Structural and functional elucidation of the trimeric autotransporter adhesin BpaC from *Burkholderia pseudomallei*

by

Andreas Reinhard Kiessling

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Biomedical Sciences
&
Astbury Centre for Structural Molecular Biology

July 2022

Intellectual property and Publication Statement

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. The literature review (1) was used in chapter 1. The material for the research article (2) is mainly incorporated in chapters 4, 5, and 6. Only the material that was produced by myself was used in this thesis to avoid any misclaiming of work. The publications in question are:

(1) Kiessling, A. R.*, A. Malik* and A. Goldman (2019). "Recent advances in the understanding of trimeric autotransporter adhesins." Med Microbiol Immunol: 1-10.

This is a literature review in which I am joint first author (*). In this paper we discuss the recent finding in the field of trimeric autotransporter adhesins with a focus on translocation novelties, sequence-to-structure relationships and domain organisation. I was responsible for the main text of the manuscript while Anchal Malik provided the figures and figure legends. Adrian Goldman contributed with editorial remarks.

(2) Kiessling, A. R., S. A. Harris, K. M. Weimer, G. Wells and A. Goldman (2022). "The C-terminal head domain of *Burkholderia pseudomallei* BpaC has a striking hydrophilic core with an extensive solvent network." Mol Microbiol **118**(1-2): 77-91.

This research article, in which I am sole first author, provides an in-depth analysis of the structural model of the C-terminal head domain of BpaC. I performed all experiments related to the structure (construct design, expression/purification optimisation, crystallisation, structure solution) and was responsible for the manuscript outline, text, and figures. The comparison with other structures was also performed by myself. Kathleen Weimer provided the genetic analysis in relation to the phylogenetic trees and the BpaC homologue BoaC from *Burkholderia oklahomensis*. Her analysis was omited from this thesis as to make recognition of individual author contribution easier to identify. Sarah Harris and George Wells contributed with their molecular dynamics simulations of the structured solvent molecules within BpaC. Adrian Goldman provided extensive editorial input. The associated structure is deposited in the PDB database with the accession code 7O23.

Acknowledgements

As part of a transeuropean training network called ViBrANT, I was able to receive plenty of advice from supervisors and students alike. Special thanks in this regard goes to my ViBrANT specific supervisory committee consisting of Prof. Volkhard Kempf, Dr. Nadia Izadi-Pruneyre, and Dr. Alex van Belkum who provided me with an outside point of view on my project and helped me define my longterm career goals.

From my fellow students, the ones working on TAAs, namely Ina Meuskens, Diana Vaca, and Arno Thibau, were all incredibly helpful. We talked about the latest TAA research, the direction the field is heading, and their own workarounds for challenging problems I encountered. All-in-all, the whole ViBrANT family made me feel part of a bigger picture and I am very grateful for that.

I also received a lot of support from several different supervisors in Leeds, owing to the different projects I participated in over my whole time in Leeds. From Prof. Peter Henderson, I learnt how diverse the field of microbial transporter can be and a newfound respect for working with radioactive material. Prof. Lars Jeuken helped me stay on track of my project, even though I tried really hard to come up with a thousand different ideas, that would have probably taken the rest of my life to explore.

A more “pastoral” support was given by Pirjo Johnson, our ViBrANT project officer, whom I shared many conversations with, mostly about non-academic related issues, and who has become a good friend of mine over the years. I will remember our little chats, in the corridor next to the printer, fondly.

Astbury 6 would have been a very boring place if not for all the different people that I had the pleasure to interact with over the years. A full list would cover a good few pages and it would risk

to dilute the individual contribution of the people who helped me most in my project and in general getting through PhD life. I had the most engaging and fun conversations about science and politics with Dr. Jack Wright that kept my spark for science going even in the darkest hours of my PhD. From Jessica Boakes I learnt resilience and the importance of respect and professionalism in the workplace, for which I am very grateful. From Dr. Eoin Leen I received valuable insights into the workings of what the Irish think of the British, helping me overcome my own cultural barriers to better integrate myself into the intricate ecosystem that is Astbury 6. Sharing a project with Maria Nikolova has taught me so many things, especially about professional communication both in terms of science but also how to be friends, yet respecting each others boundaries and needs in a way that benefits the project goal in a symbiotic way. Although LCP requires too much dexterity for my taste, Dr. Claudia Zilian managed to teach me the magic, alleviating my initial anxiety I had for this technique. I hope I can keep this confidence for my future research projects and thank her for that. Dr. Jannik Strauß helped me to remember the German way and that the oddities I encountered over the years are not only odd to me, making me feel a little bit less alien in the UK. A very big thank you goes to Dr. Stephen Muench, who was at my side from the moment I set foot in the UK in February 2018 (back then for my German MSc dissertation) and who remains a steady source of support, both academically and on a personal level, that made my time in the UK all so much easier.

I owe Prof. Adrian Goldman a great deal of gratitude as he had a great influence in my growth as a scientist. I can confidently say that I am a different person to the one that started in his lab in November 2018 and there were a lot of ups and downs that he guided me through, even before the initial Covid-19 lockdowns. I am certain, that I will never forget this time of my life.

Abstract

The trimeric autotransporter adhesin (TAA) BpaC plays a central role in the initial infection stages of the Gram-negative *Burkholderia pseudomallei*. The actual function of this protein on a biochemical level was undetermined at the start of this project. I endeavoured to identify binding partners of BpaC and associate the individual domains of the solvent-accessible part of the protein with specific binding events. To this end, I analysed the protein sequence using a combination of bioinformatic tools and TAA specific sequence rules to accurately determine secondary structure prevalences, domain borders, and likely binding partners. I designed, expressed, and purified multiple constructs of various lengths and composition, that all together cover the full solvent-accessible domain of BpaC, making every part of the protein accessible for functional or structural experiments. On a structural level, I solved parts of the C-terminal head domain of BpaC which, by extension, can be used to deduce the structural model of the full head domain. This model provided the basis for the introduction of a new subcategory of left-handed parallel β-rolls named BpaC-like head domains; based on the unique surface charge properties of the C-terminal located head domain of BpaC.

On a functional level, the identification of likely homology models for the individual domains of BpaC resulted in the prediction of binding partners of BpaC. Some of the candidates, namely extracellular matrix proteins from humans like fibronectin and collagen, were used in binding assays. These experiments revealed a preference of BpaC to bind to fibronectin, collagen I and collagen II. However, a clear domain-ligand association could not be determined. Lastly, a new hypothetical infection model is provided in which BpaC acts together with type I pili in the initial steps of adhesion of *Burkholderia pseudomallei* in a three stage process.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AHTC: Anhydrotetracycline

a.u.: Arbitrary unit

BAM: β-barrel assembly machine

BSA: Bovine serume albumine

CV: Column volume

dsDNA: Double-stranded deoxyribonucleic acid

ESP: Extended signal peptide

gDNA: Genomic deoxyribonucleic acid

GuHCl: Guanidinium chloride

HRP: Horseradish peroxidase

HRV 3C: Human rhinovirus 3C

IMAC: Immobilised metal affinity chromatography

IPTG: Isopropyl β-D-1-thiogalactopyranoside

kb: Kilobase

PDB: Protein Data Bank

PNK: Polynucleotide kinase

LB: Lysogeny broth

LPBR: Left-handed parallel β-roll

LPS: Lipopolysaccharide

T_m: Melting temperature

MES: 2-(N-morpholino)ethanesulfonic acid

mNG: mNeonGreen

NaP_i: Sodium phosphate

ORF: Open reading frame

P: Pellet (in context of cell lysis)

PCR: Polymerase chain reaction

POI: Protein-of-interest

PVDF: Polyvinylidene difluoride

RADAR: Rapid automatic detection and alignment of repeats (program, special abbreviation)

r.m.s.d.: Root-mean-square deviation

RT: Room temperature

SDS-PAGE: Sodium dodecyl sulfate–polyacrylamide gel electrophoresis

sfGFP: Superfolder green fluorescent protein

SN: Supernatant

SOC: Super optimal broth with catabolites repression

TAA: Trimeric autotransporter adhesin

TB: Terrific broth

UV280: Absorbance measured at a wavelength of 280 nm

1 Chapter 1: Introduction

1.1 Emerging resistance of Gram-negative bacteria to antibiotics

The ongoing Covid-19 pandemic has shown the world that pathogens are very much a continuous threat even with all the medical advancements and knowledge we gained over the last hundred years. While the world is slowly recovering from the effects of this pathogen, another crisis is already on the horizon: the antimicrobial resistance crisis.

Antimicrobial resistance not only encompasses the resistance to antibiotics, but also to antivirals, antifungals and antiparasitics. All of these put an immense burden on the public healthcare systems over the world (Prestinaci, Pezzotti et al. 2015). The biggest contributor to this crisis however is the multidrug-resistant bacteria (Davies and Davies 2010). An estimated 4.95 million deaths were associated globally with bacterial antimicrobial resistance in 2019 alone, showing the significance of this particular aspect of the antimicrobial resistance crisis (Murray, Ikuta et al. 2022).

In its simplest explanation, globalisation is one of the main drivers of both the Covid-19 pandemic and the antimicrobial resistance crisis. Both are worsened by the overpopulation and the increased global migration of the 21st century. A significant factor for the rise of antibacterial resistance is the increased use and more accurately misuse of antibiotics in clinical and agricultural settings (Singer, Shaw et al. 2016).

Advances to tackle the crisis by developing new antibiotics have been minimal at best. This is mainly due to high costs and low benefits of antibiotic development for pharmaceutical companies: an estimate (from 2017) for the cost of development is gauged at around US$ 1.5 billion with only an average revenue generated (in 2020) of roughly US$46 million per year (Plackett 2020).

Antibiotic courses typically last 7-14 days which is nothing compared to multi-year drug treatments for cancer or immunity-affecting diseases.

1.1.1 The importance of biofilms and adhesins in the bacterial infection cycle

Adhesion and aggregation are the first steps in bacterial pathogenicity and biofilm formation (Dunne 2002, Fux, Costerton et al. 2005). Biofilms have been proven to reduce the penetration of antibiotics significantly (Ribeiro, Felicio et al. 2016) and are responsible for the majority of chronic infections (Burmolle, Thomsen et al. 2010). This shows the significance of pathogenic proteins which are involved in the biofilm generating process. Making these bacterial populations susceptible again to certain antimicrobial agents can be achieved by targeting these proteins on a biochemical level for example by developing agents that block the binding interface, thereby preventing biofilm formation.

This thesis aims to elucidate the role of an adhesin molecule which is crucial for the pathogenic cycle of the Gram-negative bacterium *Burkholderia pseudomallei* (*B. pseudomallei)*. Understanding the individual steps in the adhesion process of *B. pseudomallei* creates opportunities to biochemically intervene at susceptible vantage points as to the necessity of this step for successful host infiltration. This would provide a useful alternative approach to treating bacterial infections, which potentially can be replicated for other species with a similar invasion mechanism.

1.2 *Burkholderia pseudomallei* as cause of melioidosis

B. pseudomallei is a Gram-negative bacterium which can be mainly found in (sub)-tropical soils and waters with most cases of infections reported in Southeast Asia, India, South America, and Northern Australia (**Figure 1**). The severity of infection can range from acute to chronic disease states. It has a high mortality rate in endemic regions with estimated melioidosis death of 89,000 in 2015 worldwide (Limmathurotsakul, Golding et al. 2016).

B. pseudomallei causes melioidosis, which is a febrile illness with acute and chronic states manifesting itself in either an asymptomatic course of sickness or with abscesses, chronic pneumonia mimicking tuberculosis, localized skin ulcers, and septic shock (Currie 2010). Infection in humans occurs via cutaneous or aerosol routes upon contact with infected animals (mainly horses) or soil (Wiersinga, van der Poll et al. 2006). It can invade a variety of epithelial cell lines and even survive in macrophage-like cells (Kespichayawattana, Intachote et al. 2004). As an intracellular pathogen, it undergoes an extensive escape from phagocytic digestion using a Type III secretion system (Gong, Cullinane et al. 2011). Once free, *B. pseudomallei* can spread through intercellular fusion using a Type VI secretion system thereby evading immune recognition (Burtnick, Brett et al. 2011). The role of the Type V secretion systems in this infection cycle, to which trimeric autotransporter adhesins (TAAs) belong, have also been studied on a systemic level (Lazar Adler, Stevens et al. 2015): they are mainly implicated in the adhesion and immune evasion process of *B. pseudomallei* in the initial steps of the host cell invasion. Unfortunately, a lack of biochemical information about the exact binding partners for this class of proteins leaves a gap in knowledge which could be exploited to interfere in this specific part of the invasion process. A closer look at individual proteins and their most likely binding partners *in vitro* is necessary to overcome this gap and provide a starting point for successful drug intervention.

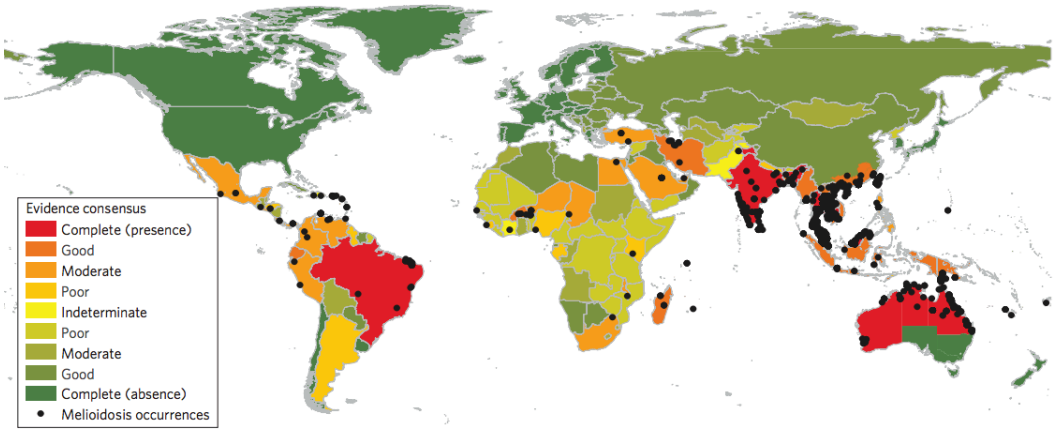


Figure 1 Occurrence of global melioidosis cases and evidence consensus on distribution data from 1910 to 2014 – Colouring represents range of consensus of *B. pseudomallei* presence from complete absence (dark green) to complete consensus (red). Strong association of *B. pseudomallei* occurrence with high rainfall, high temperature, and soil that is either rich in clay or modified heavily by human agriculture. Image obtained from (Limmathurotsakul, Golding et al. 2016) and reprinted by permission from Springer Nature under license number 5347151053987.

1.3 BpaC as research target

Individual virulence factors in *B. pseudomallei*, which can be associated with adhesion and/or immune evasion processes, were studied using knockout mutants for each protein. Three different categories were tested: the virulence in mouse infection models using intra-peritoneal doses of the pathogen, the capability of these knockout mutants to replicate in macrophage-like cells, and the capacity to survive in normal human serum (Lazar Adler, Stevens et al. 2015). The virulence factors studied, namely BoaA/B, BpaA/C/D/E, and BimA, all had different impacts in these three categories (**Figure 2**). Remarkably only one of these virulence factors, the trimeric autotransporter adhesin BpaC, was shown to be involved in all three processes, indicating a crucial role in the pathogenic cycle of *B. pseudomallei*.

Another study looked at the adherence behaviour of BpaC to human lung cell lines like A549, HEp2, and NHBE (Normal Human Bronchial Epithelial): expression of *bpaC* in *E. coli* led to a significant increase in adherence to these cell lines, the largest effect on NHBE with a five-to-sevenfold increase in adherence (Lafontaine, Balder et al. 2014). Focussing on the knockout effects of *bpaC* in *B. pseudomallei* on the adherence to the described cell lines showed that the adherence level remains intact for both A549 and HEp2, but significantly reduced for the NHBE experiment. The authors then tested the *bpaC* knockout strain of *B. pseudomallei* in an aerosol challenge experiment with mouse models and found that the virulence of the pathogen remains (Lafontaine, Balder et al. 2014). However, the author themselves describe methodological challenges, which can limit the impact of their findings: this is mainly referring to the differences in their infection model (plate grown bacteria, creation of the *bpaC* insertion via allelic exchange) to an alternative model used by another group showing different results in A549 cell binding of BpaC (Campos, Byrd et al. 2013). They suggest to use multiple infection models at the same time for future

experiments. It is clear that further studies are needed to verify the results with a much more diverse experimental setup, as outlined by the authors of Lafontaine et al. (2014), to fully understand the impact of the adhesin molecule BpaC. The very nature of an adhesion molecule can cause experimental issues as it is hard to distinguish from non-specific adhesion properties.

While BpaC has implications for various steps in the infection process of *B. pseudomallei*, other autotransporters also play a role in adhesion and internalisation and it is likely to be a combined effect which is based on redundancy to compensate for evolutionary or environmental changes in the survival of *B. pseudomallei* (Campos, Byrd et al. 2013).

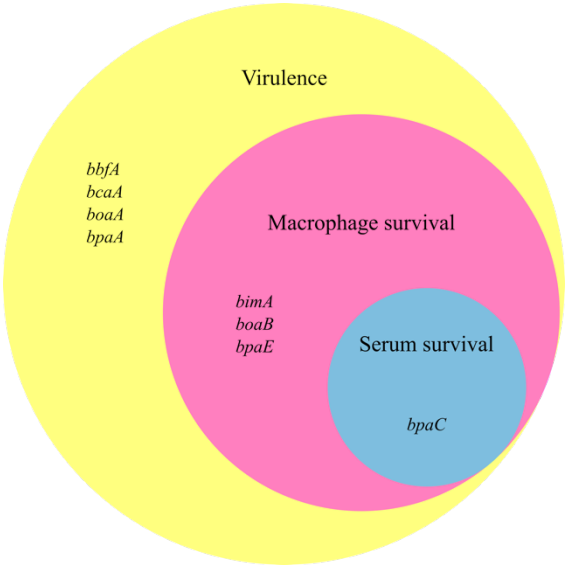


Figure 2 Venn diagram of phenotypic results of *B. pseudomallei* autotransporter influence in mouse melioidosis model – Most autotransporters of *B. pseudomallei* tested in a BALB/c intra-peritoneal mouse melioidosis model were shown to be important for virulence and/or intracellular replication in J774.2 macrophage-like cells, but only *bpaC* was also having a significant effect in survival in 45% (v/v) NHS. Gene names of autotransporters are shown in italics. Results from (Lazar Adler, Stevens et al. 2015).

1.4 Secretion systems in Gram-negative bacteria

An important attack mechanism of bacteria are virulence factors like toxins or other secreted proteins that get exported by specialised secretion systems from the cytosol of the bacterium. These secreted proteins can play many roles in promoting bacterial virulence: from enhancing attachment to eukaryotic cells, to scavenging resources in an environmental niche, to directly manipulating target cells and disrupting their functions (reviewed in (Green and Mecsas 2016)). Targeting these factors would offer an alternative, less extreme treatment pathway by halting the attack long enough for the immune system of the host to better deal with the bacterial challenge (Allen, Popat et al. 2014).

Secretion systems in Gram-negative bacteria are classified by their different structures, functions, complexity, and specificity in host efficacy (**Figure 3**). The protein complexes can either span over one or two membranes of the pathogen itself or three membranes; with the third one being the target host cell membrane. Some of them are conserved throughout all bacteria; others have a specific role to fulfil in the successful pathogenic cycle of a bacterium (reviewed in (Costa, Felisberto-Rodrigues et al. 2015)). A special interest for this thesis is the type V secretion systems (Va-Vf) which play an important role in the initial steps of host infiltration (Dunne 2002) which will be the family of secretion system that we will focus on in this chapter.

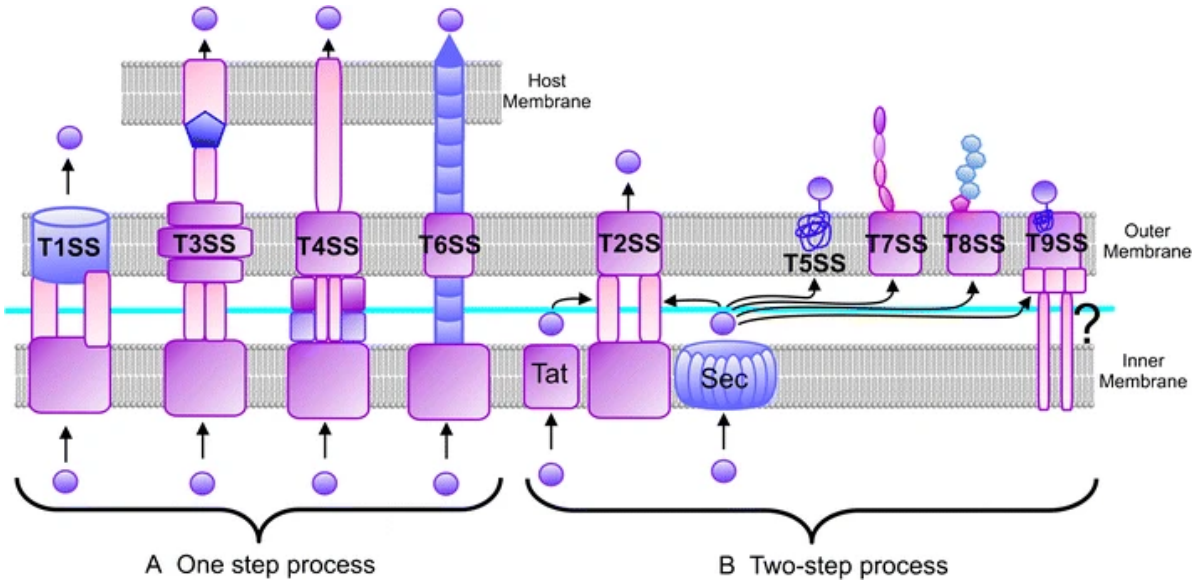


Figure 3 Overview of Gram-negative secretion systems – Secretion systems (SS) mainly transport effector proteins to the outside of the pathogen or the target cell directly. These systems can either span across one, two or even three membranes. They can be separated into one-step processes with the secretion apparatus spanning over multiple membranes (**A** - Type 1, 3, 4, and 6) or two-step processes which require the translocation of the effector protein via the inner membrane first before insertion into the outer membrane of the pathogen (**B** – Type 2, 5, 7, 8, and 9). The translocation via the inner membrane requires either the Sec translocon or the twin arginine transportation pathway (Tat). Image obtained from (Bocian-Ostrzycka, Grzeszczuk et al. 2017) distributed under an Open Access Creative Commons License by Springer Nature.

1.4.1 Overview of type V secretion systems

The type V secretion systems are often referred to as autotransporters and are part of the secretion superfamily of Gram-negative bacteria. The name originated from the initial belief of a self-sufficient secretion mechanism (Klauser, Pohlner et al. 1993) independent of other factors like chaperones or protein complexes. Later it was discovered that actually multiple systems are involved in the translocation of autotransporters: the Sec translocon, situated in the inner membrane and responsible for translocation from the cytosol to the periplasm (Leo, Grin et al. 2012), the translocation assembly machinery (Selkrig, Mosbahi et al. 2012) serving as a bridge between inner and outer membrane in a few subsystems like the classic autotransporters Va, and most importantly the β-barrel assembly machinery (BAM, (Albenne and Ieva 2017)) a highly conserved mechanism for outer membrane insertion of β-barrel containing proteins to which the type V secretion system belongs to.

The subclasses inside the type V family (Va-Vf, reviewed in (Meuskens, Saragliadis et al. 2019)) are distinguished by difference in orientation, configuration, and attachment of the passenger domain to the β-barrel domain (**Figure 4**). The first subclass described here is the so-called classic autotransporter (Va), which consists of a single polypeptide chain with a C-terminal, 12-stranded β-barrel and an N-terminal effector domain (also referred to as passenger) that usually exhibits a proteolytic function or has a self-aggregation activity. The type Vb two-partner secretion system is composed of two separate proteins, one a 16-stranded, OM integral β-barrel protein, the other released into the extracellular space without need for proteolytic cleavage from the barrel bound protein. Three monomers of the type Vc subclass, also known as trimeric autotransporter adhesins, combine to a 12-stranded β-barrel like the classical, monomeric autotransporters but have a completely different structural fold for the passenger domain due to their trimeric nature. The type

Vd autotransporters are described as a hybrid of Va and Vb with a 16-stranded β-barrel and the passenger domain being limited to only function as lipases or esterases, unlike the variety of functions that the Va passenger domain can adopt. Lastly, the type Ve inverse autotransporters, as the name implies, have a 12-stranded β-barrel on the N-terminal end of the protein rather than the C-terminal end. Type Vf autotransporters are excluded from this list as they are a recent addition to the type V family and more research is needed to accurately associate them with a certain secretion family.

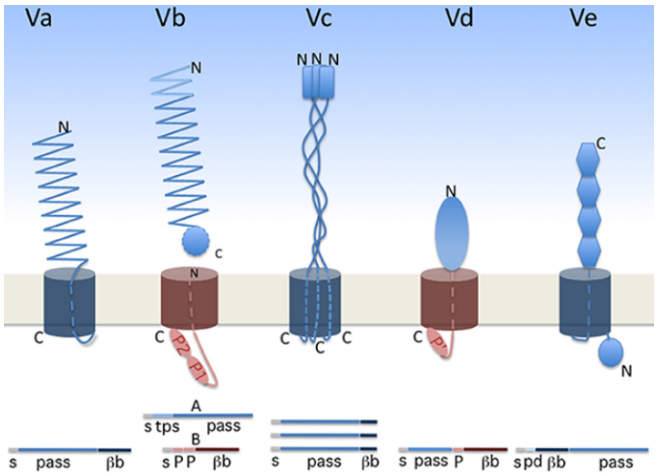


Figure 4 Overview of autotransporter family in Gram-negative bacteria – All β-barrel domains (βb) serve as a platform for passenger domain (pass) translocation and/or attachment to the outer membrane. The signal peptide (s) is recognized by the Sec machinery and the protein is translocated across the inner membrane. The classical autotransporter (Va) are monomeric, while TAAs (Vc) are trimeric. POTRA domains (P) are contained in the periplasmic side of Vb and Vc β-barrels which serve as protein-protein interaction platform. The Vb passenger domain with a TPS domain gets cleaved off after translocation. The position of the β-barrel domain in the sequence of Ve autotransporter is inverted compared to Va autotransporter, hence the name inverse autotransporter. Figure obtained from (Guerin, Bigot et al. 2017) and distributed under an Open Access Creative Commons License by Frontiers Media.

Autotransporters carry out several important functions during the invasion of a host: these can be roughly split into adhesion-related activities or immune-evasive properties (**Figure 5**, (Meuskens, Saragliadis et al. 2019)). Autotransporters can bind various components of the immune system like antibodies or parts of the complement system in order to protect themselves from the effect of the immune system. The adhesion properties of autotransporters play a vital role in the initial steps of the host cell invasion and can also direct the bacterial cell towards a certain cell type. Biofilm formation is another vital survival feature that autotransporters contribute towards. In essence, autotransporters are a highly diverse class of virulence factors for which we still require much a more detailed understanding if we are to understand pathogenicity and develop new antimicrobial strategies.

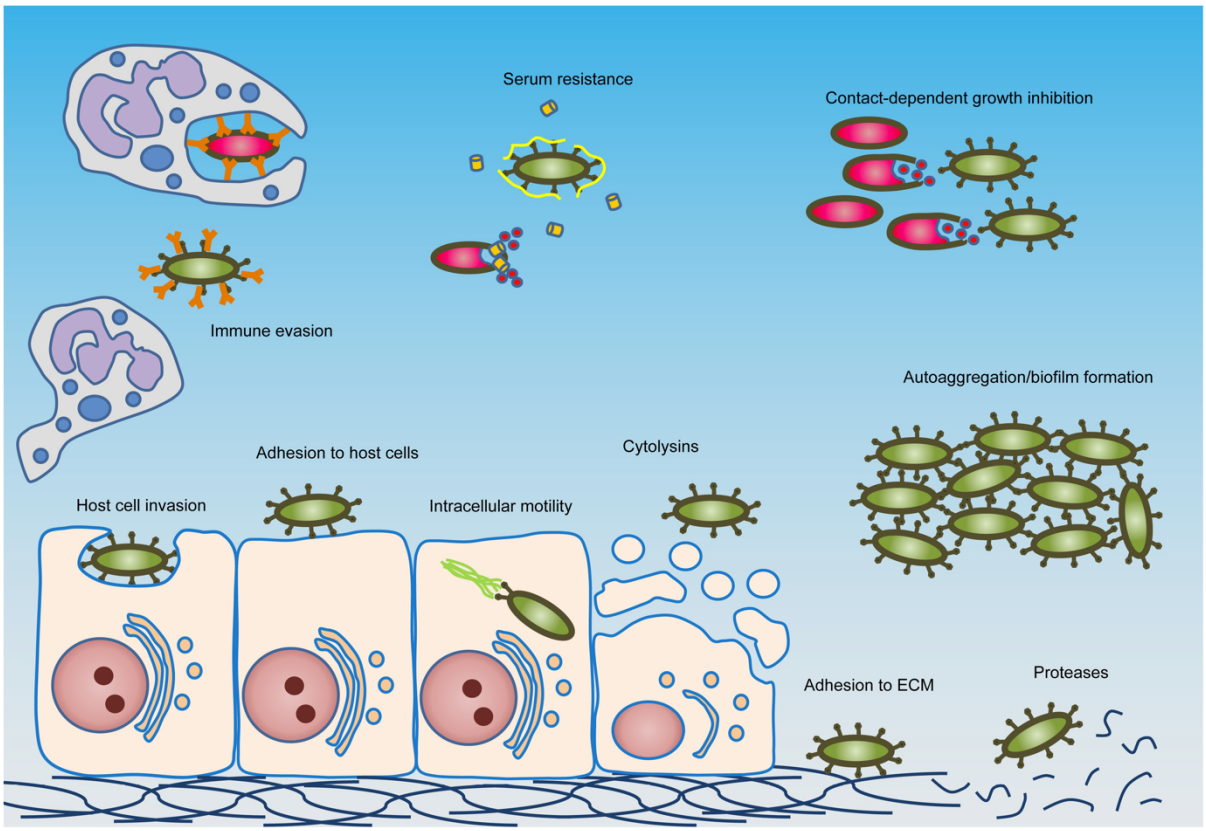


Figure 5 Overview of autotransporter functions – Schematic visualisation of autotransporter activities. Bacteria expressing autotransporters (green) possess several survival advantages over bacteria not expressing autotransporters (red): they can evade the immune system by binding antibodies and preventing opsonisation, are protected against degradation by the complement system (serum resistance), and control their growth via contact dependent inhibition. Autotransporter also are involved in biofilm formation and autoaggregation of bacterial clusters which provide another survival benefit. In terms of host interaction, autotransporter expressing bacteria can adhere either directly to the host cell or to the extracellular matrix. They can then invade the host cell and contribute to intracellular motility. Autotransporters are also involved in remodulation of their environment either by cytolytic or proteolytic activities. Figure obtained from (Meuskens, Saragliadis et al. 2019) and distributed under an Open Access Creative Commons License by Frontiers Media.

1.4.2 The type Vc subclass of trimeric autotransporter adhesins

Trimeric autotransporter adhesins (TAAs) are obligate trimers with an N-terminal extended signal peptide, a solvent-exposed passenger domain consisting of structural motifs either made up of coiled coils or β-sheet rich areas, and a C-terminal barrel domain, consisting of a β-barrel with 3 x 4 β-sheets and a coiled coil connecting the barrel to the passenger domain.

What all TAAs have in common is the highly conserved barrel domain, with the first atomic model of a β-barrel structure of the TAA Hia from *Haemophilus influenzae* solved in 2008 by Meng et al. using X-Ray crystallography (PDB: 3EMO, (Meng, St Geme III et al. 2008)). Another barrel structure was solved in 2012, this time using solid-state NMR: The structure is of the prototypical TAA YadA from *Yersinia enterocolitica* and shows the 12 β-sheets that make up the barrel domain and the connecting coiled coil residing within the barrel (**Figure 6**, (Shahid, Bardiaux et al. 2012)). TAAs are involved in many pathogenic functions like adherence (Bullard, Lipski et al. 2007), invasion (Capecchi, Adu-Bobie et al. 2005), serum resistance (Attia, Lafontaine et al. 2005), and biofilm formation (Valle, Mabbett et al. 2008). Adherence to epithelial cells is a crucial first step in the invasion of mammals, while serum resistance provides long-term survival of bacterial populations (Sandt and Hill 2001). The actual binding partners of TAAs can range from members of the extracellular matrix like collagen, fibronectin, or vitronectin (Vaca, Thibau et al. 2020) to components from the complement system like C4b-binding protein (Hovingh, van den Broek et al. 2016), to specific receptor binding events like the TAA UspA1 interacting with the human carcinoembryonic antigen-related cell adhesion molecule 1 (Conners, Hill et al. 2008). A particular challenge in this regard is the assignment of a specific binding event to a certain region of the protein, as TAAs have a general tendency to auto-aggregate and stick to even abiotic surfaces (Ishikawa, Nakatani et al. 2012).

The size of TAAs can vary dramatically from about 500 residues on the lower end of the spectrum to about 4000 residues on the upper end. The most abundant size populations are around 500 and 1500 residues (Bassler, Hernandez Alvarez et al. 2015). Translating number of residues into actual size of TAAs can be difficult to quantify as some structural motifs take up more space than others. However, negative stain electron images for two TAAs from both ends of the spectrum provide a good estimate of the total range of TAA real length distribution: for YadA with 422 residues, one can estimate about 20 nm of length from the images, while BadA with a total of 3973 residues (Thibau, Hipp et al. 2022) has an approximate length of 240 nm (**Figure 7**, (Linke, Riess et al. 2006)). Due to this accessibility on the bacterial surface and inherent immunogenicity, TAAs provide suitable vaccine targets (Thibau, Dichter et al. 2020) which is a great alternative to treating bacterial infections other than via the classic antibiotic route. Small molecule inhibitors are also being explored as an alternative treatment option, but are in the early stages of TAA research (Saragliadis and Linke 2019). However, antibiotics are still superior in terms of availability, cost, and convenience with the actual real-world impact of TAAs as vaccine or small molecule inhibitor targets yet to be determined.

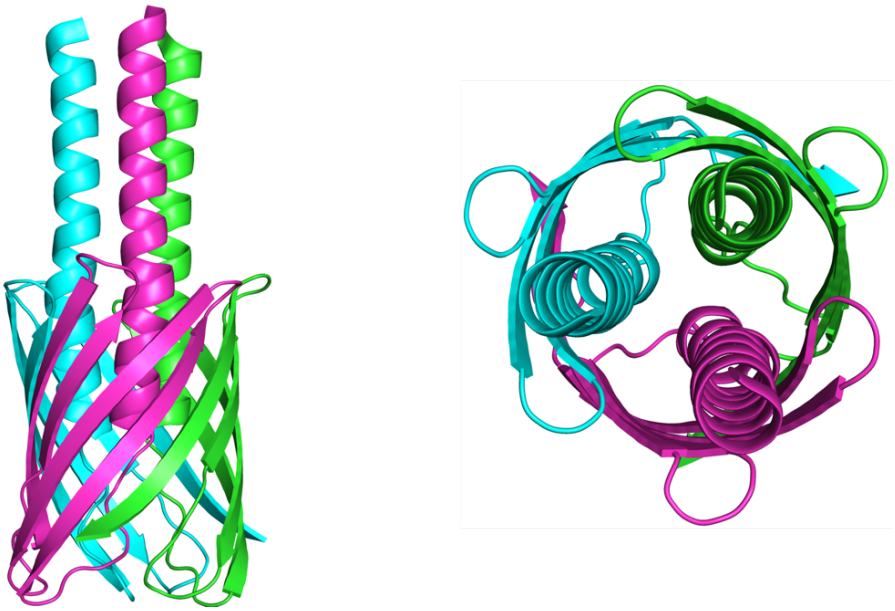


Figure 6 Solid-state NMR structure of YadA barrel domain – Cartoon representation of the barrel domain of YadA with parts of the stalk domain; as sideview (left) and topview (right). Each chain is coloured differently to highlight the contribution that each chain provides in the full assembly of the β-barrel. PDB code for the model is 2LME (Shahid, Bardiaux et al. 2012).

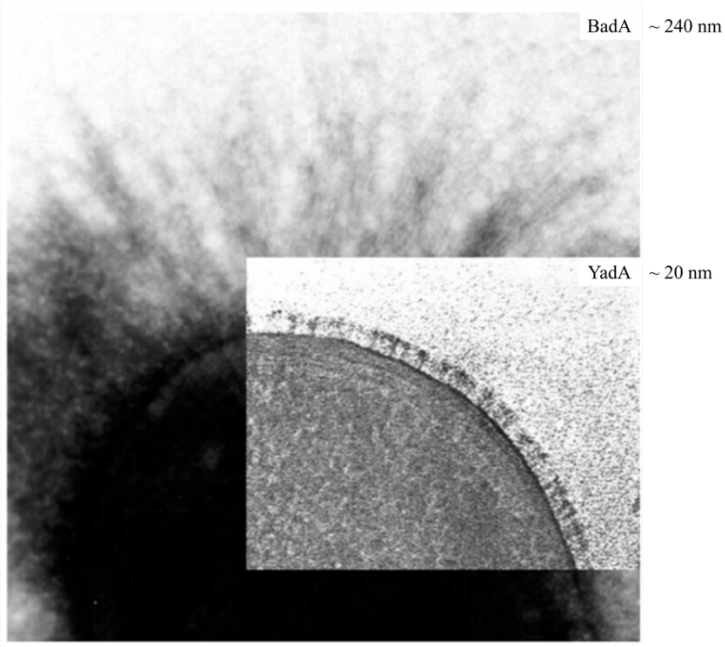


Figure 7 Real size estimation of TAAs with examples from both ends of the spectrum – Negative stain electron microscopy images of bacteria expressing YadA and BadA enable a rough size estimation for both TAAs. This can be used to set a benchmark for the lower and higher end of the TAA size spectrum (Linke, Riess et al. 2006). YadA with 422 residues (Hoiczyk, Roggenkamp et al. 2000) is one of the smallest TAAs at about 20 nm while BadA with 3973 residues (Thibau, Hipp et al. 2022) is ~240 nm and belongs to the largest TAAs that have been described in the literature. Figure obtained from (Linke, Riess et al. 2006) and reprinted by permission from Elsevier under license number 5347251321197.

1.5 Modular structures of TAAs

The comparison of two homologous proteins is usually performed via sequence alignment with programs like Clustal Omega or MUSCLE. This approach cannot be applied to TAAs as the only sequence-derived conserved region is the C-terminal anchor domain, which consists of the β-barrel domain and part of an anchoring coiled coil element (Dautin and Bernstein 2007). The trimeric nature of TAAs however narrows down the available structural motifs that individual TAA domains can adopt: the majority of the structural folds are either highly intertwined structures of β-sheet rich head domains or coiled coil containing stalk domains (Cotter, Surana et al. 2006).

These structurally imposed restraints have a major benefit: structural motifs keep reappearing in TAAs even if they have diverging sequences. One prominent example of this are the head domains of Hia and BadA, which have a near identical fold despite having no discernible sequence similarity (Szczesny, Linke et al. 2008). The order of the domains can also be switched around, almost like a set of molecular building blocks, which can again be seen for the head domains of BadA and Hia (**Figure 8**). The sequence alignment for the GIN motif also shows how the motif got its name and how it sometimes can deviate from the classical G-I-N sequence.

Some of these motifs can have highly conserved short sequence stretches that, almost like an identifier, provide a high confidence prediction for what structural motif one can expect in this area. Essentially, one can just "read" the sequence of the passenger domain without any *a priori* information about the structure of the protein to identify any of these conserved motifs. An overview of most of these motifs is given in **Table 1**. In these motifs, some names can refer to a stretch of outstanding residues in the sequence (in one-letter code), that can be found within the structural motif itself, but also can be misleading as for example the FGG motif actually reads L-G-G in the given example model (PDB: 2YO2, (Hartmann, Grin et al. 2012)). Some other names

for motifs (like Saddle) resemble the location or structural function of the motif in the 3D arrangement of the model. Other names are simply referring to the first TAA that motif was identified in (YadA-like head domain, Ylhead).

The unique connection between structure and sequence is visualised in **Figure 9** using structural examples for common head and neck motifs. This close association led to the development of the bioinformatic tool daTAA (Szczesny and Lupas 2008), that treats domains of TAAs like a dictionary: short motifs or features are identified and compared to structures of representative examples for each domain. These can then be assembled into long fibers using available structural models of connector motifs to give a preliminary homology model with an above average confidence level of prediction quality (Bassler, Hernandez Alvarez et al. 2015).

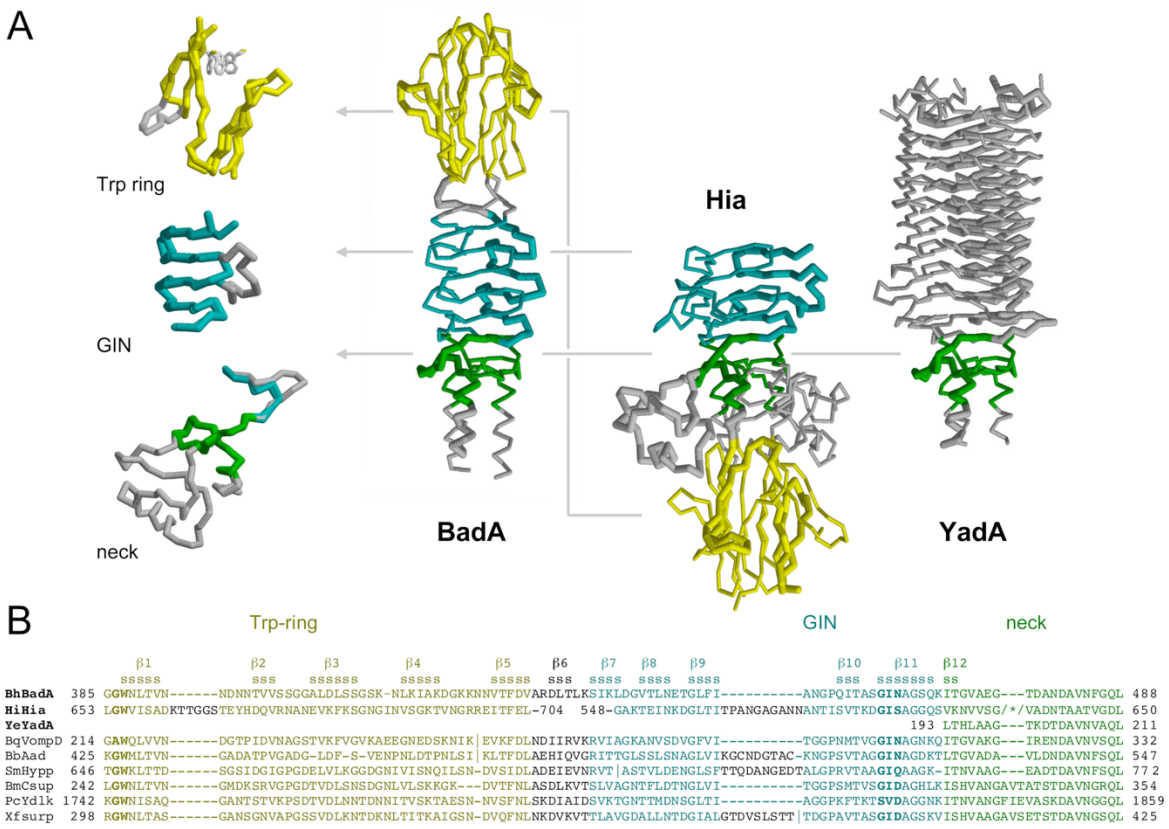


Figure 8 Reappearing structural motifs in three separate TAAs – **A** Superimpositions of the three motifs found in three separate TAAs is shown. TrpRing motif (yellow) can be found in BadA (PDB: 3D9X) and Hia (PDB: 1S7M, (Yeo, Cotter et al. 2004)), while the neck motif (green) is present in all three TAAs displayed (YadA, PDB: 1P9H, (Nummelin, Merckel et al. 2004)). **B** Sequence alignment showing the similarities between the sequences and how common the TrpRing-GIN-neck motif order is in other TAAs compared to the GIN-neck-TrpRing order in Hia. This can generally be useful as the identification of one motif can indicate the presence of the other motif in this particular example for a head domain arrangement. Figure obtained from (Szczesny, Linke et al. 2008) and distributed under an Open Access Creative Commons License by Frontiers Media.

Table 1 Overview of common head and neck domains and motifs.

Name	Description	PDB
FGG	Insertion of a 3-stranded β-meander into a coiled coil segment	2YO2
Saddle	Non-helical insertion into a coiled coil segment, exclusive to Eib proteins like EibD	2XQH
Neck	**β to α connector**	
DALL	α to β connector. Always followed by a neck domain.	2YO3, 2YNZ
HANS	Short α to β connector exclusively found before Ylheads	2YO3
Heads	**Domains of predominantly β secondary structure**	
Ylhead	Most common transversal head domain, left-handed parallel β-roll	1P9H
GIN	Transversal head, appears exclusively after interleaved head domains	3D9X
TrpRing	Most common interleaved head domain	3D9X
FxG	Variant of TrpRing	-

Note: Modified table obtained from (Bassler, Hernandez Alvarez et al. 2015) distributed under an Open Access Creative Commons License by Elsevier.

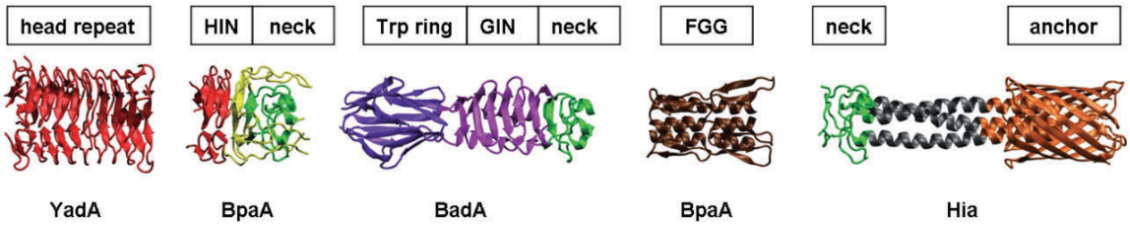


Figure 9 Representative examples of common head and neck domain folds in TAAs – YadA-like head domains (PDB: 1P9H, (Nummelin, Merckel et al. 2004)) are the most abundant structural element in TAAs, while the conserved C-terminal anchor domain of Hia (PDB: 3EMO, (Kaiser, Linke et al. 2012)) is virtually superimposable to all other TAA anchor domains. Neck domains, as seen in the crystal structure of BpaA (PDB: 3LAA, (Edwards, Phan et al. 2010)) serve as a connector between the wide β-strand rich head domains and the narrow coiled coil stalk domains. The occurrence of the amino acids GW and GIN at the N-terminal part of an uncharacterized TAA is likely an indication of the structural fold seen in the BadA head domain (PDB: 3D9X, (Szczesny, Linke et al. 2008)). FGG motifs are an example for introducing a 120° twist around the coiled coil axis as seen in the crystal structure of BpaA (PDB: 3LAA, (Edwards, Phan et al. 2010)). Modified figure obtained from (Kaiser, Linke et al. 2012) and distributed under an Open Access Creative Commons License by Blackwell Publishing Ltd.

1.5.1 Domain classification of TAAs

TAAs can be classified into repeats of head, connector, and stalk domains. This classification originated from functional association in that the head domain, rich in β-strands, is the effector of the protein that is projected by the stalk domain through the lipopolysaccharide (LPS) layer of Gram-negative pathogens in order to interact with host factors (Chauhan, Hatlem et al. 2019). While the head domains mediate a whole range of molecular interactions ranging from autoagglutination to the attachment to ECM components, the stalk domains also contribute to adhesion, which was shown for the stalk of BadA conferring attachment to fibronectin (Kaiser, Linke et al. 2012). The prototypical TAA YadA, from *Yersinia enterocolitica*, is a good example of a minimal domain organisation that can be found in a TAA (**Figure 10**). The illustration also highlights the different interaction points across the whole TAA. This mapping of molecular interactions is a crucial first step in attempting to transfer findings from one TAA to another which can ultimately be used to build a library of function-structure relations. Unfortunately, it is difficult to assign a function to a single binding location or even a single domain in a TAA, which results in the need to verify binding partners for each TAA anew.

An example for this mapping of interactions is given by Kaiser et al. who looked at the interaction between the *Bartonella henselae* TAA BadA and several extracellular matrix proteins (collagen, fibronectin, etc.) and the induction of vascular endothelial growth factor secretion (Kaiser, Linke et al. 2012): they created several deletion mutants to test the influence of the deleted domains on the binding events and concluded that the stalk can exclusively bind fibronectin, while the adherence to collagen and the induction of vascular endothelial growth factor secretion can be observed for both the head and stalk of BadA. This shows that TAAs can have redundant functions spread all over the protein to compensate for loss-of-function events (either by the host or the

pathogen itself). Function-structure associations therefore need to take the whole protein into account, not just individual domains of the protein.

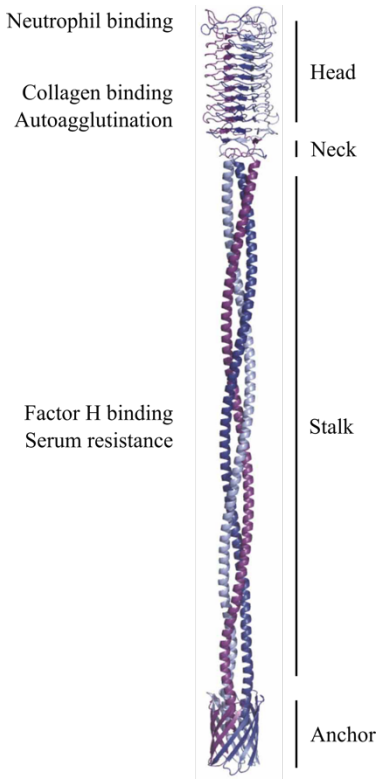


Figure 10 The prototypical TAA YadA from *Yersinia enterocolitica* – Composite cartoon model of the TAA YadA with some examples of function-domain association. YadA consists of an N-terminal head domain, which is a left-handed parallel β-roll, and a coiled coil rich stalk domain that merges with the outer membrane-bound β-barrel in the C-terminal anchor domain. Original figure was obtained from (Mikula, Kolodziejczyk et al. 2012), which is distributed under an Open Access Creative Commons License by Frontiers Media.

The domain order seen in YadA is the most minimal one that can be found in TAAs: it follows the order that has been described as “lollipop” and refers to the appearance of the protein with an N-terminal (membrane-distal) head domain, consisting of mainly β-sheet secondary structure elements, a connector motif between the wider β-sheet element and the narrower coiled coil, which is usually referred to as neck motif, a long coiled coil element named the stalk domain, and a C-terminal anchor domain that consists of a β-barrel and a connecting coiled coil residing in the inside of the barrel.

This classification into head, neck, and stalk domains is misleading and mostly kept for historic reasons: firstly, the majority of TAAs have β-sheet rich domains at multiple locations in the protein, which are all referred to as head domains. This is contrary to the intuition that one would assume that only the N-terminal domain of the protein is called the head domain, not that this refers to all β-sheet rich domains in a TAA. Secondly, the order of the different domains can be interchanged and in some extreme cases have multiple head and stalk domains, like the TAA Hsf from *Haemophilus influenzae* with a total of six stalk and head domains (**Figure 11**). Lastly, originally these classifications were assigned because it was assumed that only the head domain has a functional significance and the stalk domain fulfils only the structural purpose of projecting the head domain away from the cell surface. This hypothesis was rejected and can be seen for the previously described functional associations for YadA (Mikula, Kolodziejczyk et al. 2012) and BadA (Kaiser, Linke et al. 2012).

In short, any domain independent of location or order within the passenger domain of a TAA can be referred to as head and stalk domain and merely reflects the predominant secondary structural element within that domain.

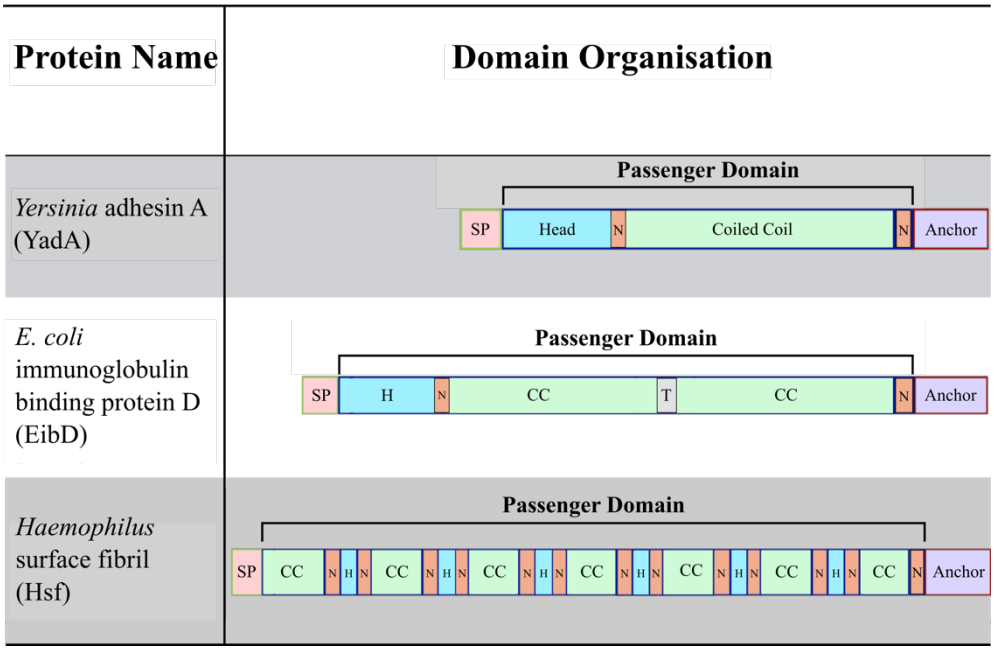


Figure 11 Complexity of domain organisation of various TAAs – Schematic showing TAAs with varying modular arrangement of functional domains within their extracellular passenger domains. TAAs always have an extended signal peptide sequence on their N-terminal end (SP, pink). β-sheet rich head domains (cyan, Head or H) are connected to coiled coil (green, CC) segments via neck motifs (orange, N). In the special case of EibD the two coiled coil elements are broken up by a transitional element (grey, T). The passenger domain is attached to the bacterial membrane via the anchor domain (purple).

1.5.2 Head domains in TAAs

Head domains in TAAs mainly consist of β-strands and can be divided into three types of structures, depending on the relative orientation of the strands to each other (**Figure 12**). A common example for the first head domain type, the β-meander, is the TrpRing motif, which can be found at the N-terminal end of BadA (PDB: 3D9X, (Szczesny, Linke et al. 2008) or in Hia (Meng, Surana et al. 2006). Another type of head domain structure is the β-prism, that consists of three walls each made of a set of five β-strands per monomer. An example for β-prism is the GIN motif found both in BadA and Hia.

The last type, and a special focus in this thesis, is the so-called left-handed parallel β-roll (LPBR): the motif is present in most TAAs and was initially described in the N-terminal head domain of YadA (Nummelin, Merckel et al. 2004). The motif consists of two β-strands from each subunit that contribute to a single superhelical turn (Kajava and Steven 2006), is highly repetitive with a set of 14 residues per repeat (with the so far only exception of 15-residue repeats found in the head domain model of UspA1; PDB: 3PR7, (Agnew, Borodina et al. 2011)). In the 14-residue repeats, the Cα positions are mostly fixed with little root-mean-square deviation (r.m.s.d.) – a feature that is very useful for molecular replacement attempts of new LPBR containing structural models or for the comparison of different LPBR motifs. A highly conserved Glycine at position 8, together with the 14-residue repeat observation, can help to identify LPBR motifs in TAAs without any structural information. These "rules" were essential for the identification and structural analysis of the C-terminal head domain of BpaC, which will be discussed in detail in this thesis.

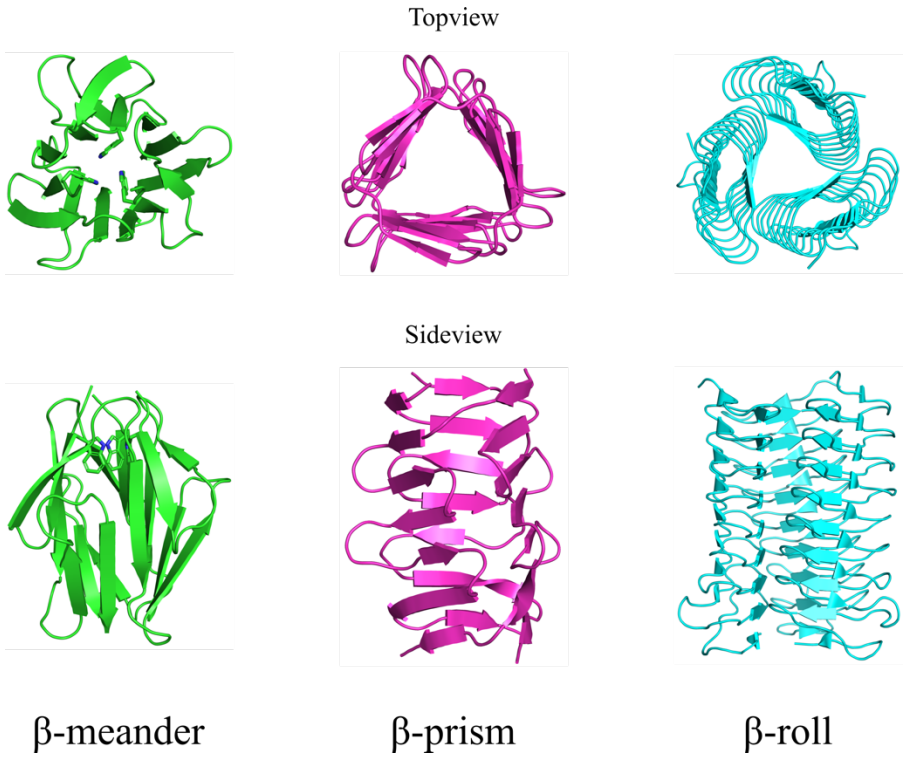


Figure 12 Overview of head domains in TAAs – Top- and sideview of structural folds) in cartoon representation) commonly found in TAA head domains. Both the β-meander (green) and the β-prism (magenta) are from the N-terminal head domain model of BadA (PDB: 3D9X, (Szczesny, Linke et al. 2008)). The example shown here for a left-handed parallel β-roll motif (cyan) is from the N-terminal head domain of YadA (PDB: 1P9H, (Nummelin, Merckel et al. 2004)).

1.6 Practical considerations for enabling TAA structure determination

The structural landscape of TAAs is still not complete and needs better annotations for common domains and structural motifs. There is also a notable drop in novel TAA structures in recent years with only about 1 TAA-related structure per year being deposited in the Protein Data Bank (PDB) between 2016 and 2019 (PDB code: 6QP4 (Mikula, Kolodziejczyk et al. 2019); 6EUN (Liguori, Dello Iacono et al. 2018); 5LNL (Wright, Thomsen et al. 2017); 3WP8 (Koiwai, Hartmann et al. 2016)) with no new TAA structure reported since 2019. This observation can be mainly attributed to the challenge that is working with TAAs.

The most basic feature of TAAs is the inherent adhesion/autoaggregation that makes this class of proteins what they are. An experiment to determine the speed of aggregation can be achieved by tracking the clumping of cells, which expressed the respective TAA, by measuring the optical density of the solution at 600 nm (OD_{600}) over a certain time frame. This assay is usually referred to as autoaggregation assay, which can also be used to identify mutations that might inhibit the translocation of the passenger domain (as shown here with A354P in YadA, **Figure 13**). However, this technique can have its limitations as the experimental setup can influence the speed of aggregation, which limits the qualitative readout from this assay. The main factors that contribute to this variability are: relative expression level of the autotransporter compared to cell density, cell density, cell viability, amount of volume used, size of reaction vessel, type of reaction vessel (plastic vs. glass), and lack of positive aggregation control. This assay can be used for simple comparisons (WT vs. mutated) of aggregation potential but should not be used for any quantification attempts.

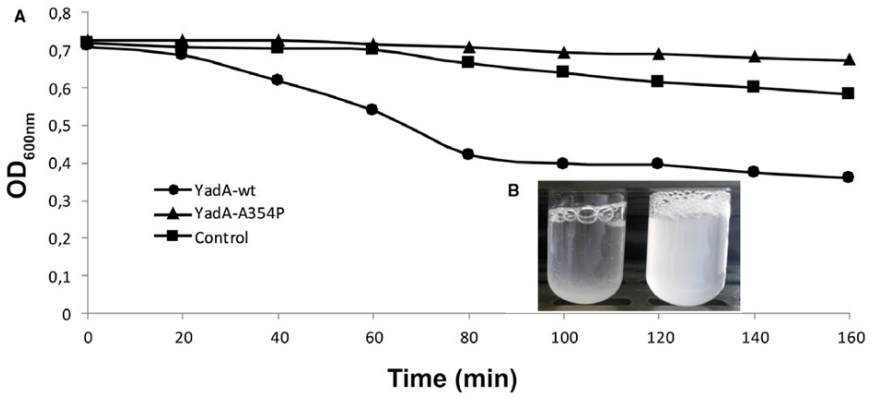


Figure 13 Autoaggregation assay of YadA and translocation mutant – **A** Series of optical density measurements (at 600 nm) of solutions containing *E. coli* expressing the TAA YadA (wild type, circle), a translocation impairing mutant (A354P, triangle), and a negative control (square). **B** Photograph of vials containing PBS with *E. coli* expressing YadA (left) and control (right) at the end of this time series. Figure taken from (Chauhan, Hatlem et al. 2019) which is distributed under an Open Access Creative Commons License by John Wiley & Sons Ltd.

Another more positive feature of some TAA domains is that of extreme stability. This is reflected by the observation, when working with TAA constructs, that the trimer band in SDS-PAGE is retained, even though the protein sample is treated in highly denaturing conditions. This high stability can be used as an advantage for alternative purification approaches: Hernandez Alvarez et al. were able to solve selected coiled coil rich structures by denaturing the protein sample, removing any contaminants and refolding the TAA construct with a very high efficiency of over 90% (Hernandez Alvarez, Hartmann et al. 2008). The success of this approach may vary for structurally more complex domains than simple coiled coils, which may be why it has only been described in the literature for these constructs.

1.6.1 *Divide-and-conquer* approach using X-Ray crystallography methods

TAAs often possess repetitive and highly modular domain arrangements, which allows the biochemist to split them up into individual segments at predefined transition points between the individual domains. Contrary to globular proteins, which may require distal interactions for the proper folding of the expressed construct, TAAs are strictly linear. Therefore, one can ignore any adjacent domains when it comes to designing smaller constructs for expression, purification, and crystallisation. This approach was coined *divide-and-conquer* (first mentioned in the context of TAAs: (Hernandez Alvarez, Hartmann et al. 2008)) and allowed the creation of composite models consisting of individual PDB models as demonstrated for the full model of SadA (**Figure 14**, (Hartmann, Grin et al. 2012)).

I employed this approach for the structural determination of BpaC which requires the accurate analysis of the sequence of BpaC using various bioinformatic tools and the domain dictionary that was previously described (Bassler, Hernandez Alvarez et al. 2015). The beauty of this method is that individual segments can be fused together to create synergistic effects just by using common transitional motifs (e.g. neck domains like DAVNxxQL) that are distributed all over the passenger domain of TAAs. This way one can create new chimeric constructs that can make use of the high solubility of one domain to compensate for the poor purifiability of another domain, even though they are originally located at opposite ends of the full passenger domain.

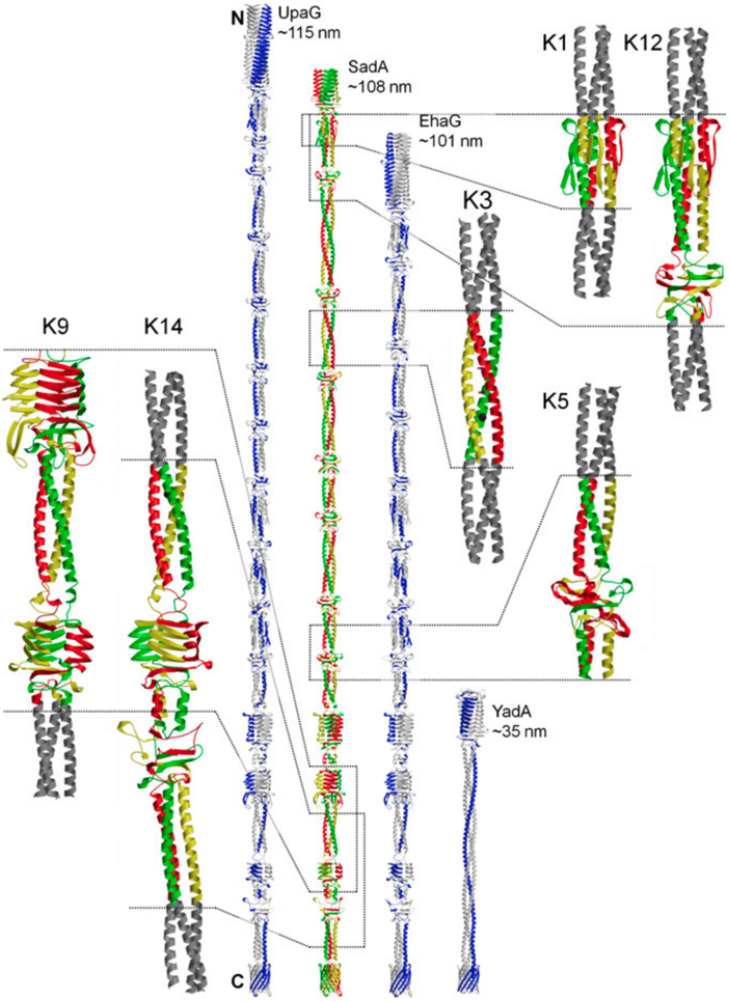


Figure 14 Composite models of four TAAs using the *divide-and-conquer* approach – The modular nature of TAAs allows the direct connection between structural models of individual segments of a TAA. Hartmann et al. solved structures of various segments of SadA (labelled with K# in the figure) and created a composite model of SadA (Hartmann, Grin et al. 2012). Using sequence alignments and other published structures further composite models were created for closely related TAAs. A size estimation for each TAA is also given. Figure obtained from (Hartmann, Grin et al. 2012) which is free to reuse for non-commercial purposes as per PNAS guidelines.

