high yield or not. However, two of the sequences were expressed despite the prediction of “no expression” but at a distinctly lower yield when compared to DAMP4. One of these two sequences showed up at a significantly higher molecular weight than expected on a SDS-PAGE gel. LC/MS analysis together with heat stability results could not unambiguously verify the target molecule’s identity, which moved slower through the gel due to a lowered charge to mass ratio because of the high number of positively charged residues. Despite this fact, this research shows that the use of molecular dynamics simulation and higher structural orders in
combination with a statistical approach has the potential for widespread use in recombinant biotechnology in predicting if a protein will be expressed at a high yield or not.
3.3 References


Chapter 4 Computational study of elements of stability of a four-helix bundle protein biosurfactant

This chapter consists of the journal article published as:

4.1 Abstract

Biosurfactants are surface-active molecules produced principally by microorganisms. They are a sustainable alternative to chemically-synthesized surfactants, having the advantages of being non-toxic, highly functional, eco-friendly and biodegradable. However they are currently only used in a few industrial products due to costs associated with production and purification, which exceed those for commodity chemical surfactants.

DAMP4, a member of a four-helix bundle biosurfactant protein family, can be produced in soluble form and at high yield in *Escherichia coli*, and can be recovered using a facile thermal phase-separation approach. As such, it encompasses an interesting synergy of biomolecular and chemical engineering with prospects for low-cost production even for industrial sectors. DAMP4 is highly functional, and due to its extraordinary thermal stability it can be purified in a simple two-step process, in which the combination of high temperature and salt leads to denaturation of all contaminants, whereas DAMP4 stays stable in solution and can be recovered by filtration. This study aimed to characterize and understand the fundamental drivers of DAMP4 stability to guide further process and surfactant design studies.

The complementary use of experiments and molecular dynamics simulation revealed a broad pH and temperature tolerance for DAMP4, with a melting point of 122.4 °C, suggesting the hydrophobic core as the major contributor to thermal stability. Simulation of systematically created *in silico* variants of DAMP4 showed an influence of number and location of hydrophilic mutations in the hydrophobic core on stability, demonstrating a tolerance of up to three mutations before a strong loss in stability occurred. The results suggest a consideration of a balance of stability, functionality and kinetics for new designs according to their application, aiming for maximal functionality but at adequate stability to allow for cost-efficient production using thermal phase separation approaches.
4.2 Introduction

Biosurfactants, surfactant molecules produced by a wide range of different microorganisms[1], have advanced as a current trend in industry, as they are a sustainable alternative to petroleum-based chemically synthesized surfactants[2-5]. As surfactants are applied worldwide in innumerable different products and industrial sectors, ranging from food, personal care and hygiene, cosmetics, cleaning and washing, and paints to the pharmaceutical industry, the advantages of biosurfactants in those products of being biodegradable, non-toxic, highly surface active and functional, renewable and ecofriendly are close at hand[3, 4]. Despite these advantages and the current movement toward industrial sustainability, biosurfactants are so far only used in a few commercial products[5-7]. This is due to the cost of production and recovery or purification, which frequently includes the use of expensive chromatographic steps and makes them at present not competitive with cheaper chemical surfactants. Further, difficulties with expression yield, and the design of molecules for a specific function, need to be overcome[6].

To address these problems, Middelberg et al. recently designed a four-helix bundle biosurfactant protein, DAMP4 (MD-[PS-MKQLADS-LHQLARQ-VSRLEHA-D]4), a highly functional surface-active molecule with unique switchable foaming behaviour[8, 9]. Recombinant expression of DAMP4 in the microbial cell factory Escherichia coli (E. coli) results in a stable product at high yield[8, 10]. DAMP4 is expressed intracellularly but can be purified cheaply in a simple two-step process with thermal treatment followed by solid-liquid separation. Costly chromatography or mechanical cell disruption is not necessary, which allows this process to be used even in low-cost industrial sectors[10]. In detail, following expression, Na₂SO₄ is added to the culture broth to a concentration of 0.5 M and the entire medium heated to 90 °C for 30 minutes. The kosmotropic salt in combination with the high temperature causes cell lysis and denaturation of host cell proteins, while DAMP4 is salted in and stays stable in solution. Two filtration steps (1.2 μm and 0.2 μm) remove precipitated impurities and allow the recovery of industrial grade target protein[10]. This process takes advantage of the above-average thermostability of DAMP4 in comparison to the majority of the bacterial protein contaminants. Understanding the structural fundamentals of stability of this four-helix bundle protein can help facilitate the design of new molecules to ensure the applicability of the low-cost purification process while maintaining functionality.

Four-helix bundles have been studied widely, as a naturally-occurring protein fold[11-14] as well as a designed structural motif[8, 15-18]. The stability of four-helix bundles can vary and is influenced by different factors. Several studies demonstrate they can have extremely high thermal stability. Schafmeister et al. (1997) reported the design of a 108 amino acid long four-helix bundle (DHP₄) with native-like structure and a melting point similar to that of proteins from thermophilic
microorganisms of 122 °C. They suggested the design with a reduced set of alternating hydrophilic and hydrophobic amino acids influenced the properties of the bundle[18]. Hecht et al. (2002) showed with their 2\textsuperscript{nd} generation combinatorial library of four-helix bundles, in which they increased the molecule length from 74 to 102 amino acids and followed their binary pattern of hydrophobic and hydrophilic residues, that it is possible to design stable and functional proteins that have a native-like structure without specifically designing the side chains. Compared to the original library, the proteins revealed considerably more stability and ordered side chains, and they were able to design a second-generation protein with a melting point of ~100 °C as shown by DSC[19]. Munson et al. (1994) studied the packing of the hydrophobic core, and reported the melting point of the widely used model four-helix bundle protein Rop wild type to be 64 °C[20]. Among suggested mechanisms influencing the stability of the four-helix bundle fold are the hydrophobic packing[21], loop length[12], electrostatic interactions including hydrogen bonding, ion pair (salt bridges) and dipole interactions[22], sequence length and polarity[21].

DAMP4 consists of four identical helices, which are arranged to build an antiparallel four-helix bundle with high helical content (see Figure 4-1). DAMP4 is 98 amino acids long, has a molecular weight of 11.1 kDa and a theoretical pI of 6.7. Its helices are derived from the carboxyl-terminal oligomerization domain of the lac repressor protein (Lac21), which self-assembles to form tetramers[23], connected with an Asp-Pro-Ser linker, Asp to promote electrostatic repulsion between adjacent helices, Proline to ensure helix termination due to its low helix propensity, and Ser for flexibility[8]. Due to energetic reasons the four helices will arrange to a bundle to prevent water penetration of the hydrophobic residues based also on known behaviour of heptad peptides containing hydrophobic residues in the a and d positions of the heptad. Considering behaviour, structure and sequence, we suggested three mechanisms, which might be important for DAMP4’s stability: (i) the pronounced hydrophobic core; (ii) ion pair interactions, and; (iii) a tightly-held hydration shell. Each of DAMP4’s helices consists of three heptads, in which every 1\textsuperscript{st} and 4\textsuperscript{th} amino acid is hydrophobic. To minimize the solvent exposure in agreement with the hydrophobic effect, these are oriented towards the inside of the bundle and prevent any solvent penetration into the core. Charged residues on the outer surface of the molecule interact electrostatically with each other and thus drive folding and stability. In DAMP4, positively-charged lysine (at positions 6, 30, 54 and 78) is interacting with negatively charged aspartic acid (at positions 10, 34, 58 and 82), giving four salt bridges, one on each helix. Previous results[8] suggested the importance of the water structure around DAMP4 for its functionality and therefore stability, which might be promoted by positively charged arginine on the solvent exposed surface creating a stabilizing hydrogen-bonded water network[24, 25]. Arginine is with five possible hydrogen-bond donor sites
unique[26, 27] and strongly interacts with water molecules, whose mobility is lowered around the guanidinium group[24].

Figure 4-1 Fold structure of the biosurfactant protein DAMP4.
The hydrophobic core is highlighted in yellow, residues forming salt bridges in green and arginine residues as strong contributors to the hydration shell in red. a) Ribbon, b) ball representations.

This study aims to understand the stability and contribution of each of these mechanisms by a combination of experiment and molecular dynamics (MD) simulation. The experiments show a high thermal stability at all pHs with a melting point of 122 °C, but average stability against chemical denaturation. In agreement, MD simulations reveal the hydrophobic core as the most important structural feature for DAMP4 stability, and further show the influence of number and location of hydrophilic mutations in the core on the overall stability of the molecule.

4.3 Material and methods
DAMP4 was expressed[8] and purified[10] in E. coli as described before.

4.4 Differential scanning calorimetry (DSC)
DSC measurements were used to assess the thermal stability of DAMP4. Measurements were performed with a MicroCal VP-DSC Microcalorimeter (GE Healthcare Life Sciences, Fairfield, CT, USA) with the VPViewer 2000 DSC software. To determine the melting point ($T_m$) the sample (1 mg ml$^{-1}$ DAMP4 in Milli-Q water) was treated with a ThermoVac degassing station at 25 °C for 5 min to remove air bubbles. In the DSC measurement a scan rate of 90 °C h$^{-1}$ was applied for a
temperature ramp from 30 °C to 130 °C, and samples analysed with the MicroCal-enabled Origin data analysis package.

4.4.1 Circular dichroism (CD)

CD spectra were determined with a J-810 Spectropolarimeter (Jasco, Easton, MD, USA) to analyse secondary structure of DAMP4 (0.25 mg ml$^{-1}$ in 1 mM HEPES buffer, 200 μM EDTA) at 20 °C and 90 °C at a pH range from 3.2 to 10.3. To determine stability against chemical denaturants, CD signals at a wavelength of 222 nm as a measure for helix content were recorded. Samples with 1 mg ml$^{-1}$ DAMP4 were prepared by ratio-based mixing of Milli-Q water and 8 M GdmCl-solution to reach final concentrations of 0 M, 0.5 M, 1 M, 2 M, 3 M, 3.2 M, 3.5 M, 3.8 M, 4 M, and 5 M GdmCl. The signal at 222 nm was measured for 300 s at 20 °C and the average for this time interval calculated for each of the GdmCl-concentrations.

4.4.2 Homology modelling (HM)

All homology modelling experiments and simulations were conducted in the MD software package YASARA structure version 12.6.28 (YASARA Biosciences GmbH, Vienna, Austria)[28-30]. As no homologous sequences using PSI-BLAST[31] for DAMP4 were found in the protein data bank (PDB) a single helix based on the aligned sequence of template 3EDC[32] (sequence alignment in figure S1 in supplementary material) was modelled first. Three copies were created and the four helices were connected manually. Therefore peptide bonds were added manually to the arranged helices, energy minimized and the side chain rotamers were refined in a following 500 ps MD simulation with the YASARA2 force field[33]. The model quality was validated with YASARA’s internal structure validation tool[34] and the external ProSA (protein structure analysis) program[35, 36]. YASARA’s structure validation tool is based on a library of gold standard reference proteins normalized in knowledge-based potentials. It compares the model to the average of these references and calculates the deviation from that standard expressed in a Z-score. This Z-score is 2.732, therefore the DAMP4 model is considered optimal. The ProSA overall model quality Z-score was -6.69, which is in the range typical for native molecules of the same size. The knowledge-based energies are below zero showing no problematic parts. A Ramachandran plot to assess the geometrical validation around the C$_\alpha$-atoms can be found in the supplementary material[37]. In silico variants were created with homology modelling with standard parameters by providing the DAMP4 structure as a template. This was followed by a refinement simulation to find the structure with the lowest energy.
4.4.3 Molecular dynamics (MD) simulation

Simulations were conducted in a cubic simulation box with an extension of 10 Å and 20 Å around all atoms for the starting structure generation and production simulation, respectively. Periodic boundary conditions, Particle Mesh Ewald for long-range electrostatics[38], the Amber03 force field[39] and a cutoff for van der Waals interactions of 7.86 Å were applied. To increase computational speed the equations of motion were integrated with a multiple timestep algorithm (impulse method[40] or VERLET-I[41]), where intramolecular forces were calculated every 1.25 fs, whereas non-bonded Van der Waals and electrostatic interactions were updated only every 2nd step and then added with a scaling factor of two. A comparison to other multiple timestep algorithms[40] revealed the most stable trajectories for this approach under the premise that non-bonded interactions are calculated frequently (e.g. every 2nd step as in this case). Validation in YASARA was performed on the dihydrofolate reductase molecule by calculating the resulting energy shift. This energy shift was reported 0.02 per degree of freedom (K_B T/ns) at 298 K, which is comparable to drifts reported for other MD programs[42]. Further validation was conducted by simulating 25 protein crystal structures (like those described by Krieger et al.[28]) and no influence on the root mean square deviation (rmsd) was observed compared to the simulations without multiple timesteps. The temperature was kept constant by rescaling atom velocities using a Berendsen thermostat and the pressure was set to 1 bar and controlled by the time averaged solvent density and adjusted by resizing the cell isotropically to the target solvent density set according to the experimental value[43]. The simulations were run in triplicates. To generate the starting conformations, the homology models were simulated at 298 K and pH 7.4 in a simulation box filled with TIP3P water molecules in a neutralization experiment[44]. Trajectories as starting structures were saved at 3 ns, 4 ns and 5 ns. For the production simulation these were energy minimized, and simulated at 423 K for 100 ns, to allow for a faster unfolding at lower computational expense[45]. The first 100 ps of the simulation were considered as equilibration and discarded from data analysis and trajectories saved every 5 ps for study purposes.

Simulation data was analysed using Matlab Version R2012a (Mathworks, Natick, MA, USA). To analyse and reduce the data set a smoothing moving average function was applied. For DAMP4 and each variant all data points of all three simulations were averaged in 50 ps simulation time intervals, in which two snapshots overlapped with the previous interval, respectively. Molecular graphics were created with YASARA (www.yasara.org) and POVRay (www.povray.org).
4.5 Results and discussion

4.5.1 Experimental characterization of DAMP4 stability

Thermal stability of DAMP4

Thermal stability was characterized using differential scanning calorimetry (DSC) to obtain the melting point ($T_m$) and to investigate tertiary structure changes with increasing temperature. The melting temperature was determined in Milli-Q water, eliminating the influence of salt ions on stability. The results are depicted in Figure 4-2. The endothermic peak in heat capacity occurs at approximately 122.4 °C and corresponds to the protein’s transition point.

![Differential scanning calorimetry showing a melting point above 120 °C.](image)

Figure 4-2 Differential scanning calorimetry showing a melting point above 120 °C.

The melting point lies above the average of proteins and is in the region of proteins in thermophilic organisms[46]. It is a single sharp peak indicating a cooperative process, as unfolding and denaturation occur simultaneously.

![Circular dichroism spectra of DAMP4 at a) 20 °C and b) 90 °C.](image)

Figure 4-3 Circular dichroism spectra of DAMP4 at a) 20 °C and b) 90 °C.

The spectra show DAMP4 is stable over the entire pH range even at elevated temperatures.
A second scan of the same sample showed no distinct peak similar to the first one in the course of temperature progression, which indicates that thermal unfolding is irreversible likely due to the formation of insoluble aggregates after denaturation. DAMP4 has a CD spectrum characteristic for helical proteins, in agreement with its designed four-helix bundle structure in bulk (Figure 4-3)[47]. This helicity is preserved over the entire pH range at 20 °C, and even at a high temperature of 90 °C, with exception of pH 3.2, where some loss of helical content occurs. It has been previously reported a low pH causes a lowering of the thermal denaturation temperature[48] and partial unfolding occurs likely due to a combination of acid denaturation due to the high net positive charge of the molecule and the high temperature. This finding of weaker helicity at lower pH may indicate a significance of charges for the thermal stability of DAMP4.

