PROTEIN STRUCTURE DYNAMICS IN GRAM NEGATIVE BACTERIAL SECRETION SYSTEMS

NICHOLAS BRAGAGNOLO

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN

PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CHEMISTRY

YORK UNIVERSITY

TORONTO, ONTARIO

DECEMBER 2024

Abstract

Bacteria use a variety of secretion systems that enhance their virulence. The conjugative type IV secretion system attributes transmission of mobile DNA elements in bacteria, whereas type IV pili are akin to the type II secretion system and are responsible for host adhesion and twitching motility. Understanding the structure and function of components in these systems has the potential to elucidate the mechanisms of bacterial commensalism and antagonism, and would aid in the development of inhibitors to bacterial pathogenesis by disrupting the tools virulent species use for stable, long-term infections. The entry exclusion protein of the F-like type IV secretion system, TraG, consists of a membrane-bound N-terminal domain and a periplasmic C-terminal domain, denoted TraG*. TraG* is essential in preventing redundant DNA transfer through interaction with a cognate TraS in the inner membrane of the recipient cell to prevent conjugation when the recipient cell carries the same plasmid. Using a multitude of biophysical methods including thermofluor, circular dichroism, collision induced unfolding mass spectrometry, and small angle X-ray scattering, the structural dynamics of TraG* was investigated. This allowed for the characterisation of the 50 N-terminal residues, the 77 C-terminal residues, and the TraS-interacting region of the protein as highly dynamic, providing evidence for the mechanism of the long-distance interactions made by TraG. This dissertation also describes the 1.3 Å crystal structure of the N-terminally truncated Type IV pilin of *Pseudomonas aeruginosa* from strain P1 (ΔP1), the first structure of its phylogenetically linked group (group Ia) to be reported. The structure was solved from X-ray diffraction data collected 20 years ago with a molecular replacement search model generated using AlphaFold. Comparisons of the ΔP1 structure to other Type IV pilins using ProSMART indicated that there are cases of higher structural homology between different phylogenetic groups of *P. aeruginosa* than there are between pilins from the

same phylogenetic group, indicating that the classification of pilins based on structural homology may be necessary in developing anti-virulence drugs and vaccines.

Acknowledgments

Firstly, I would like to thank my supervisor Dr. Gerald Audette for his guidance and his trust in me throughout my graduate schooling experience. His confidence in my independence as a scientist allowed me to undertake many challenges and thrive in them. I would also like to thank my advisors Dr. Philip Johnson and Dr. Derek Wilson for serving on my defense committee and providing me valuable feedback on my project. I thank Dr. Ozzy Mermut and Dr. Michael Suits for serving as the external advisors on my dissertation defense committee, reading my thesis thoroughly and asking great questions. I would like to thank Dr. Tamara Kelly, Dr. Aarthi Ashok and the Toronto chapter of the Canadian Consortium of Science Equity Scholars for allowing me to perform valuable pedagogical research with them. I am appreciative of the administrators in the department of Chemistry for their assistance, especially Mandy Ramnaraine.

I thank my fellow Audette lab members, past and present. Ayat Yaseen for training me in many vital techniques and for welcoming me in the lab, and Christina Rodriguez and Arnold Apostol for their friendship, advice, and conversations throughout our time in the Audette lab. I'd like to thank members of the Wilson lab; Esther, Banafsheh, and Vimanda for their thoughtful discussions and aid in experimentation, especially Cristina Lento for guiding my learning in performing mass spectrometry on my own. I appreciate the scientists at the BioCAT facility of the Advanced Photon Source who helped me perform small angle X-ray scattering. I would like to thank Dr. Katalin Hudak and the graduate students from her lab that helped cultivate my independence in a research lab and provided me the foundational skills required for graduate school. Acknowledgements are deserved by many faculty of science members and colleagues at York University and many other institutions that I'm happy to call friends, especially those in the Life Science Building; there are too many to name but know that your contributions to my academic career and to my life are innumerable.

I would like to thank my friends and family; every one of you has supported me in some way throughout my journey to get here. For my band members for always helping me to destress after a long week in the lab by giving me an opportunity to perform. For my brothers John and Mark, for teaching me that life without humour and music is like a protein crystal that does not diffract. My dog Tanner for providing me regular serotonin boosts and keeping me from going hyperglycemic by requiring lots of walks. Prominently my parents, Margie and Phil, for their

unconditional love and support, and for providing me with the opportunities to pursue any dream I wish. I would like to thank my wife, Cassandra; my inspiration, my voice of reason, my love, and the most intelligent person I've ever met. Thank you for putting up with me throughout the writing of this dissertation; I cannot wait to start a new chapter of our lives together. And finally, for my future child, I thought about you a lot while writing this dissertation. I dedicate this work to you; I cannot wait to meet you.

Table of Contents

Chapters

List of Tables

List of Figures

List of Abbreviations

aa	Amino acid	**IDT**	Integrated DNA Technologies
AF	AlphaFold	**IM**	Inner Membrane
ATP	Adenosine Triphosphate	**IMAC**	Immobilized Metal Affinity Chromatography
BCA	Bicinchoninic Acid	**Imid**	Imidazole
BSA	Bovine Serum Albumin	**IMS**	Ion Mobility Spectrometry
CD	Circular Dichroism	**Inc**	Inclusion
CE	Collision Energy	**IPTG**	Isopropyl-β-D-thiogalactoside
CIP	Calf Intestinal Phosphatase	**JCSG**	Joint Center for Structural Genomics
CIU	Collision Induced Unfolding	**kb**	Kilobase pairs
Cryo-EM	Cryogenic Electron Microscopy	**Km**	Kanamycin
Da	Dalton	**LB**	Luria Bertani
DNA	Deoxyribonucleic acid	**LLPS**	Liquid-Liquid Phase Separation
Eex	Entry Exclusion	**LPS**	Lipopolysaccharide
EOM	Ensemble Optimization Methods	**MALS**	Multi-Angle Light Scattering
EI	Exclusion Index	**MCS**	Multiple Cloning Site
ESI	Electrospray Ionization	**MCSG**	Midwest Center for Structural Genomics
fin	Fertility Inhibition	**MDR**	Multiple Drug Resistant
FPLC	Fast Protein Liquid Chromatography	**MES**	2-(N-morpholino)ethanesulfonic acid
GSH	Reduced Glutathione	**MIR**	Multiple Isomorphous Replacement
GST	Glutathione S-Transferase	**MPD**	2,4-methyl-pentanediol
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid	**Mpf**	Mating pair formation
HDX	Hydrogen-Deuterium Exchange	**Mps**	Mating pair stabilization
Hfr	High frequency of recombination	**MR**	Molecular Replacement
HGT	Horizontal Gene Transfer	**MS**	Mass Spectrometry
IDR	Intrinsically Disordered Region	**MWCO**	Molecular Weight Cut-Off
m/z	Mass-to-Charge Ratio	**rpm**	Rotations per minute

NaAc	Sodium Acetate
NEB	New England Biolabs
NTA	Nitrilotriacetic acid
OD	Optical Density
OM	Outer Membrane
oriT	Origin of Transfer
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDB	Protein Databank
PEG	Polyethylene Glycol
PMSF	Phenylmethylsulfonyl fluoride
pLDDT	Predicted Local Distance Difference Test
PNT	Protein Nanotube
PPI	Protein-Protein Interaction
PSP	Phase Separation Predictor
RB	Running Buffer
RF	RoseTTAFold
R_g	Radius of Gyration
R_h	Radius of Hydration
RT	Room Temperature
RNA	Ribonucleic acid
SAD	Single-Wavelength Anomalous Diffraction
SAXS	Small Angle X-ray Scattering
SDS PAGE	Sodium Dodecyl Sulfate Polyacrylamide gel electrophoresis
SEC	Size Exclusion Chromatography
Sfx	Surface Exclusion
Spp.	Several Species
T_m	Melting Temperature
TM	Transmembrane
Tris	Tris(hydroxymethyl)aminomethane
T4P	Type IV Pilus
T4SS	Type 4 Secretion System
UV	Ultraviolet
Vir	Virulence

Publications Associated with the Research Described in this Dissertation

Publications Associated with Chapter I

Apostol, A.J., Bragagnolo, N.J., Rodriguez, C.S., & Audette, G.F. 2024. Structural insights into the disulfide isomerase and chaperone activity of TrbB of the F plasmid type IV secretion system. **Current Research in Structural Biology.** 8, 100156. https://doi.org/10.1016/j.crstbi.2024.100156

Bragagnolo, N., Rodriguez, C., Samari-Kermani, N., Fours, A., Korouzhdehi, M., Lysenko, R., & Audette, G.F. 2020. Protein Dynamics in F-like Bacterial Conjugation. **Biomedicines**. 8(9), 362. https://doi.org/10.3390/biomedicines8090362

Audette, G.F., Yaseen, A., Bragagnolo, N., & Bawa, R. 2019. Protein Nanotubes: From Bionanotech towards Medical Applications. **Biomedicines**. 7(2),1-46. https://doi.org/10.3390/biomedicines7020046

Shala-Lawrence, A., Bragagnolo, N., Nowroozi-Dayeni, R., Kheyson, S., & Audette, G. F. (2018). The interaction of TraW and TrbC is required to facilitate conjugation in F-like plasmids. **Biochemical and biophysical research communications**. 503(4), 2386-2392. https://doi.org/10.1016/j.bbrc.2018.06.166

Bragagnolo, N., Rodriguez, C & Audette, G.F. The Structural Ensemble of the Conjugative F-like T4SS. *Advances in Medical Biochemistry, Genomics, Physiology and Pathology*, R. Bawa, E.H. Chang, G.F. Audette, A. Diwan and S.A. Faiz (eds.) **Jenny Stanford Publishing**, Singapore, Chapter 16, pp 439-456 (2022).

Publications Associated with Chapter II

Bragagnolo, N. & Audette, G.F. 2022. Solution characterization of the dynamic conjugative entry exclusion protein TraG. **Structural Dynamics.** 9, 064702. https://doi.org/10.1063/4.0000171

Bragagnolo, N. & Audette, G.F. Ensemble modeling of small angle X-ray scattering data reveals the flexible nature of interacting regions in the bacterial conjugation protein TraG. **Biochemical and biophysical research communications.** *Manuscript in Preparation.*

Publications Associated with Chapter III

Bragagnolo, N. & Audette, G.F. 2024. The 1.3 Å Resolution Structure of the Truncated Group Ia Type IV Pilin from *Pseudomonas aeruginosa* strain P1. **Acta Crystallographica Section D. Structural Biology.** 80(12), 834–849. https://doi.org/10.1107/S205979832401132X

CHAPTER I: INTRODUCTION

1.1 Topics of this Dissertation

Bacteria are under constant evolutionary pressure to compete and survive in their microenvironments. Secretion systems are ubiquitous in the bacterial kingdom due to the advantages they provide; examples of secretion system functions include enhanced motility and surface adhesion, secretion of virulence factors, production of biofilms and the ability to transmit DNA between members of a colony[1]. Novel secretion systems are perpetually being discovered as they emerge in native bacteria due to the evolutionary advantage that they provide[2,3]. A thorough understanding of the structures and dynamic movements of components in bacterial secretion systems will aid in uncovering their mechanism of function[4]. This would advance our understanding of bacterial commensalism and antagonism, and it has the potential to aid in the development of inhibitors to bacterial pathogenesis by disrupting the tools virulent species use for stable, long-term infections[3,5]. This is especially relevant in gram-negative bacterial pathogens which have an additional cell membrane, making them difficult to target with antibiotics[6].

This dissertation describes research on the structures of two proteins from two different gram-negative bacterial secretion systems. It is separated into 6 chapters; Chapter I is an introduction which summarizes the relevant topics pertaining to the other chapters. Chapter II involves the biophysical characterization of TraG, a protein from the F-like conjugative type 4 secretion system responsible for entry exclusion. This project examines the structural dynamics of TraG truncation mutants from the F and R100 plasmids using an array of biophysical techniques to characterize regions of the protein[7]. The second research project is described in Chapter III: The 1.3Å Structure of the Group Ia Pilin from *Pseudomonas aeruginosa* strain P1. This chapter describes the structural solution and data refinement of the pilin protein from the Type IV pilus,

related to the type 2 secretion system, and compares its structure to other known pilins[8]. Chapter IV is a summary of pertinent findings and an aggregate conclusion, Chapter V displays References, and Chapter VI is the Appendix.

1.2 Bacterial Secretion Systems

Bacteria use a variety of mechanisms to cooperate, compete, and subsist in their microenvironments such as invading host cells, transmitting signals within a colony, disseminating genetic information, and evading immune systems[1,3,9]. The prominent apparatus which operates these behaviours in bacteria are secretion systems, multi-component macromolecular machines ubiquitous to the bacterial kingdom capable of endowing a wide array of functions. Secretion systems are organized based on their secretion mechanism and the type of molecule secreted, of which currently discovered there are 12 types ranging from Type 0 to Type XI[1,3,10]. Of these, secretion systems 0 to VI are widespread in gram negative bacterial species, with many secretion systems having multiple subtypes and exceptions to their classification, and with analogous secretion systems in gram positive bacteria (**Figure 1.1**). Thus, the diversity of secretion systems and continual discovery of new functions, components and bacterial species has made classification of bacterial secretion systems an ever-evolving process[3,9].

Secretion systems are labeled as type 0 when there are non-proteinaceous molecules responsible for the secretion of a substance from bacteria, such as in the case of outer membrane (OM) vesicles[1,3,11]. Type I secretion systems (T1SS) are found in gram negative bacterial species, which are those with a cell wall composed of an inner phospholipid membrane surrounded by a periplasm composed of peptidoglycan and an outer phospholipid membrane[1,12]. T1SS involve a one-step secretion mechanism which involves the use of three structural proteins: an ATP Binding

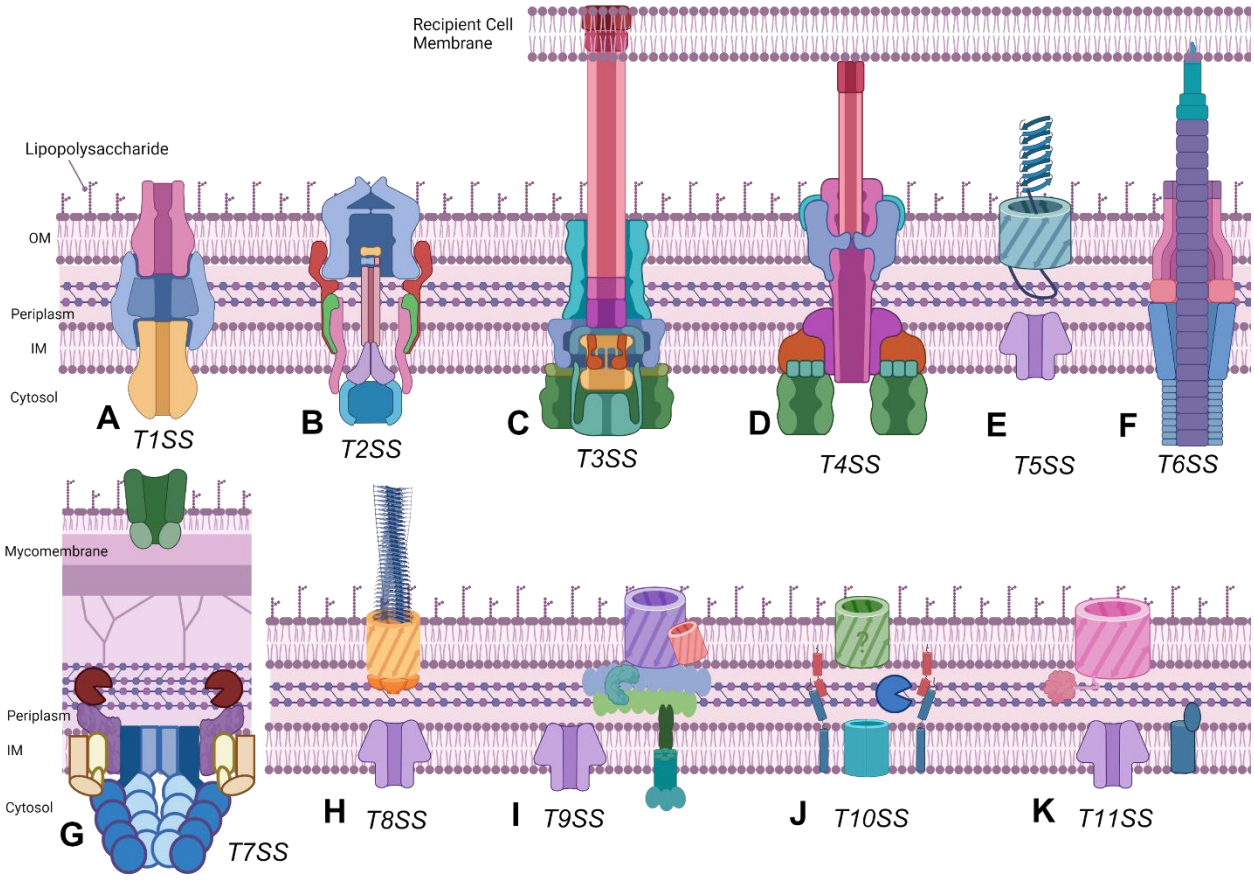
Recipient Cell Membrane
Lipopolysaccharide
OM
Periplasm
IM
Cytosol
A T1SS
B T2SS
C T3SS
D T4SS
E T5SS
F T6SS
Mycomembrane
Periplasm
IM
Cytosol
G T7SS
H T8SS
I T9SS
J T10SS
K T11SS

Figure 1.1: Bacterial Secretion Systems, from types I-XI as labeled A-K. T1SS, T3SS, T4SS and T6SS have one step secretion mechanisms, whereas all others are two step secretion mechanisms; the T5SS, T8SS, T9SS and T11SS use a Sec IM protein/protein complex (magenta) for secretion of the substrate to the periplasm[13]. These secretion systems are depicted as known in gram negative bacterial cell walls, except for the T7SS shown in the mycobacterial cell membrane, which contains peptidoglycan and arabinogalactan in its periplasm and a complex capsule known as the mycomembrane[14,15]. **A**) The hemolysin T1SS is a model secretion system and consists of HlyB in yellow as the ABC transporter in the IM, HlyD in blue as the membrane fusion protein, and TolC in pink as the OM factor[16]. **B**) Proteins shown are GspC-L of the T2SS and compose the four subassemblies[17,18]. The T3SS depicted in **C**) is in the configuration when primed for effector secretion, however the T4SS in **D**) is shown after pilus attachment; retraction and membrane fusion must occur prior to substrate secretion[4,19,20]. **E**) depicts the T5SS subtype A which is considered the simplest secretion system as it merely has a Sec transporter and an autotransporter coloured in light blue which is shown with a passenger molecule, a β-helix fold linked to the periplasmic side of the β-barrel autotransporter[21,22]. The T6SS shown in **F**) displays an elongated contractile tail made of the purple hexameric Hcp polymer with the VgrG spike protein as the effector tip in teal to puncture the recipient cell; the sheath of the T6SS is shown in the cytoplasm in blue with the base plate in the IM above it[23,24]. **G**) The T7SS require mycosin proteases (red) linked to iEccB5 (purple) to process PE/PPE effectors as they are brought into the specialized cell wall of mycobacteria[14,15,25]. The T8SS is shown secreting an amyloid curli in **H**), and the T9SS shown in **I**) has the Sov/SprA translocon in the open conformation (blue c-shaped molecule and the purple β-barrel) where effectors will then be transferred to the PorV shuttle (pink β-barrel) where they will either be released or anchored to the cell surface[26–28]. **J**) The mechanism of secretion out of the periplasm for T10SS is not yet known, it is predicted to occur via a Tol-like OM factor (green) or through membrane disruption[29]. The IM holin is shown in cyan and the peptidoglycan hydrolase is shown in blue with neighbouring anchor proteins ChiY (dark blue) and ChiZ (red). **K**) The OM protein of the T11SS (pink) is of the DUF560 family for transporting lipidated cargo, as processed by the IM localization of lipoproteins (Lol) complex (dark blue)[2,10]. Figure created with BioRender.com.