1.6.2 Creating chimeric proteins with the leucine zipper mutant GCN4pII

In TAAs, neck domains facilitate the transition between α-helix and β-sheet rich segments. This transition point can be used to create chimeric proteins with adaptors that can facilitate purification or help with crystallisation efforts.

A previously described coiled coil adaptor for TAAs is the mutated GCN4 leucine zipper domain GCN4pII or G4tri that was modified to create a stable trimer fold (Harbury, Kim et al. 1994). This mutated trimer version will be referred to as GCN4 for the rest of the thesis. GCN4 allows the replacement of otherwise labile coiled coil domains like the coiled coil segment inside the β-barrel of TAAs or coiled coil domains that would otherwise have a high tendency to aggregate; rendering structural determination unfeasible. Hernandez Alvarez et al. have shown that these chimeric fusions were able to facilitate the structural determination of several coiled coil rich domains of the TAA SadA from *Salmonella enterica* (Hernandez Alvarez, Hartmann et al. 2008).

Coiled coil proteins can be divided in characteristic seven residue repeats (a to g) with each residue at a predefined position. A common neck motif sequence in TAAs is DAVNxxQL ("x" being any residue) which signifies the start of coiled coil segments and by logical deduction can be used as a fusion point for GCN4 (here shown for the SadA structural model with the PDB code 3ZMF, **Figure 15**). The only caveat is to determine the starting point of the coiled coil register to ensure a seamless transition. In BpaC however, the most likely starting point of the register is the leucine in the DAVNxxQL motif.

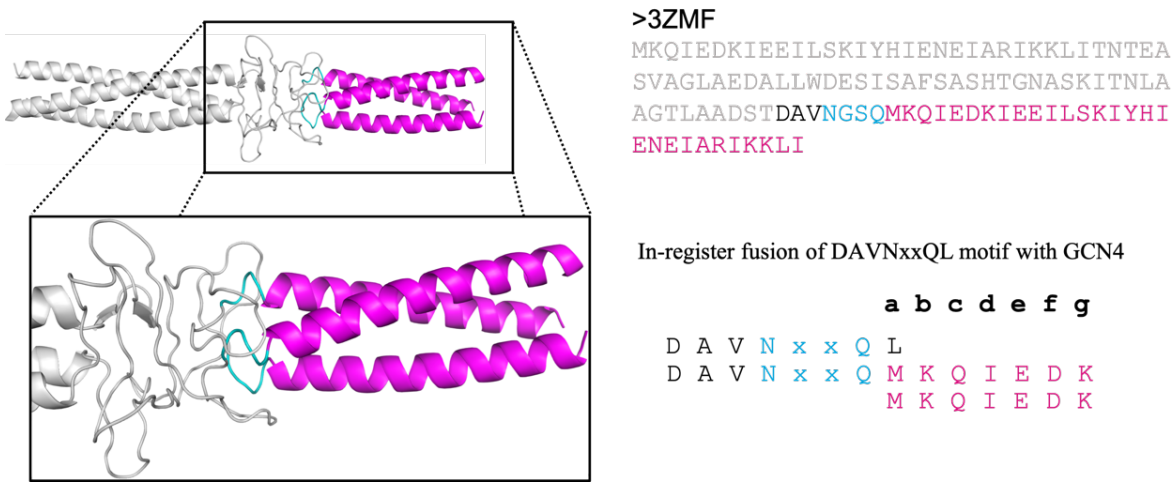


Figure 15 Fusion of SadA (303-358) with coiled coil adaptor GCN4pII – The neck motif DAVNxxQL provides a defined starting point for a subsequent coiled coil element. In this example, a DAVNxxQL motif (cyan) in SadA (PDB: 3ZMF, (Hernandez Alvarez, Hartmann et al. 2008)) was used to create a merging point with the coiled coil register (**a**-**g**) of GCN4 (magenta) at position **a** replacing the leucine of the DAVNxxQL motif with the starting methionine from the GCN4 adaptor. This motif can be found in various TAAs and guarantees the start of a coiled coil element with high accuracy.

1.7 Aim of the Thesis

The TAA BpaC was shown to be crucial in the pathogenicity of *B. pseudomallei,* but little was known about the biochemical properties and functions of BpaC. The identification of binding partners for BpaC is crucial to predict the impact that this protein has in the infection process. Unfortunately, TAAs can bind to a large variety of different proteins and without a preselection process can surmount to a time-intensive and costly detection process. This is due to the aggregation tendency of TAAs, which in turn leads to non-specific binding events that are then reported as false-positive hits. Understanding more about BpaC itself, both by looking at its sequence and eventually its structure will help narrow down the potential binding partners. The comparison with other TAAs, with similar structural motifs as BpaC, can also provide valuable binding information.

The first goal of this thesis therefore was to fully analyse the sequence of BpaC to be able to predict functional domains, and using the domain dictionary and other bioinformatic tools, even predict certain structural motifs with a high degree of accuracy. A second goal was then to use this analysis and deduce logical start and end points for individual domain constructs that then can be expressed, purified, and crystallised for structural determination. As was done for SadA before, one can then create a full composite model of the whole passenger domain of BpaC. A third goal was to characterise the binding properties of individual segments of the passenger domain of BpaC to selected extracellular matrix proteins in an attempt to associate certain binding events to specific domains. The successful accomplishment of all of these goals would provide an extensive addition to the available knowledge about BpaC, potentially leading to the development of inhibitor molecules that would abolish the pathogenic effect of BpaC and by extension *B. pseudomallei*.

2 Chapter 2: Materials and Methods

2.1 Sequence analysis of *bpaC*

Multiple genes with the name *bpaC* exist in databases from various strains of *B. pseudomallei*. The reference point for following analysis will be the 1152 residues long BpaC from *B. pseudomallei* strain 1026b which can be found as locus name BP1026B_I1575 in the *Burkholderia* genome database (burkholderia.com, (Winsor, Khaira et al. 2008)) or as entry ID A0A0H3HIJ5 in UniProt (uniprot.org, (UniProt 2019)); full (modified) sequence attached in **Appendix A**.

2.1.1 Identification of repeats and gene de-optimisation before gene synthesis

Repeat elements are very common in TAAs and need special addressing before attempting any form of cloning. This also creates the nearly impossible challenge of creating specific primer pairs within these repeats. For analysis, the **R**apid **A**utomatic **D**etection and **A**lignment of **R**epeats (RADAR, (Heger and Holm 2000)) tool from the EMBL-EBI sequence analysis tools set was used (Madeira, Park et al. 2019).

Several iterations of silent codon exchanges were performed around the start and end point of each repeat in *bpaC* by creating *in silico* primer pairs of 22-25 bp length and checking for multiple annealing sites using standard hybridization settings in SnapGene® Viewer (Vers. 9.0.4., Insightful Science). The annealing sites were chosen so one could amplify any number of repeats (e.g. repeat 2-4 for insertion into an expression vector) regardless of which repeat is selected. This also would allow the potential deletion or insertion of bases at these sites if so desired.

Gene optimisation from *B. pseudomallei* to *E. coli K-12* was ignored when ordering the finalised gene to avoid overriding the de-optimisation for the repeat segment. The final de-optimised gene

was ordered using the custom gene synthesis service from Gene Universal Inc (Newark, Delaware, USA). The product was inserted in plasmid pBluescript II SK(+) at XbaI-SacI cloning sites.

2.1.2 Prediction of domains and structural motifs

Several sequence analysis tools were used to identify domain boundaries, potential functionality, and even structural folds. PSIPRED was used to identify stretches of high α-helical content and β-sheet rich areas (Buchan and Jones 2019); HMMSCAN (Potter, Luciani et al. 2018) for the identification of Pfam associations (Mistry, Chuguransky et al. 2021); coiled coils were identified by DeepCoil (Ludwiczak, Winski et al. 2019), and TAA-specific structural motifs were assigned using the so-called domain dictionary (Bassler, Hernandez Alvarez et al. 2015). The combination of all these tools provides a good estimation of domain boundaries. Including only part of a structural fold could result in unfolded areas in the protein construct, which is a potential source of expression and purification problems. Consequently, the prediction of structural motifs greatly improves the chances of a well-folded construct by only including residue ranges with complete motifs.

2.2 Molecular Biology

Escherichia coli K-12 were grown in lysogeny broth (LB) media for protein expression and plated on LB agar plates for selection of individual colonies after transformation. For heat-shock transformation, SOC medium was used in the final resuscitation step. A list of media is displayed in **Table 2**.

Table 2 List of media.

Name	**Ingredients**
LB media (1 L)	10 g tryptone, 10 g NaCl, 5 g yeast extract
LB agar (200 mL, 10 plates)	2 g tryptone, 2 g NaCl, 1 g yeast extract, 3 g agar
SOC medium (1 L)	20 g tryptone, 0.5 g NaCl, 5 g yeast extract, 2.5 mM KCl, 10 mM $MgCl_2$, 20 mM glucose, pH 7

2.2.1 Primer design and PCR protocol

Primers were designed using SnapGene® Viewer and OligoCalc (Kibbe 2007). The former was used for checking for multiple annealing sites and bookkeeping, whereas the latter was used to calculate the annealing temperature and avoid potential hairpin/self-dimerization events. The minimum primer length was 18 bp with a minimum melting temperature (T_m) of 50 °C and a maximum difference within primer pairs of ± 5 °C. Annealing temperature (T_a) was set at T_m + 3 °C. The Q5 High-Fidelity 2X Master Mix (New England Biolabs Ltd. [NEB], Hitchin, United Kingdom) was used throughout all PCR reactions according to the manufacturer's protocol: a typical PCR reaction mix consisted of 0.75 µL 10 µM primer mix (forward and reverse), 1 ng of template DNA, 12.5 µL of twofold concentrated master mix to a final volume of 25 µL with MilliQ-grade water. The standard PCR program (**Table 3**) was carried out using a T100™ Thermal

Cycler (Bio-Rad Laboratories Inc., Watford, United Kingdom). An advanced touchdown PCR protocol was used if the standard protocol failed repeatedly (**Table 4**).

Table 3 Standard PCR program for Q5 polymerase.

Step	Temperature	Time	
Initial denaturation	98 °C	30 s	
Denaturation	98 °C	10 s	25 x
Annealing	T_a	25 s	
Elongation	72 °C	20 s per kb	
Final elongation	72 °C	120 s	
Hold	4-10 °C	∞	

Table 4 Touchdown PCR program for challenging reactions using Q5 polymerase.

Step	Temperature	Time	
Initial denaturation	98 °C	30 s	
Denaturation	98 °C	10 s	20 x
Annealing	Ta + 10 °C – 0.5 °C/cycle	25 s	
Elongation	72 °C	20 s per kb	
Denaturation	98 °C	15 s	10 x
Second elongation	72 °C	20 s per kb	
Final elongation	72 °C	240 s	
Hold	4-10 °C	∞	

2.2.1.1 PCR analysis and agarose gel clean-up

25 µL PCR mix was mixed with 5 µL of 6 x Gel Loading Dye, Purple (NEB) and loaded on a 1% agarose gel cast with TAE buffer (40 mM Tris pH 8.5, 20 mM acetic acid, 1 mM EDTA) and 10,000 x SYBR Safe DNA gel stain and run at 50 V until the lower running front (pink) reached 1/3 of the gel. 2 µL of Quick-Load 1 kb DNA ladder was run in a separate marker lane for

comparison. Bands of the desired size were excised and purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer's instructions.

2.2.1.2 Round-the-horn mutagenesis and ligation protocol

Round-the-horn site-directed mutagenesis is a PCR-based mutagenesis, which is used for doing simple insertions and substitutions (< 60 bp) and generally for deletions within constructs (Moore 2012). Insertions and substitutions are created by linear amplification of the annealing primer sequence in which the mutation is put at the non-annealing 5'-end of the primer pair. A deletion can be achieved by spacing the primers apart. The annealing part in both primers is usually adjusted to a similar melting temperature T_m of 60 °C-63 °C and a GC content of <60%. The insert sequence is split between both primers. Primers were phosphorylated at their 5'-end by T4 Polynucleotide Kinase (NEB), but instead of using the provided buffers, I used the buffer provided for the T4 DNA ligase (NEB) as this is already premixed with ATP and according to the manufacturer's instructions has equal efficiency to the PNK buffer. Following the standard PCR protocol, the methylated template DNA was digested using 0.3 µL of *Dpn*I incubated for 1 h at °37C to avoid false-positive transformation results. PCR product was cleaned up following the gel clean-up protocol as described before. Ligation of the amplified fragments was performed with the T4 DNA ligase by mixing 4.3 µL of purified PCR fragment with 0.2 µL of ligase and 0.5 µL of provided buffer and incubated for 4-6 h at RT followed by heat-shock transformation into OmniMAX cells (Thermo Fisher Scientific).

2.2.1.3 In-Fusion primer design and enzymatic ligation protocol

In-Fusion cloning requires 20 bp extensions at the non-annealing 5'-end of the primer pair that are homologous to the target sequence. The target sequence is usually a vector, which has been cut at the desired entry sites with restriction enzymes (pOPIN) or by inverse PCR amplification. These modified primers were used in the standard PCR program and following that the solution mix was subjected to a *Dpn*I digest and a gel clean-up performed as described before. For enzymatic end-to-end ligation, the NEBuilder HiFi DNA Assembly Cloning Kit (NEB) was used, which is capable of merging 2 or more fragments together in a single reaction. The recommended protocol consists of an incubation time of 60 min at 50 °C with 12.5 ng of vector DNA and a 2:1 insert:vector ratio. For larger inserts of similar size to the vector, a 1:1 ratio was employed. Heat shock transformation into OmniMAX cells was performed immediately after completion of the incubation time.

2.2.2 Transformation protocol for plasmids and ligated PCR products into *E. coli*

Plasmids or PCR products were transformed using competent cells of different strains depending on the application. An in-house protocol was used to produce these competent cells (**Appendix B**). For amplification of DNA and selection of the correct genotype, One Shot OmniMAX *E. coli* strain was used. The expression of proteins was performed in BL21Star(DE3). Transformation of plasmid DNA was carried out with 10-50 ng of DNA and transformation of PCR products was performed with 5 µL of DNA (5-20 ng). The DNA was added to 50 µL of competent cells and incubated on ice for 20 min. Heat-shock was performed at 42 °C for 30-45 sec and the tubes were immediately put on ice for 2 min. 50-200 µL of S.O.C.-medium (Invitrogen) was added to the suspension and left for resuscitation for 30 min-1 h at 37 °C. This was then plated out on agar plates for selection and left in an incubator to grow overnight at 37 °C.

2.2.3 Plasmid DNA Miniprep protocol and sequencing

After transformation in the *E. coli* strain OmniMAX, isolated colonies were picked and grown in LB medium with appropriate selection antibiotic at 37 °C overnight. DNA was prepared using the NucleoSpin Plasmid kit (Macherey-Nagel). DNA concentration was estimated by measuring the absorbance at 260 nm using the DS-11+ Spectrophotometer. DNA sequencing was carried out with the Mix2Seq kit (Eurofins Genomics). Sequencing was performed in the vicinity of mutation sites or for new constructs on the whole ORF including promoters and termination sequences.

2.2.4 Vector nomenclature and tag abbreviation

The amount of vectors and constructs used in this thesis required the use of abbreviations with a common code system for type of vector, tag type and position, residues of BpaC included in the construct, protease cleavage sites, and fusion partners.

2.2.4.1 List of abbreviations with descriptions

EV – Empty vector (no Met after promoter)

H# – His tag with number of residues

N/C – position of tag, e.g. NH6V for N-terminal His_6 tag with human rhinovirus 3C (HRV 3C) cleavage site. Order of letters relates to position within sequence, e.g. CVH6 for C-terminal HRV 3C cleavage site followed by a His_6 tag.

S – StrepII tag [WSHPQFEK] (Schmidt and Skerra 2007)

V – HRV 3C cleavage site [LEVLFQG|P]

W – Tryptophan tag [NWNWNW] (Pina, Carvalho et al. 2016)

2.2.4.2 Overview of constructs and vectors

Several different vectors were used in this study (**Table 5**). pBlueScript II SK (+) is the vector that the synthesised *bpaC* sequence was inserted into and only used for further cloning purposes. pOPIN vectors were initially used for protein expression of soluble domains of BpaC but later replaced by the low copy pET28a vector (**Table 6**).

pIBA is a low-copy vector inducible by anhydrotetracycline (AHTC) and under tight control of the *tet* promoter which was used for the SpyCatcher/SpyTag experiments and also to obtain the GCN4 sequence from the pIBA-GCN4tri vector used by Hernandez Alvarez et al. (Hernandez Alvarez, Hartmann et al. 2008).

Table 5 Overview of vector families.

Name	Promoter	Resistance	Copy number	Source
pBAD	*araBAD*	*AmpR*	High	Invitrogen
pBlueScript II SK (+)	-	*AmpR*	High	General Biosystems
pOPIN	*T7* (IPTG)	*AmpR*	High	Ray Owens (Berrow, Alderton et al. 2007)
pIBA	*tet* (AHTC)	*AmpR*	Low	IBA Lifesciences
pET	*T7* (IPTG)	*AmpR/KanR*	Low	Novagen

Table 6 Vectors for protein purification with tags.

Name	Tag	Source
pBAD*	[POI]-LysLeu-His_8	Invitrogen
pET28a	[POI]-Lys-His	Novagen
pIBA-GCN4	[POI]-[GCN4pII]-His_6	Dirk Linke (Hernandez Alvarez, Hartmann et al. 2008)
pOPINE	[POI]-Lys-His_6	Ray Owens (Berrow, Alderton et al. 2007)
pOPINF	His_6-HRV 3C-[POI]	Ray Owens (Berrow, Alderton et al. 2007)
pOPINFW	His_{10}-$(NW)_3$-HRV 3C-[POI]	Based on pOPINF
pOPINFWS	His_{10}-$(NW)_3$-HRV 3C-[POI]-StrepII	Based on pOPINFW

POI refers to protein of interest. Additional residues in three-letter nomenclature.

* Initial vector came with a different tag position but was used here for the promoter system.

2.2.4.3 Abbreviations for membrane-bound BpaC constructs

These constructs are expressed and translocated to the outer membrane of *E. coli* where the passenger domain of BpaC is facing towards the outside of the cell.

Deletions – Before beginning and after end of deleted segment is mentioned as StartAA#EndAA#del, e.g. from the full length BpaC sequence I deleted residues 160 to 432: BpaC-Q159S433del. Further abbreviation of whole domains is shown in **Table 7**.

Table 7 Deletion abbreviations for full length BpaC constructs.

Domain deleted	**Residues deleted**	**Abbreviation**
N-terminal head	Gly74 to Asp167	ΔN
Stalk I	Gln173 to Leu260	ΔCC12
Stalk II	Gln259 to Leu393	ΔCC23
Stalk I+II	Gln173 to Leu393	ΔCC13
C-terminal head	Gln392 to Leu1055	ΔCC34
Whole passenger domain	Gly74 to Leu1055	ΔP

The segment that is deleted excludes the mentioned residues, i.e. deleted residues starts after the first mentioned residue and finishes before the last-mentioned residue in a row. This is so you are able to identify the residue of the transition point even after deletion.

An example for a BpaC construct, which is expressed in the pBAD vector system and in which the N-terminal head domain has been deleted, is defined as follows: pBAD-BpaC-ΔN.

2.4.4.4 Abbreviations for BpaC domain constructs expressed in the cytosol

Constructs that include sequences taken from the solvent-accessible passenger domain and expressed in the cytosol of *E. coli* follow a similar nomenclature logic (**Table 8**). Here, the start and end of the extracted sequence chosen are shown with the appropriate start and end position and associated amino acid within the original full length BpaC sequence. Deletions are denoted as described for the full-length constructs. In the special case of the cysteine substitutions at Cys76 and Cys97 (both to a serine) an additional insertion in the abbreviation is made.

Table 8 Nomenclature examples for domain constructs of BpaC.

Example	Abbreviation	Vector example	Explanation
Standard	(Vector name)-(StartAA#EndAA#)	pET28a-T166S433	Construct in pET28a Residues taken from Thr166 to Ser433 of BpaC
Including deletion	(Vector name)-(StartAA#(ΔAbbreviation)EndAA#)	pET28a-S99(ΔCC12)T517	Construct in pET28a Residues taken from Ser99 to Thr517 of BpaC but with a deleted segment in between Gln173 to Leu260
Including GCN4 fusion and substitution	(Vector name)-(StartAA#(C#S)EndAA#)(GCN4)	pET28a-G90(C97S)Q173(GCN4)	Construct in pET28a, amino acid sequence taken from Gly90 to Gln173 of BpaC but with a cysteine-to-serine substitution at residue 97 and with a GCN4 fusion just after the Gln173.

Refers to the residue number within the sequence of *bpaC*.

2.3 Protein expression in *E. coli* BL21(DE3)

2.3.1 Small scale expression for analytical experiments and lysis tests

Initial expression trials were performed in small-scale (50 mL cultures) to estimate the ratio of soluble/insoluble protein after lysis before scaling up to large-scale expression (450 mL cultures) for protein purification.

5 mL of LB medium with 100 µg/mL of carbenicillin or 50 µg/mL of kanamycin, depending on plasmid resistance marker, and 1% glucose was inoculated with a single colony of heat shock transformed BL21Star(DE3) and incubated overnight at 37 °C and 200 rpm. OD_{600} was measured and 50 mL of LB medium with appropriate antibiotic concentration was inoculated to achieve a final OD_{600} of 0.05. Bacterial culture was incubated at 37 °C and 140 rpm and OD_{600} measurements were taken at regular intervals. Once the OD_{600} reached 0.6, protein expression was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM and bacterial culture was harvested after 3 h by centrifgation at 4500 x g for 40 min. For constructs expressed in the pBAD vector, 0.2% of L-arabinose was used to induce protein expression instead of IPTG. Supernatant was discarded and cells frozen at -20 °C until further use.

2.3.2 Large scale expression for protein purification

50 mL of LB medium with appropriate antibiotic and 1% glucose was inoculated with a single colony of heat shock transformed BL21Star(DE3) and incubated overnight at 37 °C and 200 rpm. OD_{600} was measured and 450 mL of LB medium with 100 µg/mL of carbenicillin or 50 µg/mL of kanamycin, depending on plasmid resistance marker, was inoculated with the necessary starter culture volume to achieve a final OD_{600} of 0.05. The bacterial culture was incubated at 37 °C and 140 rpm and OD_{600} measurements were taken at regular intervals. Once the OD_{600} reached 0.6,

protein expression was induced by adding IPTG to a final concentration of 1 mM and incubated for a further 3 h until the bacterial culture was harvested by centrifugation at 4000 x g for 30 min. The supernatant was discarded and cell pellets were transferred to 50 mL falcon tubes for storage at -20 °C.

2.3.3 Whole cell sample preparation for SDS-PAGE

An aliquot was taken for use in SDS-PAGE analysis while harvesting bacterial cells. A calculated volume of cells, which is the equivalent of cells in a 100 µL cell suspension at OD_{600} of 0.5, was taken and spun down in 2.0 mL Eppendorf tubes at max rpm for 10 min. Supernatant was discarded and cells were lysed by resuspension in 50 µL of 1x SDS loading buffer (62.5 mM Tris-HCl pH 6.8, 1% SDS, 0.8% Glycerol, 1.5% 2-Mercaptoethanol, 0.005% Bromophenol blue) and incubated at RT for 15 min. Optionally, a pinch of DNase I was added if the viscosity of the sample was high and hindered gel loading. A final centrifugation step for 10 min at max rpm was performed and 10 µL of sample was carefully taken from the top of the solution for loading on a 15-well protein gel (4–20% Mini-PROTEAN® TGX™ Precast Protein Gels, Bio-Rad). The gel was run according to the manufacturer's protocol along with 3 µL of protein standard marker (Color Prestained Protein Standard, Broad Range (10-250 kDa), NEB). Electrophoresis was stopped once the Bromophenol blue running front reached the bottom of the gel.

2.3.4 Lysis test for BpaC domain solubility comparison

For the cytosolically-expressed BpaC domain constructs, lysis tests were performed to estimate the solubility and the amount of unfolded species for a given construct for rapid comparison with each other before scaling up selected constructs to large scale purifications. For this, cell pellets from small scale expression tests were thawed on ice and resuspended in 20 mL of lysis buffer (50 mM Na_2PO_4 pH 8, 300 mM NaCl, 10 mM Imidazole) together with 200 µL of Proteoloc protease inhibitor cocktail (Expedeon). Lysis was performed by sonication with a microtip on ice. The program was set to 40% amplitude, 10 s/50 s on/off for a total of 1 minute on-time, which was repeated once. 8.4 µL of lysed cells were kept separate for SDS-PAGE analysis. This sample was then spun down at max rpm for 15 minutes and the supernatant (SN) transferred to a separate tube and diluted with lysis buffer to 35 µL before adding 7 µL of 6 x SDS loading buffer (375 mM Tris-HCl pH 6.8, 6% SDS, 4.8% (v/v) Glycerol, 9% 2-Mercaptoethanol, 0.03% Bromophenol blue). The pellet (P) was resuspended in 42 µL of 1 x SDS loading buffer. Both were frozen at -20 °C until SDS-PAGE analysis.

2.4 Purification of cytosolically-expressed BpaC domains

2.4.1 Sample preparation for Immobilised Metal Affinity Chromatography (IMAC)

Cells from the large scale expression harvest were thawed on ice and resuspended in either a cells wet weight to buffer volume ratio of 1:4 or a minimum of 20 mL of lysis buffer (50 mM Na_2PO_4 pH 8, 300 mM NaCl, 10 mM Imidazole) to ensure the microtip of the sonicator was fully submerged in the 150 mL polystyrene container. Lysis by sonication was performed on ice using the following program: 40% amplitude, 10 s/50 s on/off for a total of 1 minute on-time. This was repeated once or twice depending on the density of the solution. A cooling period of 5 minutes was used in between repeats for each sample. 8.4 µL of lysed cells were taken for SDS-PAGE analysis (SN+P). The remaining suspension was cleared by centrifugation at 40.000 x g (JA 25.50) for 40 min at 7 °C. The supernatant was filtered using a 0.45 µm syringe filter before transferring the supernatant to a 50 mL falcon tube for bead incubation with IMAC resin.

2.4.2 Ni-NTA purification

All buffers were made using MilliQ-grade water (Q-POD, 0.2 µm membrane filter, 18.2 Ω, Merck). Buffers were filtered using either 0.22 µm or 0.45 µm Whatman membrane filters (GE Healthcare Life Science) before use.

2.4.2.1 Standard protocol

Most purifications were performed using the same buffer compositions but with varying volumes for resin amount, wash volume, Imidazole concentration for washes, and elution volumes. Standard buffer names, compositions, and abbreviations are shown in **Table 9**.

Ni-NTA IMAC was performed using 20 mL Econo-Pac chromatography columns (Bio-Rad) packed with Ni IMAC resin (Ni Sepharose 6 Fast Flow, GE Healthcare Life Sciences). Resin was prepared with washes of 5 column volume (CV) ddH_2O and 5 CV lysis buffer before being transferred to a 50 mL falcon tube. Cleared supernatant from the lysis step was mixed with IMAC resin and put on a tube roller for a 30 min incubation step at RT. The resin-supernatant mix was transferred back to the gravity flow column and flowthrough was collected for later analysis with SDS-PAGE. Wash and elution buffers were added in various amounts to remove non-specific bound contaminants and elute the His-tagged protein (**Table 10**). Resin was recovered by adding 5 CV of 1 M Imidazole, 20 CV of ddH_2O and 2 CV of 20% ethanol for storage.

Table 9 List of buffers used for protein purifications.

Name	**Ingredients**
Crystallization buffer	20 mM Tris pH 8, 150 mM NaCl
IMAC elution buffer (E)	As Lysis buffer, 300 mM Imidazole
IMAC wash buffer 1 (W1)	As Lysis buffer, 30 mM Imidazole
IMAC wash buffer 2 (W2)	As Lysis buffer, 50 mM Imidazole
Lysis buffer (also W0 for IMAC)	50 mM NaP_i pH 8, 300 mM NaCl, 10 mM Imidazole

Table 10 Overview of different purification attempts for soluble domain constructs of BpaC.

Name	ID	Resin (mL)	W0 (CV)	W1 (CV)	W2 (CV)	E (CV)
pOPINFW-T914D1097	P001	0.5	10	20	5	8 x 1 CV
pOPINFW-D386T517	P002	4	5	3	3	0.6, 2, 4 x 1 CV
pOPINFW-T914Q1054(GCN4)	P003	2	5	3	3	0.6, 2, 5 x 1 CV
pOPINFWS-N748S947	P004	2	10	5	3	0.6,4.6, 2, 2 CV
pOPINFW-V249S433	P005	2	10	5	3	0.6, 4.6, 2, 2 CV
pOPINFW-D386T517	P006	4	15	5	7.5	0.6, 7, 1 CV
pET28a-S741Q1054(GCN4)	P008	2	20	5	10	0.6, 8, 1 CV
pET28a-G90C97SQ173(GCN4)	P010	2	20	5	20	0.6, 8, 1 CV
pET28a-S99(ΔCC12)T517	P011	2	20	5	10	0.6, 8, 1 CV
pET28a-L260T601	P013	1	25	-	30	0.6, 10, 1 CV
pET28a-S99(ΔCC12)T601	P014 P016	5	25	-	35	0.6, 10, 1 CV
pET28a-(GCN4)A261Q392(GCN4)	P017	2	25*	-	40*	0.6, 3 x 3, 5 CV*
pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4)	P018	1	25	-	40	0.6, 7 CV

* NaCl concentration was increased to 500 mM for this purification.

Abbreviations of buffers (W0-2, E)explained in Table 9.

2.4.2.2 Extended washing protocol to test for the presence of chaperone

The release of chaperones is triggered by the addition of 1 mM ATP/$MgCl_2$ in between the first and second wash step. Non-specific contaminants are removed by washing with lysis buffer as before (W0). Chaperones and contaminants attached to the chaperone-protein complex are released by addition of 5 CV of lysis buffer together with a final concentration of 1 mM ATP/$MgCl_2$ and incubated for 30 min at RT. The flowthrough was collected for SDS PAGE analysis and the remaining protein eluted with 10 CV of IMAC elution buffer.

2.4.2.3 Purification under denaturing conditions

For the stalk domain construct pET28a-(GCN4)A261Q392(GCN4), purification under denaturing conditions (Guanidinium chloride, GuHCl) was performed. In purification P017 (**Table 10**), the filtered supernatant was taken further for native IMAC while the pellet was used for purifying under denaturing conditions. For this, the pellet was resuspended in 35 mL of denaturation equilibration buffer (50 mM HEPES pH 8, 500 mM NaCl, 10 mM Imidazole, 6 M GuHCl) and dispersed using a 16-gauge luer-lock syringe needle and incubated at 30 °C for 1 h in a shaking incubator. The suspension was spun down at 40,000 x g (JA 25.50) for 45 min at 20 °C. IMAC purification steps were identical to the native purification protocol for P017 except for the presence of 6 M GuHCl in all buffers. SDS-PAGE analysis was omitted until after refolding due to the disruptive nature of GuHCl on electrophoresis performance. Refolding was achieved via dialysis (see section 2.4.4).

2.4.2.4 Temperature challenge after IMAC elution of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4)

Following the IMAC step of purification number P018 of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4), the elution sample was split into 35 µL aliquots in 0.2 mL PCR tubes. A total of 13 samples were aliquoted this way, each for a temperature point starting at 40 °C up to 95 °C in 5 °C increments along with a RT control. A T100™ Thermal Cycler (Bio-Rad Laboratories Inc., Watford, United Kingdom) was used for the temperature challenge and set to the relevant temperature in a gradient program. Multiple Cyclers were used to cover the full temperature range at the same time. The temperature challenge was performed for 30 min after

which a cooling period to RT was added to the program until use in SDS-PAGE. Samples were spun down and standard SDS-PAGE protocol was applied after that.

2.4.3 Removal of N-terminal His tags by HRV 3C protease

pOPINF and pOPINFW constructs contained an N-terminal HRV 3C cleavage site (**Table 6**). To test the ability to cleave the HRV 3C tag in the pOPINFW plasmid (addition of $(NW)_3$ in front of the HRV 3C cleavage site), a sample of pOPINFW-D386T517 purified sample (P002) was exposed to different protein:HRV 3C ratios (w/w) overnight at 7 °C. For this, the IMAC elution was dialysed against 20 mM Tris-HCl pH 7.6, 150 mM NaCl, 1 mM DTT, followed by dividing the sample up into aliquots to mix with HRV 3C in different ratios: control (no HRV 3C), 1:5 + 500 µM Urea, 1:10, 1:20, 1:40, 1:100, 1:200. The final volume was normalised to ensure comparability on SDS-PAGE. The sample with the 1:100 protein:HRV 3C ratio was further mixed with 100 µL of Ni-NTA resin and spun down before using the supernatant as input for SDS-PAGE. The resin was then incubated with 100 µL of IMAC elution buffer for further analysis.

2.4.4 Buffer exchange via two-step dialysis

Selected fractions from IMAC elution were pooled and transferred to a SnakeSkin™ dialysis tube (3.5 kDa molecular weight cut-off, ThermoFisher) of appropriate length. Dialysis was performed in two steps: the first one at a protein:buffer (v/v) ratio of 1:50 or a minimum of 500 mL of dialysis buffer for 2 h at RT, the second one at a protein:buffer (v/v) ratio of 1:100 or a minimum of 1 L overnight at 7 °C. This ensures thorough buffer exchange for downstream applications sensitive to Imidazole or different salt concentrations. In most cases, the dialysis buffer consisted of 20 mM Tris-HCl pH 8, 150 mM NaCl. One exception for this is the purification attempt P016 (**Table 10**) for pET28a-S99(ΔCC12)T601, which served as a test for optimised buffer conditions, with the buffer of choice consisting of 50 mM 2-(N-morpholino)ethanesulfonic acid (MES) pH 5.5 and

50 mM NaCl. A further exception was the refolding attempt on pET28a-(GCN4)A261Q392(GCN4) in P017 which used an excess of dialysis buffer to ensure the complete removal of GuHCl: 10 mL of IMAC elution were transferred to 2 x 2 L of refolding buffer (20 mM HEPES pH 8, 500 mM NaCl, 50 mM L-Glu).

2.4.5 Protein quantification via Nanodrop

Protein concentration was estimated by measuring 280 nm absorption (UV280) using a DS-11+ Spectrophotometer (DeNovix) assuming the theoretical extinction coefficient calculated by ExPASy ProtParam software (Wilkins, Gasteiger et al. 1999) was correct for samples with >80% purity, but for every other measurement an average E1% (g/100 mL) of 10 was assumed.

2.4.6 Protein concentration as preparation for crystallisation trials

Following the dialysis step to exchange the protein into the final buffer for crystallisation trials, the protein solution was concentrated using centrifugal concentrators. Depending on the volume of collected fractions, either 20 mL or 2 mL concentrators were used (Vivaspin 20 or 2; 3 kDa molecular weight cut-off PES, Sartorius). Concentrators were washed with MilliQ-grade water and equilibrated with target buffer (depending on downstream application) before the protein sample was concentrated to the desired concentration according to the manufacturer's description.

2.4.7 Size exclusion chromatography for analytical purposes

To check the monodispersity of the refolded species in the purification P017 (**Table 10**) of pET28-(GCN4)A261Q392(GCN4), size exclusion chromatography was performed. After extended dialysis from high GuHCl containing buffer into no GuHCl containing buffer, the protein solution was filtered with a 0.22 µm PES syringe filter (33 mm, Millipore) before concentration to 1 mL using a centrifugal concentrator as described in section 2.4.6. A HiLoad 16/60 Superdex 200 pg

column (GE Healthcare) was attached to attached to an ÄKTA pure micro system (Cytiva) kept at 9 °C. The column was prepared according to the manufacturer's protocol and equilibrated with 2 CV of refolding buffer (20 mM HEPES pH 8, 500 mM NaCl, 50 mM L-Glu). The sample was loaded on a pre-equilibrated 2 mL injection loop. The sample was injected at a flow rate of 0.25 mL/min and eluted over 1 CV using a flow rate of 0.5 mL/min. After 0.3 CV, fractions of 6 mL in 15 mL falcon tubes were collected for later analysis on SDS-PAGE. UV280 chromatogram was recorded and displayed using Excel.

2.4.8 SDS-PAGE

Samples were taken at various time points during the purification (**Table 11**). Samples from wash steps were expected to be too dilute to be visualised by Coomassie stained gels. To overcome this, samples were concentrated by acetone precipitation (indicated by * in figure descriptions, e.g. W0*). For this, 200 µL of sample were mixed with 1.7 mL of ice-cold acetone and incubated at -20 °C for at least 2 h (usually overnight). Precipitate was spun down in 4 °C centrifugation step at 20,800 x g for 45 min. The acetone was removed and the pellet was air-dried for 10 min. The sample was resuspended in 20-40 µL 1 x SDS loading buffer and incubated at 37 °C and 800 rpm until the pellets were fully dissolved again.

Samples were mixed with 6 x SDS loading buffer (375 mM Tris-HCl pH 6.8, 6% SDS, 4.8% Glycerol, 9% 2-Mercaptoethanol, 0.03% Bromophenol blue) to a final volume of 42 µL (35 µL sample + 7 µL 6 x buffer). Samples were heated to 95 °C for 5 minutes and then spun down at 11,000 x g for 10 minutes. 12 µL were loaded into either 26 well (4-20% Criterion Tris-HCl Protein Gel, 15 µL, Bio-Rad) or 15 well gels (4-20% Mini-PROTEAN TGX Stain-Free Protein Gels, 15 µL, Bio-Rad) alongside 1 µL of PageRuler Plus Prestained Protein Ladder (10 to 250 kDa, Thermo Fisher Scientific). Gels were run at 80 V until protein samples passed into the

gel and 160 V until the lowest MW marker band reached the bottom of the gel. The gel was stained with Quick Coomassie Stain (Generon) for 1 h and destained with ddH_2O overnight. Gels were imaged using a G:BOX (Syngene) for 26 well gels and a gel scanner (Perfection 2400 Photo, Epson) for 15 well gels. Images were processed with GIMP (Vers. 2.10; crop, convert to grayscale, modify contrast/brightness) and annotated with Inkscape. Intensity values were estimated with the Gel analyser tool in Fiji (Schindelin, Arganda-Carreras et al. 2012).

Table 11 Samples taken for SDS-PAGE during purification of BpaC soluble domains.

Name	Abbreviation	Amount taken (µL)
Supernatant (after lysis)	SN	8.4
Pellet (after lysis)	P	-
Flowthrough (IMAC)	FT	17.5
Wash (IMAC)	W0-3	200*
Elution (IMAC)	E1-X	35
Everything else	-	35

Final volume was always 35 µL sample with 7 µL 6 x SDS loading buffer.

* Note that values bigger than 35 µL indicate that these samples were concentrated by precipitation with ice-cold acetone.

2.5 Structural studies of cytosolically-expressed BpaC domains

2.5.1 Preparation of crystallisation trials using Sparse Matrix screens

Crystal plates (MRC Plate 96-well 3 drop UV Crystallization Plate, Molecular Dimension) were setup using the NT8 crystallization robot (Formulatrix) with a drop setting of 100 nL:100 nL (reservoir:protein) and using the JCSG Core Suite screens I-IV (Qiagen). The final setup for each trial is described in (**Table 12**). Monitoring of crystal growth and imaging thereof (normal, UV, SONICC) was achieved using the Rock Imager plate hotel (Formulatrix).

Table 12 Crystallization trials of purified BpaC domains.

Name	ID	Purification ID	Drop conc. (mg/mL)	Buffer	JCSG Core
PINFW-D386T517	C001	P001	7.6	20 mM Tris-HCl pH 8	I-IV
pOPINFW-D386T517	C002	P002	3.71, 2.97, 2.37	20 mM Tris-HCl pH 8	I-IV
pOPINFW-T914Q1054(GCN4)	C003	P003	2.05, 1.03	20 mM Tris-HCl pH 8, 150 mM NaCl	I-IV
pOPINFWS-N748S947	C004	P004	3.69, 1.85	20 mM Tris-HCl pH 8, 150 mM NaCl	I-II
pOPINFW-D386T517	C005	P006	20.89, 13.79, 2.57	20 mM Tris-HCl pH 8, 150 mM NaCl	I, II, IV
pET28a-S741Q1054(GCN4)	C006	P008	130.09, 18.32,8.96	20 mM Tris-HCl pH 8, 150 mM NaCl	I-IV
pET28a-G90C97SQ173(GCN4)	C007	P010	16.33, 8.17	20 mM Tris-HCl pH 8, 150 mM NaCl	I-IV
pET28a-S99(ΔCC12)T601	C008	P014	28.17	20 mM Tris-HCl pH 8, 150 mM NaCl	I-IV
pET28a-S99(ΔCC12)T601	C009	P016	86.79	20 mM MES pH 5.5, 50 mM NaCl	I-IV

Name of the purified construct is shown alongside the crystallization and purification ID. Drop concentrations were estimated using UV280 absorbance.

2.5.2 Harvesting of crystals and data collection parameters

Crystals grown in hit conditions were sent to Diamond Light Source beamlines I04-1 and I23 for X-Ray diffraction experiments. Crystals were harvested using Dual Thickness MicroMounts (MiTeGen) with different loop sizes (20-100 μm) and plunge-frozen in liquid nitrogen for storage and transport in Unipucks (Molecular Dimensions). Data were collected at the fixed wavelength of the corresponding beamline.

For the specific case of the dataset that was collected on the crystal hit that led to the structure of the C-terminal head domain of BpaC, the specific harvest conditions and data collection parameters were as follows: Crystals were harvested at day 2 by adding 0.2 μL cryoprotectant solution to the 0.1 μL drop and incubating the mixture for 1 min until crystals were harvested using a 0.2 mm LithoLoop (Molecular Dimensions). Crystals were plunge frozen in liquid nitrogen and sent to I04 for X-Ray diffraction experiments. The wavelength was set to 0.979 Å, the beamsize was 32 x 20 μm and 3600 images were collected over a 0.1 ° Ω angle oscillation change between images. Transmission was set to 100% with an exposure time of 0.02 s per image.

2.5.3 Data processing of S741Q1054(GCN4)

Image processing was performed using XDS. The presence of multiple lattices was detected and lattice indexing was performed in XDS to extract a single lattice for downstream processing. Scaling and integration of the dataset was performed in XDS as input for ccp4i2 model building.

2.5.4 Model building in the ccp4i2 program suite

All initial model building was performed inside the ccp4i2 program suite (Vers. 1.0.2, (Potterton, Agirre et al. 2018)). Merging step on the XDS processed dataset was performed with Aimless (Evans and Murshudov 2013). For molecular replacement, a CHAINSAW (Stein 2008) trimmed

down version (only Val11 to Thr152) of the PDB entry 3S6L (DOI: 10.2210/pdb3S6L/pdb) was used as input for Phaser (McCoy, Grosse-Kunstleve et al. 2007). The phaser solution was used as input for Buccaneer (Cowtan 2006). Clearly wrong chains and residues were removed in Coot (Vers. 0.8.9.1., (Emsley, Lohkamp et al. 2010)). Several rounds of model building in Coot followed by refinement in Refmac5 (Murshudov, Skubak et al. 2011) were performed until all residues with sufficient $2F_o$-F_c electron density (at 1 σ) were built.

2.5.5 Structure Refinement in Phenix/Coot

The final refinement steps were performed in Phenix (Vers. 1.18, (Adams, Afonine et al. 2011)) with alternating building rounds in Coot and structure validation using Molprobity (Williams, Headd et al. 2018).

2.5.6 Structure analysis of S741Q1054(GCN4)

The structure was analysed and images created using PyMOL (Vers. 2.4.1, (Schrodinger 2017)) and Inkscape (Version 1.0.1, https://inkscape.org). The full C-terminal head domain (T431-Q1054) was created by making a separate object out of the first three layers of repeats and aligning them in PyMOL to the first two layers thereby creating an additional layer with regards to the old model. This process was repeated until T431 and then merged in Coot. Varying side chains were replaced using the actual sequence and the structure information from the corresponding layers in the actual structure. Size estimation of homology models and the extended head domain model (T731-Q1054) was performed in PyMOL. The APBS Electrostatistics plugin in PyMOL was used for electrostatic surface visualisation (https://pymolwiki.org/index.php/APBS). The Angle between sheets were calculated using the AngleBetweenHelices plugin (https://pymolwiki.org/index.php/AngleBetweenHelices).

2.5.7 Structure comparison with other TAAs of similar fold

Alignment of YadA-like head repeats (G742-S1021) was performed by artificial designation of start and end point of each individual repeat. A frequency plot for these repeats was then created using the WebLogo server (Crooks, Hon et al. 2004). Similar TAA structures were identified using a shortened structure model (S741-A782, three repeats) as input for the Dali server (Holm and Laakso 2016). Top hits that belonged to the TAA protein class were assessed and structurally compared using PyMOL. Frequency plots were created as well after manually aligning the repeats as done for BpaC. pI calculations were initially performed in ExPASy (Gasteiger, Gattiker et al. 2003) and for more accurate calculations, the IPC 2.0 server was used (Kozlowski 2021). If not stated otherwise, the pI value was calculated in ExPASy.

2.6 Microbiological studies of full length BpaC and deletion/insertion mutants

2.6.1 Translocation studies

2.6.1.1 SpyCatcher/SpyTag fluorescence system

The SpyTag was inserted between residue Ala73 and Gly74, which is just after the N-terminal predicted signal cleavage site (Ala69|Ala70) but before any predicted structural fold. The SpyCatcher counterpart was purified as a sfGFP fusion construct from the expression vector pIBA3 following the large-scale expression protocol described in section 2.3.2 (using 0.1 µg/mL of AHTC instead of IPTG). pIBA3-SpyCatcher-sfGFP was a gift from Ina Meuskens with permission from Jack Leo (Addgene plasmid #107420) and purified following the protocol published in (Chauhan, Hatlem et al. 2019). Different domain deletion constructs (ΔN, ΔCC12, ΔCC3) were tested along with a double cysteine mutant (Cys76Ser, Cys97Ser), and an alanine-to-proline substitution mutant (Ala1085Pro).

Each construct was expressed following the small scale expression protocol in section 2.3.1. After harvesting, cell pellets were resuspended in 1 x PBS and the equivalent of 1 mL of an OD_{600} of 1 was transferred to a 2 mL Eppendorf tube. The cells were then spun down and resuspended in 1 mL of 1 x PBS. 20 µL of 1 mg/mL of SpyCatcher-sfGFP was added and incubated for 1 h at RT while shaking at 300 rpm. Cells were again spun down and washed with 1 mL of PBS, repeating this step three times in total. Final resuspension in 75 µL of 1 x PBS was transferred to 0.5 mL for fluorescence measurement using a DeNovix QFX Fluorometer with an excitation emitter of 470 nm and an emission filter of 514-567 nm.

2.6.1.2 Periplasmic stress reporter system

The construct pUA66-P*rpoE*-mNG (a kind gift of M. Steenhuis, (Steenhuis, Abdallah et al. 2019)) was transformed in BL21Star(DE3) cells (*KanR*) and competent cells were created using an in-house protocol, attached in **Appendix B**. These competent cells were then used for heat shock transformation with constructs contained in pBAD (*AmpR*): empty vector (pBAD-EV), BpaC wild type (pBAD-BpaC-WT) and Ala1085Pro substitution mutant (pBAD-BpaC-A1085P). Transformed bacteria were plated on LB agar plates containing both carbenicillin and kanamycin at appropriate concentration levels and incubated overnight at 37 °C. Individual colonies were picked and used for inoculation of 5 mL LB overnight cultures (37 °C) with both antibiotics present. OD_{600} was determined and a starting OD_{600} of 10 mL of LB culture (37 °C, 200 rpm, both antibiotics at appropriate concentrations) was set at 0.05. Protein expression was induced at OD_{600} of 0.6 using 0.1% L-arabinose. The equivalent of 120 µL at an OD_{600} of 1 was taken at regular intervals. Each sample was spun down and resuspended in 300 µL of 1 x PBS and again spun down for final resuspension in 100 µL 1 x PBS. The final OD_{600} was measured using Nanodrop for more accurate normalisation to mNeonGreen (mNG) fluorescence signal. Fluorescence was measured using a DeNovix QFX Fluorometer with an excitation emitter of 470 nm and an emission filter of 514-567 nm.

2.6.2 Autoaggregation assay

Due to the adhesive properties of TAAs, a common assay to check for correct expression, folding and translocation of the protein is an autoaggregation assay. BpaC was expressed in *E. coli* BL21Star(DE3). Protein expression was induced for a set amount of time (here: OD_{600} = 0.6, 3 h, 37 °C, 1 mM IPTG) in 10 mL of LB medium. The final OD_{600} was measured and the culture pelleted at 3000 x g for 5 min before resuspending in 1 x PBS matching all cultures to a final OD_{600} of 1.0 and transferring an equal amount of each culture in glass tubes and incubating at RT overnight, while taking an OD_{600} measurement every twenty minutes for the first two hours and at the end point after 24 hours.

2.6.3 Binding of BpaC to extracellular matrix proteins

Binding of full length BpaC and BpaC-ΔN mutant, and empty vector control in BL21Star(DE3) was tested using a commercially available kit (ECM cell adhesion array kit, ECM545, Sigma-Aldrich) that comes with a detection system (lysis followed by detection of dsDNA with a fluorescent dye) and a 12 x 8-well removable strip system. Each row is coated with a different extracellular matrix protein: collagen I, collagen II, collagen IV, fibronectin, laminin, tenascin, vitronectin, BSA.

As a first test, the dynamic range of the fluorescent dye was explored by lysing *E. coli* cells at different cell densities (from OD_{600} of 1.0 to 0.001) and adding lysis/dye solution followed by fluorescence measurement with a DeNovix QFX Fluorometer (470 nm excitation, emission filter 514-567 nm). As a control, 100 mM NaOH was added to the cell suspension that was split prior to the addition of both normal and +NaOH lysis/dye solutions. Linearity of correlation between fluorescence and cell density was estimated.

Small scale expression of pBAD-EV, pBAD-BpaC-WT, and pBAD-BpaC-ΔN in BL21Star(DE3) was performed as described before. After harvesting, cells were resuspended in 20 mL of 1 x PBS and a target value of an equivalent of 1 mL at an OD_{600} of 1 was estimated. The appropriate volume was taken, spun down and resuspended in 1 mL of 1 x PBS.

Wells of the strips from the ECM kit were rehydrated with 100 µL of 1 x PBS for 10 minutes at RT. Solution was removed and an additional blocking step was performed using 200 µL of blocking buffer (1 x PBS, 0.05% Tween-20, 2% BSA) incubating for 2 h at 37 °C. Strips with blocking buffer were emptied by turning over and gently shaking onto a tissue until most liquid was removed. Immediate addition of 100 µL of cells at an OD_{600} value of less than 1 was performed to prevent dehydration of the wells. Wells were incubated for 3 h at 37 °C in a closed-off 96 deep-well block with additional water in empty wells to prevent dehydration. Cells were removed as described earlier and immediate addition of 200 µL of wash buffer (1 x PBS, 0.05% Tween-20, 0.2% BSA) was done. This step was repeated four times. In the last step 150 µL of wash buffer was added and 50 µL of lysis buffer/dye solution was added according to the manufacturer's protocol. Lysis was performed in 15 min and solution in wells was resuspended several times before transferring 150 µL into 0.5 mL Eppendorf tubes for fluorescence measurement. For each test condition triplicates (separate colonies) were performed. Statistical analysis was performed in GraphPad Prism (Vers. 9.3.0 for Windows) using a one-way ANOVA with Tukey and Dunnett tests as follow-up.

3 Chapter 3: Gene analysis, manipulation and biochemical characterisation of full length BpaC

3.1 Sequence preparation for creating highly specific primer annealing sites

Trimeric autotransporter adhesins consist of similar structural folds due to the restrictions that the trimeric nature imposes on possible or allowed folds. This often can lead to repetitive segments within a domain which help to enlarge a particular binding interface or project another functional domain away from the cell surface. In BpaC, the C-terminal head domain consists of 14-residue long repeats of high sequence identity from residue 434 to 1021. This equates to a total of 42 repeats within the C-terminal head domain. The majority of the repeats can be assigned to a certain family that have been designated “GDN”, “GEN”, and “GSN”, highlighting the importance of the residue at position 2/14 which will be further discussed in **Chapter 5**. On the C-terminal end of the domain, one can find repeats with much less sequence identity that do not pose a challenge for cloning purposes due to the high variance of codons. However, the repeats assigned to the three families (GDN/GEN/GSN) are almost identical, which limits the choice of codons for a particular amino acid. This heavily impacts the specificity of any primer that would be designed within this area and would make further manipulations to the sequence almost impossible. Therefore, the first challenge in this project was to (de)optimise the nucleotide sequence around the transition points in between repeats enabling the design of primers that have unique binding sites or at least a low chance for off-site annealing. Secondly, as at the beginning of this project there was no information on the individual domains of BpaC nor on the likely structural folds present within these domains,

I had to perform a full bioinformatic analysis on the protein sequence of BpaC to determine domain borders, likely structural folds and potential homologue structures as a basis for a *divide-and-conquer* approach for structural elucidation of these domains by X-Ray crystallography. Lastly, deletion points in BpaC were defined that allowed the bridging of domains that were far apart in sequence. For this, the structural transition motif DAVNxxQL, which exists in the transition areas between coiled coils and β-sheet rich areas, was used as the fusion point. The structural conservation on both sides of this motif guaranteed a low structural impact on the final merged construct.

3.1.1 Gene (de)optimisation of repeat elements

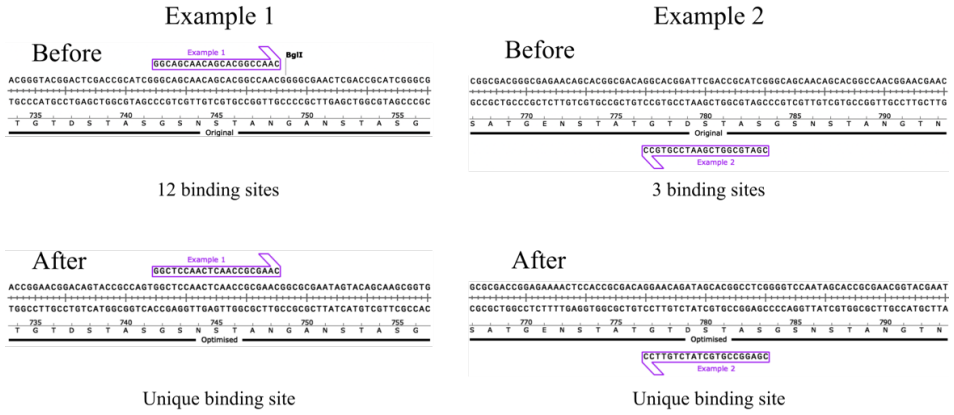
The high sequence identity between repeat elements present in the C-terminal head domain of BpaC was a major hurdle to be overcome before any cloning attempts could be made. One way to increase the diversity in the nucleotide sequence is to introduce controlled silent mutations *in silico*. First, primers with a length of 18-22 nucleotides were created in SnapGene and potential multiple binding sites were listed and compared individually. Secondly, the individual codons were diversified so that each binding site had a unique nucleotide sequence. This was an iterative process, which required going back and forth between identifying new binding sites that arose due to the change in nucleotide sequence and changing these codons but without returning to previously eliminated binding sites. After this process, a total of 600 nucleotides were changed, which is about 17% of all nucleotides in BpaC. An excerpt of the alignment between original and (de)optimised sequence and two examples for the improvement in primer binding specificity is shown in **Figure 16**.

1) Identify high identity repeats in BpaC

2) Increase primer specificity by implementing silent codon mutations

```
742  G   S   N   S   T   A   N   G   A   N   S   T   A   S  755
Original   GGC AGC AAC AGC ACG GCC AAC GGG GCG AAC TCG ACC GCA TCG
Optimised  GGC TCC AAC TCA ACC GCG AAC GGC GCG AAT AGT ACA GCA AGC
           ***  *  ***     **  **  *** **  *** **      **  ***
756      D               S       T       A   S       T  769
Original   GGC GAT AAC AGC ACG GCG AGC GGC ACG AAC GCA TCG GCG ACG
Optimised  GGT GAT AAC TCC ACA GCT AGT GGG ACC AAT GCC AGC GCG ACC
           **  *** *** *  **  **  **  **  **  **      *** **
770      E               T       T   D           S  783
Original   GGC GAG AAC AGC ACG GCG ACA GGC ACG GAT TCG ACC GCA TCG
Optimised  GGA GAA AAC TCC ACC GCG ACA GGA ACA GAT AGC ACG GCC TCG
           **  **  *** *  **  *** *** **  **  *** **  **  **  ***
```

3) Analyse potential primer binding sites in SnapGene



4) Repeat 2) and 3) until no more improvement is possible

Figure 16 Examples of the (de)optimisation process for repeat sequences of BpaC – **1)** WebLogo output of repeats 1 to 32 each with a length of 14 residues shown as frequency plot (Crooks, Hon et al. 2004). **2)** Nucleotides are grouped in codons with the translation shown above the codon as one-letter code. Empty translation spaces indicate an identical residue to the one shown in the top row of the aligned sequences. Indicator for identical match between original and (de)optimised sequence is shown per row (*). Alignment excerpt shown for residues 742 to 783. **3)** Two examples of the effect of codon optimisation on primer annealing specificity. **4)** This cycle was repeated until sufficiently optimised for downstream applications.

3.1.2 Identification of structural domains

The more accurate the domain prediction, the less likely is the chance of obtaining a misfolded protein when splitting a full-length protein into smaller parts. In the case of TAAs this is even more specific as there are no distally interacting domains due to the strict linear nature of the trimeric arrangement. This means that isolated TAA domains can fold independently of each other. PSIPRED (Buchan and Jones 2019) provides an overview of the secondary structure prevalence of BpaC and was used as a first orientation point to identify the more α-helical stalk domain(s) and the more β-sheet rich head domains (**Figure 17 + 18**). This is particularly accurate for some areas that are very similar in TAAs, like the obligate trimeric coiled coil that is inside the membrane-bound β-barrel or the β-barrel consisting of a total of 12 β-sheets. However, there are also shortcomings to this method, which presents itself in contradicting results from different bioinformatic methods. Combining all of the information from multiple bioinformatic methods is needed to increase the accuracy of the secondary structure prediction. PSIPRED struggles with the detection of continuous α-helical segments (orange box, **Figure 17 + 18**). A more precise detection of coiled coils is given by the DeepCoil prediction software (Ludwiczak, Winski et al. 2019), which was used to identify coiled coil sites in BpaC (**Figure 19**). PSIPRED prediction either did not include a complete α-helical assignment for a coiled coil segment (Stalk I), predicted an area as having no secondary structure assignment even though it is clearly coiled coil (Stalk II) or even had the wrong assignment as rich in β-sheets (Transition to head). The coiled coil within the barrel domain was predicted correctly by both programs.

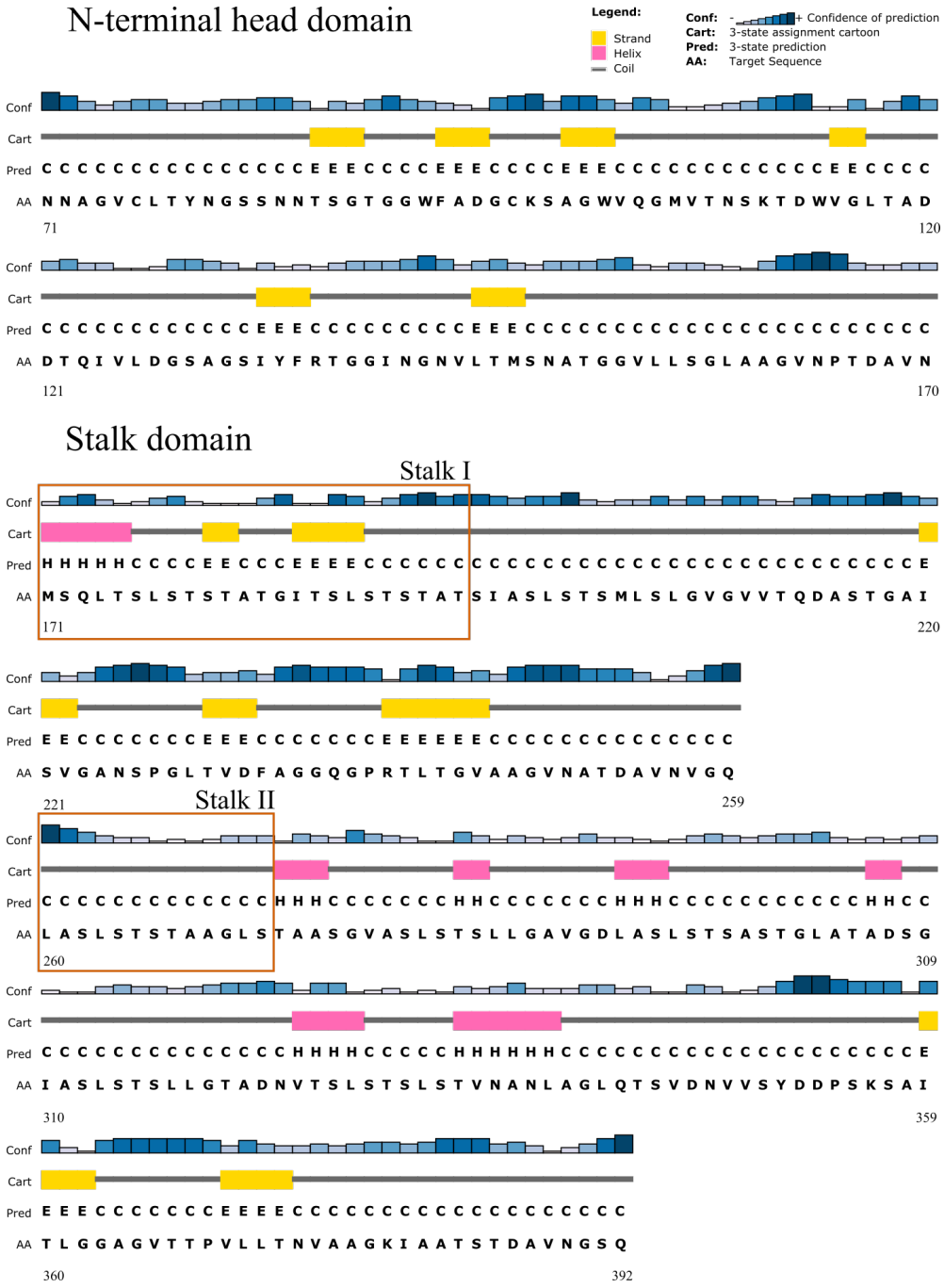


Figure 17 PSIPRED prediction results for first half of BpaC – Secondary structure prediction is shown with confidence scale (Conf), 3-state cartoon (Cart) as coil (grey line), helix (pink), and strand (yellow), and as letter code (Pred) with coil (C), helix (H), and strand (E). The sequence is shown in the bottom row (AA). Residue number is shown at beginning and end of each row. Major deviations that are identified later by more elaborate coiled coil prediction methods are highlighted and labelled to correspond to the findings for the DeepCoil prediction results (orange box).