**Stability against chemical denaturants**

To test DAMP4’s stability against chemical denaturation a number of CD spectra with varying guanidinium chloride (GdmCl)-concentration were measured.

![Figure 4-4 Stability of DAMP4 at different concentrations of GdmCl at 20 °C.](image)

Stability against chemical denaturants is typical of that for proteins.

The sigmoidal transition curve shows unfolding of DAMP4 in a simple two-state folded - unfolded process. Transition of DAMP4 occurs between 3 M and 4 M GdmCl, whereby 50% of the native structure is lost at 3.23 M elucidating a moderate stability against chemical denaturants.
4.5.2 Investigation of DAMP4 stability with molecular dynamics simulation

Sequence design of stability variants for MD simulation

To test which of the previously-suggested stabilization mechanisms applies for the high thermal stability observed experimentally, a number of different stability variants of DAMP4 were created. To account for the influences of salt bridges, the hydrophobic core and water structure, variants in which each of these mechanisms was eliminated completely while maintaining the other effects to the greatest possible extent were created in silico. The variants are all labelled with mechanism (hydrophobic core (hc), salt bridge (sb), water shell (ws)), number of mutations in comparison to DAMP4, and the number of the variant with a capital letter when several variants with the same number of mutations exist. The sequences of the variants can be found in Table 4-1.

Table 4-1 Sequence design for DAMP4 in silico stability variants for MD simulation. In these variants one of the suggested stability mechanisms (hydrophobic core, salt bridges, water shell) is eliminated or weakened. Red marks site of mutations in comparison to DAMP4.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence features used</th>
<th>Sequence</th>
<th>Suggested stability mechanism removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAMP4</td>
<td></td>
<td>MD-[PS-MKQLADS-LHQLARQ-VSRLHEA-D]₄</td>
<td></td>
</tr>
<tr>
<td>Hydrophobic variant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hc8</td>
<td></td>
<td>MD-[PS-MKQNADS-LHQNARQ-VSRLHEA-D]₄</td>
<td>Hydrophobicity</td>
</tr>
<tr>
<td>Salt bridge variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sb4A</td>
<td></td>
<td>MD-[PS-MKQLAKS-LHQLARQ-VSRLHEA-D]₄</td>
<td>Repulsion between positive charges</td>
</tr>
<tr>
<td>sb4B</td>
<td></td>
<td>MD-[PS-MDQLADS-LHQLARQ-VSRLHEA-D]₄</td>
<td>Repulsion between negative charges</td>
</tr>
<tr>
<td>sb4C</td>
<td></td>
<td>MD-[PS-MKQLAQS-LHQLARQ-VSRLHEA-D]₄</td>
<td>Neutral amino acid inserted</td>
</tr>
<tr>
<td>Water shell variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ws8A</td>
<td></td>
<td>MD-[PS-MKQLADS-LHQLAKQ-VSKLEA-D]₄</td>
<td>Arg has five proton donor groups, Lys only three</td>
</tr>
<tr>
<td>ws8B</td>
<td></td>
<td>MD-[PS-MKQLADS-LHQALAAQ-VSALEHA-D]₄</td>
<td>Ala has a high helix propensity but does not interact with water, simulated at pH 5 to maintain net charge</td>
</tr>
</tbody>
</table>

To disrupt the hydrophobic core, two hydrophobic leucines (GRAVY index[49] 3.8) on each helix were replaced with hydrophilic asparagine (GRAVY index -3.5), which has a similar side chain geometry and reduces the risk for destabilization due to steric or packing reasons (variant hc8). To break salt bridges three different variants were designed: two, where the replacement with the same charge, lysine with aspartic acid (variant sb4A) in the first heptad of each helix and vice versa (variant sb4B) causes a destabilization due to charge repulsion, and one with the effect neutralized by insertion of uncharged glutamine for aspartic acid (variant sb4C). Glutamine was selected as the side chains of glutamine and aspartic acid have the same capability to build H-bonds to water and the amino acids have the same GRAVY indices. To cause a distortion in the hydration network around the molecule, two arginine residues on each helix, which have five hydrogen-bond donor groups, were replaced with lysines, which maintains the charge structure, but the interactions with
water molecules are less (variant ws8A). Biswas et al. described that Lys→Arg substitutions can lead to increased thermostability, likely due to an improvement in electrostatic interactions and hydrogen-bonding network[25]. A second water shell variant contains alanine residues in the same location, as alanine is known to have a high helix propensity[50] but does not interact with water (variant ws8B). To maintain the net charge this variant was simulated at pH 5.

**Simulation results for stability variants**

A simulation at a high temperature, well above the experimentally measured melting point of DAMP4, was run to investigate the differences in stability of the variants compared to DAMP4. To elucidate the progress of destabilization and unfolding, the content of α-helices and the rmsd of the Cα-atoms were analysed (Figure 4-5).

![Simulation results of DAMP4 stability variants.](image)

**Figure 4-5 Simulation results of DAMP4 stability variants.**

a) Helical content and b) rmsd Cα. These results are the moving average values in 50 ps intervals of all three starting structures.

87
At the beginning of the simulation the $\alpha$-helical content is around 68% for DAMP4 and its variants, with exception of sb4B, with 65% lower compared to the others likely due to some loss of secondary structure by repulsion of similar charged residues, and ws8B, with 71% higher consistent with the high helix propensity of alanine. After a similar course of $\alpha$-helical content for all structures up to around 50 ns, the differences emerge towards the end of the simulation. DAMP4 has the highest helical content, around 72%. Destabilization occurs during the 100 ns simulation time mainly around the turn regions, but the structure stays mainly intact with minor loss of tertiary structure. The same effects can be seen for the salt bridge and water shell variants, for which only minor loss in helical content occurs. Thereby variant sb4B has the second lowest helical content, with average 5% difference to the other two salt bridge variants. This observation can be explained by the position of the mutation at the beginning of each helix and the lower helix propensity of aspartic acid (0.69) compared to lysine (0.26)[50]. As seen from the simulation conformations, loss of secondary structure starts from the turn regions and open beginning or end of the strands, which makes this variant particularly susceptible for lower helical content. However, hc8 clearly reveals the greatest influence of the temperature on the structure, with a drop in secondary structure starting after approximately 60 ns to 50-55%. The same trends can be seen considering the rmsd $C_\alpha$. The salt bridge and water shell variants behave within the general fluctuations very similar to DAMP4, especially from 60 ns to 100 ns, whereas hc8 noticeably sets off with rmsd values intermittent double that of DAMP4, in particular in this time frame. DAMP4 has a rather stable rmsd, whereas the other variants are more subjected to structural fluctuations, indicating the lower stability of the variants compared to DAMP4. These results are also reflected in the variance, relative variance and mean rmsd values which can be found in the supplementary material, where variant hc8 has distinctively higher values. Conclusively, these results clearly show the hydrophobic core as most important factor contributing to DAMP4’s stability, whereas salt bridges and the hydration shell contribute in only a minor way to the overall stability.

Sequence design of hydrophobic stability variants for MD simulation

To investigate the previous findings further and to study the influence of number and position of hydrophilic mutations in the core, another set of DAMP4 variants was designed. 1 to 8 hydrophobic leucine residues were replaced systematically with hydrophilic asparagine residues (Table 4-2).

Hc1 contains a single hydrophilic point mutation on the first helix. Hc2A has two mutations on helix 1, one in the first heptad and one in the second. To see the difference of mutations on different helices, hc2B was designed with one mutation on helix 1 and one mutation on helix 2. Hc3 has three mutations, one on each of the first three helices. Hc4A and hc4B contain four mutations each;
in the first case these mutations are segregated within the core, and in the latter they are clustered.

Figure 4-7 shows a schematic illustration of the allocation of the mutations.

Table 4-2 Sequence design for hydrophobic variants with 1 to 8 Leu to hydrophilic Asp mutations in the hydrophobic core. Red marks site of mutations in comparison to DAMP4.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAMP4</td>
<td>MD- [PS-MKQLADS-LHQLARQ-VSRLEHA-D]₄</td>
<td>-</td>
</tr>
<tr>
<td><strong>Hydrophobic variants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hc1</td>
<td>MD- (PS-MKQNADS-LHQLARQ-VSRLEHA-D [PS-MKQLADS-LHQLARQ-VSRLEHA-D]₃)</td>
<td>Single mutation on the first helix</td>
</tr>
<tr>
<td>hc2A</td>
<td>MD- (PS-MKQNADS-LHQNARQ-VSRLEHA-D [PS-MKQLADS-LHQLARQ-VSRLEHA-D]₃)</td>
<td>Two mutations on the first helix</td>
</tr>
<tr>
<td>hc2B</td>
<td>MD- (PS-MKQNADS-LHQLARQ-VSRLEHA-D PS-MKQLADS-LHQNARQ-VSRLEHA-D [PS-MKQLADS-LHQLARQ-VSRLEHA-D]₂)</td>
<td>Two mutations, one on helix 1, one on helix 2</td>
</tr>
<tr>
<td>hc3</td>
<td>MD- ([PS-MKQNADS-LHQLARQ-VSRLEHA-D]₃ PS-MKQLADS-LHQLARQ-VSRLEHA-D)</td>
<td>Three mutations, one on helices 1 to 3</td>
</tr>
<tr>
<td>hc4A</td>
<td>MD- ([PS-MKQNADS-LHQLARQ-VSRLEHA-D]₄ PS-MKQLADS-LHQNARQ-VSRLEHA-D)</td>
<td>Four mutations, one on each helix, segregated</td>
</tr>
<tr>
<td>hc4B</td>
<td>MD- (PS-MKQLADS-LHQNARQ-VSRLEHA-D PS-MKQNADS-LHQLARQ-VSRLEHA-D PS-MKQLADS-LHQNARQ-VSRLEHA-D PS-MKQLADS-LHQLARQ-VSRLEHA-D)</td>
<td>Four mutations, one on each helix, clustered</td>
</tr>
<tr>
<td>hc8</td>
<td>MD- [PS-MKQNADS-LHQNARQ-VSRLEHA-D]₄</td>
<td>Eight mutations, two on each helix</td>
</tr>
</tbody>
</table>

**Simulation results for hydrophobic variants**

The hydrophobic variants were simulated under the same conditions as the previous stability variants, and the results for α-helix content and rmsd Cα are displayed in Figure 4-6. The simulations reveal that variants hc1, hc2A and hc2B lose only minor helical content, and their rmsd Cα values are as well in the same region as DAMP4. This shows, their structure experienced some distortions and movement of the helices and is marginally affected in their tertiary structure, but the secondary structure remains mainly unaffected in the evaluated time span. It has to be noted, in comparison to salt bridge and water shell variants, rmsd values are comparable or even higher for the hydrophobic variants with 1 to 3 hydrophilic replacements, despite having a lower number of mutations (see mean rmsd values in supplementary material). This shows that the effect of mutations in the hydrophobic core is stronger than mutations disrupting salt bridges or altering the water shell. Variants hc4A and hc4B are more affected by the high temperature and have distinctly lower helical content than DAMP4, whereby hc4B reaches the same extent as hc8, whose core is completely disrupted. Rmsd Cα values are approximately 5 Å and 6 Å after 100 ns simulation for hc4A and hc4B, respectively, lower than for hc8 with 7 Å. This deviation shows that the effects for hc8 are locally (secondary structure) and globally (tertiary structure) whereas for hc4A and hc4B the local effects are stronger.
Figure 4-6 Simulation results of DAMP4 and hydrophobic core variants.

a) Helical content, and b) rmsd Cα. These results are the moving average values in 50 ps intervals of all three starting structures. They show strong influence on stability when three or more hydrophobic residues are mutated.

To further study the influence of the mutations on the destabilization pathway and the structure, the final simulation conformations were analysed (Figure 4-7). To better characterize these changes, Figure 4-8 additionally contains the top views, distances between the four helices at the bottom (Cα 5, 46, 53, 94) and the top (Cα 22, 29, 70, 77) as a numerical way to capture tertiary movement in the course of the simulation, and the change in solvent accessible surface area (SASA) compared to DAMP4 for selected variants after 100 ns. DAMP4 loses some secondary structure in the proceeding simulation, and the structure becomes disordered mainly in the turn regions and at the beginning and end of the strands. The four-helix bundle fold is not affected and only little distortion occurs. In the final structure of hc1 helix 1 clearly stands out by a shift versus the remaining helices.
Helix 1 contains the single hydrophilic mutation, which is likely the trigger for this movement and detachment. The extent of this effect is also shown in the top view of hc1 (Figure 4-8) and the top distances (Figure 4-8c)) for helix 1-helix 3, and helix 1-helix 4, increase. Defined secondary structure is lost around the turn regions, but to a low extent. Exposure of residues which are still buried in DAMP4 occurs on helix 4 at the bottom part due to the movement of helix 1, but the effect is minor. Hc2A has, compared to hc1, an additional mutation on the second heptad of helix 1. Considering the three final conformations of this variant, this clearly enhances the destabilization and the influence on the tertiary four-helix bundle fold structure is stronger. Helix 1 shows the same shift as observed for hc1 but due to the stronger nature of this shift it influences helix 2, which gets displaced. The top view of hc2A (Figure 4-8a)) shows the displacement and influence on the four helices and the loss of their ideally parallel arrangement. The lower red coloured residue on helix 1 corresponds to the mutated Asn 8 and the middle red coloured residue on the same helix to Asn 15 (Figure 4-8d)), showing a twist of helix 1 exposing these residues to the surrounding water which is not observed for DAMP4. The same shift and extent of distortion is seen in hc2B, with the movement of helix 1 influencing the other three helices, whereas the loss of defined secondary structure is low. With four mutations, one on each helix segregated (hc4a), the effect and the structure become more spherical. The destabilization is stronger on the less protected ‘outer’ helices 1 and 4, and more loss of helical content occurs. The distance plots show how the structure becomes more flexible after around 50 ns. SASA data shows a twist in helix 1 whereby normally buried hydrophobic residues become exposed, and in turn the relative hydrophilic residues are then buried on the opposite site of the helix. Due to the complete loss of helical structure on the lower part of
helix 4, this part becomes completely exposed to the surrounding water. In variant hc4B four mutations, one on each helix, are clustered at the upper part of the four-helix bundle (see Figure 4-7). These mutations have a strong global effect and the four-helix bundle fold is mostly lost. The structure reveals a bursting from the top, in which region the helical structure is lost, showing the local and global effects.

Figure 4-8 a) Top view, designated \( C_{\alpha} \) distances between the four helices at the bottom b) and top c) of four-helix bundles and d) difference in solvent accessible surface area compared to DAMP4 for selected hydrophobic variants.

The distance plot of the top region reflects this trend, with strong fluctuations in the distances, showing the flexibility of the structure. In hc8, the variant with eight mutations, two on each helix, the effect on the tertiary and secondary structure is obvious. The four helices completely lose their parallelism, as seen in the top view in Figure 4-8a), and defined fold. Starting from the turn regions, major parts of helical content are lost. Strong fluctuations in the distance plots in upper and lower part of the bundle (Figure 4-8b) and c)) suggest the high flexibility and instability of this variant. These results demonstrated the strong impact on stability of DAMP4 when three or more hydrophobic core residues are mutated. Investigation of the position of the mutations revealed an influence on the unfolding pathway, and that the effect is stronger when the hydrophilic residues are clustered compared to segregation within the core.
4.5.3 *Further discussion*

In this study we presented a four-helix bundle biosurfactant protein, DAMP4, with extraordinary stability within a broad pH range and even at temperatures above 90 °C.