Casette (ABC) transporter in the inner membrane (IM) which is bridged to a membrane fusion protein that passes the secreted molecule to the OM factor (**Figure 1.1A**). The T1SS allows for the secretion of ions, carbohydrates, small molecules, peptides and in some cases larger proteins. Many bacterial pathogens use T1SS to secrete virulence factors such as toxins and adhesins.

T2SSs are a conserved group that transport folded proteins from the periplasm into the extracellular environment[1,17,18]. The T2SS is composed of four subassemblies: the OM complex, the IM platform, the secretion ATPase, and the pseudopilus (**Figure 1.1B**). The secretion system itself is found only in the OM and requires protein delivery into the periplasm via the Sec or Tat secretion pathway prior to secretion. As the T2SS secretes folded protein substrates, proteins must be shaped by additional factors in the periplasm prior to export through the T2SS. There is a broad array of factors secreted by the T2SS, and in some species the T2SS secretes multiple substrates, while in others only a single protein can be transported. These secreted proteins can have a range of biological functions and can be enzymes such as proteases, lipases, phosphatases, and proteins that process carbohydrates, or virulence factors. The T2SS is part of the type IV pilus (T4P) superfamily; T4P are filaments found on bacterial cell surfaces important for virulence in many human pathogens and are evolutionarily related to the T2SS pseudopilus with proteins that are structurally similar[30–32].

T3SSs are described as injectosomes due to their structure and function; they secrete proteinaceous substrates – effector proteins – across the IM and OM of bacteria and transport them across a target eukaryotic cell membrane[1,19,20]. They are composed of a core of nine proteins that are highly conserved among all known systems, with an additional 10 to 20 proteins that play an important role in their function (**Figure 1.1C**). The T3SS is evolutionarily related to flagellin; eight of the core proteins are found in the flagellar apparatus. Pathogenic species can typically excrete

a few effector proteins using T3SSs, however some species can secrete more than 30 different effector proteins, all of which contribute to the virulence of the pathogen by remodelling the normal cellular functions of the host cell or mammalian tissue.

The T4SS is the most pervasive secretion system in prokaryotes; natural selection has favoured conjugation as donor-mediated horizontal gene transfer (HGT) is a necessity for survival based on the requirement for rapid evolution in competitive environments[4,6,33,34]. The T4SS is a large and dynamic multi-protein complex; in gram-negative bacteria it spans the entirety of the cell wall (**Figure 1.1D**)[34–37]. T4SSs can secrete single proteins, protein-protein and protein-DNA complexes into a wide range of target cells, including other bacteria and eukaryotic cells. Despite the diversity in the T4SS substrates and functions, all T4SS are evolutionarily related and therefore operate in a similar pathway and share common core components. They feature a pilus that extends via polymerisation of secreted pilin monomers to attach to a neighbouring recipient cell, which then depolymerizes at the base to retract the pilus and bring the cells in proximity prior to transfer of a biomolecule. In *Legionella pneumophila*, *Brucella suis*, and *Heliobacter pylori* the T4SS is used to aid in infection by translocating effector proteins into host cells, while the T4SS of *Agrobacterium tumefaciens* is used to transfer oncogenic DNA into plant cells[1,38]. The most common T4SSs in bacterial species function in the interbacterial transfer of a circularized double-stranded DNA, a conjugative plasmid which encodes the T4SS[4,34,39]. As these plasmids can perform recombination with the bacterial chromosome and can harbour virulence genes such as antibiotic resistance genes, the conjugative T4SS allows for the dissemination of virulence genes horizontally in a host-controlled manner, thereby increasing the rate of bacterial evolution.

The T5SS is unique in that the protein substrate which is transferred to the OM of the bacteria has a β-barrel that forms a channel to allow for either the remainder of the β-barrel protein

or a separate protein to be secreted (**Figure 1.1E**)[1,21,22]. There are three classes of T5SSs based on the type of and number of proteins involved in the secretion process: autotransporter secretion, two-partner secretion and chaperone usher secretion. T6SSs differ from T5SS in that they share structural homology to phage tails and transport effector proteins in a contact-dependent manner[1,24,40]. They are made up of large multiprotein complexes which perform secretion in five stages: membrane complex assembly, baseplate complex positioning, elongation of the contractile tail, contraction and sheath disassembly and effector translocation (**Figure 1.1F**). The T6SS protein effectors are different from those of the T5SS in that many of them are important for competition with other bacterial species for a host niche, rather than in virulence of a host mammalian cell as is common for T5SS. There has been some evidence to suggest that T6SS effectors can aid in bacterial inter-strain communication in a process akin to quorum sensing (QS)[41].

Some gram-positive bacterial species such as *Mycobacteria* spp. have a specialized cell wall called a mycomembrane, which is heavily lipidated[1,14,15]. These species have evolved a specialized T7SS to transport effector molecules with roles in pathogenesis, such as the model proline-glutamate/proline-proline-glutamate (PE/PPE) effectors, and in some cases to perform conjugation (**Figure 1.1G**)[25]. The T7SS has five core structural components, four of which are transmembrane (TM) in the IM and interact with chaperones in the cytoplasm and periplasm.

Secretion systems of type VIII, IX, X, and XI all secrete proteins exclusively; however, they differ in the method of secretion and the function of the secreted polypeptide. The T8SS is found in gram negative bacterial species, namely *Proteobacteria* and *Bacteroidetes*, and secrete curli fibrils, an amyloid important in biofilm formation and host colonization (**Figure 1.1H**)[26]. T9SS can secrete protein effectors that can be either delivered extracellularly or anchored to the

membrane surface, which are important in pathogenesis, metal acquisition and carbohydrate degradation (**Figure 1.1I**). They are distributed in the phylum *Bacteroidota* and use a motor complex and a translocon to transport proteins[27,28,42]. The T10SS is a specialized secretion system used by gram negative bacteria to secrete hydrolytic enzymes and toxins; the model T10SS is found in the *Serratia* genus and is involved in the secretion of chitinases (**Figure 1.1J**)[29]. The system is best characterized by its use of a holin membrane protein paired with a peptidoglycan hydrolase to secrete the effector across the OM in a two-step pathway. The T11SS is the most recently discovered and is similar to the T2SS in that it is capable of transporting lipoproteins and soluble proteins responsible for facilitating bacterial symbiosis[2,10]. The surface lipoprotein assembly modulator (Slam) mechanism of lipoprotein surface tethering first identified in *Neisseria meningitidis* was determined to be secreted by an OM protein with a β-barrel, originally named "domain of unknown function 560", which was later re-categorized as the first discovered T11SS (**Figure 1.1K**).

There is a vast wealth of evolutionary, structural and functional information to be gleaned from studying bacterial secretion systems. Emerging discoveries regarding interactions between different secretion systems displays the need for a thorough understanding of the functions of all components of these systems to describe mechanisms underlying bacterial symbiosis, commensalism, antagonism, virulence and host pathogenesis[3,9].

1.3 Bacterial Evolution by Horizontal Gene Transfer

Prokaryotes reproduce asexually through binary fission; vertical gene transfer to daughter cells provides the bacterial kingdom with a rapid rate of evolution relative to most eukaryotes[43,44]. However, horizontal gene transfer (HGT) enhances the rate of bacterial evolution such that they

can quickly adapt to new microenvironments by increasing the plasticity of the bacterial genome[4,33,45,46]. Routes of HGT include transformation through the import of extracellular DNA, viral transduction through the cellular entry of bacteriophage-packaged genes, and conjugation (**Figure 1.2**). These mechanisms are utilized by bacteria to acquire and integrate novel genes into their host chromosome through homologous recombination. If these integrated genes provide bacterial species a selective advantage, HGT often increases the rate of dissemination of these virulence genes[46–49]. Extrachromosomal DNA such as conjugative plasmids can have broad species-host ranges and do not require gene transposition events for proper gene function, thus amplifying the propagation of virulence genes in bacterial communities[50–52]. Conjugation is enabled by a T4SS transcribed and translated from the transfer (*tra*) or virulence (*vir*) gene region of a conjugative plasmid; plasmids typically carry all the genes required for HGT during conjugation from the donor to the recipient cell, and for their maintenance during vertical transfer[6,52,53]. As genes enhancing bacterial virulence, including antibiotic resistance genes, can integrate into conjugative plasmids, conjugation is considered to be the prominent contributor to the spread of virulence genes in bacterial pathogens[54–56].

Multidrug-resistant (MDR) bacterial pathogens, classified as strains that exhibit a phenotype with resistance to three or more antibiotic classes, are a substantial threat to public health as the number of accessible antibiotics that can effectively eliminate infections of MDR species are perilously limited[57]. Genes conferring antibiotic resistance evolve and persist as a result of selective pressures in bacterial populations; rapid cellular reproduction and the plasticity of the bacterial genome attributed to HGT methods enhances their propagation through colonies[45]. These antibiotic resistance genes include those that code for drug efflux pumps, antibiotic inactivating enzymes, biofilm formation, and many other virulence factors[58]. Conjugation is the only route of

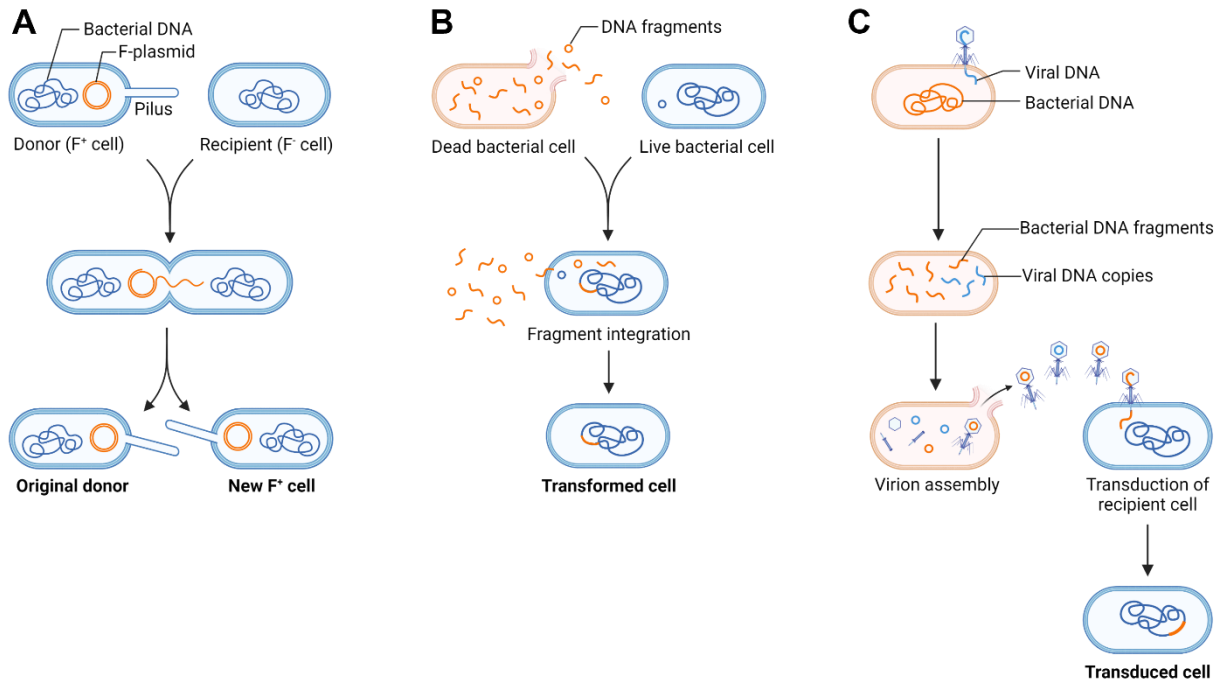

Figure 1.2: Routes of Horizontal Gene Transfer in Bacteria. Novel genes can be acquired by bacteria horizontally through **A**) conjugation mediated by a donor cell carrying a conjugative F-plasmid (F^+ cell), **B**) transformation of extracellular DNA, and **C**) virus-mediated transduction of DNA[4,33,45,46]. Figure created with BioRender.com and adapted from a BioRender template by Parker I.M. and Ona S.

HGT where genes are transferred by bacteria in a donor-controlled fashion; mobile genetic elements such as plasmids or chromosomally integrated conjugative elements (ICEs) are replicated in a donor cell and transported into a recipient cell using a T4SS[4,59]. As conjugation systems are ubiquitous in the bacterial kingdom, inhibition of plasmid conjugation has been suggested as a tool to prevent the spread of antibiotic resistance genes, thus preventing the proliferation of MDR pathogens[5,60,61].

1.4 Bacterial Conjugation in Gram Negative Bacteria

The process of bacterial conjugation is complex, and the mechanisms of the many stages involved in the transfer of conjugative DNA from a donor to a recipient cell are poorly understood

as they vary in gram-negative and gram-positive bacterial species and in different plasmid families[36,56]. Plasmids code for the genes required in the assembly of a T4SS and the transfer of the plasmid DNA; in incompatibility (Inc)F-type plasmids these genes are located in the *tra* operon, and expression is regulated by proteins and RNA upstream of and within the operon[62,63]. Once the transfer proteins (Tra and Trb) from the T4SS are expressed and assembled, pilin processing occurs, followed by pilin assembly and extension[64]. After a recipient cell has been contacted, pilus retraction proceeds, and in some systems such as the F-like T4SS, this is followed by a mating pair stabilization step which allows for the close contact of cell membranes[65]. Exclusion events then occur to prevent redundant transfer of conjugative DNA to a recipient cell that has the same plasmid[7,66–68]. A mating bridge is formed between the cells by the fusion of lipid bilayers in the OM to allow for the transfer of the plasmid by a Tra protein complex called the transferosome[69–71]. The conjugative plasmid is then replicated simultaneously in both cells via rolling circle replication, following which the mating junction closes forming two conjugative donor cells. In summary, the role of the conjugative T4SS is to process and secrete self-polymerizing pilin monomers to form a pilus that will extend to contact a neighbouring recipient cell and depolymerize at the base to retract the pilus to bring the cells in proximity, allowing for cell membrane fusion and DNA transfer[4] (**Figure 1.3A**). The T4SS of the F plasmid and its conjugative process will be detailed in the following sections.

1.4.1 Initiation of Conjugation and Pilin Processing/Pilus Tip Assembly

The transcription of genes coding for conjugative T4SSs is strictly regulated as its production and subsequent operations are metabolically costly processes for bacterial cells[72,73].

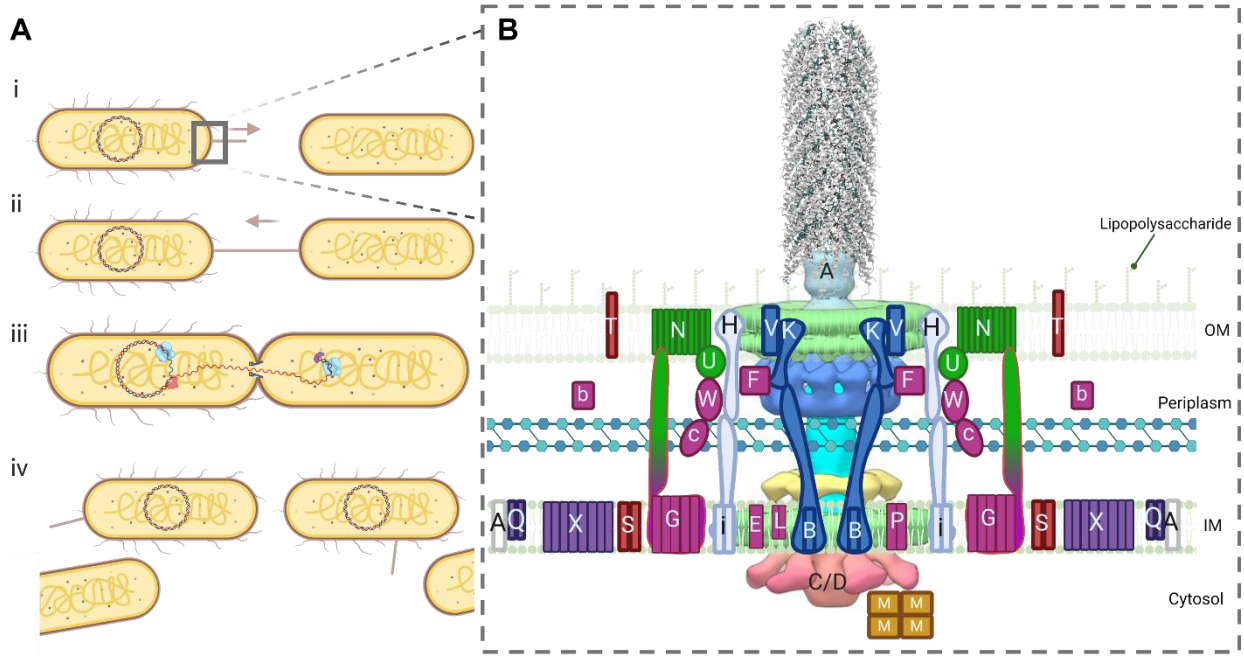

Figure 1.3: Bacterial Conjugation and the F-like T4SS. **A**) (i) Bacteria harbouring plasmids coding for conjugative T4SS are piliated and will repeatedly extend and retract conjugative pili to contact a neighbouring bacterium[53]. (ii) When the pilus tip recognizes a neighbouring cell wall, retraction occurs to bring the cells in proximity. (iii) After surface and entry exclusion steps, a mating junction forms resulting in the fusion of the lipid bilayers. The synchronous transfer and replication of a single strand of nicked plasmid DNA from the donor to recipient is then mediated by the relaxasome (pink) and transferosome (purple), respectively. Rolling circle replication by DNA replication machinery (light blue) in the recipient allows for synthesis of the complimentary strand of the plasmid DNA. (iv) Bacterial conjugation ends after the single stranded (ssDNA) has been transferred to the recipient cell and the mating junction has closed. The recipient cell is now a host cell for the conjugative plasmid. **B**) A model of the F-like T4SS based on available structural information, including a cryo-ET model of the pED208 core complex and a cryo-EM model of the assembled pED208 pilus (PDB ID: 5LEG)[37,74]. Tra proteins are labeled with capital letters and Trb proteins are shown with lower case letters. Proteins are colored based on function; the pilin TraA is white, pilin processing proteins are purple, proteins responsible for pilus assembly/extension are colored in fuchsia, the core complex is colored in dark blue, while pilus retraction proteins are colored in light blue, green colored proteins are responsible for mating pair stabilization, and the red proteins are responsible for exclusion. ATPase proteins TraC and TraD are complexed in the cytosol based on the cryo-ET image but are present in the IM as well. The OM complex in blue is composed of TraK and TraV in an outer core complex that creates 13-fold symmetry, with TraB complexed in the inner core with 17-fold symmetry (not shown in image)[75]. Figure created with BioRender.com[4].