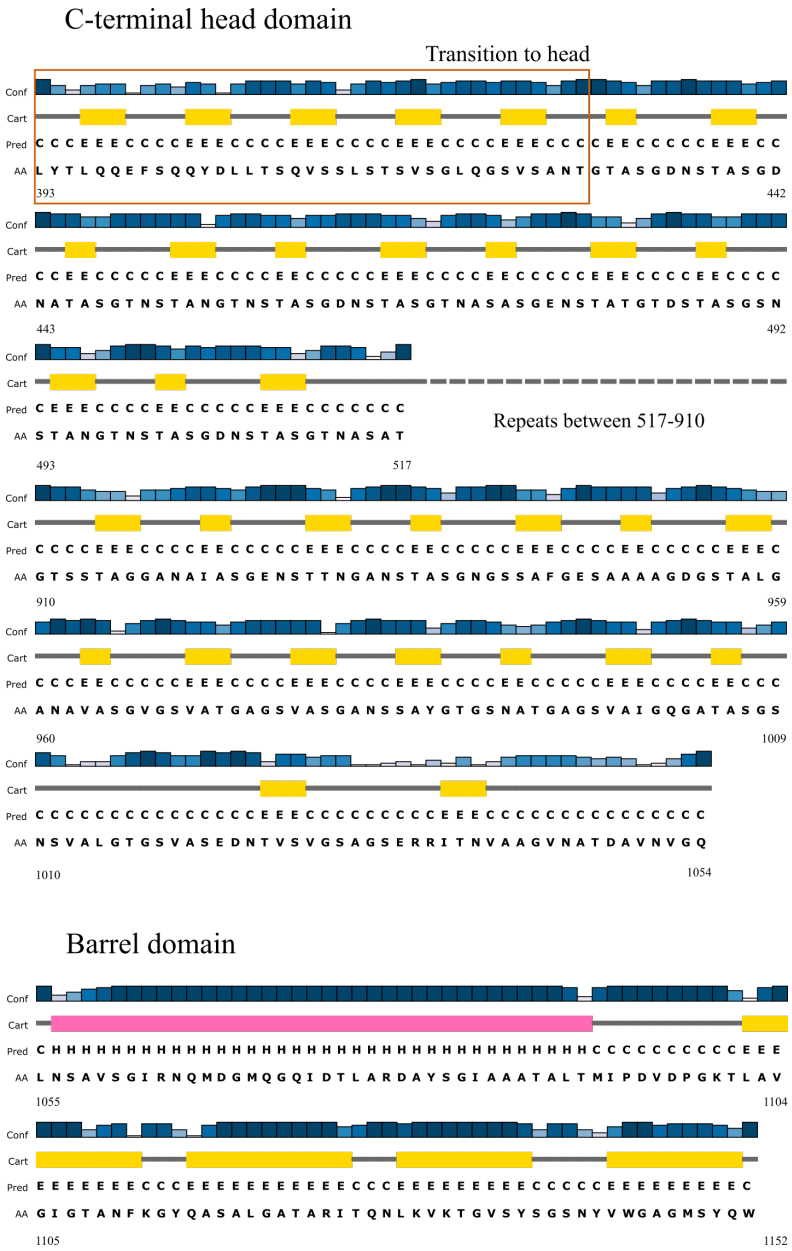


Figure 18 PSIPRED results for parts of the C-terminal head domain and barrel domain of BpaC – Secondary structure prediction is shown with confidence scale (Conf), 3-state cartoon (Cart) as coil (grey line), helix (pink), and strand (yellow), and as letter code (Pred) with coil (C), helix (H), and strand (E). The sequence is shown in the bottom row (AA). Residue number is shown at beginning and end of each row. Assignment for residues 517-910 were skipped as these are identical to the remaining C-terminal head domain as discussed before. Major deviations that are identified later by more elaborate coiled coil prediction methods are highlighted and labelled to correspond to the findings for the DeepCoil prediction results (orange box).

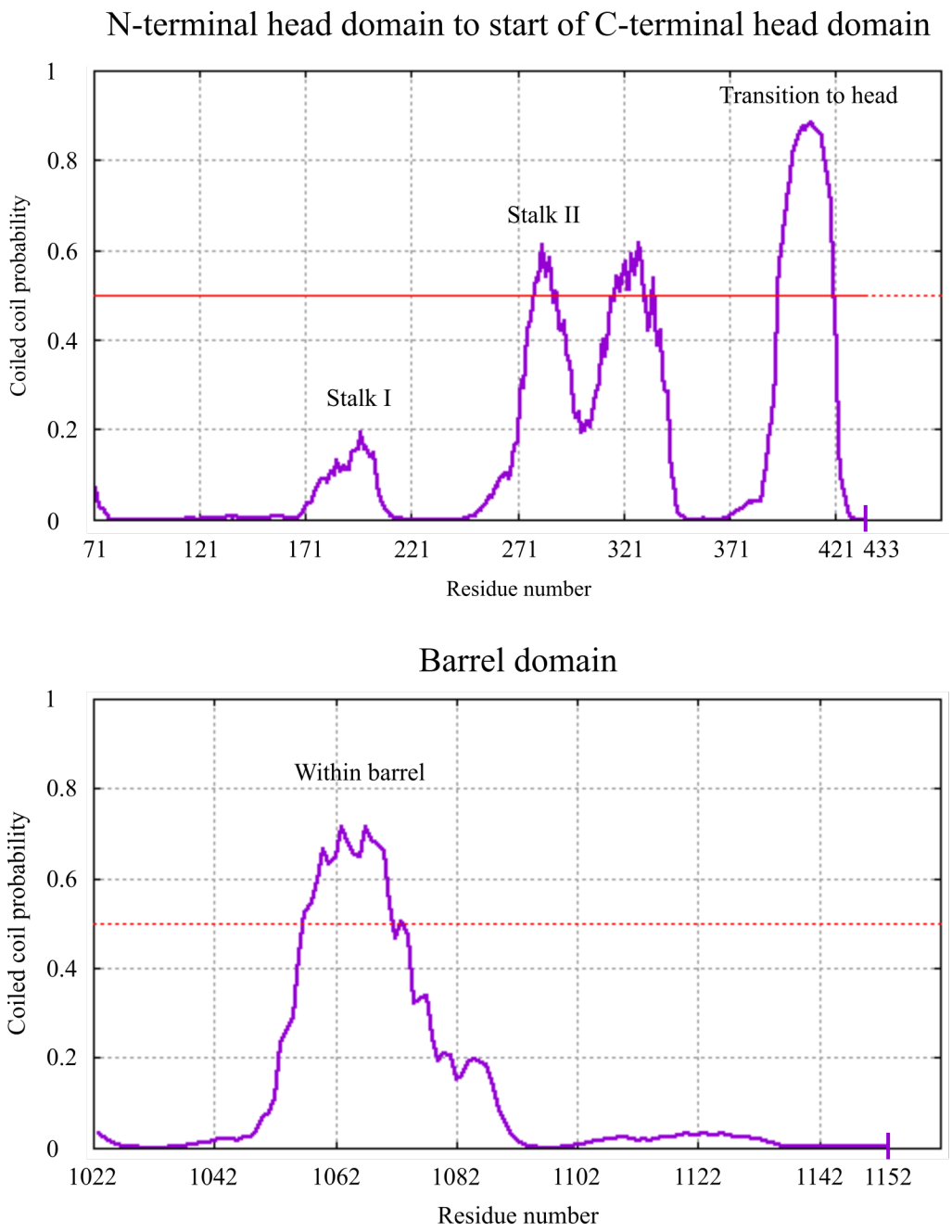


Figure 19 DeepCoil analysis of BpaC – Coiled Coil probability is given as purple line with an arbitrary threshold set at 0.5 (red). Residue numbers from 71-433 and 1022-1152 are shown, which exclude the cleaved residues of the signal peptide sequence (1-70) and the C-terminal head domain repeats (434-1021). Designations of individual peaks correspond to the labelled areas in the PSIPRED prediction.

3.1.3 Applying TAA rules for more accurate domain prediction

The structural nature of TAAs is unique in that the protein sequence can contain enough information to allow the prediction of structural models with high accuracy especially within a previously identified structural motif. The confidence of assignments lacking clear TAA motifs can be increased by other secondary structure prediction software like PSIPRED or DeepCoil. Searching the PDB for homologues by sequence can be challenging due to the sequence divergence of individual residue positions within a certain domain but a manual comparison of structural features and identified motifs of all published TAA structures and BpaC can give a good starting point for later verification by data-driven structure solutions. Using this information, I defined domain borders, assigned domain names, and provided a likely homology model that can be used for example for molecular replacement later on in the structure prediction pipeline (**Table 13**)

Table 13 Assignment of domain borders for BpaC.

Name	Domain boundaries	Identified motifs	Suggested homology model	Reference
Extended Signal peptide region (ESPR)	(ATG)-N71	ESPR	AMA|AN (Cleavage site)	(Campos, Byrd et al. 2013)
N-terminal head domain	N72-L174	TrpRing (GW), GIN motif (GIN), Neck (DAVNxxQL)	BadA Nhead, 3D9X	(Szczesny, Linke et al. 2008)
Stalk domain I	T175-L260	Coiled Coil (DeepCoil), Neck (DAVNxxQL)	AtaA stalk, 3WPO (3418-3465)	(Koiwai, Hartmann et al. 2016)
Stalk domain II	A261-V418	Coiled Coil (DeepCoil), FxG (L361GG), Neck (DAVNxxQL), Coiled Coil (DeepCoil)	SadA stalk, 2YO2	(Hartmann, Grin et al. 2012)
C-terminal head domain	T431-L1055	LPBR repeats, Neck (DAVNxxQL)	BoaA Chead, 3S6L*	(Edwards, Gardberg et al. 2011)
Barrel domain	N1056-(End)	Coiled Coil (DeepCoil), β-barrel (PSIPRED)	YadA barrel, 2LME	(Shahid, Bardiaux et al. 2012)

Domain predictions were made by combining PSIPRED, DeepCoil, and motif assignments. Motifs have been identified using a TAA rules library.

3.1.4 Creation of deletion mutants

The structural assignment of the functional domains of BpaC allows for the creation of domain deletion mutants that are unlikely to impact the folding and successful translocation of the remaining passenger domain. This is due to the presence of the neck transition motif DAVNxxQL at several locations in the passenger domain of BpaC: N-terminal to stalk I, stalk II, transitional coiled coil before the C-terminal head domain, and before the barrel domain. In essence, the Gln of the start point of the deletion (first DAVNxxQL) is fused together with the Leu of the end point of the deletion (second DAVNxxQL) for a structurally complete DAVNxxQL motif. This method was used to create a deletion mutant for every solvent-accessible domain (N-terminal head domain, stalk domain, C-terminal head domain) with an additional mutant for the subdomain named "stalk I" with a suspected additional aggregation potential compared to "stalk II" alone (**Figure 20**). This aggregation potential is explored in more detail in **Chapter 4** which is covering the expression and protein purification of soluble BpaC domains. These deletion mutants, alongside with the full-length WT construct, were used in biochemical assays (binding to extracellular matrix proteins).

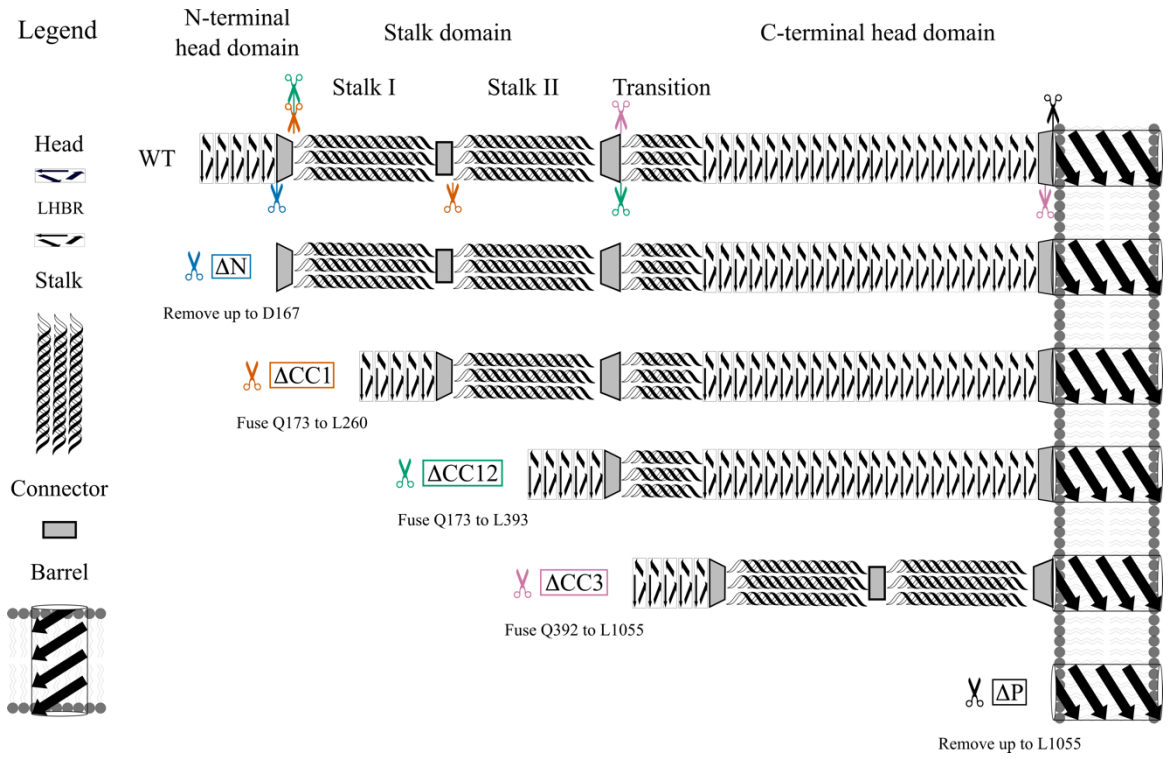


Figure 20 Schematic of BpaC WT and deletion mutants used in biochemical assays – Domains of BpaC are shown with simplified secondary structure assignments. Start and end point of individual deletion mutants are marked as scissors with corresponding colours in the wild type construct (WT). Deletion mutants designed to exclude every functional domain: deletion of N-terminal head domain (ΔN), stalk I domain (ΔCC1), stalk I+II domain (ΔCC12), C-terminal head domain including transitional coiled coil segment (ΔCC3), and full passenger domain deletion (ΔP).

3.1.5 Creation of substitution mutant A1085P

A good control construct was needed for the translocation assays as there is no published data on BpaC for this kind of experiment. A wider literature search revealed a study on the TAA YadA from *Yersinia enterocolitica* that explored the effect of introducing single proline substitutions at the C-terminal end of the coiled coil located inside the β-barrel (Chauhan, Hatlem et al. 2019). The structural similarity of the β-barrel domain inside the TAA family (Dautin and Bernstein 2007) enables the approximate transfer of the substitution location in YadA to BpaC. This is done via sequence alignment in Clustal Omega (**Figure 21**), comparing the location of A354P in YadA with the corresponding BpaC position, and cross-checking the secondary structure prediction of BpaC using PSIPRED with the secondary structure obtained from the barrel domain model of YadA (PDB: 2LME).

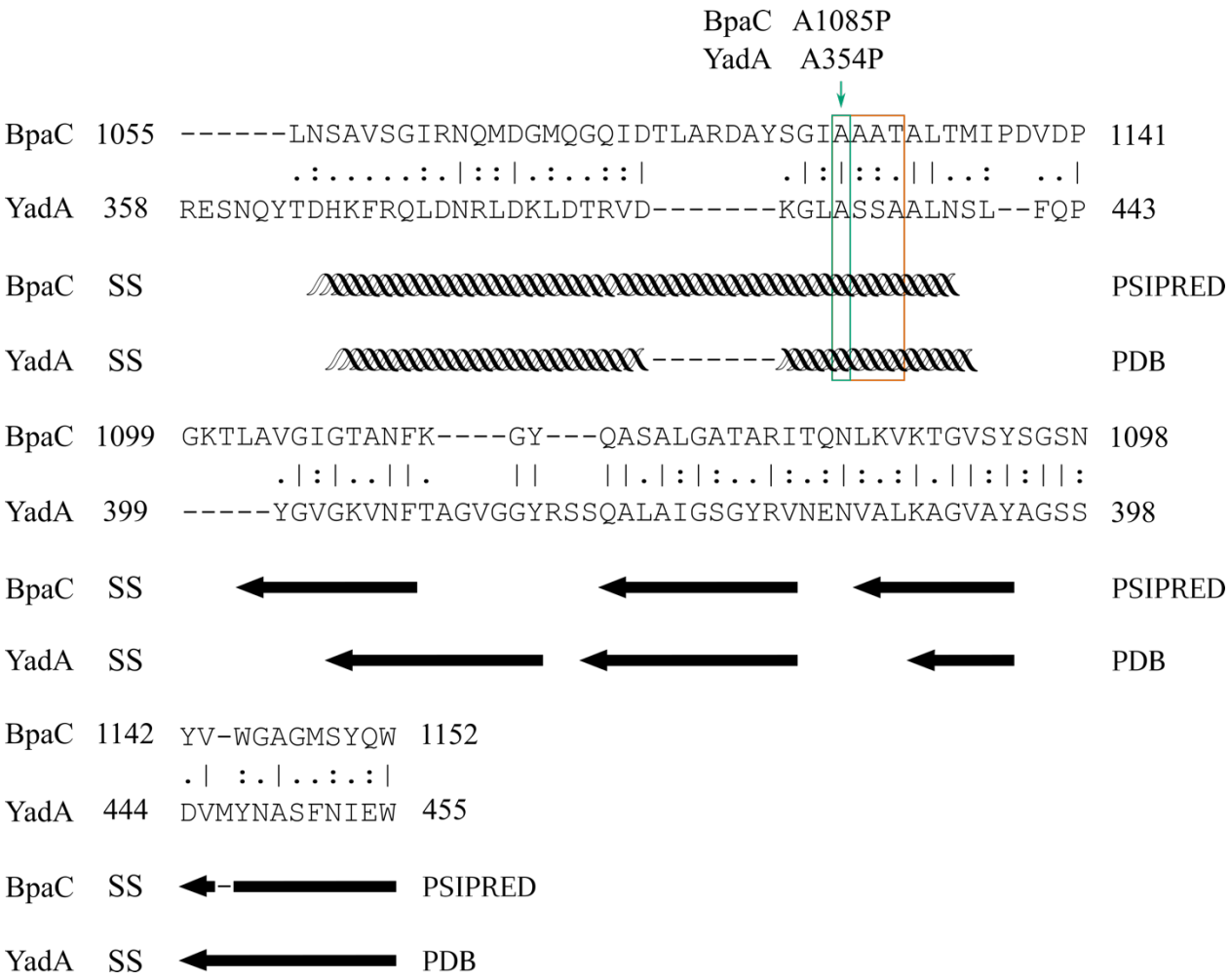


Figure 21 Alignment of barrel domain of BpaC and YadA – Sequence was aligned in EMBOSS Needle using a EBLOSUM62 matrix (Needleman and Wunsch 1970) with start and end residue numbers shown for each line. Similar residues are marked with one or two dots depending on degree of similarity and identical residues with a connecting line. Secondary structure (SS) assignment was performed using PSIPRED for BpaC and for YadA using the solid-state NMR model of the barrel domain (PDB: 2LME, (Shahid, Bardiaux et al. 2012)). The coiled coil within the β-barrel is indicated by a helical schematic with a gap in the alignment for YadA. The four β-sheets present in each TAA monomer are shown as arrows. The "ASSA" motif in YadA is highlighted (orange box) and the residue selected for substitution by proline indicated with a green arrow and a green box.

3.2 Translocation studies on full-length BpaC and deletion mutants

Carrying out biochemical interaction assays on a protein that has only ever been studied on a systemic level requires a variety of different control experiments as TAAs are by nature highly adhesive and tend to have a low affinity for any binding partner. This leads to a high rate of false-positive hits in any form of identifying new potential binding partners. The first question addressed was if the mutants created had an impact on the translocation of the passenger domain. A false-negative outcome would be the consequence of a stalled translocation, which would lead to the misinterpretation of the importance of a certain (deleted) domain on a binding event. Therefore, two orthologous experimental setups were designed to confirm the successful translocation of all deletion mutants: one method, which uses a SpyTag/SpyCatcher system, has been successfully used before to study the impact of proline substitution mutants within the barrel domain of YadA on the translocation of the passenger domain (Chauhan, Hatlem et al. 2019). The other method was adapted from a periplasmic stress assay that was initially used to study the impact of small drug inhibitors on the function of the BAM (Steenhuis, Abdallah et al. 2019). Since TAAs also are translocated via the BAM this seemed like a reasonable orthologous experiment to see if the deletion mutants can be measured via the periplasmic stress that is created during overexpression.

3.2.1 Insertion of the SpyTag into full-length BpaC and deletion mutants

The SpyTag/SpyCatcher system has been used before to study the translocation of TAAs, specifically in YadA (Chauhan, Hatlem et al. 2019). Here, a 13-residue long sequence called SpyTag (AHIVMVDAYPTK) is inserted at the N-terminal end of the protein-of-interest (POI, (**Figure 22**, **A**)). The counterpart to this is a fusion construct consisting of the SpyCatcher protein itself (about 15 kDa) and a superfolder GFP (sfGFP) attachment (about 27 kDa). Both of these have reactive residues that form a covalent isopeptide bond when interacting with each other (**Figure 22**, **B**).

In TAAs, an extended signal peptide is located at the N-terminal end and is cleaved off during the translocation process at a given cleavage site that is different in each TAA. Luckily, the exact cleavage site in BpaC has been published (Lafontaine, Balder et al. 2014) and was used as information to find the most suitable insertion site for the SpyTag. The N-terminal signal peptide is cleaved between N71 and N72, so I inserted the SpyTag (along with 7 flexible G/S linker residues) in between A73 and G74. This insertion was performed for the WT and following mutants: one for each domain deleted (ΔN, ΔCC12, and ΔCC3), one substitution mutant with all cysteines mutated to serines (C76S, C97S) and an expected negative control (the A1085P mutant) described in section 3.1.5 (**Figure 22, C**).

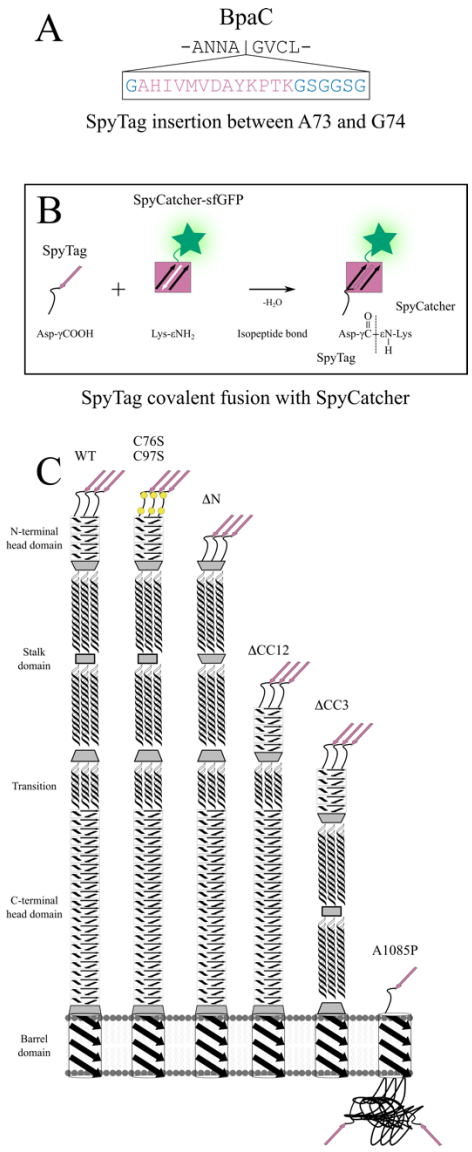


Figure 22 Adaptation of SpyTag/SpyCatcher translocation assay for BpaC and deletion mutants – **A** The 13-residue long SpyTag (magenta) was inserted between A73 and G74 of BpaC with seven additional residues to create a flexible spacer between BpaC and the SpyTag sequence (blue). **B** SpyTag and SpyCatcher (about 15 kDa) spontaneously form a covalent isopeptide bond between Asp-γCOOH within the SpyTag and Lys-εNH inside SpyCatcher. Attached to SpyCatcher is a sfGFP moiety (about 27 kDa) for detection of the complete complex via fluorescence. **C** Schematic of WT BpaC and all deletion mutants and their expected folding and translocation profile. Domain description is shown for WT along with simplified secondary structure profile for each domain: Arrows indicating β-sheet rich area, helix representing areas with likely coiled coil content. SpyTag attached to the N-terminal end of the protein (magenta arrow). The substitution mutant A1085P has not been confirmed as a negative control for the translocation of the passenger domain of BpaC and just highlights the expected outcome *a priori*, which is a misfolded passenger domain halted at the translocation step.

3.2.2 Purification of pIBA3-SpyCatcher-sfGFP

pIBA3-SpyCatcher-sfGFP was purified following standard IMAC purification protocol as used for the BpaC soluble domain purifications with small deviations. 3.67 g of *E. coli* BL21 Star cells grown in 900 mL of LB medium were used in this purification. Final purity after IMAC was 75% (estimated by band intensity comparison in SDS-PAGE), which is enough for biochemical assays (**Figure 23**, **E**). 2 mL of 0.56 mg/mL was the final protein amount and concentration at the end of the purification. This equals a final yield of 1.24 mg/L culture.

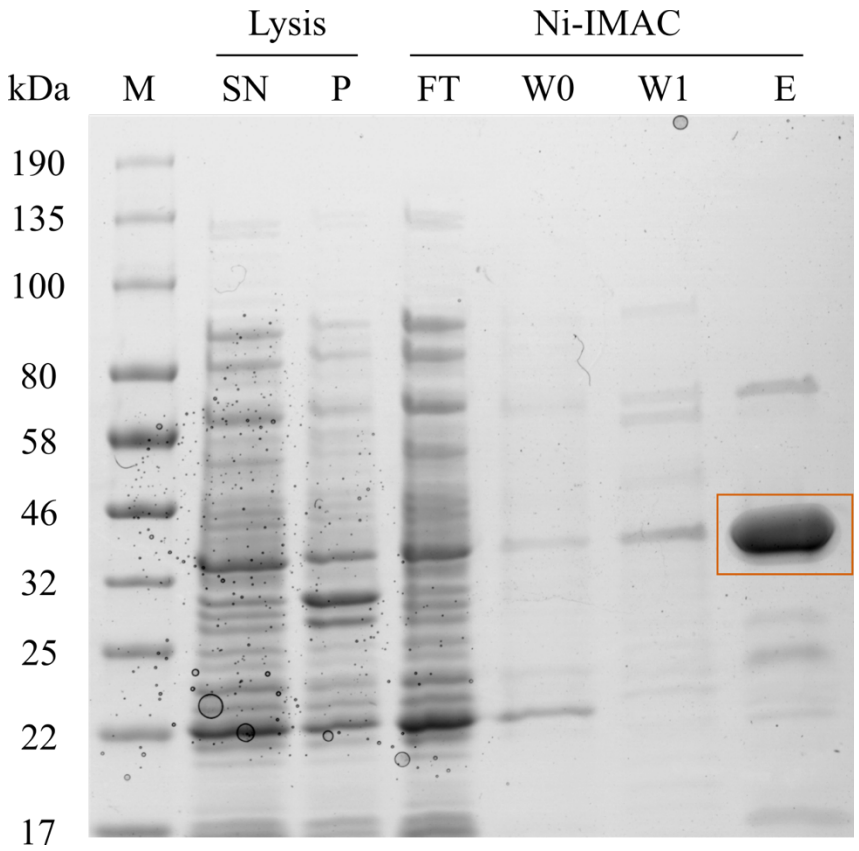


Figure 23 SDS-PAGE of Ni-IMAC purification of pIBA3-SpyCatcher-sfGFP – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), wash 0-1 (W0-W1), and elution (E). Expected size is about 41 kDa, which matches well with the apparent band (orange box).

3.2.3 Translocation efficiency measured with the SpyTag/SpyCatcher system

Only successfully translocated passenger domains expose the SpyTag to the outside of the cell thereby allowing the formation of a covalent bond between the SpyTags on the N-terminal end of the passenger domain and the SpyCatcher-sfGFP molecules in solution. Since the cells are still intact, one can remove free SpyCatcher-sfGFP in solution by several rounds of centrifugation and replacing the supernatant with 1 x PBS until only covalently-bound SpyCatcher-sfGFP is left attached to the cells. Finally, the total fluorescence of the resuspended solution is measured (excitation 488 nm/detection 510 nm) along with the OD_{600} value. This is to estimate the cell density in order to normalise the fluorescence signal to the cell amount. Each construct with a SpyTag was measured and compared to a WT sample that was treated in the same way but did not carry a SpyTag (**Figure 24**). The expected negative control of A1085P produced a lower value than the rest of the SpyTag constructs but higher than the expected value for a negative control. Deletion of the stalk domain (ΔCC12 Spy) seems to increase the amount of translocated passenger domains/cell by almost threefold compared to WT Spy. This is consistent with the expectation of the stalk domain being the main contributor to the aggregation properties of the passenger domain of BpaC.

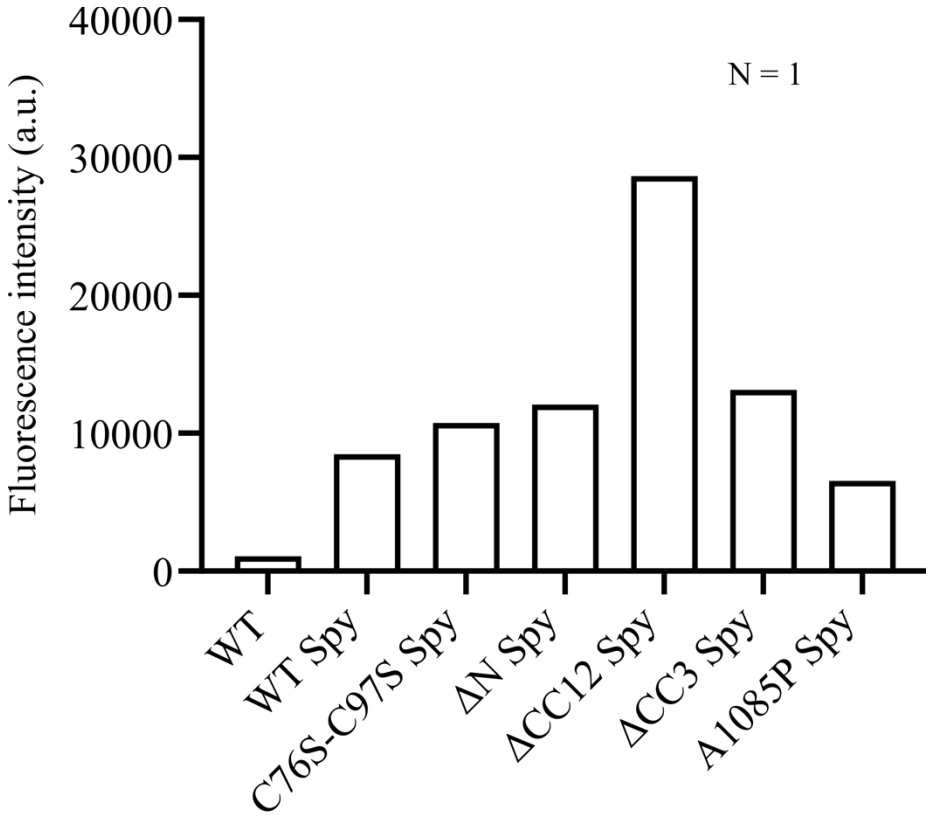


Figure 24 Fluorescence readout of SpyTag/SpyCatcher translocation assay – Fluorescence was excited at 488 nm and measured at 510 nm. Values were normalised to the cell density (OD_{600}). Only a single value for each sample was taken for this graph due to Covid-19 lab access restrictions. Therefore, no statistical analysis was performed with these values. WT without SpyTag (WT) was used as negative control for the SpyTag/SpyCatcher reaction. The additional negative translocation control of the A1085P substitution mutant shows a higher value than expected, albeit being lower than for WT Spy.

3.2.4 Periplasmic stress assay as orthologous analysis method

The results for the SpyTag/SpyCatcher assay exposed the supposed negative control A1085P as inconclusive as a control for this type of experiment. Therefore, an orthologous approach was needed to verify the positive results obtained from that experiment. Instead of focussing directly on the translocation of the passenger domain, one can indirectly estimate the success of the translocation by measuring the periplasmic stress, that is the accumulation of misfolded protein in the periplasm which leads to an overexpression of chaperones and proteases. Some of the periplasmic stress response factors are under the control of the so-called P*stress* promoter in *E. coli*. In this periplasmic stress assay, a fluorescent fast-folding probe called mNeonGreen was genetically engineered to be under P*stress* promoter control and as a result can be used as a direct measure of periplasmic stress. This system, which was originally intended to measure the effect of small molecule inhibitors targeting BAM, was adapted by me to measure if there is a correlation between BpaC expression (WT and mutants) and periplasmic stress (**Figure 25**).

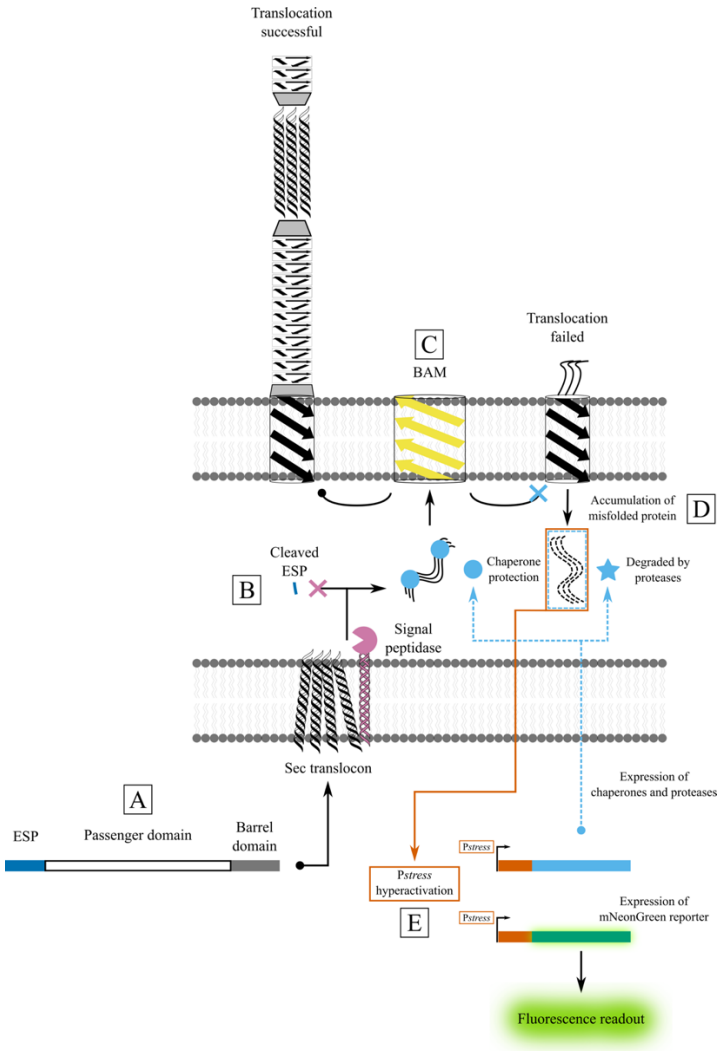


Figure 25 Periplasmic stress assay adapted for BpaC translocation study – **A** Construct of interest is expressed in the cytoplasm of *E. coli* BL21Star cells and passed onto the Sec translocon in the bacterial inner membrane in an unfolded manner. **B** The extended signal peptide (ESP) at the N-terminal end of the protein is recognised by the Sec translocon and as consequence the whole peptide chain is passed through the inner membrane. The ESP is then cleaved off by a membrane-bound signal peptidase associated with the Sec translocon complex. Chaperones keep the peptide chain in an unfolded but protected state. **C** BAM recognises the TAA chains and with a still poorly understood mechanism incorporates the TAA as a trimer into the bacterial outer membrane. Successful translocation leads to a fully functional and properly folded passenger domain. **D** Failed translocation leads to the accumulation of misfolded protein in the periplasm which is recognised by chaperones and proteases present in the periplasm. In a feedback loop, the expression of chaperones and proteases is stimulated using the P*stress* promoter system. The same promoter system has been fused upstream of *mNeonGreen*. In essence, the creation of periplasmic stress leads to the overexpression of *mNeonGreen* which can be measured via fluorescence readout at 488 nm/510 nm.

As an initial trial, the periplasmic stress that the induction of pBAD EV (no gene inside ORF), pBAD-BpaC WT, and pBAD-BpaC A1085P, produced was compared measuring the fluorescence readout after 0, +1, +2, and +3 h of induction with 0.1% L-arabinose (488 nm excitation/510 nm detection). These values were measured in triplicate and normalised to the OD_{600} value measured at the same time as the fluorescence readout. I performed a one-way ANOVA with a follow-up Dunnett test in GraphPad Prism to assess the statistical validity of these values. The results show a clear induction of periplasmic stress for the A1085P mutant over the EV negative control at both +2 h and +3 h post-induction (**Figure 26**). Additionally, the WT construct also induces an increase in periplasmic stress compared to the control, but not as much as the A1085P mutant. These results suggest that this assay can be used for the intended purpose and is ready to be used to test other mutants or used with other TAAs.

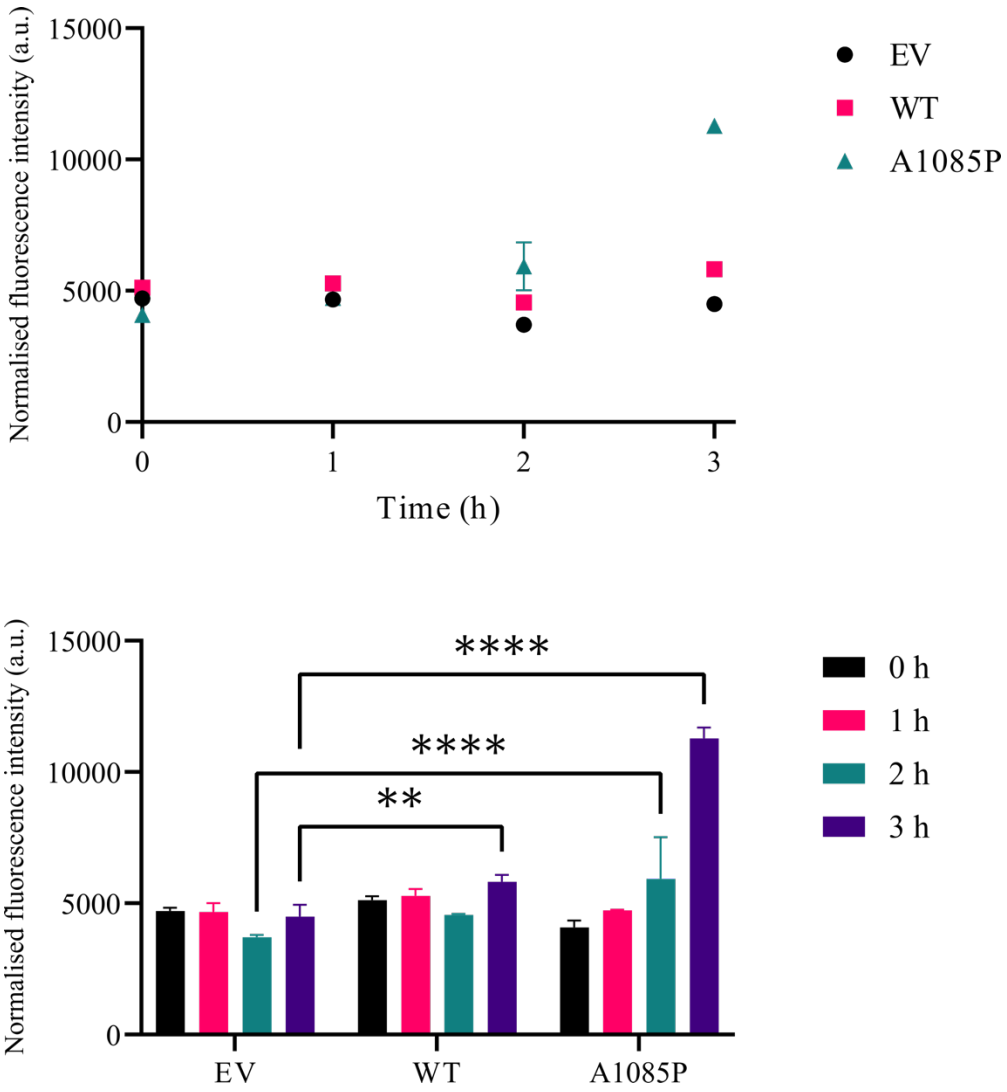


Figure 26 Periplasmic stress assay results – **Top** Fluorescence intensity values (N = 3) correlating to periplasmic stress are normalised to cell density using the OD_{600} value and plotted against time after induction (black 0 h, magenta 1 h, turquoise 2 h, purple 3 h) of protein expression of a negative control sample (EV), BpaC WT (WT) and BpaC A1085P mutant (A1085P). **Bottom** The same results are shown as column graph to better visualise the statistical analysis results for the ANOVA test with follow up Dunnett test, ** $P \leq 0.01$, **** $P \leq 0.0001$, as reported by GraphPad Prism.

3.3 Extracellular matrix assay to screen for potential binding partners

No information on the potential function of BpaC was given at the start of my PhD, so the identification of potential binding partners was a high priority beside elucidating the structure of the protein. A good starting point was the summation of known binding partners in other TAAs. Apart from members of the complement system, a common category of binding partners was that of extracellular matrix proteins, which are easily available in commercialised screens, albeit for a different purpose: Originally these were meant to study the interaction of human cell surface integrins, *in vitro* cell differentiation, and screening potential cell adhesion promoters/inhibitors. One such kit (Sigma-Aldrich, ECM545 kit) is a 96-well plate that comes precoated with different ECM proteins: collagen I, collagen II, collagen IV, fibronectin, tenascin, laminin, vitronectin, and BSA as control. The experimental design was adapted as follows (**Figure 27**): first the POI is expressed under a *pBAD* promoter with 0.1% L-arabinose for 3 h at 37 °C. Second, the plate is activated with 1x PBS, any excess liquid removed, and then the cells are added and incubated for another 3 h at 37 °C. An extensive washing step is performed with 1 x PBS and repeated four times to reduce non-specific binding. Finally, the cells are lysed using a lysis buffer/dye mixture with the dye binding to free dsDNA. This can then be measured using a fluorometer at the standard GFP setting (488 nm excitation/510 nm detection).

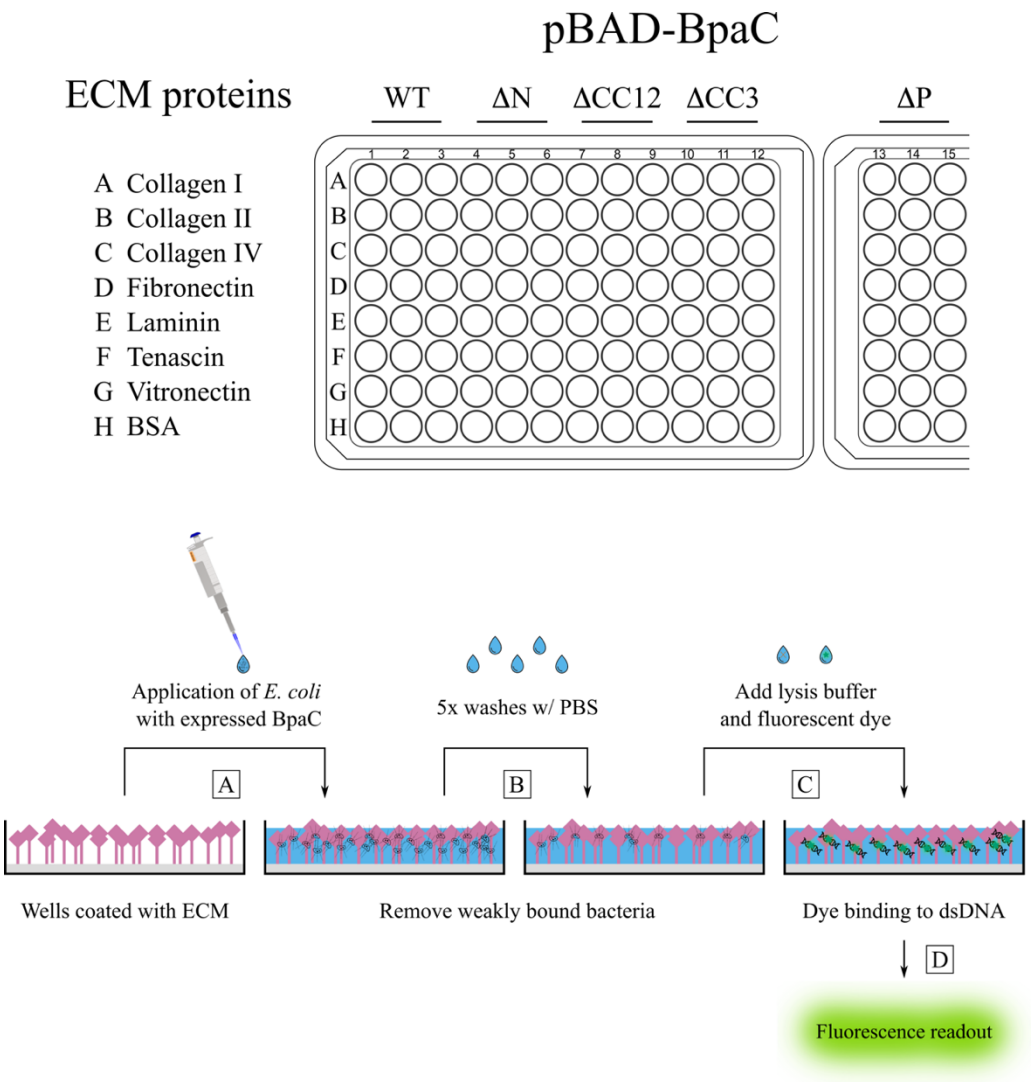


Figure 27 Schematic of adapted ECM cell adhesion array kit for BpaC – **Top** 96-well plate is precoated with different ECM proteins (A-H). Experiments were performed on pBAD-BpaC WT, major deletion mutants for all functional domains (ΔN, ΔCC12, ΔCC3), and the control mutant, which lacks all solvent-accessible domains (ΔP). All experiments were performed in triplicate with repeated biological triplicates (3 x 3). **Bottom A** Wells that are precoated with ECM proteins and activated with 1x PBS are incubated with *E. coli* BL21Star cells expressing BpaC WT or mutants. **B** Weakly-bound bacteria are removed by continuous washes with 1x PBS and a gentle tap of the plate on an absorbing sheet after each step. **C** Premixed lysis/dye buffer is added to the remaining cells, which are chemically lysed. **D** The fluorescent dye binds to the now free dsDNA, which can be quantified by a fluorescence readout at standard GFP settings (488 nm excitation/510 nm detection).

3.3.1 Initial lysis test to check the kit adaptation to bacterial cells

Since the ECM cell adhesion array kit was originally designed for mammalian cells the chemical lysis may not be as efficient for bacterial cells. Incomplete lysis could have ramifications for the interpretation of the resulting data so a complete and reproducible lysis performance at a variety of different cell densities was required as a first test before using the kit for ECM binding studies with *E. coli*. A simple initial test was to follow the protocol as outlined before but using different predefined cell densities (OD_{600} values) and measuring fluorescence (standard GFP setting, 488 nm excitation/510 nm detection) at different time points (**Figure 28**). If there is linear correlation between cell density and fluorescence readout then this means the lysis is reproducible at different cell densities and therefore complete or at least plateauing at a similar level. Time points (15 and 60 min) were also compared to see if the lysis is already completed after 15 min or if a longer incubation time is required for *E. coli* cells.

In conclusion, the goodness-of-fit measures for the linear regression models (R^2) for both observed time points were above 0.99 which indicates a complete and/or reproducible lysis procedure for the lysis/dye buffer provided by the ECM assay kit. Furthermore, the largest change observed of relative fluorescence was less than 0.1 which suggests that an incubation time of 15 min was sufficient for complete and/or reproducible lysis for *E. coli* with the provided lysis/dye buffer. Overall this shows that the lysis/dye buffer in question is suitable for the following microbiological experiments.

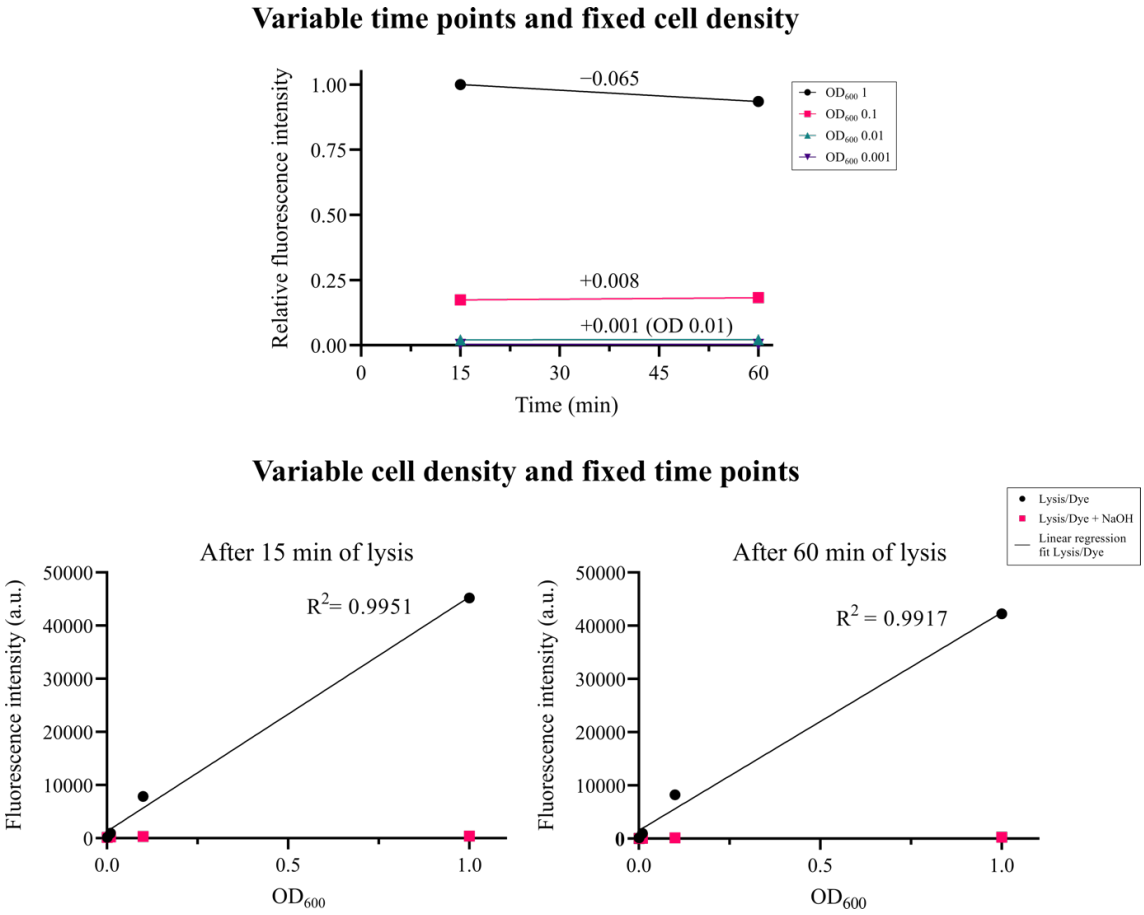


Figure 28 Lysis test of *E. coli* cells with ECM assay kit buffer – Incubation of lysis/dye buffer (which was part of the ECM assay kit) was tested on *E. coli* cells and displayed with different variable and fixed factors to highlight the different conclusions drawn from this test series. **Top** Highest fluorescence value measured was set as 1 and the remaining values calculated as relative fluorescence intensity. Relative fluorescence intensity is shown with fixed cell density and variable time points (15 and 60 min). Cell density (OD_{600}) range is covered by a factor of 10x from 1 down to 0.001. Relative change of fluorescence intensity is shown between 15 min and 60 min (Result for OD_{600} 0.001 not shown as there was no change observed for this cell density). **Bottom** Fluorescence values given as arbitrary units and plotted against cell density (OD_{600}) at fixed time points of observation. Standard lysis/dye buffer is compared to lysis/dye buffer with added NaOH as negative control (denaturates DNA which the dye is binding to). A linear regression model was applied to the data points for the standard lysis/dye buffer condition. The goodness-of-fit for this model is given as fraction of 1 (R^2) with 1 describing a perfect fit of the model to the given data points.

3.3.2 Binding of pBAD-BpaC WT and mutants to ECM cell adhesion array kit

The data obtained from the binding of pBAD-BpaC WT and relevant deletion mutants to the ECM cell adhesion array kit can be viewed and interpreted in different ways. Focussing on the ECM proteins themselves first, one can observe a clear trend for the binding of all BpaC constructs tested to collagen I, II, and fibronectin (**Figure 29**). A more detailed look at the different interactions to the individual ECM proteins is given on the following page.

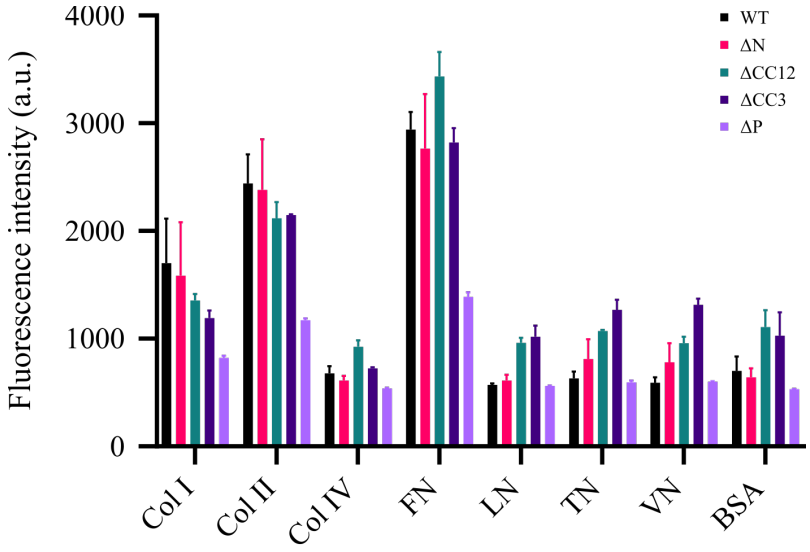


Figure 29 Overview of results of BpaC WT and mutants binding to ECM proteins – Fluorescence intensity of dsDNA binding dye for pBAD-BpaC WT and relevant deletion mutants is plotted against the different ECM proteins coated in each row of the assay (WT black, ΔN red, ΔCC12 green, ΔCC3 dark purple, ΔP light purple). The different ECM proteins are: collagen I (Col I), collagen II (Col II), collagen IV (Col IV), fibronectin (FN), laminin (LN), tenascin (TN), vitronectin (VN), and BSA. No statistical test results are shown in this overview to minimise visual noise but a clear binding preference for fibronectin, collagen II, and to a lower degree collagen I can be observed.

To create a more direct comparison of the effects of WT and deletion mutant binding, the individual values were split into separate graphs. Here, the fluorescence intensity of the individual constructs is plotted against all ECM proteins, either grouped by BpaC construct (**Figure 30**) or by ECM protein (**Figure 31**). All values that are underlying these graphs have been subject to ANOVA tests with follow-up Dunnett tests (setting BSA as control group). The largest effect can be observed for the binding of all constructs to Fibronectin with a large mean difference compared to the BSA mean value (N = 3 x 3, with appropriate error propagation; Dunnett test, **** $P \leq 0.0001$). Even for the supposed control construct of ΔP, which lacks the complete passenger domain, one can observe a statistical mean difference of fibronectin binding compared to BSA (Dunnett test, * $P \leq 0.05$). The second largest mean difference effect was observed for collagen II, with BpaC WT and ΔN having the largest effect (Dunnett test, **** $P \leq 0.0001$), and BpaC ΔCC12 and ΔCC3 with a reduced but still significant mean difference (Dunnett test, ** $P \leq 0.01$). For collagen II, ΔP did not show a significant mean difference (Dunnett test, $P > 0.05$). Lastly, BpaC WT and deletion mutant binding to collagen I was observed to have a large binding effect only for WT and ΔN (Dunnett test, ** $P \leq 0.01$). In conclusion, not a single domain could be identified to contribute the most or least to a specific binding event, especially since they are expected to be structurally diverse. The implications of these observations and the consequences for future approaches to this problem are discussed in section 6.1.5.

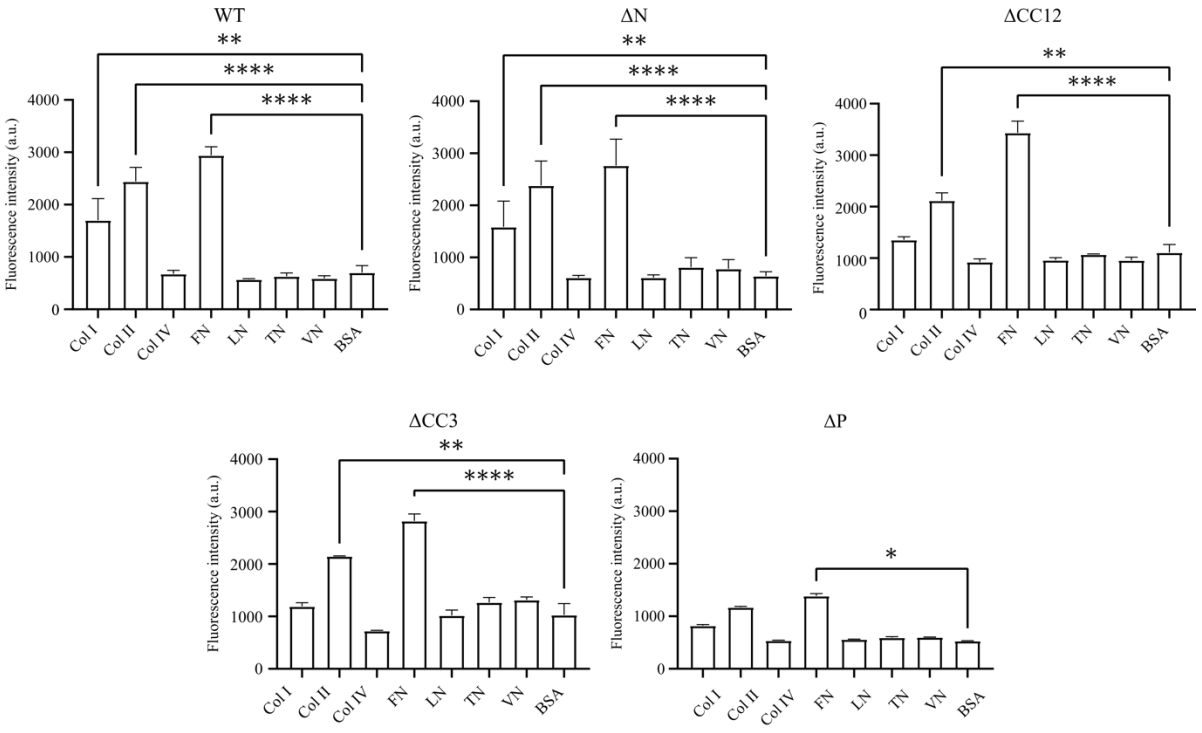


Figure 30 Individual results of ECM binding assay grouped by BpaC construct – Fluorescence readout of dye binding to dsDNA released from lysed cells, expressing different BpaC constructs, that were bound to different ECM proteins. Values are split into individual graphs to show the impact of a specific construct on the binding of all ECM proteins tested. ANOVA tests with follow-up Dunnett tests were performed for each group of values (* $P \leq 0.05$, ** $P \leq 0.01$, **** $P \leq 0.0001$). Each sample was tested with three technical replicates and three biological replicates (N = 3 x 3) with averaged mean and standard deviation calculated with appropriate error propagation.

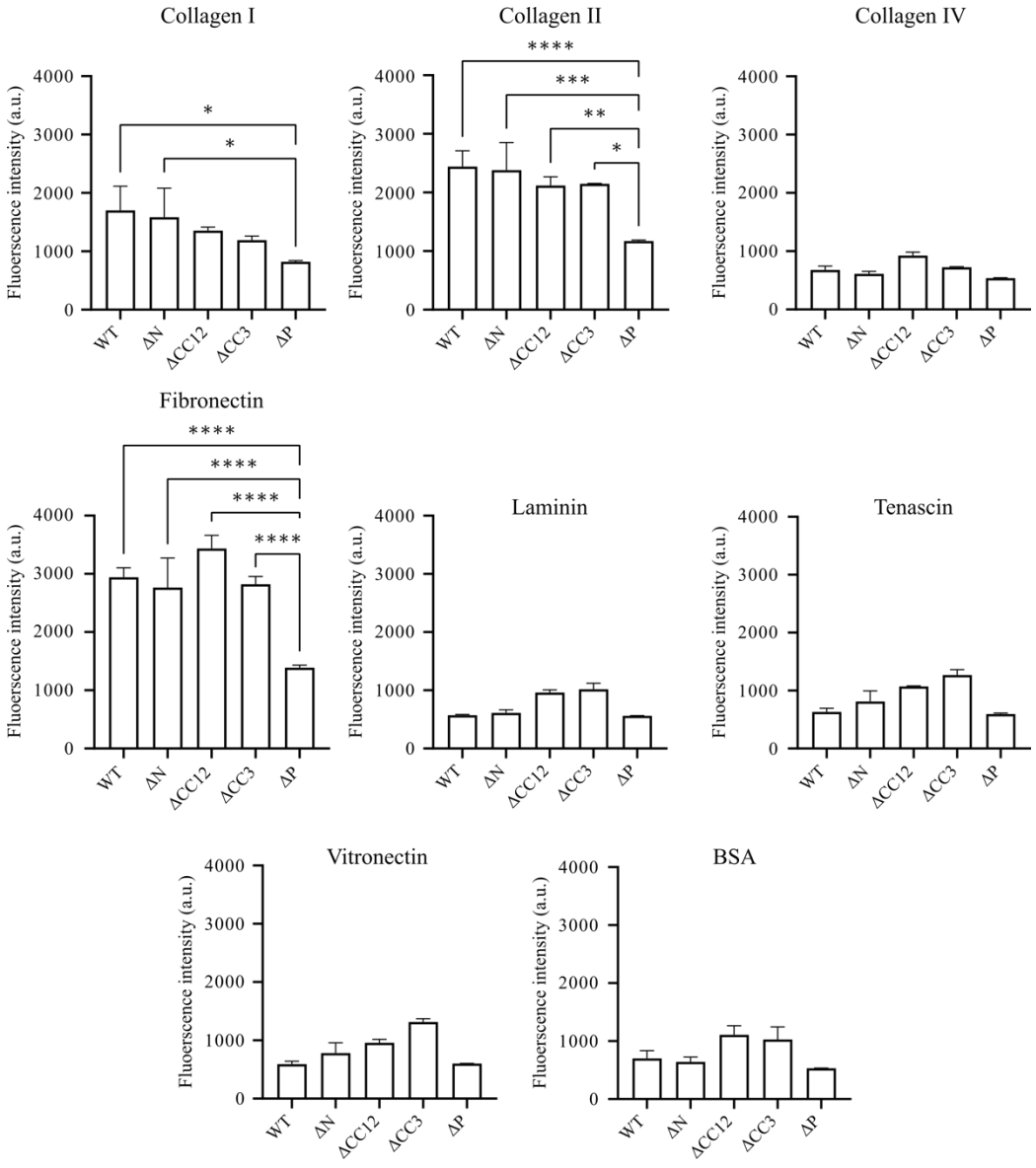


Figure 31 Individual results of ECM binding assay grouped by ECM protein – Fluorescence readout plotted against BpaC WT and deletion mutants, grouped by ECM protein. ANOVA tests were performed with follow-up Dunnett test (ΔP as control group). For collagen I, a significant mean difference was observed only for BpaC WT and ΔN compared to ΔP (* $P \leq 0.05$), while collagen II and fibronectin all had significant mean differences of all constructs observed compared to ΔP (* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$). No statistically significant difference was observed for the means within the remaining ECM protein data groups compared to ΔP.

3.4 Pathogenicity island of *bpaC*

To complete the picture of bioinformatic analysis on *bpaC*, I used a further resource that was available to me: the *Burkholderia* database (Winsor, Khaira et al. 2008). This is a gene database with curated entries for various *Burkholderia* species and subspecies. Especially the linkage to adjacent genes and functional annotations was interesting in this context as to elucidate the potential role that *bpaC* is playing in the wider infection context of *B. pseudomallei*. An overview of the most relevant genes with locus tag, gene product and functional prediction is given in **Table 14**. Another layer of information is provided by the IslandViewer 4 tool (Bertelli, Laird et al. 2017), which is available for some entries in the *Burkholderia* database: It provides information about functionally related genes, which are grouped into so-called pathogenic islands that are relevant during the infection process of the pathogen. For *bpaC*, a group of genes is reported to be linked together: members of an expression regulation system likely controlling the expression of the surrounding genes including *bpaC* and genes that are part of a chaperone/usher fimbril adhesion machinery (**Figure 32**). The implications of these findings are discussed in section 6.2.

Table 14 Overview of adjacent genes of *bpaC*.

Locus tag	Refseq ID	Gene product	Function/GO term
BP1026B_I1569	WP_004521426	Fimbrial subunit	Fimbrial-type adhesion domain
BP1026B_I1570	WP_004527078	Fimbrial chaperone protein	PapD-like
BP1026B_I1572	WP_004527078	Usher protein	PapC-like
BP1026B_I1573	WP_004554117	Type-1 fimbrial protein	Pilin (type 1 fimbria component)
BP1026B_I1575	**WP_014696818**	**BpaC**	**Pathogenesis**
BP1026B_I1577	WP_004193126	DNA-binding response regulator	DNA-binding response regulator
BP1026B_I1578	WP_004531338	Two-component regulatory system, sensor kinase protein	Phosphorylation, signal transduction
BP1026B_I1579	WP_004550403	EAL domain-containing protein	EAL domain, signalling protein

Gene annotation of adjacent functional clusters of *bpaC* which either are selected as part of the predicted pathogenicity island (I1569-I1575) or the possible association is inferred by literature reference (I1577-1579). Most information was retrieved from the *Burkholderia* genome database with entry ID mentioned as locus tag (Winsor, Khaira et al. 2008). Sequence cross-reference to the NCBI Reference Sequence Database via Refseq ID (O'Leary, Wright et al. 2016). Functional assignment was retrieved from InterPro (Blum, Chang et al. 2021) and summarised.