A wide pH and temperature tolerance is important for industrial processes and applications, and stability in harsh process conditions may be required particularly in downstream processing[51]. Among advantages associated with elevated process temperatures are higher solubility, lower viscosity, higher diffusion rates and a speed increase of kinetic reactions[51]. Further, a highly stable protein is less susceptible to proteolytic degradation, and thus a higher expression yield in a heterologous expression host system is to be expected[15]. In the case of DAMP4, recombinant expression in *E. coli* results in high yield and good solubility, and DAMP4’s stability against high temperatures can be used for a simple and cheap purification strategy[10]. These characteristics are of particular interest considering the current demand for low-cost products which are industrially sustainable, eco-friendly, and renewable. Effectiveness at wide pH and temperature range is an advantage compared to chemical surfactants[5].

In a combination of experiments and MD simulation, this study revealed the hydrophobic core as the most important structural feature for the high thermal stability of the four-helix bundle biosurfactant protein. The strong hydrophobic interactions in the core explain why DAMP4 has above-average thermostability, but is only moderately stable against chemical denaturation with GdmCl. Several studies suggest a weakening of hydrophobic interactions by GdmCl[52, 53] and an interaction of the guanidinium molecules with hydrophobic residues in the protein core and thus a shift in equilibrium towards the unfolded state[53-55]. In contrast, kosmotropic salts, for example Na₂SO₄ as used in the thermal process, are known to strengthen hydrophobic interactions, as the energy to solvate hydrophobic residues in the salt solution is even more unfavourable[56]. This results in a decrease in solubility[56], as reinforced hydrophobic interactions on protein surfaces lead to aggregation, as might be seen for bacterial contaminant proteins in the DAMP4 purification process. In contrast, DAMP4 only contains hydrophobic residues in the core, thus Na₂SO₄ stabilizes DAMP4, but leads to a faster denaturation of the bacterial host cell proteins at the elevated temperature. The thermal process is carried out at neutral pH where DAMP4 charges are neutralized and equally distributed in the bundle. However, CD results showed that DAMP4 is stable over the entire pH range at 20 °C, even at extremely low pH where strong electrostatic interactions occur and the molecule is positively charged. It is known that repulsive charges strongly oppose folding[56], confirming the importance of hydrophobic over electrostatic interactions for DAMP4 stability. From previous studies it is known that DAMP4 folds into a four-helix bundle structure in bulk, whereas DAMP4’s monomer DAMP1 (PS-MKQLADS-LHQLARQ-VSRLEHA-D) does not.
adopt a defined secondary fold and remains unstructured[57]. Whereas electrostatic attraction makes only a minor contribution to the folding into three-dimensional structures of proteins, the main driving forces are polar, van der Waals and hydrophobic interactions[56]. In DAMP4 the hydrophobic interactions are the major contributors to the formation of secondary and tertiary structure through the gain in free energy upon formation of α-helices with a hydrophobic and a hydrophilic plane, and the burial of the water-averse residues in the protein core by folding of the four helices into a bundle. The hydrophobic and hydrophilic planes, built by α-helix formation, provide DAMP4 its surface active characteristics. DAMP4 adsorbs to hydrophobic-hydrophilic interfaces, for example air-water or oil-water, as a monolayer with the thickness of a single helix, indicating an unfolding of the four-helix bundle into four single helices[57]. However, compared to DAMP1, the absorption kinetic of DAMP4 to the interface is slow, slower than expected for a diffusion controlled process, indicating a large energy barrier for DAMP4 that does not exist for DAMP1[57]. This barrier is very likely associated with the free energy of unfolding which has to be overcome for adsorption for DAMP4 and is influenced by the strong hydrophobic interactions in the core. This balance of different interactions and forces has to be considered according to the application of new designs which are stable enough for high expression yield and purification, but highly functional with fast adsorption kinetics.

Once adsorbed to the interface, DAMP4 is able to build a stable foam in the presence of kosmotropic salt (Na2SO4), but not in the presence of chaotropic salt (NaSCN)[8]. As suggested previously, kosmotropic Na2SO4 may share DAMP4’s hydration layer and electrostatically stabilize the thin film, and at the same time increase hydrophobic interactions, so the barrier of desorption is higher and the foam more stable. Compared to that, chaotropic NaSCN may weaken DAMP4’s hydration layer and hydrophobic interactions, which leads to a decreased barrier for desorption from the interface and almost no foaming is observed.

As can be seen from DSC and MD, DAMP4 unfolding is a cooperative process, in which secondary and tertiary structure are lost simultaneously. Once its structure is denatured, unfolding in bulk is not reversible, as upon unfolding, exposed hydrophobic residues of different molecules interact strongly with each other, which ultimately leads to the formation of strong aggregates. It is not clear if this process operates only in bulk, and whether interfacial desorption might be associated with a refolding process.

The MD results suggest, the mutation of more than three hydrophobic core residues leads to a rapid loss of stability. However, up to three hydrophilic residues in the core can be tolerated in terms of thermostability. This can be particularly useful in new designs, when higher solubility or faster adsorption kinetics are required.
4.6 Conclusion

This research uniquely linked sequence design to stability, process behaviour, and functionality and aimed to broaden the application of sustainable biosurfactants in more industrial products including low-cost industrial sectors. We investigated the stability of a unique biosurfactant four-helix bundle protein, DAMP4, using experimental techniques and molecular dynamics simulation of the protein and different stability variants. The results suggested that the hydrophobic core is the most important feature for DAMP4’s stability and leads to a melting point of 122.4°C. This fact is exploited in a simple “bake-to-break and precipitate” (BPP) process which allows quick and economical purification to industrial grade DAMP4. Other results, namely the stability against GdmCl, adsorption behaviour, and pH tolerance, support the importance of the hydrophobic core for DAMP4’s characteristics. The broad pH and temperature tolerance demonstrated is of high significance for industrial processes and applications and helps to target current problems of biosurfactants. The MD simulations showed an influence of number and location of mutations in the hydrophobic core on stability and unfolding behaviour which may be considered in future sequence designs.

4.7 Acknowledgments

The authors would like to gratefully acknowledge funding by an Australian Research Council Discovery Grant (DP120103683). A.S. was financially supported by an UQI Tuition Fee Scholarship and an AIBN RHD Living Scholarship. A.P.J.M. thanks the Queensland Government award of the 2010 Smart State Premier’s Fellowship. We acknowledge Lei Yu for executing DAMP4 expression and purification. Further, we would like to thank the HPC support from the University of Queensland Research Computing Centre (RCC), the Queensland Cyber Infrastructure Foundation (QCIF) and NCI National Facility for their support and the supercomputer resource allocation.

4.8 Supplementary material

Figure 4-9 Multiple sequence alignment[58] for monomer DAMP1 and homology modelling template 3EDC of E. coli’s lactose operon repressor.
Figure 4-10 Ramachandran plot for DAMP4 backbone atoms.

99% of residues are in the favoured, 1% in the allowed region[37].

Table 4-3 Relative variance, variance and mean rmsd $C_\alpha$ as numerical descriptors for rmsd $C_\alpha$ plots depicted in Figure 4-5 and Figure 4-6.

<table>
<thead>
<tr>
<th></th>
<th>Relative variance rmsd $C_\alpha$ $\frac{STD}{Mean}$ [%]</th>
<th>Variance rmsd $C_\alpha$ $STD^2$</th>
<th>Mean rmsd $C_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAMP4</td>
<td>10.0858</td>
<td>0.1439</td>
<td>3.7547</td>
</tr>
<tr>
<td>hc1</td>
<td>17.0660</td>
<td>0.4933</td>
<td>4.1155</td>
</tr>
<tr>
<td>hc2A</td>
<td>10.0039</td>
<td>0.1688</td>
<td>4.1067</td>
</tr>
<tr>
<td>hc2B</td>
<td>8.4170</td>
<td>0.0883</td>
<td>3.5311</td>
</tr>
<tr>
<td>hc3</td>
<td>14.5264</td>
<td>0.4130</td>
<td>4.4239</td>
</tr>
<tr>
<td>hc4A</td>
<td>16.3894</td>
<td>0.6279</td>
<td>4.8350</td>
</tr>
<tr>
<td>hc4B</td>
<td>21.3206</td>
<td>1.1751</td>
<td>5.0844</td>
</tr>
<tr>
<td>hc8</td>
<td>18.3966</td>
<td>1.0552</td>
<td>5.5839</td>
</tr>
<tr>
<td>sb4A</td>
<td>13.1367</td>
<td>0.3057</td>
<td>4.2088</td>
</tr>
<tr>
<td>sb4B</td>
<td>10.7227</td>
<td>0.2012</td>
<td>4.1837</td>
</tr>
<tr>
<td>sb4C</td>
<td>12.2776</td>
<td>0.1928</td>
<td>3.5766</td>
</tr>
<tr>
<td>ws8A</td>
<td>10.4293</td>
<td>0.1674</td>
<td>3.9227</td>
</tr>
<tr>
<td>ws8B</td>
<td>9.9407</td>
<td>0.1357</td>
<td>3.7060</td>
</tr>
</tbody>
</table>
4.9 References

Chapter 5  Adsorption behaviour of a four-helix bundle biosurfactant to an air/water interface

5.1  Introduction

The folding/unfolding events occurring due to the structure of a protein are closely related to its function[1]. Understanding these relationships, including when, how, and why structural rearrangements occur, can facilitate the design of new and improved, and currently highly demanded, sustainable molecules for uses in industrial applications.

One key area for the use of designed biomolecules produced in microorganisms such as yeast, or bacteria, in particular *Escherichia coli* (*E. coli*), is the use of proteins as biosurfactants, the sustainable alternative to petroleum-derived chemical surfactants. Surfactants lower the interfacial tension, reaching values close to that of the solvent. The dynamical surface tension, and how the molecules influence it, are very important factors as it determines many biological, as well as industrial processes and the understanding and specific modification are inevitable for new and improved product designs in particular for biosurfactants[2]. This involves an understanding of the drivers, pathways and mechanisms of interfacial adsorption, conformations and conformational changes upon adsorption, timeline and approaches to influence these characteristics.

The structure of protein or peptide biosurfactants varies from that of classical surfactants with their facial amphiphilicity rather than the conventional hydrophilic headgroup - hydrophobic tail setup, leading to a broader variety of possible arrangements and interactions. Due to the flexibility in the design of their side chains they offer a vast array of design opportunities, with the possibility to specifically create certain interactions, such as switchability and reversibility of their function depending on the environmental conditions[3]. However, a better fundamental understanding of underlying mechanisms, conformations and interactions is needed on a molecular base to further drive the use of this green alternative in industrial products, and this chapter aims to establish the links between structural or sequential features and of the events occurring at the interface for DAMP4[4].

As described in Chapter 4 DAMP4, produced intracellularly in *E. coli*, can be purified in a simple and cost-efficient two-step process[5]. This is possible due to DAMP4’s extraordinary stability in bulk, with a melting point of 122 °C, likely caused by the tight hydrophobic interactions in the protein’s core[6]. DAMP4 is special in a way that it contains all hydrophobic residues buried in the core, and none of them are exposed to the water phase in its’ native structure. Despite the obvious
tightness and stability of the four-helix bundle in bulk, experiments suggest an unfolding of the molecule upon adsorption to an interface. Neutron reflectometry shows DAMP4 adsorbs to an air/water interface as a monolayer, with an interfacial film thickness of 10-14 Å, which corresponds to the thickness of an α-helix[7]. Further, adsorption time for pure DAMP4 solutions is slower than expected for a simple diffusion controlled process, suggesting an energetic barrier for adsorption which e.g. does not exist for its monomer DAMP1. Circular dichroism (CD) reveals, compared to DAMP4, DAMP1 is unstructured in bulk, however likely folds into a helical structure when adsorbed to the interface[7].

These structural changes are closely related to the molecule’s surface activity. Drivers for unfolding in a protein due to a change in conditions, such as the proximity to an interface, are governed by van der Waals and electrostatic interactions, hydrogen bonding, the hydrophobic effect and favourable energetic changes due to conformational entropy[8].

In this chapter, the unfolding of DAMP4 upon adsorption to an air/water interface will be confirmed, and the drivers and conformational arrangements of DAMP4 investigated using molecular dynamics simulations. This is an alternative to experimental methods in which the molecules and the adsorption process can be monitored in atomic detail including conformations and/or conformational changes. Due to a lack of effective experimental methods on the atomic scale, only few studies on interfacial unfolding behaviour of amphiphilic proteins and peptides, and in particular other four-helix bundles, have been conducted. Molecular self-assembly at interfaces is to a great extent unexplored and a full understanding of the biomolecules’ chemistry at interfaces is still lacking[3, 9]. Experimental techniques such as Langmuir or spectroscopic measurements or atomic force microscopy can be used to study interfacial systems[10]. However, the direct experimental analysis of conformations at the interface is problematic, as for example crystal structure determination at the interface using NMR spectroscopy, electron microscopy or X-ray diffraction remains difficult, in particular for aggregate-forming molecules[11], and the conformation at the interface can vary widely from that observed in bulk. Different experimental techniques can infer tertiary structures and lead to hypotheses of conformation and assembly of molecules. Mackenzie et al.[12] determined the 3D structure of Ranaspumin-2 (Rsn-2), a naturally occurring surfactant protein in frog foam nests, which in bulk adopts a globular fold with four antiparallel β-sheets and one α-helix across them. Based on their results from secondary structure analysis with infrared reflection absorption spectroscopy, and the layer thickness determined by neutron reflection they suggest a conformation of Rsn-2 at the interface with a clamshell-like unfolding of this compact structure with β-sheets and α-helix in one plane and thus maintaining its secondary structure when incorporated into the air/water interface. However they emphasized that
these findings of the structure at the interface are purely speculative at that stage of their study and only inferred from experimental observations[12]. It is known that chemistry and composition of the aqueous phase close to an interface is quite different from the bulk and characteristics such as water orientation, ion concentration or pH can vary strongly. However the kind and extent of these differences is an area of active research and results thus far are somewhat controversial. For example it has been claimed that the pH at the air/water interface shows a discrepancy from the bulk pH of several units, with an acidic pH<4.8[13-15] and a preference of H3O+ hydronium ions in elevated concentrations at the interface using results from simulation[14] and experiment[16]. Other studies suggest the surface has a negative charge with the augmented presence of hydroxide OH−-ions[17]. This is why in this study different interfacial pHs and conditions were tested on their potential to be a trigger for DAMP4’s unfolding. A combination of mechanisms leading to this event will be suggested and the event of unfolding and the final conformation at the interface elucidated.