The expression of conjugative systems is controlled by basal levels of constitutive gene expression or signaling molecules (such as autoinducers or sex pheromones used in QS), or both[76]. The

regulation of conjugation in plasmids from the IncF family of gram-negative bacteria occurs solely by the former process and has been termed fertility inhibition (fin)[77]. In the canonical F-plasmid, the gene coding for the activator of the *tra* operon expression, *traJ*, is controlled by the *finOP* regulatory genes. TraJ expression is downregulated by the binding of *traJ* transcripts with antisense RNA *finP*. The protein FinO binds and protects *finP* RNA from RNase E degradation and promotes duplex formation between *finP* and *traJ* RNA in a predicted kissing motif[77–83]. This antisense RNA control mechanism reduces F transfer 10–2000 fold[84]. Many plasmids from the F-like family have an insertion sequence 3 (IS3) element in the *finO* gene and low constitutive expression of FinP; therefore, expression of TraJ is not inhibited, resulting in high constitutive expression of the *tra* operon[77].

While these proteins are not explicit components of the T4SS, their function is intimately linked to the regulated production of their respective conjugative system[77–82]. Therefore, novel antibiotics could be developed to interact with these proteins, transcripts, or genes to prevent or overexpress the production of the T4SS. Such interactions could result in the reduction of the spread of virulence genes by decreasing the expression of conjugative T4SS proteins or cause cell death through overproduction of the large macromolecular complex[72,73].

In bacterial mating, conjugative pili are assembled and extended from the T4SS to bind a neighbouring cell, then retracted to bring the potential recipient cell in close proximity to the donor cell such that a mating bridge can be formed (**Figure 1.3B**)[85]. The first step in pilus production is the post-translational processing of propilin subunits. Within the F-like T4SS, proTraA is synthesized with a leader sequence and is embedded within the IM using proton motive force in a Sec-independent mechanism[86,87]. The IM protein TraQ interacts with proTraA to ensure stabilization within the membrane, following which the leader sequence is acetylated by TraX at

Ala52 and the preceding N-terminus is cleaved to produce the mature pilin[88]. Mature TraA is then extracted from the IM for pilus assembly.

Five proteins are involved in the formation of the pilus tip: TraE, -C, -G, -K and -L, all of which are also responsible for pilus extension as well[34,89]. TraE is a mainly hydrophobic protein found within the IM, thought to be involved in the process of pilus tip formation as well as the assembly of the IM complex. TraL is a mainly hydrophobic protein localized within the IM thought to limit the number of F pili produced by a cell[64,89,90]. The TraL family of proteins range in size from 91 to 105 aa with a shared homology to VirB3 from the tumor inducing (Ti) plasmid of *A. tumefaciens* and TrbD from the RP4 plasmid[91,92]; TraL from the F plasmid is 91 aa[93]. TraK is part of the pilus stalk, TraC is an ATPase and TraG is multifunctional. These proteins will be discussed in detail in later sections. Mutations in these *tra* and *trb* genes cause various disruptions in the assembly of the pilus, from alterations in length, to completely inhibiting the formation of a pilus. Although these proteins are known to be involved in the assembly and/or extension of the pilus, the structure and function of each is largely unknown.

The pilus from the T4SS is composed of mature TraA subunits, which assemble into a helical structure with an average length of approximately 20 µm, an outer diameter of 8–10 nm, and an inner lumen of 2–3 nm in diameter (**Figure 1.3B**)[36,94]. On average, an *Escherichia coli* cell will have anywhere from 1–5 pili, with some variants having up to 20[34]. Pili can range in length and flexibility; cells expressing rigid pili use these appendages to manifest an aggregate network of rigid polymers, thus, encouraging non-specific clumping of donor and recipient cells to form mating pairs[39]. Some pili, such as the F pilus, are capable of extension and retraction allowing for close contact between a donor and recipient cell for bacterial conjugation. Other pili will use retraction and extension as a manner of maneuverability termed twitching motility[95]. During phage

infection, prior to DNA or RNA being injected into a bacterial cell, a major site of adsorption is the pilus[74,96]. As the pilus is an important aspect in virulence, as well as providing bacteria adaptivity to their environment it is deemed an attractive target for novel therapeutics.

The structure of the F-like pilus has been determined by scanning transmission electron microscopy (STEM), where the subunits were found to be mainly α-helical and the assembled pilus exhibited a 5-fold rotational symmetry[96,97]. A separate cryogenic electron microscopy (cryo-EM) structure provided a higher resolution image of the five-start helical structure composed of ~12.8 TraA pilins per helical turn (**Figure 1.4D**)[74]. To promote assembly of TraA subunits from the base of the pilus and extend the nanotubule, the structural stalk and channel proteins TraB, TraK and TraV are required[98]. These proteins compose the core complex, which has been shown by cryogenic electron tomography (cryo-ET) to form an outer ring with 13-fold symmetry composed of TraV and TraK, with an inner ring composed of TraB with 17-fold symmetry (**Figure 1.4A-C**)[75]. Symmetry mismatches such as these are common in several types of secretion systems and are considered to provide flexibility between different protein complex layers or subassemblies. The role of this asymmetry for the function and mechanism of secretion machinery has not been established for any characterized secretion system.

1.4.2 Pilus Extension and Retraction

Extension and retraction processes are different in conjugative and non-conjugative pili in several key regards, specifically when comparing to the model system, the type IV pilus. Fluorescence microscopy imaging has shown F pili rotate throughout the process of pilus extension

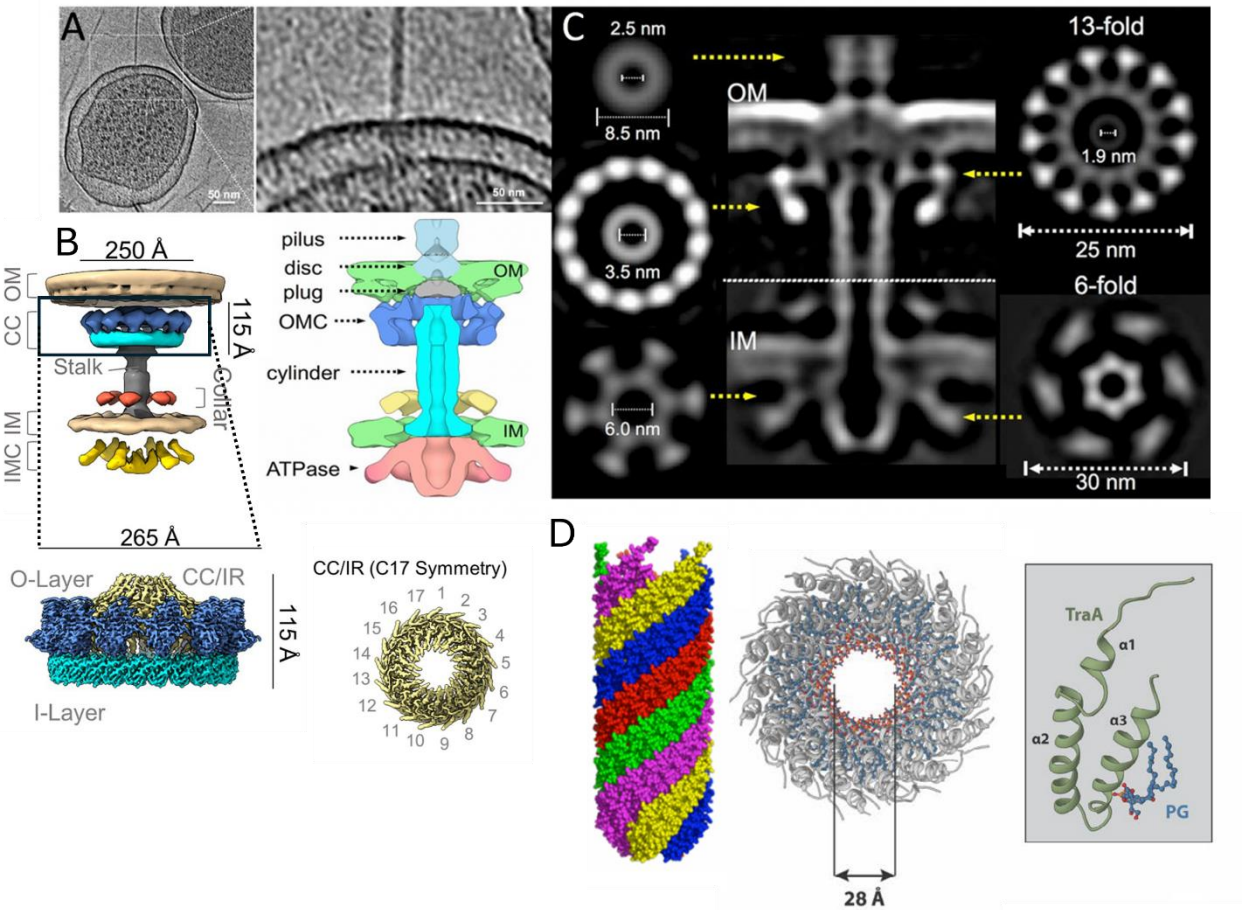

Figure 1.4: The F-like Pilus and T4SS. Figures **A**–**C** are *in situ* structures of the F-like pED208 T4SS by cryo-ET while **D** depicts the overall architecture of the pED208 pili by cryo-EM. **A**) A tomographic slice of an *E. coli* minicell showing an extended F pilus attached to the OM. **B**) A three-dimensional depiction of the pED208 core T4SS from ATPase to pilus. At the OM, the pilus narrows and connects to the disc at the junction where the pilus meets the OM complex; a luminal space approximately the same diameter as the pilus is present at this junction, which is sealed by a plug domain. The core complex has been solved to 3.3 Å resolution by cryo-EM which allowed for a more accurate view of the inner ring composed of TraB and TraV in 17-fold symmetry[99–101]. **C**) A central slice of the averaged structure of the core T4SS with cross-sectional views of the separate domains. The IM complex (the VirB4-like ATPase, TraC) forms a 6-fold rotational domain with a central lumen of ~60 Å with the entire domain spanning 300 Å. The OM complex (the VirB7, -B9, and -B10 homologues TraV, TraK, and TraB) has a 13-fold symmetry and spans 250 Å with a central lumen of 19 Å. **D**) A view of the surface representation of the pED208 pilus, as well as an image of the TraA monomer structure with phosphatidylglycerol (PG) (PDB ID: 5LEG). The pilus is composed of a 5-start helical assembly with each helical strand depicted in a separate color. This 3.60 Å resolution cryo-EM structure aligns with the architecture of the cryo-ET pilus structure in that it has an outer diameter of ~87 Å and has a central lumen of 28 Å. Figures adapted from Hu *et al.* 2019, Bragagnolo *et al.* 2020, and Costa *et al.* 2021[4,37,99].

while type IV pili do not, indicating that the molecular motors involved in the F pilus remain stationary leading to pili rotation (**Figure 1.5**)[102,103]. Furthermore, type IV pili encode two different hexameric adenosine triphosphatases (ATPases), PilF (for extension) and PilT (for retraction), in contrast to the F-like T4SS family in which retraction is considered to occur in an energy-independent manner; the ATPase proteins TraC and TraD are proposed to supply energy to the system for pilus extension and transferosome transport cooperatively[37,103]. This intricate process of assembling TraA pilin monomers requires the aid of many proteins, including TraB, -C, -E, -F, -G, -H, -K, -L, -U, -V, -W, and TrbC[36]. TraF, -H, -U, and -W are specific to the F T4SS, all others have structural and functional homologs in other T4SSs (**Figure 1.6**)[64].

When the pilus contacts and attaches to a neighbouring cell, a signalling event occurs and the retraction of this pilus is performed by subunits TraF, TrbB, TraH, and TrbI[35,36,98]. Many of these proteins have multiple proposed functions; TrbB and TraF both contain thioredoxin-like domains and are thought to be responsible for chaperone activity, and aid in the correct conformational folding of TraH, TraU, and TraN in forming the transferosome complex in the final mating pair formation step[65,104,105]. Pilus retraction is still a major point of intrigue as it is hypothesized that it occurs in an energy independent manner[36]. Once initiated, F-pilus retraction occurs at an average of 15.8 nm/s, which is less than half the mean extension rate of 39.5 nm/s[102]. TrbI can directly interact with TraH, and this interaction is considered to initiate pilus retraction[106]. As TrbI localizes in the IM and TraH assembles into the OM when in the context of the T4SS complex, the TrbI:TraH pair is hypothesized to be part of a second envelope-spanning structure analogous to the TraB, -K, -V core complex[104].

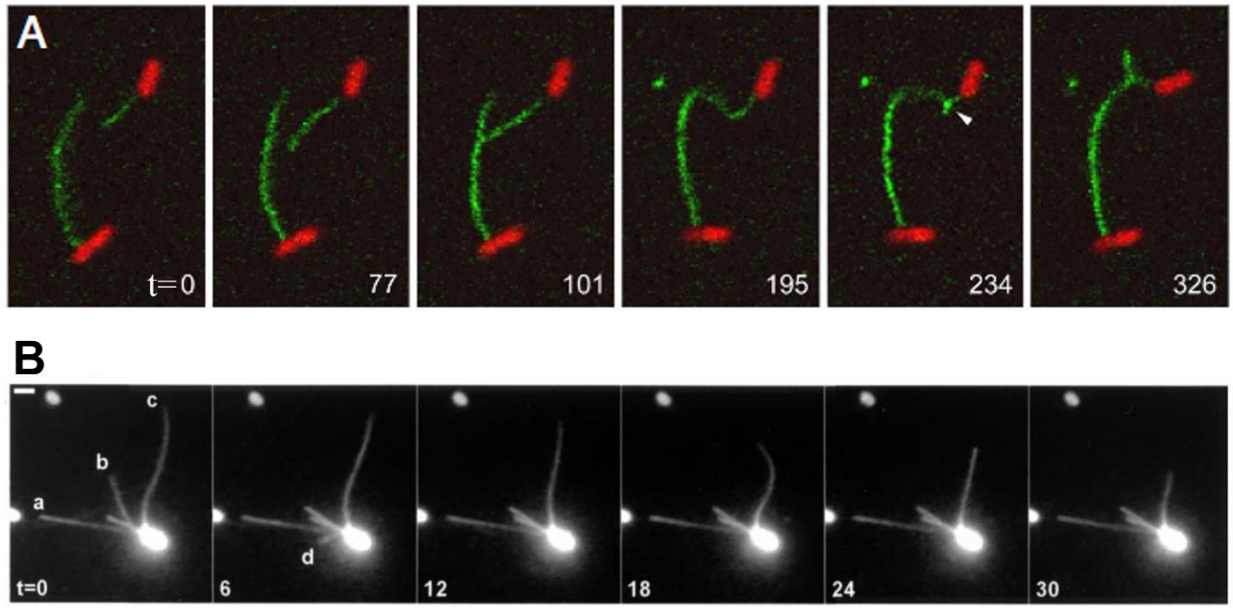

Figure 1.5: Fluorescence Microscopy Images of Pilus Extension and Retraction, demonstrating speed and direction of polymerisation. **A**) displays *E. coli* K-12 strain HfrH cells continuing to extend F-pili after the distal portion has become immobilized by binding to another pilus. As indicated by the arrow, the pili form a supercoil as each pilus continues to elongate, and retraction has been shown to shorten and remove the supercoil. This indicates that the F-pilus rotates as it is polymerized, whereas this phenomenon is not seen for *P. aeruginosa* strain PAK producing type 4 pili shown in **B**), providing evidence that the pilus remains stationary during extension and retraction. The timescale shown in the bottom of each image demonstrates the higher rate of extension/retraction in T4P compared to the F-T4SS. Figures adapted from Clarke *et al.* 2008 and Skerker and Berg 2001[102,103].

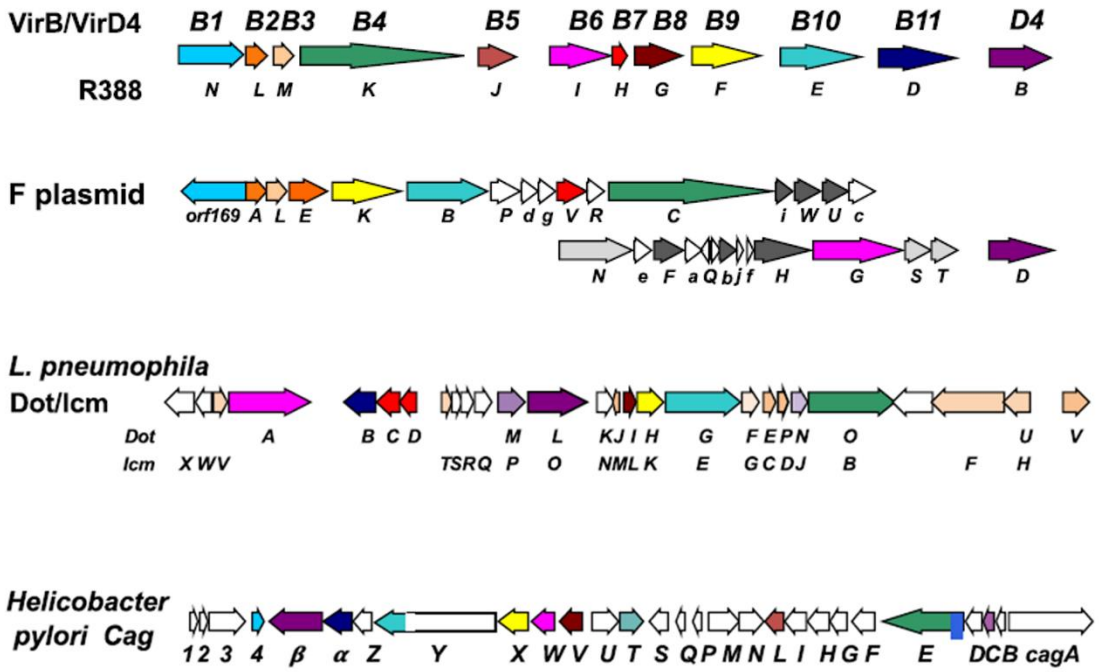

Figure 1.6: Gene Organisation of Operons for Model T4SS Plasmids. VirB/VirD4 from *Agrobacterium tumefaciens* and R388 are two minimal T4SS with similar gene organization and are lacking accessory proteins seen in the three extended systems. Genes with functions specific to each extended system are white, whereas all others are colour coded by similar function. For a more detailed view of the *tra* operon from the F-T4SS with colour coding matching Figure 1.3, see **Figure S1**. Figure adapted from Costa *et al.* 2021[99].