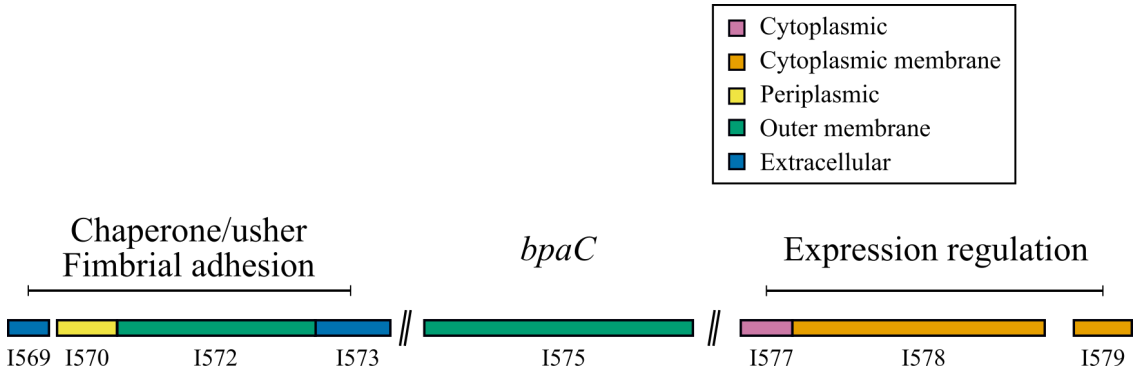


Figure 32 Pathogenic island surrounding *bpaC* – Data was extracted from IslandViewer 4 entry associated with *bpaC* in the Burkholderia database (*B. pseudomallei* strain 1026b, locus tag BP1026B_I1575). Bars representing individual genes are coloured by cellular localization of expressed product and can be identified by their locus tag which has the shortened format BP1026B_X (X = number shown in figure). Bigger space in between the genes (>1 kb) was shortened for better visualisation (double slash). Genes are grouped together by functional annotation using InterPro.

4 Chapter 4: Protein purification optimization of cytosolical expressed BpaC domain constructs for structural studies

Large quantities (mg range) of a well-folded, highly pure protein-of-interest (POI) is needed to have a good chance of success in obtaining protein crystals that diffract to less than 3 Å resolution in structural determination attempts by X-Ray crystallography. The domain border definitions and sequence characterization performed in the last chapter are the basis of creating various constructs that take a few considerations into account on the way to an optimal construct for each domain of BpaC: what is the ideal tag position? What is the ideal size range for a construct also in relation to the included domain residues? What are domain-specific features that can help or hinder the stability/solubility of the POI? Can you fuse different domains together to get an even better purification result? Are there alternative purification methods, other than IMAC, that improve the purity and/or yield of a given construct?

At the beginning of this project there were no available constructs of BpaC or even a report of the successful expression of parts of the protein, other than full length expression for *in vivo* studies. This led to the so-called *divide-and-conquer* approach that was applied to the individual domains of BpaC: I performed a series of *trial-and-error* expression and purification attempts, which ultimately led to a well-purifiable construct for each domain in a reasonable yield and that does not degrade or significantly aggregate during purification. This process also led to a domain construct that produced crystals that diffracted to 1.4 Å and generated a structural model of parts of the C-terminal head domain of BpaC, revealing a new subcategory of the LPBR fold and a novel

insight into the relation of position and surface charge of this TAA domain category that has not been reported before.

4.1 Expression test reveals vector preference

Affinity tag position in novel protein constructs can make the difference between producing a soluble, well-folded protein or a protein that is targeted by cellular proteases during expression. If the protein is targeted by proteases it can also potentially remove important residues, including the affinity tag itself, making the downstream purification very difficult to complete. For historic reasons, several different vectors and vector families were available in our lab stock at the beginning of this project. Frustratingly, most of these stock vectors proved unreliable at best when it came to finding reproducible expression conditions. These vectors also came with severely lowered PCR and cloning success rates which slowed down the progress of the project significantly. Ultimately, I decided to use a common expression vector, pET28a, which was not present in our original lab stock. This switch was the basis of high cloning and expression reproducibility that allowed me to continue with my actual research interests.

4.1.1 Identification of problematic vector stocks

At the beginning of this project a variety of different vectors were available that were suitable for expression in *E. coli*. Our lab had a preference for the pOPIN vector suite due to established cloning protocols and the variety of tags available within this family of vectors. Due to historic reasons, I chose the pBPTS vector for the insertion of the domain of interest, as it was used for YadA expression (another TAA) in our lab before. In order to fuse a GCN4 anchor to the first C-terminal head domain fusion construct of BpaC – T914-Q1054(GCN4) – the pIBA vector was chosen for insertion of the domain of interest, as the available vector in our lab already came with the desired GCN4 sequence (gifted by Ina Meuskens/Jack Leo).

Modifying available tags for these vector constructs expanded the optimization space for later purification steps: this included the addition of protease cleavage sites that enabled an increase in purity by including a reverse IMAC step after successful cleavage of the His-tag, addition of repeats of NW_3 to increase UV280 visibility (especially for UV280 invisible areas of BpaC like parts of the C-terminal head domain), additional His residues to increase binding affinity to IMAC resin (His_6 to His_{10}), and linker residues between tag elements or between the POI and the tag.

Already at this first step, a clear variability in cloning efficiency showed that the quality and purity of the available stock vectors was low: even a simple insertion (creating pOPINFW from pOPINF) was a very time-consuming undertaking with unclear reasons for the variability in transformation efficiency. Additionally, a highly variable and unusual growth behaviour could be observed (much longer doubling times compared to expected standard *E. coli* doubling times). After overcoming these initial hurdles, expression tests were performed to scout the preferred expression conditions for the soluble domain constructs of BpaC. In line with previous observations of inconsistency within the available stock vectors, expression test results either were inconclusive (no expression)

or associated with very low expression levels – too low for efficient downstream purification use. An example SDS-PAGE of this series of expression tests is given (**Figure 33**) covering different expression conditions, BpaC residues, and expression conditions. For most of the tested conditions and samples, no expression of the POI could be observed. Only for pBPTS-N487S947, an apparent monomer band could be observed after 4 h of expression at 37 °C.

The creation of the pOPINFW variant underwent a lengthy quality control process, which mainly consisted of several cloning attempts using different PCR conditions and polymerases/primers until finally obtaining a clear DNA sequencing with a high accuracy of nucleotide assignment. This is a strong difference to the consistently poor sequencing results that were returned for each pOPINF and pOPINE vector that was created. To show that it really is the quality of the initial vectors provided, an expression test using stock pOPIN vectors (pOPINF/pOPINE) is shown together with the pOPINFW variation (**Figure 34**). Both pOPINF-S99V208 and pOPINE-S99V208 showed poor expression profiles, probably due to the range of residues used rather than the vector being used. However, the time needed from design-of-experiment to obtaining a reliable expression test result was too long and hindered high throughput attempts which were ultimately needed to test different residue ranges of BpaC and obtain reliable expression results. Although pOPINFW-D386T517 expressed well and reliably, later purification attempts showed that C-terminal tags are highly preferred for TAA soluble domains. This expression comparison only is included here to show the difference between the original stock vectors and vectors that underwent a quality control improvement.

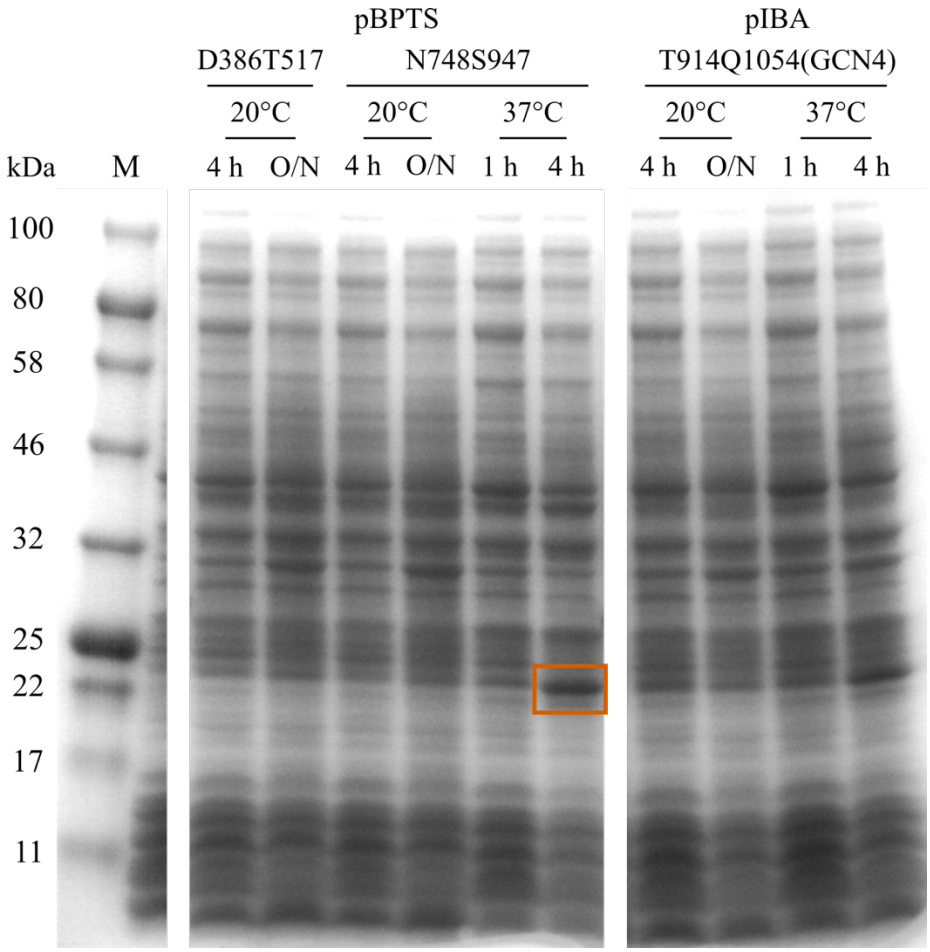


Figure 33 SDS-PAGE of initial expression test for different lab stock vectors – Marker lane (M) with molecular weight standards. Samples are *E. coli* BL21 Star whole cells normalised to each other by an OD_{600} value of 1 at a defined volume and lysed in 1 x SDS loading buffer before being loaded directly onto the gel after a centrifugation step. Top line describes vector backbone name, second line describes the construct cloned into the vector, third line is the expression temperature used, and last line is the expression time in hour or overnight (O/N). Expected molecular weight of D386T517 with associated tag is about 17 kDa as monomer with no apparent band identified; for N748S947 with the same tag, the molecular weight for the monomer is about 22 kDa which matches well with the apparent band identified in the 4 h/37 °C lane (orange box). Lastly, the expected molecular weight for T914Q1054(GCN4) with associated tag as monomer is about 17 kDa with no apparent band identified.

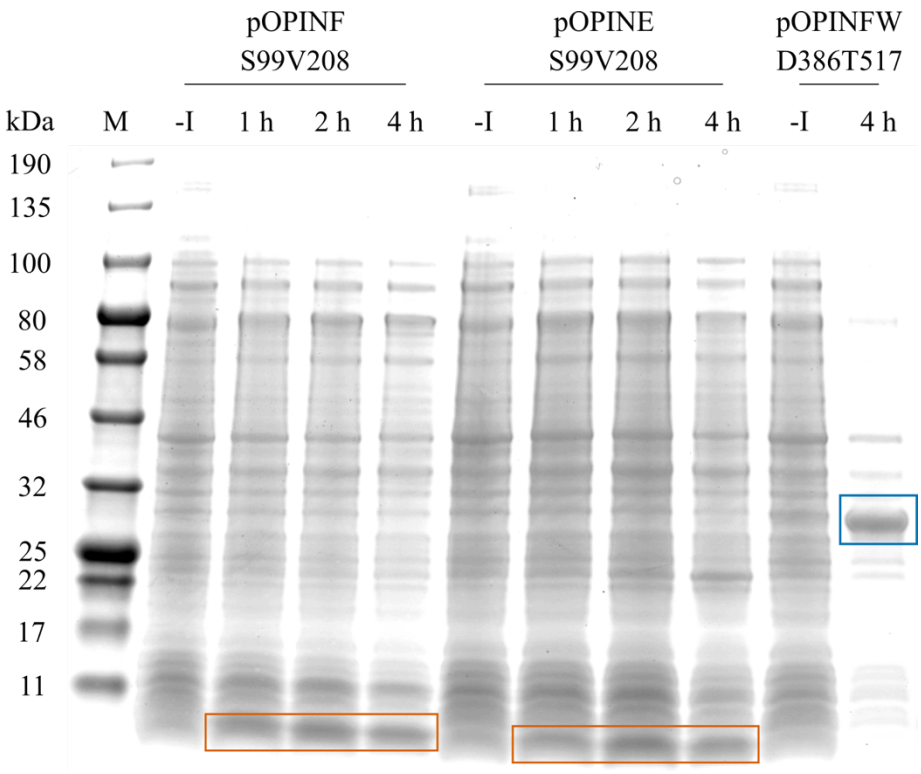


Figure 34 SDS-PAGE of expression test of different pOPIN vector constructs – Marker lane (M) with molecular weight standards. Top line is describing the vector backbone name, second line the name of the construct cloned into the vector, and last line describing the expression time with an additional uninduced control sample (-I). All constructs were expressed in *E. coli* BL21 Star cells at 37 °C and whole cells are normalised as described before and lysed in 1 x SDS loading buffer and directly loaded onto the gel after a centrifugation step. Expected molecular weight of pOPINF-S99V208 is about 13 kDa for the monomer (pOPINE equivalent with about 12 kDa) with apparent bands of what is likely a degradation product below the 11 kDa marker lane (orange box). Expected molecular weight of pOPINFW-D386T517 is about 16 kDa as a monomer with an apparent oligomer band of undefined configuration at about 30 kDa (blue box).

4.1.2 Apparent difference between N- and C-terminal tags for soluble TAA domain constructs

Going through the literature I noticed a strong preference for C-terminal His-tags in the vectors used for TAA soluble domain expression. This led into a simple yet effective comparison between N-terminal and C-terminal tags for a given residue range (**Figure 35**). Degradation of the POI is a clear sign of folding issues, which was observed for the expression profile of pOPINF-F93(C97S)V208. For reference, pOPINF contains an N-terminal His_6 tag with an HRV 3C cleavage site compared to pOPINE with a simple C-terminal Lys-His_6 tag. In comparison, no obvious degradation was seen in the expression profile of pOPINE-F93(C97S)V208. A likely explanation for this is the intricate structural nature of the trimeric arrangement of TAAs, which might be influenced by a nascent peptide chain that is not part of the original sequence that is known to eventually form a trimer. Moving the tag residues interfering with trimerization to the end of the nascent peptide chain would be enough to overcome this issue.

In conclusion, the unreliability of the available vector stocks and the clear preference for a simple C-terminal His-tag made me choose a common bacterial protein expression vector – pET28a. This proved to be essential for obtaining a reliable expression profile and growth properties of the *E. coli* cells transformed with this vector backbone. Ultimately, this allowed a much faster generation time (weeks to merely days) from going from designing a new construct to testing the expression and purifiability of the POI.

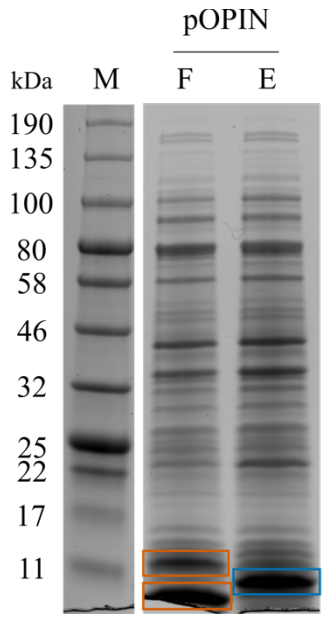


Figure 35 SDS-PAGE of expression test for different pOPIN vectors – Marker lane (M) with molecular weight standards. Samples are expressed in *E. coli* BL21 Star cells at 37 °C for 4 h and normalised via OD_{600} values to each other before lysis and loading onto the gel as described before. The construct cloned within both pOPIN vectors (F for N-terminal tag, and E for C-terminal tag) is F93(C97S)V208 with an expected molecular weight of about 14 kDa (pOPINF)/13 kDa (pOPINE) for the monomer. The upper band (orange box) for pOPINF-F93(C97S)V208 seems to correlate roughly with the expected actual monomer size at about 12 kDa with the lower band being of stronger intensity. For pOPINE-F93(C97S)V208 one can identify a single band for the monomer (blue box), which is slightly lower than the upper band for the pOPINF equivalent of the suspected monomer band.

4.2 IMAC resin type optimisation for BpaC constructs

Different resin types were tested to compare purity versus yield to estimate a general tendency for all constructs.

4.2.1 Purification of pOPINFW-D386T517 with Co-NTA resin

pOPINFW-D386T517 was purified from 846 mg of *E. coli* BL21 Star cells grown in 450 mL of LB medium. For this purification, 500 µL of Co-NTA resin was used. It has a higher specificity for His-tagged proteins but a lower binding capacity than Ni-NTA. This resulted in a purity for the IMAC elution fraction of over 0.99 (**Figure 36, E2**). The estimated size of the monomer is about 21 kDa. The apparent size of both monomer and trimer in SDS-PAGE is higher than expected at about 27 and 81 kDa. The trimer band is visible but the majority of the protein is in the monomer band. The supernatant:pellet ratio was 1.07. The final yield was 0.14 mg/L culture. Although the purity was high the yield was far below what would be needed for attempting protein crystallisation trials. A switch to Ni-NTA resin for further purifications seemed logical.

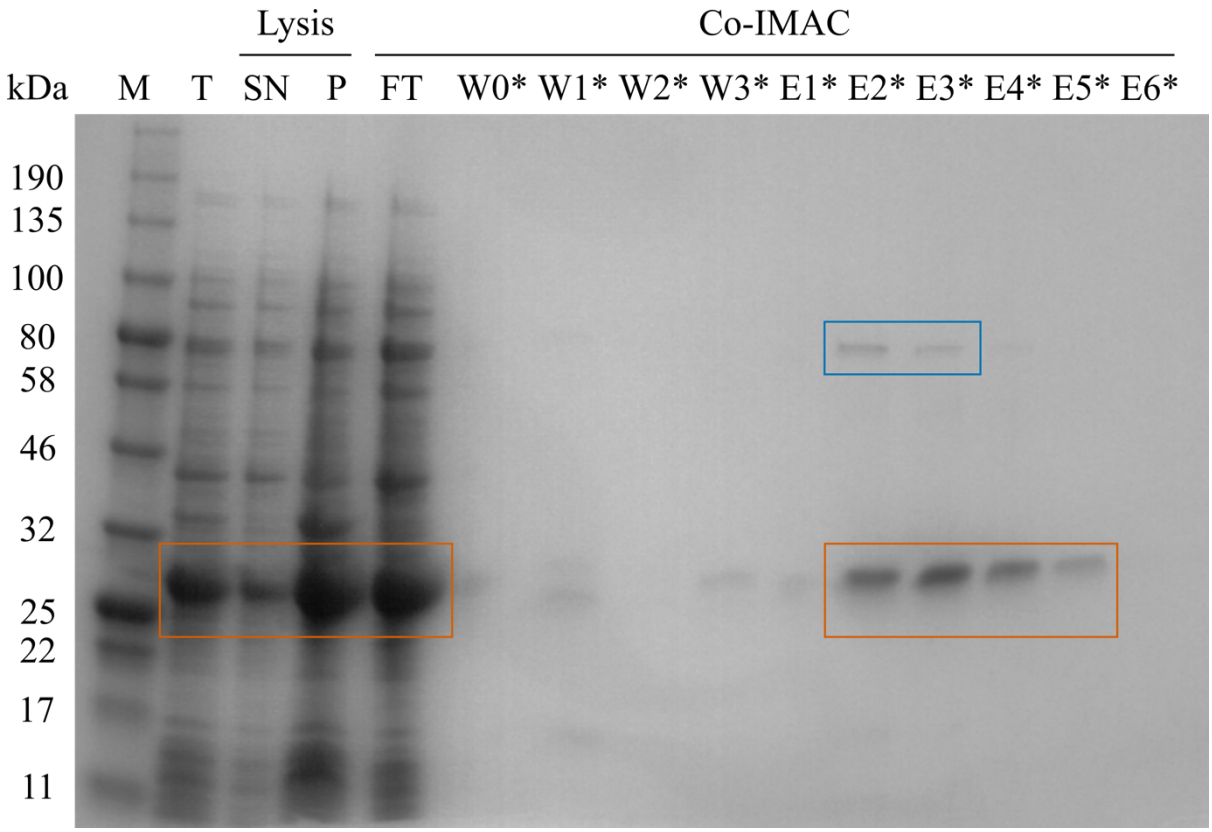


Figure 36 SDS-PAGE of Co-IMAC purification of pOPINFW-D386T517 – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Total cells (T), Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-2 (W0-W2), elution 1-6 (E1-E6). Apparent monomer band at + 6 kDa than expected size (orange box), apparent trimer band at three times the apparent monomer size (blue box). Both wash and elution fractions have been concentrated using the acetone precipitation method (*).

4.2.2 Purification of pOPINFW-D386T517 with Ni-NTA resin

An attempt to extract non-bound protein from the previous purification attempt was performed. For this, the flowthrough of the Co-NTA IMAC purification described in the previous section was reapplied to a Ni-NTA resin. 1 mL of resin was used to achieve a higher yield than before. This resulted in a purity for the IMAC elution fraction of about 0.8 with additional bands around 80 kDa and below the monomer size band (**Figure 37, E2**). Final yield was 0.84 mg/L culture after concentrating the dialysed IMAC elution fractions for crystallisation attempts. This was a sixfold increase in yield compared to the Co-NTA purification attempt but with a significant drop in purity.

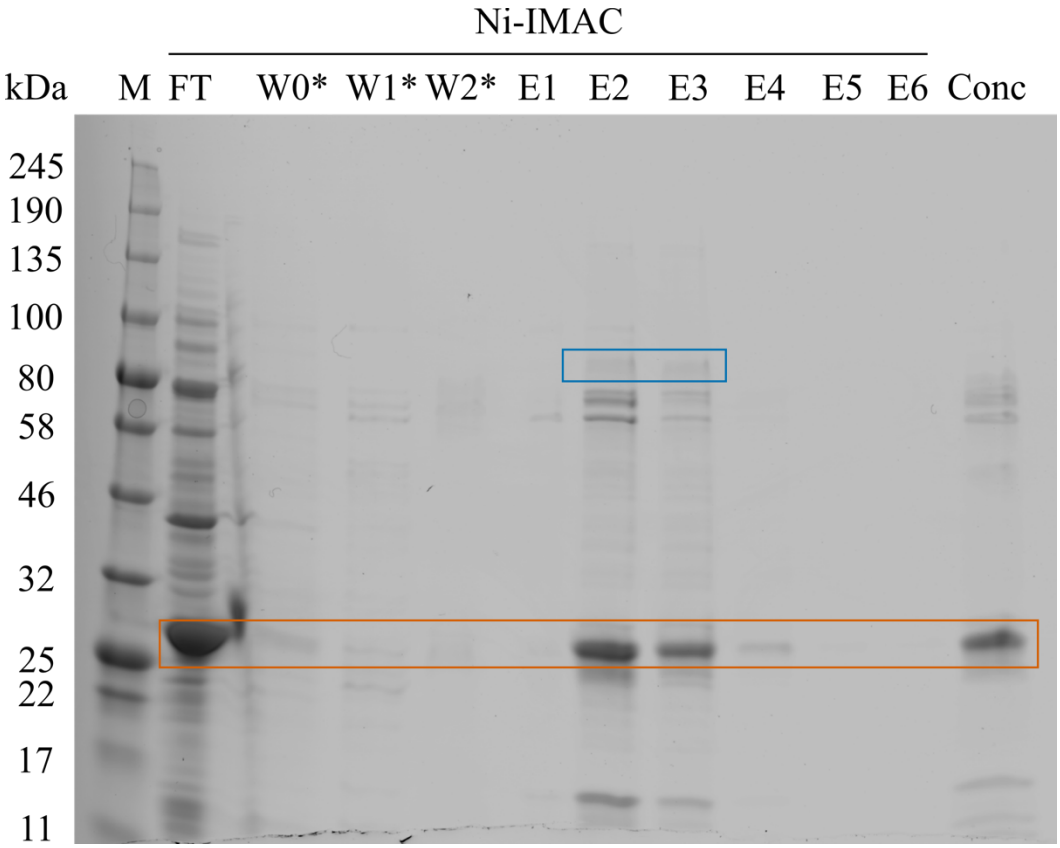


Figure 37 SDS-PAGE of Ni-IMAC purification of reapplied pOPINFW-D386T517 – Marker lane (M) with molecular weight standards. IMAC: Flowthrough (FT), washes 0-2 (W0-W2), elution 1-6 (E1-E6), concentrated sample used for crystallisation (Conc). Apparent monomer band (orange box) is shown with apparent trimer box (blue box). Additional impurities compared to Co-IMAC purification. Wash fractions were concentrated using acetone precipitation (*).

4.3 HRV 3C tag accessibility of pOPINFW constructs

The N-terminal tag of pOPINFW (my modification of pOPINF) carries an HRV 3C cleavage site. Here, the accessibility of this cleavage site was tested either as part of a purification scheme or in a separate, more systematic approach. Cleavage of the tag can improve the success rate of crystallisation attempts as it removes an unstructured part of the construct. It also is a way to improve the purity of the purification by the addition of a reverse IMAC step after cleavage, retaining contaminants that bind to the IMAC resin in an non-specific way, yet allowing the cleaved product to flow through.

4.3.1 Purification of pOPINFW-D386T517 with on-column HRV 3C cleavage

A further purification attempt on pOPINFW-D386T517 was performed on 633 mg of *E. coli* BL21 Star cells grown in 450 mL of LB medium. The focus in this purification was on the accessibility of the HRV 3C tag and how this affects purity and yield. 4 mL of Ni-NTA resin was used for this purification. The purity of the final concentrated sample (**Figure 38, C**) after HRV 3C cleavage and reverse Ni-IMAC was over 0.99 with no visible trimer band. The results suggest that the tag is only partially accessible for HRV 3C cleavage, as a large amount of protein was retained on the resin (**Figure 38, Reverse IMAC E**). The supernatant:pellet ratio was 0.71. The final yield was 1.65 mg/L culture.

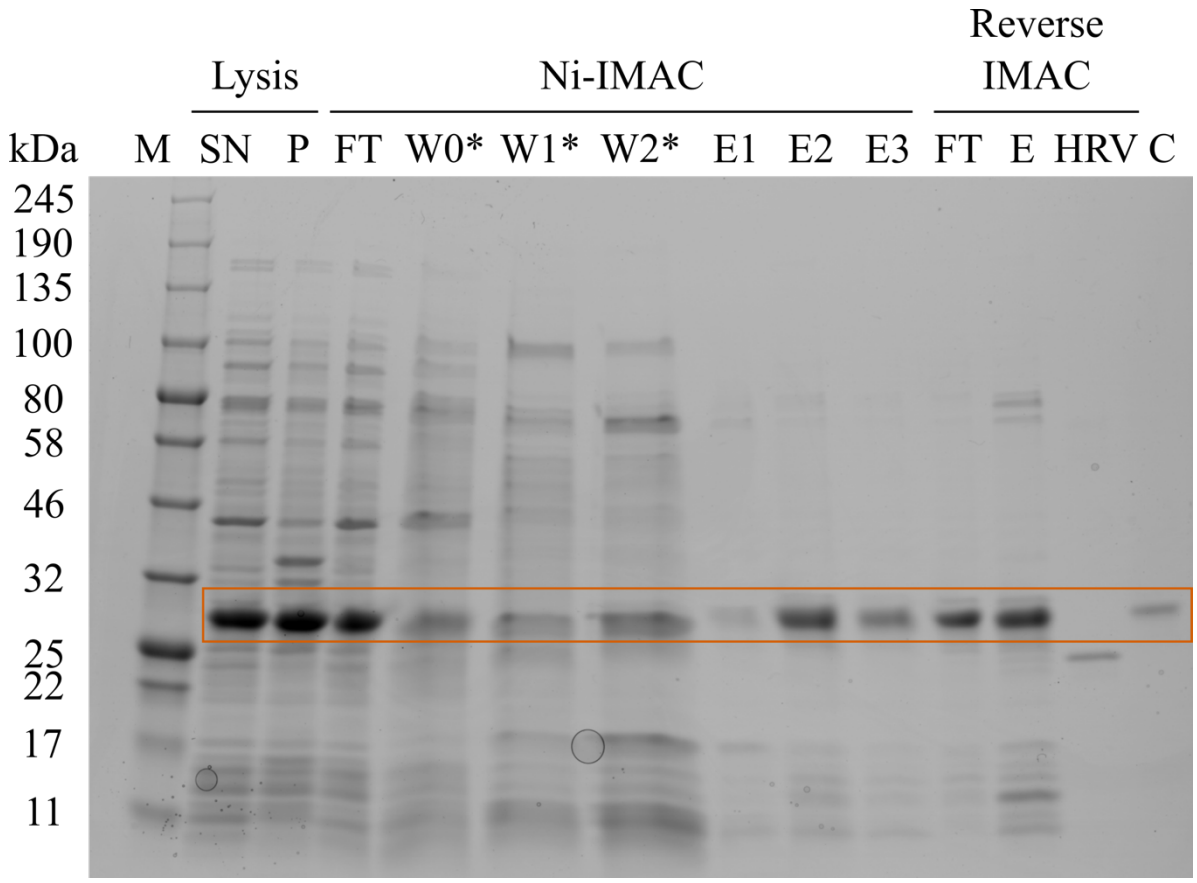


Figure 38 SDS-PAGE of Ni-IMAC purification of pOPINFW-D386T517 including reverse IMAC step – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), washes 0-2 (W0-W2), elution 1-3 (E1-E3). Reverse IMAC after on-column HRV 3C cleavage: Flowthrough (FT), elution (E), HRV 3C only (HRV). Concentrated samples of reverse IMAC flowthrough (C). Apparent monomer band is highlighted (orange box) with no visible trimer band. Same double band contamination around 75 kDa in reverse IMAC elution sample as observed in previous purification. Wash fractions were concentrated using acetone precipitation (*).

4.3.2 HRV 3C cleavage test of pOPINFW-D386T517

The flowthrough of the reverse IMAC from the purification in section 4.3.1 was used for this test. Different ratios of protein:HRV 3C were tested to estimate the optimal ratio for further experiments. No discernible band shift or change in band intensity was observed in the SDS-PAGE for the POI (**Figure 39**). The mass difference of cleaved to uncleaved construct is ~4 kDa, which should be visible on SDS-PAGE. There was no significant level of cleavage even at a protein:HRV 3C ratio of 1:5 with added 500 µM Urea, which was added to improve tag accessibility.

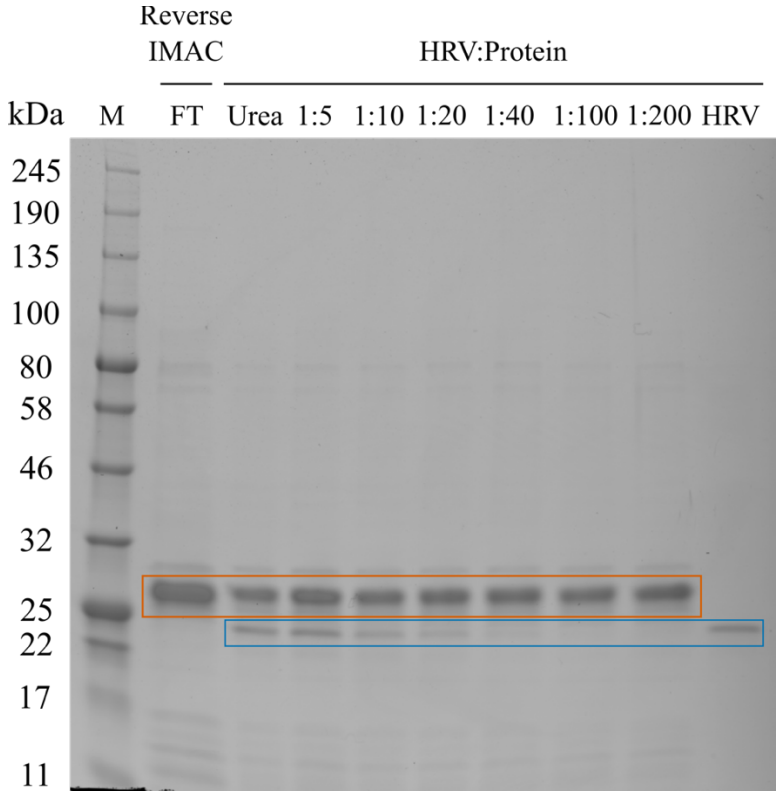


Figure 39 SDS-PAGE of HRV 3C cleavage test of pOPINFW-D386T517 – Marker lane (M) with molecular weight standards. Flowthrough of reverse IMAC (FT) was used for cleavage test. HRV 3C was added at different protein:HRV 3C ratios (1:5 to 1:200). One sample contained a ratio of 1:5 protein:HRV 3C and additionally 500 µM Urea (Urea). HRV 3C was loaded for comparison (HRV, blue box). The estimated monomer band is highlighted (orange box).

4.3.3 Purification of pOPINFWS-N748S947 with in-solution HRV 3C cleavage

Another example for HRV 3C cleavage accessibility was observed for the purifification of pOPINFWS-N748S947, which contains parts of the C-terminal head domain and an additional C-terminal StrepII tag. The construct was purified from 662 mg of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 2 mL of Ni-NTA resin was used for this purification. The expected monomer size is about 23 kDa. The supernatant:pellet ratio was 0.40 with a purity of the IMAC elution fraction of 0.44 for the POI (**Figure 40**). The final yield was 0.28 mg/L culture. This construct stays stable as a trimer in SDS-PAGE but suffers greatly from low purity, likely due to the folding difficulties observed for pOPINFW as seen in other constructs. Cleavage was also not successful even though the cleavage was performed in solution after dialysing out the excess imidazole, which would interfere with cleavage efficiency. This indicates a general problem with the pOPINFW vector, which is likely due to a steric hindrance by the Trp residues in the tag sequence that were added for UV280 visibility.

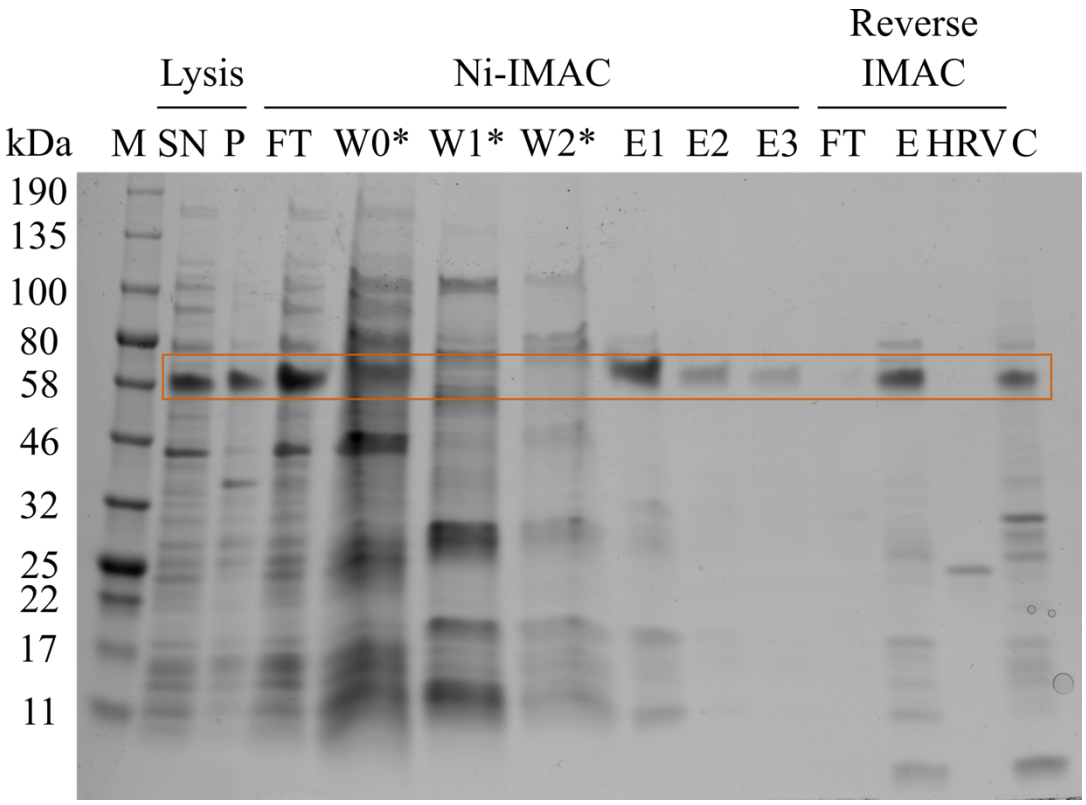


Figure 40 SDS-PAGE of Ni-IMAC purification of pOPINFWS-N748S947 including reverse IMAC step – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), washes 0-2 (W0-W2), elution 1-3 (E1-E3). Reverse IMAC after HRV 3C cleavage of IMAC elution after removal of Imidazole by dialysis: Flowthrough (FT), elution (E), HRV 3C only (HRV). Concentrated samples of reverse IMAC flowthrough (C). Apparent trimer band is highlighted (orange box) with a shift of about − 9 kDa compared to the expected trimer weight of around 69 kDa. Same double band contamination around 75 kDa in reverse IMAC elution and concentrated flowthrough sample as observed in previous purification. Wash fractions were concentrated using acetone precipitation (*).

4.4 Preserving C-terminal neck motif by replacing anchor helix with GCN4

BpaC contains a coiled coil bundle, starting at L1055, residing within the β-barrel. A preceding neck motif was identified that carried the DAVNxxQL motif also seen in other parts of the structure. In this experimental series a construct that included part of the coiled coil bundle (pOPINFW-T914D1097) was compared to a construct that has the native coiled coil replaced by the trimeric mutant of the leucine zipper GCN4: pOPINFW-T914Q1054(GCN4).

4.4.1 Purification of pOPINFW-T914D1097

pOPINFW-T914D1097 was purified from 937 mg of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 500 µL of Co-NTA resin was used for this purification. A lower resin volume was selected in an attempt to increase the purity of the sample. Uninduced versus induced cells are shown on the SDS-PAGE (**Figure 41, left**) to demonstrate the expression of the construct and the location of the monomer band for comparison with the purification gel. Most of the protein was detected in the pellet and a small amount can be seen in the wash fractions of the purification on the purification gel (**Figure 41, right**). The expected molecular weight for the monomer is about 13 kDa. No trimer band was detected and no protein was found in the elution fractions. The presence of high molecular weight bands in the pellet and flowthrough fraction of the purification gel indicates a high aggregation potential in the sample.

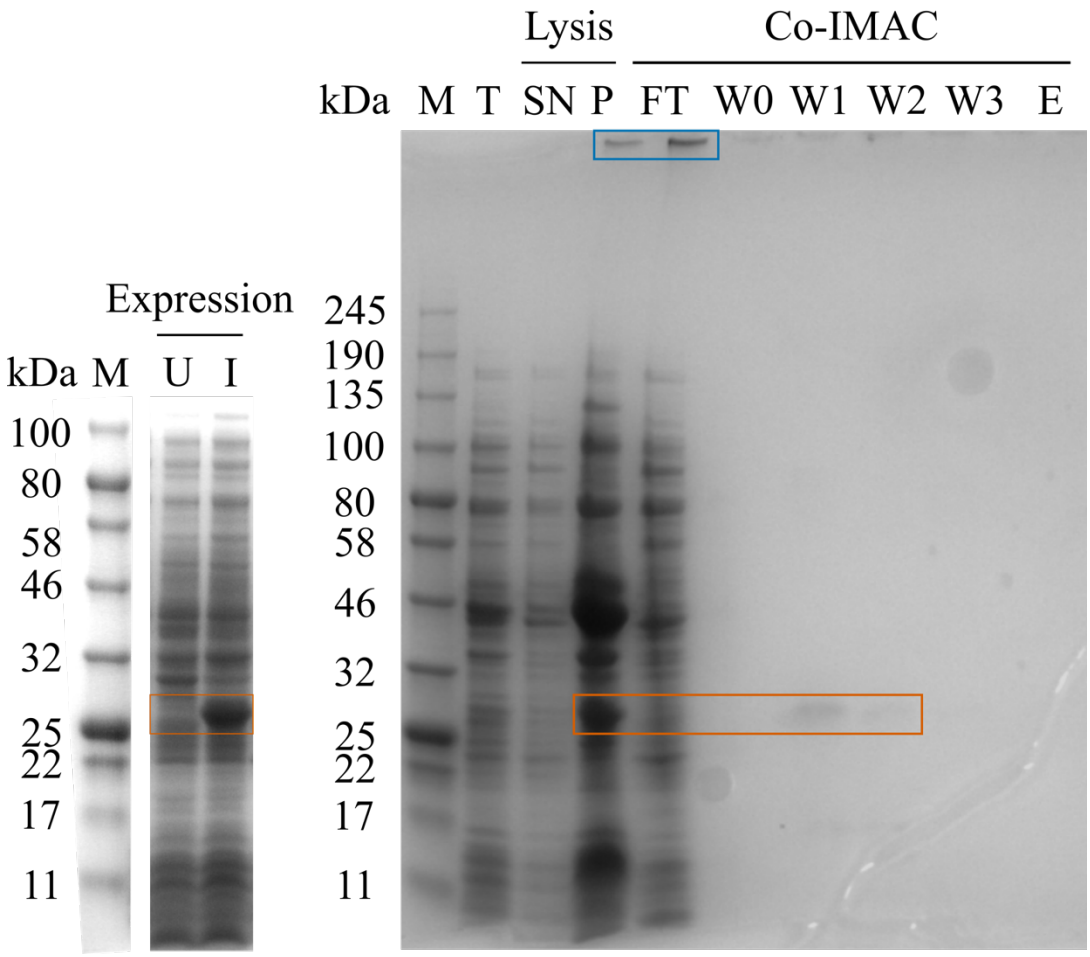


Figure 41 SDS-PAGE of expression and Co-IMAC purification of pOPINFW-T914D1097 – Two marker lanes (M) with molecular weight standards. Left SDS-PAGE showing expression test with uninduced (U) and three-hour post induction (I) BL21 Star cells transformed with pOPINFW-T914D1097 normalised by cell density before loading. Right SDS-PAGE was loaded with total cells (T), supernatant after lysis (SN), pellet after lysis (P). Fractions analysed in the Co-IMAC step: Flowthrough (FT), washes 0-3 (W0-W3), and elution (E). Expected monomer size is 13 kDa, apparent band of undefined oligomeric state is highlighted (orange box). Contaminants which were unable to enter the gel in the pellet and flowthrough fraction indicate a highly aggregating tendency of the construct (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.4.2 Purification of pOPINFW-T914Q1054(GCN4)

pOPINFW-T914Q1054(GCN4) was purified from 1.16 g of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 2 mL of Ni-NTA resin was used for this purification. The expected monomer size is about 20 kDa. The supernatant:pellet ratio was 0.29 with a purity of the IMAC elution fraction of 0.19 for the POI (**Figure 42**). The final yield was 0.18 mg/L culture. This was an improvement compared to pOPINFW-T914D1097 but still had an overall low purity and yield. This is likely due to the use of the N-terminal tag system of pOPINFW and further proof of the unsuitability of this tag for cytosolically expressed TAA constructs.

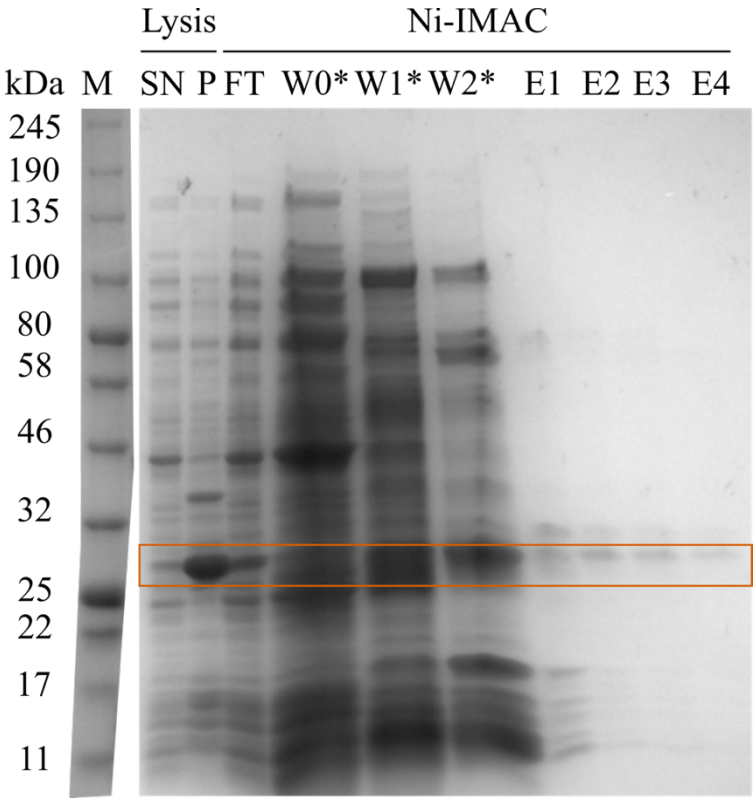


Figure 42 SDS-PAGE of Ni-IMAC purification of pOPINFW-T914Q1054(GCN4) – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), washes 0-2 (W0-W2), elution 1-4 (E1-E4). Expected monomer size is about 20 kDa with an apparent POI band at around 27 kDa (orange box). Wash fractions were concentrated using acetone precipitation (*).

4.5 Purification of optimised pET28a-S741Q1054(GCN4)

Major changes in this construct compared to pOPINFW-T914Q1054(GCN4) are the switch from an N-terminal tag with several features (His_{10}-$[NW]_3$-HRV 3C cleavage site) to a simple C-terminal His_6-tag and the addition of 173 more C-terminal head domain residues. The tag switch became necessary as the HRV 3C cleavage test with pOPINFW-D386T517 (section 4.3.2) showed the inaccessibility of the tag for cleavage. The addition of more residues to the construct was a consequence of the good results obtained for the purification of pOPINFWS-N748S947 (section 4.3.3).

pET28a-S741Q1054(GCN4) was purified from 1.07 mg of *E. coli* BL21 Star cells grown in 900 mL of LB medium and 2 mL of Ni-NTA resin was used for this purification. The expected monomer size is about 32 kDa. For this construct, only the trimer band is visible in SDS-PAGE, which indicates high stability of the construct. The supernatant:pellet ratio is 0.72 with the purity of the IMAC elution fraction of over 0.99 for the POI (**Figure 43**). This results in a final yield of 11.79 mg/L culture medium, which is 2-10 x higher than all the other purifications included in this thesis. The final concentration for crystallisation attempts was above 130 mg/mL, which shows the extreme solubility of this construct.

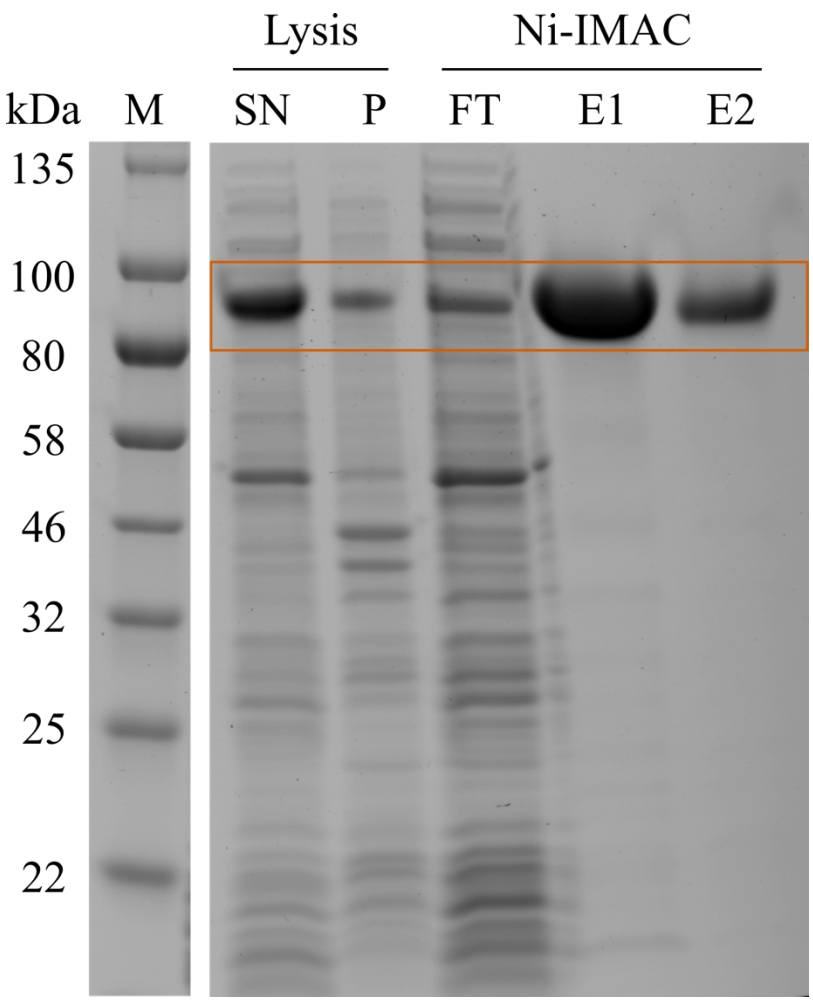


Figure 43 SDS-PAGE of Ni-IMAC purification of pET28a-S741Q1054(GCN4) – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), elution 1-2 (E1-E2). Expected monomer size is about 32 kDa with no monomer band visible. The expected trimer size is 96 kDa, which matches well with the band of the POI (orange box).

4.6 Folding restraints of N-terminal head domain constructs

The challenges, that working with the N-terminal head domain constructs has been, are presented in the following purification series. This is mainly apparent by the low yield of the constructs but also by the appearance of chaperones visible in the SDS-PAGE of both pET28a-G90(C97S)Q173(GCN4) and pET28a-S99(ΔCC1)T517 (blue boxes in **Figure 45** and **Figure 46**), indicating folding issues in these constructs due to their repetitive nature. The neck motifs at the end of the N-terminal head domain and within the stalk domain were also used to make deletions that are highly unlikely to have an impact on proper folding of the construct. These motifs can also be used as fusion point for GCN4 as done before for the C-terminal head domain construct.

4.6.1 Purification of pOPINF-S99V208 and pOPINE-F93(C97S)V208

pOPINF-S99V208 and pOPINE-F93(C97S)V208 differ in tag position (pOPINF = N-terminal His_6-HRV 3C cleavage site, pOPINE = C-terminal His_6) and N-terminal residues as the exact end of the N-terminal domain is undefined (see PSIPRED prediction for N-terminal domain, **Figure 17** in section 3.1.2). A cysteine-to-serine mutation (C97S) was introduced to prevent potential non-specific disulphide bridges. pOPINF-S99V208 was purified from 206 mg of *E. coli* BL21 Star cells grown in 50 mL of LB medium. pOPINE-F97(C97S)V208 was purified from 1.03 g of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 3 mL of resin was used for both purifications. The expected monomer size for pOPINF-S99V208 is about 13 kDa; for pOPINE-F93(C97S)V208 it is also about 13 kDa. For both constructs, almost all of the POI was in the pellet fraction (**Figure 44**). No purity value is given for pOPINF-S99V208 as the IMAC elution fraction is empty in SDS-PAGE. The purity of pOPINE-F93(C97S)V208 is below 0.2 and likely due to non-specific protein that was still bound to the resin after the washes. No yield was determined as there was either no protein in the IMAC elution overall or the purity was far below tolerance for further use.

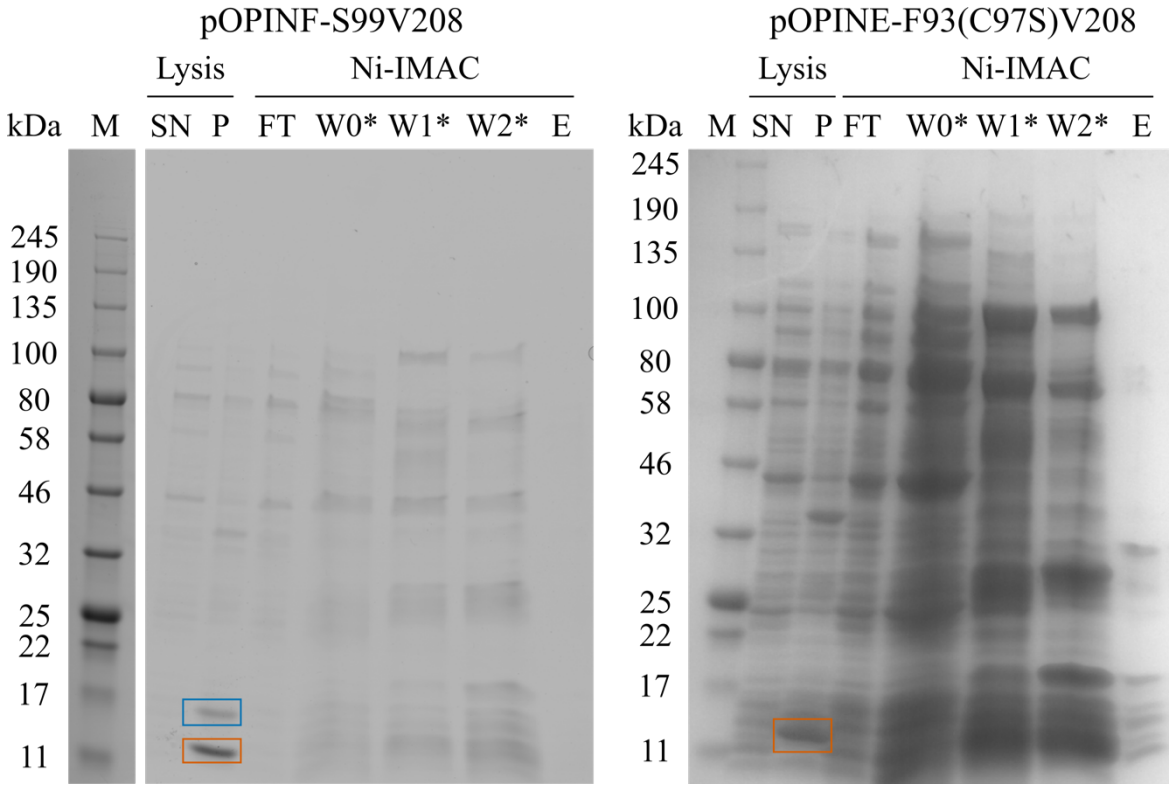


Figure 44 SDS-PAGE of IMAC purifications of pOPINF-S99V208 and pOPINE-F93(C97S)V208 – Two marker lanes (M) with molecular weight standards. Left SDS-PAGE showing Ni-IMAC purification of pOPINF-S99V208. Right SDS-PAGE showing Ni-IMAC purification of pOPINE-F97(C97S)V208. For the lysis step: supernatant after lysis (SN), pellet after lysis (P). Fractions analysed in the Ni-IMAC step: Flowthrough (FT), washes 0-2 (W0-W2), and elution fractions (E). Expected monomer size is about 13 kDa for both constructs with the apparent monomer band highlighted (orange box). An additional band was identified for pOPINF-S99V208 that is either a contaminant or related to the expressed protein (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.6.2 Purification of pET28a-G90(C97S)Q173(GCN4)

pET28a-G90(C97S)Q173(GCN4) was purified from 1.66 g of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 2 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 13 kDa. The supernatant:pellet ratio is 0.79 with a final purity of 0.71 in the IMAC elution fraction (**Figure 45**). The final yield was 0.68 mg/mL with over 90% losses during final concentration step. The presence of a characteristic double band at about 80 kDa in the wash and elution fractions indicates a contamination that is tightly bound to the target protein.

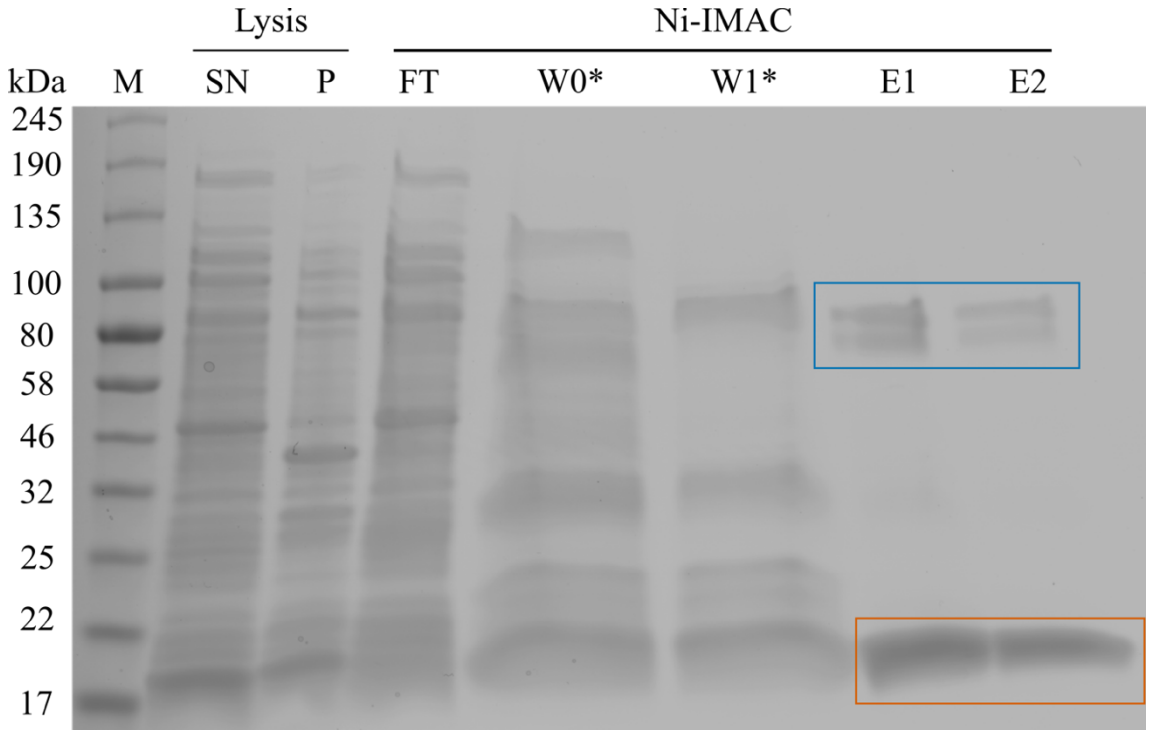


Figure 45 SDS-PAGE of Ni-IMAC purification of pET28a-G90(C97S)Q173(GCN4) – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), washes, elution 1-2 (E1-E2). Expected monomer size is 13 kDa with the apparent band of the POI highlighted (orange box). Double band contamination later identified as chaperones also marked (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.6.3 Purification of pET28a-S99(ΔCC1)T517

pET28a-S99(ΔCC1)T517 was purified from 1.45 g of *E. coli* BL21 Star cells grown in 850 mL of LB medium. 2 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 33 kDa. Similar to pET28a-G90(C97S)Q173(GCN4), a characteristic double band around 80 kDa was identified (**Figure 46**). The supernatant:pellet ratio is 0.61 with a IMAC elution purity of the POI of less than 0.2. No yield was determined as purity was below acceptable threshold for further use.

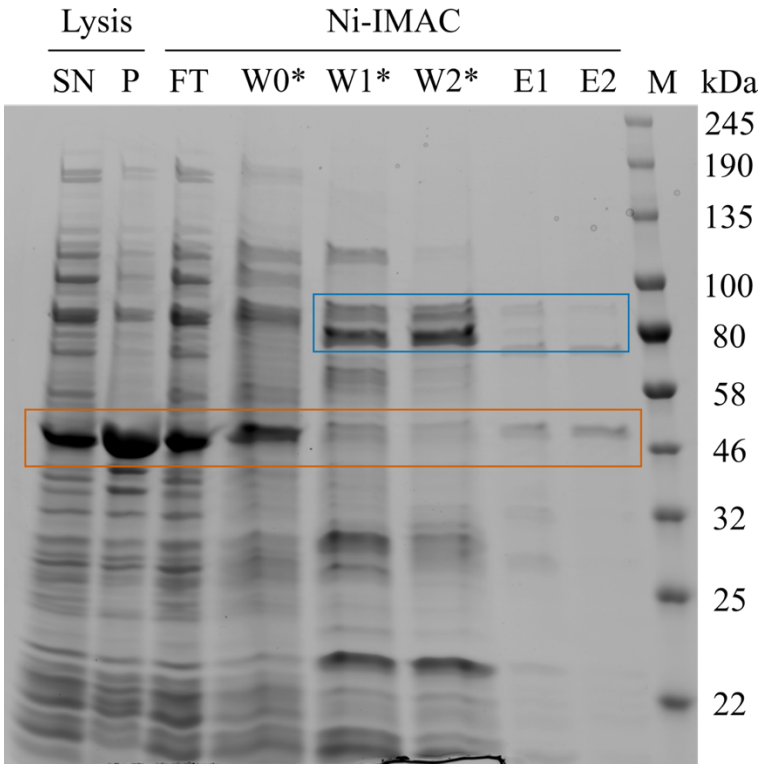


Figure 46 SDS-PAGE of Ni-IMAC purification of pET28a-S99(ΔCC1)T517 – Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC: Flowthrough (FT), washes, elution 1-2 (E1-E2). Marker lane (M) with molecular weight standards. Apparent monomer or undefined oligomer band of the POI (orange box) with a shift of + 17 kDa compared to the expected size of the monomer of about 33 kDa. Double band contamination later identified as chaperones also marked (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.6.4 Identification of contaminating chaperones in N-terminal head domain constructs

The flowthrough of the previous purification of pET28a-S99(ΔCC1)T517 was reapplied to 2 mL of Co-NTA resin to test the presence of chaperones using an $MgCl_2$/ATP wash step. An initial wash step without ATP was performed to serve as comparison and to remove all previous contaminants due to the difference between Ni-NTA and Co-NTA (**Figure 47**). The subsequent wash step, which included $MgCl_2$/ATP, shows the removal of most contaminants including the double bands at 80 kDa. This would indicate that these proteins are likely chaperones that bound to the partially unfolded target protein. The function of chaperones is to help proteins reach their native (folded) state. They do this by binding to unfolded, usually hydrophobic patches of the protein, unwinding the polypeptide chain in an isolated environment, using ATP (which requires Mg^{2+} ions to function) in the process. Chaperones can get "stuck" in this cycle if the protein does not fold and the lysate is devoid of additional ATP to complete the cycle. The additional proteins that came off in this wash step are likely additional contaminating proteins that were adhering to the chaperone/protein complex. The IMAC elution for this purification step shows over 0.99 purity. The identification of the previous double bands as chaperones would also explain why there were great losses during the concentration step. The partially unfolded protein, which had hydrophobic patches protected by chaperones, would have likely still had areas of hydrophobicity that then stuck to the concentrator membrane and futher lead to a large loss of the POI by precipitation onto the membrane. In essence, even by removing the chaperones, the ability to purify these constructs was not restored, likely due to the fact that the protein was unable to fold properly.

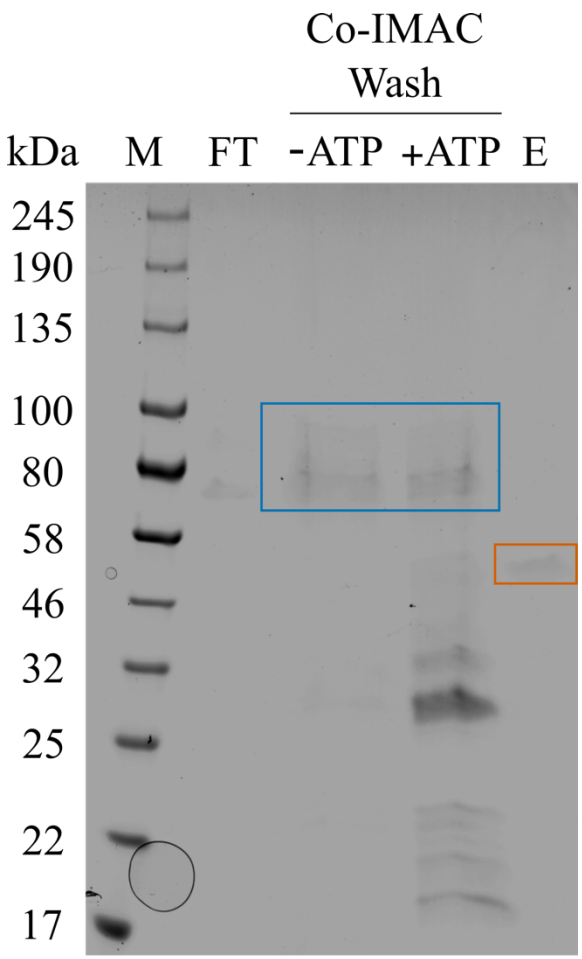


Figure 47 SDS-PAGE of Co-IMAC purification of rebound pET28a-S99(ΔCC1)T517 – Marker lane (M) with molecular weight standards. IMAC step: Flowthrough (FT), first wash step without $MgCl_2$/ATP (-ATP) and second wash step with $MgCl_2$/ATP (+ATP), elution (E). Apparent monomer band or undefined oligomer band of the POI (orange box). Characteristic chaperone double band around 80 kDa (blue box). Note that all fractions except the marker lane have been concentrated using the acetone precipitation method.

4.7 Challenging purification of stalk domain constructs

The purification of constructs containing parts of the stalk domain, particularly stalk I (T175-T252), proved challenging. This was due to the fact that most purifications either failed because all of the POI was in the pellet fraction after lysis or the protein has not been expressed in the first place; likely due to the toxic nature of the constructs on the cells. In this section, two examples are described that showcase the challenge to obtain well-purifiable, high yield constructs for this domain.