5.2 Material and methods

5.2.1 Molecular dynamics simulation

All molecular dynamics simulations were conducted with the software package YASARA structure version 14.2.12 (YASARA Biosciences, Vienna, Austria)[18, 19]. The simulations were run in the NVT ensemble with periodic boundaries, a Berendsen thermostat which rescales atom velocities for temperature control and Particle Mesh Ewald for long-range electrostatics[20]. Time steps of 2 fs and 2.5 fs were used for intra- and intermolecular forces, respectively. The force field Amber03[21] for proteins and biomolecules was selected and simulation snapshots saved every 25 ps. Amber based force fields have been successfully used before for interfacial simulations of proteins[22]. A cutoff for van der Waals interactions of 10.48 Å was used. For the simulations, the molecule was placed in a cubic simulation cell and the cells filled with TIP3P water molecules and the protonation state adjusted according to pH[23]. Counter ions to neutralize the cell are placed at locations of minimum/maximum electrostatic potential. Then a steepest descent energy minimization to remove conformational stress was applied, followed by a simulated annealing minimization with a time step of 2 fs and the scaling of atom velocities by 0.9 at every 10th step until the potential energy converges, which means it changes less than 0.05 kJ/(mol atom). To create an interface, the simulation cells were extended to three times the size of the original box in z-direction[10, 24, 25]. This generates a water slab with two interfaces which can be referred to as a water liquid-vapour interface[10]. The interface is treated as a vacuum, which is a reasonable approximation because of
the small size of the system and the low probability of finding air molecules such as O₂ or N₂ in this volume[25]. In the case of simulations in the presence of hydronium or hydroxide, water molecules in proximity to distinct residues (Table 5-1) were replaced with the respective ions. Counter ions for cell neutralization were added, followed by another energy minimization. Snapshots were analysed in YASARA and data evaluated with Matlab R2014b (Mathworks, Natick, MA, USA).

5.2.2 Unfolding simulations of DAMP4 in the presence of an air/water interface with ions

Three basis simulation scenes were set up for the unfolding simulations of DAMP4 at the air/water interface to investigate the drivers for unfolding and the influence of pH, ions and location. DAMP4 was loaded into a simulation box of dimensions 100 Å x 40 Å x 100 Å. The molecule was placed at the upper part of the simulation cell with helices 1 and 4 facing upwards. The simulation box was filled with water molecules and energy minimized. Like this the three basic scenes vary slightly in location of DAMP4 in respect to the interface and location of individual water molecules. For interfacial simulations the simulation cell was increased by three times in z-direction to create the interface, resulting in the final dimensions 100 Å x 120 Å x 100 Å. Each of these scenes was then adjusted in protonation state for the three different pH conditions tested here, namely pH 7.4 for neutral, pH 4.8 for acidic, and pH 10.5 for basic environment. This corresponds to unchanged interfacial settings, or a shift towards acidic or basic environment at the interface with consistently neutral pH in bulk, respectively. Further, at each pH four different conditions were considered: the presence of four H₃O⁺, four OH⁻, the combination of four OH⁻ and four H₃O⁺, and without any ions. Therefore selected water molecules were replaced by respective OH⁻ or H₃O⁺ ions close to specifically selected residues based on previous results (see Table 5-1). In previous test simulations of DAMP4 at the interface with a high number of ions, the location of these turned out to be important, and the corresponding residues were chosen accordingly for the simulations with a low number of ions to increase chances of an unfolding event to occur. Depending on protonation state and ions added, various numbers of Cl⁻ and Na⁺ ions were added to neutralize the cell and the energy of the system was minimized.

Table 5-1 Overview of different starting ion locations for DAMP4 unfolding simulations for the three basic systems. Each of these systems was simulated three times with different starting conformation.

<table>
<thead>
<tr>
<th>System</th>
<th>H₂O⁺ location</th>
<th>OH⁻ location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Asp 10, Glu 23, Glu 95, Asp 98</td>
<td>Arg 17, Arg 21, Arg 89, Arg 93</td>
</tr>
<tr>
<td>2</td>
<td>Asp 2, Glu 23, Asp 74, Glu 95</td>
<td>Arg 17, Arg 21, Arg 89, Arg 93</td>
</tr>
<tr>
<td>3</td>
<td>Asp 4, Asp 10, Glu 23, Glu 95</td>
<td>Arg 17, Arg 21, Arg 89, Arg 93</td>
</tr>
</tbody>
</table>
This set up of production runs was chosen as the pH at an air/water interface is still unknown, and preliminary studies of DAMP4 simply in the presence of an air/water interface did not show any signs of unfolding within periods of hundreds of nanoseconds. Similar was observed by a variation in amino acid sequence or a rise in temperature. However unfolding at the interface was promoted by the addition of ions at random locations around the molecule sitting at the interface. By slow reduction in ions from high concentrations we observed that a minimum of four ions was necessary for unfolding to occur within a simulation time of approximately nanoseconds, and that at least three of these ions need to be in specific turn residues and charged side chains.

The simulations were analysed in terms of parameters indicative of conformational changes, including hydrophobic solvent accessible surface area, secondary structure content, and distances between specific residues representing the distances between helix 1 and 4 at the top (Pro 3-Leu 94), the centre (His 13-His 85), and the bottom (Glu 23-Ser 76). The simulation data for each of the three starting scenes and the three ion locations, resulting in nine simulations, was averaged at each time step. This was repeated for each combination of pH and ions. Further the simulation snapshots were visually analysed for evidence of unfolding.

5.2.3 Potential of mean force (PMF) calculation

For the PMF calculations the starting conformations and compositions were set up as described previously using Packmol[26]. Simulation force field parameters were used as derived from the
Amber03 force field by using Antechamber[27]. The simulations were run in triplicates with the LAMMPS molecular dynamics simulator[28] in the NVT ensemble with periodic boundary conditions, a Lennard Jones cutoff of 9.8 Å and long-range Coulombic interactions. Bonds and angles were constrained with the SHAKE algorithm[29]. The timestep was 1 fs, and an energy minimization was followed by a 10 ns equilibration of the entire system. Thereby, the centres of mass of the upper (Met 77 - Leu 87) and lower (Leu 87 - Ala 97) parts of helix 4, not including the turn regions, were fixed to their position after minimization to prevent the molecule from rotation on the interface. Further the distance of the centre of mass of the entire protein, and the centre of mass of the layer of water molecules was constrained to 16.9 Å to ensure the location of DAMP4 at the interface. For the bulk simulations the size of the simulation box was changed to 90 Å x 90 Å x 90 Å to prevent self-interaction of DAMP4 upon unfolding. In addition to the above constraints in the presence of ions at pH 4.8 and pH 10.5 the distance between the centre of mass of helices 1 and 4 was fixed at 9.2 Å in x-direction as spontaneous unfolding was to be expected. Ions were restrained with a spring constant of 5 kcal/mol/Å² to the respective oppositely charged residues in all directions. For the production simulations, the distance between the centres of mass of the two helices was increased from 9.5 Å to 27.95 Å in 10 steps in the x-direction. At each distance, stepwise increasing springs with force constants of 5 kcal/mol/Å², 10 kcal/mol/Å², and 25 kcal/mol/Å² were applied and with each the system was equilibrated for 200 ps before moving on. After the 200 ps equilibration time at 25 kcal/mol/Å², the spring force was collected for 500 ps with block averaging, whereby the force at every 10th timestep was averaged over 5000 values, resulting in ten averaged data points per distance and corresponding standard error of the mean. These averaged potentials were used to construct the free energy for unfolding by numerical integration of the averaged potential along the reaction coordinate. A stiff spring approximation was used for the pulling, with a harmonic force that was tested to ensure it was sufficiently strong[30].

5.2.4 Calculation of water distribution around DAMP4

To identify the water distribution around DAMP4, a distribution function, g_r, was determined. The number of water molecules N was determined in shells of 1 Å steps around DAMP4 and was normalized by the available volume as approximated by the volume of a cylindrical shell V with respect to an outer radius r_2 and the inner radius of the shell r_1.

\[ V = \pi h (r_2^2 - r_1^2) \] (27)
For DAMP4 the values of radius \( r \) and height \( h \) were estimated to 12 Å and 40 Å, respectively. The values were averaged over ten equally distributed time spots from 1 ns to 10 ns simulation time and normalized by the particle density of H\(_2\)O \( n_{\text{water}} \).

\[
g_r = \frac{\sum_{t=1}^{10} \left( \frac{N_{r,t}}{V_r} \right)}{10 \, n_{\text{water}}} \tag{28}
\]

Based on a density of 1 g/cm\(^3\) \( n_{\text{water}} \) can be determined as 0.0335 H\(_2\)O molecules/A\(^3\).

**5.2.5 Calculation of dipole moments and quantification of orientation of water molecules**

Dipole moment vectors of water molecules were used to identify the orientation of water molecules. Electric dipole moments were calculated as the length of the vector \( \vec{P} \) determined by

\[
\vec{P} = \sum_{i} q_i \ast (\vec{R}_i - \vec{C}) \tag{29}
\]

with charge \( q \) and cartesian coordinates \( R \) of atom \( i \). \( C \) corresponds to the geometric centre of the water molecule. Magnitude of dipole moments is dependent on the current force field.

To obtain information on the water structure a simulation of a 80 Å x 80 Å x 80 Å TIP3P water slab in the absence of a biomolecule was run for 10 ns with the same parameter and conditions used as for the peptide interfacial simulations. Snapshots were saved every 250 ps for analysis purposes and the simulation box divided in 240 1 Å slices along the z-axis. To estimate the order of the water molecules in proximity to the interface in comparison to bulk in the simulation, two different parameters based on the dipole moment vector were used. Firstly, the angle between the xy-plane and the dipole vector was used as a measure to compare the direction in which the dipole moments were pointing and thus the direction of the water molecules in each of the 1 Å bins. These were conflated into groups of interfacial water (distance < 3 Å from the interface), intermediate water (4-8 Å from the interface) and bulk water (9-13 Å from the interface). A polar plot with the angle and the length of the vector corresponding to the instantaneous dipole moment was created to visualize the results. As a second measure for the ordering of the molecules, an average vector of all dipole vectors, normalized by their length, in each bin was calculated. The magnitude of this average vector is enhanced if the water molecules are pointing in the same direction, is cancelled out for vectors pointing in opposite direction, and reaches 1 for perfectly aligned and ordered water molecules. The values were averaged over 10 snapshots of an equilibrated simulation taken at 1 ns intervals from 1 ns to 10 ns.
5.3 Results and discussion

Simulations were used to investigate the likelihood that the water structure or other factors (pH, ions) would influence the structural changes occurring in DAMP4 at the air/water interface.

5.3.1 Investigation of the water structure at the interface and around DAMP4 for unfolding

Numerous studies have confirmed the importance of structured water around proteins and its influence on protein structure, function and stability[31, 32]. DAMP4 is likely to hold a tightly packed and highly oriented water shell because of its unique structure with a firmly packed hydrophobic core and a high number of arginine residues on its surface. Previous studies indicated the protein’s topology and charge were some of the most important factors determining the surface hydration states and sites[31, 32], and that water molecules are less tightly bound in narrow crevices due to surface tension water forces[31]. DAMP4’s four-helix bundle fold offers a rather smooth surface without any deep crevices. Its solvent accessible surface comprises exclusively hydrophilic residues, and water shows preferred interactions with polar side chains with which they built specific H-bonds, restricting their motion[31]. Further, DAMP4 has eight positively charged arginine residues on its surface. Arginine is unique among all amino acids, possessing five possible H-bond donor sites, and strongly interacts with water molecules, thereby providing a strong hydrogen-bonded network around DAMP4[33, 34]. This network of structured water might play a key role in overcoming the high energy barrier for DAMP4 in proximity to the air-water interface. The air-water interface impacts a range of properties in the water molecules below, including a lowering of the diffusion coefficient compared to bulk[35], different hydrogen-bond patterning[36], and structuring and orienting of water molecules as first suggested by Hansen in 1961 when studying adsorption kinetics of aliphatic acids and alcohol as the process seemed barrier limited at low times and diffusion limited at intermediate times[37]. This mechanism inferred a breaking of the oriented water structure at the interface by the earlier molecules, releasing the diffusion barrier through ordered water, whereas later molecules seemed solely limited by their diffusion to the interface as the interfacial ordered water barrier had been overcome already.

Besides experimental methods to determine protein hydration, such as neutron and X-ray diffraction, simulation methods have been proven useful to help identify and understand the role of water molecules on protein function and structure[31]. Henchman and McCammon used molecular dynamics simulation to investigate the water structure around acetylcholinesterase and investigated different static and dynamic water properties around the enzyme including occupancy, neighbouring waters, H-bonds, dipole moments as well as residence times. They found variations in these properties depending on the proximity of the water molecules to the protein and the hydration site
such as decreasing numbers of neighbouring waters upon approximation, together with more water-residue H-bonds. Equally, deeper buried waters revealed longer residence times, in dependence of the topology. They concluded the importance of treating proteins as hydrated complexes instead of single molecules[32]. In recent decades a lot of effort experimentally as well as computationally has been put in the understanding and influence of the water structure in the vicinity of different interfacial systems[35, 38].

In this work a 10 ns MD simulation of a pure water slab without any protein molecules at 298 K as well as a 10 ns simulation with minimized and fixed DAMP4 in pure flexible bulk water were investigated for some kind of structuring of water molecules which could assist in the unfolding event of DAMP4 occurring at the interface.

**Hydration around DAMP4 in the simulation**

Hydration around DAMP4 was analysed as previous results indicated an importance of the water shell around DAMP4. Firstly the water molecule distribution function, defined by equation (28), of water molecules around DAMP4 was determined. Results showed the first hydration shell with the peak in normalized waters per volume within a distance of 3 Å from DAMP4 (Figure 5-2). The second hydration shell is barely visible and much less pronounced compared to the first shell, and located at a distance of around 6.5 Å from DAMP4. The figure shows a classical water distribution around a protein, and no particular strong hydration or different behaviour compared to other standard proteins could be observed.

![Figure 5-2](image)

**Figure 5-2 Distribution function of water molecules around DAMP4 as defined by equation (28).**

To further evaluate the water distribution around DAMP4, the location of water molecules around fixed DAMP4 was monitored during 10 ns simulation at room temperature. Figure 5-4 shows the results, with each blue spot corresponding to a water location in the superposition of all 400
conformations captured within the simulation. Clusters of blue dots indicate a longer residence time of the waters in these locations, or a higher water density around the corresponding residues. For better visualisation this plot is the cross-sections along the z-axis in 5 Å slices (see Figure 5-3). Although the water molecules in the bulk with a distance further than approximately 7 Å from the interface are very evenly distributed, a high density of water molecules can be found in a distance up to 3 Å from the molecule’s surface, corresponding to the first hydration shell. This is followed by an area of low density of water molecules, before the next hydration shell is lightly visible.

As previously seen in the radial distribution function of the water molecules in Figure 5-2, the second hydration shell is much less pronounced than the first shell. It is possible that a simulation longer than 10 ns might have revealed a greater difference in water density between the different hydration shells and bulk water as high water density spots will become clearer with increasing sampling time. The figures indicate that some residues have higher water density or longer residence times than others. Among these residues are as due to their potential as hydrogen bond donors expected arginines (Arg 89, Arg 41, Arg 21), but as well glutamines (Gln 79, Gln 14, Gln 86, Gln 55) or other, in particular charged, residues (Asp 98, Asp 2, Lys 54, Glu 95). Charged residues with long water residence time will later prove important for DAMP4 unfolding in the presence of an air/water interface.

**Figure 5-3 Visualization of location of cuts of DAMP4 along the z-axis in 5 Å slices to identify hydration sites.**

In a 10 ns long simulation of fixed DAMP4 in flexible water molecules the location of each single water molecule was captured and visualized in Figure 5-4. DAMP4 Cα-atoms are pictured in balls, side chains as sticks.
Figure 5-4 Cross sections of 3D plot of location of water molecules around DAMP4 in 10 ns simulation with the protein fixed at its initial location in 5 Å slices from a) z=-20 to h) z=+20 as visualized in Figure 5-3. Red dots correspond to Cα-atoms of the respective residues; each blue dot represents the location of a water molecule throughout the simulation.
It was shown that the Amber03 force field does promote protein hydration with different hydration shells and hydrated spots, however simply the formation or removal of a strong hydration shell does not seem sufficient to unfold DAMP4 in the simulation, and additional drivers such as for example charges might be needed to overcome the energetic barrier for unfolding of DAMP4 in MD simulations.