1.4.3 Mating Pair Formation and Mating Pair Stabilization

After a potential recipient cell is brought in proximity to the conjugative donor cell, mating pair formation (Mpf) and mating pair stabilization (Mps) events occur. Two sets of genes are necessary for conjugation to proceed; mobility (MOB) genes which process the conjugative plasmid and will be discussed later, and Mpf genes[107]. In general, Mpf genes are those responsible for the formation of a trans-envelope channel, which is formed to transfer the nucleoprotein complex (**Figure 1.2A**). A mating pair is formed during pilus retraction events when donor and recipient cells are close enough such that they are no longer separated by exocellular material; the OMs then fuse by a currently unknown mechanism[85]. Mpf between donor and recipient cells is

thought to be a complex process that is not well understood for F-like systems[36]. Other well-studied T4SS, such as the P-type T pilus from *Agrobacterium tumefaciens,* have a relatively simple Mpf system, whereas F-type systems have the previously described core T4SS, as well as additional proteins involved in pilus extension and retraction (TraF, -H, -U, -W, and TrbI), and additional Mps proteins, TraN and -G, which confer properties of efficient mating in liquid and on solid media.

Mps occurs after pilus retraction and Mpf events, yet these genes are not often seen outside of the F-like plasmid family; stabilization of contact between donor and recipient cells is essential for the formation of the mating pore and therefore essential for conjugation by F-like T4SSs[36,108]. Other T4SS families have alternative methods for stabilizing a mating pair, such as the production of aggregation proteins, or compensation by the production and recipient-contacting of many pili in IncI plasmids[36,109,110]. The process of Mps for F-like T4SS provides properties of resistance to SDS and shear forces within a mating pair[111]. Unlike other major steps of conjugation, such as pilus assembly and retraction events, Mps is thought to be much simpler as it involves only two proteins, TraN and TraG[112]. These two proteins are predicted to span the entirety of the gram-negative bacterial cellular envelope; TraG contains an IM bound NTD and a C-terminal periplasmic domain that is hypothesized to contact TraN bound in the OM, which has loops extending into the extracellular matrix (ECM)[36,75]. This interaction has been deemed critical for the formation of the mating pore, as both proteins are essential for DNA transfer to occur. The first step in Mps involves interaction between TraN on the OM of the donor cell with OM protein A (OmpA) and lipopolysaccharide (LPS) moieties on the recipient cell to maintain close contact during conjugation[112].

Mps is a multistep process; the mating pore is stabilized as mentioned, then exclusion events occur to prevent redundant donor-donor transfer[36]. The first exclusion event is surface exclusion (Sfx) as mediated by the OM lipoprotein TraT; if the cognate TraT is expressed in the recipient cell's membrane indicating the presence of the same plasmid in the recipient cell, then an interacting partner (currently not established) is theorized to prevent mating pore formation and conjugation will be prevented[67]. If this step fails, entry exclusion (Eex) is relied upon to prevent donor-donor DNA transfer. The IM protein TraS and the periplasmic C-terminal half of TraG are interacting partners in this event; if the potential recipient has the same plasmid TraG in the donor cell is thought to scan the IM of the recipient cell and interact with the cognate TraS to arrest conjugation[113]. If the recipient cell does not have a cognate TraS in its IM, and thus does not have the same plasmid, the C-terminus of TraG is predicted to extend to contact TraN in the OM and further the formation of the mating junction[75,112].

1.4.4 Conjugative DNA Transfer

As part of the conjugation process, the DNA to be transferred is processed through the formation of a nucleoprotein complex using TraM and the relaxase TraI, which nicks circularized plasmid DNA at the origin of transfer (oriT) to form a ssDNA topology. The genes which code these proteins are MOB genes which are common in plasmids transferred through other methods of HGT[107]. TraM binds as a tetramer to aid in DNA unwinding and chaperones the complex to the IM mating pore formed by TraD, a hexameric ring ATPase that pumps the nucleoprotein complex through the pore into the recipient cell when all other T4SS have performed their function for Mpf[114,115]. Rolling circle replication allows for asymmetric replication of the ssDNA in the donor cell, thus reforming the plasmid while the mobile ssDNA is being transferred to the recipient cell[71].

In the recipient cytoplasm the nucleoprotein complex is removed, and the ssDNA is recircularized and replicated, thus converting the recipient cell into a donor for the newly integrated genes[36,116].

1.5 Incompatibility and Exclusion Systems in Conjugative Plasmids

The result of exclusion mechanisms was discovered early in the history of conjugation research; Lederberg *et al*. discovered that cells harbouring the F plasmid were poor F conjugation recipients, and later the inefficient transfer was termed superinfection immunity[117,118]. In attempting to mate cells with a high frequency of recombination (Hfr) conjugative plasmid with recipient cells without an F-exclusion system, less than 5% of the recipient cells would survive compared to canonical F mating[119]. The primary cause of cell death was determined to be extensive damage in the recipient cell membrane resulting in cell permeability due to extensive cell-cell contacts mediated by pili during conjugation, a process named lethal zygosis[67,120]. Metabolic costs due to excessive DNA transfer contribute to lethal zygosis, but in a lesser extent. F-mediated superinfection immunity is attributed to exclusion genes *traS* and *traT*, and the exclusion process is essential in preventing lethal zygosis through excessive mating[119,121].

There exists an association between conjugative DNA exclusion and sex pilus type; plasmids and ICEs carrying the same sex pilus type would be excluded, however plasmids with different sex factors permit conjugation even if the pilus type is identical[122,123]. When conjugative plasmids of the same type are transferred to a donor by transduction, this phenomenon did not occur in cells despite an active Eex system, however even when Eex was abolished it was observed that any second F factor in an F-containing strain would be removed[124,125]. Therefore, it was

concluded that two phenomena contribute to superinfection immunity: plasmid incompatibility and Eex[67,124,126].

1.5.1 Plasmid Incompatibility

Incompatibility in mobilizable elements can be defined as the inability of two plasmids to be propagated stably in the same cell line[127]. It is a manifestation of relatedness; elements such as MOB genes that are involved in plasmid replication control and partitioning are identical, causing malfunction in the inheritance of both plasmids into the daughter cells. Incompatibility is also linked to copy number, as the number of origins of replication of the same grouping affects whether they are replicated[62]. Speed of replication and toxicity of the plasmid also affects the survivability of these plasmids, as outgrowth during replication is theorized as the mechanism for the loss of one plasmid in a similar pair[128]. As incompatibility is a universal property of plasmids, it allows for the formation of a suitable classification scheme[62,129]. The main families are considered to be IncF, IncP, IncA/C, IncH, IncI, and IncN based on the plasmid multilocus sequence typing database[130]. An indeterminate multitude of subfamilies exist due to the continuous discovery and synthesis of novel plasmids.

1.5.2. Surface and Entry Exclusion Systems

Plasmid inclusion families are seen to have differences in their exclusion systems in terms of gene organization and the structure of the encoded proteins[67]. The previously described exclusion system is that of the IncF-like group; the genetic organisation for Eex and Sfx genes on other plasmid inclusion families is depicted in **Figure 1.7**. In comparing the genetic organisation

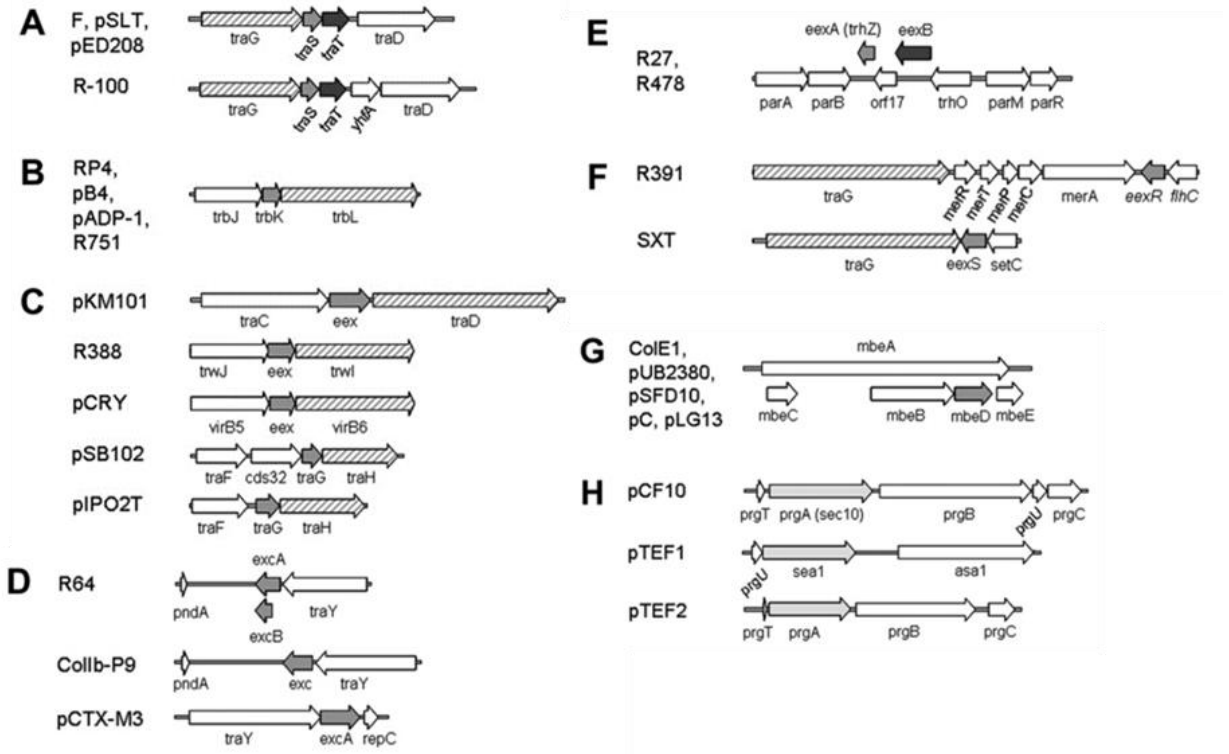

Figure 1.7: Exclusion Genes of Plasmid Incompatibility Groups. The exclusion gene-containing regions are organized differently in each plasmid incompatibility group: **A)** IncF-like; **B)** IncP-like; **C)** IncN-like; **D)** IncI-like; **E)** IncH-like; **F)** ICEs R391 and SXT; **G)** ColE1-like; **H)** Pheromone-responding conjugative plasmids, more commonly found in gram positive bacteria. Arrows corresponding to the direction of transcription in exclusion genes are shadowed depending on the system; exclusion genes of pheromone-responding conjugative plasmids are coloured light grey, genes encoding IM exclusion proteins similar to *traS* are intermediate grey, and genes encoding OM exclusion proteins similar to *traT* are dark grey. *traG* homologs are dashed and genes with functions not related to exclusion systems are filled in white. Figure from Garcillán-Barcia *et al.* 2008[67].

of plasmid groups from gram-negative bacteria, synteny (the physical localisation of genetic loci) is maintained for Eex genes in relation to other transfer-related genes[67,68]. Plasmid families have some differences in their genetic organisation; many have their *traS*-like gene before their *traG*-like gene in the locus, and most plasmids do not have a *traT*-like Sfx gene at all. Only a few groupings lack a *traG*-like gene for exclusion; IncI-like, IncH-like and ColE1-like plasmids only have a *traS*-like exclusion gene, and the pheromone responding conjugative elements have their own type of exclusion system (**Figure 1.7**)[131–135]. The most common plasmids in gram-negative

bacteria are from the inclusion families shown to have both TraG- and TraS- like systems; IncF-like, IncP-like, and IncN-like (includes IncW)[63,136]. There exists the potential for evolutionary selection for TraG/TraS-like Eex systems against TraT-like Sfx systems, as supported by the observed increase in exclusion index (EI) in Eex in comparison to Sfx[67]. EI is calculated as the frequency of transfer of a plasmid given to a plasmid-free recipient divided by the frequency of transfer to a recipient carrying the same plasmid; for the F-plasmid in *E. coli* the EI was 100-300 because mating was observed at 100-300 times higher frequency when the plasmid was not present in the recipient in comparison to donor-donor exchange[111]. In comparing EI of donor-recipient exchanges for the F plasmid, *traT* point mutants showed a smaller reduction in mating than the *traS* point mutants[137]. An increased exclusion activity was seen in the Eex system, which was attributed an EI of 200, while an EI of 20 was associated to the Sfx system, indicating a higher reliance on the Eex system for preventing erroneous donor-donor transfers. The exclusion process performed by these systems is also seen to be gene-dose dependent; when *traS* and *traT* were cloned into a multicopy plasmid the EI increased to 10,000[111].

In comparing *traS* genes for different incompatibility groupings, low sequence homology is observed even in members of the same inclusion family. TraT from the plasmid pED208 ($TraT_{pED208}$) of the IncF family has the lowest inter-family protein homology to $TraT_F$ at 80% aa identity, however $TraS_{pED208}$ has no identity to $TraS_F$ or $TraS_{R100}$[67]. Closely related F and R100 plasmids show a TraS homology of only 17% with no region displaying greater identity[113].

Exclusion genes have proven to be important for the stability of a conjugative plasmid; F donor cells with both *traS* and *traT* mutations are unstable and have not been isolated[111,137]. This is supported by F plasmid-mediated superinfection immunity, which was confirmed to be mediated by TraT and TraS[119]. The presence of TraS was deemed to be more important for colony survival

than TraT as indicated by the described EI from Eex compared to Sfx[67,119]. It is theorized that the disruption of Eex systems would be severely detrimental to bacterial colony survival[67].

1.6 Structure and Function of TraG

1.6.1 TraG and Entry Exclusion

Eex is an important function in conjugative systems to prevent the inefficient transfer of a plasmid from a donor cell to another cell that has the same plasmid[89]. As mentioned previously, after Sfx occurs and cells are proximal, Eex is performed before completion of the mating bridge. In the F-like T4SS superstructure, two IM-bound proteins TraG and TraS are responsible for Eex; TraG in the donor cell is thought to scan the IM of the recipient cell and interact with the cognate TraS to prevent conjugation (**Figure 1.8A**)[113]. If the TraG-TraS interaction does not occur then TraG interacts with TraN in the OM of the donor cell, resulting in Mps and advancing the conjugation process[36,112]. Eex systems of the closely related plasmids F and R100 retain plasmid specificity; TraG present in donor cells will interact with a cognate TraS in the recipient cell to prevent redundant conjugation but will not interact with TraS from a different plasmid[113]. This would allow conjugative DNA transfer between the two donors such that they harbour both conjugative plasmids. Additionally, TraG will not interact with TraS in the same IM, it functions in a trans fashion only.

TraG from the R100 plasmid ($TraG_{R100}$) is a 102.9 kilodalton (kDa) protein with an N-terminal TM domain bound in the IM and a periplasmic CTD from residues Ala452-Glu940 denoted TraG* (**Figure 1.8B**)[113]. It is one of the largest proteins encoded by the *tra* region of F-

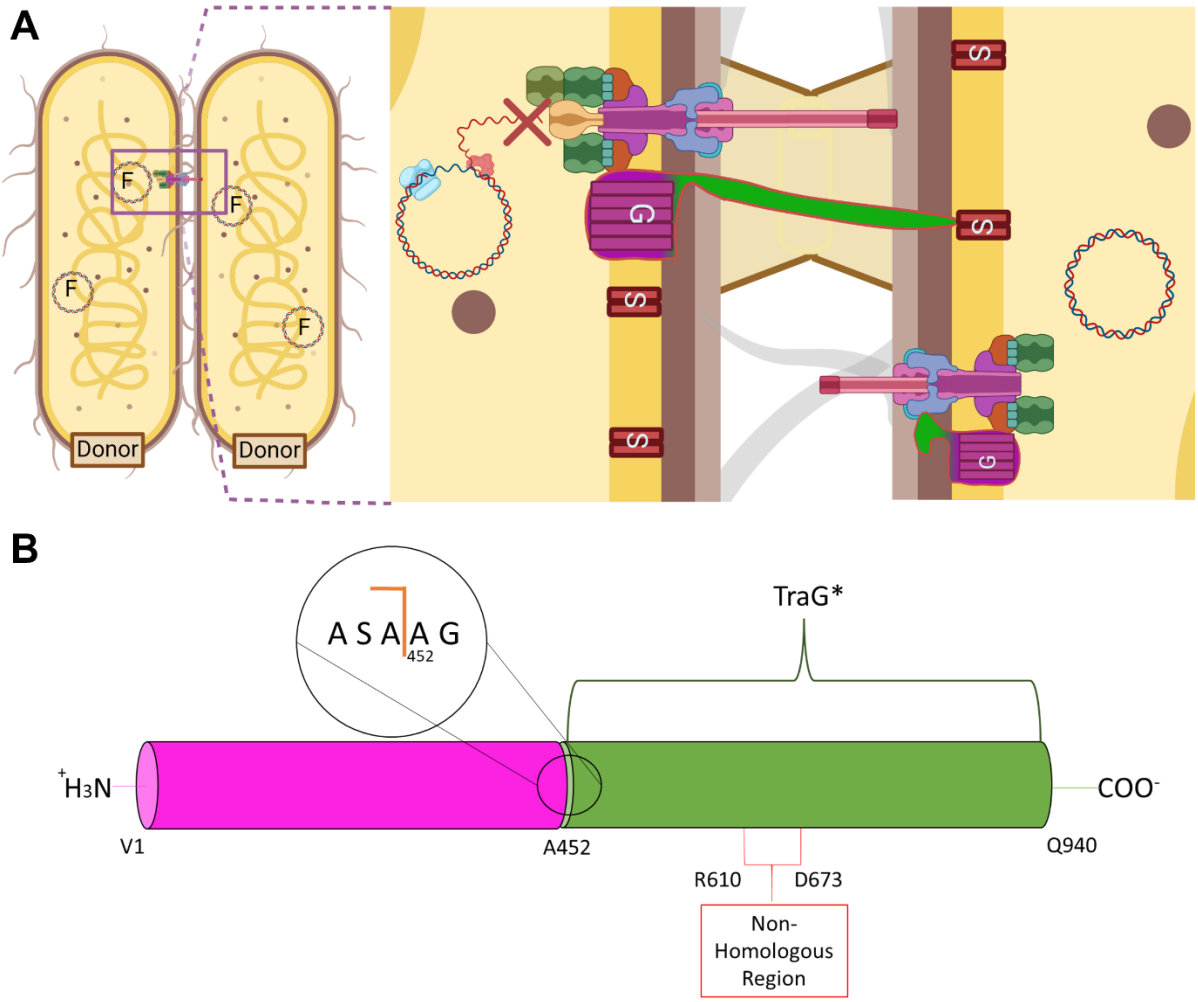

Figure 1.8: The Entry Exclusion Process and a Schematic of TraG. **A)** A simplified depiction of the Eex process in gram-negative bacterial conjugation in F-like T4SSs. In a colony of bacteria carrying plasmids with identical *tra* regions, conjugation between donor cells is prevented through Eex, proposedly in the event of faulty surface exclusion[67]. The formation of a complete mating pore and the transfer of redundant plasmid DNA is prevented through interaction of TraG in the donor cell with a cognate TraS embedded in the recipient's IM. Figure created with Biorender.com. **B)** Generalized domain organization of TraG, with residues and numbering of $TraG_{R100}$. The N-terminal TM region of the protein responsible for aiding in pilus assembly is coloured in magenta and the C-terminal periplasmic region termed TraG* is coloured in green. These regions are defined by a signal peptidase I cleavage site after residue A451 which indicates the beginning of TraG*[113].

like plasmids. Frameshift mutations in the NTD of the protein have been shown to affect the polymerization of pilus subunits, while the whole protein must be intact for Mps and Eex events to occur[113,138]. It was previously thought that the presence of a signal peptidase I cleavage site after

residue Ala451 releases TraG* into the periplasm to fulfill its role in Mps, however it appears the CTD must be anchored to the IM for proper TraG* function[113,139]. TraG from the F ($TraG_F$) and R100 plasmids have an overall sequence identity of 93%, however the region in TraG* from residues 610 to 673 shows only 17%. This region of TraG was shown to be responsible for plasmid specificity in entry exclusion using *in vivo* conjugative mating assays, indicating the role of the region for interacting with their cognate TraS[113].