4.7.1 Purification pET28a-G90(C97S)Q259(GCN4)

pET28a-G90(C97S)Q259(GCN4), which covers the stalk I area, was purified from 1.22 g of *E. coli* BL21 Star cells grown in 900 mL of LB medium. 2 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 21 kDa. No supernatant:pellet ratio was calculated as almost all of the POI is in the pellet fraction (**Figure 48**). No purity or yield value can be given as well as the IMAC elution fraction was empty. The most likely explanation for this behaviour is that the construct contained significant portions of aggregating areas and was toxic to the cells.

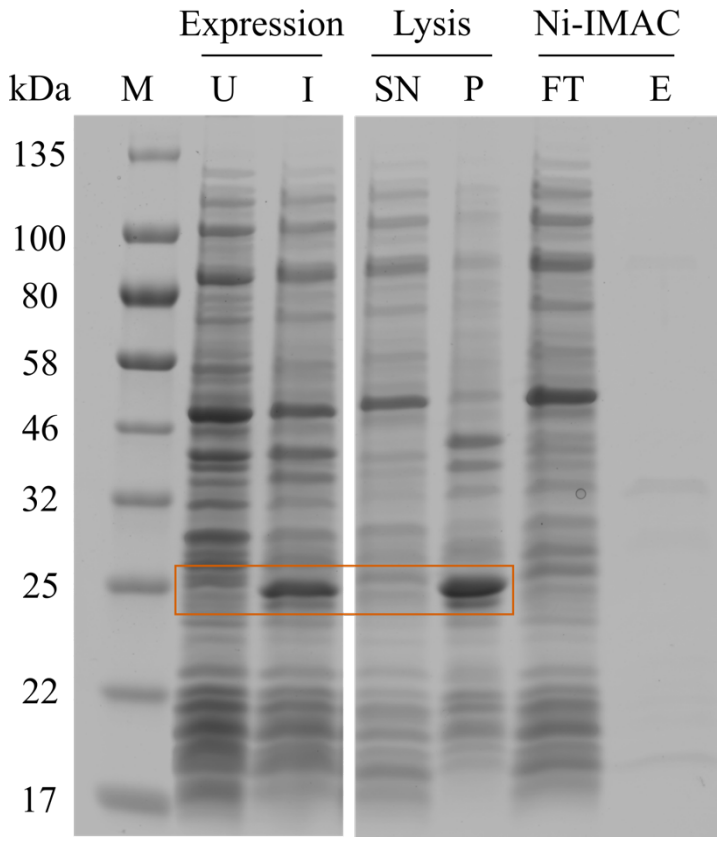


Figure 48 SDS-PAGE of Ni-IMAC purification of pET28a-G90(C97S)Q259(GCN4) – Marker lane (M) with molecular weight standards. Left SDS-PAGE showing expression test with uninduced (U) and three-hour post induction (I) BL21 Star cells transformed with pET28a-G90(C97S)Q259(GCN4) normalised by cell density before loading. Right SDS-PAGE was loaded with supernatant after lysis (SN) and pellet after lysis (P). Fractions analysed in the Ni-IMAC step: Flowthrough (FT) and elution (E). Expected monomer size is about 21 kDa, apparent monomer band with a shift of + 4 kDa is highlighted (orange box).

4.7.2 Purification of pOPINFW-V249S433

pOPINFW-V249S433 was purified from 341 mg of *E. coli* BL21 Star cells grown in 450 mL of LB medium. 2 mL of Ni-NTA resin was used for this purification. The expected monomer size is about 19 kDa. No supernatant:pellet ratio was calculated as almost all of the POI is in the pellet fraction (**Figure 49**). No yield was determined as the purity of the IMAC elution fraction was below the threshold of 0.2. This also made the concentration of the wash and elution fractions for SDS-PAGE necessary as otherwise the bands would have been below the detection limit of Coomassie Blue.

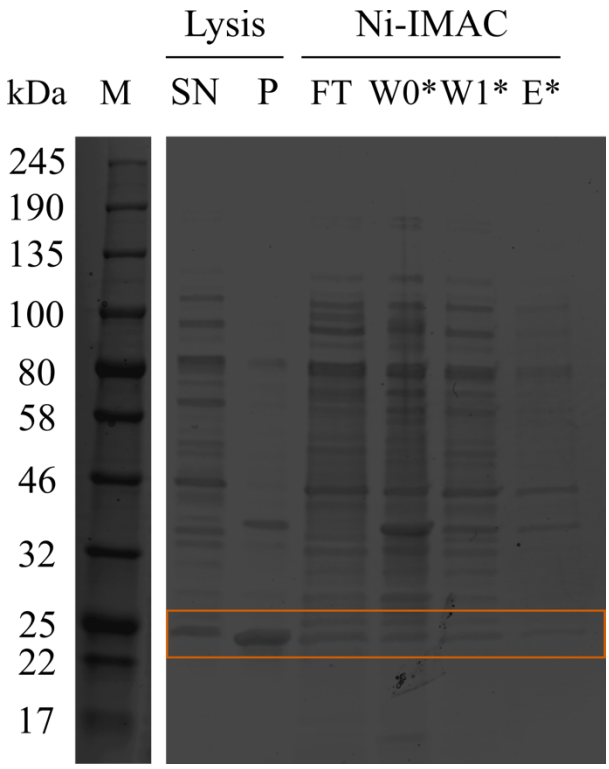


Figure 49 SDS-PAGE of Ni-IMAC purification of pOPINFW-V249S433 – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution (E). Expected monomer size is about 19 kDa with an apparent monomer band a shift of + 5 kDa than expected size (orange box). Note that wash and elution fractions have been concentrated using the acetone precipitation method.

4.8 Native versus denaturing purification of a stalk domain construct

A method that was previously used for the successful purification of TAA domains rich in coiled coils consists of the addition of a strong denaturation agent like 6 M GuHCl to the purification steps to completely denature the protein and then refold it after IMAC via rapid dialysis (Hernandez Alvarez, Hartmann et al. 2008). Exploring this option for the stalk domain constructs of BpaC was tested via a comparative native versus denaturing purification on pET28a-(GCN4)A261Q392(GCN4).

4.8.1 Native IMAC purification of pET28-(GCN4)A261Q392(GCN4)

pET28a-(GCN4)A261Q392(GCN4) was purified from 4 g of *E. coli* BL21 Star cells grown in 1 L of TB medium. 2 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 20 kDa. The supernatant:pellet ratio is 0.11 (**Figure 50**). The purity of the IMAC elution fraction is 0.39. The yield of the POI in the elution fraction was 0.34 mg/L culture.

The low supernatant:pellet ratio shows the high aggregation tendency of the stalk domain of BpaC. The cytosolical expression and native IMAC purification seems to attract non-specific proteins binding to the construct. Denaturation of the protein would cancel out the aggregation potential and allowing the POI to refold in an isolated environment unable to attract further non-specific contaminants.

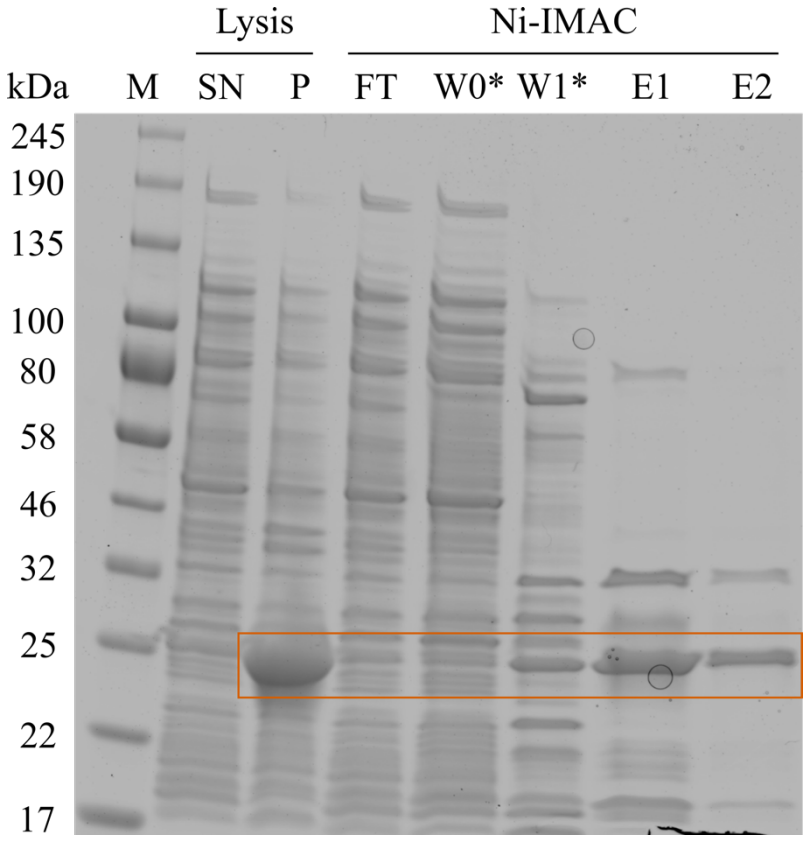


Figure 50 SDS-PAGE of native Ni-IMAC purification of pET28a-(GCN4)A261Q392(GCN4) – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution fractions 1-2 (E1-E2). Apparent monomer band of the POI (orange box) with a shift of + 4 kDa compared to the expected size of the monomer of about 20 kDa is highlighted. Wash fractions were concentrated using acetone precipitation (*).

4.8.2 Denaturing purification and refolding of pET28a-(GCN4)A261Q392(GCN4)

The pellet fraction of the native purification of pET28a-(GCN4)A261Q392(GCN4) was used to extract the POI using a denaturation buffer containing 6 M GuHCl. 2 mL of Ni-NTA resin was used for this purification and all buffers for IMAC contained 6 M GuHCl as well. The expected monomer size is about 20 kDa. No supernatant:pellet ratio is given as GuHCl heavily interferes with the SDS-PAGE. As a consequence, only the dialysed IMAC elution fraction (devoid of GuHCl) was analysed via SDS-PAGE (**Figure 51**). The purity of the elution fraction was 0.93 and after SEC over 0.99. The final yield after SEC was 0.59 mg/L culture. A clear trimer/monomer distribution was observed in SDS-PAGE for the SEC elution fractions. Most of the POI was observed around the peak of about 70 mL elution volume. The absorbance values at 280 nm for the SEC trace are very low due to the low extinction coefficient of 4470 $M^{-1}cm^{-1}$ of the POI and due to the use of an incorrect column size (HiLoad 16/60 Superdex 200 pg is too big for only 1 mg of protein) which makes it hard to see a clear protein peak.

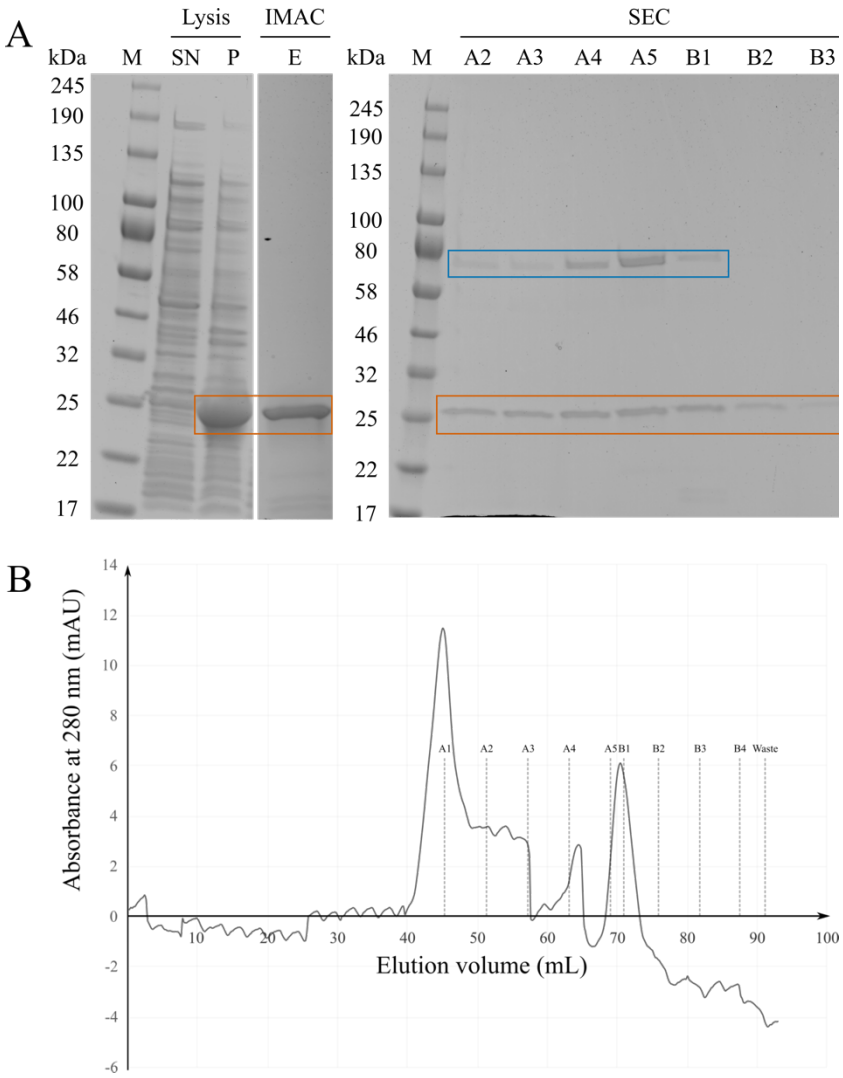


Figure 51 SDS-PAGE of denaturing Ni-IMAC purification of pET28a-(GCN4)A261Q392(GCN4) – **A** Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC elution fraction after dialysis (E). Apparent monomer band of the POI (orange box) with a shift of +4 kDa compared to the expected size of the monomer of about 20 kDa is highlighted. No flowthrough and wash fractions were collected as GuHCl heavily reduces SDS-PAGE performance. SEC elution fractions (A2-B3) were analysed as well. **B** SEC chromatogram of the SEC run using a HiLoad 16/60 Superdex 200 pg column is shown with the elution fractions marked at the relevant elution volume corresponding to the fractions analysed in **A**. Apparent trimer band at about 72 kDa is highlighted (blue box) which includes a + 12 kDa shift compared to the expected trimer size of about 60 kDa.

4.9 Improving constructs by fusion with C-terminal head domain

The next strategy to overcome the low solubility of the N-terminal head domain and the stalk domain constructs was to merge these domains using the DAVNxxQL neck motif at the end of the N-terminal head domain, stalk I domain and stalk II domain so there is no impact on the overall structure by introducing discrete deletions.

4.9.1 Purification of pET28a-S99(ΔCC12)T601

pET28a-S99(ΔCC12)T601 was purified from 1.69 g of *E. coli* BL21 Star cells grown in 850 mL of LB medium. 5 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 28 kDa. The supernatant:pellet ratio is 1.40 and the purity of the IMAC elution is over 0.99 (**Figure 52**). The final yield is 1.49 mg/L culture. The protein solution turned into a gelatinous solid state at about 56 mg/mL concentration in the final concentration step as preparation for crystallisation screens and had to be diluted to less than 30 mg/mL to become liquid again. This is likely due to the addition of charged residues in this construct that led to an apparent increase in solubility while still retaining the aggregation potential of the N-terminal head domain.

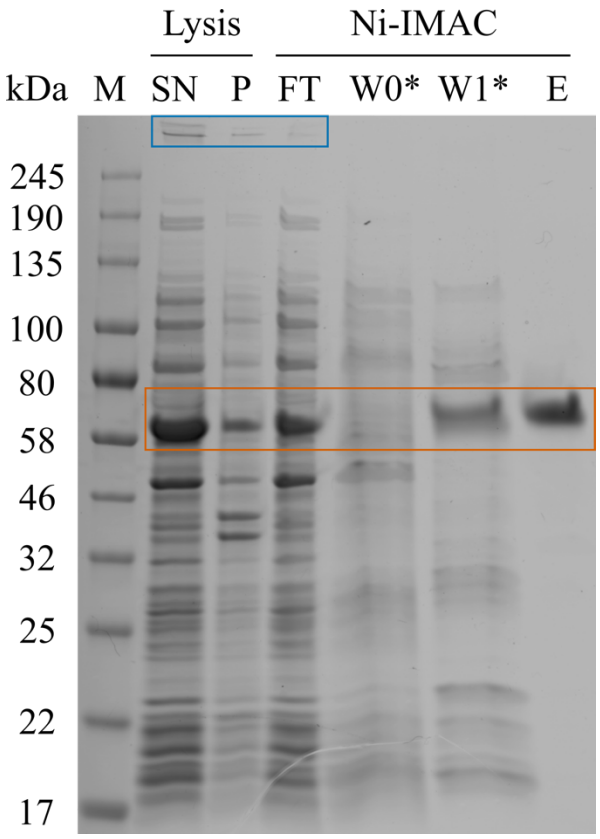


Figure 52 SDS-PAGE of Ni-IMAC purification of pET28a-S99(ΔCC12)T601 – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution (E). The expected monomer size is about 28 kDa. No monomer band was detected but an undefined oligomeric state (likely trimer) is highlighted at about 60 kDa (orange box). Additional bands at the top of the gel indicating the tendency to form higher oligomeric complexes for this construct (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.9.2 Purification of pET28a-L174(ΔCC12)T601

pET28a-L174(ΔCC12)T601 was purified from 816 mg of *E. coli* BL21 Star cells grown in 425 mL of LB medium. 1 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 28 kDa. The supernatant:pellet ratio is 0.90 (**Figure 53**). The purity of the IMAC elution fraction is 0.65. Final yield was not determined as this was an initial test purification to check the viability of this construct. The IMAC flowthrough fraction started to precipitate overnight at 7 °C, which did not occur in the flowthrough fraction of pET28a-L260T601, which was purified in parallel with this construct.

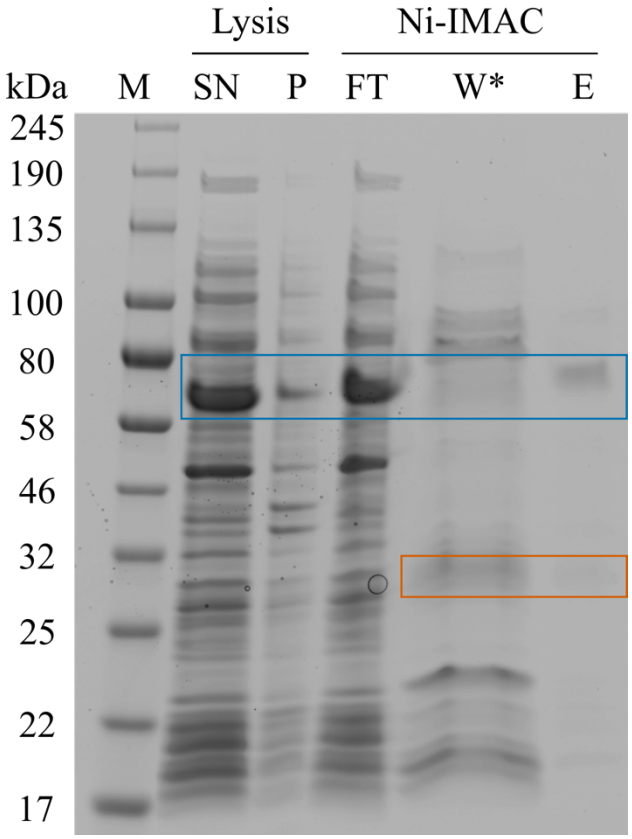


Figure 53 SDS-PAGE of Ni-IMAC purification of pET28a-L174(ΔCC12)T601 – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), wash (W), elution (E). The expected monomer size is about 28 kDa. The apparent monomer band is visible around the expected size (orange box). An undefined oligomeric state (likely trimer) is highlighted (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.9.3 Purification of pET28a-L260T601

pET28a-L260T601 was purified from 768 mg of *E. coli* BL21 Star cells grown in 425 mL of LB medium. 1 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 33 kDa. The supernatant:pellet ratio is 1.46 (**Figure 54**). The purity of the IMAC elution fraction is 0.82. The final yield was not determined as this was an initial test purification to check the viability of this construct. The characteristic double bands at around 90 kDa, previously identified as chaperones, are also present in the SDS-PAGE of this purification.

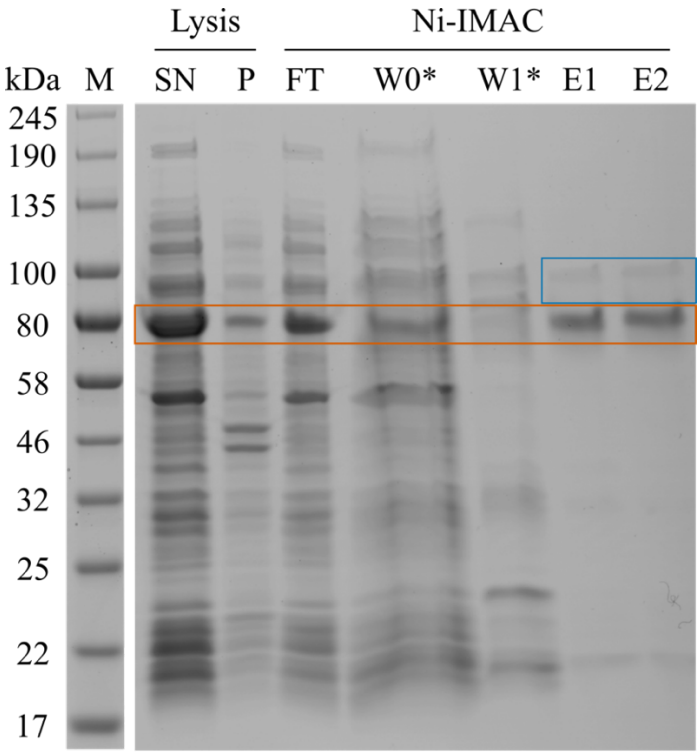


Figure 54 SDS-PAGE of Ni-IMAC purification of pET28a-L260T601 – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution fractions 1-2 (E1-E2). The expected monomer size is about 33 kDa. The apparent monomer band is barely visible and indistinguishable from contaminants. An undefined oligomeric state (likely trimer) is highlighted (orange box). Characteristic double band around 90 kDa for previously identified chaperones is also present in the elution fractions (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.10 Buffer optimisation of N-terminal fusion construct

There are several factors that could contribute to the gel formation during the previous concentration attempt of pET28a-S99(ΔCC12)T601. One very prominent attribute of most BpaC passenger domains is the low pI of the domains. pET28a-S99(ΔCC12)T601 has a pI of about 4.1, which makes it highly negatively charged in a buffer of pH 8. In this purification attempt the pH of the final buffer after IMAC (before concentrating for crystallisation screens) was lowered to a pH of 5.5 to see if this is the main cause of (soluble) aggregation.

4.10.1 Optimised purification of pET28a-S99(ΔCC12)T601

pET28a-S99-delCC12-T601 was purified from 1.80 g of *E. coli* BL21 Star cells grown in 850 mL of LB medium. 5 mL of Ni-NTA resin was used in this purification. The expected monomer size is about 28 kDa. The supernatant:pellet ratio is 1.07 (**Figure 55**), slightly lower than for the last purification which was 1.24. The purity of the IMAC elution fraction is 0.91, similar to the purity of the previous purification of this construct of 0.96. The final yield was 6.13 mg/L, the second highest yield of all tested constructs. The final concentration for crystallisation screens was about 87 mg/mL and without turning into a gel. This is a major improvement over the previous purification which had the protein solution turn into a gel at about 56 mg/mL.

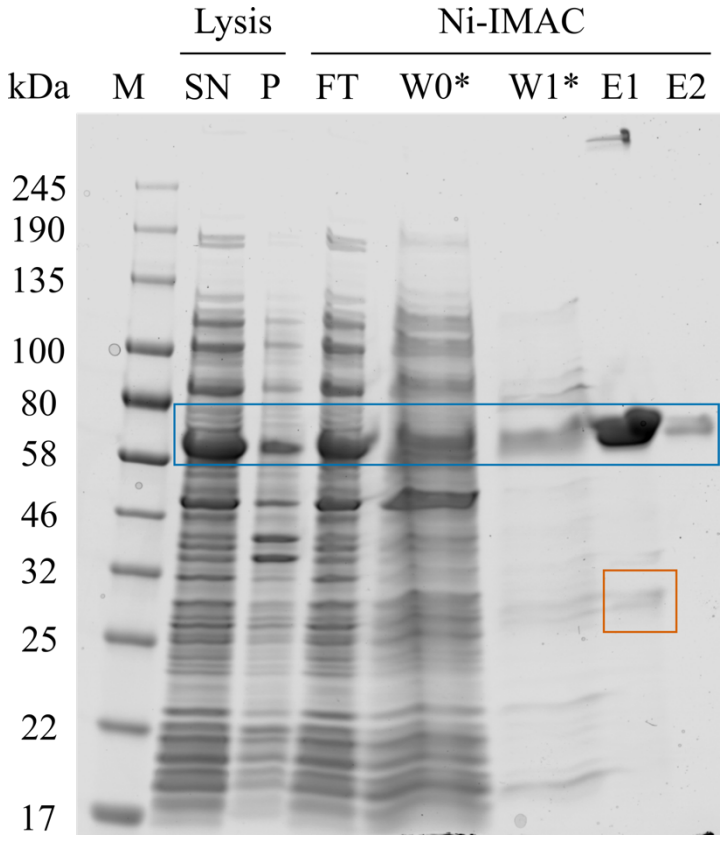


Figure 55 SDS-PAGE of Ni-IMAC of pET28a-S99(ΔCC12)T601 – Marker lane (M) with molecular weight standards. Samples taken before the IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution fractions 1-2 (E1-E2). The expected monomer size is about 28 kDa which matches with the apparent monomer band (orange box). An undefined oligomeric state (likely trimer) is highlighted at about 60 kDa (blue box). Wash fractions were concentrated using acetone precipitation (*).

4.11 Improving purification purity using a temperature gradient

For this final N-terminal head domain construct I was curious to see how using the residues that were involved in the crystallisation interface of pET28a-S741Q1054(GCN4) would change the outcome of the purification and crystallisation if combined with a different domain. However, the purity of the IMAC elution fraction was not as high as expected and I wanted to explore an alternative way of improving the purity for this construct by incubating aliquots of the IMAC elution fraction for 30 min in a step-wise temperature gradient using a PCR machine.

4.11.1 Purification of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4)

pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4) was purified from 6.4 g of *E. coli* BL21 Star cells grown in 1 L of TB medium. 1 mL of Ni-NTA resin was used for this purification. The sample lysate was very viscous, which made the initial IMAC purifications steps exceptionally time-consuming (more than 2 h were needed for applying the sample onto the resin and collecting the flowthrough). The expected monomer size of the POI is about 40 kDa. The supernatant:pellet ratio was 0.55 (**Figure 56**). The purity of the IMAC elution fraction was 0.37 and improved to about 0.78 for the highlighted fraction in the temperature gradient SDS-PAGE analysis (**Figure 57**). The final yield was 0.63 mg/L culture. This is the first time bands for each possible oligomeric state of a BpaC construct were observed in SDS-PAGE. No clear dimer band has been observed before, and it is unclear if this is a method artefact or a real dimer band. This was considered when determining the purity of the POI for the temperature gradient experiment with the dimer band being considered as contaminant. The best trade-off between purity improvement and appearance of higher oligomeric bands (aggregates) was between 50 and 55 °C. A consecutive analytical SEC run would give the necessary information on the quality of this sample and its usability for crystallisation trials.

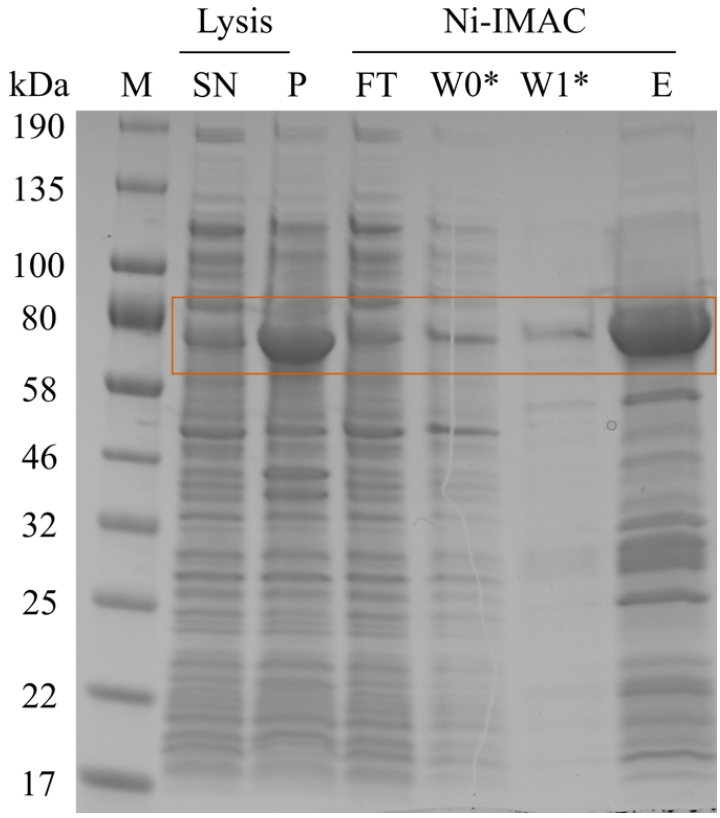


Figure 56 SDS-PAGE of Ni-IMAC purification of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4) – Marker lane (M) with molecular weight standards. Samples taken before IMAC step: Supernatant after lysis (SN), pellet after lysis (P). IMAC step: Flowthrough (FT), washes 0-1 (W0-W1), elution fraction (E). Undefined oligomer band of the POI (orange box) with a shift of + 37 kDa compared to the expected size of the monomer of about 40 kDa is highlighted. Wash fractions were concentrated using acetone precipitation (*).

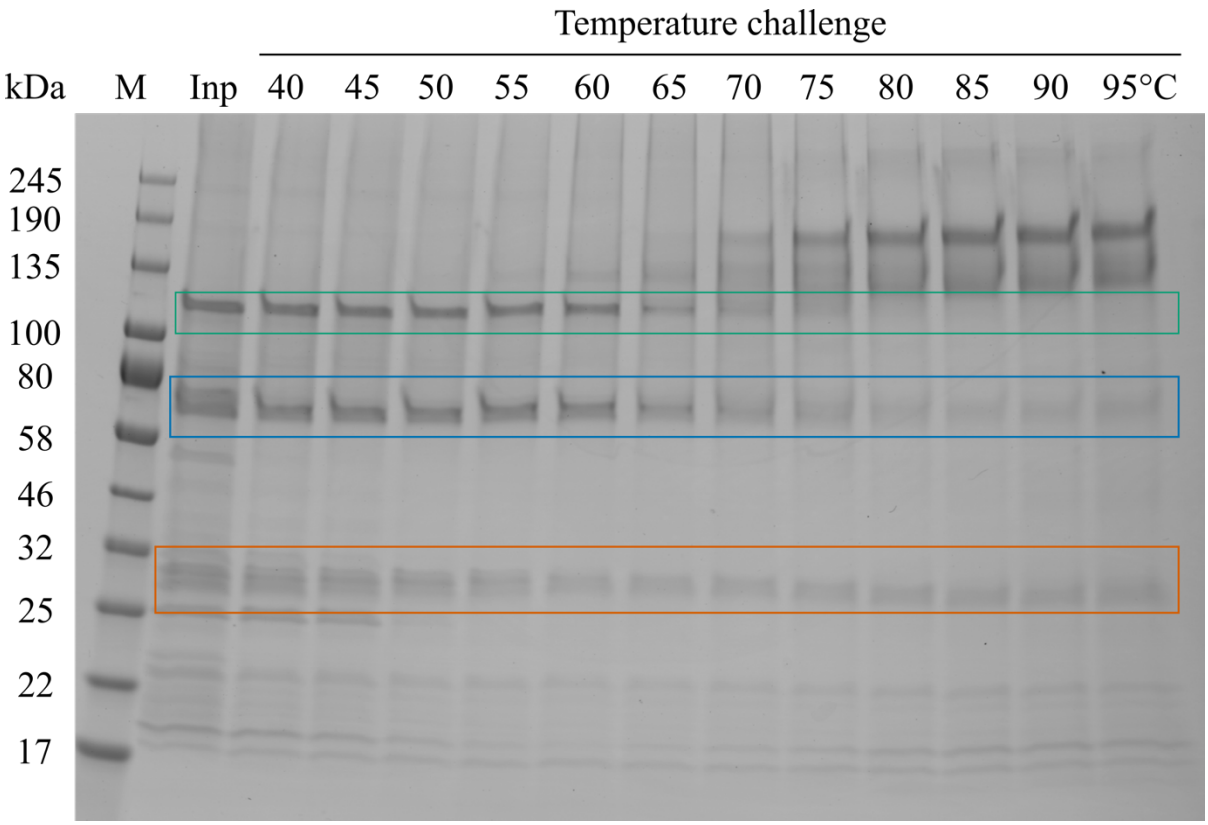


Figure 57 SDS-PAGE of 30 min temperature challenge of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4) – Marker lane (M) with molecular weight standards. IMAC elution fraction kept at 7 °C as control (Inp). Temperature challenge in 5 °C intervals from 40 to 95 °C (40, 45, …, 95 °C). Apparent monomer band with a shift of -10 kDa compared to the expected monomer size of about 40 kDa is highlighted (orange box). Possible apparent dimer band at about 75 kDa (blue box). Likely apparent trimer band at about 120 kDa (green box) which matches the expected trimer size. Higher oligomeric aggregates start appearing after 60 °C. Lower molecular weight contaminants below 25 kDa could not be removed using this method.

4.12 Overview of all purifications

An overview of all purifications is given in table form to show the progression in a construct series for each domain in terms of supernatant:pellet ratio, purity and yield observed (**Table 15**). This is graphically supported by a flowchart that also highlights the order of discoveries that fed into designing the next generation of constructs for each domain (**Figure 58**).

Table 15 Overview of purification statistics for N-terminal head domain constructs of BpaC.

Construct name	SN:P ratio	Purity	Yield (mg/L)	Comment
N-terminal head domain				
E-S99V208				All in pellet
F-F93(C97S)V208				All in pellet
28a-G90(C97S)Q173(GCN4)	0.79	0.71	0.68	Chaperones
28a-G90(C97S)Q259(GCN4)				All in pellet
28a-S99(ΔCC1)T517	0.61	0.19		Chaperones
28a-S99(ΔCC12)T601	1.40	0.99	1.49	Turned into a gel
28a-S99(ΔCC12)T601	1.07	0.91	6.13	Optimised
28a-V75(C76S)(C97S)(ΔCC12)(S447G826del)Q1054(GCN4)	0.55	0.37 (IMAC) 0.78 (Heat)	0.63	Viscous lysate

Constructs are sorted by order shown in **Figure 58**. Construct vectors are abbreviated as follows: pOPINE/F (E/F), pET28a (28a). Important parameters are supernatant:pellet ratio (SN:P ratio), purity, and yield given in mg of POI per L culture.

Table 16 Overview of purification statistics for stalk domain constructs of BpaC.

Construct name	SN:P ratio	Purity	Yield (mg/L)	Comment
Stalk domain				
FW-T166L271				All in pellet
FS-V249S433	<0.2	<0.2		Below purity threshold
E-T166S433				No expression
28a-(GCN4)A261Q392(GCN4)	0.11	0.39 (Native) 0.93 (Denaturation)	0.34 (N) 0.59 (D)	High refolding efficiency
28a-L174(ΔCC2)T601	0.90	0.65		Improve purity
28a-L260T601	1.46	0.82		Improve purity

Constructs are sorted by order shown in **Figure 58**. Construct vectors are abbreviated as follows: pOPINE/F (E/F), pET28a (28a). Important parameters are supernatant:pellet ratio (SN:P ratio), purity, and yield given in mg of POI per L culture.

Table 17 Overview of purification statistics for C-terminal head domain constructs of BpaC.

Construct name	SN:P ratio	Purity	Yield (mg/L)	Comment
C-terminal head domain				
FW-T914D1097				All in pellet
FW-T914Q1054(GCN4)	0.29	0.9	0.18	Low yield
28a-S741Q1054(GCN4)	0.72	0.99	11.79	Best purification

Constructs are sorted by order shown in **Figure 58**. Construct vectors are abbreviated as follows: pOPINE/F (E/F), pET28a (28a). Important parameters are supernatant:pellet ratio (SN:P ratio), purity, and yield given in mg of POI per L culture.

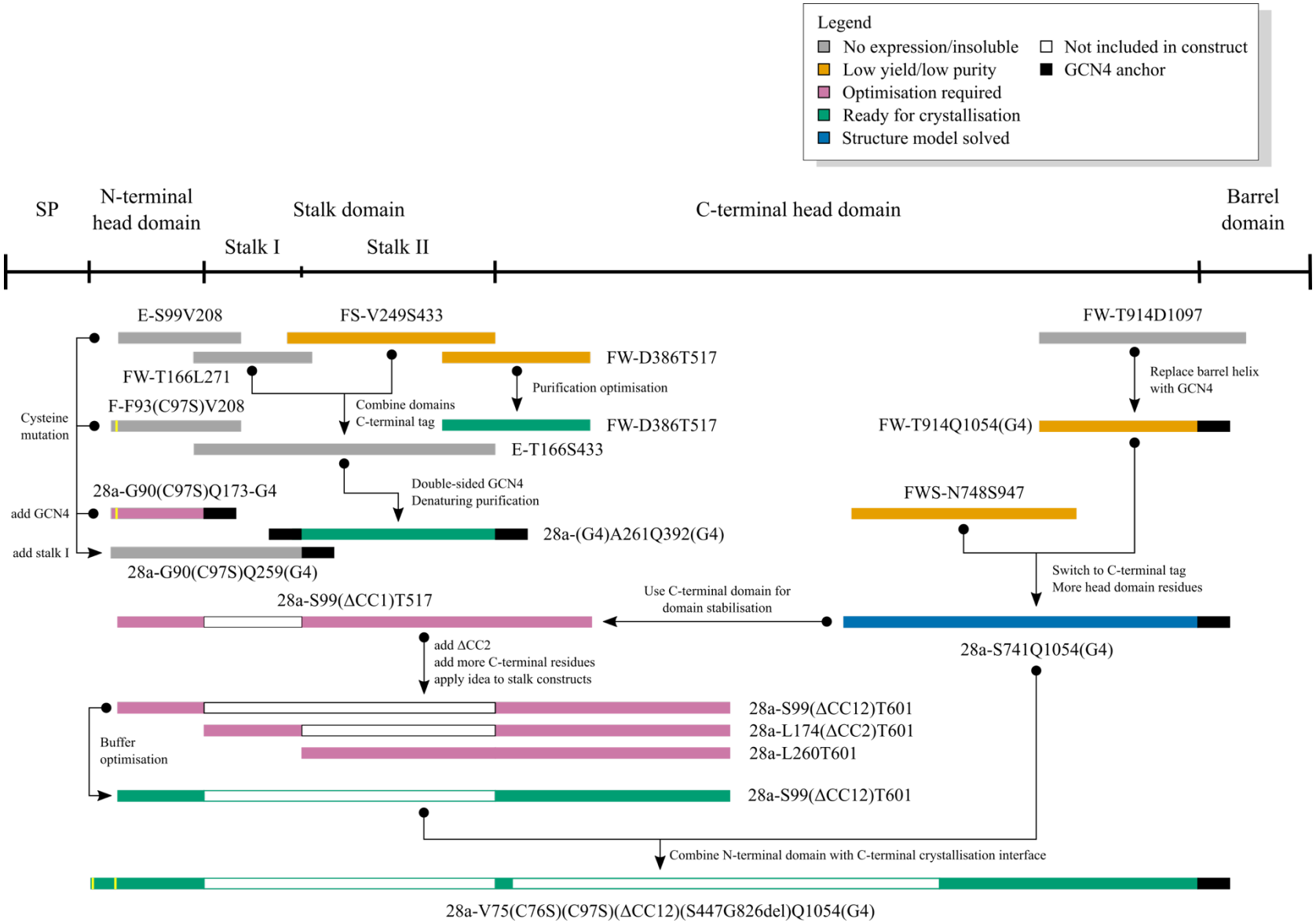

Legend
No expression/insoluble
Low yield/low purity
Optimisation required
Ready for crystallisation
Structure model solved
Not included in construct
GCN4 anchor
SP
N-terminal head domain
Stalk domain
Stalk I
Stalk II
C-terminal head domain
Barrel domain
E-S99V208
FS-V249S433
FW-D386T517
FW-T166L271
Purification optimisation
Combine domains
C-terminal tag
F-F93(C97S)V208
Cysteine mutation
FW-D386T517
E-T166S433
Double-sided GCN4
Denaturing purification
28a-G90(C97S)Q173-G4
add GCN4
28a-(G4)A261Q392(G4)
add stalk I
28a-G90(C97S)Q259(G4)
FW-T914D1097
Replace barrel helix with GCN4
FW-T914Q1054(G4)
FWS-N748S947
Switch to C-terminal tag
More head domain residues
Use C-terminal domain for domain stabilisation
28a-S99(ΔCC1)T517
28a-S741Q1054(G4)
add ΔCC2
add more C-terminal residues
apply idea to stalk constructs
28a-S99(ΔCC12)T601
28a-L174(ΔCC2)T601
28a-L260T601
Buffer optimisation
28a-S99(ΔCC12)T601
Combine N-terminal domain with C-terminal crystallisation interface
28a-V75(C76S)(C97S)(ΔCC12)(S447G826del)Q1054(G4)

Figure 58 Schematic overview of all BpaC constructs tested – BpaC was split into structural domains from N- to C-terminus (left to right): extended signal peptide (SP), N-terminal head domain, stalk domain with the subdomain stalk I and stalk II, C-terminal head domain, and the barrel domain which includes the coiled coil residing within the β-barrel anchoring the protein into the membrane. The ticks used to divide the domains are relative to the overall length of the protein. Constructs are represented by coloured bars which relate to the performance of that construct during expression and purification and described in the legend box. The length of the construct bars also scales with the number of residues within the construct. An empty box inside the construct bar represents residues left out in that construct compared to the actual sequence. The yellow line within some constructs represents a cysteine-to-serine mutation.

4.12 Overview of all crystallisation attempts

In the beginning of the project, almost every purification was followed by a crystallisation attempt using the Formulatrix© NT8 robot together with the JSCG Core Suites that were available in our crystallisation facility. The purity and quality of the underlying purification was not checked and led to lots of precipitation and poorly diffracting hits (**Table 18**). In the next stage of the project, the focus was shifted to improving the quality of the POI first and trying to reach higher protein concentrations with a target purity of over 95%. This led to the creation of pET28a-S741Q1054(GCN4), which produced protein crystals that diffracted to up to 1.4 Å at a very high protein concentration of about 130 mg/mL. The resulting structure covered about 50% of the protein sequence (by structural expansion due to identical repeats in between solved structure and remaining C-terminal head domain). Still, the constructs covering the N-terminal head domain and stalk domain did not produce crystals diffracting to better than 10 Å, which meant a change of design approach had to be undertaken for a significant benefit towards successful crystallisation results. I decided to use the apparent high solubility of the C-terminal head domain by fusing parts of the domain to residues covering the N-terminal head domain and parts of the stalk domain, both using the neck motifs as fusion points between even far away residues. Initially this produced great results and a viable construct was obtained for the N-terminal head domain, the stalk I domain, and the stalk II domain. However, taking a closer look at the SDS-PAGE of the purified constructs and taken into account the viscosity problems – especially for pET28a-S99(ΔCC12)T601 – an optimisation step was necessary to counteract the aggregation tendency of these constructs. Buffer optimisation led to an increase from a final concentration of about 28 mg/mL to 87 mg/mL for pET28a-S99(ΔCC12)T601 by a simple reduction of pH to negate the charge effect that the C-terminal head domain introduces with its low local pI. Unfortunately the latter experiments all did not produce significantly diffracting protein crystals (<10 Å) and the end of the project was reached at that time.

Table 18 Overview of crystallisation attempts.

Vector	Residues	JSCG Suite	Purification gel	Protein buffer	Protein concentrations	Precipitation?	Hits	Diffracted?
pOPINFW	D386T517	I+II	**Figure 37**, section 4.2.2	20 mM Tris-HCl pH 8	7.6 mg/mL	>75%	2	X
pOPINFW	D386T517	I-IV	**Figure 38**, section 4.3.1	20 mM Tris-HCl pH 8	3.7 mg/mL, 3.0 mg/mL, 2.4 mg/mL	>50%	5	X
pOPINFWS	N748S947	I+II	**Figure 40**, section 4.3.3	20 mM Tris-HCl pH 8, 150 mM NaCl	3.7 mg/mL, 1.9 mg/mL	>50%	10	>10 Å
pOPINFW	T914Q1054 (GCN4)	I-IV	**Figure 42**, section 4.4.2	20 mM Tris-HCl pH 8, 150 mM NaCl	2.1 mg/mL, 1.0 mg/mL	>50%	3	X
pOPINFW	D386T517	I+II+IV	-	20 mM Tris-HCl pH 8, 150 mM NaCl	20.9 mg/mL, 13.8 mg/mL, 2.6 mg/mL	>50%	3	>10 Å
pET28a	S741Q1054 (GCN4)	I-IV	**Figure 43**, section 4.5	50 mM Tris-HCl pH 8, 150 mM NaCl	130.1 mg/mL, 18.3 mg/mL, 9.0 mg/mL	>25%	>20	1.4 Å
pET28a	G90(C97S) Q173(GCN4)	I-IV	**Figure 45**, section 4.6.2	20 mM Tris-HCl pH 8, 150 mM NaCl	16.3 mg/mL, 8.2 mg/mL	>75%	4	>10 Å
pET28a	S99(ΔCC12)T601	I+II+IV	**Figure 52,** section 4.9.1	20 mM Tris-HCl pH 8, 150 mM NaCl	28.2 mg/mL, 14.1 mg/mL	>75%	2	>10 Å
pET28a	S99(ΔCC12)T601	I-IV	**Figure 55,** section 4.10.1	20 mM MES pH 5.5, 50 mM NaCl	86.8 mg/mL, 57.3 mg/mL, 28.6 mg/mL	>50%	11	>10 Å

Associated purification gels are linked with figure number and section location (except for the last attempt on pOPINFW-D386T517 where the elution profile looked very similar to the one in **Figure 38**). Overall percentage of conditions with precipitation in the drop was estimated and most promising hits were counted and passed further to X-Ray diffraction experiments.

5 Chapter 5: Structural determination and analysis of the C-terminal head domain construct S741Q1054(GCN4)

The elucidation of a protein structure can provide new biochemical information that broadens our understanding of the potential function of the POI. In the context of TAAs, structural determination can be used to verify the initial hypothesis of a structural prediction for a certain domain. The high degree of structural conservation between TAA motifs then allows best-guess associations between known binding partners in other TAAs and the domain in question. While the backbone atom positions of TAA motifs can be highly superimposable with a low root-mean-square deviation (r.m.s.d.) differences in residue composition especially facing towards the solvent can indicate a change in binding partner preference.

In this particular case, parts of the C-terminal head domain of BpaC (residues S741 to Q1054) produced hexagonal crystals that diffracted to 1.4 Å at the Diamond Light source beamline I04. I used the C-terminal head domain model of BoaA (PDB: 3S6L) for molecular replacement. A starting model was created using BUCCANEER, an automated model building tool within the CCP4 suite. Several iterations of manual model building in Coot and refinement in REFMAC5 were performed until no significant improvement of R-factors were observed. However, these R-factors were too high for the given resolution raising some questions about the validity of the model or the underlying data. The latter is more likely as the associated difference electron density map (F_o-F_c) contained unexplained negative density peaks that looked like data artefacts rather than model building issues. Manually reprocessing in XDS indeed changed the interpretation of

the data set. Following up on this with some more elaborate refinement attempts in PHENIX (including addition of alternative confirmations, split occupancy assignment for solvent molecules and symmetry related processing challenges), a much-improved model could be obtained with acceptable R-factors, an improved difference density map, and refinement statistics within reasonable parameter ranges.

Due to the identical structural repeats in the N-terminal end of the model one can use this information to obtain a model for the full C-terminal head domain of BpaC by some simple copy and alignment operations in PyMOL. This means that after this project, structural information for almost 54% of all the residues in BpaC is available for further research.

5.1 Crystallisation trials with purified S741Q1054(GCN4)

Purified S741Q1054(GCN4) was used at a final concentration of about 130 mg/mL as estimated by UV280 absorption. At this concentration, the accuracy of protein concentration can vary drastically and is therefore only an approximate value. This is reflected in the dilution series that was setup for crystallisation which was aimed at 0.67x and 0.33x of the original concentration but the absolute values were 18 mg/mL, and 9 mg/mL. Four crystallisation screens (JCSG Core Suite I-IV) were used together with a 3 drop plate configuration (drop 1: 130 mg/mL, drop 2: 18 mg/mL, drop 3: 9 mg/mL). Crystals appeared in various conditions 12 h after setup in drop 1. Selected conditions were cryoprotected by adding 200 nL of protectant solution consisting of the buffer components of the well solution adjusted to 1.5x of the original concentration plus 35% Glycerol. This was added to the original drop volume of 100 nL to ensure that the final ionic strength of the now 300 nL drop stayed the same with a Glycerol concentration of about 23%. Crystals were then harvested and plunge frozen in liquid nitrogen and sent to the Diamond Light Source for X-Ray diffraction experiments. A time course of the crystal growth which provided the diffraction data for the structure model of S741Q1054(GCN4) is displayed (**Figure 59**). Crystals had to be harvested after two days as the original hexagonal crystal form started to degrade after 36 h. Buffer components at harvest were: 0.1 M HEPES pH 6.5, 0.8 M $(NH_4)_2SO_4$, and 23.34% Glycerol.

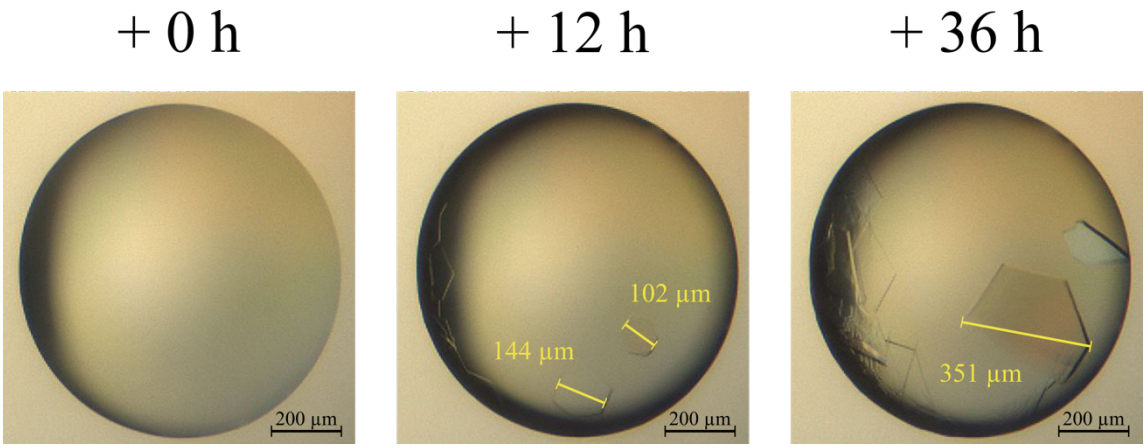


Figure 59 Time course of crystal growth of successful diffraction condition of S741Q1054(GCN4) – Purified S741Q1054(GCN4) at a concentration of 130 mg/mL, 18 mg/mL, and 9 mg/mL was mixed 1:1 with the precipitant solution and prepared in a vapour diffusion plate setup using the NT8 crystallisation robot in a 96-well 3 drop format. Selected condition shown here is in drop 1 (130 mg/mL; JCSG core suite IV, E10; 0.1 M HEPES pH 6.5, 0.8 M $(NH_4)_2SO_4$) over a time course of 36 h. Size estimation of mostly hexagonal crystals is shown with yellow bars and numbers.

5.2 Model building of S741Q1054(GCN4)

Protein models obtained by X-Ray crystallography methods only represent a snapshot of the current position of all the atoms within the protein crystal (organic and non-organic alike) and the accuracy of this model is limited by the quality of the underlying data and the ability of the structural biologist to correctly interpret the resulting electron density map in a biochemical meaningful way. The structure solution for S741Q1054(GCN4) at first looks straightforward as the ability to correctly place residues and their sidechains into density peaks is vastly improved by collecting high resolution diffraction data (below 2 Å). However, the occurrence of an unusually large cell axis (*c* greater than 500 Å) led to spots on the diffraction screen that were very close to each other, which made automated spot picking difficult and error-prone. Initially, this led to a misinterpretation of the cell axis length and consequently affected the accuracy of the model when comparing to the underlying data. A manual indexing step with adjusted parameters in XDS was necessary to more accurately interpret the cell parameters, which ultimately led to an improved model represented by lower R values for both R_{work} and R_{free} alike.

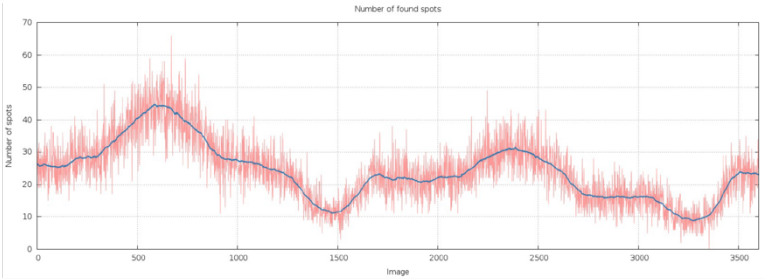
Special attention was also paid to alternate conformations and solvent molecules with and without split occupancies. This is additional information resulting from the high resolution of the diffraction data and the low B-factors (less than 10 $Å^2$) in some areas of the protein model. The challenge of having the crystallographic axis coincide with the central trimer axis (biological) created some ambiguity of the exact location and number of solvent molecules and alternate confirmations within that area. The final model has an R-factor of 18.39%/21.71% (R_{work}/R_{free}) with further refinement statistics shown in section 5.2.6.

5.2.1 Data collection parameters and initial indexing attempts

3600 images were collected from X-Ray diffraction experiments of S741Q1054(GCN4) crystals shown in **Figure 59**. These images were subjected to automated data processing in autoPROC, which is part of the image processing pipeline at the Diamond Light source beamline I04-1. An overview of the autoPROC process for S741Q1054(GCN4) highlights the challenges that an usually long cell axis can bring to the spot finding algorithm within autoPROC (**Figure 60**). Only 4% of the initial spots identified by XDS were used for finding a space group and unit cell solution owing to the low number of spots identified per image. Normally at least 50% of the spots are required to find a reliable solution. Nevertheless, autoPROC produced an output .mtz file with good data statistics (**Table 19**) which was taken forward to CCP4i2 for molecular replacement and model building.

A

Finding spots in images 1-3600



B

Indexing attempt using 4% of the intitial 86266 spots

WARNING

The selected indexing solution uses less than 50% of initial spots - please check this carefully for ice-rings, multiple lattices or problematic spot searching (due to poor initial background estimation).

LATTICE-CHARACTER	BRAVAIS-LATTICE	QUALITY OF FIT	UNIT CELL CONSTANTS (ANGSTROEM & DEGREES) a	b	c	alpha	beta	gamma
44	aP	0.0	57.4	57.5	177.8	90.2	99.2	119.9
17	mC	0.7	99.4	57.6	177.9	90.0	100.9	90.0
31	aP	2.0	57.4	57.6	177.8	80.7	80.8	59.9
27	mC	2.7	99.4	57.6	177.9	90.0	100.9	90.0
39	mC	6.0	99.6	57.5	177.8	90.2	100.7	89.8
10	mC	8.9	99.6	57.5	177.8	89.8	100.7	90.2
9	hR	9.5	57.6	57.4	524.1	90.1	90.0	120.1
30	mC	123.2	57.6	350.9	57.4	85.5	120.1	90.0
29	mC	123.5	57.6	99.4	177.8	84.8	99.3	90.0

BRAVAIS-TYPE	POSSIBLE SPACE-GROUPS FOR PROTEIN CRYSTALS [SPACE GROUP NUMBER,SYMBOL]	
aP	[1,P1]	
mC,mI	[5,C2]	
hR	[146,R3]	[155,R32]

C

Analysis of data with POINTLESS and indexing refinement

```
Cell:   57.53   57.53  525.99   90.00   90.00  120.00
Wavelength:  0.97950 A
Run number:    1 consists of batches          1 to       3600
   Resolution range for run:    87.66    1.42
   Phi range:                 0.00 to             360.00
```

Figure 60 Overview of X-Ray diffraction image processing with autoPROC for S741Q1054(GCN4) – 3600 diffraction images were collected with 0.1° oscillation per image during data collection up to a total of 360°. These images were then subjected to the autoPROC pipeline for automated data processing integrating all images into a single .mtz datafile that can be used for molecular replacement and model building. **A** Graph showing the distribution of number of spots identified by XDS in each diffraction image (image numbers 1-3600). **B** Indexing was performed on 4% of the initial spots identified in the previous step which triggered a warning from autoPROC as it uses less than the normally desired 50% or more of spots used for indexing. Identification of most likely space group and cell parameters was performed with the chosen solution highlighted (orange box). **C** The data were then analysed by POINTLESS and the initial cell parameters refined to the final values produced by autoPROC (blue box).

Table 19 X-Ray diffraction processing results produced by autoPROC.

Parameter	Value
Space group	R32
Unit cell	
a, *b*, *c* (Å)	57.53, 57.53, 525.99
α, β, γ (°)	90, 90, 120
Resolution (Å)	87.66-1.42 (1.45-1.42)*
Reflections (observed/unique)	884392/55885
Mean(I/s(I))	7.0 (1.1)
Completeness (%)	100.0 (100.0)
$CC_{1/2}$	99.7 (43.9)
Multiplicity	15.5

Diffraction images of the S741Q1054(GCN4) data collection were automatically indexed, scaled, and integrated in autoPROC. The most important values from this processing job are shown here.
*Values in parenthesis represent the highest-resolution bin.

5.2.2 Model building using 3S6L as molecular replacement template

The most common method of solving a structure using X-Ray crystallography is molecular replacement. This method, however, requires a suitable template model that is close enough to the search model in order to solve the so-called phase problem of X-Ray crystallography. An initial BLAST search of the PDB revealed a good candidate for molecular replacement: a model covering parts of the C-terminal head domain of the *B. pseudomallei* TAA BoaA, which is closely related to BpaC, with PDB accession code 3S6L (Edwards, Gardberg et al. 2011). For BpaC, the estimation of the contents of the asymmetric unit in CCP4i2 revealed a solvent content of about 53% with a 32.1 kDa protein molecule together with a Matthews coefficient of 2.61. The probability of this coefficient occurring for a single molecule in the asymmetric unit is 99.6%. Next, the input model was reduced to only include chain A of the PDB file of 3S6L using PyMOL (**Figure 61**, **A**). A sequence alignment was carried out between 3S6L and S741Q1054(GCN4) using CLUSTALW for CHAINSAW to be able to trim the mutated residues to a common Cγ atom

of the sidechain if available (**Figure 61**, **B**). The trimmed input model was used as search model for PHASER along with the reflections provided by the .mtz output file from autoPROC. The solution found by PHASER is associated with a translation function Z score of 37.21 (PHASER considers a valid solution to have a translation function Z score of above 8 and a refined log-likelihood gain value of 3788 (**Figure 61**, **C**). An electron density map was generated in the form of a $2F_o$-F_c map, which was used by BUCCANEER together with the trimmed input model from CHAINSAW for automated model building (**Figure 61**, **D**). A total of 339 residues were built this way with associated R_{work}/R_{free} factors of about 32%/34%. Additional residues that were not part of the largest chain were removed as deemed wrong before starting several cycles of manual model building. Residues were fit into the density using real space refinement within Coot and usually 5-10 residues were built at a time before a new refinement cycle was initiated in REFMAC5. A convenient output of this refinement cycle is the difference map that is generated by REFMAC5 (F_o-F_c). This is a great tool to help with model building as it shows over-or underbuilt areas in the model, allowing for a visual feedback in the model building cycle. Several rounds of this process were carried out until a total of 314 residues out of 351 were built. During this process, solvents were added automatically by REFMAC5 using the Babinet bulk solvent scaling option with an explicit solvent mask. Clearly wrong solvent molecules were removed in Coot and the remaining molecules further refined. In total, 236 solvent molecules were added this way and the final R-factors for this intermediate model was 22.5%/25% (R_{work}/R_{free}).

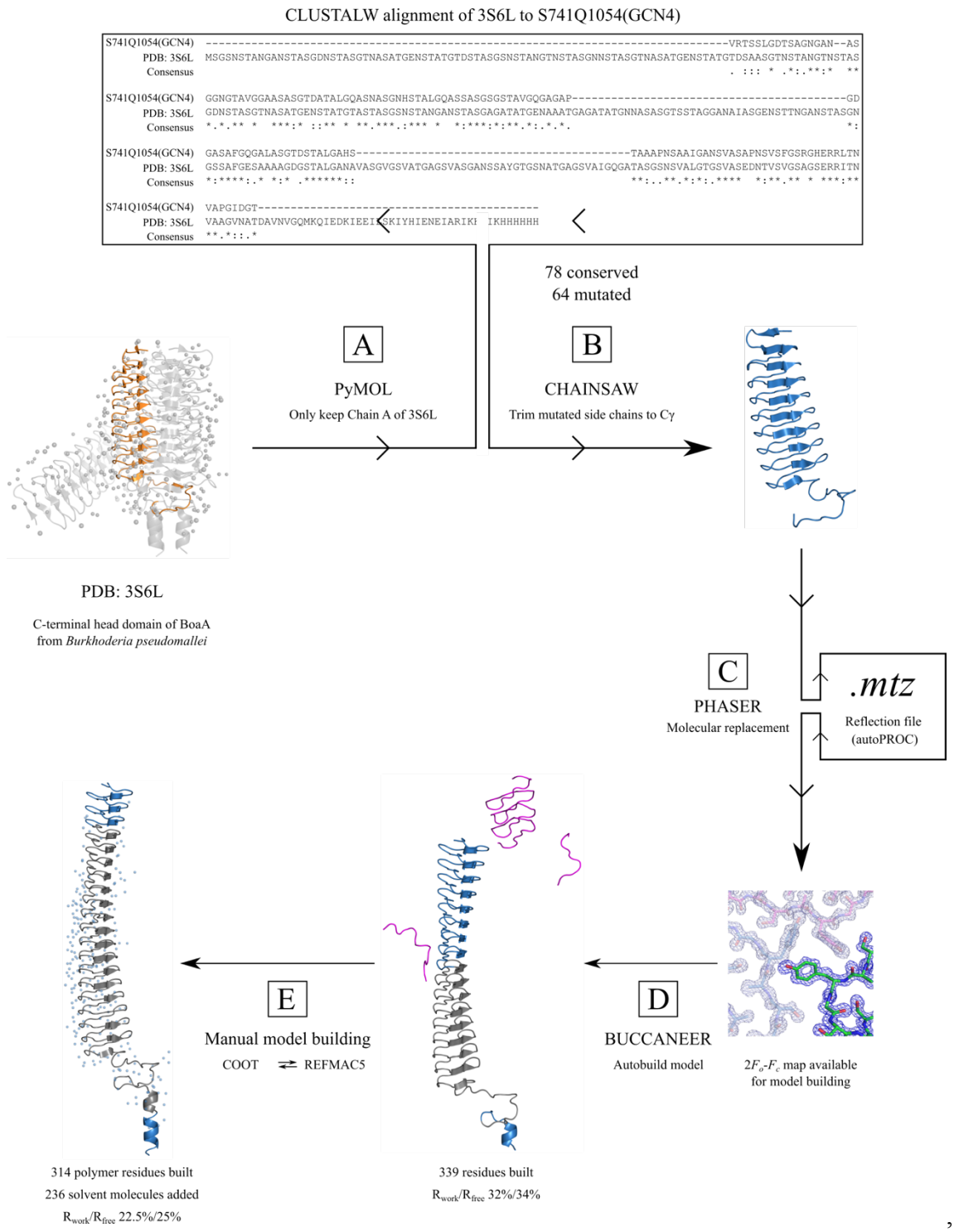

CLUSTALW alignment of 3S6L to S741Q1054(GCN4)
S741Q1054(GCN4) ----------VRTSSLGDTSAGNGAN--AS
PDB: 3S6L MSGSNSTANGANSTASGDNSTASGTNASATGENSTATGTDSTASGSNSTANGTNSTASGNNSTASGTNASATGENSTATGTDSAASGTNSTANGTNSTAS
Consensus . ::: * .*:.**:* **
S741Q1054(GCN4) GGNGTAVGGAASASGTDATALGQASNASGNHSTALGQASSASGSGSTAVGQGAGAP----------GD
PDB: 3S6L GDNSTASGTNASATGENSTATGTASTASGSNSTANGANSTASGAGATATGENAAATGAGATATGNNASASGTSSTAGGANAIASGENSTTNGANSTASGN
Consensus *.*.** * ***:* ::** * **.***.:*** * *:***:*:**.*:.*.*. *:
S741Q1054(GCN4) GASAFGQGALASGTDSTALGAHS----------TAAAPNSAAIGANSVASAPNSVSFGSRGHERRLTN
PDB: 3S6L GSSAFGESAAAAGDGSTALGANAVASGVGSVATGAGSVASGANSSAYGTGSNATGAGSVAIGQGATASGSNSVALGTGSVASEDNTVSVGSAGSERRITN
Consensus *:****:.* *:* .******:: **:..**.*:*:.**** *:**.** * ***:**
S741Q1054(GCN4) VAPGIDGT----------
PDB: 3S6L VAAGVNATDAVNVGQMKQIEDKIEEIESKIYHIENEIARIKK IKHHHHHH
Consensus **.*:::.*
78 conserved
64 mutated
A
PyMOL
Only keep Chain A of 3S6L
B
CHAINSAW
Trim mutated side chains to Cγ
PDB: 3S6L
C-terminal head domain of BoaA
from Burkhoderia pseudomallei
C
PHASER
Molecular replacement
.mtz
Reflection file
(autoPROC)
D
BUCCANEER
Autobuild model
2Fo-Fc map available
for model building
E
Manual model building
COOT ⇌ REFMAC5
339 residues built
Rwork/Rfree 32%/34%
314 polymer residues built
236 solvent molecules added
Rwork/Rfree 22.5%/25%

Figure 61 Workflow of S741Q1054(GCN4) initial model building attempt – Input model for molecular replacement was found with an evolutionary closely related TAA (BoaA) from *B. pseudomallei*. **A** Only chain A of the input model (PDB: 3S6L) was kept as input for molecular replacement by selective removal of all other chains and solvent atoms in PyMOL. **B** Chain A of 3S6L was then passed on to CHAINSAW which performs a sequence alignment with S741Q1054(GCN4) and retains all conserved sidechains while trimming the remaining sidechains to Cγ atoms if this is a shared part of the sidechains. **C** This model is then used in PHASER together with the reflection file (.mtz) generated by autoPROC to find the phase solution using molecular replacement. **D** The resulting electron density map ($2F_o$-F_c) was used by BUCCANEER, together with the CHAINSAW model, to attempt to build all the missing residues and sidechains of the model for S741Q1054(GCN4). **E** Repetitive rounds of manual model building in Coot followed by refinement cycles in REFMAC5 led to an intermediate model of S741Q1054(GCN4).