**Ordered water at the air/water interface in the simulation**

In the slab simulations of pure water without protein, the dipole moments of water were used as a measure for the orientation of molecules. In highly ordered water the average dipole moment approaches the instantaneous dipole moment of water which is 2.4 D, as no cancellation due to the effect of water molecules facing a different orientation occurs[32].

![Figure 5-5 Dipole moments of all waters in 1 Å bins along the z-coordinate normalized by the number of molecules. Z-coordinates of the simulation box of 80 Å to 160 Å correspond to the water slab.](image)

Figure 5-5 displays the average dipole moments of all water molecules in 1 Å bins along the z-direction of the water slab. In the case of ordered water at the interface one would expect waters in a distance of 0-3 Å from the interface (in Figure 5-5 this belongs to z-coordinates 80-83 Å and 157-160 Å) to be close to the instantaneous value of 2.4 D. However, in this simulation this value is only reached at the lowest distance from the air/water interface, when there are very low numbers of water molecules and conclusively no mutual interaction because of their far distance from each other. In the respective interfacial interval the value reaches around 0.86 D and 0.70 D, which is still above the average value of 0.34 D for the bulk region. This could indicate some ordering of the water molecules at the interface. To further investigate the orientation of the water molecules the angle of the dipole moment vectors were figured in a polar plot with the distance of the points from
the centre corresponding to the dipole moment. Figure 5-6 shows the result of this for the waters after 10 ns. The results showed that the water molecules exhibit angles within the full range of $90^\circ$ to $-90^\circ$ towards the plane.

Figure 5-6 Angle of the dipole moment vector of each water molecule and the $xy$-plane.

a) Interfacial water (0-3 Å distance), b) intermediate water (4-8 Å distance from the interface), and c) bulk water (9-13 Å). Colours mark 1 Å bins that compose the different layers in the order of blue, red, yellow, purple, and green.

No augmented formation of clusters could be observed for the interfacial region which was arbitrarily defined up to a distance of 3 Å from the ideal interface. The angles for the interfacial region in Figure 5-6a) seem to span a slightly lower range of angles compared to intermediate b) and bulk water c), however this is likely to be an artefact of the lower number of molecules in the respective bins.

Figure 5-7 Magnitude of average dipole moment vectors.

Calculation as the mean of ten simulation snapshots over 10 ns simulation time along the $z$-coordinate of the simulation box. The interface is located at 80 Å and 160 Å, respectively. The inlet shows the magnitudes with increasing distance from the interface.
Conclusively an average dipole vector of all single water molecule vectors in each 2 Å-slice, normalized by their length, was calculated as an average of ten simulation snapshots. For ordered water, with molecules facing the same direction at the interface, the length of the average vector would be larger, as the vector would be enforced, whereas oppositely facing water molecules will decrease the length due to cancellation. The results are showing the expected trend of larger magnitudes at the interface (z-coordinates 80 Å and 160 Å), however this trend is not continued further into the bulk phase, as only the 158 Å bin shows values above the average for bulk and not the other interface with the 82 Å bin. One can expect the interface to have effects on the water structure and pKa values up to a distance of 5 Å[39] into the water and the orientation of the water molecules in bulk here seems rather random with an average length of the vector of 0.1 compared to perfectly aligned water which would have a value of 1. Altogether the results presented here on the structuring and orientation of water molecules close to the air/water interface in the MD simulation lead to the conclusion of no to little influence of the interface on the water structure. This can either be an artefact of the applied force field (Amber03) or water model (TIP3P). This includes the lowering of the energetic penalty due to ordered water molecules and the influence of the strong hydration shell of DAMP4.

5.3.2 Other factors influencing DAMP4 adsorption to the interface: Drivers, mechanism of unfolding and conformation of DAMP4 at the interface

Physical and chemical characteristics of the interface: A controversial subject

The air/water interface is an area of active research; however its physical and chemical characteristics are still poorly understood, and even controversial[13, 17, 40, 41]. A consensus seems to have been reached however, on the fact that the composition and characteristics of the water surface are quite distinct from the ones in bulk water[13, 17, 42, 43]. One area of debate is related to the surface pH, and the concentrations of hydroxide and hydronium ions due to the autoionization of water. Even though the entire water phase is neutral of charge with a pH of 7, the location of ions, their concentration and resulting interfacial pH is an area of dispute and spectroscopic experimental and computational results seem to contradict macroscopic observations. Computational studies point towards an acidic surface of water and an increased concentration of hydronium ions. Jungwirth et al.[13] performed computational work on the nature of the air/water interface. They used MD simulations of single H₃O⁺ molecules in an 11 Å thick water slab and found a 150-fold increase of the ion concentration at the surface compared to bulk at 300 K. This corresponds to a pH of 4.8 at the top water surface layer and they concluded the surface of pure water is acidic. Their explanation for the surface propensity and stabilization of hydronium is the proton’s characteristics as a strong H-bond donor, but poor acceptor and a preferential adsorption at
the interface prevents a disruption of the water network. They performed experiments with infrared spectroscopy on ice nanoparticles to prove their surfaces are slightly acidic and found significant isotopic exchange (D$_2$O$+\text{H}_2$O$\rightarrow$2HDO) due to proton transfer, indicating the acidity[13]. Petersen and Saykally came to the same conclusion after performing UV second harmonic generation (SHG) experiments. They measured the surface propensity of different ions by means of their concentration on the surface vs. bulk and modelled the Gibbs free energy of adsorption and excluded a surface enhancement of hydroxide ions[44]. Mundy et al. agree with an excess of protons at the interface, however their MD simulations showed that the excess of H$_3$O$^+$ over OH$^-$ ions is significantly less than stated by Jungwirth and his group. They found a stabilization of hydroxide at the interface by $k_BT$ and thus concluded slightly enhanced surface population[45]. A range of other experimental[16, 46, 47] and computational[13, 15, 48-50] studies were consistent with these molecular-scale results.

Macroscopic experiments indicate opposite findings to molecular-level studies, with the adsorption of hydroxide ions at the interface. For example Beattie and co-workers strongly disagreed with the computational findings by Jungwirth et al. and suggest a neat water surface is basic and negatively charged due to the preferential adsorption of hydroxide ions[17]. They based their theory on experimental observations, as gas bubbles in water move towards the positive electrode in electrophoreses, implying a negative charge at neutral pH. The same has been observed for oil bubbles in water in absence of surfactants. Beattie et al. confirmed these earlier results by pH measurement of hexadecane in NaCl-water emulsion as the protons released from the hydroxide ions adsorbed at the interface caused a measurable drop in pH unless buffered with NaOH solution[51]. In addition, hydrophobic solid surfaces without ionisable groups in water showed the same characteristics of a negative surface charge. However they noted, there was no spectroscopic indication that there are hydroxide ions within a pH range of 2 to 13[17]. Gray-Weale and Beattie recently published a theory which explains the affinity of the OH$^-$ ions for the interface with the suppression of water dipole moment fluctuations in vicinity to ions. This causes an attractive force towards regions with lower than bulk dipole-fluctuation density on the ions, which is particularly strong for hydroxide as the oxygen strongly binds four water molecules. The interface reduces the ions constrained solvent region and the free energy cost becomes smaller. They based their theory on experimental observations which showed the driving mechanism is independent of the nature of the hydrophobic surface or other ions present, and simulated the effect of single hydroxide ions on nearby dipole fluctuations and fitted the results to experimental data revealing the size of this reduction. They showed their theory is not opposing earlier spectroscopic results as the hydroxides are mainly below the outermost molecular layers[52].
Even though the extent of variation and physicochemical characteristics of the interface cannot be determined with certainty, it is very likely that this change in environment upon diffusion to the interface plays a fundamental role for the unfolding process of DAMP4 at the interface with a destabilization of the protein which is extraordinarily stable in the bulk. As literature on the direction of the change is ambiguous both conditions, an acidic or a basic interface, were tested in our MD simulations here.

**Protein adsorption and conformational changes at an air/water interface**

Experimental data suggested that DAMP4, which in bulk is arranged to a four-helix bundle, is undergoing conformational changes and unfolding when adsorbed to the interface[7]. Other studies previously proposed the amphiphilic character of such molecules as providing the thermodynamic driver for this process, where polar and apolar termini pull in different directions when proximity to the interface is reached. This leads to a stretching force which opens the tertiary fold without changing its secondary structure, as anticipated for DAMP4[12]. In addition Leon et al.[53] proposed hydrophobic interactions as the driving force for adsorption and unfolding of apolipoprotein-III to lipopolysaccharides and a number of different other structures. In our previous study we found that hydrophobic interactions play a fundamental role for DAMP4’s stability and characteristic behaviour in general[6]. This led to our hypothesis that hydrophobic interactions are a key factor in overcoming the large energy barrier for unfolding of the four-helix bundle into four single helices at the interface. As results above did not show any layers of ordered waters at the air/water interface in the simulation, it was unlikely that the simple presence of the interface will lead to DAMP4 unfolding within a reasonable time frame, and so additional triggers in agreement with interfacial literature were introduced. In the next few sections, it will be shown, that unfolding of the four-helix bundle at the interface is possible in the MD simulation, the conformation of DAMP4 adopted at the interface will be discussed, together with the possible unfolding mechanisms into this state from the compact four-helix bundle in bulk, as well as the conditions required for these conformational changes.

**Unfolding conditions**

DAMP4 consists of four identical helices (helix 1 (H1) to helix 4 (H4)) connected with a DPS linker. Molecular dynamics simulation was used to investigate the behaviour of this bundle in proximity to, and at an air/water interface. Due to the system size required and the large energy barrier present for DAMP4 unfolding due to the strong hydrophobic interactions in the molecule’s core[6] different strategies were applied to succeed in unfolding the molecule at the interface. Firstly, DAMP4 was positioned directly at the interface (see Figure 5-8) as the complete simulation
of the process of DAMP4 as four-helix bundle in bulk, which means in the middle of the water slab and surrounded by water molecules, to adsorption as four single helices did not succeed after five simulations of around 100 ns length. Preliminary simulations of DAMP4 in a completely extended but α-helical conformation with the hydrophobic side chains facing upwards and only covered by one layer of water molecules, suggested a parallel arrangement of the helices at the interface, thus the molecule was placed as exemplary depicted in Figure 5-8 with the plane formed by H1 and H4 parallel to the interface. We hypothesized that this conformation would lead to unfolding of the four-helix bundle in the most simple, fastest and energetically most favourable way by opening up H1 and H4 and like this exposing the hydrophobic core (Figure 5-9), ending up in a structural arrangement which we labelled as “W-conformation”.

When DAMP4 was simulated as a four-helix bundle placed in the middle of the water slab formed by the two interfaces (100 ns) no unfolding was observed in the time scale simulated. This may be due to various reasons, for example the simulation time required for unfolding to occur was not feasible with the computational resources available and/or the molecule might have been stuck in a local energy minimum and did not lead to the observation of an unfolding process but only to the solvation of DAMP4 within the bulk water. Alternatively, the force field applied could not properly represent the physical events happening at the interface (see 2.2.8) and realistically depict the ordering of water molecules as observed experimentally, or different requirements and conditions.

Figure 5-8  Starting configuration for DAMP4 unfolding simulations in the presence of an air/water interface.

a) side view, b) top view. Around 90% of the molecule’s volume is placed in the water. Helices 1 and 4 are facing upwards.
needed for unfolding were not fulfilled in the simulation time applied as the unfolding at the interface is a rare and complex event involving a combination of factors.

Secondly, methods to lower the energy barrier for unfolding by influencing hydrophobic interactions including the design of DAMP4 mutants with hydrophilic residues introduced into the hydrophobic core, increasing the temperature to extreme 498 K[54] or using one different pH-value were applied, but did not succeed in overcoming this barrier without applying an additional trigger. Lastly, in addition to a change in pH value due to a predicted shift in pH in proximity to the interface[13, 17, 39] specifically placed hydronium (H$_3$O$^+$) and hydroxide (OH$^-$) ions were used as the additional driver. As literature on interfacial pH is controversial, extreme interfacial pHs in both directions, pH 4.8 for acidic and 10.5 for basic, as well as neutral pH 7.4, reflecting unchanged conditions at the interface, were considered in this study for a universal unfolding mechanism for DAMP4.

Figure 5-9 Suggested unfolding pathway of DAMP4 into “W-conformation” when an interface is approached.

A shows the unfolding pathway from the starting conformation, B is the intermediate step into the final “W-conformation” C. In this figure the interface is hypothetically located parallel to the two rear end helices (A) and displayed as a blue plane. Hydrophobic parts of the molecule are highlighted in yellow, other residues are coloured in red. Numbers refer to helix labels.

Simulations in the presence of an air/water interface were set up at three pHs, with the addition of 4 H$_3$O$^+$, 4 OH$^-$ or 4 H$_3$O$^+$ and 4 OH$^-$ ions. This number of ions was in previous test simulations identified as the minimum number necessary to allow for unfolding to occur within the time frame investigated here. Three slightly different starting scenes with DAMP4 located at the interface (Figure 5-8) were used in the production simulations. The ions were placed as close to oppositely charged residues on locations as outlined in Table 5-1, so that direct interactions are likely to occur. In most cases, these direct interactions were enforced in order to induce shielding of the charges, leading to a higher net charge and allowing the repulsive charges within the molecule to help trigger unfolding. In the majority of the simulations the hydronium ions directly bond to the respective oppositely charged residues during the simulation and remain at this site with long residence times. In three out of four ions this interaction is permanent throughout the entire simulation. In case of hydroxide ions the interaction is less strong, and residence times of the freely moving ions at their
residue of initial placement are usually less than 1 ns, indicating a lower affinity for DAMP4 compared to hydronium in the simulation. The results for DAMP4 in the presence of an air/water interface are presented in Figure 5-10. These results are the average of three different starting scenes, each with three different starting locations of the respective ions (Table 5-1). The hydrophobic solvent accessible surface area is a measure of the extent of unfolding, as with unfolding this value will increase to a maximum of around 4000 Å² for completely unfolded DAMP4 in W-conformation (see Figure 5-10). In the absence of any ions, no unfolding at any pH can be observed, consistent with previous results, suggesting the need for additional triggers.

Figure 5-10 Averaged hydrophobic solvent accessible surface areas of DAMP4 in the presence of an air/water interface.

pH 4.8 (blue), pH 7.4 (yellow), and pH 10.5 (green) in the presence of a) no ions, b) 4 H₃O⁺ ions, c) 4 OH⁻ ions, and d) 4 H₃O⁺ and 4 OH⁻ ions.