1.6.2 Proposed Mechanisms for TraG Function

The mechanism for the assembly of pilin subunits into polymers as aided by the N-terminus of TraG is not completely understood; mutations in TraE, -K, -B, -V, -C, -W, -U, -F, -H, and the N-terminal region of TraG all result in accumulation of mature pilin in the IM[139,140]. In performing Eex, TraG is predicted to scan the neighboring recipient IM to interact with TraS, which would require the protein to interact over two bacterial periplasms and OM bilayers. This distance can be approximated to 35–45 nm, assuming donor and recipient OMs are proximal after surface exclusion (Sfx)[141–144]. As this distance is likely unachievable by a single globular protein of which only 488 residues are found in the periplasm, the steps of Mps are thought to occur first. The C-terminal region of TraG is hypothesized to undergo a reversible conformational unfolding and thrust into the mating pore formed after TraN interacts with OmpA of the recipient cell, mediated by the TraG-TraN interaction that is predicted to occur during this Mps step[113,145]. In this theoretical model, the N-terminal region of TraG maintains an intact structure and TraG* is extended through the mating pore to perform its role in Eex. This model would have TraG* cross an approximate distance of 15–25 nm (equating to the space from the top of the donor IM, through the periplasm, the mating pore and the periplasmic space of the recipient cell) to contact TraS in

the recipient cell IM; a more feasible proposal for a protein-protein interaction (PPI)[7,141–144]. A TraS-TraG complex has not, however, been detected in mating cells through crosslinking or immunoprecipitation experiments, nor interaction detected using a bacterial two-hybrid system[113]. As well, an interaction between TraN and TraG has not been detected using similar methods[108]. The likely cause for the lack of detectable assembly is due to the small number of interacting protein partners in a mating pair, as well as the transient nature of these PPIs.

1.6.3 Insights on TraG Structure from Homologs

There is a lack of structural data for TraG from the F-like T4SS family; however, some structural information can be gleaned from other conjugative systems. The canonical T4SS of the IncN type plasmids is the T pilus from *A. tumefaciens*, and is well characterized, largely due to the utility of this system in the genetic modification of plants mediated by the Ti plasmid[38,146]. pCRY, a similar plasmid from the same family, contains the gene product VirB8, a protein with a membrane-bound domain orthologous to the N-terminal region of TraG. There is evidence indicating the evolution of the *traG* gene in the F plasmid involved the fusing of an ancestral homolog of genes *virB6* and *virB8* to produce a single gene[34]. The structure of VirB8 has been solved in numerous species; *A. tumefaciens*, *Brucella suis*, *Bartonella quintana*, *Bartonella tribocorum*, and *Rickettsia typhi*[147–149]. As well, the structure of a VirB8 homolog from the plasmid pIP501 (from IncN family, similar to pIPO2T), $TraH_{pIP501}$, in *Enterococcus faecalis* was determined[150]. These structures feature a nuclear transport factor 2-like fold and show high structural homology despite low sequence homology, with backbone root mean square deviation (RMSD) values ranging between 2.6–2.8 Å in comparing all structures[150]. These VirB8 structures indicate the protein functions as a dimer; however, $TraH_{pIP501}$ is monomeric. A VirB8 homolog

from pKM101, TraE was shown to form hexamers when isolated and when interacting with the VirB6 homolog from this plasmid, TraD[151]. This implies the diversity in oligomerization possible in the TraG-like proteins of these systems. There are no high-resolution structures of VirB6, however structural predictions have shown there are 4–9 TM helices in VirB6 from the Ti plasmid, while the NTD of TraG is predicted to span the membrane 3–5 times[113]. Despite the proposed ancestral relation between the proteins, the function of VirB6 is quite different from TraG as it behaves as an IM channel. There is also low confidence for similarity between the structures of the C-terminal periplasmic domain of TraG and VirB8, as crystal structures of VirB8 have shown the domain assembles into a stack of octameric ring-like structures, which are predicted to function in the conveyance of substrates from the IM complex into the channel housed by the OM complex[34,152]. As TraG* orthologues are commonly associated in oligomers from their IM-bound N-terminal domain, it is difficult to predict whether TraG* from F-like T4SS will oligomerize into its proper complex without the IM-bound N-terminal domain.

1.7 The Human Pathogen *Pseudomonas aeruginosa*

Pseudomonas aeruginosa is the most common pathogen of human and warm-blooded mammals from the group of *Proteobacteria* known as the non-fermenting gram-negative bacteria (NFGNB)[153,154]. As *P. aeruginosa* is a waterborne, obligative aerobic, non-fastidious bacterial species that can survive in a wide temperature range of 4-40°C, it often colonizes the respiratory system of mammals[153,155,156]. It is a well-known opportunistic pathogen of the lungs of immunocompromised patients such as those with healthcare-associated pneumonia, chronic obstructive pulmonary disease (COPD), and cystic fibrosis (CF)[155–157]. *P. aeruginosa* infections are associated with high morbidity and mortality, and with the emergence of many novel MDR

strains found in nosocomial settings it has been included in the "critical" category of the World Health Organization's (WHO) priority list of bacterial pathogens for which research and development of new antibiotics are urgently needed[157]. The prevalence of novel MDR *P. aeruginosa* strains in the healthcare setting has been further exacerbated by the COVID-19 pandemic, which saw many uncharacterized co-infections of *P. aeruginosa*, likely due to overprescription of broad-spectrum antibiotics and the use of ventilators which are common fomites for this species[158–160].

Important characteristics of *P. aeruginosa* which make it a successful and adaptable pathogen include its high metabolic flexibility, which allows it to persist in diverse environments and harsh conditions[153]. *P. aeruginosa* has a large genome (~5MB to ~7MB depending on the strain) characteristic of a classical "secretor", as 3% of open reading frames (ORF) are of secreted virulence factors and proteases[153,157,161]. There is an expansive arsenal of mechanisms of secretion that *P. aeruginosa* use to adapt to various environments; chief among them are T4P, biofilm formation systems (exopolysaccharides, proteins and extracellular DNA), QS systems (Las, Rhl and Pqs), flagellins (FliC and FliD), pyoverdine (PVD) siderophore as an iron uptake system, LPS and outer membrane proteins (OMPs), the T3SS, the T6SS, and the T2SS (**Figure 1.9**)[157].

1.8 The Type 4 Pilus from *P. aeruginosa*

1.8.1 Subtypes of T4P

T4P are responsible for a variety of bacterial functions; they confer the main method of attachment for *P. aeruginosa* by binding to a receptor on host mucosal cells for pathogenicity,

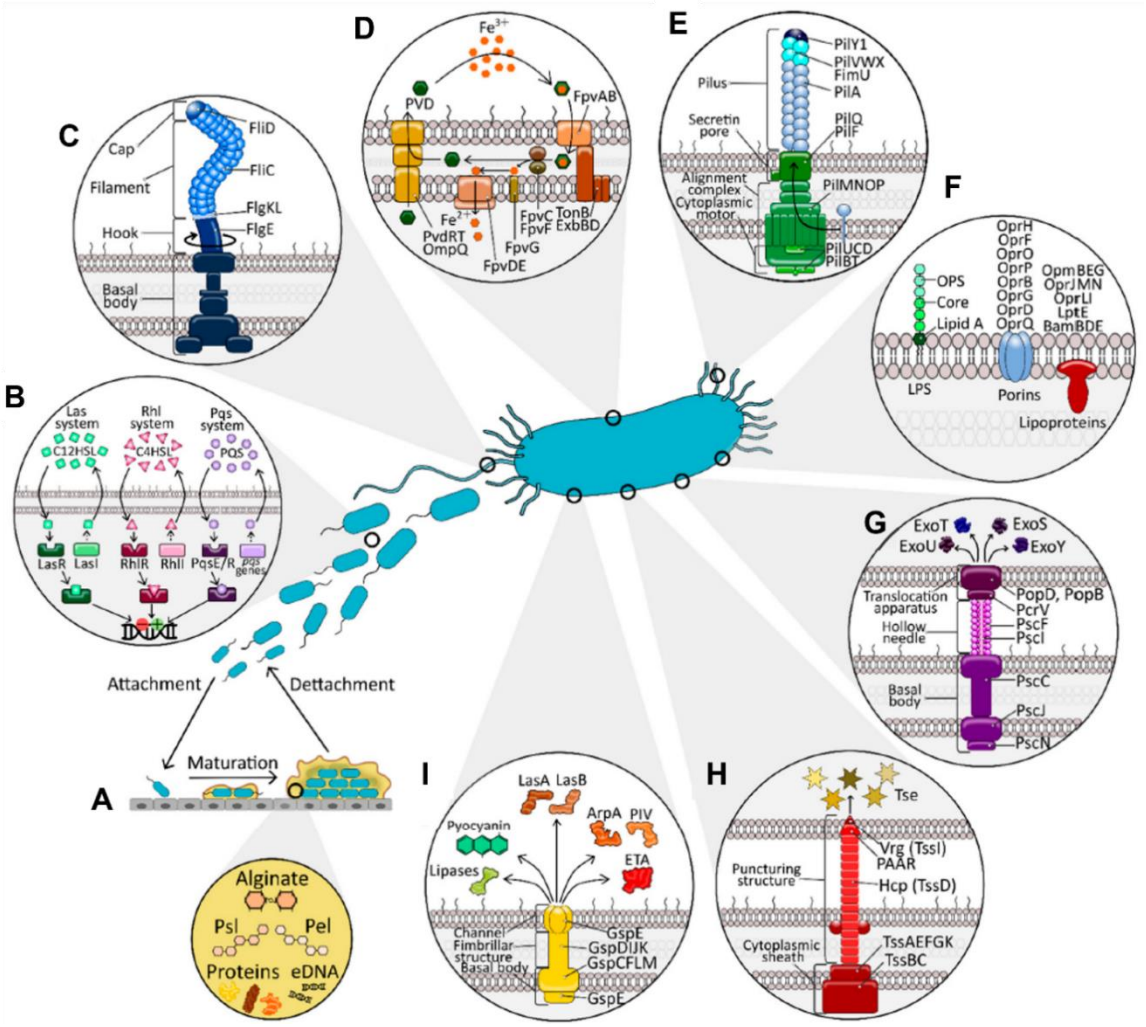

Figure 1.9: *P. aeruginosa* Virulence Factors Used During Respiratory Infections. **A**) Biofilm formation ability and the composition of the extracellular matrix (exopolysaccharides, proteins, and extracellular DNA) allows for protection against the host immune system and for the establishment of long-term infections. **B**) The three main QS systems (Las, Rhl and Pqs) aid in virulence gene regulation and rapid colony adaptation to environmental changes. **C**) Flagellins FliC and FliD incorporated within the flagellar structure allow for chemotaxis in the aqueous mucosal layer common in the respiratory tract. **D**) Pyoverdine (PVD) siderophores function as an iron uptake system for nutrient acquisition. **E**) Type 4 pili (T4P) are often the initiator of the infection as they allow for adhesion and twitching motility. **F**) The specialized lipopolysaccharide (LPS) of *P. aeruginosa* limits entry of harmful compounds and OMPs provide a variety of functions including antibiotic resistance. **G**) The type III secretion system (T3SS) and its four main effectors are secreted directly into the host cell for virulence, while **H**) the type VI secretion system (T6SS) secretes effectors which target the host microbiome for bacterial competition. **I**) The type II secretion system (T2SS) is important for the virulence factors it releases to the extracellular milieu: lytic enzymes (lipases, proteases (AprA and PIV) and elastases (LasA and LasB)), exotoxin A (ETA), and pyocyanin. Adapted from Jurado-Martin *et al.* 2021[157].

provide twitching motility across surfaces, and aid in microcolony and biofilm formation, cell signaling, DNA uptake, and for regulating virulence gene expression[105,162–164]. The T4P is a filamentous polymer made of individual type IV pilin protein monomer PilA, that are polymerized by several chaperones and minor pilins to form fibers that can extend micrometers in length with an outer diameter of ~5-8 nm[32,105,165]. Type IV pilin monomers are assembled into a pilus and disassembled by a membrane-spanning complex known as the Type 4 pilus machinery that is evolutionarily and structurally related to the type II secretion system. There are two main T4P sub-types based on sequence homology in the pilin protein: the type IV a (T4aP) and the type IV b (T4bP) pili. The T4aP are present in a broad species range including *P. aeruginosa*, whereas T4bP are a heterogeneous group common in enteric species such as *E. coli* and *Vibrio cholerae* [32,166]. Disruption of the T4P has been shown to greatly reduce the virulence of gram-negative bacterial pathogens, such that the type 4 pilin proteins have been considered promising targets for the development of anti-virulence drug targets and vaccine/antibody therapy candidates[31,105,167–169]. Another relevance to the biomedical field is the use of T4Ps in the formation of self-assembling protein nanotubes (PNTs) that have significant potential as targeted drug delivery vehicles due to their low immunogenicity and the ability to design the receptor binding domain of the pilin to target various surfaces[170–172].

1.8.2 Structure of the T4P Secretion System

The machinery which regulates the extension and retraction of T4P is a multi-protein complex which spans the length of the bacterial cell wall (**Figure 1.10A**)[173]. In T4aP from *P. aeruginosa* it consists of four main subcomplexes: the motor, the alignment subcomplex, the OM pore, and the fibre[32]. The motor is present in the cytoplasm and IM and consists of the ATPases

PilB for extension, two hexameric PilT complexes for retraction, and PilU with a proposed function in aiding pilus retraction[174,175]. FimX and PilZ are coupled to PilB to mediate T4P biogenesis, as the protein complex forms a receptor for the signal transduction molecule cyclic-bis(3′ → 5′)-dimeric GMP and has been shown to alter the conformation of PilB when bound[176]. PilC is a 3-pass IM protein with two homologous cytoplasmic globular domains; it is predicted to localize in the pore formed by PilB and PilT and is proposed to be rotated by these ATPases to catalyze the polymerization of the PilA pilin[32].

The alignment subcomplex is composed of PilM, -N, -O, and -P and is essential for connecting the cytoplasmic motor to the periplasmic secretin[177]. PilM binds to the cytoplasmic side of the IM and forms a ring which binds it to PilN, PilB and PilT. PilN and PilO are structurally similar and form an oligomeric cage-like ring in the IM and periplasm. These proteins are anchored to the secretin PilQ at the N0 domain by the IM lipoprotein PilP[32]. The OM pore is mainly composed of the 14-PilQ-subunit secretin which has C7 symmetry based on a 7 Å resolution cryo-EM structure and is responsible for forming the channel for the pilus fibre[178]. PilQ assembles into 6 different domains numbered N0-N5 based on a structure of the *Thermus thermophilus* T4P solved by cryo-ET (**Figure 1.10B**)[173]. Domains N0-N3 are seen to be separate from N4 and N5 by 30 Å when a pilus is present in the channel; a change is seen to a closed conformation in which N0-N3 shifts up to contact the other domains when no pilus is present (**Figure 1.10C**). Other proteins which compose the OM pore include PilF, the pilotin lipoprotein responsible for proper multimerization and localization of PilQ, as well as FimV and TsaP which are peptidoglycan-binding proteins which have unconfirmed functions but are proposed to aid in localizing and stabilizing the periplasmic PilQ domains[32].

The pilus fibre is composed mainly of the pilin PilA, however many minor pilins (PilV, -W, -X, -E and FimU) and an adhesin PilY are involved in forming the pilus tip[179–181]. The minor pilins are predicted to self-assembling into a PilVWX trimer similar to the GspIJK system of the T2SS; minor pilins are thought to prime pilus assembly by reducing the energy barrier to extraction of pilins from the membrane[182]. FimU and PilE are additional minor pilins which aid in sequestering PilVWX to the PilA polymer and may aid in melting of the PilA N-terminal helix during pilus assembly[183]. PilY is also predicted to localize with the minor pilins at the pilus tip and is responsible for surface recognition and aids in surface-binding of the pilin[184]. As well, PilA must be processed prior to pilus assembly; the prepilin peptidase PilD hydrolyzes the 6-7 residue leader sequence of nascent PilA and then methylates the new N-terminus[32,173].

T4P machinery is highly similar to that of the T2SS; they are both formed by at least 10 proteins, with the T4aP machinery consisting of proteins nearly identical to the T2SS[185]. The only differences are that the protein PulL in T2SSs corresponds to a fusion of the T4P ATPase PilM and connector PilN[186]. PulN is unique to T2SS but is only found in a subset of species as its role has been shown to be non-essential[187]. The last difference is the absence of PilY1 in T2SS, which in T4aP interacts with the PilVWX complex at the tip[179]. The VWX complex is present in both systems as it is responsible for interacting with secreted effectors and is predicted to play a role in recognizing them[184]. As well, the eight Pil proteins sufficient for pilus assembly are conserved in both T4P and T2SS[185].