5.2.3 Alternative indexing with XDS changes cell parameters

The initial challenge when trying to solve the structure of this construct was the large R-factor discrepancy between calculated and expected R-factor range: as a rule of thumb - the higher the resolution, the lower the R-factors. There were also some regularly occurring positive and negative peaks in the difference density map (F_o-F_c) that hinted at a data artefact problem (**Figure 62**). This anomality in the difference density map was even seen in areas where there were very low B-factors associated with the main chain atoms (less than 10 $Å^2$). Normally peaks in the difference density map indicate model building errors but in this specific case a well-defined $2F_o$-F_c electron density map makes model building errors very apparent compared to the almost randomly distributed difference map peaks in that same area. Taken together this led to an artificial calculation barrier for the R-factors that could not be decreased further, even with several optimisation attempts of the underlying structure model. This strongly indicated that the underlying data was misinterpreted in the initial indexing step by autoPROC culminating in a wrong unit cell parameter estimation.

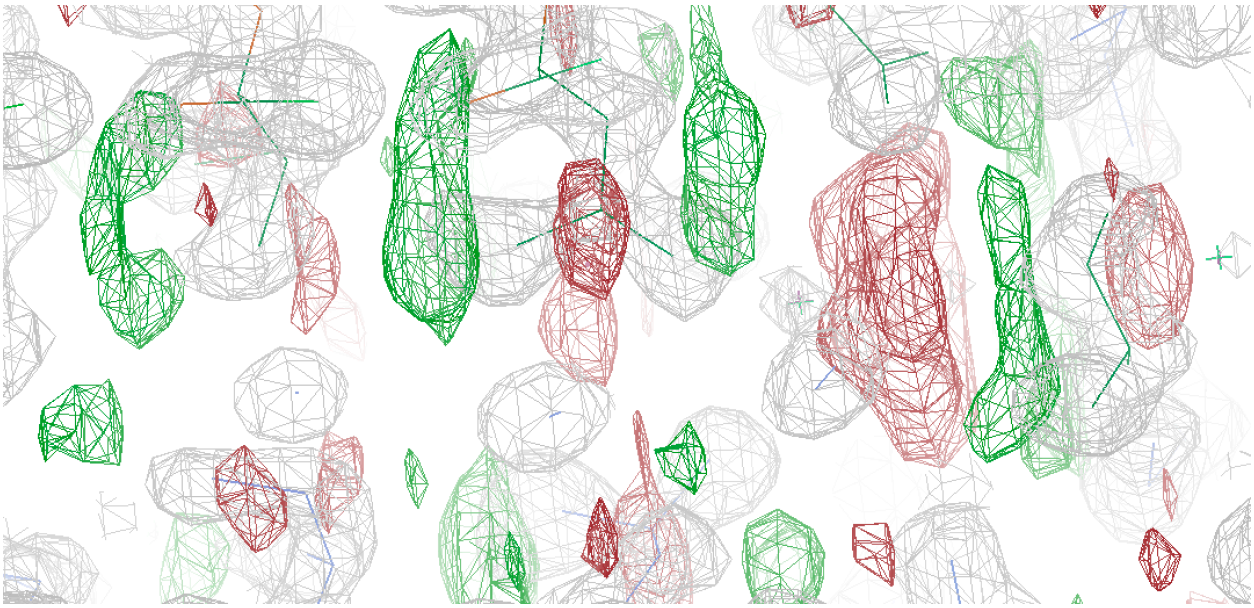


Figure 62 Difference map obscurities for initial model based on autoPROC data – Example area selected of the initial model of the head domain of BpaC, which was processed as shown in **Figure 61**. Difference map was contoured at 3 σ (F_o-F_c, green mesh representing positive difference density and red mesh as negative difference density). $2F_o$-F_c map was contoured at 1.5 σ (grey mesh) and residues shown in stick representation.

Manual reprocessing of the diffraction images was necessary to address the poor initial indexing attempts by autoPROC. Diffraction images 1-3600 were fed into XDS as input for indexing. The default parameters for spot finding were not sufficient to accurately discriminate spots due to the high *c* cell axis value of more than 500 Å, which has an inverse relationship to the distance of the spots on the diffraction image: they appear very close to each other, which makes them hard to be read as separate spots by the software. XDS provides several options to overcome this problem: first the "STRONG_PIXEL" parameter was increased to 6 (default is 3), which increased the threshold of the intensity of reflections required to be acknowledged as a spot in XDS; second the distance sensitivity of spots ("SEPMIN") and the spot cluster radius ("CLUSTER_RADIUS") were reduced significantly from 7 to 2 for "SEPMIN" and 3.5 to 1 for "CLUSTER_RADIUS". Although this improved the initial output, it was still difficult for XDS to find an accurate solution

for the unit cell parameter and space group problem. Looking at the images themselves one could identify a clear angular relation between individual indexed and non-indexed spot groups. This indicates a potential non-merohedral twinned crystal. Following the instructions for separating indexed and non-indexed spots into lattice groups (found on the XDS Wiki; strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Indexing) I was able to identify two to three separate lattices. Only the lattice covering the largest fraction of the data (lattice 1) was selected for further processing as the remaining lattices consisted of a marginal fraction of spot numbers (less than 10% of overall spots). Applying all these reprocessing changes lowered the identified spots to 8007, down from 86266 reported by autoPROC. While this might be a lower number for the identified spots than before, the percentage of indexed spots dramatically increased to around 99.5%. In absolute numbers 7966 spots were indexed by manual reprocessing in XDS, while autoPROC reported 3451 spots (4% of identified 86266 spots) as indexed. The indexed spots of all 3600 diffraction images were then taken forward to the integration step in XDS. Upon inspection of the result statistics in XDS for this step, the *c* cell axis parameter stood out with deviating by around 25 Å over the course of all images processed (**Figure 63**). An easy solution to this problem was to only include the frame number range that had an apparent stable *c* cell axis length for another attempt at integration. This reduced the variation from 25 Å to around 2 Å over the selected frame number range 800 to 2300. A final CORRECT step in XDS was applied before producing the output .mtz file with a number of altered statistics compared to the previous autoPROC attempt (**Table 20**). Of all the noticeable changes, the most dramatic difference between the output values of autoPROC and the manual XDS reprocessing run could be observed for the *c* cell axis change of about – 9.46 Å. Distributing this change in cell axis over the whole structure model attributes to a local average distortion of about 0.4 Å which might explain the regular dispersion of difference

map peaks in between the layers of the initial data set – like the stretching and contraction of an accordion.

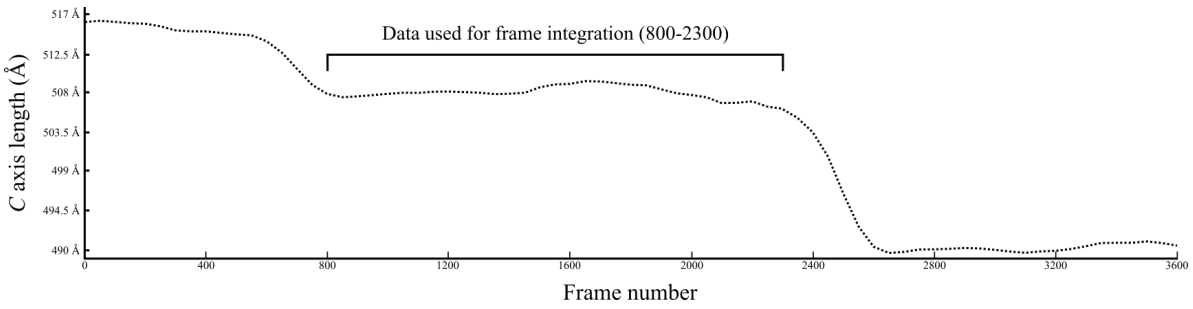


Figure 63 *C* axis length estimation for complete diffraction image set – Initial data integration attempt in XDS for complete diffraction image set (3600 images in total). Only unit cell axis of *c* is shown as *a* and *b* remained almost identical over all frames. Frame number range 800 to 2300 was selected for the final integration step.

Table 20 X-Ray diffraction manual reprocessing results in XDS.

Parameter	**autoPROC**	**XDS reprocessing**
Space group	R32	R32
Unit cell		
a, *b*, *c* (Å)	57.53, 57.53, 525.99	57.39, 57.39, 516.53
α, β, γ (°)	90, 90, 120	90, 90, 120
Resolution (Å)	87.66-1.42 (1.45-1.42)*	44.79-1.40 (1.45-.1.40)
Spots (identified/indexed)	86266/3451	8007/7966
Reflections (observed/unique)	884392/55885	523981/65682
Mean(I/σ(I))	7.0 (1.1)	7.1 (1.4)
Completeness (%)	100.0 (100.0)	99.53 (99.28)
$CC_{1/2}$	99.7 (43.9)	99.8 (68.9)
Multiplicity	15.5	2.0

Processing values reported from autoPROC are shown alongside the results from the manual reprocessing of the diffraction images in XDS. Values that deviated to a noticeable extent are highlighted in magenta.

*Values in parenthesis represent the highest-resolution bin.

The new .mtz file was imported into CCP4i2 and another molecular replacement PHASER run initiated; this time with the solvent-free intermediate model of S741Q1054(GCN4) described in **Figure 61**. A unique solution was found with a translation function Z score of 97.78 and a refined log-likelihood gain value of about 29071. A REFMAC5 job was queued for this PHASER solution and the reported R_{work}/R_{free} values after ten cycles was 27.3%/29.6%. At this point, the model was exported to be further refined in PHENIX.

5.1.5. Influence of B-factor distribution on model building difficulty

Even though the overall resolution of the underlying data for S741Q1054(GCN4) is high with 1.4 Å, the ease of model building varies dramatically in different parts of the model. This is reflected in the high B-factor variance (also called temperature factor, a measure of the uncertainty of an atom position) from 5.36 $Å^2$ to 67.94 $Å^2$ that comes with an increased difficulty of building the correct side chain orientation the higher the B-factors of the associated main chain atoms are. Especially at the N- and C-terminal end of the model of S741Q1054(GCN4), there is a clear correlation between B-factor and local real-space correlation (**Figure 64, A+C**), the latter being a measure of how well the model matches with the $2F_o$-F_c map. The area with the lowest B-factor was also the one that had the most alternate conformers and even a "hole in the ring" for example with Y986 in layer 40 of S741Q1054(GCN4) (**Figure 64, B**). This would explain why it was more difficult to build the remaining N-terminal layers and parts of the GCN4 anchor after BUCCANEER managed to build most of the layers with acceptable $2F_o$-F_c map features. Copying main chain position from the layers below proved to be a good solution for building the remaining three layers at the N-terminal end of S741Q1054(GCN4). Only the GCN4 anchor remains incomplete.

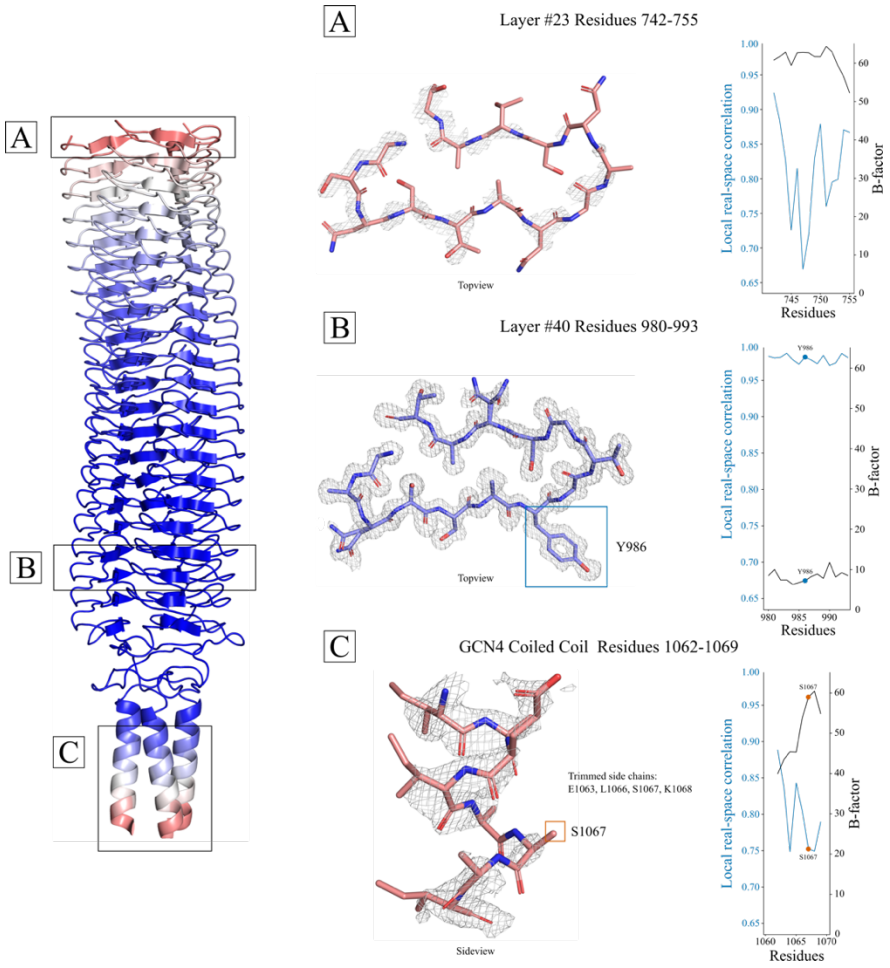


Figure 64 B-factor distribution of the model of S741Q1054(GCN4) – B-factors ranging from 5.36 Å^2 to 67.94 Å^2 (red to blue) are mapped onto main chain atoms shown here in cartoon representation of S741Q1054(GCN4). Corresponding layers highlighted in the full structure (left) are shown alongside the stick representation of the individual residues (right, **A**-**C**). Atoms in stick representation are coloured by B-factor values (except for oxygen and nitrogen atoms) and overlaid with the $2F_o$-F_c map (grey mesh; contoured at 1.5 σ). Graph on the right side showing local real-space correlation (blue; scale starts at 0.64) and B-factor average (black) of relevant residues. **A** Top view of stick representation of residues 742-755 with corresponding local real-space correlation and B-factor graph. **B** Top view of stick representation of residues 980-993. Y986 highlighted both in stick representation and graph to emphasise quality of underlying model-to-map relation (blue). **C** Top view of stick representation of residues 1062-1069. S1067 was highlighted in the stick representation and the graph (red) to show the consequence of trimming residues in areas with poor map quality.

5.2.5 Special refinement considerations including alternative confirmations

The crystal structure of S741Q1054(GCN4) contains one polymer chain (chain A) inside the asymmetric unit. The extension of this polymer chain to a full biologically relevant trimer and the adjacent trimer molecules affects different special refinement considerations along crystallographic inflection points and axes.

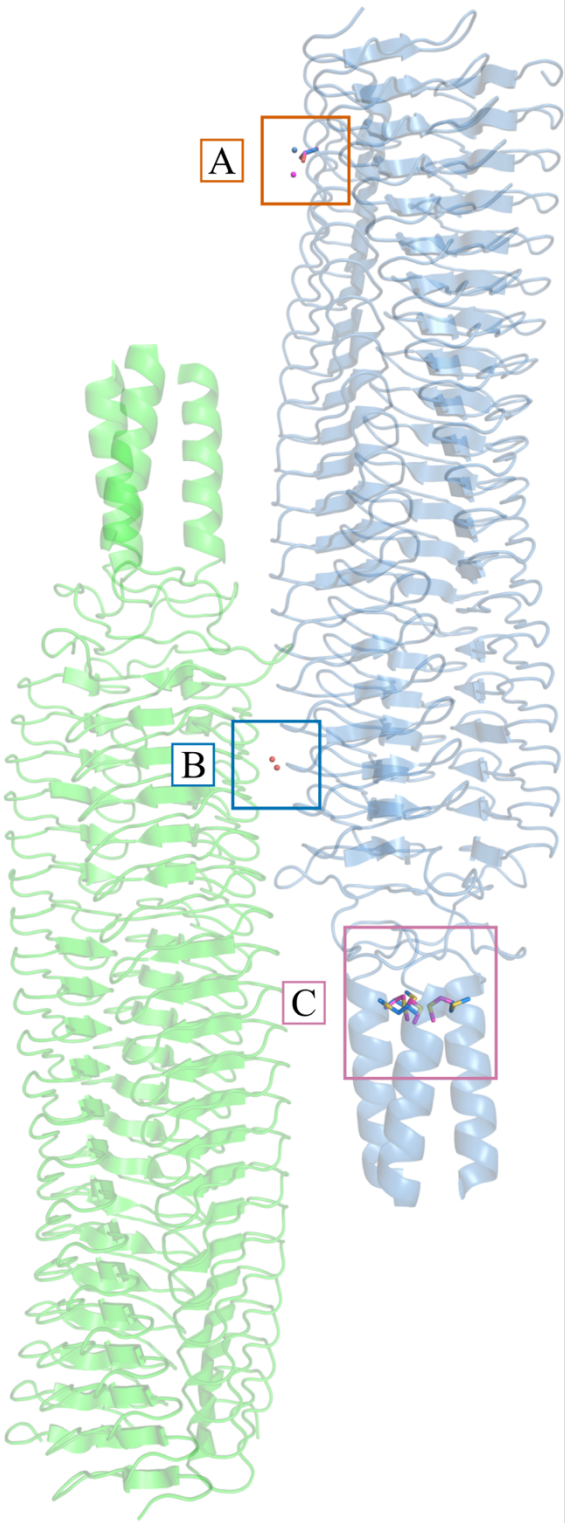
One consequence of a resolution of less than 2 Å for model building is the possibility of alternate conformations for side chain rotamers. Within the model of S741Q1054(GCN4) there were alternate conformers that had interacting solvent molecules (**Figure 65, A**). The occupancy of these solvent pairs were coupled to the occupancy of the side chain atoms of the interacting conformers, adding to a total occupancy of 1 for both alternate conformations.

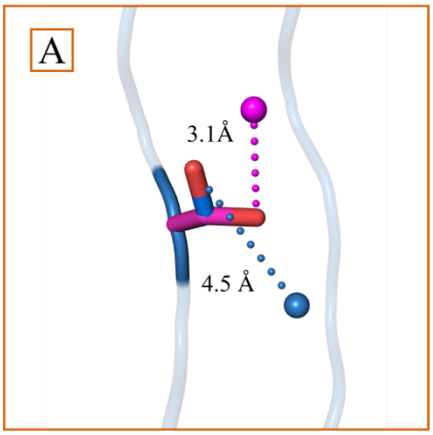
The asymmetric unit in the crystal of S741Q1054(GCN4) contained one monomer in space group R32, which had a few consequences for electron density map calculations around the symmetry points (crystallographic axes) as well as crystallographic contact areas to a second trimer within the crystal. These led to distortions in the $2F_o$-F_c map providing a challenge for accurate model building of solvent molecules or side chain rotamers in these areas.

Originally a single solvent molecule was placed at the centre of a large distorted peak located at the interface between two trimers. During refinement however, the solvent molecule position proved to be unstable and constantly jumped to one side of the feature, leaving a clear positive difference map (F_o-F_c) peak at the other side of it. A solution to this problem was to split the relevant solvent molecules into pairs with split occupancy of 0.5 to more accurately reflect the data at this location (**Figure 65, B**).

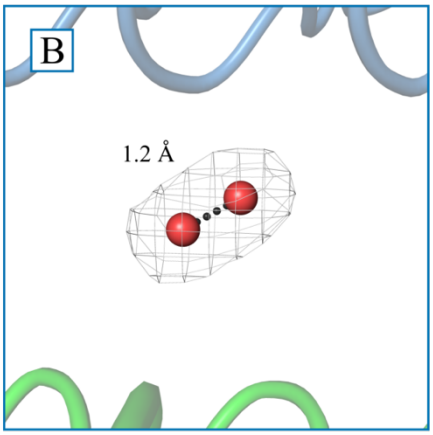
The last major refinement consideration before bringing the structure solution to a conclusion was the challenge that the crystallographic axis posed to the accuracy of the electron density peaks

around that area. For example, even in areas with a very well defined $2F_o$-F_c map of the surrounding main chain atoms, the map appearance closest to the crystallographic axis was much poorer resolved. A good example for this is the methionine which is part of the GCN4 helix: the $2F_o$-F_c map reveals a clear density feature for a single sidechain but leaves additional minor density peaks around the central axis (**Figure 65, C**). Initial attempts to fill these peaks with solvent molecules failed and the best solution found was to split the side chain into two conformers. The proximity of the second conformer closest to the axis was unstable during refinement until the occupancy was fixed to 0.33 and the symmetry related clashes were acknowledged in the refinement options in PHENIX. This provided a stable solution for this problem and likely reflects the underlying data the best. Overall 14 alternate conformations were built with most of them in the areas with low B-factor values.

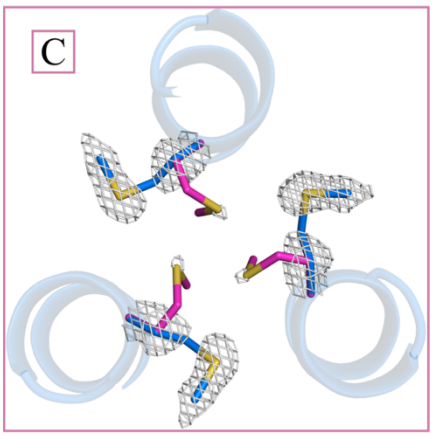




Example 1: Alternate conformations



Example 2: Crystallographic contacts



Example 3: Proximity to crystallographic axis

Figure 65 Overview of special refinement considerations in S741Q1054(GCN4) – Two biological trimers are needed to show the locations of the different refinement considerations before passing these on to PHENIX. **A** An example of one of many different alternate conformations in S741Q1054(GCN4): Ser767 is shown together with an interacting solvent molecule and split into atom groups (blue, conformer group A; magenta, conformer group B) during refinement that sums up to a total occupancy of 1 (conformer A + B). Distances between serine γOH and solvent molecule are displayed next to the coloured dots of the related conformer group. **B** An example electron density map peak distortion in areas of crystallographic contacts. Some density peaks (grey mesh, map contour set at 1.5 σ) are close to contact points between the different polymer chains (here blue trimer and green trimer) which affects the accuracy of the map calculation in this area. In this case, two solvent molecules would be too close to each other so the occupancy for both was reduced to 0.5 to more accurately describe the model-map discrepancy in this location. **C** The proximity of side chains and solvents to the crystallographic axis can lead to clashes during refinement if not specifically acknowledged in the refinement options. In this case the only methionine (yellow for sulphur atom) in the structure had two different conformers: conformer B (magenta) was assigned a fixed occupancy of 0.33 so that the total occupancy of the atoms around the axis would not exceed 1. By logical extension the atoms in conformer A had a fixed occupancy of 0.67. $2F_o$-F_c electron density map peaks (grey mesh, map contour set at 1.5 σ) are shown to highlight the difficulty of building the correct rotamer for conformer B.

5.2.6 Final refinement parameters

Several steps during model building were challenging but ultimately solved: The first problem that needed to be addressed was the poor map quality that could be found at the N- and C-terminal end of the model. Luckily a lot of structural information could just be copied from other parts of the model. For example, as GCN4 is an α-helix, one can extrapolate the general main chain atom locations as the backbone interactions are well-defined for α-helices. For the LPBR layers, the main chain is virtually identical so was essentially "copied" from the layers below to provide a framework for the N-terminal layers. The second hurdle during refinement originated from the high resolution of the X-Ray data. This resulted in the necessity of building alternate conformers in some areas of the model. These needed optimisation as the exact orientation of side chain atoms and the distribution of the occupancy between conformer A and B needed several refinement rounds to come to a conclusion. The F_o-F_c map was very useful in this case as it provided an indication of which residues needed additional alternate conformers. The B-factor variance within the side chain atoms of these conformers were a good indicator if for example a side chain flip (Asn) was needed to create a more accurate model. The last big refinement hurdle was the solvent molecules that were especially difficult to build around areas of electron density map distortions as for example close to pseudosymmetry inflection points or on symmetry axes. This was by far the most time-consuming part of the refinement and needed several rounds of checking solvent-residue interactions, distances between solvents, occupancy of solvent pairs, or logical solvent networks on the inside of the protein. A total of 282 solvent molecules were built in the end. Finally, the structure was submitted for deposition at the PDB and passed all the quality control checks. The PDB code and the final refinement statistics are listed (**Table 21**).

Table 21 Final refinement statistics for S741Q1054(GCN4).

Parameter	Value
PDB accession code	7O23
R_{work}/R_{free} (%)	18.39/21.71
Protein/solvent/ligand atoms	2095/280/13
Average B (Å^2)	
Protein	22.42
Solvent	29.11
Ligand	34.61
R.m.s.d., bonds (Å)	0.006
R.m.s.d., angles (°)	0.82
Ramachandran plot (%)	
Most favoured regions	98.47
Additionally allowed regions	1.53
Rotamer outliers (%)	0.93
Clashscore	4.21
Molprobity score	1.20

Evaluation in PHENIX was carried out together with MolProbity which is part of the refinement suite of PHENIX.

*Values in parenthesis represent the highest-resolution bin.

PHENIX also provides the option to compare several model and refinement values with other structures in the PDB with similar resolutions (POLYGON, reference (Urzhumtseva, Afonine et al. 2009)). This is mainly a visual tool to show how well the model of S741Q1054(GCN4) compares with other structure models (**Figure 66**). The overall clashscore and both r.m.s.d. values (angles and bonds) are in the better end of the available histogram bins, likely as these are mostly unaffected by the quality of the underlying electron density map and more a result of the refinement constraints set within PHENIX. The R-factors and average B-factor of the model of

S741Q1054(GCN4) however are dependent on the quality of the map and ended up with higher values than the histogram bin with the most structures. This discrepancy is likely due to the map distortions and data challenges that were described before which need to be considered when judging the R-factors associated with this model.

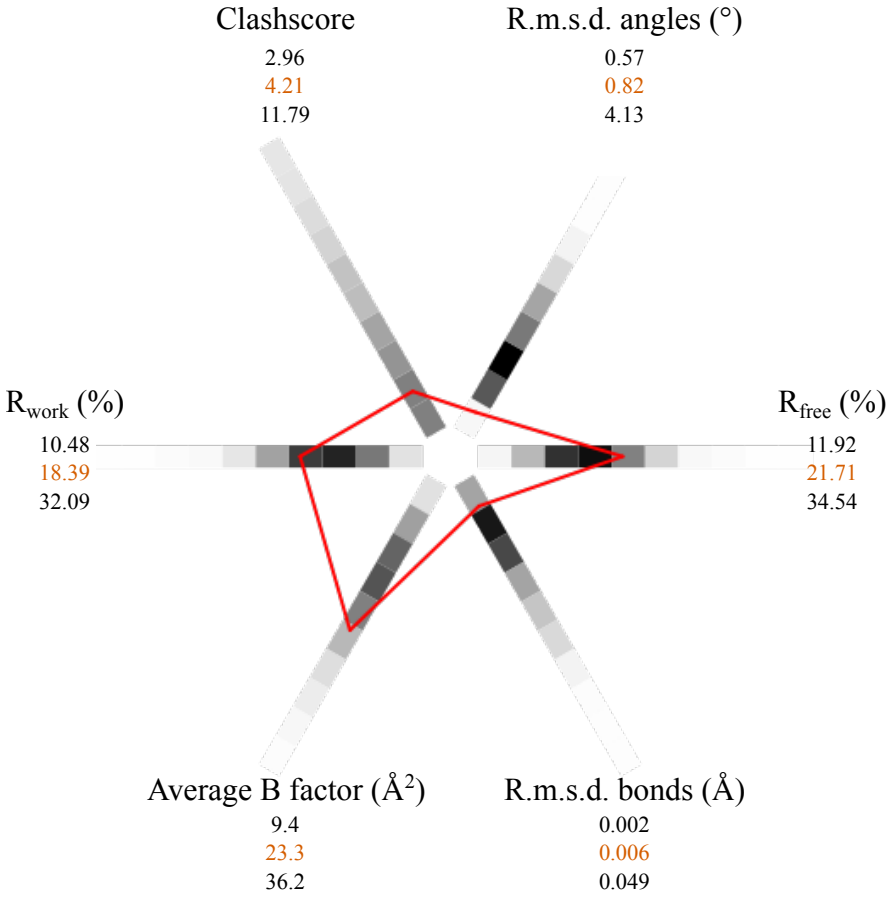


Figure 66 Statistical comparison of the structure solution of S741Q1054(GCN4) provided by POLYGON as part of PHENIX – This is a histogram of the distribution of selected refinement statistics, either given by PHENIX or by MolProbity and comparing it to 2034 PDB entries of similar resolution (Urzhumtseva, Afonine et al. 2009). The range of the values is shown as the best value available within these entries (top of the number groups), the value associated for the model of S741Q1054(GCN4) (orange), and the worst value associated with these PDB entries (bottom of the group). Red line indicates the location of the value for S741Q1054(GCN4) alongside the histogram distribution. The closer to the centre, the better the value is. Greyscale boxes indicate the size of the histogram bins (black containing 735 entries to white with 0 entries).

5.3 Structural expansion of S741Q1054(GCN4) by alignment of repeats

The exceptional connection of sequence-to-structure in TAAs opens up a lot of model prediction avenues with very high accuracy – something that is normally not possible in Structural Biology without at least some form of data backing this up. After solving the structure of parts of the C-terminal head domain of BpaC, I realised that the remaining residues of the head domain follow some strict rules about folding and side chain orientation. In a later section of this chapter, a comparison between different LPBR structures will be made, which shows how closely related the individual 14 residue repeats are (very low r.m.s.d. of the main chain atoms).

From sequence data alone, I was able to identify 14 residue repeats throughout the whole C-terminal head domain. These can be further divided into three families named after the first three residues within the different repeat families: GDN/GEN/GSN. Layer 1 (starting at G434) to layer 32 (ends at S881) are almost identical within their three respective families, differing in only a few residues. While there are differences between each family, the most important position in the repeat is the second residue as it contributes the most to the surface charge profile of the C-terminal head domain in BpaC (**Figure 67, A**). The importance of this finding is discussed in section 5.5.6 onwards. Arranging the individual layers into their respective families gives a good overview of how repetitive this part of the protein is (**Figure 67, B**).

B

	GDN	GEN	GSN	
	Position			Layer
	123456789-1234	123456789-1234	123456789-1234	
Extended domain 434-740	GDNSTASGDNATAS		GTNSTANGTNSTAS	1-2
	GDNSTASGTNASAS	GENSTATGTDSTAS	GSNSTANGTNSTAS	3-5
	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGTNSTAS	6-8
	GDNSTASGTNASAS	GENSTATGTDSTAS	GSNSTANGTNSTAS	9-11
	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGTNSTAS	12-14
	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGTNSTAS	15-17
	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGTNSTAS	18-20
Solved structure 741-1021	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGANSTAS	21-23
	GDNSTASGTNASAT	GENSTATGTDSTAS	GSNSTANGTNSTAS	24-26
	GNNSTASGTNASAT	GENSTATGTDSAAS	GTNSTANGTNSTAS	27-29
	GDNSTASGTNASAT	GENSTATGTASTAS	GSNSTANGANSTAS	30-32
	GAGATATGENAAAT	GAGATATGNNASAS	GTSSTAGGANAIAS	33-35
		GENSTTNGANSTAS		36
	GNGSSAFGESAAAA	GDGSTALGANAVAS	GVGSVATGAGSVAS	37-39
	GANSSAYGTGSNAT	GAGSVAIGQGATAS	GSNSVALGTGSVAS	40-42

Figure 67 Sequence alignment of the C-terminal head domain of BpaC – **A** Frequency plot of residues 1-14 of each repeat (created in WebLogo) for layers 3 to 26. **B** Individual layers with 14 residues can be assigned to an overall repetitive sequence of 42 residues. These can be further split into three families that differ noticeable in the second residue in the repeat, hence why we named these GDN/GEN/GSN type repeat in our publication (Kiessling, Harris et al. 2022). The sequence that covers the parts the structural model of S741Q1054(GCN4) containing the LPBR repeats is highlighted (yellow). The 2 x 42 repeats that were used for the structural expansion to the full C-terminal head domain model (**Figure 68**) is boxed in blue. The majority of the GDN/GEN/GSN type repeats follow a consensus sequence within each family with deviations thereof highlighted in red. Layers 37-42 were deemed as too different from the original families and henceforth excluded from this alignment (grey).

This sequence conservation is the basis of the structural expansion of the model of S741Q1054(GCN4) to a full C-terminal head domain model. The process can be done in PyMOL in the form of a simple copy-align job: First, the most N-terminal two 42-residue repeat segment (G742-S824) was copied into a new independent object. Then, the bottom repeat of this object was aligned to the top repeat of the template model (**Figure 68, A**). The now overlapping part of the aligned segment was then removed, leaving only the newly added repeat object thereby expanding the overall model by a single 42-residue repeat. This process was repeated for a total of seven times to cover the remaining sequence of the head domain (**Figure 68, B**). After this point the sequence transitions to a predicted coiled coil segment with unknown structure, which meant additional residues had to be trimmed to T431, that is the last residue that can be assigned to the model with high confidence (**Figure 68, C**). All the separate objects in PyMOL needed to be combined to a single model by fusing the split peptide bonds, which was performed in Coot (**Figure 68, D**). In addition, the residues that differed from the consensus sequence of the individual repeat family were mutated to match the sequence (**Figure 68, E**). The last remaining challenge was choosing the correct rotamers for these differing residues. To keep it simple, either a rotamer was copied from a similar residue position from the data-driven model of S741Q1054(GCN4) or the most likely rotamer orientation was chosen. The impact of this decision is negligible as this model is only used for electrostatic charge mapping and the actual residues that differ from the sequence is only 6 out of 306. The solved part of S741 to S1021 is already the largest LPBR structure to date, but with the structural expansion to T431 it far exceeds any size expectations for a TAA related structure model using X-Ray crystallography.

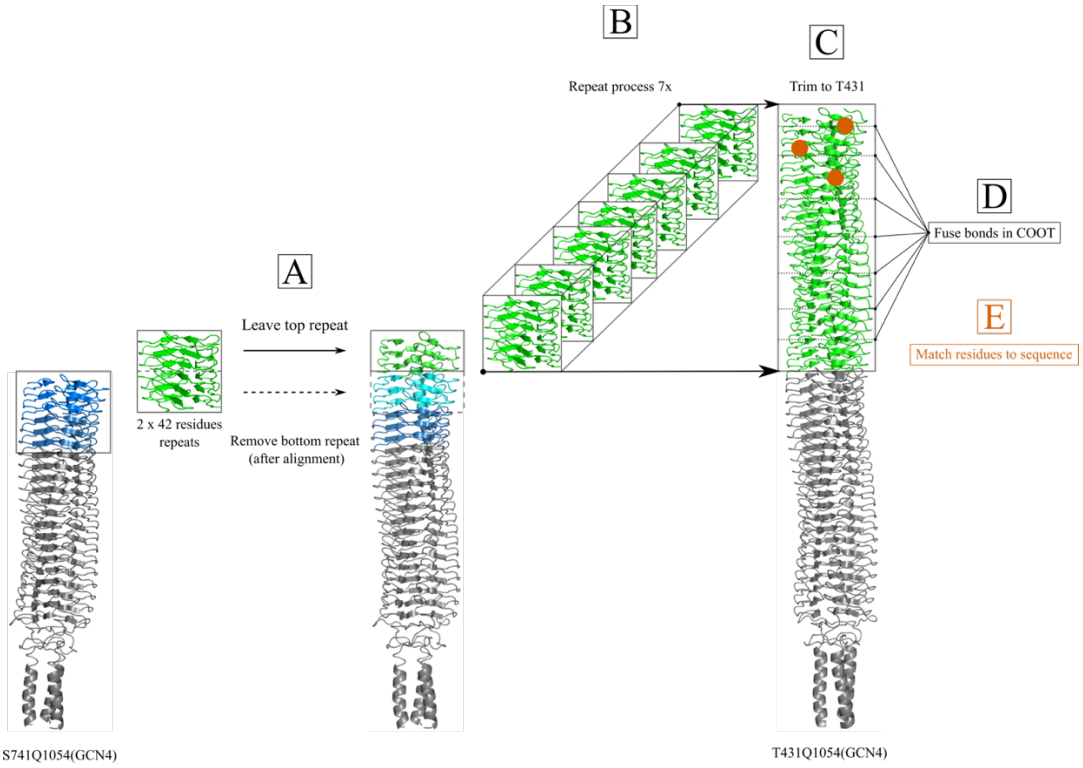


Figure 68 Creation of full C-terminal head domain model of BpaC – The exceptional nature of structural repeats within the C-terminal head domain allows for a structural expansion of the original model of S741Q1054(GCN4) to cover all the way up to the first residue of the LPBR motif at T431. This extended model was created in PyMOL as follows: **A** Starting at the first complete layer in the template model at G742, two full 42-residue repeats were copied into a new object (green) in PyMOL. The bottom 42-residue repeat of this object was then used to align to the former top repeat (turquoise) of the template model. The bottom repeat was then removed, leaving the newly added 42-residue repeat in its correct position. **B** This process of copying two repeats into a new object, aligning the bottom part of the new object to the top part of the template, and removing the bottom repeat, was repeated seven times until the N-terminal end of the C-terminal head domain was exceeded. **C** The extra residues in the last object alignment were then removed until T431 was the actual start of the C-terminal head domain model. **D** The complete model with all its still separate repeat objects was transferred to Coot in which the peptide bonds between the individual repeat objects were fused. **E** Although most of the sequence in that area of the protein was identical to the repeats below, some sequence deviations had to be addressed by copying the rotamer configuration of the same residue at this location in a layer below or simply choosing the most-likely rotamer of this residue (crosscheck with **Figure 67**).

5.4 Structural analysis of S741Q1054(GCN4)

The model building of S741Q1054(GCN4) was a challenging process, mainly due to the additional effort that was put into the accuracy of the solvent network and the alternate conformers. A positive consequence of this process however was the ability to look very closely into the intricate interactions that make this domain different when compared to similar LPBR folds. Initially, this domain was thought to be "just another" LPBR with no new features. However, careful analysis opened up so much new insights which ultimately led to a full-fledged story, worthy of publication. This section will describe the structural model of S741Q1054(GCN4) in detail, highlighting the structural features that provide BpaC with stability, the abnormal hydrophilicity of the trimer core facing residues, and the exceptionally negative charge of the solvent-facing side of the model.

5.4.1 Structural highlights of S741Q1054(GCN4) with a likely impact of stability

The C-terminal head domain of BpaC has a remarkable overall length of about 20 nm, as was estimated from the extended model described in section 5.3. How such a long structure can be maintained is partially due to a variety of stabilising elements of the C-terminal head domain as can be observed in the structural model of S741Q1054(GCN4) (**Figure 69**). The highly repetitive nature of the almost identical repeats in the GDN/GEN/GSN families gives an idea of how important the continuation of a stabilising element throughout the whole C-terminal head domain is.

A plethora of hydrogen bond interactions can be identified in S741Q1054(GCN4) that produce interconnectivity between the LPBR layers either on the outside of the protein, in between individual monomer layers, and in between trimer core facing residues. Starting from the outside of the protein, there is a long, continuous hydrogen bond network consisting exclusively of contributions from the side chain of N@3 (**Figure 69, A**) that is conserved from layer 32 all the way up to layer 1 at the N-terminal end of the C-terminal head domain. Only layers 23 to 32 are visible in the model of S741Q1054(GCN4), but the high sequence identity in the rest of the layers allows to draw accurate conclusions of the remaining structure of the C-terminal head domain. Another example of an external hydrogen bond network can be found at the repetitive sequence of N/D/N@10 (compare with **Figure 67**), which relies on the flipping of asparagine to form the necessary hydrogen bond interactions to the aspartate below and above the motif shown (**Figure 69, B**). There are also additional solvent molecules that help to bridge too long distances between the side chains of aspartate and asparagine (middle and bottom side chain in **Figure 69, B**).

Moving more inwards into the structure, the LPBR layers are what likely contribute most to the stability of this domain. The side chains of the core facing residues at position 5 and 7 fit together

nicely – like a zipper made out of three parts. BpaC has mostly hydrophilic residues (T, N, S) at these positions and an example motif is displayed with T@5 and T@7 together with the central solvent molecule that is shared by the γOH of each T@7 (**Figure 69, C**). The general backbone structure of each layer has a lot of interconnecting hydrogen bonds that also can be observed in other LPBR structures like the head domain of YadA (PDB: 1P9H, (Nummelin, Merckel et al. 2004)). One can find hydrogen bonds in between main chain atoms within a single repeat (**Figure 69, D, h1-h3**) or in between the different layers above and below (**Figure 69, D, h4-h12**).

One last feature to highlight is the occurrence of a salt bridge in the transitional neck motif at the far C-terminal end of BpaC. This neck motif is needed to transition between the broader C-terminal head domain, made out of primarily β-sheets and the narrow coiled coil within the β-barrel of BpaC. What is interesting in this particular salt bridge is the fact that each component of this three-part bridge originates from a different monomer: in the example of the figure, the negatively charged D1048 of the first monomer bridges over to the positively charged R1036 of a second monomer until it reaches the negatively charged E1022 of the last monomer completing the salt bridge (**Figure 69, E**). This strengthens the interconnectivity of the individual monomers in the neck motif area.

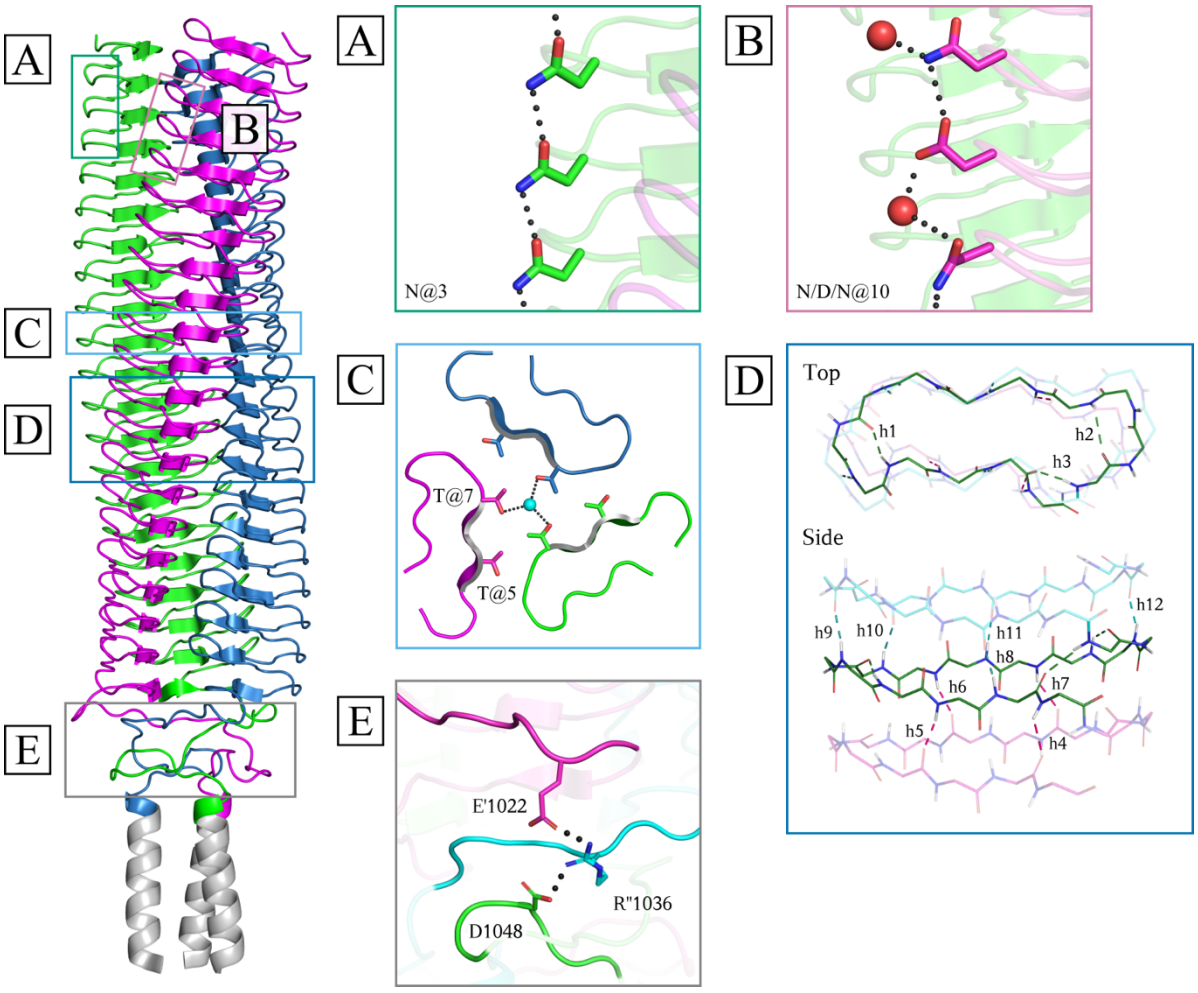

A
B
C
D
E
A
N@3
B
N/D/N@10
C
T@7
T@5
D
Top
h1
h2
h3
Side
h9
h10
h11
h12
h8
h6
h7
h5
h4
E
E'1022
R"1036
D1048

Figure 69 Structural highlights of the model of S741Q1054(GCN4) – Several features that contribute to the stability of S741Q1054(GCN4) are displayed. **A** Hydrogen bonds in between N@3 sidechains forming a continuous network from layer 1 to layer 32 (only layers 23-32 included in model), highlighted for monomer A (green box). **B** Another hydrogen bond network of two N@10 residues surrounding a central D@10 (magenta box, highlighted for monomer B) stretching from layer 1 to 29 (layers 23-29 visible). Additional solvent molecules helping to bridge the larger gaps between sidechain atoms are shown (red spheres) but may differ in other areas of the protein depending on rotamer orientation (e.g. flipping of one or both Asn sidechains creating attraction/repulsion forces). **C** An example of a common LPBR core motif in BpaC with T@5 and T@7 (turquoise box) as the only core facing residues of the repeat, with an additional interaction to a central solvent molecule shared by all three threonine residues. **D** Example of the stabilising effects of parallel β-sheet stacking within and between layers (blue box). Hydrogen bond interactions between main chain atoms of a single 14 residue repeat of BpaC (top view, h1-h3) and between main chain atoms of the surrounding top and bottom repeat (side view; h4-h7 middle to bottom, h8-h12 middle to top). **E** Salt bridge inside the most C-terminal neck motif of BpaC (grey box).

5.4.2 Trimer core motifs of S741Q1054(GCN4) are mostly hydrophilic

The C-terminal head domain of BpaC is made up of 42 repeats, each consisting of 14 residues, that are referred to simply as layers. While only layers 23-42 are visible in the structural model of S741Q1054(GCN4), they cover all the possible core motifs in the C-terminal head domain of BpaC. A core motif refers to residues 5 and 7 in the 14 residue repeat, both facing towards either the adjacent monomers or to the trimer axis itself (**Figure 70, A**). Most of the core motifs in BpaC consist of hydrophilic residues (T@5 and S/T/N@7 for the most part) which is a novelty in itself as YadA, the first LPBR structure published, exclusively consists of hydrophobic residues (V/I@5 and I/V@7). The structural model of S741Q1054(GCN4) was trimmed to G434S1021 to only show the LPBR layers 23-42 for better illustration of core motif distribution (**Figure 70, B**). An interesting cooccurrence in the hydrophilic core motifs in BpaC is the appearance of solvent molecules around the trimer axis, something that has not been observed before in any LPBR motif to date. For the motifs with S@7, one solvent molecule per sidechain could be observed in some but not all of the layers with this residue at that position. A similar observation can be made for T@7, which shares a single solvent molecule for all three core facing side chains, but not all of the layers with this residue-position combination have solvent molecules at the centre. A likely explanation for this is the changing map quality especially towards the more N-terminal end of the model of S741Q1054(GCN4). Another interesting core motif can be observed with N@7 in the respective layer: the sidechain orientation for this is important for the ability to form hydrogen bonds to the other asparagine sidechains in the adjacent monomers. An attempt of splitting the sidechain orientations was performed: a split with 2/3 in the original position and 1/3 with a flipped sidechain as an alternate conformer would more accurately reflect the likely "real" distribution of asparagine sidechains in this motif (as shown in **Figure 70, B** for "T N"). PHENIX refinement

however was not stable and would always flip the sidechain back, leaving all of the alternate conformers of the affected asparagine identical to the original rotamer. This is likely due to the fact that only a single monomer per asymmetric unit existed in the crystal. This would average out any trimeric arrangement around the trimer axis, which is also the crystallographic axis in this case.

The biggest assortment of central solvent molecules can be found in the core motifs of layers 35 and 36. Layer 35 has a unique G@7, which does not occur in any other core motif in BpaC nor in any other solved LPBR structure. The consequence of the absence of the sidechain at the crucial residue position 7 is a large cavity that is filled by a tetrahedral arrangement of solvent molecules with three solvent molecules stabilised by the N@7 from the layer below (layer 36) and the central solvent molecule supported by the surrounding three solvent molecules with an additional interaction from the T@7 central solvent from above (the latter not shown in **Figure 70** for clarity). The core motifs of layers 37 to 42 are more closely related to the ones present in YadA and other LPBR structures as these contain hydrophobic residues at position 7 of the 14-residue repeat, however some of them still contain hydrophilic residues at position 5 (e.g. the "S F" core motif). This also coincides with the strongly reduced sequence similarity of these layers compared to the virtually identical layers 1 to 32. This further supports the hypothesis that the combined biochemical features of the almost identical layers 1 to 32 (or to 36 to some extent) are essential for the exceptional length of the C-terminal head domain of BpaC.

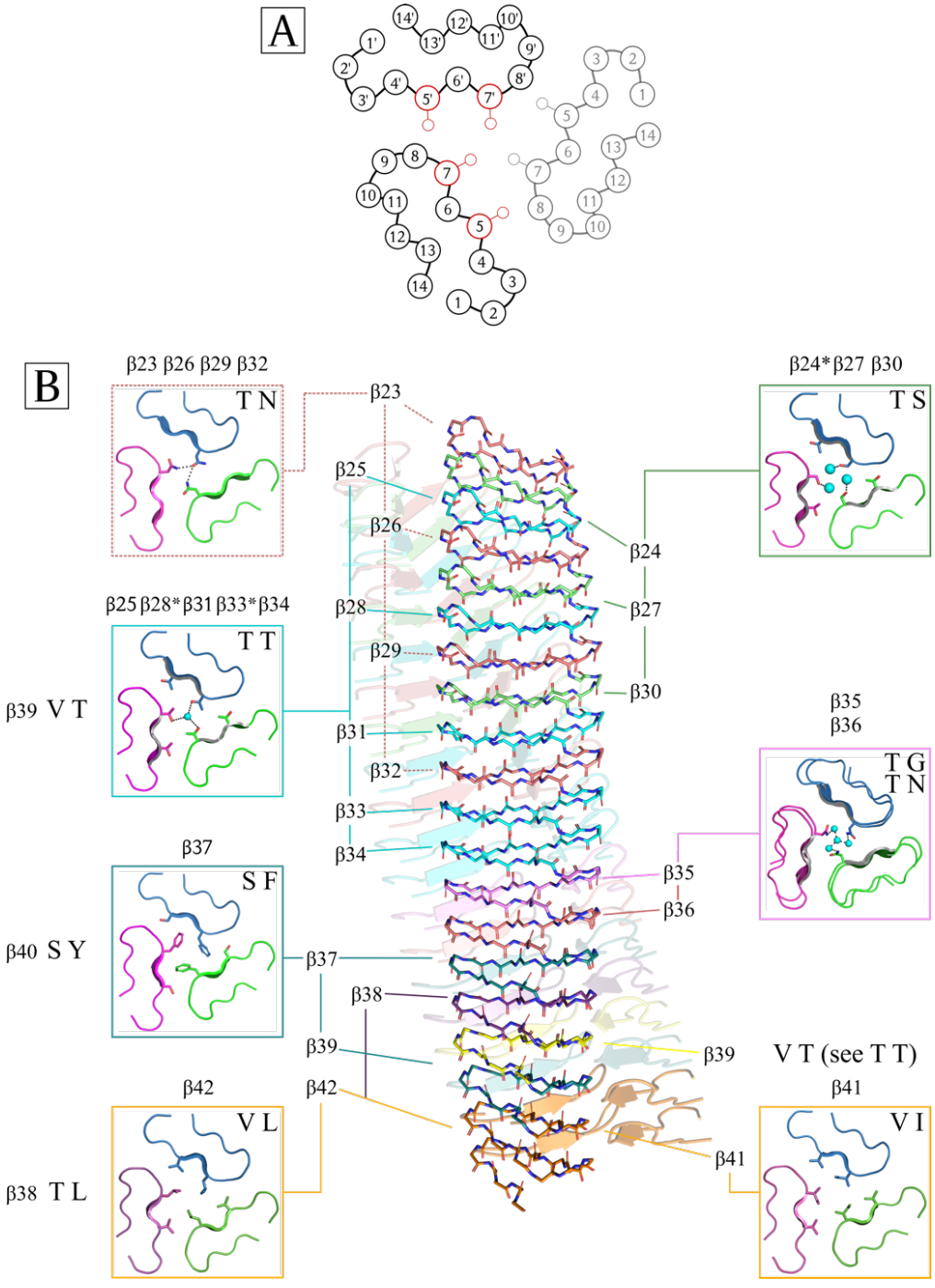

A
B
β23 β26 β29 β32
T N
β24*β27 β30
T S
β25 β28* β31 β33* β34
T T
β39 V T
β35
β36
T G
T N
β37
S F
β40 S Y
V T (see T T)
β42
V L
β38 T L
β41
V I
β23
β25
β26
β24
β28
β27
β29
β30
β31
β32
β33
β34
β35
β36
β37
β38
β39
β42
β41

Figure 70 Overview of all LPBR core motifs visible in S741Q1054(GCN4) – LPBR layers are numbered from the start of the C-terminal head domain (G434) to the last residue in the last layer (S1021, β42). Only layers 23 to 42 (β23 to β42) are visible in the structural model of S741Q1054(GCN4) (shown are only residues 434 to 1021 in this figure). **A** Top view of a schematic of an LPBR layer with all three monomers shown. Residues within a repeat are numbered at the position of the Cα atom. Position 5 and 7 within the repeat are shown with the position for the Cβ atom (red). **B** An overview of core motifs in BpaC with layers, shown in stick representation for all main chain atoms, associated to the respective core motif by connecting lines and numbers with matching colours. Letters within the boxes represent the residues at position 5 and 7 of the motif. Hydrogen bond interactions between side chains and to solvent molecules are shown (grey dots). Structurally similar motifs have been grouped together and the additional motifs are only indicated with letters and layer position next to the motif that is similar to them (e.g. S@5 and Y@7 only has an additional oxygen atom compared to S@5 and F@7 with the same rotamer configuration but only the "S F" motif is shown in full). Some layers come with additional solvent molecules (cyan spheres) but the model of S741Q1054(GCN4) does not show solvent molecules in all of the motifs (the ones without solvent molecules at the central position marked with *). A special case is shown with the layers 35 and 36: the G@7 of layer 35 (β35) creates a unique cavity inside the trimer core of BpaC which is filled by solvent molecules that likely exist because of an interaction to the sidechain of N@7 from the layer below (β36), supporting an extensive central solvent network of a total of four tetrahedral solvent molecules present in layer 35.

5.4.3 S741Q1054(GCN4) is highly negatively charged at neutral pH

One of the most striking features of the C-terminal head domain of BpaC is the abundance of negatively charged residues at the outside of the trimer. A more exact comparison with other LPBR models is given in the following sections.

The very regular 14-residue repeats allow the creation of frequency plots (WebLogo, (Crooks, Hon et al. 2004)), i.e. a plot that shows the relative distribution of amino acids at a given position in the repeat. Creating this frequency plot for all LPBR layers in BpaC (G434 to S1021) provides a good overview of the impact of certain residue positions on the biochemical importance of the C-terminal head domain for the rest of the protein or the surrounding area. The exact position of charged residues in relation to its adjacent monomer should be highlighted as well: D/E@2 is situated at the edge of the outside loop of one monomer, while residues D/E@9 and D@10 are located on the opposite side of the gap between both monomers (**Figure 71**, **A**). This likely has an impact on the interaction between adjacent monomers. Mapping the charged side chains on the cartoon representation of all LPBR layers in BpaC better illustrates this observation (**Figure 71**, **B**). A better graphical representation of the consequence of having that many charged residues spread over the C-terminal head domain can be achieved by plotting the electrostatic surface charge volume onto the model of T431S1021 (APBS plugin in PyMOL; **Figure 71**, **C**).

The charge distribution also correlates well with the sequence conservation within the C-terminal head domain of BpaC: most charged residues are found in layers 1-29 with a decrease in the more C-terminal layers 30-42 (which also have a mostly hydrophobic trimer core). This adds another correlative observation: not only is the negative charge of the C-terminal head domain something not observed before for LPBR motifs but it correlates with the occurrence of residues with

hydrophilic sidechains facing towards the inside of the trimer. How exactly this correlation has a causative connection remains unanswered but may be of functional importance.

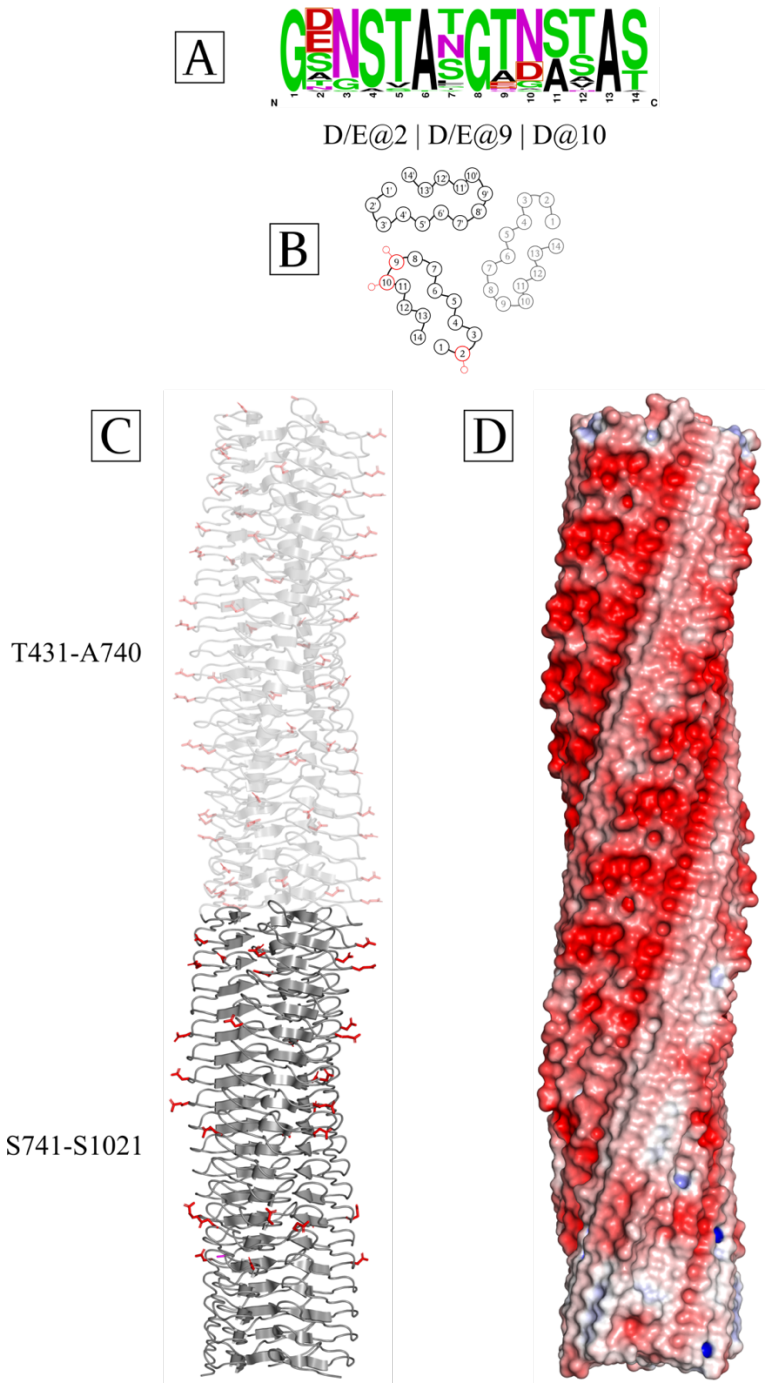


Figure 71 Solvent-accessible charged residues in the C-terminal head domain of BpaC – **A** Frequency plot of residues 1-14 of all LPBR layers in the C-terminal head domain of BpaC (G434 to S1021) made in WebLogo. Amino acids are coloured by biochemical properties. Negatively charged residues at pH 7 are highlighted (red box around red letters). D or E at residue position 2, 9, and 10 all contribute to the negative surface charge of the C-terminal head domain of BpaC. **B** Schematic of 14 residue repeat with positions 2, 9, and 10 highlighted in monomer A (red). **C** Cartoon representation of T431S1021 with negatively charged residue side chains shown in stick representation (red). T431 to A740 are shown as transparent to emphasise the uncertainty of that model as these repeats are inferred from the template model of S741Q1054(GCN4). **D** Electrostatic surface charge representation of T431S1021 performed by the APBS plugin in PyMOL (Schrodinger 2017). Colour correlates to charge distribution at pH 7 (red to blue, negative to positive).

5.4.4 Discovery of multiple solvent channels in S741Q1054(GCN4)

Solvent molecules are often an underrepresented part in the analysis of structural models solved by X-Ray crystallography. This is mainly due to the fact that the solvation shell around protein molecules is always present but the main focus in terms of functional importance is usually on the protein chain itself. However, when realising how unusual the hydrophilic core motifs are in the wider context of LPBR folds, I also identified an array of solvent molecules around the trimer axis and in between monomers that are highly repetitive and organised. As some of these solvent molecules run all the way from the first LPBR layer in the model to the last one, I decided to call them “solvent channel”, as they are water molecules running along a certain pre-defined path just like a channel. Three separate channels have been defined: the “outer solvent channel”, the “inner solvent channel”, and the “central solvent channel” (**Figure 72, A & B**). In all of these channels: interactions within a single LPBR layer can be found in an almost identical manner in all the other layers with the same residues at the same position. Both the inner and core channel are not as

continuous as the outer channel but have well-defined interaction points to the protein chain, mostly via hydrogen bonds stemming from hydrophilic residues (**Figure 72, C**).

The outer solvent channel can be identified by four hydrogen bond/dipole interactions per solvent molecule in between two monomer chains: two conserved interactions stem from backbone atoms while the other two are from hydrophilic residues like T@5 (**Figure 72, D**). The outer solvent channel is highly conserved in all LPBRs, likely due to the residue-independent nature of two of the four interactions, but has not been described before in the literature.

BpaC is the first LPBR structure that has additional solvent channels described: the inner and central solvent channel. The inner solvent channel only exists because of the hydrophilic nature of the very regular S/T/N@7 repeat that can be found throughout most of the C-terminal head domain of BpaC (for reference, see GXN alignment in **Figure 67** in section 5.3). Focussing on a single S/T/N/S repeat, there are hydrogen bonds from the γOH of Ser and Thr stemming from monomer A, while the interaction from the side chain atoms of N@7 to the central solvent molecule is from monomer B respectively (**Figure 72, E**). This "break" in the repeat sequence is needed to make full use of all interactions from N@7 passing over to the start of the next repeat with S@7 starting all over again. Additional interactions originate from backbone carbonyls belonging to residue 6 of the 14-residue repeat.