Distances between selected residues on the top, centre and bottom of helices 1 and 4 on DAMP4 were also used as a measure for unfolding. It can be seen that all distances in Figure 5-11a), which shows the results with no ions, remain fairly constant, with minor fluctuations at the bottom. However, the his-his distance at pH 4.8 is within 8 Å on average larger than at other pHs with a standard distance of 6 Å. At pH 4.8 histidine is protonated, causing direct repulsion between His 13 and His 85, which are located in the middle oppositely to each other on helices 1 and 4 (see Figure 5-1). This repulsion likely causes the increased distance, and might facilitate the overcoming of the
energetic barrier for unfolding independent of any ions as this observation is only dependent on the pH. With four H$_3$O$^+$ ions and pH 4.8 unfolding occurs with high likelihood as can be seen by the fact that the maximum hydrophobic surface area is reached (Figure 5-10b). However the ions do not manage to induce unfolding at other pHs. At pH 4.8 DAMP4 has a net positive charge of +7. By adding hydronium ions close to four negatively charged residues this charge can even increase up to hypothetically +11 by shielding of the charges. This extreme net charge, together with the central charged histidines pushing apart, likely results in unfolding. Distance plots confirm the importance of the central histidine residues, as the opening starts with a large increase in his-his distance (Figure 5-11b) in the centre of folded DAMP4, followed by opening at the bottom (Glu 23-Ser 76), and delayed by around 1 ns the top starts to open. 4 OH$^-$ lead to the same results as without the presence of ions, with increased his-his distance at pH 4.8 and minor fluctuations in top and bottom distance. However results change in the presence of a combination of H$_3$O$^+$ and OH$^-$ ions.

![Figure 5-11](image)

**Figure 5-11** Top (Pro 3-Leu 94), bottom (Glu 23-Ser 76) and middle His-His (His 13-His 85) distances for DAMP4 unfolding simulations at different ion- and pH conditions to follow the pathway of opening.

The addition of OH$^-$ seems to counteract the effect of the H$_3$O$^+$ ions as the likelihood of unfolding to occur at pH 4.8 decreases. This can be due to the fact that the charge balance is being reversed and the discrepancy not strong enough anymore to promote unfolding enough, or the OH$^-$ ions have
an additional effect such as for example on the water network around the molecule. We tested to attain the same unfolding effect at pH 4.8 by simple protonation of the respective negative residues, which would make the molecule even more positive than the interaction with hydronium, with the result that the simple change in charge state is not sufficient to cause unfolding, indicating the hydronium ions influence a different factor in addition to simply changing the protonation state. As shown earlier the charged residues are sites of longer residence times of water molecules, and ions at these locations might have an influence on the stabilizing hydrogen network around the molecule. The same results show that a combination of ions could hypothetically lead to unfolding at pH 10.5. At this pH the mechanism of unfolding could be similar, but with an overall negative net charge of DAMP4. Without ions the net charge of DAMP4 is -5 at pH 10.5, because of the side chains of lysine deprotonating at this pH range. The simulation results showed that with the addition of ions, both H$_3$O$^+$ and OH$^-$, unfolding is as well possible at basic pH, however which of the ions was responsible for unfolding cannot be determined here. In additional simulations, unfolding occurred as well at pH 10.5 with the addition of 8 OH$^-$ ions, showing it is theoretically possible to cause interfacial opening simply with a change in pH to 10.5 and only the presence of OH$^-$ ions. However this was only successful in a single one of the simulations tested, and with a concentration increase to double the number tested in this study.

**Conformational changes upon adsorption**

Molecular dynamics simulation provided an atomic insight into the conformational changes taking place when DAMP4 adsorbs to the interface. This is the first study of a four-helix bundle unfolding at room temperature at an interface and thus provided a unique picture of the pathway of the structural changes. Prévost and Ortmans conducted a MD-study in 2001 of four-helix bundle Apolipoprotein E, which unfolds upon binding to lipids, in the presence of a water/organic phase interface to mimic lipid binding. However, due to computational limitations they argued these large conformational changes in ApoE when binding to lipids could not be realized. They explained the experimental observations related to unfolding back to in the simulations identified changes in the water structure in proximity to the interface[55]. However, direct unfolding of a four-helix bundle protein at a hydrophobic/hydrophilic interface in atomic detail has to the best of our knowledge not been reported yet. As anticipated earlier, the orientation towards the interface plays a critical role for the unfolding mechanism in the simulation, and the parallel helices 1 and 4 have to be facing up with the larger hydrophobic plane being oriented upwards, which might happen readily through random diffusion at the interface. This mechanism allows for an easy trigger and fast and short ways to the kinked W-conformation with a minimum energy barrier to overcome.
Figure 5-12 DAMP4 unfolding pathway at an air/water interface as observed in MD simulation.

$\text{H}_2\text{O}^+$ ions are highlighted in yellow. Hydrophobic core residues are colored in magenta. Note that intervals between snapshots are not constant.
The entire process is presented in Figure 5-12. This mechanism is reproducible and the same in all simulations in which unfolding occurred, independent of pH or ions. The complete process from compact four-helix bundle to interfacially extended DAMP4 in W-conformation is on average finished after 1-3 ns, dependent on the start conditions.

Previous results showed an importance of the speed of the initial step, as slow unfolding at the beginning, e.g. due to the presence of a lower number of ions, frequently led to a short opening of the helices. However this was followed by the invasion of water molecules and consecutively the folding back in of the helices. The unfolding starts from the center of the molecule, with the central histidines pushing apart and slightly starting to expose the first hydrophobic residues to the air within 12.5 ps. Next, the exposed helices start to bend slightly towards the outside, adopting an oval structure with more central hydrophobic side chains in contact with the air phase as the interactions in the turn regions are still quite strong partially due to the connecting ions. Finally the pushing forces together with the urge of the hydrophobic residues to favorably interact with the air phase, leads to the breaking of these turn interactions, and the entire helices start to move apart at around 125 ps. From then, helices 1 and 4 simultaneously move apart and internally rotate by 180° to fully expose the hydrophobic side chains as well as hydrate the polar residues. This process finishes after around 1 ns and the entire molecule lies flat and embedded on the surface. During the remainder of the simulation this conformation is maintained. Helices 2 and 3 remain in their position during this process without much movement. The conformational changes observed only comprise tertiary structure changes, whereas the secondary structure is maintained with minor losses and regains of helical structure (Figure 5-13).

Figure 5-13 Secondary structure contents of DAMP4 during the unfolding simulation are maintained.

Radius of gyration ($r_g$) shows tertiary structure changes.
In the W-conformation all helices are parallel and can interact with neighbouring hydrophobic or hydrophilic residues in the air or water phase, respectively. This arrangement of the helices is beneficial due to packing reasons with a surface coverage of around 40 nm²/molecule, as the concentration at the interface can be higher while occupying smaller space. The separation of hydrophobic and hydrophilic residues allows for optimal functionality and hydrophobic residues are shielded from solvation.

Free energy calculations of unfolding under different environmental conditions
To validate previous findings about DAMP4’s unfolding behaviour under different pH and ion conditions we used the potential of mean force (PMF) to calculate free energies of unfolding for selected conditions tested in the unconstrained molecular dynamics simulations.

![Graphs showing potential of mean force and mean force for unfolding of DAMP4 in varying interfacial conditions.](image)

Figure 5-14 Potential of mean force relative to the initial distance or free energy and mean force for unfolding of DAMP4 in varying interfacial conditions.
In the figures, a negative mean force means the molecule wants to open up and unfold, whereas a positive mean force indicates that the DAMP4 molecule would prefer to be more compact. It has to be considered that the distance between helices 1 and 4 after unfolding in W-conformation with no constraints imposed is around 25 Å. The free energies were calculated by numerical integration of the averaged potential along the reaction coordinate, which in our case was defined as the distance between helices 1 and 4. This reaction coordinate was selected due to our previous findings and the knowledge of the unfolding pathway. PMF and free energies were averaged over five independent runs.

The free energy curves followed in general a particular path, which can be seen in Figure 5-14a). Exceptions from the standard pathway are pH 7.4 bulk and pH 4.8 interface, which will be discussed in detail later. Between a helix distance of approximately 10-20 Å the free energy barrier, which needs to be overcome for unfolding, can be seen. As expected from previous results this barrier is highest for pH 7.4 and pH 10.5 without the presence of ions. At a distance between the helices of around 18-22 Å the optimal spacing is reached, as indicated by the local energy minimum. At this conformation the molecule is lying flat at the interface with all four helices in parallel. By increasing the helix distance further, the molecule is overstretched as indicated by the once again rise in ΔG.

The most drastic effect of the addition of ions to a given pH can be seen for pH 4.8, where the barrier for unfolding is decreased by 10.2 kcal/mol from 15.5 kcal/mol to 5.3 kcal/mol. pH 4.8 in the presence of four H3O+ ions is the only condition under which the initial potential at a distance of 9.5 Å is negative, indicating a strong desire of the helices to open up at the beginning. In this environment the low barrier begins in the second reaction coordinate distance (11.55 Å). At the other conditions a pulling force needs to be applied at the start in order to overcome the barrier. Thereby the molecule needs to open up the helices to 13-15 Å, so that initially buried hydrophobic residues are in contact with the air phase and the unfolding becomes favourable. This finding is in agreement with the previous simulations, where at pH 4.8 with hydronium ions present the chances for unfolding were very high. It is important to point out that for the free energy simulations a different software (LAMMPS) as well as a new setup of the same system were used. This confirms the previous results with the unfolding of the four-helix bundle were not an artefact of the simulation software used or the initial setup of the system.

For pH 10.5 the energy barrier is also significantly lowered. A decrease from approximately 10 kcal/mol to 5 kcal/mol shows the barrier is halved by the addition of hydroxide ions. This barrier is comparable to the one at pH 4.8, despite the difference compared to the same pH condition with no ions present being smaller. It shows that the barrier is lowered by the addition of hydroxide ions, however not as much as observed for pH 4.8. This lowered barrier shows that unfolding under these
conditions becomes possible, and should under basic environment be as likely as in the acidic one. However the low number of unfolding events observed in the unconstrained conditions (Figure 5-11) can eventually be attributed to the higher barrier that needs to be overcome in the first step of unfolding, possibly related to the lower charge difference in basic conditions compared to the acidic environment.

It is obvious that at pH 7.4 without an air/water interface (e.g. in bulk) the unfolding of the molecule is highly unfavourable. The more the helices are being pulled apart, the more water molecules penetrate the hydrophobic core. As no distribution of hydrophobic and hydrophilic side chains into the air and water phases is possible, the unfolding in bulk water is not beneficial for the four-helix bundle, resulting in a continuously increasing free energy. This result is as expected and can be considered as a negative control for this set of simulations.

The second exception from the standard curve of the free energies is observed for pH 4.8 with an air/water interface, but without the presence of hydronium ions. Compared to the other conditions the unfolding is highly unfavourable until approximately 17.7 Å helix to helix distance, after which the free energy stagnates. Considering the potential of mean force, this pathway is mainly caused by the strongly unfavourable potential of the first three steps of unfolding. In the other cases, unfolding starts to become favourable in the third step, one step earlier than for pH 4.8. The main structural difference compared to pH 7.4, is the protonation of histidine side chains. Considering the changed charge structure (Figure 5-15), this enables the formation of salt bridges at the upper and lower turn regions between Glu 95 and His 96, and Glu 23 and His 24. In the previous set of simulations we already observed the importance of the location of the hydronium ions in turn regions.

Figure 5-15 Structure of DAMP4 at pH 4.8.

Protonated histidine residues are highlighted in yellow; residues at which hydronium ions were placed are highlighted in green.
Despite the fact that these newly formed salt bridges are not on different helices, they might cause a stronger coherence in the turn regions. The potential importance of these residues can be reinforced by the significant lowering of the unfolding barrier when hydronium ions are present. These ions were placed close to Glu 23 and Glu 95, potentially shielding these charges and lowering the electrostatic interactions with the histidines. Figure 5-11b) additionally shows the opening of DAMP4 in the unconstrained simulations consistently started from the centre of the molecule, whereas the opening of helices 1 and 4 in the turn regions was delayed, probably due to the stronger coherent forces present.

The free energy results show that a change in environment plays an important role for unfolding to occur, as the barrier for unfolding is lowered compared to unchanged conditions at pH 7.4. Significant changes can thereby be observed for acidic conditions, with a very small barrier to overcome. A comparable effect can be seen for basic conditions, where the ions cause halving of the energy barrier for unfolding. Note that halving of the free energy barrier results in a much greater chance of the activation energy being overcome. These results were in good agreement with the previous set of unconstrained unfolding simulations of DAMP4 in different interfacial environments.

**Suggested drivers for DAMP4 unfolding at the air/water interface as determined by MD simulation**

MD simulations showed that a number of factors play together when unfolding of DAMP4 at an air/water interface occurs. We suggested that a change in environment and thus protein charge state might play a crucial role for the conversion from the stable four-helix bundle in bulk, to an unfolded chain of helices in W-conformation upon adsorption to the interface. One factor frequently described to vary considerably in bulk compared to the interface is the pH. However, the extent and direction of this change, and its subsequent influence on other factors such as e.g. ion concentration are uncertain and a field of active debate. The simulation results presented here show a universal mechanism, which allows unfolding of DAMP4 by a change in environmental pH, either in basic or acidic direction due to the shielding of charges which causes repulsive intramolecular forces which are, in combination with several other premises, enough to induce the conformational changes observed experimentally. In this mechanism, DAMP4 is diffusing as a stable four-helix bundle in bulk, and is randomly approaching the interface. As the interface is approached, a pH gradient occurs - either as a shift in basic or acidic direction away from bulk. This leads to an increased concentration of ions, either H_3O^+ in acidic, or OH^- in basic environment, and perturbations of the molecule.
In addition, the chances of protonation or deprotonation increase or decrease in these environments, respectively, resulting in an increasing net charge compared to designed neutrality at pH 7.4. Assuming a shift towards a lower pH, simulated by a pH of 4.8, the probability of histidine residues on DAMP4 to be protonated are high and there is an increasing concentration of H$_3$O$^+$ ions in proximity to the interface. DAMP4 is moving randomly by Brownian motion throughout the layers getting closer to the interface, and by chance will pass these ions, enabling interaction with negatively charged residues on the surface. This interaction is particularly important on helix 1 and helix 4, as these will start the unfolding process. It was reported earlier that surface active proteins can drag charged molecules such as ions to an interface[24].

![Figure 5-16 Summary of the trigger and pathway of DAMP4 unfolding at an air/water interface.](image)

DAMP4 is diffusing randomly in bulk 1) whereby it passes an increasing concentration of hydronium ions due to a decrease in pH towards the interface. The ions will 2) in equilibrium adsorb and desorb. Once the interface is approached and the molecule has the right orientation 3), the unfolding process starts 4) until the final W-conformation is reached 5).

As the interface is close enough for DAMP4 to “see” it, the movement might be directed to the air-phase and the hydrophobic residues inside the molecule start to interact with the air-phase. The change in environmental conditions might cause small fluctuations on the structure, and the molecule starts to unfold and expose the hydrophobic core as described in the previous section of this chapter. The interaction with the ions is a dynamic process and ions adsorb and desorb, possibly
leaving the adsorbed molecule with no direct ion-interactions at some point. We showed that this mechanism is possible in the simulation with the DAMP4 protonation state of pH 4.8 and caused unfolding by covering four negative charges with H₃O⁺ ions on the already positively charged (+7 net) molecule. In particular the positively charged central histidines seem to play an important role in speeding up the unfolding process at this pH. A similar mechanism was observed by Golkaram et al., who suggest a transition from helix to coil of model peptide Ala13 under different pH conditions due to a proton transfer from HCl to charged side chains as investigated by MD simulation with the reactive force field reaxFF. This proton binds more strongly to the oxygen of the peptide bond and causes a split from the chlorine ion and denaturation of the α-helix[56].