1.8.3 General Structure of Type IV Pilin Monomers

The general architecture of the type 4 pilin protein is an extended N-terminal α-helix connected by an α-β loop to a 4- to 7- stranded β-sheet, termed the head domain, which packs onto

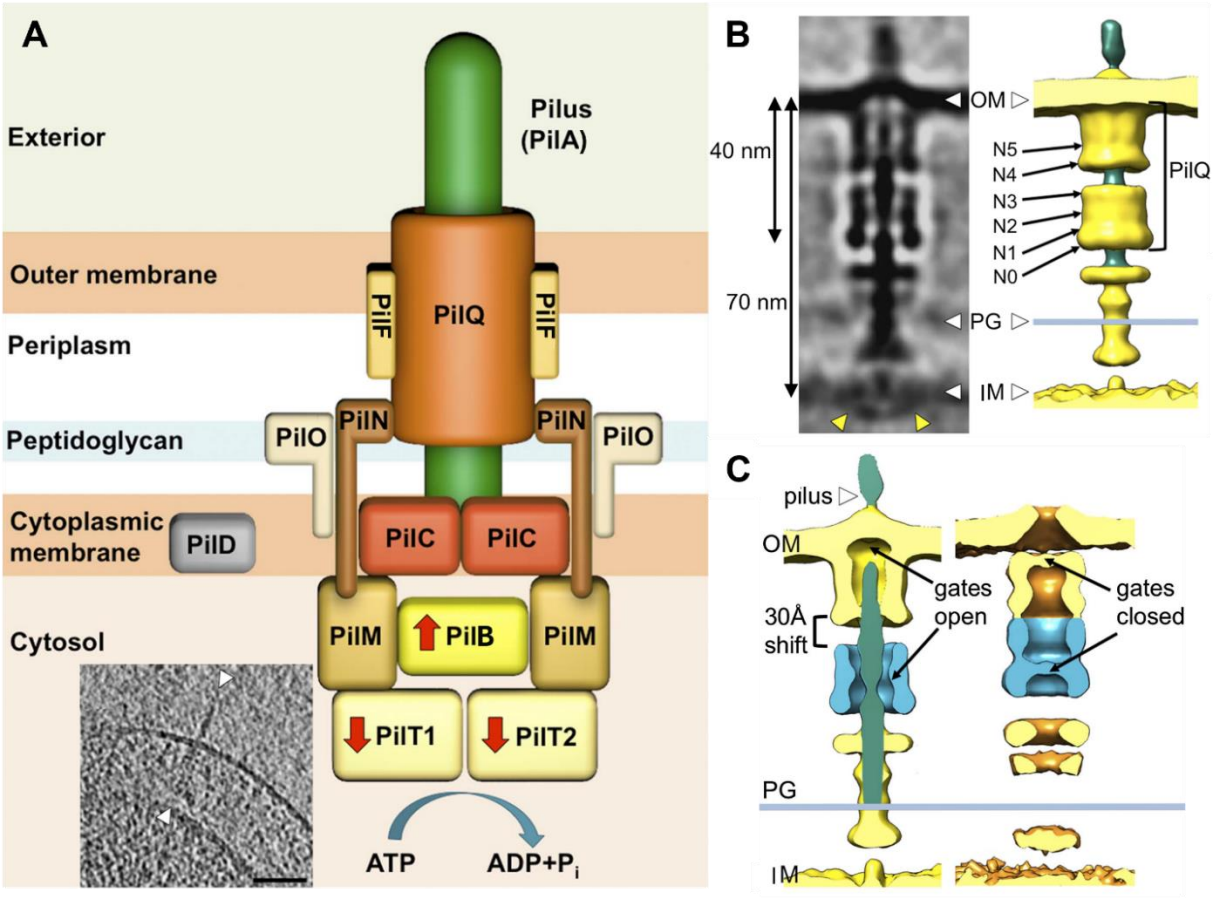

Figure 1.10: Structure of the T4P Secretion System. **A**) A schematic of the proteins essential to piliation in the *P. aeruginosa* T4aP, with a cryo-ET slice microscopy image of the analogous core T4aP complex in *T. thermophilus*. The schematic displays the cytoplasmic ATPases PilB (bright yellow) and PilT1/PilT2 (pale yellow), which drive pilus extension and retraction, respectively as indicated with red arrows. The dimeric complex PilC (red) is located in the IM and is essential for pilus formation. PilM (light brown), PilN (dark brown), PilO (beige) and PilP (not pictured) form the IM assembly platform and connect the periplasmic and cytoplasmic sides of the complex. The PilQ secretin (orange) forms a channel in the OM for secretion of the pilus forming protein PilA (green), which is processed by the prepillin peptidase PilD (grey). The membrane protein PilF (light orange) plays a role PilQ assembly. **B**) displays the ~35 Å cryo-ET structure of the *T. thermophilus* T4P core complex which spans the 70 Å distance between the IM and OM. Domains of PilQ are numbered N0-N5 and in **C**) are seen shifting from an open to a closed conformation depending on the presence of the pilus. This figure was adapted from Gold *et al.* 2015[173].

one half of the α-helix[31,188]. The N-terminal helix is split into two subdomains; the α1-N region is highly hydrophobic and protrudes from the globular C-terminal domain, while the α1-C region embedded in the C-terminal globular head domain is amphipathic and packs against the head

domain. The α1-N region is multifunctional, it serves as a TM anchor to hold pilin subunits in the cytoplasmic lipid bilayer prior to assembly, it is the protein interaction domain for pilin-pilin interactions to form the central core in the assembled pilus polymer, and it has been predicted to serve as a regulatory domain for transduction of the pilus attachment signal to induce virulence gene expression[31,163]. The N-terminal helix is highly conserved in *P. aeruginosa* pilins; sequence variations in the α-β loop and the hypervariable loops connecting the β-sheets in the head domain provide structural diversity between pilin proteins of different bacteria. While there are variations in the protein sequence, the β-sheet of type IV pilin proteins is a conserved structural motif composed of a β-meander with 3-to-4 β-strands having nearest-neighbor connectivity. The C-terminal globular head domain faces the exterior of the assembled pilus, and in *P. aeruginosa* pilin proteins it features a conserved disulfide-bonded loop region responsible for cellular and surface adhesion termed the D-region (**Figure 1.11A**).

1.8.4 T4a Major Pilin Groups

Five different groups of T4a major pilins have been identified in *P. aeruginosa*, classified by sequence length, the size of the D region, and the identity of the pilin accessory protein encoded downstream of the pilin[189]. Group I and II were the first to be identified as they are from common laboratory strains. Group II pilins are the simplest as the lack accessory genes between $pilA_{II}$ and the downstream $tRNA^{Thr}$ gene, whereas Group I pilins are slightly larger proteins and have an ORF between the two accessory genes that is responsible for glycosylation of the C-terminal Serine, termed *pilO* or *tfpO*. Group I is also split into two subgroups (Ia and Ib) based on sequence homology of the $pilA_I$ and *tfpO* genes. *P. aeruginosa* strains PAO1, PAK, CD, KB7 and K122-4 express group II pilins, however strains P1 and 1244 express group Ia pilins and strain SBI-N

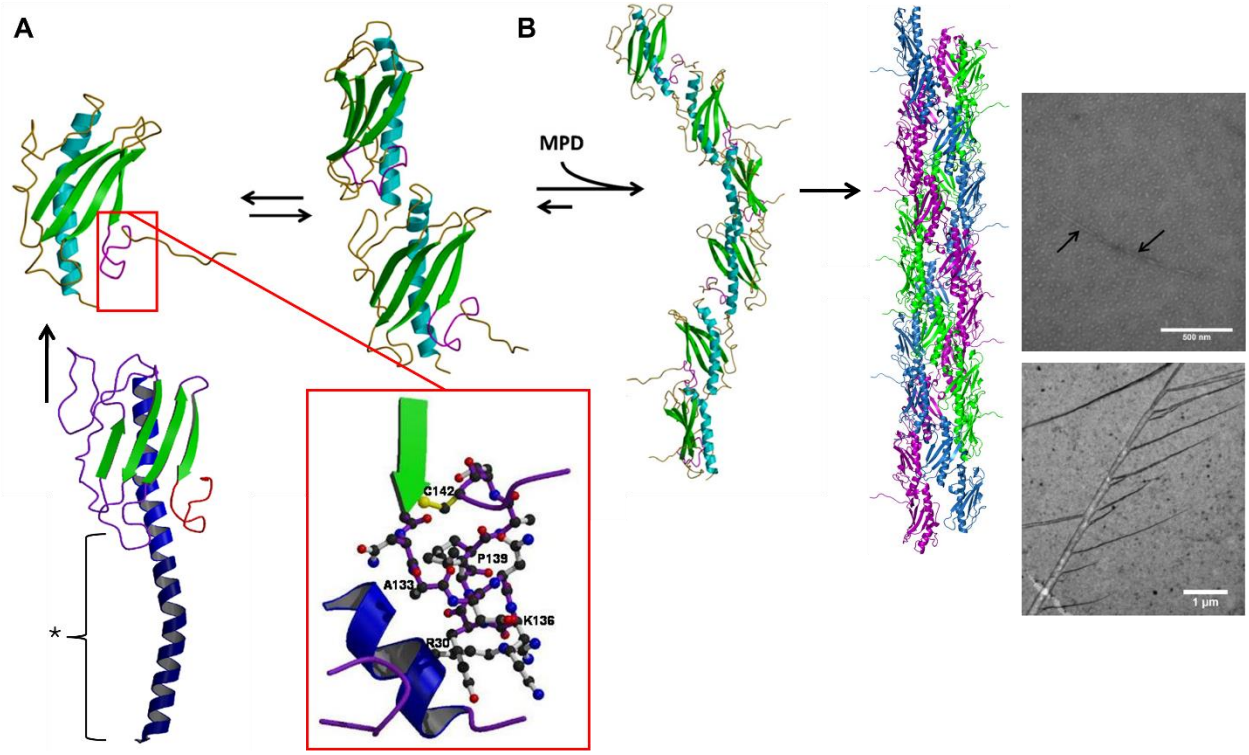

Figure 1.11 The Architecture of the Type IV Pilin from *P. aeruginosa*, specifically the group II pilin from strain K122-4188. **A)** The full-length pilin is shown with the α1-N hydrophobic helix (in the starred bracket). Cleaving this region or expressing the pilin without this portion of the helix does not appreciably change the crystal structure of the head domain of pilins187,189. ΔK122 is shown with the α-1C helix coloured cyan, the β-sheet is in green, the loop regions are in gold, and the receptor binding D-region is coloured magenta, boxed in red and enlarged to show the conserved main chain architecture of the disulfide bonded loop. A monomer-dimer equilibrium is observed for ΔK122, which proposedly serves as an initiator for the nucleation of longer chains upon the addition of a hydrophobic trigger molecule such as 2,4-methylpentane-diol (MPD), as shown in **B)**[105,171]. These fibrils can form protein nanotubes that assemble via a three-start helical model as is seen in native T4P. The ΔK122 monomer/dimers are seen as aggregates in transmission electron microscopy, and form pilin fibrils (highlighted by arrows in the upper image)[171]. However, these pilin fibrils associate into PNTs, which can then coalesce into larger PNT bundles as seen in the lower TEM image. Figures adapted from Petrov *et al.* 2013[171,190].

expresses a group Ib pilin. Group III pilins such as those found in PA14 are longer at 173 residues and have an ORF immediately downstream of the $pilA_{III}$ gene called *tfpY* responsible for enhanced twitching motility and increased number of surface pili[192]. Group IV pilins share higher sequence homology to the minor pilin PilE from *N. meningitidis* than to other *P. aeruginosa* pilin proteins

from other groups, and two accessory proteins from genes *tfpW* and *tfpX* are encoded immediately downstream of $pilA_{IV}$, with functions of a glycosyltransferase/oligosaccharyltransferase and an unknown function similar to that of TfpY, respectively[192–194]. Group V pilins such as the pilin from *P. aeruginosa* strain 110594 are a single residue shorter than group III pilins, and they have a single accessory protein encoded by *tfyZ* downstream of the pilin gene, which is homologous to TfpY and TfpX in sequence and function[31,189,192,195].

1.8.5 T4Ps as Drug Delivery Vehicles

The term bionanotechnology refers to the use of biological molecules engineered to form nanoscale building materials[170]. The assembly of small molecules into more complex higher ordered structures is referred to as the "bottom-up" process, in contrast to nanotechnology which typically uses the "top-down" approach of making macroscale devices smaller. These biological molecules include DNA, lipids, peptides, and proteins[170,196–198]. The molecular self-assembly observed in protein-based systems is mediated by non-covalent interactions such as hydrogen bonds, and electrostatic, hydrophobic and van der Waals interactions. When taken on a singular level these bonds are relatively weak, however combined they are responsible for the diversity and stability observed in many biological systems. The use of PNTs for biomedical applications is of particular interest due to their well-defined structures, assembly under physiologically relevant conditions, and manipulation through protein engineering approaches; such properties of proteins are difficult to achieve with carbon or inorganically derived nanotubes[170,199,200].

Work on the K122-4 pilin from *P. aeruginosa* has revealed that the protein oligomerizes into nanotubes in the presence of hydrophobic surfaces or compounds[105,171,201]. When generated *in vitro*, the pilin-derived PNTs share a similar morphology and diameter (~5–6 nm) to *in vivo* T4P,

the former can reach a length of several hundred micrometers compared to native pili that typically have a length of 10 μm[165,191,202]. From a bionanotechnology perspective, T4P form robust nanofibers with the ability to bind biotic and abiotic surfaces via their tips. These interactions have been mapped to the D-region of the pilin. It has been estimated that the attractive force between the native T4P tip and steel is in the range of 26–55 pN/molecular interaction, and 78–165 pN/molecular interaction for nanotubes derived *in vitro* (**Figure 1.11B**)[203].

The β-sheet and connecting loops of type IV pilins form the hydrophilic surface of the assembled pilus, and in the context of using a T4P-based PNT as a drug delivery vehicle this surface would be exposed to the immune system[170]. These regions show significant sequence variability between T4P from different bacterial systems, which permits mutagenesis of the region to design fibers with altered surface properties, proposedly with little effect on the propensity for polymerization. Protein engineering of the pilin monomer has the potential to lead to nanofiber attachment to other abiotic surfaces; addition of a polyhistidine tag to the C-terminus of the protein can potentially direct binding to nickel and copper surfaces or nanoparticles. The D-region of the pilin is responsible for forming specific interactions with cellular glycolipids and thus has the potential to bind biotic surfaces[172]. This receptor-specific interaction provides evidence that T4P-based PNTs have great therapeutic potential as they could allow for mediated drug delivery upon receptor binding of the synthetic nanofibers.

Functional nanostructures have been generated from native bacterial pili and explored for their potential use as biological nanowires. For example, the type IV pili of *Geobacter sulfurreducens* reduces Fe^{3+} oxides by transporting electrons over long distances and has potential applications for use in microbial-based fuel cells[204,205]. Further studies have shown that cultures of *G. sulfurreducens* produce biofilms that exhibit high current densities when incorporated into

microbial fuel cells. *G. sulfurreducens* pili are also capable of long-range metallic-like conductivity and supercapacitor behavior, and therefore are proposed to be a low-cost and environmentally sustainable form of energy storage[170,205,206].

1.9 *P. aeruginosa* Strain P1

In 1984, sputum samples of CF patients with *P. aeruginosa* infections were cultured from the University of Minnesota Hospital, Minneapolis, and a strain was serotyped P1 for a study on the effect of nonopsonic phagocytosis on different bacterial strains[207]. Four years later, Pasloske *et al.* (1988) reported on cloning and sequencing the pilin genes from strains P1 and K122-4, identifying their sequence features as differing from the model *P. aeruginosa* pilins, the PAK, PAO and CD pilins[208]. The sequence of P1 was determined to be nearly identical, save for one residue (K92Q) to that of strain 1244, another clinical isolate[208,209]. In 2004, the crystal structure of the truncated K122-4 (ΔK122) pilin [PilA(Δ1-28)] was solved; crystals of the N-terminally truncated P1 (ΔP1) pilin [PilA(Δ1-35)] were obtained and diffracted in a similar manner but the structure could not be solved due to a lack of anomalous signal in heavy-atom-soaked crystals, nor could the structure be solved using molecular replacement (MR) due to insufficient similarity between proteins with solved structures[190,210].

1.10 Protein Structure and Dynamics

1.10.1 Importance of Flexibility in Proteins

The initial consensus on protein structure-function relationships relied on the premise that the amino acid sequence of a protein dictates how it will fold into a highly specific three-dimensional structure which directly relates to its activity[211]. The importance of dynamics in protein structure was initially emphasized by the induced fit and the Monod-Wyman-Changeux models which theorized movement in enzymes are required for substrate binding, product formation, and product release[212–214]. Flexibility and dynamics in a protein's structure is known to be integral to its function as the molecular motions possible in a protein (transporter, chaperone, enzyme, etc.) dictates its substrate specificity/promiscuity, affinity and kinetics[214,215]. Un-structure in proteins can also have important functions, as many transiently and intrinsically disordered regions of proteins are commonly involved in PPIs, protein-nucleic acid binding, phosphorylation, or to serve as flexible linkers between functioning domains[211,216,217]. A continuum exists to the extent of disorder intrinsic to a protein's structure, from a well-defined single domain fold to a multidomain protein with disordered flexible regions, to disordered molten globules, and to extended random coils (**Figure 1.12A**)[216,218,219]. As well, within a cell the same protein can exist in an array of states as influenced by the peculiarities of its amino acid sequence and its chemical environment; all proteins emerge from ribosomes in an unfolded state, fold as they are translated, and have the potential to misfold and aggregate (**Figure 1.12B**)[219,220]. Proteins with conformational heterogeneity can go through cycles of folding and unfolding and often have multiple active conformations leading to multiple functions.