The interactions that hold the central solvent molecules in place are due to the hydrophilic nature of the residues at position 7, mainly S/T@7. These residues only differ by a γCH_3 group, but have different interactions with the solvent molecules associated with the structurally identical γOH. In most layers with T@7 in the structural model of S741Q1054(GN4), the γOH of all three T@7 interact with a single central solvent molecule (**Figure 72, F**). The special case here is that it sits exactly on the trimer axis, which is also the crystallographic axis. Out of the three possible

hydrogen bonds only one "real" hydrogen bond can exist at any given time from the γOH of a T@7 to this central solvent molecule. Consequently, the occupancy of this solvent molecule was set to 0.33 in the final model to reflect the correct situation in the (crystallographically extended) trimer model. As mentioned before, although Ser and Thr only differ by a CH_3, the additional space around the trimer axis at S@7 leads to a rearrangement of solvent molecules in that area: one central solvent molecule per S@7 can be observed, also again interacting with the γOH of S@7 (**Figure 72, G**). In this particular layer a network connection from the central solvent channel to the inner solvent channel is present via a shared interaction to the γOH of S@7. The consequence of this interconnectivity can only be hypothesised without any bond energy calculations but one can suspect it is very critical to the stability and possibly function of the C-terminal head domain of BpaC.

The least regular channel of the three channels, the central solvent channel, required a more detailed illustration to show the specific layers that harbour central solvent molecules (**Figure 73, A-C**). The detailed interactions for S/T@7 are already discussed but included in the figure to show the difference to a special motif worth mentioning: a unique G@7, so far only found in BpaC and additionally to that only present in one of the 42 LPBR layers. What is interesting here is that the cavity created by the absence of a side chain in Gly led to an influx of central solvent molecules around the trimer axis, which supported by hydrogen bond interactions from N@7 from the layer below (**Figure 73, D**).

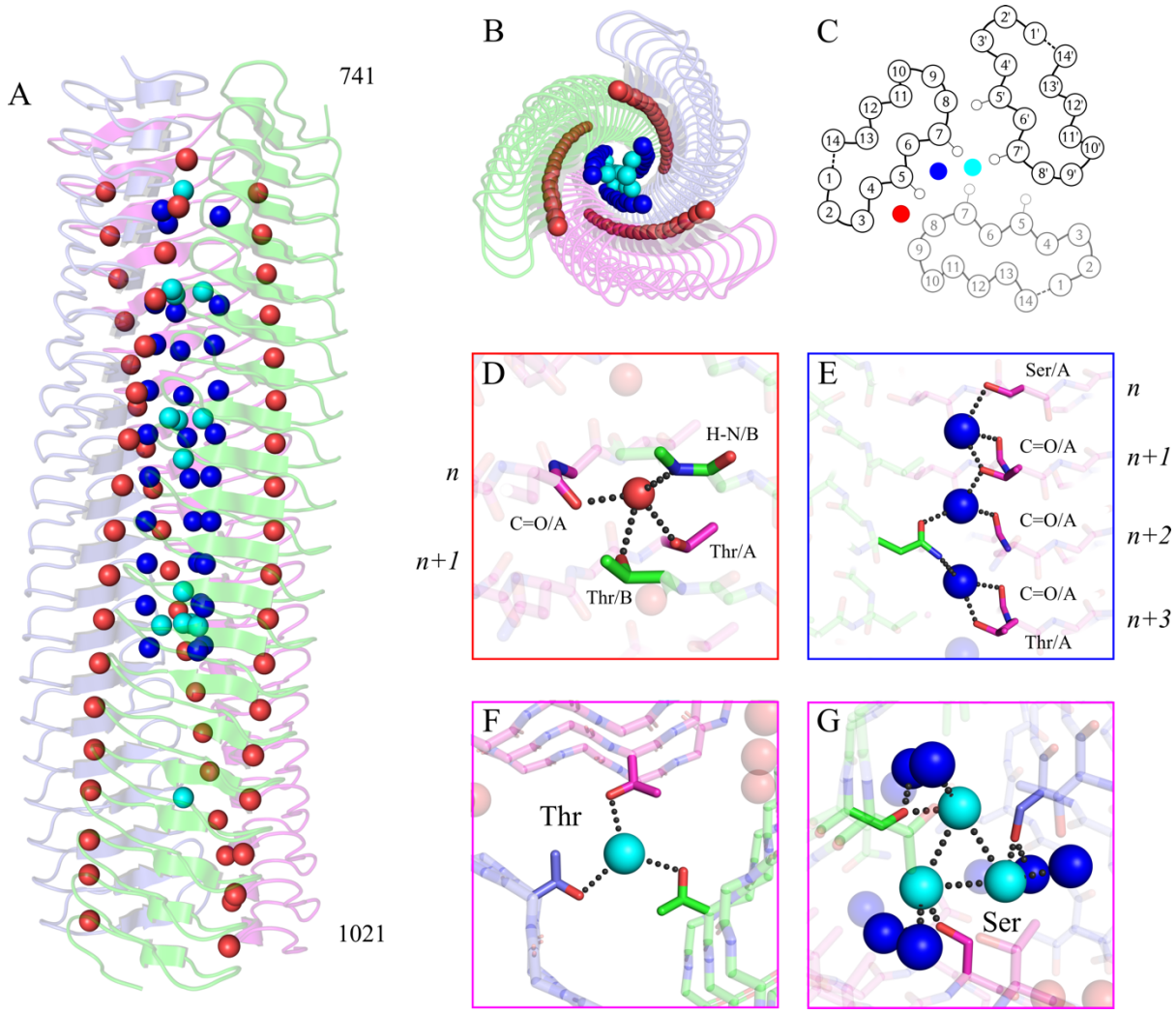

A
741
1021
B
C
D
H-N/B
C=O/A
Thr/A
Thr/B
n
n+1
E
Ser/A
C=O/A
C=O/A
C=O/A
Thr/A
n
n+1
n+2
n+3
F
Thr
G
Ser

Figure 72 Overview of solvent channels in S741Q1054(GCN4) – Solvent molecules that are structurally relevant are separated into different "channels" depending on their relative location to the trimer core. Solvent molecules that are not part of these channels are omitted in this illustration. Monomer A-C are coloured for a better explanation of the relevant hydrogen bond interactions (magenta = A, green = B, blue = C). Hydrogen bond interactions shown as black dots and layers numbered in relative position (*n* = any layer number). **A** Side view of cartoon representation of S741S1021 with outer (red), inner (dark blue), and central (cyan) solvent molecules shown as spheres. **B** Top view of cartoon representation explained in **A**. **C** Schematic of 14 residue repeat with location of Cα atoms shown with position number and an additional display of the position 5 and 7 Cβ atoms. Location of solvent channels indicated by coloured spheres. Symmetry equivalents of the channels in between the other monomers are not shown but can be logically inferred. **D** Hydrogen bond interactions (black dots) displayed for an example case of a solvent molecule belonging to the outer solvent channel. **E** The hydrogen bond interactions with the inner solvent channel molecules are shown. **G** In the layers with S@7 one can observe not only hydrogen bond interactions from the γOH of S@7 to central solvent molecules (cyan) but also to the inner solvent channel (shown in **E**) creating an extensive network in between all these solvents. Figure used in publication without modifications (Kiessling, Harris et al. 2022).

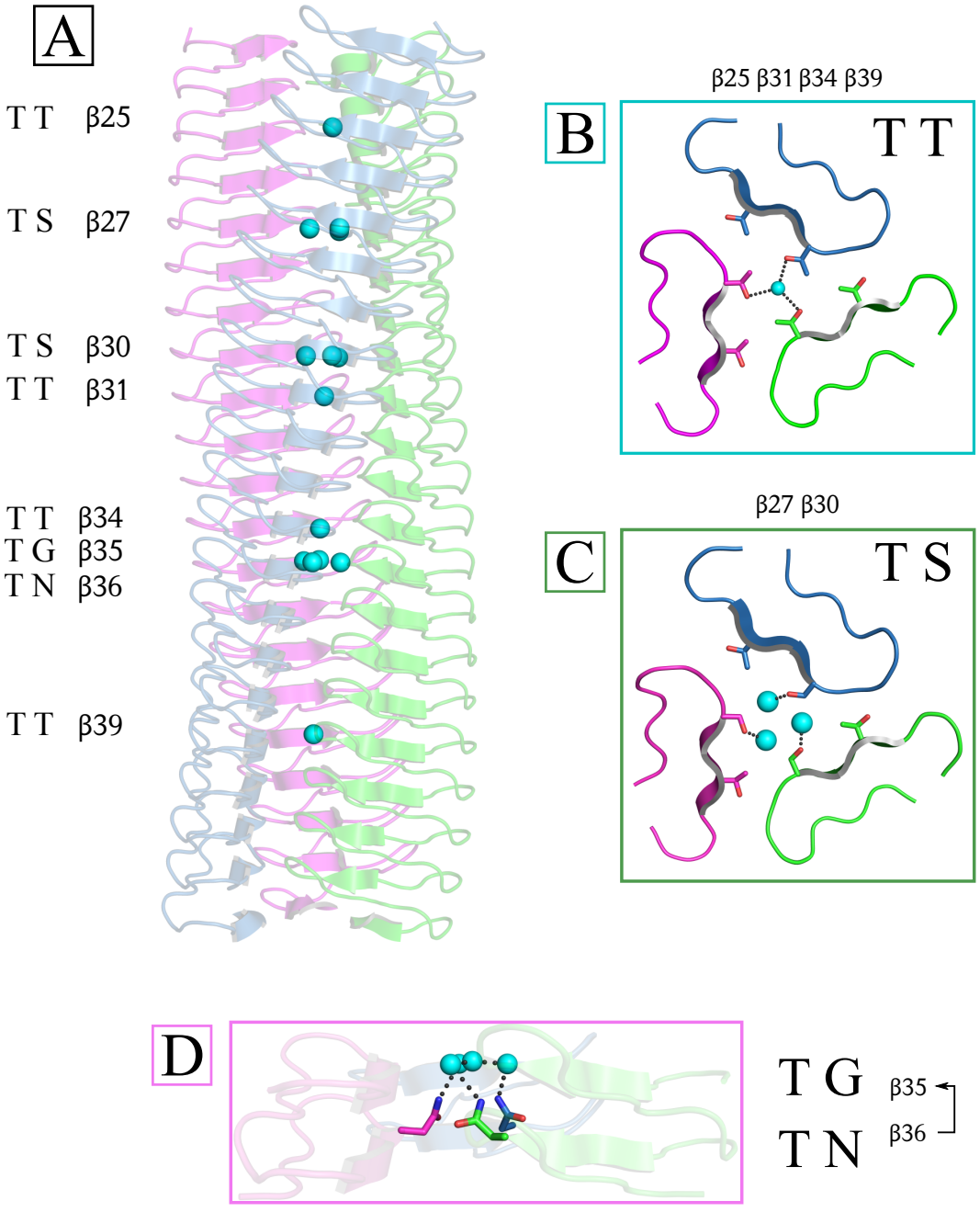

A
T T β25
T S β27
T S β30
T T β31
T T β34
T G β35
T N β36
T T β39
β25 β31 β34 β39
B
T T
β27 β30
C
T S
D
T G β35
T N β36

Figure 73 Central solvent molecules of S741S1021 – Cartoon representation of S741S1021 (LPBR only) with coloured monomers. Central solvent molecules coloured as cyan spheres with hydrogen bond interactions shown as black dots. Stick representation for selected side chain residues. **A** Layers with central solvent molecules labelled with one-letter code representation of residues 5 and 7. Layers are numbered (βX) from the first (predicted) LPBR in the C-terminal head domain of BpaC (G434) downwards. **B** Layers with the core motif T@5 and T@7 and a single central solvent molecule are listed and an example layer is displayed with the relevant sidechain interaction to the central solvent molecule. **C** Layers with the core motif T@5 and S@7 and three central solvent molecules are shown with an example layer that illustrates the hydrogen bond interaction between the γOH of each S@7 and the associated central solvent molecule. **D** The special central solvent network in the layers 35 and 36 is displayed. The cavity that the lack of a sidechain in layer 35 at position 7 creates (G@7) is filled in by solvent molecules that are stabilised from N@7 from the layer below. Figure used in publication without modifications (Kiessling, Harris et al. 2022).

5.5 Comparison of S741Q1054(GCN4) with all available LPBR structures

The outstanding features of the C-terminal head domain of BpaC became even more remarkable when comparing the structural model of S741Q1054(GCN4) to all the other LPBR containing structures in the PBD: it is by far the largest structure of all published (even before the logical structural extension to T431), has a uniquely hydrophilic trimer core with an extensive solvent network, and the charged residues at the surface of the trimer are all negatively charged. An overview of LPBR containing structures, used for comparison is given in **Table 22**.

Table 22 Overview of TAAs with LPBR containing structures.

Name	Total length of TAA (aa)	PDB accession code	Residues in trimmed PDB	Reference
AtaA	3630	3WP8	2989-3104	(Koiwai, Hartmann et al. 2016)
BoaA	1626	3S6L	1396-1506	(Edwards, Gardberg et al. 2011)
BpaC	1152	–	434-1021	This thesis
EibD	511	2XQH	160-273	(Leo, Lyskowski et al. 2011)
UspA1	863	3PR7	68-255	(Agnew, Borodina et al. 2011)
YadA	455	1P9H	65-188	(Nummelin, Merckel et al. 2004)

The total length of each TAA is given to estimate the relative location of the domain within the sequence. PDB accession code associated with the structure is listed. Structural models within the PDB were trimmed to only contain the relevant LPBR motifs.

I compared the different core motifs in all available LPBR containing models and observed a progression from hydrophobic core motifs to hydrophilic core motifs. I further compared the solvent channels of BpaC with that of the aforementioned models. This section also unveils a new subcategory suggestion for LPBR containing structures that was the consequence of a simple, yet important observation: the relative position of a LPBR head domain within the solvent-accessible passenger domain of a TAA correlates with its surface charge in an almost linear fashion. In other

words: the more N-terminal the head domain is located, the higher is the positive-to-negative charged residue ratio. The new generation of structure prediction methods also allowed to test this hypothesis on selected LPBR motifs (no experimentally obtained structure associated) to see if one can anticipate the surface charge of the structural model just by its domain position.

5.5.1 Superposition of main chain trace of all available LPBR structures

Comparison of the LPBR containing structures shows how well the main chain atoms of the other structures overlap with the LPBR motifs of BpaC, with a typical r.m.s.d. value of 0.22 Å to 0.57 Å for the Cα atoms (**Table 23**). The main chain atoms of the LPBR motifs in BoaA (PDB: 3S6L) and BpaC showed the closest similarity. They are the most closely-related of those studied with both being from *B. pseudomallei* and is likely the reason molecular replacement with the PDB model parts of the C-terminal head domain of BoaA (3S6L) was immediately successful.

Table 23 Alignment of main chain atoms of selected TAA models.

Name	PDB ID	Residues	Alignment method		
			super (Å)	*cealign* (Å)	*minimal* (Å)
BpaC	7O23	431-1021	–	–	–
BoaA	3S6L	1397-1504	0.32	0.37	0.22
AtaA	3WP8	2994-3106	0.94	1.13	0.57
EibD	2XQH	160-273	0.38	0.98	0.25
UspA1	3PR7	66-283	1.10	1.68	0.42
YadA	1P9H	65-188	2.66	1.50	0.28

Structural models trimmed to the residues listed were individually superimposed onto the full C-terminal head domain model of BpaC (model creation described in section 5.2). The *super* method was mostly superior to *cealign*, with the exception of YadA. All results using both methods are listed. Trimming the individual models further to a minimal 14-residue repeat (or 15-residue repeat for UspA1) and using the *super* alignment again gave the best results (named *minimal*).

A visual representation of this alignment is provided (**Figure 74**) which shows that all main chain traces are overlapping to the point that they cannot be distinguished easily. Most of the LPBR motifs, even the ones without a structural model available, have this 14-residue repeat pattern seen in YadA. The only exception for this is the 15-residue LPBR repeat within the structural model associated with the LPBR motifs of UspA1 (PDB: 3PR7). An additional G@9 forces the main chain to leave the regular LPBR pattern at position 5-6 of the 15 residue-repeat merging back into the aligned pattern at position 10 (**Figure 74, A**). Trying to align the range of residues containing all visible LPBR motifs in the UspA1 model (66-283, PDB: 3PR7) to the full C-terminal head domain model of BpaC (residues 431-1021) led to an overlap with very little contact points (**Figure 74, C**). It seems however that this 15-residue repeat is the exception. All r.m.s.d. values are within expected variances given that the data quality is different for each model and might reflect normal uncertainty of model building. This helps to put these values into a reasonable framework for comparison.

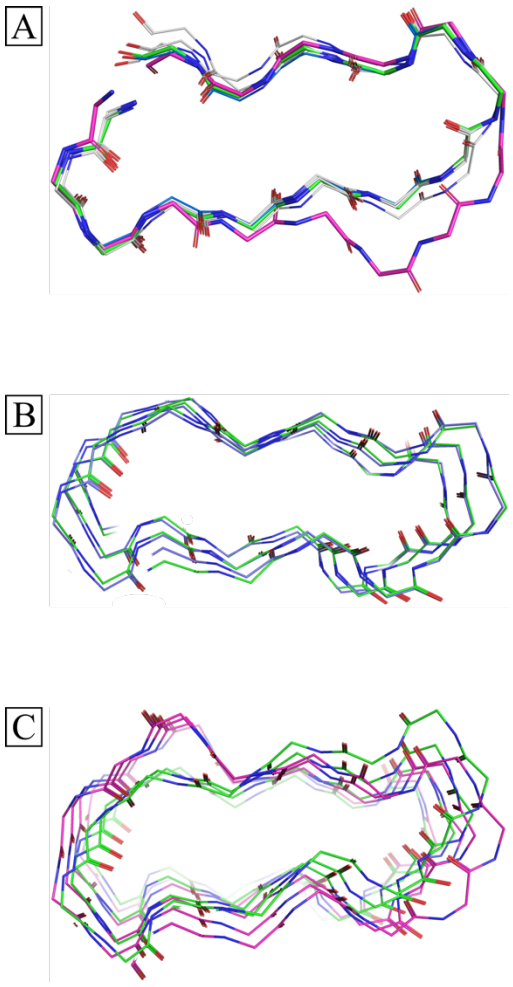


Figure 74 Superposition of a 3-layer LPBR motif of all available LPBR structures – The main chain atoms (Ca) of all available LPBR containing structures (**Table 23** for reference) is shown as stick representation. Only a selected range of residues are shown for better visibility. **A** Superposition of main chain trace of a single layer from all six LPBR containing structures. Main chain carbonyl atoms coloured in red, backbone amides coloured blue. For better visibility, only certain Cα atoms were coloured (BpaC, green; UspA1, magenta) while the remaining were left grey. **B** Superposition of main chain trace of BpaC (green) and BoaA (blue). The latter was used as input model for molecular replacement. **C** Superposition of main chain trace of BpaC (green) and UspA1 (magenta).

5.5.2 Comparison of trimer core motifs

The repetitive nature of the LPBR motifs allows the creation of a frequency plot for the most common 14-residue repeats (and the exceptional 15-residue repeat for UspA1). This frequency plot is a simple illustrative tool to show the relative occurrence of a certain residue for a given position in all LPBR repeats of a selected TAA. Frequency plots were created for all six TAAs which had a PDB structure available that contained LPBR repeats (**Figure 75, A**). This revealed that BpaC has a highly conserved G@1, unlike all the other LPBR containing structures. However, no particular structural consequence could be observed for this observation. Secondly, hydrophilic residues at position 7 are an exclusive feature of the C-terminal head domain of BpaC. Some LPBRs have hydrophilic T@5 but the extensive inner and central solvent network unique in BpaC requires a hydrophilic residue at position 7 (**Figure 75, B**). Lastly, focussing on the charged residues in the solvent-accessible positions of the repeats (1-3, 9-10, 12, 14 in a 14-residue repeat; 1-3, 10-11, 13, 15 in the 15-residue repeat of UspA1) shows that the charged residues in BpaC are exclusively negatively charged (**Figure 75, C**). Even the model that was used for molecular replacement (BoaA, PDB: 3S6L) has some positively charged residues. The ratio of charged residues is hard to gauge from just these frequency plots and will be discussed in detail in section 5.5.6.

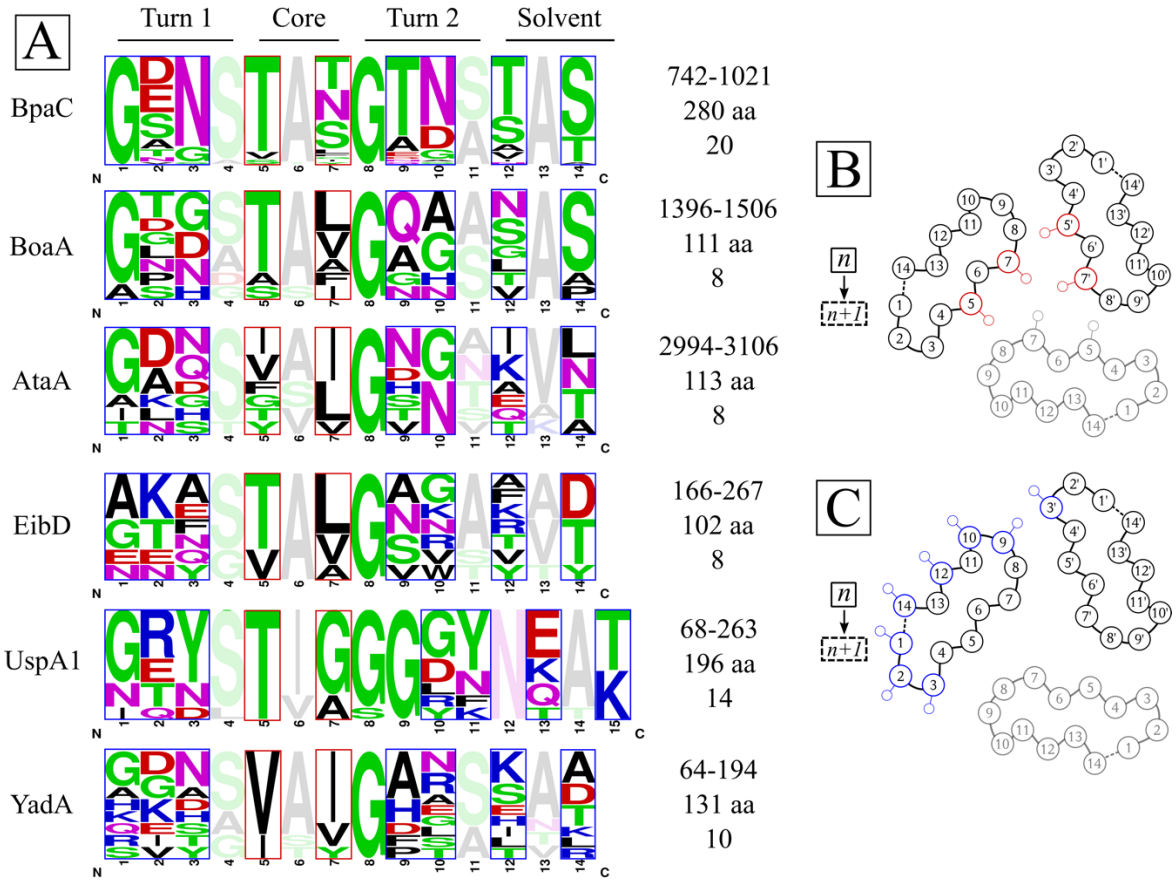

A
Turn 1
Core
Turn 2
Solvent
BpaC
BoaA
AtaA
EibD
UspA1
YadA
742-1021
280 aa
20
1396-1506
111 aa
8
2994-3106
113 aa
8
166-267
102 aa
8
68-263
196 aa
14
64-194
131 aa
10
B
C
n
n+1

Figure 75 Frequency plots of LPBR layers in selected TAAs – The highly repetitive nature of the LPBR motifs allows for a mapping of residue frequency per position within the most common 14-resdiue repeat or the special 15-residue repeat of UspA1. The higher the frequency of an individual amino acid in a repeat position, the larger the letter. **A** Frequency plots are shown with coloured residues indicating the biochemical category they belong to. Residue range used for motif creation in WebLogo (Crooks, Hon et al. 2004) shown (middle column, top), along with the absolute number of residues included in the plot (middle), and lastly the rounded number of layers included in the frequency plot (bottom). Residue ranges are split into four areas that show the relative position of the residue in the trimer arrangement: core facing (red box) and solvent facing (blue box). Residues facing towards the inside of a monomer are made transparent to better highlight the structurally important residues. Note that the G@8 is extremely conserved to the point that the S@8 observed in UspA1 was identified as an error in model building (explained separately). **B** Schematic of a single LPBR layer of a 14-residue repeat. Circles roughly correlate with position of Ca atom in the main chain. Repeats end with residue 14 in layer *n* and continues to residue 1 in the layer below (*n+1*). Position of Cβ atom shown for residues 5 and 7 as these are the core facing residues (red). **C** Same schematic as in **B** but with Cβ atom position shown for solvent-facing residues (blue). Figure used in publication without modifications (Kiessling, Harris et al. 2022).

5.5.3 G@8 is highly conserved in all LPBR motifs

Comparing all the different frequency shows an irregular residue occurrence, namely that the G@8 is highly conserved in all LPBR motifs. By lacking a side chain, glycine can adopt certain angles along its peptide bonds that other amino acids cannot, which can be plotted in a Ramachandran plot. The torsion angles found for all G@8, conserved in all LPBRs, falls into an area which is much more preferred (or at least allowed) by glycine residues at around 130 ° for the φ angle and 140 ° for the ψ angle. This would indicate that these torsion angles are required to maintain the structural integrity of an LPBR layer and essential to keep along the evolutionary track of LPBR motifs (**Figure 76**).

G@1 is highly conserved in BpaC, and also largely present in BoaA at the same position. However, it does not fall into in an area of the Ramachandran plot that would only favour glycine, meaning there is no specific benefit gained with regards to the freedom of torsion angles that can be adopted. Position 1 is also a solvent-facing residue so there wouldn't be any steric hindrance that would prohibit a side chain at this position, however it might have a relevance for a potential ligand binding pocket that requires space in that exact area.

The special case of the 15-residue repeats for the N-terminal head domain of UspA1 (PDB: 3PR7) also affects the torsion angles that are adopted at G@8, falling into a different area on the Ramachandran plot as compared to G@8 in the other LPBR motifs. G@9 is highly conserved in UspA1 and the flexibility and the small real space that glycine takes up is probably needed to compensate for the deviation from the 14-residue LPBR structural pattern (compare with **Figure 74, C**).

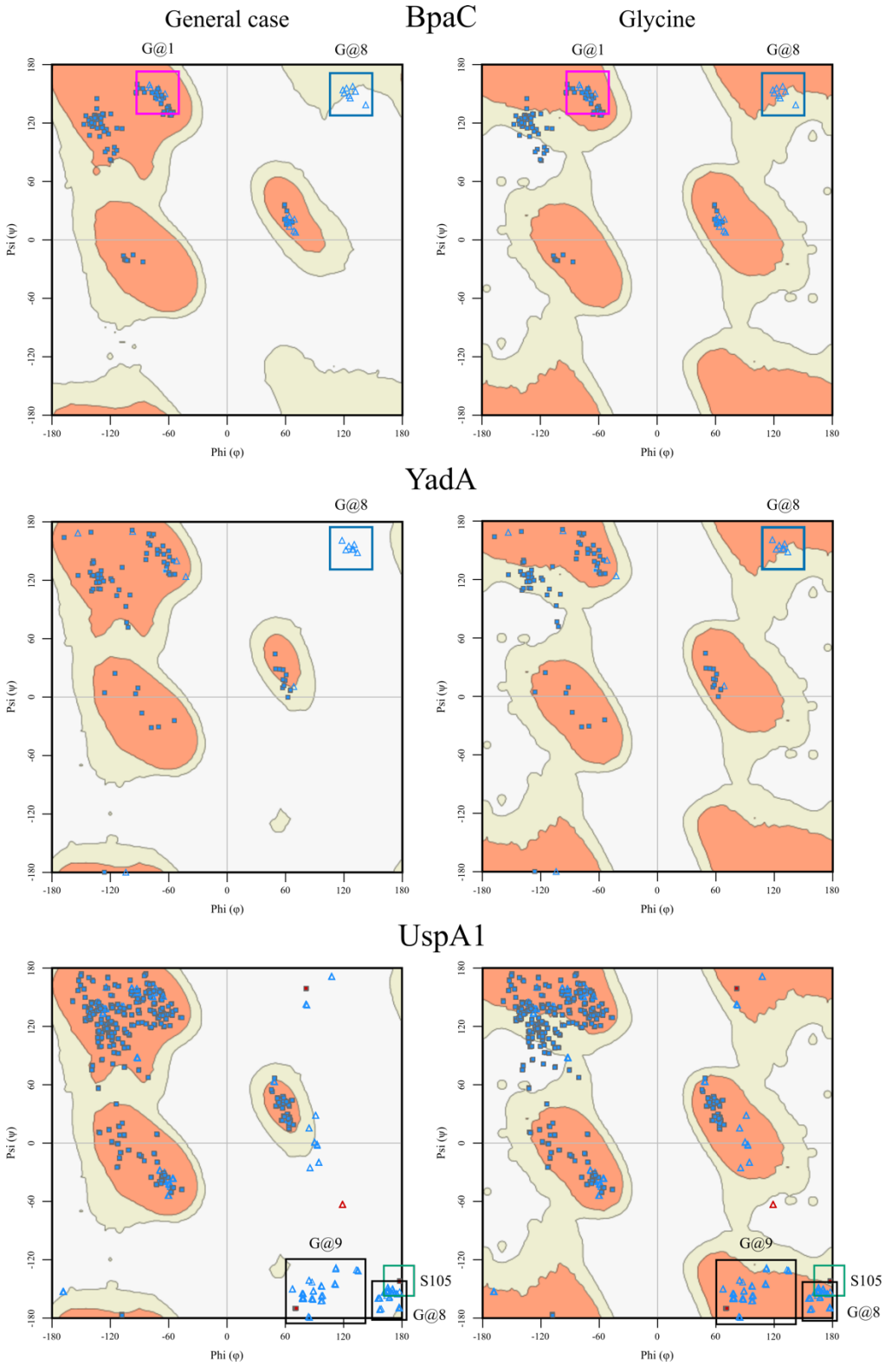

General case
BpaC
Glycine
G@1
G@8
G@1
G@8
Psi (ψ)
Phi (φ)
YadA
G@8
G@8
UspA1
G@9
S105
G@8
G@9
S105
G@8

Figure 76 Ramachandran plot for BpaC and YadA head domain models – Ramachandran plot was created in Coot using the C-terminal head domain model of BpaC (S741Q1054(GCN4)), the N-terminal head domain model of YadA (PDB: 1P9H, residues 68-263), and the N-terminal head domain model of UspA1 (PDB: 3PR7, residues 54-332). The torsion angle around the peptide bond atoms Cα-C (psi, ψ) is plotted against the torsion angle around the peptide bond atoms N-Cα (phi, φ) of each residue (blue squares; glycine: blue triangle). Areas in the plot are separated into three categories as displayed in Coot: favoured (dark orange), allowed (beige), and disallowed. The torsion angles for the highly conserved G@8 within the LPBR motifs of both BpaC and YadA are highlighted (blue box). G@1 are highlighted for BpaC (magenta box) which are conserved only in this structure. Note that G@1 falls into the favoured region for the general case Ramachandran plot, while G@8 is only really favoured for the glycine type plot. G@8 and G@9 in UspA1 was marked separately (black box) as it is part of an unusual 15-residue repeat that also falls into a different area of the Ramachandran plot as G@8 in BpaC and YadA. S105 from UspA1 is highlighted (green box) to show that it falls into a disallowed area. This is discussed in more detail in the following section.

5.5.4 The special case that is UspA1

The N-terminal 15-residue repeat head domain of UspA1 is an interesting exception to the standard 14-residue LPBR repeat found in all the other identified LPBR motifs. The available model (PDB: 3PR7) is poorly refined with a number of strong peaks in the difference map (F_o-F_c) indicating the model is missing several atoms or molecules. A prominent example for this is the almost completely unbuilt outer solvent channel in 3PR7, that I discovered by comparing the structure with the outer solvent channel of BpaC (**Figure 77**). Moreover, the MolProbity report (**Figure 78**) shows several categories that are far off from acceptable geometries. This makes it challenging to understand the exact structural consequence of the 15-residue repeat abnormality and limits the range of possible interpretation. This also makes one question if S105, which would be the only serine at this otherwise extremely conserved G@8 position, is actually what the model builders claim to be. As highlighted in the Ramachandran plots in section 5.5.3, the torsion angles for S105 are highly improbable and taken this further, the CaBLAM analysis performed as part of the MolProbity suite also shows a clear Cα geometry outlier for this residue. Finally, the steric hindrance around this location does not allow for any extensive side chain atom volume. In short, this residue is a likely to have been modelled wrong and the serine should be replaced by a glycine.

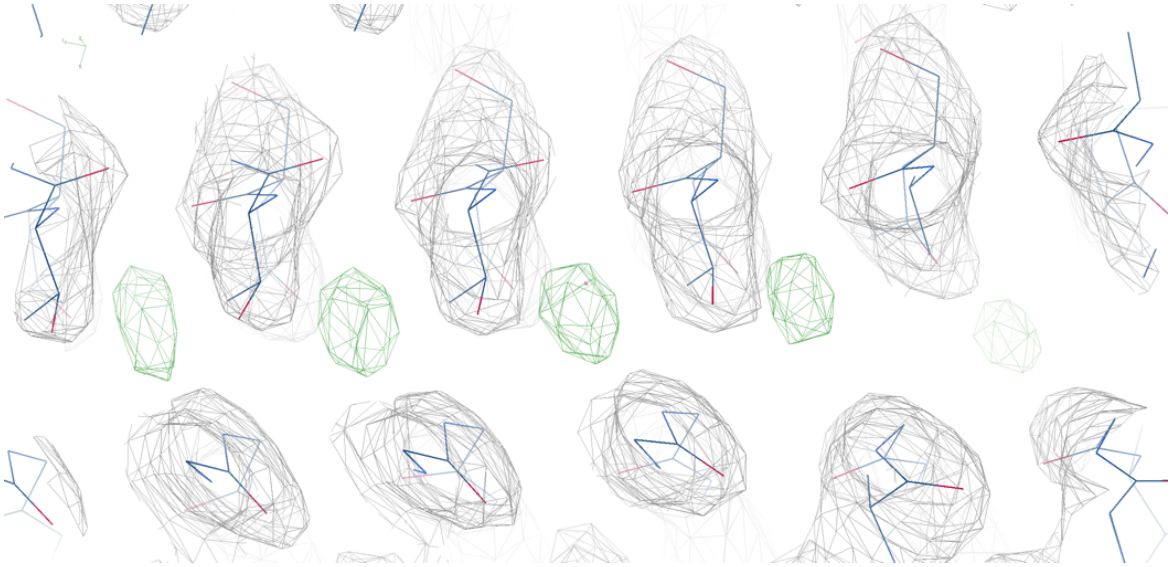


Figure 77 Missing solvent molecules in the structural model of UspA1 – Both the $2F_o$-F_c electron density map (grey mesh) and the difference map (F_o-F_c, green mesh representing positive difference density) are shown overlaid with parts of the structural model of the N-terminal head domain of UspA1 (PDB: 3PR7). The contour level for the $2F_o$-F_c map was set at 1.5 σ while the difference map was contoured at 5 σ. The location of the positive difference density peaks are at a very similar location relative to the protein chain as for the outer solvent channel in all the other LPBR motif structures.

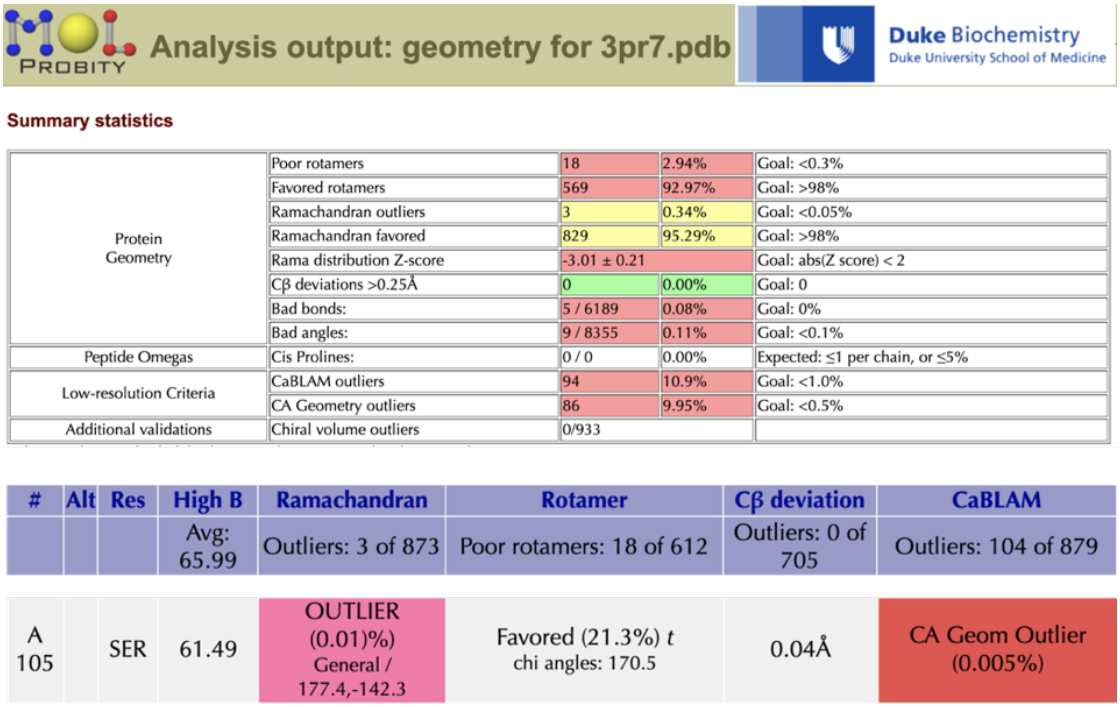


Figure 78 MolProbity result for PDB entry 3PR7 of UspA1 – The MolProbity analysis result for PDB accession code 3PR7, which encompasses the N-terminal head domain of UspA1, is shown alongside a more detailed analysis for residue S105; the only S@8 found in all LPBR motifs. The summary statistic is displayed with absolute numbers (left, usually the amount of the analysed object) and percentage values (middle), with the goal percentage value shown on the right side. The extended analysis of S105 shows the stark deviation from the expected value for both Ramachandran plot (torsion angle expectation) and CaBLAM (secondary structure expectation).

5.5.5 Comparison of solvent channels in LPBR structures

Solvent channels have not been discussed in earlier TAA structures. However, once I identified the three solvent channels in BpaC, I was able to identify the outer solvent channel as a conserved structural feature in all LPBR containing structures (**Figure 79**). Due to some waters not being assigned in the published structures there are either some gaps in the channel in certain layers (AtaA, 3WP8; YadA, 1P9H) or most of the solvent molecules are missing altogether (UspA1, 3PR7) even though there is clear evidence for it in the underlying electron density maps (discussed in section 5.5.4 for UspA1). Importantly, the inner and core solvent channels are unique to BpaC due to the hydrophilic nature of core-facing residues.

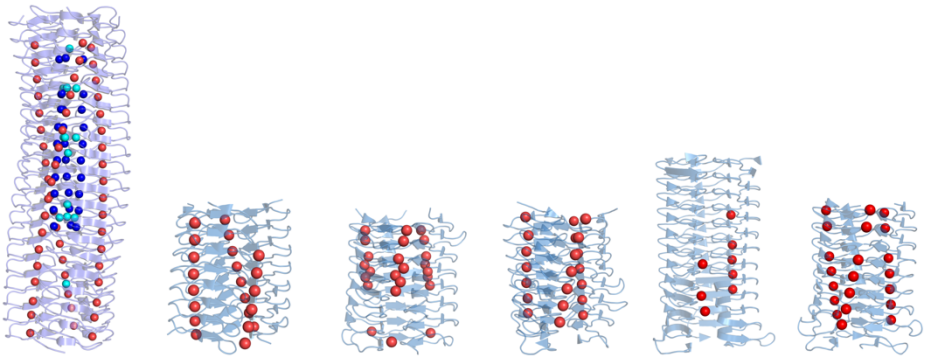


Figure 79 Comparison of solvent channels in selected LPBR structures – Cartoon representation of trimmed protein models are shown with solvent channels shown as spheres: outer, red; inner, blue; central, cyan. No inner and central solvent molecules were observed outside of BpaC. Name of the protein, PDB code, range of residues included in trimmed model, and total length of protein are listed. The model for the LPBR layers in UspA1 is marked as it misses a large amount of the solvent molecules in the outer solvent channel, although there is clear evidence for it in the difference map (F_o-F_c) provided in the PDB (*, **Figure 77**).

5.5.6 Comparison of electrostatic charge surface profiles suggest a new subcategory for LPBR classification

The most important observation I made during the structural analysis of S741Q1054(GCN4) was the identification of a multitude of negatively charged, solvent-facing residues on the outside of the trimer covering the whole of the C-terminal head domain of BpaC. The consequence of these residues are an extremely charged head domain at a standard physiological pH of 7, which is the most accurate pH assumption as this part of BpaC is extracellular and humans are one of the hosts for *B. pseudomallei*. The extreme charge of the domain is also likely the reason why S741Q1054(GCN4) was so soluble (>130 mg/mL) with a buffer pH of 8.

I was wondering if the high number of charged residues in the C-terminal head domain of BpaC is a common occurrence for all LPBR motifs, which started the frequency plot investigation (described in section 5.5.2): it showed not only that the charged residues are often located at the same solvent-facing positions within the LPBR repeat, but also that the amount and ratio of positively-to-negatively charged residues vary dramatically for each selected TAA with an LPBR motif. In an effort to find an explanation for this variation, available information was collected and summarized (**Table 24**). Upon inspection of this data, I noticed that LPBR containing structures are located at either extreme of the passenger domain with a distinct link to the related pI of the head domain. Focussing on the domains located more towards the C-terminal end, one can find negatively charged LPBR head domains with an average pI range from 2.3 to 4.7 and the domain end located between 7 and 14% of the total TAA length (count starts at C-terminus). A different observation can be made for the LPBR domains at the more N-terminal end of the TAA: the average pI ranges from 8.9-9.2 with the domain end location being between 59-70% of the total

TAA length. EibD is a bit of a borderline case as the LPBR domain is in the exact middle of the full length sequence and it has an almost neutral average pI of 6.4.

Table 24 Statistics for LPBR containing structures relevant for subcategory assignment.

	BpaC	**BoaA**	**AtaA**	**EibD**	**UspA1**	**YadA**
PDB	7O23	3S6L	3WP8	2XQH	3PR7	1P9H
Head position	741-1021	1396-1506	2994-3106	124-266	166-267	64-194
Length (aa)	1152	1626	3630	511	863	455
Relative end position from C-terminus	11%	7%	14%	47%	70%	59%
Positively charged	0	0	3	8	17	11
Negatively charged	34	3	6	7	14	10
Positive %	0%	0%	33%	53%	55%	52%
Negative %	100%	100%	67%	47%	45%	48%
Average pI	2.3	3.9	4.7	6.4	8.9	9.2

Best guess head domain position containing LPBR motifs in various TAAs is shown, together with total length of TAA, and the position of the last residue within the head domain relative to its C-terminus (in percentage). Absolute number of solvent-accessible charged residues visible in the available PDB models for each head domain are listed. Note that there are some residues that were excluded as they were facing into the inside of the protein monomer which makes them inaccessible to the solvent and do not contribute to the outside surface charge. Relative percentage of charged residues is given for a better visualisation of subcategory assignment (split indicated by dashed lane). The average pI was calculated in IPC 2.0 using the sequence of the full head domain of the respective TAA (first row).

A visual representation of the charged side chains can be achieved by mapping the electrostatic surface charge volume onto the protein model using the APBS plugin in PyMOL (**Figure 80**). The extended model of the C-terminal head domain of BpaC (T431S1021) was used as part of this visualisation to point out the full biochemical impact of the negative surface charge. The total size of the domain with about 20 nm amplifies the effect of all the negatively charged side chains along the domain, significantly increasing the importance of this domain for the potential function of BpaC.

Together with the statistical observations from **Table 24**, a natural split can be observed between certain models. For historical reasons all of these models, except YadA, would have been called YadA-like head domains. The unique biophysical properties of BpaC and the observed split between N- and C-terminal located LPBR domains sprung the idea of a new subcategory for these head domains: the BpaC-like head domain. In this case, the C-terminal head domain of BoaA and AtaA would be classified as BpaC-like, whereas the N-terminal head domain of EibD and UspA1 would be assigned to the YadA-like subcategory. The reason for this division is still unclear because the binding partners of many TAAs, such as BpaC, are still not known.

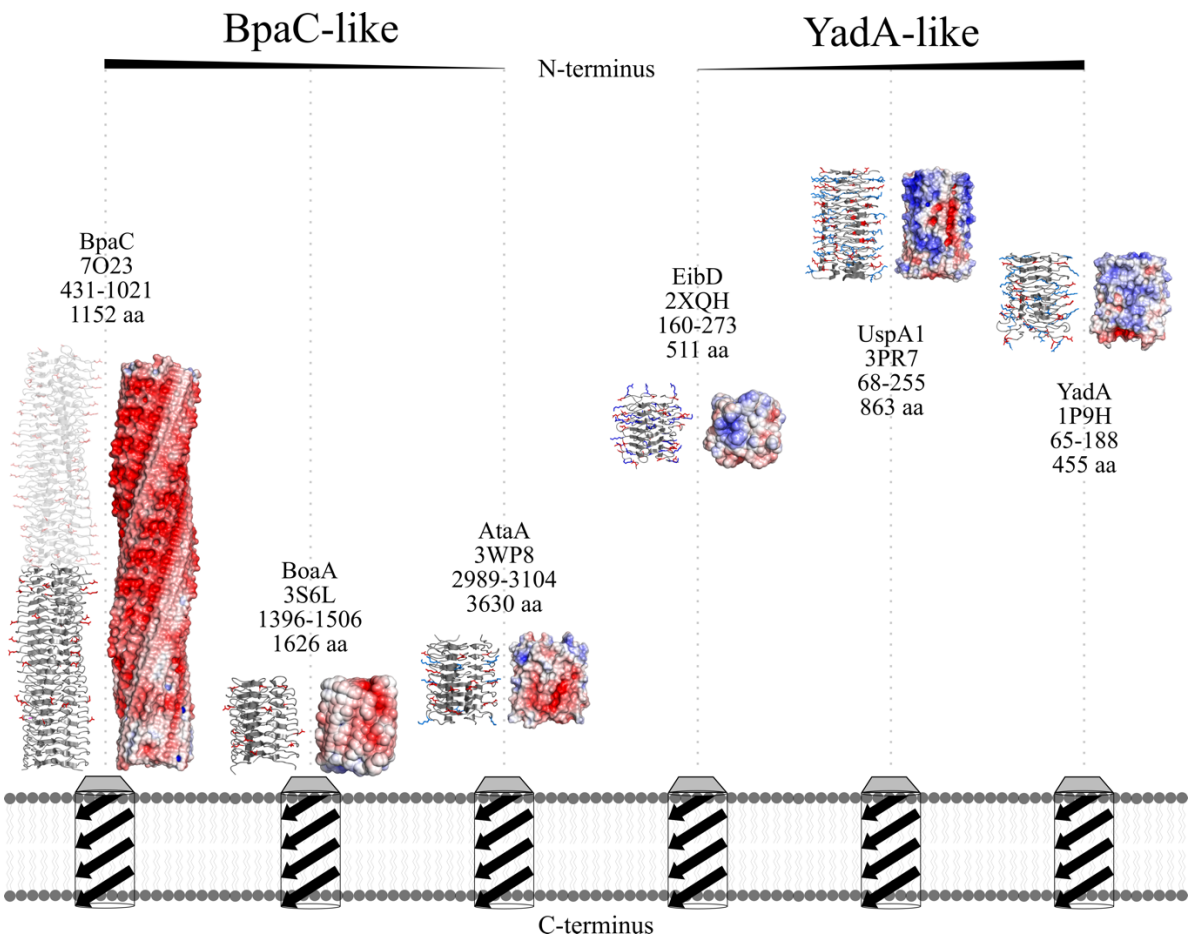


Figure 80 Comparison of LPBR models by electrostatic surface charge representation – Schematic showing all the different LPBR motif containing models both in cartoon representation and electrostatic surface charge representation (APBS plug in PyMOL). Cartoon representation is shown with charged residues in stick representation and coloured either as negatively charged (red) or positively charged (blue) at pH 7. Colour scheme for the electrostatic surface charge representation follows the same principle. Relative position of the models along the full protein sequence is indicated but the actual position will differ and only serves illustrative purposes (more accurate description in **Table 24**). Name, PDB code, residue range of displayed model, and total length of TAA are displayed for each model. Head domains are assigned either to be more BpaC-like or YadA-like depending on surface charge or relative position of the domain along the protein chain.

5.5.7 RoseTTAFold structure prediction of selected LPBR domains verifies the subcategory hypothesis

I have used structural prediction methods (AlphaFold (Jumper, Evans et al. 2021) or RoseTTAFold (Baek, DiMaio et al. 2021)) to test my hypothesis about the correlation between domain position and charge. I searched for additional LPBR containing head domains by carefully analysing all TAA sequences with any associated structural models for repetitive elements and was able to identify two LPBR motifs in AtaA and BoaA with no structural data available for them. The G@8 alignment method was a reliable approach to at least narrow down the range of residues only containing the LPBR layers for input for RoseTTAFold. Each residue in the generated model was assigned a local Cα error estimate (in Å) that reflects the accuracy of the prediction. A cut-off of larger than 2 Å was set for the removal of inaccurate residue predictions. A single LPBR layer of each of the two models was structurally aligned to a single LPBR layer of the C-terminal head domain of BpaC using the *super* command in PyMOL (described before in section 5.5.1). The resulting r.m.s.d. for the AtaA to BpaC alignment was 0.53 Å and for BoaA to BpaC it was 0.50 Å. Recalling the r.m.s.d values for the structural comparisons of the "real" structures with BpaC, starting at 0.22 Å and the highest at 0.57 Å, shows that these deviations are well within the acceptable range considering these are models generated by a structural prediction software without any additional experimental data fed into it.

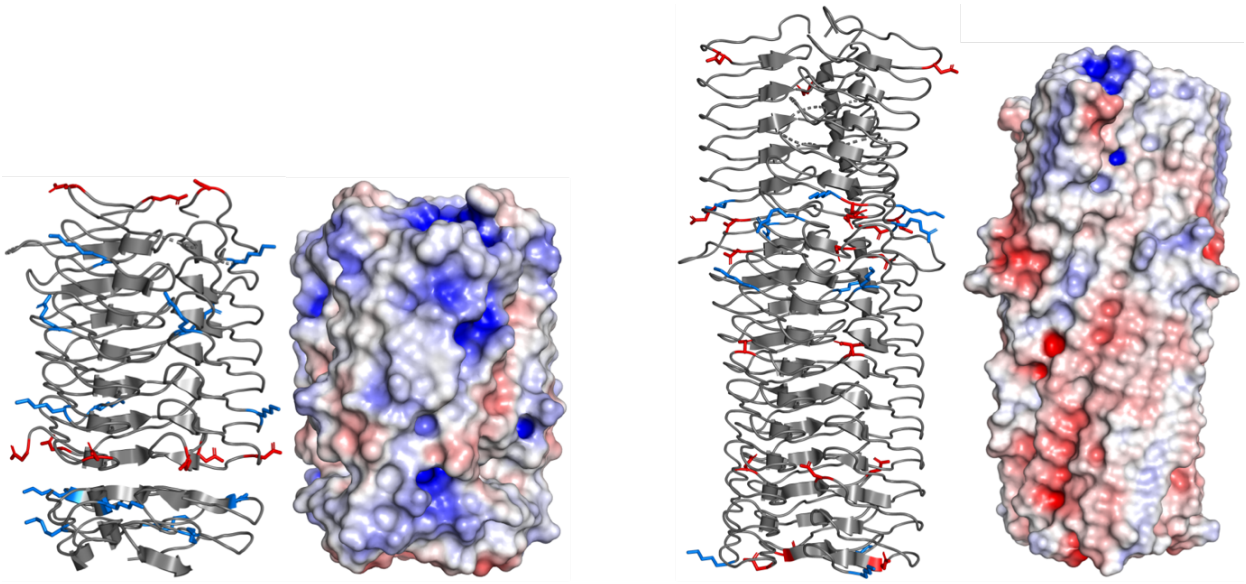
The expansion to trimeric models was possible by copying the monomeric models and aligning them to the trimeric model of S741Q1054(GCN4) in PyMOL. The model was trimmed to match the exact number of layers visible in the other two models respectively. That means the model of S741S1021 of BpaC was trimmed to ten layers for AtaA alignment and to 17 layers for BoaA

alignment. I then mapped the electrostatic surface charge and highlighted the charged residues in stick representation (**Figure 81**).

Analysing the electrostatic surface charge representation reveals a lot of positively charged patches for AtaA, as expected. This is also reflected in the average pI of about 8.4 as estimated by IPC 2.0. The end point of this head domain is located at a relative position of 93% of the complete protein chain when counting from the C-terminal end backwards. For comparison, the closest values from the previous analysis in section 5.5.6 are from the UspA1 model, with an average pI of 8.9 and a relative domain end point of 92%.

The predicted model for BoaA however is the first case that might prevent the hypothesis from becoming a generally accepted fact: with the location of the N-terminal domain at a relative end position of 73% along the complete protein chain, one would expect the pI to be at least above 7, making it a positively charged domain at a pH of 7. The visual representation of the electrostatic surface charge and the calculated pI of about 4.1 indicate the opposite. A possible explanation for this is that there are seven aspartates per monomer that contribute to the pI calculation but only two of them are solvent-facing, whilst the other five are either facing towards another monomer or facing inwards into a monomer chain shielding the sidechain charge from solvent access. This means that the actual contribution of charged residues to the pI calculation would stem from four positively charged residues and two negatively charged residues resulting in a net plus of positively charged residues. This would elevate the pI to above 7, thus restoring the original hypothesis. Clearly a structural model based on experimental data is needed to verify this assumption. The consequence of this outlier observation is that one cannot accurately predict the actual charge of an LPBR containing head domain just by knowing its relative location along the TAA chain. Further information is needed to solidify that claim. This shows yet again the importance of

experimentally validated structural models and the superiority of these well-researched methods over models generated by algorithms, sophisticated as they may be. Nevertheless, the connection between head domain position and surface charge is established and will be a good starting point for further research into the functional consequence of this observation.



AtaA	BoaA
K7ZP88	Q3JFX2
110-265	180-437
3630 aa	1626 aa
Δ120-124	Δ224-230
Δ219-234	Δ244-248
8.4	4.1*

Figure 81 Structural models of selected LPBR motifs created by RoseTTAFold – The RoseTTAFold structure prediction method (Baek, DiMaio et al. 2021) was tested on two predicted LPBR containing head domains. These are separate from the head domains previously described in AtaA and BoaA as they are on the N-terminal end of the protein. Both N-terminal head domains do not have any experimentally generated structural model associated with them so far. Structural models are shown as cartoon representation with charged residues shown in stick representation (red, negative; blue, positive) and the electrostatic surface charge volume created by the APBS plugin in PyMOL (same colouring scheme). Name of TAA, associated UniProt code, residue range of selected LPBR motifs, and length of TAA are listed for each model. The models generated by RoseTTAFold were filtered for a local error estimate of below 2 Å, meaning that residues that had a higher error associated in the final model were removed. The range of residues that were excluded from the final model are listed (Δ). Lastly the average pI was calculated in IPC 2.0. Note that some of the charged residues in the model of BoaA actually face into the inside of the monomer likely affecting the real value of the pI of this domain (*).

6 Chapter 6: Conclusions, Discussion, and Future Perspective

6.1 Discussion of key findings from this thesis

6.1.1 Analysis of *bpaC* to identify domain borders and structural motifs

A sequence annotation of *bpaC* was first performed by Lafontaine et al. with the description of different sequence motifs and some rudimentary secondary structure recognition (Lafontaine, Balder et al. 2014). While this provided a good starting point to aim for a rough domain association, it was incomplete and also provided no real information about how many structurally diverse domains there are. Building on this work, I conducted a full sequence analysis considering the unique trimeric nature of TAAs, the already extensive domain dictionary and structural motif rules, and the most recent bioinformatic tools like DeepCoil and PSIPRED (with advanced additions). The creation of this TAA specific sequence analysis approach was crucial in providing more accurate predictions that were the basis of all the construct designs in this thesis. Purely based on this sequence analysis, a total of four functional domains could be identified in the passenger domain of BpaC: an N-terminal head domain that has structural motifs that relates it to the head domain of BadA, two separate stalk domains of which one had slightly higher aggregation potential than the other, and a very large C-terminal head domain consisting of 42 LPBR layers that almost makes up half of the sequence of the protein. The identification of the neck motifs (DAVNxxQL) provided a very accurate manipulation point for either removing individual domains or fusing them together in order to increase their solubility. They also enabled the precise definition of domain borders which was the original goal of this analysis. To ensure this analysis

reached a wide audience I also requested an update to the UniProt entry of BpaC (A0A0H3HIJ5) at the start of my project and on the 10th of April 2019 it got a much more curated entry with "reviewed" status, making it more accessible to the wider research community.

6.1.2 Identification of a pathogenicity island surrounding *bpaC*

The genetic context of a POI can give a plethora of information on the potential role this protein plays in the wider cellular context. The well-curated genome database for *Burkholderia* species (Winsor, Khaira et al. 2008) provided a lot of information on adjacent genes to *bpaC* (*B. pseudomallei* strain 1026b, identifier BP1026B_I1575) and even suggested that it is part of a wider pathogenicity island, as defined by IslandViewer 4 (Bertelli, Laird et al. 2017). It revealed an expression system that has striking similarities to a two-component response regulator system described for TAAs in *B. cencocepacia* by Pimenta et al. (Pimenta, Mil-Homens et al. 2020): here, the authors describe the histidine kinase BCAM0218 that phosphorylates a response regulator after sensing an external stimulus. This in turn changes the DNA binding affinity of the regulator leading to a decrease in TAA expression. This two-component system can also be found next to *bpaC*: the response regulator (BP1026B_I1577) and the sensor histidine kinase (BP1026B_I1578). This expression system is likely to impact not only the expression of *bpaC* but also the more upstream chaperone/usher fimbrial adhesion system. However, one can only speculate on the exact nature of how this expression is regulated and what external stimulus is required for a change of expression.

The upstream gene cluster of the chaperone/usher fimbrial adhesion system provides crucial information to the overall role of BpaC in the pathogenicity of *B. pseudomallei*. This gene cluster contains type I pili that are typically involved in the formation of microcolonies, biofilms, and the receptor-mediated adhesion to host cells (Thanassi, Bliska et al. 2012). They are mannose-sensitive

bacterial surface fibers that have their adhesive subunit on the tip of the fiber which bind primarily to glycan receptors of the host cells (Zav'yalov, Zavialov et al. 2010). Their size (submicron to about 2 micron) means that they are larger than most TAAs (Forero, Yakovenko et al. 2006), while BpaC is expected to be a lower medium sized TAA with about 55 nm. This supports my hypothesis of BpaC helping to form more short-ranged connections, mainly with extracellular matrix components with the host cell that tighten the initial connections which was formed with the more specific but longer range glycan-pili interactions. The implications of this connection is further discussed in the context of the overall role of BpaC later on in this chapter.

6.1.3 Preparation of *bpaC* for cloning experiments

No vector containing *bpaC* was available on Addgene and the extraction of gDNA from *B. pseudomallei* (a BSL-3 strain) could simply not be performed at the University of Leeds or in a reasonable time frame from other institutions. In the end, the complete gene was ordered as a synthesised product from General Biosystems, Inc. (Durham, USA). This had the major advantage that the sequence could be manipulated to reduce redundancy, which would have been very cumbersome with classic PCR mutagenesis approaches. The highly repetitive nature of the C-terminal head domain of *bpaC* resulted in multiple stretches of identical nucleotides, leaving areas of the sequence inaccessible for cloning experiments. Available optimisation tools account for the codon usage of a particular gene for expression in the specified host organism. This is how I came up with the idea of "deoptimisation": essentially introducing silent mutations all over the repeat sequences but going back and forth by designing primers in SnapGene and seeing how much these mutations improved the specificity of these primers. This process resulted in a much more accessible sequence that enabled the precise annealing of primers at the repeat segment of choice. This ultimately allowed the creation of soluble domain constructs with variable numbers of repeats

or the deletion of segments in order to close the gap between two rather distant repeat elements as was needed for the creation of pET28a-V75(C76S)(C97S)(ΔCC12)(ΔS447G826)Q1054(GCN4), which combined the N-terminal end of the C-terminal head domain with the C-terminal end of the head domain.

6.1.4 Creation of deletion mutants and soluble domain constructs

Experiments with BpaC in the literature were all restricted to the full-length protein or knockout mutants. With the full sequence analysis performed during this project came the ability to create deletion mutants that had a much higher chance of folding correctly than their randomly selected counterparts by using the neck motifs as connection and transition point for the deletion. These deletion mutants were verified by using the SpyTag/SpyCatcher assay and the periplasmic stress assay to confirm the successful translocation of all created mutants for the full-length experiments. Using the extended sequence analysis, I created a range of cytosolically-expressed constructs for each domain, most of which needed optimisation before being able to use them properly for either biochemical experiments or crystallisation attempts: the simplest change was including or removing certain residues at the start or the end of the construct, especially for the constructs on the far N-terminal end of the protein. This was mainly because, while the cleavage site for the extended signal peptide was known, the extend of the unstructured linker element between the signal peptide and the N-terminal head domain was unclear. The next step up in the optimisation effort was the fusion of constructs with the GCN4 adaptor at the neck motif transition, which ultimately led to the structural solution of S741Q1054(GCN4). This fusion for the N-terminal constructs and the stalk domain constructs also had some benefits, but ultimately did not yield any protein crystals. Using the C-terminal head domain as solubility enhancer proved to be essential to obtain working constructs for both N-terminal head domain and stalk domain. A side effect of

this fusion was the need to optimise the final protein buffer to compensate for the extreme charge of the C-terminal head domain, preventing the concentration of the protein sample to the high concentrations desired in crystallisation attempts. In summary, for each domain there exists a membrane-bound deletion mutant and at least one well-purifiable construct either by native IMAC or denaturation/refolding approaches (**Figure 58** in section 4.12). This provides a complete and extensive toolset for functional and structural studies on BpaC.

6.1.5 Identification of extracellular matrix proteins that bind to BpaC

In vivo assays performed in 2015 provided the first evidence of the involvement of BpaC in several pathogenesis-associated phenotypes of *B. pseudomallei* (Lazar Adler, Stevens et al. 2015). This phenotypic association is limiting as it does not provide a specific subset of proteins that could be implicated in the related processes, leaving the potential targets plentiful. Looking at the most likely binding partners in other TAAs, one can identify protein classes in humans that keep reoccurring throughout the literature: extracellular matrix proteins (summarised in (Vaca, Thibau et al. 2020)), parts of the complement system like C4b-binding protein (several examples listed in (Hovingh, van den Broek et al. 2016)), or specific receptors like UspA1 binding to CEACAM1 (Conners, Hill et al. 2008). The easiest class to work with was the extracellular matrix proteins as they were commercially available in ready-to-use kits (purified and coated onto wells), albeit these kits were originally designed for the use with human cell lines. The adaptation of this commercial kit to *E. coli* allowed the identification of three extracellular matrix proteins that preferentially bound to BpaC: Collagen II, fibronectin, and to a lesser degree collagen I. While the binding effect was much more pronounced for collagen II and fibronectin throughout all deletion mutants, the deletion mutants ΔCC12 (stalk domain) and ΔCC3 (C-terminal head domain) had a reduced binding effect to collagen I that was below the significance threshold of the applied statistical test.

Deletion of individual domains did not reduce the binding to collagen II and fibronectin, suggesting that there is either a cooperative effect of different domains or the binding is a result of general adhesiveness of BpaC to these proteins. This shows the limitations of this particular assay for determining the exact domain that is responsible for binding. Nevertheless, this is the first evidence that BpaC binds to the extracellular matrix in humans and more interestingly to specific proteins that have been reported before in the context of TAA binding.

From a structural point of view this provides an interesting connection point for all three ECM proteins: the N-terminal head domain of YadA binds to both collagen I and II (Leo, Elovaara et al. 2008) which is an LPBR motif type domain just like the C-terminal head domain of BpaC. The deletion of the C-terminal head domain (ΔCC3) reduced the binding propensity to collagen I and II in the assays performed during this project, connecting these two structural motifs and their shared binding partners. Structural motif identification revealed the similarity between the N-terminal head domain of BpaC and the head domain of BadA (both containing a TrpRing motif, a GIN motif and similar neck motif). Binding experiments with deletion mutants of BadA showed a cooperative binding effect of the head and the stalk to fibronectin (Kaiser, Linke et al. 2012) which fits very well with the observations made in this thesis for BpaC. Both of these points support the hypothesis that structural conservation between TAA motifs can predict functional binding.