Geierstanger et al. investigated the acid denaturation of native sperm whale apomyoglobin and confirmed the importance of the protonation of histidine, aspartic and glutamic acid residues for the conformational change of the molecule to a stable transition state intermediate at acidic pH 4[57]. The same general proceeding can be suggested for a shift at the interface towards basic pH. An excess of negatively charged OH⁻ ions facilitates a larger charge difference and the net negative charge of -5 in the simulation due to lysine deprotonation is enforced with a combination of H₂O⁺ and OH⁻ ions, or simply the interaction with 8 OH⁻ ions (some direct contacts, other just interact with the water layer) also led to unfolding. However, the net charge in the basic pH simulations is lower than at the acidic pH investigated here, and thus unfolding in the simulation might be less likely as the unfolding process might not happen fast enough. No unfolding was observed assuming an unaltered pH at the interface of 7.4, indicating that the simple presence of an air/water interface and to some extent suppressed hydrophobic interactions due to a decreased area of surrounding water molecules are not enough trigger to cause DAMP4 unfolding. This indicates the environmental changes toward the interface likely to be a strong contributor to the forces leading to unfolding. However it is known that molecular dynamics simulation cannot capture all physical effects happening at the interface, and results for pH 4.8 and pH 10.5 were shown to be states which might have forced the molecule into conditions which were not represented otherwise.

Even though an increased number of charges facilitates the actual unfolding process considering a single molecule, it has been shown that higher charges slow down adsorption kinetics due to intermolecular repulsion due to a charging of the interface. In the presence of an interfacial excess of positive or negative ions this effect will be even stronger in the case of shielding of charges for DAMP4, compared to previous studies which describe an acceleration of adsorption rate due to the shielding of charges in the presence of both ions[58]. This shows the stronger effect of charge repulsion between molecules over the speed of the unfolding of single molecules for adsorption kinetics as from simulations one would anticipate a slower unfolding at neutral pH.
In this study we only investigated the unfolding of DAMP4 in pure water of neutral pH, assuming a shift in pH towards the interface to either acidic or basic conditions, or no change in pH from bulk. As literature on this topic is already diverse, other more complex pH conditions, which involve further ionic interactions and concentrations of e.g. buffer ions, were not considered here. From the results obtained in this study we can infer a change in conditions which is important for DAMP4’s unfolding and surface activity, as well as a faster unfolding under acidic conditions consistent with other MD studies, however the mechanism suggested here is universal and can as well be applied under basic conditions. This means, a conclusion on how the interface is distinct from bulk, cannot be drawn from the results here.

In summary, several factors enable the event of DAMP4 unfolding at an air/water interface to occur. Among them are the right orientation towards the interface, a change in environment leading to a destabilization and a corresponding change in protonation state and increased ion concentration acting as a natural trigger to speed up the unfolding process by shielding of charges in particular in the turn regions, together with an additional factor which is influenced by the presence of the ions, such as the water structure. The results presented here show one possible mechanism which might lead to an unfolding of DAMP4 at an air/water interface which worked in the MD simulation. However other mechanisms might apply and more studies are necessary in order to confirm these findings.

**Limitations of the system and force field used**

Two major bottlenecks existed when DAMP4 unfolding was simulated at the air/water interface. The first one was the force field used, which is a very critical parameter in order to realistically reproduce the investigated systems. The second one is the available time scale for the simulation as the time scale in which a lot of biological events happen is frequently not accessible for MD simulation and thus different methods needed to be applied to speed up these processes.

The Amber03 force field is one of the classical, nonreactive all-atom force fields, particularly developed for the simulation of proteins and other biomolecules. Nonreactive force fields are capable of properly simulating a broad range of macromolecules and different environmental conditions and their effects on these molecules. However one disadvantage of this kind of force field is their lack of simulating fluctuating pH. The simulations run with a fixed, pre-defined pH which does not change in response to environmental conditions, e.g. a change to acidic pH due to the proximity to the air/water interface and thus an induced proton transfer between the solution and protein. Techniques such as density functional theory (DFT) can include systems where the pH changes itself, however they are very computationally demanding[56]. Alternatively, reactive force fields such as ReaxFF[59] have been used to overcome this limitation. Unfortunately, the YASARA
software is to date limited to Amber-based force fields, which is why the entire system in this study was considered interfacial area with a constant pH. This is possible, as the water slab is with a thickness of 40 Å a very small system. Further, nonreactive force fields do not include reactions such as proton transfer between ions and protein depending on pH or natural events such as the self-ionization of water. As such the H$_3$O$^+$ ion is a transition state and the ion would rather release a H$_2$O molecule, while the proton stays attached to the respective negative residue when interaction with DAMP4 occurs. However even in the case of proton transfer, the respective H$_2$O would stay around at the transfer site for a while and not diffuse into bulk straight away, which is why the H$_3$O$^+$ ion is a reasonable approach.

In our simulations, the DAMP4 molecule was directly placed at the air/water interface with the majority of the molecule embedded in the bulk water, but the previously suggested face of helix 1 and helix 4 is exposed to the air phase. This is not the most realistic approach, however turned out to be necessary for the unfolding event to occur. This may also be attributed to the limitations of the force field, which was not optimized for the use in air/water simulations and might be underestimating the driving force for a protein to adsorb to the interface and thus the water molecules pushing DAMP4 back into bulk are stronger than in reality, where the driving force pushes towards the interface. The positioning of the molecule directly at the interface lowers the energetic barrier by removal of water molecules on one side just enough to allow for unfolding to happen. Likewise the timescale for the unfolding is lowered to a scale accessible for MD simulation.

All unfolding tests were conducted as single-DAMP4 simulations. It is well known that higher charged molecules in general have a higher electrostatic adsorption barrier because already adsorbed molecules charge the surface and repel further adsorbing ones from the interface[58]. Due to this the adsorption kinetics are slowed down. A similar result was observed for the dynamical surface tension of DAMP4. At pH 4 and pH 9 adsorption kinetics are slower than at pH 7 due to the intermolecular repulsion from the interface. This effect is not captured in MD, as only single molecule unfolding simulations were conducted. As such, the intermolecular influence of already adsorbed DAMP4 on newly approaching molecules in terms of speed cannot be assessed here.

5.4 Conclusion

In summary, no particular role for water molecules due to higher ordered structures around DAMP4 or the interface could be observed in the simulations. Further changes to the system corresponding to different interfacial scenarios revealed possible drivers for unfolding of DAMP4 at an air/water interface. The results showed that, despite the high thermal stability of DAMP4, the environmental
changes towards the interface likely trigger the conformational changes observed experimentally. This includes a change in DAMP4’s protonation state due to a pH gradient towards the interface, either acidic or basic direction, leading to an increased net charge of the molecule, which is further increased due to shielding of charges with $\text{H}_3\text{O}^+$ or $\text{OH}^-$ ions, respectively. In addition, with general fluctuations of the molecules and the right orientation of DAMP4 to the interface, these effects lead in combination to the starting of the unfolding process, with helices 1 and 4 which are facing upwards opening up and pushing the molecule apart. In the final arrangement at the interface, all four helices are in parallel, with the hydrophobic residues facing the air, and the hydrophilic residues the water phase, generating an optimal surface coverage and possibilities for intermolecular interactions.
5.5 References


Chapter 6  Discussion

Biosurfactants are becoming an increasingly popular alternative to petroleum-derived surfactants due to the progressing environmental awareness of producers and consumers. This thesis aimed to create a fundamental understanding of a promising family of four-helix bundle biosurfactant proteins, and in particular one member, DAMP4, by applying molecular dynamics simulation. These molecules are renewable, eco-friendly, non-toxic, and sustainable, and have the added advantage of allowing an infinite set of design opportunities. The simulation method gave a more detailed view on the links between molecular sequences and expression and purification behaviours, linked by the stability, and their behaviour at an air/water interface on the atomic scale. This is of particular interest as this family addresses several problems which are frequently associated with biosurfactants, and understanding the basics of their characteristics will enable the design of more and improved biosurfactants with similar features in the future.

6.1 Key research findings

This part will provide a further discussion and outline the key research findings of this thesis in respect to DAMP4 and the family of four-helix bundle biosurfactants. Key research findings will be underlined.

One of the main advantages of peptide or protein biosurfactants is their large flexibility and versatility. Proteins are macromolecules consisting of a variety of monomer building blocks, connected via peptide bonds. These building blocks are 20 naturally occurring amino acids. They consist of the backbone carbon, an amino group, a carboxyl group, and a side chain, which is the part that varies for every amino acid. The side chain can be polar or nonpolar, hydrophilic or hydrophobic, charged or uncharged. The amino acid sequence, the order and type of amino acids in a protein, determines its characteristics and three-dimensional structure[1].

This means protein biosurfactants offer a huge variety of design possibilities through the exploitation of the different characteristics of the amino acids. It allows them to be specifically tailored to a particular application by purposefully designing the amino acid side chains to meet the needs for this particular product, addressing a large variety of functions. In addition it enables the creation of particular interactions depending on environmental conditions, actively controlling the functionality. For example DAMP4 was designed to be switchable with a pH trigger. It allows its foaming properties to be actively turned on and off, simply by changing the pH by one unit. This is because of DAMP4’s charge structure. The foam is stabilized through electrostatic repulsion
between the charged side chains of the molecule, and if the pH is close to the isoelectric point, the net charge is zero and an electrostatically stabilized foam cannot be built[2]. Another example is the design of peptides to self-assemble into higher-dimensional structures for the fabrication of new biomaterials. This includes peptides or proteins self-assembling into nanofibers through ionic interactions which can then form scaffold hydrogels[3]. In these cases the proteins are specifically engineered to enable particular intermolecular interactions. However for the successful design of biosurfactants we need a better understanding of the fundamentals of these molecules, their properties, interactions, as well as how they relate to a particular structure or function. While the vast design opportunities give an endless variety of theoretical molecules, this might also result in problems. Protein interactions, folding, stability, and structure are complex, and combinations of effects cannot always be predicted, resulting for example in dysfunctional, misfolded, or unexpressable proteins. There is a need for further study of their physicochemical properties on the molecular level, in particular with focus on the structure-function relationships[4].

Hence, it is necessary to establish a more thorough fundamental understanding of this family of four-helix bundles, which allows them to be designed freely according to the desired application, while keeping the costs to a minimum to allow for industrial incorporation and competitiveness to chemical surfactants.

The aim of this thesis is depicted in Figure 6-1. It was to understand the link between the primary sequences of the molecules to the stability of the three-dimensional four-helix bundle fold, which in turn can be related to the expression outcome, the process behaviour, as well as the functionality by adsorption to a hydrophilic/hydrophobic interface. An increased understanding of how the sequence

---

**Figure 6-1 Flow chart of the aim of this thesis.**

The aim of this thesis is to establish the link between the biosurfactants’ sequences, their corresponding three-dimensional structures and their stabilities in order to make predictions about their expression outcomes, purification- and process behaviours, as well as functionalities.

Hence, it is necessary to establish a more thorough fundamental understanding of this family of four-helix bundles, which allows them to be designed freely according to the desired application, while keeping the costs to a minimum to allow for industrial incorporation and competitiveness to chemical surfactants.

The aim of this thesis is depicted in Figure 6-1. It was to understand the link between the primary sequences of the molecules to the stability of the three-dimensional four-helix bundle fold, which in turn can be related to the expression outcome, the process behaviour, as well as the functionality by adsorption to a hydrophilic/hydrophobic interface. An increased understanding of how the sequence
can govern these characteristics will allow the rational design of biosurfactants that can overcome three main problems identified by Banat and Marchant[5]:

- Low yield
- High production and purification cost
- No fundamental understanding

The approach used to establish these links was molecular dynamics simulation, as it gives a new view of the molecules in addition to experiments. It is a novel method, which might give new insights into the possibilities of protein biosurfactants.

**Exploration of the link between sequence and expression yield**

Four-helix bundles have been studied widely, as naturally occurring proteins such as Cytochrome C or Apoferrittin, or de-novo designed proteins[6]. This is because it is a simple, yet well-defined, fold with only one type of defined secondary structure. They can be easily designed by following a pattern of polar and nonpolar amino acids, providing the flexibility to choose these residues freely as long as this pattern motif is followed[7]. However when creating new biosurfactants from our family of four-helix bundles to meet a particular demand, the design of the primary sequence does not only affect its possible future function but also the ability to express them at high yield. As the cost needs to be minimized for biosurfactants to be economically viable, a high yield is desirable. Experimentally a new sequence needs to be cloned into a plasmid, transformed into *E. coli*, and overexpressed in the bacteria. This process is time-consuming and expensive, and there is no guarantee that it will be a success. In the best case the protein is expressed at high yield, in the worst case there is no accumulation of soluble protein. This may result from the degradation of the product by bacterial proteases, expression of the product in an insoluble or aggregated form, or poor expression resulting in an uneconomically low yield. In order to avoid this, we generated a simple tool, described in Chapter 3, which provides a reasonably quick check if a designed four-helix bundle is expected to be problematic upon overexpression in *E. coli* or not.

For this tool, we applied molecular dynamics simulation to predict experimental expression outcome in *E. coli* based on the primary sequence of the molecule by homology modelling it into the 3D structure.

We conducted 25 ns long high temperature MD simulations of ten four-helix bundle sequences to destabilize them, and used stability-related data from the simulations to create a statistical model with a support vector machine classifier. This approach was chosen as research suggests a relationship between the successful expression of a particular amino acid sequence, and the stability of the native structure of the protein in bulk[7-10]. The primary amino acid sequence plays an
important role for this, as it determines the molecular fold, as well as any possible amino acid interactions – stabilizing or destabilizing. Thereby, the substitution of a single amino acid can lead to a change in stability of the native structure strong enough to lead to no expression yield through weakening their ability to withstand proteolytic degradation[7]. This has been exploited for many computational overexpression prediction tools in the past, but it was yet to test if a method, such as molecular dynamics as used in this study, was capable of capturing these small changes in the sequences and relate them to a different expression outcome.

This study led to the key research finding that combining higher structural data and behaviour in a MD simulation with a statistical classifier can be useful to determine if a sequence is likely to be expressed at high yield in *E. coli*.

With stability data from triple runs of each of the ten four-helix bundle sequences and a test set of five additional sequences, we identified a prediction model which was capable of predicting the different expression outcomes for seven independent sequences (five test set sequences and two case study sequences), leading to 100% prediction accuracy. This was a great success, as previously developed prediction tools were not suitable for this family of four-helix bundles (Table 3-6). In a validation with five newly designed sequences three of them were predicted correctly. For the other two sequences, despite the fact that no accumulation of soluble protein was suggested by the model, soluble protein could be detected by SDS page. However this was at distinctly lower yield than the other molecules, and one of them at significantly higher molecular weight than expected, indicating problems during the expression process. These results show that the model is capable of identifying highly overexpressed molecules, and ruling out more problematic ones with lower to no yield. The support vector machine classifier has been found to be successful in classifying the data and creating the prediction model. The model used for future predictions was selected out of all possible models with two to five descriptors based on performance, simplicity, and ease of application.

The prediction model, which was able to best distinguish between soluble expression or not with a support vector machine classifier included the descriptors total energy, energy due to bonded interactions, as well as electrostatic energy, which was in good agreement with previous findings.