The occurrence of unstructured regions greater than 50 residues in length is common in many functional proteins[212,216,221]. These intrinsically disordered regions (IDRs) are more

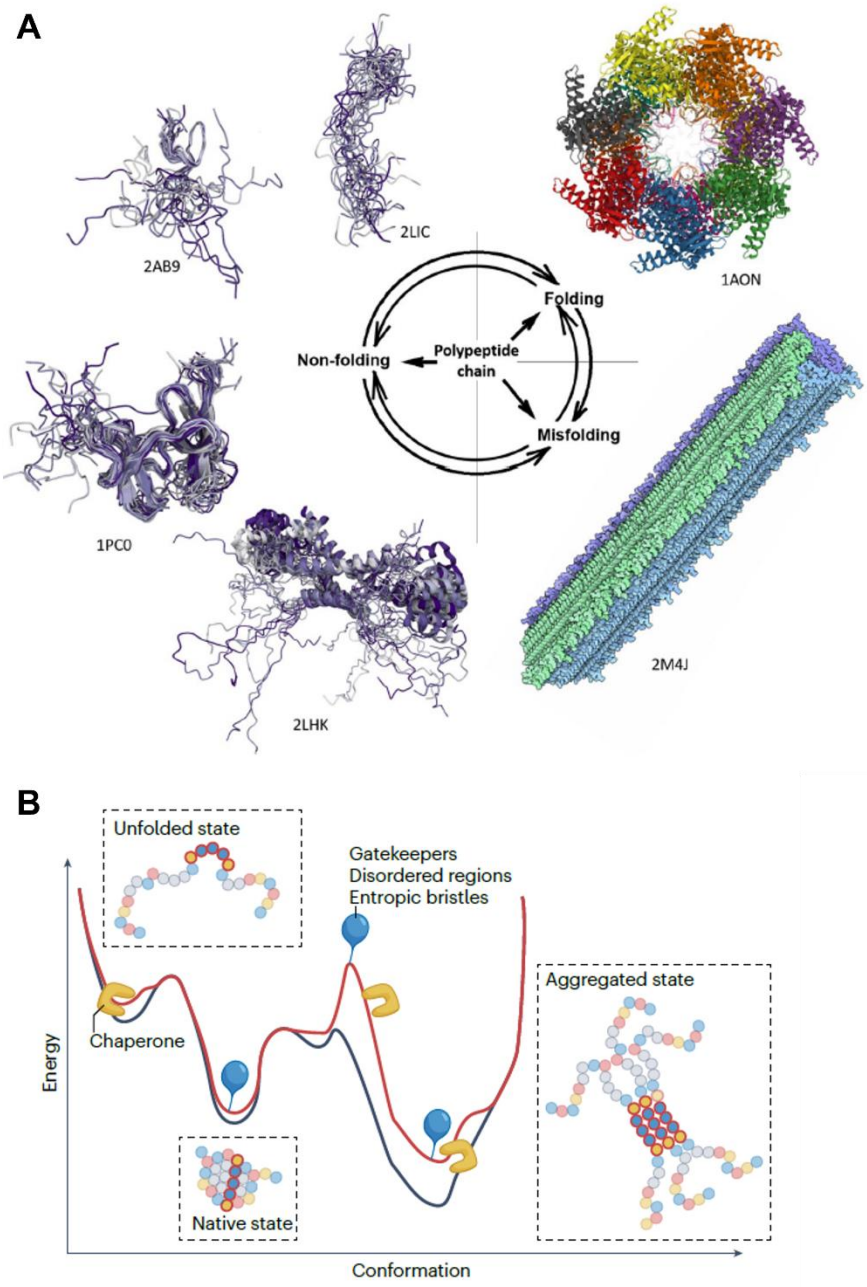

Figure 1.12 The Continuum of Protein Structural Conformation. **A**) The fate of a polypeptide in a cell can be represented by the types of structures seen by X-ray crystallography and nuclear magnetic resonance (NMR). These structures are of different proteins, however many proteins in a cell can exist in this perpetual folding-non-folding-misfolding cycle depending on the disorder intrinsic to the primary sequence of the protein. 1AON: GroEL/GroES/(ADP) chaperonin complex[222], 2M4J: β-amyloid fibril[223], 2LHK: T3SS chaperone[224], 1PCO: porcine pancreatic prolipase[225], 2AB9: precursor protein that produces the cyclic trypsin inhibitor SFTI-1[226], 2LIC: Polyserine Tract of *Apis mellifera* Vitellogenin[227]. Figure adapted from Uversky 2021[219]. **B**) A free-energy landscape demonstrating how proteins can exist in a variety of states, and how the presence of disordered regions can prevent aggregation by increasing the free energy barrier required to alter the conformation. Adapted from Louros *et al.* 2023[220].

commonly found in eukaryotic and viral proteins rather than prokaryotes due to the requirement of complexity in their morphologies. In eukaryotic cells it is suggested that there exists a link between intrinsic disorder and evolution, as various cellular compartments are formed by proteins with IDRs[228–230]. IDRs in prokaryotic proteins have been found frequently in proteins which participate in pathogenesis, interspecies interactions, DNA modification, and the generation of precursor metabolites and energy[231,232]. FlgE from the bacterial flagellum is an example of a bacterial protein that constitutes a large multiprotein complex and features a necessary IDR that permits multiple conformations for important PPIs[233]. As well, several bacterial regulatory proteins have conserved short intrinsically disordered linker regions called Q-linkers[216,217,234]. Therefore, it is not uncommon for disordered regions in bacterial proteins to be important flexible linkers of folded domains.

1.10.2 Protein Phase Separation by Intrinsically Disordered Regions

Liquid-liquid phase separation (LLPS) involves the exchange of macromolecule/water interactions for macromolecule/macromolecule and water/water interactions under conditions for which this process is energetically favorable, resulting in the formation of a dense phase (a condensate) and a dilute phase [235–237]. It is a de-mixing of a homogeneous solution such that distinct regions are occupied by distinct species as promoted by the reversable self-concentration of a macromolecule, and can manifest as liquid-liquid, to liquid-gel, to liquid-solid phase separations. This compartmentalization has been shown to occur in proteins and nucleic acids and can manifest *in vivo* and *in vitro*. LLPS has been increasingly implicated in the formation of membrane-less organelles for cellular organization, signaling transduction, transcription activation, RNA processing, polymerization of biomolecules and pathological aggregation (including those by

cancer, neurodegeneration, and by pathogens)[238–242]. Numerous physical principles and chemical interactions drive LLPS in macromolecules, including multivalent, electrostatic, cation-pi, pi-pi, and hydrophobic interactions[235,243–246]. In the case of protein LLPS, the dominant interactions driving condensation are multivalent interactions between disordered regions or modular interacting domains[242,247]. Many phase separating proteins have large IDRs that do not form a stable folded structure, and their interactions with solvent are typically outcompeted by intermolecular non-covalent bonds with other protein molecules[236,248].

Solvation is typically high in proteins with IDRs, and in those regions, there is a high correlation between the number of contacts a residue makes with water molecules and the probability that a residue is involved in a pi-pi interaction. Pi-pi interactions are non-covalent bonds that involve p orbitals of atoms consisting of one electron interacting with another single occupancy p orbital, where one orbital is electron rich, and the other is electron poor as caused by the atom's proximal covalent bonds[249,250]. These interactions are commonly associated with aromatic rings where induced quadripolar interactions provide the non-covalent bonds formed by π (pi) orbitals; these orbitals are present in any bound sp^2 hybridized system. Every residue in a protein has the potential to perform pi-pi interactions as all 20 amino acids have a peptide bond, which features sp^2 hybridized orbitals throughout the atoms of the planar amide[246,251]. However, those amino acids with small side chains that leave the peptide backbone exposed; Gly, Ser, Thr and Pro, are residues in which amide pi orbitals commonly perform this interaction in proteins. As mentioned, aromatic rings stack to perform pi-pi interactions and therefore amino acids Tyr, Phe and Trp also participate using their side chains[246,251,252]. Other side chains with conjugated pi systems in their side chains include His, Gln, Asn, Glu, Asp and Arg[246,251].

While any combination of pi-pi interaction is possible between residues in an IDR, it has been found that many of such contacts involve one of the six amino acids containing side chains with non-aromatic sp^2 hybridized atoms, or the peptide bond[246]. 36% of all observed pi-pi interactions occurring throughout solved protein structures in the protein databank (PDB) do not involve an aromatic residue, and face-to-face contacts of peptide bonds occur as often as aromatic face-to-face contacts. It was also found that planar pi-pi interactions occur more often at positions associated with disorder, as they have an overall secondary structure with less rigidity; this is especially true of regions responsible for phase separation. Of the phase separating IDRs solved in the PDB the single property of long-range (≥5 residues apart or different chains) pi contact propensity is sufficient for marking them as outliers in comparison to the rest of the proteome. Through these principal structural properties, many researchers have made software packages to predict IDRs in a protein based on the primary sequence[246]. The work of Vernon *et al.* (2018)[246] showed that most phase separating proteins can make significant non-local planar pi-interactions, and IDRs recognized with their primary sequence analysis method called the Phase Separating Predictor (PSP) or PScore engine have been shown to have a high propensity for phase separation *in vitro*. Newer software for the analysis of protein disorder from primary sequence such as MolPhase[253] and PhaSePred[242,254] have appropriately integrated algorithms for the potential for a primary sequence to perform long-range pi contacts as part of their scoring algorithms.

1.10.3 Methods of Study for Protein Structure Determination and Structural Dynamics

1.10.3.1 Thermofluor Assays

Differential scanning fluorimetry, often referred to colloquially as Thermofluor, involves the use of a fluorescent hydrophobic probe to determine the optimal solvent for protein stability using the melting temperature (T_m) of the protein as an indicator of conformational unfolding (**Figure 1.13**)[255–258]. The probe is naturally quenched upon addition to an aqueous sample of protein in a buffer by interaction with water in the solvent; the fluorescence signal is measured spectrophotometrically[259]. Heat is applied to the samples in a steady, linear fashion, and as the protein begins to unfold it exposes hydrophobic residues that the probe binds to and fluoresces. As the temperature continues to increase, the protein begins to aggregate and hydrophobic residues become excluded from solvent, resulting in a loss of probe binding and a decrease in fluorescence. The produced profile is a melting curve that can allow for characterization of thermal stability through comparison of the melting temperature (T_m), which is the midpoint of the unfolding transition, as determined through the highest value on the second derivative of the function. The assay can be altered to optimize buffer conditions to aid in protein stability, characterize the effect of mutations on the thermal stability of proteins, predict the crystallisation propensity of proteins, and determine quantitative binding affinities between a protein and a ligand[259–261].

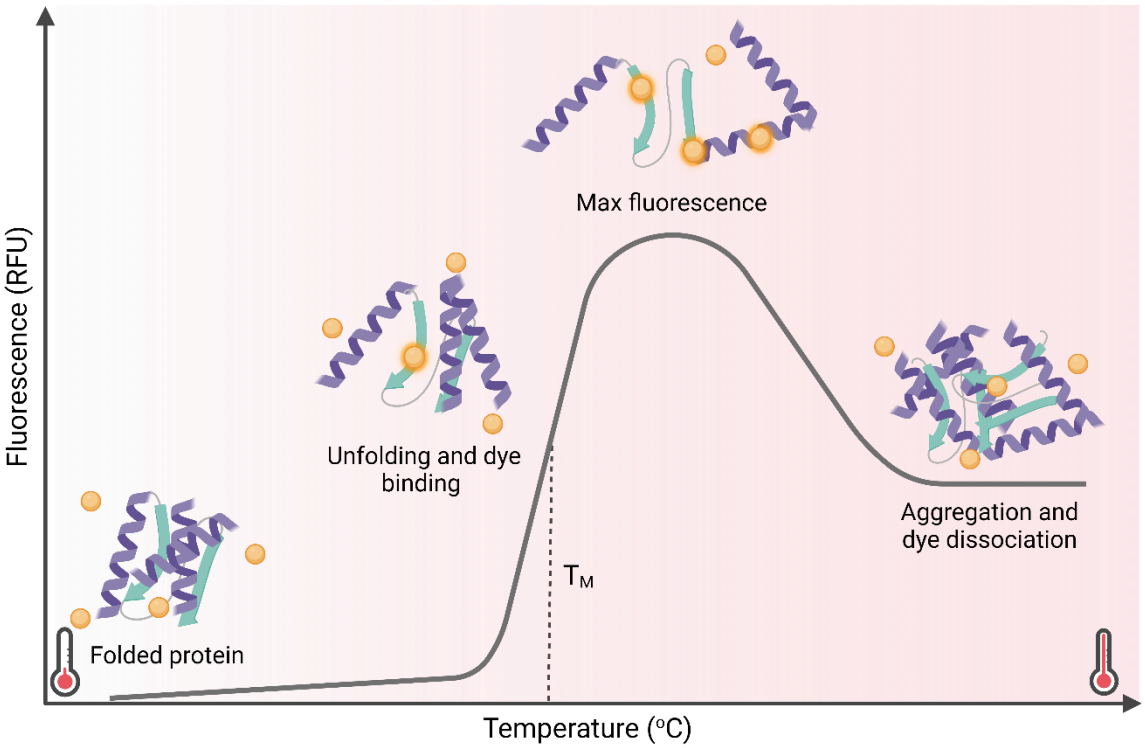

Figure 1.13 Optimal Output Graph from a Thermofluor Experiment. A protein (in purple and green) in solution with a hydrophobic fluorescent dye (yellow) is placed in a room temperature (RT) chamber with a fluorescence detector[262]. The sample is heated at a consistent rate while monitoring fluorescence; at first the protein should be completely folded, and as the protein begins to unfold due to thermal denaturation, hydrophobic residues are exposed to the solvent allowing the hydrophobic dye molecules to bind and fluoresce. This fluorescence signal increases as temperature is continually increased and all protein molecules in solution completely unfold, following which the hydrophobic regions of the protein bind together, forming aggregates that exclude the dye molecules. This results in a lowering of the fluorescence signal, creating a sigmoidal curve in which you can detect the melting temperature (T_m) when the fluorescence signal is increasing at its most rapid rate. Figure created with BioRender.com and adapted from a BioRender template by Turner, M.

1.10.3.2 Circular Dichroism

Circular dichroism (CD) is a widely used and well-developed technique for the determination of protein secondary structure[263,264]. The technique relies upon the differential absorption of left- and right-hand circularly polarized light (CPL) (**Figure 1.14**). In this case, the electric field of a photon has a circularly rotational direction relative to the direction of

propagation, where the photon vector remains constant in magnitude[265,266]. When these photons pass through an asymmetric (chiral) molecule, the speed, absorbance and the wavelengths differ depending on the direction of their polarization. As the sum of the vectors for the right- and left-handed polarized light forms an ellipse, the change in ellipticity (measured as Δε) of a substance

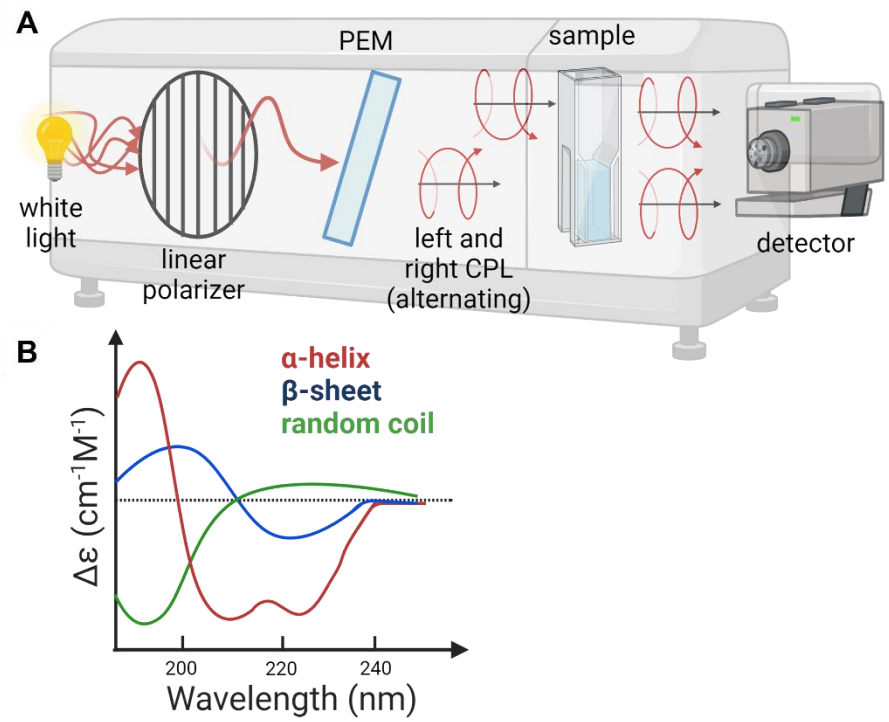

Figure 1.14 Schematic of a CD Spectrometer and Prototypic CD Spectra for Protein Secondary Structure. **A**) White light is passed through a linear polarizer and a photoelastic modulator (PEM) to generate left and right CPL in an alternating fashion[267]. Chiral molecules absorb left and right circularly polarized light, resulting in a detected amplitude changes. As the wavelength of light used is changed, the structural features of a chiral molecule in solution can further affect these amplitude changes. **B**) Prototypic CD spectra for proteins with secondary structures composed of all α-helices (red), β-sheets (blue), or no significant secondary structural features (green). Figures created with BioRender.com.

as a function of the wavelength of incident light is reported in performing CD experiments, although the change in absorbance of the differentially polarized light is what is measured in the CD spectrometer. As proteins are chiral molecules and secondary structures result in specific absorbance patterns at certain emission wavelengths, CD is widely used to determine the secondary structure content of proteins and can be used to determine the structural stability of proteins in solution[265,268,269].

1.10.3.3 Protein Mass Spectrometry and Collision Induced Unfolding

Mass spectrometry (MS) has been adapted from identifying organic and inorganic small molecules to identifying proteins from a mixture and studying the structure and structural dynamics of purified proteins via a variety of techniques[270,271]. Mass spectrometers ionize molecules such that they can be influenced by an electromagnetic field which accelerates the ionized particles to separate them by their mass-to-charge ratio (m/z) before being measured by a detector (**Figure 1.15A**)[271,272]. Typically, electrons are used to ionize molecules in MS, resulting in positively charged mass fragments. The two most common techniques for ionization in protein MS is electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). ESI is a soft ionization method that allows for evaporation of the solvent using a combination of voltage, heat, and air such that molecules in solution are desorbed as intact ions in individual nanodroplets, following which they enter the mass spectrometer for analysis (**Figure 1.15B**)[273–275]. MALDI involves the mixing of the protein solution with a matrix material and applying it to a metal plate where pulsed laser light is used for desorption; the heated nanoparticles are then ionized and enter the spectrometer[276,277].

Collision induced unfolding (CIU) MS is a reliable technique for studying and comparing a protein's conformational stability[278–281]. CIU-MS employs an ion mobility spectrometry (IMS) cell within the mass spectrometer that is filled with an inert gas and provides a weak electric field gradient that filters ionized species separated in the previous trap cell (**Figure 1.15C**). Drift time in the IMS cell, which is defined by the time an ion takes to travel the distance from the drift tube to the detector, is dependent on mass and size as well as the conformational shape of the fragment. In CIU-MS, the trap collision energy (CE) is increased in a stepwise fashion, allowing for breakages of non-covalent bonds that can be visualized by IMS as a shift to a slower moving, unfolded species. Therefore, CIU provides an indication of the conformational stability of proteins; a protein that has a lower number of drift time distributions (representing different conformations) and/or remains in a lower drift time distribution at a higher CE has higher conformational stability.