6.1.6 Towards a complete structural model of the passenger domain of BpaC

The structural information that was present at the start of this project was confined to a rudimentary sequence motif analysis performed by Lafontaine et al. in 2014 which was not enough to predict any structural homology models (Lafontaine, Balder et al. 2014). The extensive sequence analysis performed on *bpaC* in this thesis revealed several structural motifs that could be used to more accurately predict a possible homolgy model for a certain area than a standard sequence alignment to a published structure could provide. Specifically looking at the sequence of known structural motifs, their secondary structure preference as determined by PSIPRED and DeepCoil, and a BLAST search against the PDB database, resulted in four homology models that together cover most of the passenger domain of BpaC (**Figure 82**): the N-terminal head domain of BpaC is likely to resemble the N-terminal head domain of BadA (PDB: 3D9X, (Szczesny, Linke et al. 2008)) as both contain a TrpRing motif, followed by a GIN motif, and finally a neck motif leading to a coiled coil sequence. The BLAST search against the PDB, together with secondary structure analysis, revealed that parts of the stalk I subdomain resembles parts of the AtaA stalk domain (residues 3418-3465, PDB: 3WPO, (Koiwai, Hartmann et al. 2016)). Combining the results of the coiled coil prediction (DeepCoil) with the discovery of an FxG motif (residues 361-3, LGG) and DAVNxxQL neck motif in the stalk II subdomain results in a high confidence for the relation of this segment to the SadA stalk model (PDB: 2YO2, (Hartmann, Grin et al. 2012)) containing all the previously mentioned elements in the exact same order. The homology model resembling the C-terminal head domain of BpaC (PDB: 3S6L, (Edwards, Gardberg et al. 2011)) was used as molecular replacement model for S741Q1054(GCN4). The structural model of S741Q1054(GCN4) supersedes the homology model, which is why it is not shown in the illustration. The extension of this model to the N-terminal end of the domain results in a structural

model that covers about 63% of the sequence of the passenger domain. Taken together with the homology models this provides an extensive structural insight into the different domains of BpaC with the exception of parts of the stalk domain. This is due to the fact that a variety of different coiled coil structures exist in TAAs and the exact length of the individual elements can vary depending on unstructured breaks in the sequence making it hard to compromise on a specific published model for homology.

The advent of next-generation structure prediction, namely AlphaFold (Jumper, Evans et al. 2021) and RoseTTAFold (Baek, DiMaio et al. 2021), is particularly interesting for TAA structure prediction as the specific rules seem to be picked up quite well by the algorithms. As a case in point, the sequence of S741S1021 of BpaC, which covers the LPBR repeats within the structural model of S741Q1054(GCN4), was submitted to RoseTTAFold for structural prediction. The resulting structural model was aligned to the trimmed version of BpaC (S741S1021), which resulted in a low r.m.s.d. for individual LPBR layers but a strong deviation for the overall alignment (**Figure 83**): The alignment very clearly shows that the characteristic superhelical twist present in the experimentally deduced structural model is not present in the RoseTTAFold generated model. Only comparing the individual LPBR layers in both models one obtains an r.m.s.d. of less than 1 Å, a similar value to the ones observed when I compared the LPBR layers of different TAA models. This example shows the power of the structure prediction algorithms of the current generation but also their limitations: the folding of the individual segments may be accurate but ignoring the trimeric nature of the protein leads to an error propagation that increases by the size of the predicted model.

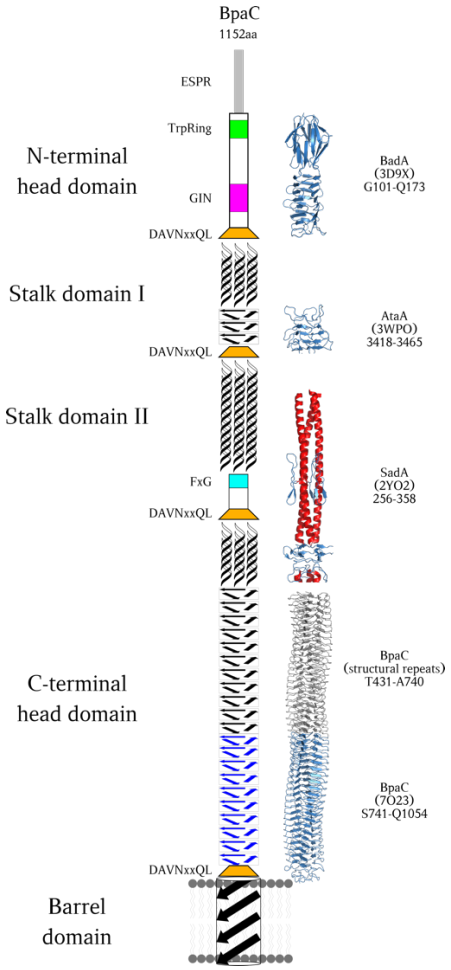


Figure 82 Mapping of structural models onto passenger domain of BpaC – Cartoon representation of structural models from either related TAAs (structural motif homology) or based on the solved structural model of S741Q1054(GCN4) with extended C-terminal head domain repeat (T431A740, grey). Best-guess homology models are deducted from the observance of a specific sequence of structural motifs in conjunction with secondary structure preference for the area. For example, the N-terminal head domain of BpaC is likely similar to the N-terminal head domain of BadA (PDB: 3D9X) due to the occurrence of three consecutive structural motifs (TrpRing-GIN-neck motif) found in both TAAs. Structural motifs in BpaC are indicated left of the schematic model of BpaC (TrpRing, green; GIN, magenta; DAVNxxQL neck motif, orange; FxG, cyan) alongside the extended signal peptide region (ESPR) at the N-terminal end of the protein which is cleaved off during translocation. Schematic of BpaC shows secondary structure preference, except for the N-terminal head domain, with a rough distinction between coiled coil heavy areas and β-sheet rich areas.

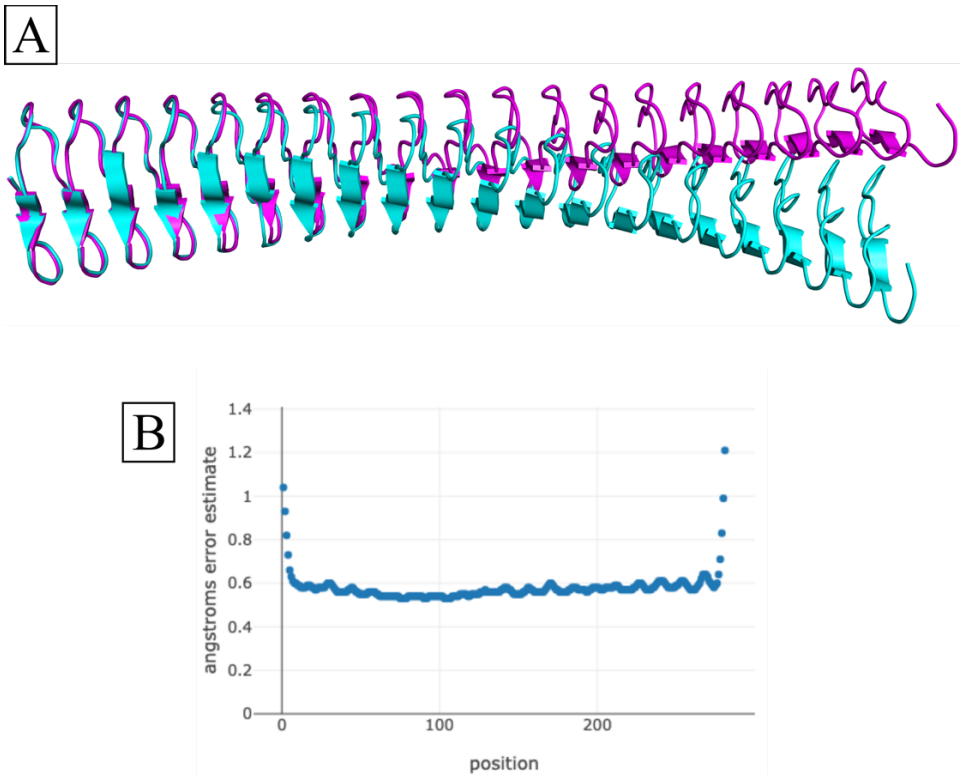


Figure 83 Structural alignment of S741S1021 of BpaC and RoseTTAFold model equivalent – **A** Cartoon representation of experimentally deduced structural model of S741S1021 of BpaC (cyan) and computationally generated model by RoseTTAFold (magenta). Alignment was performed in PyMOL using the *super* command on the last 42 residues of each model (C-terminal end on the left). **B** Plot showing local Cα error estimate (in Å) per residue for the model generated by RoseTTAFold.

The published AlphaFold model for BpaC (Identifier: AF-A0A0H3HIJ5-F1, accessed on 13/06/2022) has some parts that reflect the secondary structure prediction that was performed in the sequence analysis but is also far off the true trimeric nature of a TAA (**Figure 84**). This is especially true for the N-terminal head domain and parts of the stalk domain with a large degree of unfolded areas. The N-terminal head domain, which is predicted to resemble the head domain

of BadA (Szczesny, Linke et al. 2008), is completely unstructured in the AlphaFold model. The trimeric dependency of the fold recognition is likely the biggest source of error in this prediction.

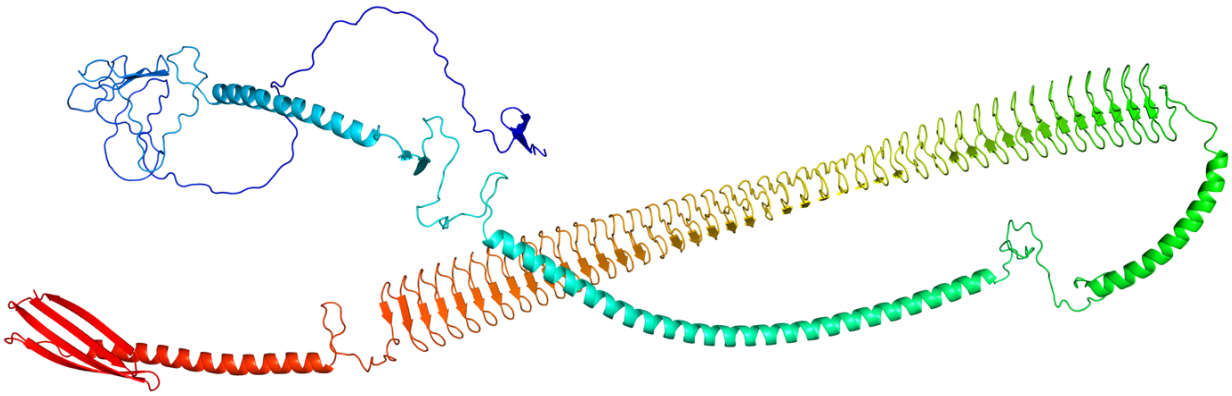


Figure 84 Structural model of BpaC generated by AlphaFold – Cartoon representation of structural model of BpaC generated by AlphaFold coloured in classic spectrum colours (N- to C-terminal, blue to red) with identifier AF-A0A0H3HIJ5-F1 as deposited in UniProt.

6.1.7 Extending the definition of LPBR motifs by introducing a new subcategory

When I first solved the structure of S741Q1054(GCN4), it initially seemed that it was similar to other YadA-like head domain structures with little new insight into the structural landscape of TAAs. Looking solely at the Cα trace of both this model and the model of covering the head domain of YadA (PDB: 1P9H) one could indeed believe that these are just copies of each other. However, when comparing the sequence, the individual residues per position within the 14-residue repeat, the structured solvents within the trimer molecule of S741Q1054(GCN4), and most importantly the surface charge on the outside of the trimer molecule, it quickly became apparent that this is a complete new subclass of LPBR motif containing domains.

Comparing all published structures that contain LPBR motifs, one can identify a trend that streches from the C-terminal head domain of BpaC to the N-terminal head domain of YadA: the relative

positon of the head domain along the sequence of the passenger domain correlates with the surface charge propensity of the domain, expressed by the pI of the sequence, in an almost linear fashion. That is the further C-terminal an LPBR motif containing domain is, the lower is the average pI of the domain. These kind of head domains can then be split into either more BpaC-like or YadA-like, although there are some borderline cases like the head domain of EibD, which sits in the middle of the passenger domain and has an almost neutral pI of 6.4. In hindsight, rather than defining two subcategories with specific (subjective) thresholds, a more appropriate description would be that of a spectrum starting from BpaC all the way to YadA. The identification and structural elucidation of more head domains with LPBR motifs will help to obtain more data points for the correlation between domain location and pI, increasing the confidence of the spectrum hypothesis.

6.2 Potential role of BpaC in the pathogenic cycle of *Burkholderia pseudomallei*

In 2020 and 2021, three publications provided evidence that glycosylation of the target protein is an important component of TAA binding. This has been massively overlooked by the community and is likely playing a much bigger role in the infection process than previously thought.

In the first set of publications, Pimenta et al. showed that mRNA levels for a set of TAAs in *Burkholderia (B.) cenocepacia* are upregulated upon contact of the pathogen with certain human bronchial epithelial cell lines. Especially interesting was the apparent depencency of this increase of mRNA levels on the presence of *O*-linked glycan structures (Pimenta, Mil-Homens et al. 2020). This publication alone however does not provide any evidence of a direct binding event between TAAs and glycan structures and mainly serves as a pointer to where future research should be taken.

Next, Pimenta et al. raised antibodies against an N-terminal stretch of the TAA BCAM2418 from *B. cenocepacia*. This part of the protein contains YadA-like head repeats that can be identified by the conserved G@8 alignment approach. Upon addition of the antibody, a reduction of adhesion of *B. cenocepacia* to the host was demonstrated. A specific set of glycan structures were identified as binding partners which hints at a cell-specific recognition mechanism (Pimenta, Kilcoyne et al. 2021). It is unclear which residues actually contribute to this binding profile and what the role of the overall surface charge to this interaction is. This would be of particular interest to see if the C-terminal BpaC-like head domains also are able to bind glycosylated structures, the importance of the glycan-sensitive pili in this context, and if the shared specificity reflects the intended host target cell environment during *B. pseudomallei* invasion.

Another more direct example is given by Tram et al., who looked at *Acinetobacter baumannii* Ata (UniProt: A3M3H0; not to be mistaken with AtaA, UniProt: K7ZP88) binding to glycan arrays and found specific glycan interaction preferences for both N-terminal head domain and full length TAA (Tram, Poole et al. 2021). To my knowledge, this is the first systematic investigation of the interaction of TAAs and glycan structures. They also estimated K_d values for both a purified head domain construct and the membrane-bound full construct (in *E. coli*) with differing values, mostly in the nM range. They then demonstrated binding of the head construct to fibronectin which required glycosylation for binding to the TAA. In the ECM binding studies performed in this thesis, all proteins were from human origin, meaning that fibronectin in particular existed in glycosylated form. While the control experiment for this ECM study is missing, glycosylation of fibronectin may also be important for the binding recognition by BpaC.

Piece-by-piece we are getting closer to a more detailed model of how BpaC might interact with the host cell. A final aspect to be considered comes from a biophysical angle that might explain the relevance of pili in the context of TAA expression a bit better: the initial attachment of the pathogen to the host usually occurs under flow conditions as the pathogen is circulating through the host organism. In type I pili from *E. coli* a so-called catch bond binding mode was described for the mannose-binding FimH adhesin that reinforces binding under mechanical stress. This way the pathogen can explore multiple binding sites during low flow conditions while retaining strong binding during high sheer forces when latched onto the correct target (Mathelie-Guinlet, Viela et al. 2021). A similar binding mode also has been described in TAAs like BCAM0224, which binds to collagen forming heterophilic bonds (El-Kirat-Chatel, Mil-Homens et al. 2013). This low-affinity binding also behaves like a force spring with a large binding strength under mechanical stress. Both of these molecule classes allow the pathogen to scan the host surface for binding sites

under low flow conditions while engaging in high strength attachment once engaged with target receptors/ECM proteins.

In general polyadhesive fibers, like TAAs and pili, provide a polyvalent surface for attachment of the bacterial cell to the host (Galván, Chen et al. 2007). This pulls the bacterium closer to the host cell in a zipper-like formation (Leo, Lyskowski et al. 2011) which in turn leads to the aggregation of host-cell receptors and trigger immunosuppressive and proinflammatory responses (Schmid, Grassl et al. 2004).

Connecting all these points together, the following model can be suggested (**Figure 85**): in **Stage I** of the attachment of *B. pseudomallei* to the host environment (usually epithelial cells) the long pili fibers "scan" surfaces for specific binding partners, comparable to *E. coli* FimH binding to a mannosylated receptor protein. Once a highly-specific binding partner is found, the pili latches the pathogen onto the host surface in a still fairly loose fashion.

This binding event triggers the expression of the pathogenicity island surrounding the pili, which includes *bpaC*, in **Stage II** of the attachment process. I have termed this expression module a pathogenicity cassette as one can imagine different pathogenicity islands like this present in the genome of *B. pseudomallei* that all have different surface targets and specifcities and are "loaded" like a module in a software program depending on the environment the pathogen is in.

In the final **Stage III** of host-pathogen attachment, BpaC engages with various members of the extracellular matrix, predominantly fibronectin and collagen II, and forms strong multivalent interactions.

After these three intial stages, internalisation of the pathogen occurs as described in the introduction. The pH during the phagocytosis step drops dramatically to a pH of about 5.5 in the late phagosome (Uribe-Querol and Rosales 2017), yet *B. pseudomallei* survives this internalisation

process and is released into the host cell (Galyov, Brett et al. 2010). A question that I asked myself in this context was the role that BpaC plays in this late stage of invasion. As we have seen, the C-terminal head domain of BpaC is highly negatively charged with a pI of about 2.3. The relative charge of the domain would decrease upon entering a more acidic environment. This could hypothetically alleviate any charge-based interactions that BpaC has with the host cell, releasing the pathogen from the phagosomal surface. The role of pH-dependent deglycosylation could also be of interest in this context, as glycosylation of fibronectin was found essential for the binding of the TAA Ata (Tram, Poole et al. 2021).

Overall BpaC is involved in both initial cell attachment and immune evasion (serum resistance). The latter still has to be explored in a molecular context.

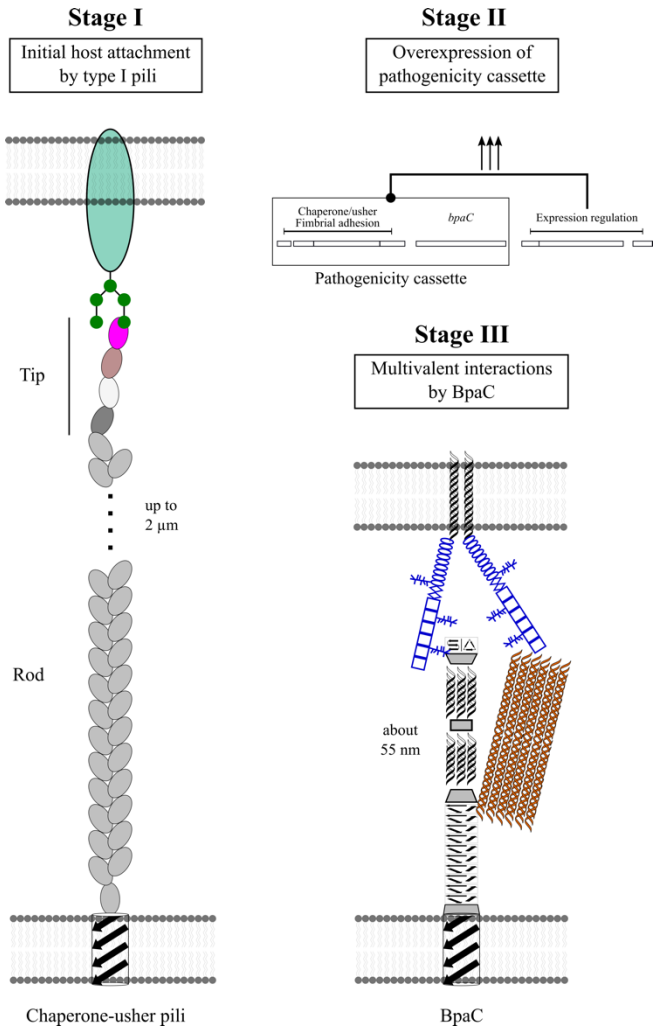


Figure 85 Proposed model of the involvement of BpaC in the pathogenicity of *B. pseudomallei* – In the first stage of attachment of the pathogen to the host cell, the chaperone-usher pili system, present in the expression cassette upstream of *bpaC*, binds via the *E. coli* equivalent of FimH (magenta) to a mannose-coated (green) target protein on the host surface. This initial binding is reinforced under high flow conditions via a catch-bond mechanism. In parallel, the second stage of the attachment cycle is initiated by the overexpression of the whole pathogenicity cassette containing both pili system and BpaC. In the final stage of the attachment of the pathogen to the host cell, BpaC binds via a multivalent interaction mechanism to the ECM proteins fibronectin (blue) and collagen I & II (orange), which introduces a very strong Velcro-like binding force between host cell and pathogen. The glycosylation of fibronectin could play a crucial role in the overall binding to BpaC and is indicated in the proposed model. However, this has not been confirmed yet for BpaC and based on observation in other TAAs that bind to fibronectin.

6.3 Suggestions for future perspectives

The information now available for BpaC provides an excellent point for future researchers to connect to what was discovered during this project. The development of a soluble domain construct for each functional domain of BpaC makes functional studies much more accessible than compared to the unoptimised starting constructs. The structural determination of the remaining parts of the passenger domain of BpaC has still room for optimisation and exploration as the actual crystallisation attempts on the final optimised constructs were quite limited due to Covid-19 time constraints. Especially the area of the stalk domain with high aggregating potential (stalk I) may contain hidden structural motifs that could provide more insight into the functional landscape of BpaC binding partners.

On the functional side of exploration, BpaC still harbours some secrets waiting to be unraveled, mainly because of its involvement in several processes related to immune evasion, all of which need specific effector proteins to bind to. A very likely candidate for serum resistance are components from the complement system like C3d or C4b-binding protein, as seen in other TAAs with similar structural motifs like UspA1 or YadA. A truncated version of C4b-binding protein exists and could be used for these binding experiments (Buffalo, Bahn-Suh et al. 2016).

Studying TAA binding interactions under more native conditions, in particular flow conditions, is underrepresented in the literature. This is however a necessary set of experiments to estimate the real binding profile of TAAs *in vivo*, because they are taking into account the multivalent binding surface of TAAs. One solution to overcome the more static binding assays is the switch to dynamic flow conditions like Müller et al. have demonstrated (Müller, Kaiser et al. 2011). The “modern” equivalent to this would be to switch to microfluidic sensor applications which maintain the flow aspect but increase the control one can exert on the experimental parameters, improving

reproducibility of these findings. Another solution might be to look at the specific bond interactions via atomic force microscopy, which allows the measurement of force constants and the identification of bond binding modes as discussed for the TAA BCAM0224 (El-Kirat-Chatel, Mil-Homens et al. 2013).

Furthermore, a more sophisticated and systematic approach to finding new binding partners is needed as currently the discovery of binding partners of TAAs is mostly performed by time and cost intensive guess work. Mass spectrometry based methods can provide new avenues by generating large datasets of possible interaction partners using whole cell pull down assays, for example. These hits can then be verified using standard biochemical interaction assays like surface plasmon resonance biosensors with purified proteins or domains.

The role of the lipopolysaccharide layer of Gram-negative bacteria in the context of TAA function has also not been extensively explored. One might wonder if the length of the C-terminal head domain of BpaC has some form of connection to the thickness of the lipopolysaccharide layer of *B. pseudomallei* and the composition thereof given the recent novel importance of glycosylation on TAA binding. This is especially interesting given the charge propensities of the head domain and the occurrence of shorter variants of BpaC in *B. mallei* and other *Burkholderia* subspecies.

Lastly, the development of small molecule inhibitors has been explored for YadA adhesion to collagen (Saragliadis and Linke 2019) and likely has similar success chances for BpaC, given that both bind to collagen and have a LPBR containing head domain. As glycosylation may be important for the binding of ECM proteins to TAAs, glycan mimetics might be worth exploring. In my opinion, this is the most promising research outlook for the ultimate goal of preventing *B. pseudomallei* infection with BpaC as target molecule and should be the first thing to be explored in the next generation of BpaC related experiments.

6.4 Conclusions

The goal of this project was to gather as much information as possible about the function of BpaC in the infection process of *B. pseudomallei*. For this, I made use of the unique sequence-to-structure relationship that is inherent to each TAA. The prediction of domains and even structural motifs *in silico* is a powerful tool to narrow down the potential binding partners, speeding up this process significantly. However, as the structural determination of the C-terminal head domain of BpaC has shown, the structural landscape of TAA motifs is still incomplete and deserves a closer look at, especially when trying to determine the importance of individual surface residues to certain binding propensities.

Both the prediction of an N-terminal head domain that resembles the head domain of the fibronectin binding BadA from *Bartonella henselae*, and a C-terminal head domain, which has the same Cα makeup as the collagen binding head domain of YadA from *Yersinia enterocolitica*, pointed towards both molecules as potential binding partners for BpaC. Both partners were confirmed for BpaC by the ECM assays carried out in this thesis and shows the strength of the structure-to-function connection of similar motifs, even if these are present in TAAs in different species.

The tendency of TAAs to autoaggregate was the largest challenge to overcome for the purpose of protein crystallisation and required a large degree of optimisation. This was mainly done by fusing the aggregation-prone domains to parts of the solubility-increasing C-terminal head domain or to the GCN4 adaptor. Using neck motifs as adaptors between different TAAs, one could imagine turning the LPBR repeats of BpaC into a general solubility tag for other TAAs, modulating the length and switching out the neck motif depending on the insolubility of the attached TAA domain(s).

Overall, this project helped to close the knowledge gap between the apparent phenotypic consequence of *bpaC* expression and the lack of biochemical information that was available at the start of it. The sequence has been fully analysed and modified to a degree that makes this protein highly accessible to the community for future cloning attempts, be it for purifying individual domains for functional assays or using full-length variations for *in vivo* assays. The structural insight gained, both from the homology model predictions, as well as the elucidation of the complete C-terminal head domain of BpaC will serve as a useful basis for future functional associations. The full impact of the surface charge propensities and novel solvent networks identified in the C-terminal head domain is yet to be determined but likely key for the overall function of the protein.

References

Adams, P. D., P. V. Afonine, G. Bunkoczi, V. B. Chen, N. Echols, J. J. Headd, L. W. Hung, S. Jain, G. J. Kapral, R. W. Grosse Kunstleve, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart (2011). "The Phenix software for automated determination of macromolecular structures." Methods **55**(1): 94-106.

Agnew, C., E. Borodina, N. R. Zaccai, R. Conners, N. M. Burton, J. A. Vicary, D. K. Cole, M. Antognozzi, M. Virji and R. L. Brady (2011). "Correlation of in situ mechanosensitive responses of the *Moraxella catarrhalis* adhesin UspA1 with fibronectin and receptor CEACAM1 binding." Proc Natl Acad Sci U S A **108**(37): 15174-15178.

Albenne, C. and R. Ieva (2017). "Job contenders: roles of the beta-barrel assembly machinery and the translocation and assembly module in autotransporter secretion." Mol Microbiol **106**(4): 505-517.

Allen, R. C., R. Popat, S. P. Diggle and S. P. Brown (2014). "Targeting virulence: can we make evolution-proof drugs?" Nat Rev Microbiol **12**(4): 300-308.

Attia, A. S., E. R. Lafontaine, J. L. Latimer, C. Aebi, G. A. Syrogiannopoulos and E. J. Hansen (2005). "The UspA2 protein of Moraxella catarrhalis is directly involved in the expression of serum resistance." Infect Immun **73**(4): 2400-2410.

Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millan, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker (2021). "Accurate prediction of protein structures and interactions using a three-track neural network." Science **373**(6557): 871-876.

Bassler, J., B. Hernandez Alvarez, M. D. Hartmann and A. N. Lupas (2015). "A domain dictionary of trimeric autotransporter adhesins." Int J Med Microbiol **305**(2): 265-275.

Berrow, N. S., D. Alderton, S. Sainsbury, J. Nettleship, R. Assenberg, N. Rahman, D. I. Stuart and R. J. Owens (2007). "A versatile ligation-independent cloning method suitable for high-throughput expression screening applications." Nucleic Acids Res **35**(6): e45.

Bertelli, C., M. R. Laird, K. P. Williams, G. Simon Fraser University Research Computing, B. Y. Lau, G. Hoad, G. L. Winsor and F. S. L. Brinkman (2017). "IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets." Nucleic Acids Res **45**(W1): W30-W35.

Blum, M., H. Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman and R. D. Finn (2021). "The InterPro protein families and domains database: 20 years on." Nucleic Acids Res **49**(D1): D344-D354.

Bocian-Ostrzycka, K. M., M. J. Grzeszczuk, A. M. Banas and E. K. Jagusztyn-Krynicka (2017). "Bacterial thiol oxidoreductases - from basic research to new antibacterial strategies." Appl Microbiol Biotechnol **101**(10): 3977-3989.

Buchan, D. W. A. and D. T. Jones (2019). "The PSIPRED Protein Analysis Workbench: 20 years on." Nucleic Acids Res **47**(W1): W402-W407.

Buffalo, C. Z., A. J. Bahn-Suh, S. P. Hirakis, T. Biswas, R. E. Amaro, V. Nizet and P. Ghosh (2016). "Conserved patterns hidden within group A *Streptococcus* M protein hypervariability recognize human C4b-binding protein." Nat Microbiol **1**(11): 16155.

Bullard, B., S. Lipski and E. R. Lafontaine (2007). "Regions important for the adhesin activity of *Moraxella catarrhalis* Hag." BMC Microbiol **7**(1): 65.

Burmolle, M., T. R. Thomsen, M. Fazli, I. Dige, L. Christensen, P. Homoe, M. Tvede, B. Nyvad, T. Tolker-Nielsen, M. Givskov, C. Moser, K. Kirketerp-Moller, H. K. Johansen, N. Hoiby, P. O. Jensen, S. J. Sorensen and T. Bjarnsholt (2010). "Biofilms in chronic infections - a matter of opportunity - monospecies biofilms in multispecies infections." FEMS Immunol Med Microbiol **59**(3): 324-336.

Burtnick, M. N., P. J. Brett, S. V. Harding, S. A. Ngugi, W. J. Ribot, N. Chantratita, A. Scorpio, T. S. Milne, R. E. Dean, D. L. Fritz, S. J. Peacock, J. L. Prior, T. P. Atkins and D. Deshazer (2011). "The cluster 1 type VI secretion system is a major virulence determinant in *Burkholderia pseudomallei*." Infect Immun **79**(4): 1512-1525.

Campos, C. G., M. S. Byrd and P. A. Cotter (2013). "Functional characterization of *Burkholderia pseudomallei* trimeric autotransporters." Infect Immun **81**(8): 2788-2799.

Capecchi, B., J. Adu-Bobie, F. Di Marcello, L. Ciucchi, V. Masignani, A. Taddei, R. Rappuoli, M. Pizza and B. Arico (2005). "Neisseria meningitidis NadA is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells." Mol Microbiol **55**(3): 687-698.

Chauhan, N., D. Hatlem, M. Orwick-Rydmark, K. Schneider, M. Floetenmeyer, B. van Rossum, J. C. Leo and D. Linke (2019). "Insights into the autotransport process of a trimeric autotransporter, Yersinia Adhesin A (YadA)." Mol Microbiol **111**(3): 844-862.

Conners, R., D. J. Hill, E. Borodina, C. Agnew, S. J. Daniell, N. M. Burton, R. B. Sessions, A. R. Clarke, L. E. Catto, D. Lammie, T. Wess, R. L. Brady and M. Virji (2008). "The *Moraxella* adhesin UspA1 binds to its human CEACAM1 receptor by a deformable trimeric coiled-coil." EMBO J **27**(12): 1779-1789.

Costa, T. R., C. Felisberto-Rodrigues, A. Meir, M. S. Prevost, A. Redzej, M. Trokter and G. Waksman (2015). "Secretion systems in Gram-negative bacteria: structural and mechanistic insights." Nat Rev Microbiol **13**(6): 343-359.

Cotter, S. E., N. K. Surana, S. Grass and J. W. St Geme, 3rd (2006). "Trimeric autotransporters require trimerization of the passenger domain for stability and adhesive activity." J Bacteriol **188**(15): 5400-5407.

Cowtan, K. (2006). "The Buccaneer software for automated model building. 1. Tracing protein chains." Acta Crystallogr D Biol Crystallogr **62**(Pt 9): 1002-1011.

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

Currie, B. J. (2010). "*Burkholderia pseudomallei* and *Burkholderia mallei*: melioidosis and glanders." Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases. 7th edn. Philadelphia: Churchill Livingstone Elsevier: 2869-2885.

Dautin, N. and H. D. Bernstein (2007). "Protein secretion in gram-negative bacteria via the autotransporter pathway." Annu Rev Microbiol **61**(1): 89-112.

Davies, J. and D. Davies (2010). "Origins and evolution of antibiotic resistance." Microbiol Mol Biol Rev **74**(3): 417-433.

Dunne, W. M., Jr. (2002). "Bacterial adhesion: seen any good biofilms lately?" Clin Microbiol Rev **15**(2): 155-166.

Edwards, T. E., A. S. Gardberg, E. R. Lafontaine and Seattle Structural Genomics Center for Infectious Disease (SSGCID) (2011). "Crystal structure of a YadA-like head domain of the trimeric autotransporter adhesin BoaA from *Burkholderia pseudomallei*."

Edwards, T. E., I. Phan, J. Abendroth, S. H. Dieterich, A. Masoudi, W. Guo, S. N. Hewitt, A. Kelley, D. Leibly, M. J. Brittnacher, B. L. Staker, S. I. Miller, W. C. Van Voorhis, P. J. Myler and L. J. Stewart (2010). "Structure of a *Burkholderia pseudomallei* trimeric autotransporter adhesin head." PLoS One **5**(9): 1--9.

El-Kirat-Chatel, S., D. Mil-Homens, A. Beaussart, A. M. Fialho and Y. F. Dufrene (2013). "Single-molecule atomic force microscopy unravels the binding mechanism of a *Burkholderia cenocepacia* trimeric autotransporter adhesin." Mol Microbiol **89**(4): 649-659.

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Evans, P. R. and G. N. Murshudov (2013). "How good are my data and what is the resolution?" Acta Crystallogr D Biol Crystallogr **69**(Pt 7): 1204-1214.

Forero, M., O. Yakovenko, E. V. Sokurenko, W. E. Thomas and V. Vogel (2006). "Uncoiling mechanics of *Escherichia coli* type I fimbriae are optimized for catch bonds." PLoS Biol **4**(9): e298.

Fux, C. A., J. W. Costerton, P. S. Stewart and P. Stoodley (2005). "Survival strategies of infectious biofilms." Trends Microbiol **13**(1): 34-40.

Galván, E. M., H. Chen and D. M. Schifferli (2007). "The Psa fimbriae of *Yersinia pestis* interact with phosphatidylcholine on alveolar epithelial cells and pulmonary surfactant." Infect Immun **75**(3): 1272-1279.

Galyov, E. E., P. J. Brett and D. DeShazer (2010). "Molecular insights into *Burkholderia pseudomallei* and *Burkholderia mallei* pathogenesis." Annu Rev Microbiol **64**(1): 495-517.

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch (2003). "ExPASy: The proteomics server for in-depth protein knowledge and analysis." Nucleic Acids Res **31**(13): 3784-3788.

Gong, L., M. Cullinane, P. Treerat, G. Ramm, M. Prescott, B. Adler, J. D. Boyce and R. J. Devenish (2011). "The *Burkholderia pseudomallei* type III secretion system and BopA are required for evasion of LC3-associated phagocytosis." PLoS One **6**(3): e17852.

Green, E. R. and J. Mecsas (2016). "Bacterial Secretion Systems: An Overview." Microbiol Spectr **4**(1).

Guerin, J., S. Bigot, R. Schneider, S. K. Buchanan and F. Jacob-Dubuisson (2017). "Two-Partner Secretion: Combining Efficiency and Simplicity in the Secretion of Large Proteins for Bacteria-Host and Bacteria-Bacteria Interactions." Front Cell Infect Microbiol **7**: 148.

Harbury, P. B., P. S. Kim and T. Alber (1994). "Crystal structure of an isoleucine-zipper trimer." Nature **371**(6492): 80-83.

Hartmann, M. D., I. Grin, S. Dunin-Horkawicz, S. Deiss, D. Linke, A. N. Lupas and B. Hernandez Alvarez (2012). "Complete fiber structures of complex trimeric autotransporter adhesins conserved in enterobacteria." Proc Natl Acad Sci U S A **109**(51): 20907-20912.

Heger, A. and L. Holm (2000). "Rapid automatic detection and alignment of repeats in protein sequences." Proteins **41**(2): 224-237.

Hernandez Alvarez, B., M. D. Hartmann, R. Albrecht, A. N. Lupas, K. Zeth and D. Linke (2008). "A new expression system for protein crystallization using trimeric coiled-coil adaptors." Protein Eng Des Sel **21**(1): 11-18.

Hoiczyk, E., A. Roggenkamp, M. Reichenbecher, A. Lupas and J. Heesemann (2000). "Structure and sequence analysis of *Yersinia* YadA and *Moraxella* UspAs reveal a novel class of adhesins." EMBO J **19**(22): 5989-5999.

Holm, L. and L. M. Laakso (2016). "Dali server update." Nucleic Acids Res **44**(W1): W351-355.

Hovingh, E. S., B. van den Broek and I. Jongerius (2016). "Hijacking Complement Regulatory Proteins for Bacterial Immune Evasion." Front Microbiol **7**: 2004.

Ishikawa, M., H. Nakatani and K. Hori (2012). "AtaA, a new member of the trimeric autotransporter adhesins from *Acinetobacter sp*. Tol 5 mediating high adhesiveness to various abiotic surfaces." PLoS One **7**(11): e48830.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.

Kaiser, P. O., D. Linke, H. Schwarz, J. C. Leo and V. A. Kempf (2012). "Analysis of the BadA stalk from *Bartonella henselae* reveals domain-specific and domain-overlapping functions in the host cell infection process." Cell Microbiol **14**(2): 198-209.

Kajava, A. V. and A. C. Steven (2006). "The turn of the screw: variations of the abundant beta-solenoid motif in passenger domains of Type V secretory proteins." J Struct Biol **155**(2): 306-315.

Kespichayawattana, W., P. Intachote, P. Utaisincharoen and S. Sirisinha (2004). "Virulent *Burkholderia pseudomallei* is more efficient than avirulent *Burkholderia thailandensis* in invasion of and adherence to cultured human epithelial cells." Microb Pathog **36**(5): 287-292.

Kibbe, W. A. (2007). "OligoCalc: an online oligonucleotide properties calculator." Nucleic Acids Res **35**(Web Server issue): W43-46.

Kiessling, A. R., S. A. Harris, K. M. Weimer, G. Wells and A. Goldman (2022). "The C-terminal head domain of *Burkholderia pseudomallei* BpaC has a striking hydrophilic core with an extensive solvent network." Mol Microbiol **118**(1-2): 77-91.

Klauser, T., J. Pohlner and T. F. Meyer (1993). "The secretion pathway of IgA protease-type proteins in gram-negative bacteria." Bioessays **15**(12): 799-805.

Koiwai, K., M. D. Hartmann, D. Linke, A. N. Lupas and K. Hori (2016). "Structural Basis for Toughness and Flexibility in the C-terminal Passenger Domain of an *Acinetobacter* Trimeric Autotransporter Adhesin." J Biol Chem **291**(8): 3705-3724.

Kozlowski, L. P. (2021). "IPC 2.0: prediction of isoelectric point and pKa dissociation constants." Nucleic Acids Res **49**(W1): W285-W292.

Lafontaine, E. R., R. Balder, F. Michel and R. J. Hogan (2014). "Characterization of an autotransporter adhesin protein shared by *Burkholderia mallei* and *Burkholderia pseudomallei*." BMC Microbiol **14**(1): 92.

Lazar Adler, N. R., M. P. Stevens, R. E. Dean, R. J. Saint, D. Pankhania, J. L. Prior, T. P. Atkins, B. Kessler, A. Nithichanon, G. Lertmemongkolchai and E. E. Galyov (2015). "Systematic mutagenesis of genes encoding predicted autotransported proteins of *Burkholderia pseudomallei* identifies factors mediating virulence in mice, net intracellular replication and a novel protein conferring serum resistance." PLoS One **10**(4): e0121271.

Leo, J. C., H. Elovaara, B. Brodsky, M. Skurnik and A. Goldman (2008). "The *Yersinia* adhesin YadA binds to a collagenous triple-helical conformation but without sequence specificity." Protein Eng Des Sel **21**(8): 475-484.

Leo, J. C., I. Grin and D. Linke (2012). "Type V secretion: mechanism(s) of autotransport through the bacterial outer membrane." Philos Trans R Soc Lond B Biol Sci **367**(1592): 1088-1101.

Leo, J. C., A. Lyskowski, K. Hattula, M. D. Hartmann, H. Schwarz, S. J. Butcher, D. Linke, A. N. Lupas and A. Goldman (2011). "The structure of *E. coli* IgG-binding protein D suggests a general model for bending and binding in trimeric autotransporter adhesins." Structure **19**(7): 1021-1030.

Liguori, A., L. Dello Iacono, G. Maruggi, B. Benucci, M. Merola, P. Lo Surdo, J. Lopez-Sagaseta, M. Pizza, E. Malito and M. J. Bottomley (2018). "NadA3 Structures Reveal Undecad Coiled Coils and LOX1 Binding Regions Competed by *Meningococcus B* Vaccine-Elicited Human Antibodies." mBio **9**(5): e01914-01918.

Limmathurotsakul, D., N. Golding, D. A. B. Dance, J. P. Messina, D. M. Pigott, C. L. Moyes, D. B. Rolim, E. Bertherat, N. P. J. Day, S. J. Peacock and S. I. Hay (2016). "Predicted global distribution of *Burkholderia pseudomallei* and burden of melioidosis." Nature Microbiology **1**(1): 15008.

Linke, D., T. Riess, I. B. Autenrieth, A. Lupas and V. A. Kempf (2006). "Trimeric autotransporter adhesins: variable structure, common function." Trends Microbiol **14**(6): 264-270.

Ludwiczak, J., A. Winski, K. Szczepaniak, V. Alva and S. Dunin-Horkawicz (2019). "DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences." Bioinformatics **35**(16): 2790-2795.

Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn and R. Lopez (2019). "The EMBL-EBI search and sequence analysis tools APIs in 2019." Nucleic Acids Res **47**(W1): W636-W641.

Mathelie-Guinlet, M., F. Viela, D. Alsteens and Y. F. Dufrene (2021). "Stress-Induced Catch-Bonds to Enhance Bacterial Adhesion." Trends Microbiol **29**(4): 286-288.

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

Meng, G., J. W. St Geme III and G. Waksman (2008). "Repetitive architecture of the Haemophilus influenzae Hia trimeric autotransporter." J Mol Biol **384**(4): 824-836.

Meng, G., N. K. Surana, J. W. St Geme, 3rd and G. Waksman (2006). "Structure of the outer membrane translocator domain of the Haemophilus influenzae Hia trimeric autotransporter." EMBO J **25**(11): 2297-2304.

Meuskens, I., A. Saragliadis, J. C. Leo and D. Linke (2019). "Type V Secretion Systems: An Overview of Passenger Domain Functions." Front Microbiol **10**: 1163.

Mikula, K. M., R. Kolodziejczyk and A. Goldman (2012). "*Yersinia* infection tools-characterization of structure and function of adhesins." Front Cell Infect Microbiol **2**: 169.

Mikula, K. M., R. Kolodziejczyk and A. Goldman (2019). "Structure of the UspA1 protein fragment from Moraxella catarrhalis responsible for C3d binding." J Struct Biol **208**(2): 77-85.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman (2021). "Pfam: The protein families database in 2021." Nucleic acids research **49**(D1): D412-D419.

Moore, S. (2012). "Round-the-horn site-directed mutagenesis." Open-WetWare http://openwetware.org/wiki/'Round-the-horn_sitedirected_mutagenesis.

Murray, C. J., K. S. Ikuta, F. Sharara, L. Swetschinski, G. R. Aguilar, A. Gray, C. Han, C. Bisignano, P. Rao and E. Wool (2022). "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis." Lancet **399**(10325): 629-655.

Murshudov, G. N., P. Skubak, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long and A. A. Vagin (2011). "REFMAC5 for the refinement of macromolecular crystal structures." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 355-367.

Müller, N. F., P. O. Kaiser, D. Linke, H. Schwarz, T. Riess, A. Schafer, J. A. Eble and V. A. Kempf (2011). "Trimeric autotransporter adhesin-dependent adherence of *Bartonella henselae*, *Bartonella quintana*, and *Yersinia enterocolitica* to matrix components and endothelial cells under static and dynamic flow conditions." Infect Immun **79**(7): 2544-2553.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.

Nummelin, H., M. C. Merckel, J. C. Leo, H. Lankinen, M. Skurnik and A. Goldman (2004). "The Yersinia adhesin YadA collagen-binding domain structure is a novel left-handed parallel beta-roll." EMBO J **23**(4): 701-711.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy and K. D. Pruitt (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic Acids Res **44**(D1): D733-745.

Pimenta, A. I., M. Kilcoyne, N. Bernardes, D. Mil-Homens, L. Joshi and A. M. Fialho (2021). "*Burkholderia cenocepacia* BCAM2418-induced antibody inhibits bacterial adhesion, confers protection to infection and enables identification of host glycans as adhesin targets." Cell Microbiol.

Pimenta, A. I., D. Mil-Homens and A. M. Fialho (2020). "*Burkholderia cenocepacia*-host cell contact controls the transcription activity of the trimeric autotransporter adhesin *BCAM2418* gene." Microbiologyopen: e998.

Pimenta, A. I., D. Mil-Homens, S. N. Pinto and A. M. Fialho (2020). "Phenotypic characterization of trimeric autotransporter adhesin-defective bcaC mutant of Burkholderia cenocepacia: cross-talk towards the histidine kinase BCAM0218." Microbes Infect **22**(9): 457-466.

Pina, A. S., S. Carvalho, A. M. Dias, M. Guilherme, A. S. Pereira, L. T. Caraca, A. S. Coroadinha, C. R. Lowe and A. C. Roque (2016). "Tryptophan tags and de novo designed complementary affinity ligands for the expression and purification of recombinant proteins." J Chromatogr A **1472**: 55-65.

Plackett, B. (2020). "No money for new drugs." Nature **586**(7830): S50-S52.

Potter, S. C., A. Luciani, S. R. Eddy, Y. Park, R. Lopez and R. D. Finn (2018). "HMMER web server: 2018 update." Nucleic acids research **46**(W1): W200-W204.

Potterton, L., J. Agirre, C. Ballard, K. Cowtan, E. Dodson, P. R. Evans, H. T. Jenkins, R. Keegan, E. Krissinel, K. Stevenson, A. Lebedev, S. J. McNicholas, R. A. Nicholls, M. Noble, N. S. Pannu, C. Roth, G. Sheldrick, P. Skubak, J. Turkenburg, V. Uski, F. von Delft, D. Waterman, K. Wilson, M. Winn and M. Wojdyr (2018). "CCP4i2: the new graphical user interface to the CCP4 program suite." Acta Crystallogr D Struct Biol **74**(Pt 2): 68-84.

Prestinaci, F., P. Pezzotti and A. Pantosti (2015). "Antimicrobial resistance: a global multifaceted phenomenon." Pathog Glob Health **109**(7): 309-318.

Project, I. (2020). Inkscape.

Ribeiro, S. M., M. R. Felicio, E. V. Boas, S. Goncalves, F. F. Costa, R. P. Samy, N. C. Santos and O. L. Franco (2016). "New frontiers for anti-biofilm drug development." Pharmacol Ther **160**: 133-144.

Sandt, C. H. and C. W. Hill (2001). "Nonimmune binding of human immunoglobulin A (IgA) and IgG Fc by distinct sequence segments of the EibF cell surface protein of *Escherichia coli*." Infect Immun **69**(12): 7293-7303.

Saragliadis, A. and D. Linke (2019). "Assay development for the discovery of small-molecule inhibitors of YadA adhesion to collagen." Cell Surf **5**: 100025.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak and A. Cardona (2012). "Fiji: an open-source platform for biological-image analysis." Nat Methods **9**(7): 676-682.

Schmid, Y., G. A. Grassl, O. T. Buhler, M. Skurnik, I. B. Autenrieth and E. Bohn (2004). "Yersinia enterocolitica adhesin A induces production of interleukin-8 in epithelial cells." Infect Immun **72**(12): 6780-6789.

Schmidt, T. G. and A. Skerra (2007). "The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins." Nat Protoc **2**(6): 1528-1535.

Schrodinger, L. (2017). The PyMOL Molecular Graphics System, Version 2.0.

Selkrig, J., K. Mosbahi, C. T. Webb, M. J. Belousoff, A. J. Perry, T. J. Wells, F. Morris, D. L. Leyton, M. Totsika, M. D. Phan, N. Celik, M. Kelly, C. Oates, E. L. Hartland, R. M. Robins-Browne, S. H. Ramarathinam, A. W. Purcell, M. A. Schembri, R. A. Strugnell, I. R. Henderson, D. Walker and T. Lithgow (2012). "Discovery of an archetypal protein transport system in bacterial outer membranes." Nat Struct Mol Biol **19**(5): 506-510, S501.

Shahid, S. A., B. Bardiaux, W. T. Franks, L. Krabben, M. Habeck, B. J. van Rossum and D. Linke (2012). "Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals." Nat Methods **9**(12): 1212-1217.

Singer, A. C., H. Shaw, V. Rhodes and A. Hart (2016). "Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators." Front Microbiol **7**: 1728.

Steenhuis, M., A. M. Abdallah, S. M. de Munnik, S. Kuhne, G. J. Sterk, B. van den Berg van Saparoea, S. Westerhausen, S. Wagner, N. N. van der Wel, M. Wijtmans, P. van Ulsen, W. S. P. Jong and J. Luirink (2019). "Inhibition of autotransporter biogenesis by small molecules." Mol Microbiol **112**(1): 81-98.

Stein, N. (2008). "CHAINSAW: a program for mutating pdb files used as templates in molecular replacement." J Appl Crystallogr **41**(3): 641-643.

Szczesny, P., D. Linke, A. Ursinus, K. Bar, H. Schwarz, T. M. Riess, V. A. Kempf, A. N. Lupas, J. Martin and K. Zeth (2008). "Structure of the head of the *Bartonella* adhesin BadA." PLoS Pathog **4**(8): e1000119.

Szczesny, P. and A. Lupas (2008). "Domain annotation of trimeric autotransporter adhesins--daTAA." Bioinformatics **24**(10): 1251-1256.

Team, T. G. D. (2019). GIMP.

Thanassi, D. G., J. B. Bliska and P. J. Christie (2012). "Surface organelles assembled by secretion systems of Gram-negative bacteria: diversity in structure and function." FEMS Microbiol Rev **36**(6): 1046-1082.

Thibau, A., A. A. Dichter, D. J. Vaca, D. Linke, A. Goldman and V. A. J. Kempf (2020). "Immunogenicity of trimeric autotransporter adhesins and their potential as vaccine targets." Med Microbiol Immunol **209**(3): 243-263.

Thibau, A., K. Hipp, D. J. Vaca, S. Chowdhury, J. Malmstrom, A. Saragliadis, W. Ballhorn, D. Linke and V. A. J. Kempf (2022). "Long-Read Sequencing Reveals Genetic Adaptation of *Bartonella* Adhesin A Among Different *Bartonella henselae* Isolates." Front Microbiol **13**: 838267.

Tram, G., J. Poole, F. G. Adams, M. P. Jennings, B. A. Eijkelkamp and J. M. Atack (2021). "The *Acinetobacter baumannii* Autotransporter Adhesin Ata Recognizes Host Glycans as High-Affinity Receptors." ACS infectious diseases **7**(8): 2352-2361.

UniProt, C. (2019). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Res **47**(D1): D506-D515.

Uribe-Querol, E. and C. Rosales (2017). "Control of Phagocytosis by Microbial Pathogens." Front Immunol **8**: 1368.

Urzhumtseva, L., P. V. Afonine, P. D. Adams and A. Urzhumtsev (2009). "Crystallographic model quality at a glance." Acta Crystallogr D Biol Crystallogr **65**(Pt 3): 297-300.

Vaca, D. J., A. Thibau, M. Schutz, P. Kraiczy, L. Happonen, J. Malmstrom and V. A. J. Kempf (2020). "Interaction with the host: the role of fibronectin and extracellular matrix proteins in the adhesion of Gram-negative bacteria." Med Microbiol Immunol **209**(3): 277-299.

Valle, J., A. N. Mabbett, G. C. Ulett, A. Toledo-Arana, K. Wecker, M. Totsika, M. A. Schembri, J. M. Ghigo and C. Beloin (2008). "UpaG, a new member of the trimeric autotransporter family of adhesins in uropathogenic *Escherichia coli*." J Bacteriol **190**(12): 4147-4161.

Wiersinga, W. J., T. van der Poll, N. J. White, N. P. Day and S. J. Peacock (2006). "Melioidosis: insights into the pathogenicity of *Burkholderia pseudomallei*." Nat Rev Microbiol **4**(4): 272-282.

Wilkins, M. R., E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel and D. F. Hochstrasser (1999). "Protein identification and analysis tools in the ExPASy server." Methods Mol Biol **112**: 531-552.

Williams, C. J., J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall, 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson and D. C. Richardson (2018). "MolProbity: More and better reference data for improved all-atom structure validation." Protein science : a publication of the Protein Society **27**(1): 293-315.

Winsor, G. L., B. Khaira, T. Van Rossum, R. Lo, M. D. Whiteside and F. S. Brinkman (2008). "The *Burkholderia* Genome Database: facilitating flexible queries and comparative analyses." Bioinformatics **24**(23): 2803-2804.

Wright, J., M. Thomsen, R. Kolodziejczyk, J. Ridley, J. Sinclair, G. Carrington, B. Singh, K. Riesbeck and A. Goldman (2017). "The crystal structure of PD1, a Haemophilus surface fibril domain." Acta Crystallogr F Struct Biol Commun **73**(Pt 2): 101-108.

Yeo, H. J., S. E. Cotter, S. Laarmann, T. Juehne, J. W. St Geme, 3rd and G. Waksman (2004). "Structural basis for host recognition by the Haemophilus influenzae Hia autotransporter." EMBO J **23**(6): 1245-1256.

Zav'yalov, V., A. Zavialov, G. Zav'yalova and T. Korpela (2010). "Adhesive organelles of Gram-negative pathogens assembled with the classical chaperone/usher machinery: structure and function from a clinical standpoint." FEMS Microbiol Rev **34**(3): 317-378.

Appendix A Optimised *bpaC* sequence

ATGAACAGGATTTTCAAATCGATCTGGTGCGAACAGACGCGTACGTGGGTTGCGGCATCGGAGC
ATGCCGTAGCGCGCGGTGGCCGCGCGTCGAGCGTCGTCGCGTCCGCCGGCGGATTGGAGAAAGT
GCTCAAGCTGTCGATTCTGGGCGCGGCATCGCTGATTGCGATGGGCGTGGTCGGACCGTTTGCC
GAGGAGGCAATGGCGGCGAATAACGCCGGTGTGTGTTTGACGTACAACGGTAGTAGCAACAATA
CATCAGGTACTGGCGGCTGGTTCGCTGATGGTTGTAAATCGGCCGGCTGGGTGCAGGGCATGGT
TACGAATAGCAAGACGGATTGGGTCGGGCTGACCGCGGACGACACGCAGATCGTGCTCGACGGT
AGCGCGGGCAGCATTTACTTCCGGACGGGCGGCATAAACGGCAACGTGTTGACGATGTCGAACG
CGACCGGCGGCGTATTGCTCAGCGGCCTCGCGGCCGGCGTCAATCCGACCGATGCGGTCAACAT
GTCCCAGTTGACCTCATTGAGCACTTCAACAGCAACCGGCATCACCTCGCTTAGTACATCTACT
GCTACCAGCATCGCTAGCTTAAGTACTTCAATGCTGTCGCTCGGCGTGGGCGTCGTGACGCAAG
ACGCCTCGACCGGCGCGATCAGCGTCGGCGCCAATTCGCCGGGCCTGACGGTGGATTTCGCGGG
GGGCCAGGGCCCGCGCACGCTGACGGGCGTCGCTGCAGGTGTAAACGCGACTGATGCAGTGAAT
GTAGGCCAGTTGGCGTCACTGTCAACGTCAACTGCAGCTGGGCTTTCCACTGCCGCGAGCGGCG
TCGCAAGCTTATCGACGAGCTTATTGGGTGCGGTGGGCGATCTGGCAAGCTTGAGCACATCTGC
ATCGACGGGGCTCGCCACTGCGGATAGCGGCATCGCGTCGTTGTCCACGTCGCTGCTCGGCACC
GCGGACAACGTGACTAGCTTAAGTACGAGCCTCAGCACGGTCAACGCGAATCTGGCCGGCCTGC
AGACCTCGGTGGACAACGTCGTGTCATACGACGATCCGTCGAAGTCGGCGATCACCCTCGGCGG
TGCGGGCGTCACGACGCCCGTCCTGCTGACGAACGTGGCTGCGGGGAAGATCGCCGCGACCTCA
ACTGATGCAGTGAACGGTTCGCAGCTTTACACGCTCCAGCAGGAGTTCTCGCAGCAGTACGATC
TGCTGACGTCGCAAGTCTCGTCGCTCAGCACCTCGGTGTCGGGTCTCCAAGGCAGCGTCTCGGC
AAATACGGGAACCGCAAGCGGTGATAATAGCACTGCAAGTGGCGACAACGCCACCGCGTCAGGC
ACCAACTCCACGGCGAACGGCACGAATAGCACCGCCAGCGGCGACAATAGCACCGCAAGTGGCA
CTAATGCATCCGCCTCGGGCGAAAACTCAACCGCAACCGGGACGGACAGCACAGCTTCAGGAAG
CAACAGTACGGCGAACGGCACAAATAGCACTGCCTCAGGCGATAACAGCACCGCCTCCGGCACG
AACGCGTCGGCGACCGGCGAGAACAGCACGGCCACGGGCACGGACAGCACGGCAAGTGGCTCGA
ACAGCACCGCAAATGGCACGAACTCCACCGCCAGTGGTGATAATAGCACTGCCTCGGGCACTAA
TGCCTCAGCAAGCGGTGAGAATAGCACTGCAACTGGGACCGATAGCACTGCAAGCGGCTCGAAC
TCCACGGCCAACGGGACGAACAGCACGGCCTCCGGTGACAATAGTACCGCATCGGGTACTAATG
CAAGTGCCACAGGCGAGAACAGCACTGCAACAGGGACGGACAGCACGGCGTCAGGTTCGAATAG
TACCGCAAACGGAACGAATAGTACTGCGAGCGGTGATAACTCTACGGCCTCGGGAACAAACGCC
AGCGCCACTGGGGAAAACAGCACGGCGACTGGTACCGATTCAACGGCGTCGGGCTCGAACTCCA
CAGCGAACGGCACGAATTCAACCGCAAGCGGGGACAATTCGACTGCATCTGGTACCAACGCGAG
CGCTACCGGCGAGAACTCGACAGCGACGGGGACCGACAGCACCGCAAGTGGTAGCAACTCCACG
GCAAACGGCACCAACAGCACAGCGTCTGGGGATAATAGTACTGCTAGCGGGACTAACGCGTCCG
CGACGGGCGAGAACTCCACGGCGACCGGAACGGACAGTACCGCCAGTGGCTCCAACTCAACCGC
GAACGGCGCGAATAGTACAGCAAGCGGTGATAACTCCACAGCTAGTGGGACCAATGCCAGCGCG
ACCGGAGAAAACTCCACCGCGACAGGAACAGATAGCACGGCCTCGGGGTCCAATAGCACCGCGT
CGGGTTCAAATAGCACCGCCTCGGGAAATAATTCTACCGCTAGCGGAACAAATGCATCAGCCAC
TGGTGAAAATAGCACTGCAACAGGCACCGACTCGGCGGCATCGGGCACGAATTCCACTGCCAAT
GGCACCAACTCCACGGCTAGCGGCGACAACAGCACCGCTAGTGGCACTAATGCGAGCGCGACCG
GTGAAAATTCAACTGCGACGGGCACGGCGAGCACGGCGAGCGGCTCGAACAGCACGGCGAACGG
CGCAAACAGCACGGCCTCAGGGGCGGGCGCGACTGCAACCGGGGAGAACGCTGCCGCAACAGGA
GCAGGCGCCACAGCAACGGGCAACAATGCAAGTGCTTCAGGCACCTCGTCGACGGCAGGTGGTG

CAAATGCAATCGCCTCCGGCGAAAACAGTACGACGAATGGTGCAAATTCCACCGCGTCGGGCAA
CGGCTCCAGCGCCTTCGGCGAGTCGGCGGCCGCCGCGGGCGATGGAAGCACGGCGTTGGGTGCA
AATGCTGTCGCGAGCGGTGTGGGAAGCGTTGCGACGGGCGCGGGCTCGGTGGCGTCAGGTGCAA
ATAGTAGTGCATATGGCACCGGCAGCAACGCGACCGGGGCAGGTAGCGTTGCGATCGGCCAAGG
CGCGACGGCCTCGGGATCGAACTCGGTCGCGCTTGGCACCGGTTCTGTCGCGTCGGAGGACAAC
ACGGTATCGGTCGGCTCCGCAGGCAGCGAGCGCAGGATCACCAACGTCGCCGCCGGCGTCAATG
CAACCGACGCCGTCAACGTCGGCCAGTTGAACAGCGCCGTGTCGGGCATCCGGAATCAGATGGA
CGGCATGCAAGGCCAGATCGATACGCTTGCACGCGATGCGTATTCCGGTATCGCGGCCGCGACC
GCGTTGACGATGATTCCGGACGTGGATCCGGGCAAGACGCTGGCCGTGGGCATCGGCACGGCCA
ATTTCAAGGGCTACCAAGCCTCCGCGCTCGGCGCGACCGCACGTATCACCCAGAACCTCAAGGT
GAAGACGGGCGTGAGCTACAGCGGCAGCAACTACGTGTGGGGCGCGGGCATGTCGTATCAGTGG

Appendix B In-house protocol for the creation of competent *E. coli* cells

Day 1

[1] Aseptically streak the required strain, from a glycerol stock, onto an LB-agar plate containing the antibiotic to which the strain is resistant (tetracycline). Grow overnight at 37 °C in a stationary incubator.

Day 2

[2] Aseptically pick a single colony from the LB-agar plate into 2 x 5 mL of LB medium (containing the appropriate antibiotic - tetracycline). Grow overnight at 37 °C in a loosely-capped tube in an orbital incubator at 200 rpm.

Day 3

[3] Incubate 400 mL of pre-warmed LB medium (plus antibiotic) with ~800 μL of the overnight culture for an OD_{600} of 0.08-0.09. Grow at 37 °C in an orbital incubator at 200 rpm until the OD_{600} is between 0.4-0.6. Take half-hourly samples in order to monitor the OD_{600}. Do so aseptically, in the presence of a Bunsen burner, to prevent contamination of the cultures. Expect the required attenuance to be reached after 2-4 h. Start at an OD_{600} of 0.08-0.09.

Working on ice throughout, where possible:

[4] Transfer the culture into a 8 x 50 mL Falcon Tube and chill on ice for 5 min.

[5] Centrifuge the cells in a swinging-bucket rotor for 10 min at 3000 rpm Discard the supernatant.

[6] Gently re-suspend the cell pellet in 20 mL of ice-cold Tfb1 buffer. Use a 5 mL pipette tip, which has a wide opening, to avoid damage to the cells. Do not vortex: the cells are fragile. This step may take several minutes so periodically return the tube to the ice to keep it cold.

[7] Incubate on ice for 5 min.

[8] Centrifuge the cells in a swinging-bucket rotor for 10 min at 3000 rpm. Discard the supernatant.

[9] Gently re-suspend the cells in 1 mL Tfb2 buffer per 50 mL of culture, as above. Keep the tube cold throughout the re-suspension. Do not vortex.

[10] Incubate on ice for 15 min.

[11] Transfer 50 µL aliquots into sterile, cooled Eppendorf tubes.

[12] Freeze on dry ice and store at –80 °C.

SOLUTIONS

Luria-Bertani (LB) Medium (1 L)

Compound	Amount	Molar mass	Manufacturer	Catalogue number	Final concentration
Tryptone	10 g	-	Melford	T1332	1%
Yeast extract	5 g	-	Melford	Y1332	0.5%
NaCl	10 g	58.44 g/mol	Lancaster	L13268	1%

Dissolve the above in 950 mL milliQ H_2O, adjust the pH to 7.0 with several drops of 5 M NaOH, make up to 1 L, transfer to 100 mL – 500 mL bottles, then autoclave for 20-30 min at 15 psi on liquid cycle.

Transformation buffer 1 (Tfb1, 100 mL)

Compound	Amount	Molar mass	Manufacturer	Catalogue number	Final concentration
KAc	294.4 mg	98.14 g/mol	Sigma	P1147	30 mM
$RbCl_2$	1.21 g	120.92 g/mol	Fluka Biochemicals	83979	100 mM
$CaCl_2$ $2H_2O$	147 mg	147.02 g/mol	BDH	100704Y	10 mM
$MnCl_2$ $4H_2O$	989.6 mg	197.91 g/mol	Sigma	M8266	50 mM
Glycerol	15 mL	-	Sigma	G6279	15%

Dissolve the above in 80 mL milliQ H_2O, then adjust the pH to 5.8 using 0.2 M acetic acid (Fisher Scientific A/0400/PB17). Be very careful as you approach 5.8; if the pH drops below 5.8, a black precipitate may form. Make up the volume to 100 mL then filter-sterilise.

Transformation buffer 2 (Tfb2, 100 mL)

Compound	Amount	Molar mass	Manufacturer	Catalogue Number	Final concentration
MOPS	0.21 g	209.3 g/mol	Fisher Scientific	BPE308	10 mM
$CaCl_2$ $2H_2O$	1.1 g	147.02 g/mol	BDH	100704Y	75 mM
$RbCl_2$	0.12 g	120.92 g/mol	Fluka Biochemicals	83979	10 mM
Glycerol	15 mL		Sigma	G6279	15%

Dissolve the above in 95 mL milliQ H_2O, then adjust the pH to 6.5 using KOH (Fisher Scientific P/5640/60). Make up the volume to 100 mL, then filter-sterilise 5 mL aliquots.