As the outcome of heterologous protein expression is a complex process which involves the combination of different effects and interactions, the use of a single descriptor was not possible, urging for a statistical classifier. One of the parameters selected for the classifier due to its strong descriptive power is electrostatic energy of the protein, which can be directly linked to the protein charge. The charge, in particular negative charges, has been the most frequently described characteristic which promotes high expression yield[11-14]. This is because the electrostatic repulsion between the residues leads to less or slower aggregation up to a certain point. After exceeding this limit the intramolecular repulsion leads to such a strong destabilization of the native
fold that the chances of degradation are high and overexpression is hampered. This is particularly the case when charges are clustered. This phenomenon could also be observed in the simulation, where a ‘successful expression electrostatic energy slot’ was seen, agreeing well with previous literature and experiments. The total energy can be seen as an approximation for the enthalpy of the molecules. The energy between directly covalently bonded atoms reflects for example unusual bond lengths or distorted geometries, both indirectly related to a stable, proper fold as required for an accumulation of soluble protein, matching the idea of the approach. These energy parameters were among others chosen as they represent values which represent an intrinsic energetic state. As such they equilibrate after some time, allowing for a shortening of the original simulations as confirmed in section 3.2.

A simulation time as short as 1 ns is sufficient to gather prediction input data that will allow us to predict successful expression. Even though the current trend goes to longer and more detailed simulations[15], we were able to drastically cut the simulation time from 25 ns to 1 ns. This will allow for faster conclusions for future predictions without the loss of predictive power. Thereby the predictions are still based on stability-related characteristics, but the protein is not forced to undergo unfolding or structural rearrangements to assess the stability. This is possible as the variables selected were all related to energies, which means they are an intrinsic state of the protein which does not constantly change, but rather oscillate around a value in accordance with small changes in temperature. The savings in computational time and resources in all future applications will be considerable and with time more data sets can be added to the training set to improve the predictive power further.

Through the model we demonstrated the successful link between the biosurfactants’ primary structure, to their theoretical 3D structure and the experimental expression outcome. The benefits of new computational science for modern recombinant biotechnology have been demonstrated, as well as the high potential to predict expression outcome by using higher-dimensional structural information in a simulation.

**Exploration of the link between sequence and stability and influence on the purification process**

A key advantage of DAMP4 and its variants is the possibility for a cost-effective purification, addressing another point of current problems of biosurfactants identified by Banat and Marchant[16]. As they are produced by microorganisms there is the need to separate the product from the producing host cells. As summarized in Figure 6-1 the second link investigated and established in this thesis is that from the primary sequence through the tertiary fold to the purification- and processing behaviour. Thereby, the stability plays a major role. The stability, and
in particular the thermal stability, is an important factor in biotechnological processes during production, purification, storage, and delivery[17]. A high stability can prevent product loss due to aggregation, and prolong the shelf-life of the product and/or its components[10]. The ability to withstand harsh process conditions is an advantage of biosurfactants over chemical surfactants, because if processes can be performed at elevated temperatures it can increase solubility and speed up particular reactions. In the case of DAMP4 and the other four-helix bundles, the high thermal stability is a major process advantage, enabling a cost-effective purification strategy through thermal treatment followed by solid/liquid separation[18]. DAMP4 is more stable than the average of proteins, including the bacterial contaminants, which denature upon heating above 90 °C. Because DAMP4 remains soluble it can be easily separated from the contaminants, such as bacterial proteins and cell debris, through cheap process steps like filtration or centrifugation. As biosurfactants are produced by microorganisms, either intracellularly or extracellularly, they always need to be separated from their microbial production host cells. This step is one of the main contributors to the overall cost, as expensive chromatographical steps might be necessary, which is not profitable for the usually low-cost industrial products surfactants are used in. The development of DAMP4’s purification process is therefore an important step toward decreasing biosurfactant production cost. As the high thermal stability is vital for the applicability of this process, Chapter 4 provided the experimental characterisation of DAMP4’s stability with respect to temperature, pH, and chemical denaturants. In addition, the sequential and structural fundamentals responsible were revealed through molecular dynamics simulation.

DAMP4 has a melting point of 122.4 °C, high thermal stability at all pHs, and moderate stability against chemical denaturation with GdmCl. Four-helix bundles consist of four helices, connected via three loops, and are easily designed by following a particular hydrophobic-hydrophilic pattern[6]. Despite folding into the same three-dimensional structure, there might not be a very high sequence homology[6, 19], which is why these molecules can vary considerably in their characteristics and offer large design versatility. It was described before that the four-helix bundle fold can be very stable, but DAMP4’s thermal stability is on an extraordinarily high level. It is in the region of proteins from hyperthermophilic bacteria, which grow at temperatures up to 113 °C and their proteins are adapted to these conditions. In vitro these proteins showed significant half-lives in the range of 90-100 °C[20]. It has been of interest for many researchers to reveal the mechanisms of these proteins in order to be able to adapt other proteins to be functional in these regions[20]. To find out why our molecule is that stable, a variety of different in silico stability variants were created for which three stabilization mechanisms were tested; salt bridges, the hydration shell, and the hydrophobic core, and their contributions to stability compared.
The simulations showed the hydrophobic core as most important factor for DAMP4’s stability; salt bridges and the hydration shell contribute in only a minor way to the overall stability. This means, the charges forming the salt bridges contribute to the interactions leading to DAMP4’s switchability, but are not necessarily needed to maintain the stability for purification purposes. These charges can be varied, for example to shift the pl resulting in switchability in a different pH range. The same is valid for the arginine residues related to the hydration shell. The finding of the importance of the hydrophobic core for four-helix bundles matches well with literature, where an optimal packing of the hydrophobic core was described as critical[19]. But how strong does the core have to be without loss of stability, and how much variation can we introduce there in new constructs? To answer this we tested another set of DAMP4 variants, with one to four hydrophilic mutations in the hydrophobic core at various locations. Three or more hydrophilic mutations in the core lead to a strong loss in stability; these mutations influence the unfolding pathway, and are stronger when mutations are clustered. In terms of design flexibility this means stability can be maintained even if one or two key hydrophobic core residues are changed. This kind of strategy, to purposefully destabilize the molecule, can be useful if a more flexible molecule is desired rather than a very rigid one. In nature, protein functions are often linked to the protein flexibility[21]. A less stable structure facilitates conformational changes, such as DAMP4 is undergoing upon adsorption to a hydrophilic/hydrophobic interface as described in Chapter 5. DAMP4 has to go through these changes to reveal its surface active characteristics and partition its hydrophobic and hydrophilic residues. This means, hydrophilic residues in the core can weaken the adsorption strength to an interface and cater for easier desorption. Networking with other molecules can be facilitated, in some cases proteins change their conformation as triggered by flexible regions by interaction with other proteins[21]. This may influence DAMP4’s adsorption behaviour as well, as already adsorbed molecules impact the adsorption of further surfactants. The strength of the hydrophobic core influences adsorption kinetics, as it induces a high barrier for unfolding and slows down adsorption speed. Staying within these stability limits, not more than three mutations in the core, should not affect the molecules’ stability to an extent where the heat purification process cannot be applied anymore. This is essential, as stronger loss of stability can result in problems upon overexpression as described earlier, as well as a decrease in melting point similar that of the proteins of the bacterial cell.

A delicate balance of the forces and interactions in the core has to be considered in new designs to create biosurfactants which are stable enough for high expression yield and purification, but highly surface active and functional with fast adsorption kinetics.
Exploration of how the sequence influences the physico-chemical properties of four-helix bundles and how these govern interfacial adsorption behaviour

Despite the high stability in bulk DAMP4 must be undergoing drastic conformational changes to reveal its surface active characteristics. Relating back to the sequence-structure-function relationship, Chapter 5 aimed at elucidating the processes happening upon adsorption to an air/water interface in atomic detail and explained with which mechanism unfolding is possible. Experiments suggested the unfolding of the molecules into a chain of four single helices, and a tertiary structure rearrangement is vital for their surface active properties. But what are the sequential features which allow DAMP4 to change its conformation so drastically and what does an interfacial conformation look like? This part of the research was impeded by the controversial literature about the physicochemical characteristics of an air/water interface. Briefly, is the interfacial area more acidic with an excess of hydronium ions or more basic with increased concentration of hydroxide ions compared to the bulk at neutral pH? As we did not want to pick either side, the challenge was to find out if there was a mechanism, which would be valid for both conditions, only driven by a change in environmental conditions rather than the specific conditions itself. The hydration shell around DAMP4 and ordered water molecules at the air/water interface as possible mechanisms to aid in the conformational rearrangement upon adsorption has been tested thoroughly, however no indication of a correlation between water and unfolding could be made based on the simulation.

MD simulations show no to little influence of the interface on the water structure under the conditions tested here. This does not necessarily exclude this phenomenon, but rather requires further testing with different software and force fields to see if the lack of structuring observed here can be attributed to the limitations of the method. Van Buuren et al. were investigating a decane/water interface under changing parameters and describe that a sharpening of the interface through parameter variation lead to formation of ordered water layers[22]. However they used different software and parameters, and it is unclear how these parameters influence an air (vacuum)/water interface in comparison to an oil/water interface, and how and if the structuring of the waters would have an influence on the molecules’ unfolding at all. Focusing on other phenomena described at the air/water interface, a shift in pH and an increased concentration of ions were successful in showing the unfolding pathway of DAMP4 as one example of the four-helix bundles. This pathway as well as the events leading to it delivers important fundamental knowledge on how and why these molecules are so effective and unique in their interfacial properties. DAMP4 unfolds at the air/water interface with the first and the fourth helix facing upwards and pushing apart; this leads to the opening of the hydrophobic core, and helices two and three following up to the interface. In the final conformation DAMP4 is located at the interface with all four helices in
parallel, the hydrophobic core residues facing towards the air, the hydrophilic surface residues oriented towards the bulk water. This time and energy-efficient pathway for four-helix bundles allows the changing of the tertiary structure without any loss in secondary structure at any time, consistent with experimental observations. It has to the best of our knowledge not been described in literature before. The knowledge of the interfacial conformation and simulations may help in creating particular intermolecular interactions which can lead to the actively controllable reversibility and switchability of the molecule’s function.

The conformational changes at the interface occur due to a combination of effects; including the right orientation towards the interface, internal fluctuations, an alteration of the protonation state due to the changing environment and a shielding of charges by oppositely charged ions as a natural trigger to speed up the unfolding process. This is a universal mechanism which works either way, if the interfacial environment is acidic or basic. In our simulations the mechanism indicated a preference for acidic conditions with increased hydronium concentration; however it is unlikely that this can be seen as evidence for the conditions found in real systems. This preference might be founded on the larger charge difference under these conditions, as well as the simulation setup. In agreement, previous simulations revealed a preference of hydronium ions for the interface compared to hydroxide ions[23]. The simulations showed one possible mechanism for the molecular drivers for an unfolding of the stable four-helix bundle DAMP4, however more studies are necessary to confirm these findings; eventually leading to a clearer understanding about the complex conditions at these interfaces.

6.2 Outlook

In Chapter 3 we presented a quick-check prediction tool for heterologous overexpression of four-helix bundle biosurfactant proteins in E. coli. This prediction model and its accuracy are restricted by the number of sequences included in the model building process. To confirm its validity further it will be beneficial to test any new four-helix bundle variant with the model prior to expression and further use their data to later improve the tool through the addition of more data sets that become available as research on these molecules continues.

The prediction model is so far limited to four-helix bundles, but this novel approach of using higher-dimensional structures in combination with the computer simulation and a machine learning algorithm showed promise to be advanced to other types of proteins in the future. Sources for information on other molecules’ sequences, structures, and expression outcome can be easily derived from literature, data bases[24] or from other researchers directly who are working in this field[7], as failed expression attempts may not always be reported. These proteins will have to be
simulated as previously described and the collected data sets used to build a new classifying prediction model. The more data sets are included in the model building process the more accurate the prediction with the new model will be.

Furthermore the principles of this general approach of using structural and behavioural data from the simulation to predict an outcome with a statistical classifier can be applied to different biological problems. It can be useful to test the therapeutic potential in any commercial high throughput pipeline. This can be for example in the establishment of a screening tool to predict the soluble expression of potential constructs for virus-like particles (VLPs). Therefore a particular data base for the wild type structure and existing constructs from a library can be used to generate a model based on the simulation of particular VLP-subunits. Molecular dynamics simulation has been successfully used to explain the failure of the expression of some VLP-constructs through the loss of a particular structure[25]; however a more generalised screening tool will be of interest for both, academia and pharmaceutical or biotechnological industry.

Experiments and simulations are a synergy, in which simulation results can be used to understand experimental outcomes and experiments can help validate simulations. To confirm the simulation results from Chapter 4 it would be beneficial to determine experimental melting points for all the in silico stability variants created. A simple validation would be to compare the stability ranking of the experimental results to the observations and conclusions made in the simulation. This includes variants for each mechanism, water shell, hydrophobic core, and salt bridges, as well as hydrophobic core variants. This will allow a control for the contributions to the stability of the mechanisms, but as well a check for the conclusions drawn for the purification heat process, namely if a particular variant is still stable enough to withstand the process. Another interesting aspect would be to investigate the adsorption behaviour of the hydrophobic variants, as we predicted different adsorption kinetics depending on the variation of hydrophobic core residues. This can be extended to desorption studies and structural investigations of the molecules after a foaming cycle, as it is currently not clear if desorption can be associated with a refolding process. This is an unanswered question as after unfolding in bulk DAMP4 seems to form strong aggregates due to the exposed hydrophobic residues, not showing signs of a refolding process in differential scanning calorimetry (DSC) measurements. Circular dichroism (CD) is a simple method to determine secondary structure, as is Fourier transform infrared spectroscopy (FTIR). Both of which could determine any refolding.

In Chapter 5 our simulations indicated that there was no enhanced, ordered water structure at the air/water interface. However, as structured water molecules are a previously described phenomenon at the air/water interface this might be an artefact of the simulation software and/or parameter. It would be useful to further investigate this with different software, force field, and water model to
see if more ordered structures can be observed, and how these influence the adsorption behaviour of DAMP4 as there is still an unknown factor influenced by the ions in our theory of potential drivers. Alternatively another similar hydrophilic/hydrophobic system could be tested. As van Buuren et al. used a decane/water interface and describe the phenomenon of ordered water layers this could be another suitable test environment, as DAMP4 should reveal the same adsorption behaviour to an oil/water interface[22]. In this chapter we elucidated the final arrangement of DAMP4 unfolded at the interface as “W-conformation”. With this knowledge it will be possible to simulate the interactions and processes of several four-helix bundle molecules adsorbed to an interface, as well under different environmental conditions such as the pH. This knowledge would help in the design of specific interactions within different molecules.

### 6.3 Conclusion

In summary, this thesis contributed to the knowledge and fundamentals of the design of four-helix bundle biosurfactant proteins which are promising candidates to compete with the low-cost chemical surfactants but with all the advantages of biosurfactants, including the vast design flexibility of proteins. This encompasses the molecular basis for a stable design which allows for cheap purification, as well as the pathway of interfacial adsorption behaviour of four-helix bundles and the responsible molecular and environmental drivers. In addition we presented a tool which enables a quick check for new protein designs on their overexpression success likelihood. With this knowledge future biosurfactant designs, which can be quickly adapted to an application, will be facilitated, ensuring high expression yield, low-cost purification, and functionality of the molecules. Thereby molecular dynamics simulation has been proven a useful and valuable tool to give new insights from the theoretical atomic perspective on the sequence-structure-function relationships of this unique group of molecules.
6.4 References

1. R. Winter, F. Noll, in Methoden Der Biophysikalischen Chemie; Springer-Verlag, 2013.
Appendix A: Supplementary Information for Chapter 5

Figure S-1  Average force in x-direction of five independent constrained unfolding simulations under different environmental conditions.