1.10.3.4 X-ray Crystallography

Of all techniques for determining the high-resolution structure of macromolecules, X-ray crystallography is the most impactful for structural biology and remains the most powerful method for determining the structure of proteins, nucleic acids, small molecules, lipids, and complexes of any combination of the aforementioned molecules[282–285]. Of the over 224,000 structures currently in the Protein Data Bank (PDB), 84% are from protein crystal X-ray diffraction, and 25 Nobel prizes have been awarded related to X-ray crystallography and the structures of molecules solved by it[282]. Crystallisation of proteins is typically approached in an empirical manner as crystallisation conditions for proteins that have never been crystallized cannot currently be predicted[286,287]. A protein solution of high concentration and purity is combined with a variety of precipitants,

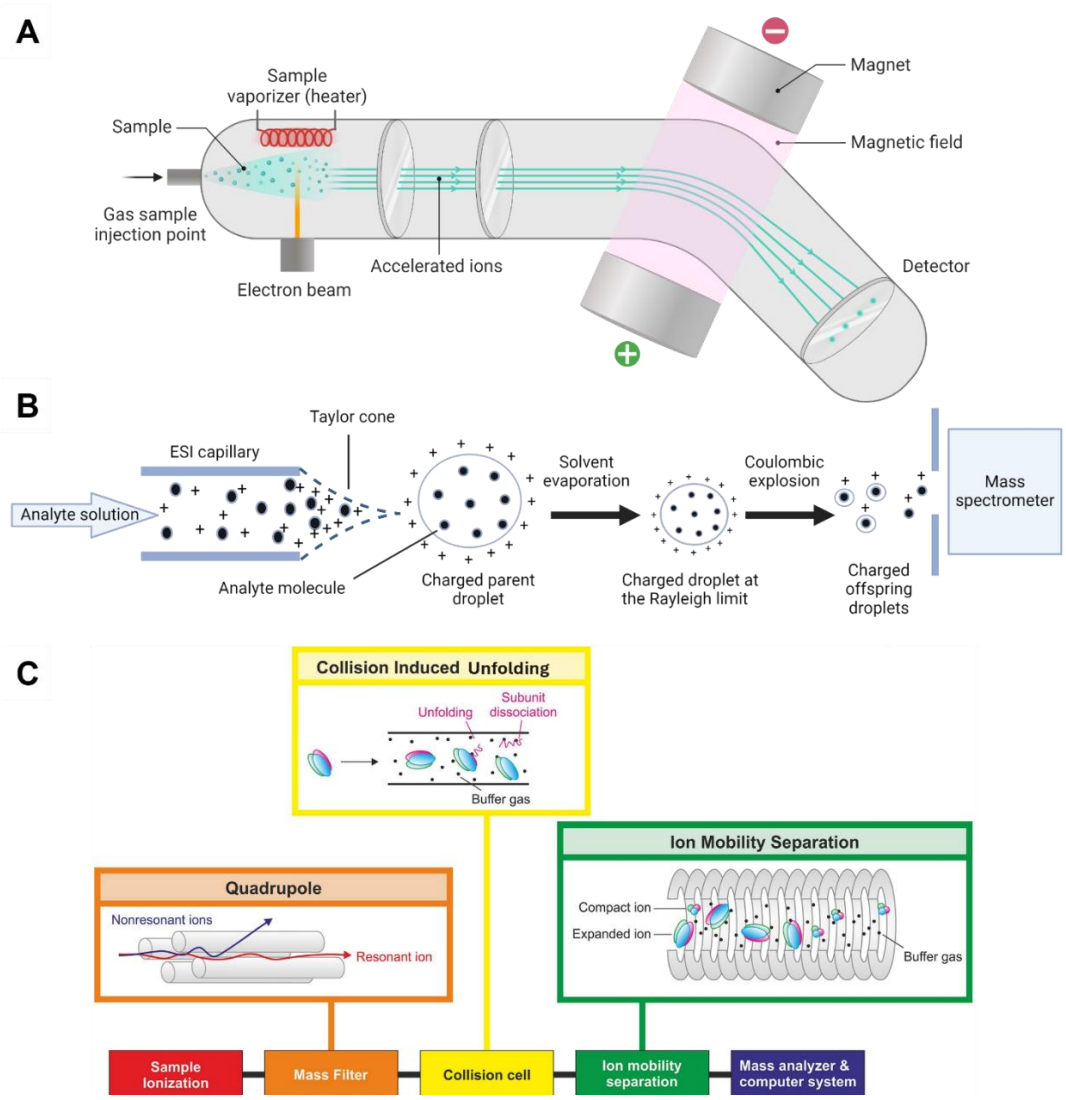

Figure 1.15 Simplified Diagram of a Mass Spectrometer, ESI, & CIU-MS. **A**) A simplified diagram of a mass spectrometer, displaying the acceleration and separation of ions by magnets to alter their time to the detector based on their m/z ratio. Figure created with BioRender.com. **B**) ESI involves the use of a capillary carrying a potential difference of ~3kV through which the sample of interest is pumped; the high voltage combined with the surface tension of the liquid causes the formation of an electrospray cone (known as the Taylor cone)[288]. Fine drops from the liquid jet are created when the charge exceeds the Rayleigh limit, and a heated inert gas further aids in aerosolization allowing for the formation of nanodroplets containing single analyte molecules due to Coulombic explosion caused by charge repulsion by ions in the droplets[289]. Figure created with BioRender.com, adapted from a template by Lamichhane, D. **C**) Components required for CIU-MS; firstly, sample ionization occurs typically through ESI, then desired ions are filtered through a quadrupole before entering the collision cell where the protein is unfolded through collisions with large inert gas molecules such as argon[290]. The following drift tube contains a small inert gas (typically N_2) which allows a uniform electric field to be induced throughout the tube to separate analyte ions based on size and conformation, such that larger and more unfolded molecules have longer drift time[291]. Figure adapted from Calabrese and Radford, 2018[292].

buffers, salts, and other additives in attempt to find a suitable solvent which stabilizes the molecule such that it is amenable for molecules to arrange in an ordered lattice. This is achieved in the context of vapour diffusion, where the protein solution is mixed with the crystallisation condition thus diluting both protein and precipitant by a proportion dependent on the volumetric ratio of the mixture, and the crystallisation solution is placed into a reservoir before the experimental chamber is sealed (**Figure 1.16A**)[282,287]. This forces water to evaporate from the experimental drop as it diffuses to the higher concentration of precipitant in the mother liquor, thus increasing the concentration of protein and precipitant in the experimental drop over time until vapour diffusion between the liquids in the chamber results in a concentration equilibrium. If the concentration of protein and precipitant in the experimental drop reaches a chemical space (which can be described by a solubility phase diagram) suitable for crystal nucleation, and if after initial crystals form vapour diffusion reaches equilibrium such that the protein remains in the metastable zone, single, large crystals suitable for X-ray diffraction will likely be produced (**Figure 1.16B**). Common experimental configurations for protein crystallisation by vapour diffusion include hanging drop, where the experimental drop is placed on the underside of the lid of the sealed container and hangs above the mother liquor due to surface tension, and sitting drop where a platform is used to harbour the experimental drop and the mother liquor is placed at the bottom of the container prior to sealing.

Crystals are a 2-dimensional (2D) or 3-dimensional (3D) array of identical unit cells arranged in a crystal lattice which can exhibit many symmetry elements[283]. There are 230 possible combinations of rotational and translational symmetry that can exist in a 3D lattice, resulting in the space groups used to categorize symmetry in 3D space[293,294]. These space groups can be further categorized into the 32 crystallographic point groups that are compatible with a 3D lattice; they are grouped based on which symmetry operations in space groups would provide the same result

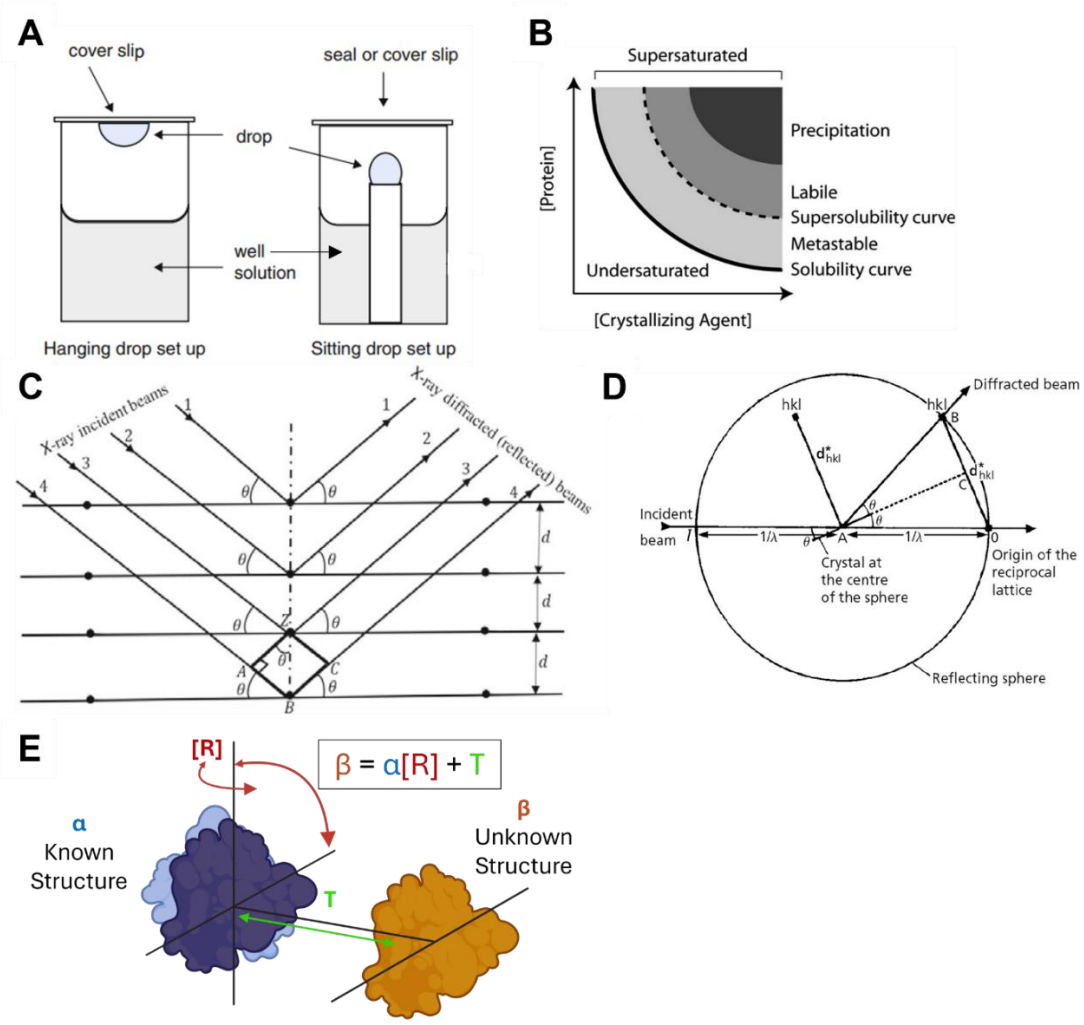

Figure 1.16 Protein Crystallisation and X-ray Diffraction. **A)** Schematic of common vapour diffusion experimental setups for protein crystallisation. These experiments require an experimental drop consisting of the protein solution of interest mixed with the mother liquor, typically containing a precipitating agent such as ammonium sulphate or polyethylene glycol, as well as a reservoir (labelled 'well') in which an exceeding volume of the mother liquor is placed. In the hanging drop configuration, the experimental drop is placed on a coverslip and inverted to hang above the mother liquor in the well; the coverslip is typically sealed to the container with vacuum grease. The sitting drop setup has a separate platform for the drop to lay apart from the mother liquor and is typically sealed with tape. Figure from Ranatunga *et al.* 2014[295]. **B)** A solubility phase diagram for crystallisation. The labile zone is the optimal range of protein-precipitant concentrations which allow for crystal nucleation; for many proteins, the labile zone is a very narrow range in crystallisation experiments and difficult to achieve[296]. Figure from Luft *et al.* 2011. **C)** A depiction of how X-rays would diffract from the planes of a crystal lattice and form constructive interference according to Bragg's Law. Figure from Ameh 2019[297]. **D)** The conditions which yield a crystal diffraction pattern is fulfilled by Ewald's sphere. Figure from Ameh 2019[297]. **E)** A depiction of the molecular replacement process, in which a molecule with a known structure is rotated in both planes via the rotation function and then translated to fit in the unit cell of the unknown molecule, thus solving the phase problem[298]. Figure created with BioRender.com.

in the same lattice, and on restrictions to the symmetry available in crystal lattices[297]. Space groups are categorized into crystal systems based on their point groups and into their lattice groups by their Bravais lattices. There are seven crystal systems, from those with the least to the most symmetry they are: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic[297]. The lattice systems are named similarly except there is no trigonal lattice, however rhombohedral lattices can exist. Of the 230 space groups, protein crystals typically only pack in 65 as they are chiral molecules and cannot properly form the necessary crystal contacts to pack into the other 165 space groups, which contain mirror planes or centers of inversion[299]. Knowing the crystallographic space group of the diffracted crystal aids in solving the structure of the molecule in the crystal; only a portion of the crystal needs to be diffracted with damaging X-rays, rather than all frames in a 360° rotation, to solve the contents of a single asymmetric unit as dictated by the symmetry of the space group[282,283]. Non-crystallographic symmetry (NCS) can also exist within a unit cell, where the asymmetric unit contains multiple copies of the crystallized molecule.

When X-rays are impinged on a crystal they are diffracted by the electrons in the atoms of the molecule in the unit cell, and due to the regularity of the unit cell the X-rays diffracted by atoms in one unit cell interfere with X-rays that are diffracted by equivalent atoms on the proximal unit cell, and all other proximal unit cells[300,301]. This interference can be constructive or destructive, and the conditions which result in constructive interference is defined by Bragg's Law, $n\lambda = 2dsin(\theta)$ (**Figure 1.16C**). This equation states which crystal orientation and illumination angles result in constructive interference, as λ is the wavelength of X-rays, *d* is the distance between the planes of the lattice which are defined by Bragg indices *hkl*, and θ is the incidence angle of the X-ray beam. The *d* spacing represents the resolution of the reflection, as reflections with larger *h, k,* and *l* have smaller *d* values and therefore higher resolution. Ewald's sphere

describes how the wavevector of the incident and diffracted beam relates to the diffraction angle for a reflection and the reciprocal lattice of the crystal[297,300]. As diffraction detectors measure the elastically scattered waves reflected by the crystal, the wavevectors which satisfy Bragg diffraction and produce a constructive interference spot lie on Ewald's sphere and are therefore reciprocal lattice points (**Figure 1.16D**).

To derive the 3D structure of a molecule from the diffraction pattern of its crystal, real space information must be derived from the reciprocal space data provided by each diffraction image[300]. This is done by performing a Fourier transform on the structure factor, which itself is a Fourier summation arising from the overall scattering contributed by every electron in the atoms which compose the molecule in the unit cell of the crystal. The structure factor represents the sum of all amplitudes of waves scattered by all electrons at each particular coordinate in the unit cell, however this overall scattering is discretized into individual reflections due to Bragg's law because a crystal is being diffracted rather than a single object. The electron density can be calculated by the equation $\rho(x,y,x) = \frac{1}{V}\sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)}$, where $\rho(x,y,x)$ is the electron density with real space coordinates x, y, z; V is the volume of the unit cell, and F_{hkl} is the structure factor with reciprocal space coordinates h, k, l. As indicated by the imaginary number i present in the equation, F_{hkl} is a complex number with an amplitude $|F_{hkl}|$ and a phase α_{hkl}. The electron density equation can be rewritten as $\rho(x,y,x) = \frac{1}{V}\sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i[hx+ky+lz-\alpha_{hkl}]}$, which presents the phase problem in crystallography as only the amplitude of the reflected beams can be determined during the diffraction experiment; the phase information is lost. The phase problem in macromolecular X-ray crystallography can be solved by a number of methods: multiple isomorphous replacement (MIR) which uses heavy atoms introduced to the crystal to determine the coordinates based on the intensely scattering atom[302,303]; multi-wavelength anomalous diffraction (MAD) which uses X-

rays of different wavelengths and therefore requires synchrotron radiation[304,305]; single wavelength anomalous diffraction (SAD), which is similar to MAD except only one additional wavelength of X-rays is chosen for crystal diffraction based on knowledge of an appropriately scattering atom dispersed throughout the protein, typically done using selenium through incorporation of selenomethionine and the appropriate 0.979Å wavelength[306]; and molecular replacement (MR)[307]. If there is a structural model with homology to the molecule in the unit cell of the crystal, MR can be used to solve the phase problem for the crystal's diffraction data if the structures are of a sufficiently similar fold (**Figure 1.16E**). The MR model must be rotated using a function that coincides the Patterson vectors of the model with those of the unknown molecule. When sufficient agreement is reached the rotated model is then translated into the position of the unknown molecule in the unit cell, thus allowing for phasing; if the MR model does not display sufficiently high structural homology to the molecule in the crystal the rotation function will not be able to match the Patterson vectors of each molecule, and the calculation will fail. For information regarding the parameters used for the validation of X-ray diffraction data, see **Section A.651**.

1.10.3.5 Small Angle X-ray Scattering

Small angle X-ray scattering (SAXS) is similar to X-ray crystallography in that it involves the interaction of X-rays with the electrons in a sample to derive its 3D structural information[308]. In SAXS experimentation, a monochromatic X-ray beam incident on a solution sample of a protein scatters diffusely onto a 2D detector (**Figure 1.17A**). The produced scattering pattern represents an average of the scattering from all possible orientations of all molecules in the solution, therefore the scattering from the background solution must be subtracted from the sample to determine the scattering from the protein of interest[309]. As the protein in solution is randomly oriented, the

scattering intensity recorded depends only on the magnitude of the scattering vector and not the direction, and can be expressed by the equation $q = (4\pi\sin\theta)/\lambda$. In this equation, the scattering angle is defined as 2θ, and the wavelength of the incoming X-rays is λ. The scattering vector describes the change in direction of elastically scattered X-rays, therefore many images of the sample and the subtraction buffer must be collected and averaged to properly represent the molecular motions which occur in the solution[310]. Each image is then integrated about the beam center such that a scattering profile is produced; these 1D curves are defined as scattering intensity *I* versus scattering vector *q*. To obtain the scattering profile of the protein, the buffer scattering profile must be subtracted from the sample scattering profile. Scattering profiles can provide the molecule's radius of gyration (R_g) and molecular weight (MW) through Guinier analysis, a function which allows for observation of scattering at the smallest q values. Scattering profiles can also be adapted to Kratky plots, which are used to qualitatively assess the flexibility and/or the degree of unfolding in samples and are mathematically expressed as $q^2I(q)$ vs. q[311–313]. The shape of a Kratky plot dictates the protein's folding, globular proteins will have a Gaussian peak, unfolded proteins will have a plateau at high q, while prolate or partially unfolded proteins will have a combination of bell-shape and plateau. A normalized Kratky plot normalizes scattering profiles by mass and concentration by plotting $q^2I(q)/I(0)$ vs. q. Performing an inverse Fourier transform on the scattering profile provides the real space distribution function P(r), and the maximum dimension Dmax (the largest paired distance achieved between points on the molecule as determined by the P(r) function)[313]. In SAXS the P(r) function is used to describe the distances between all scattering points on the macromolecule, grouped into paired sets, in order to predict the overall shape of the molecule[314]. If sample purification and buffer subtraction is done properly, all of this information

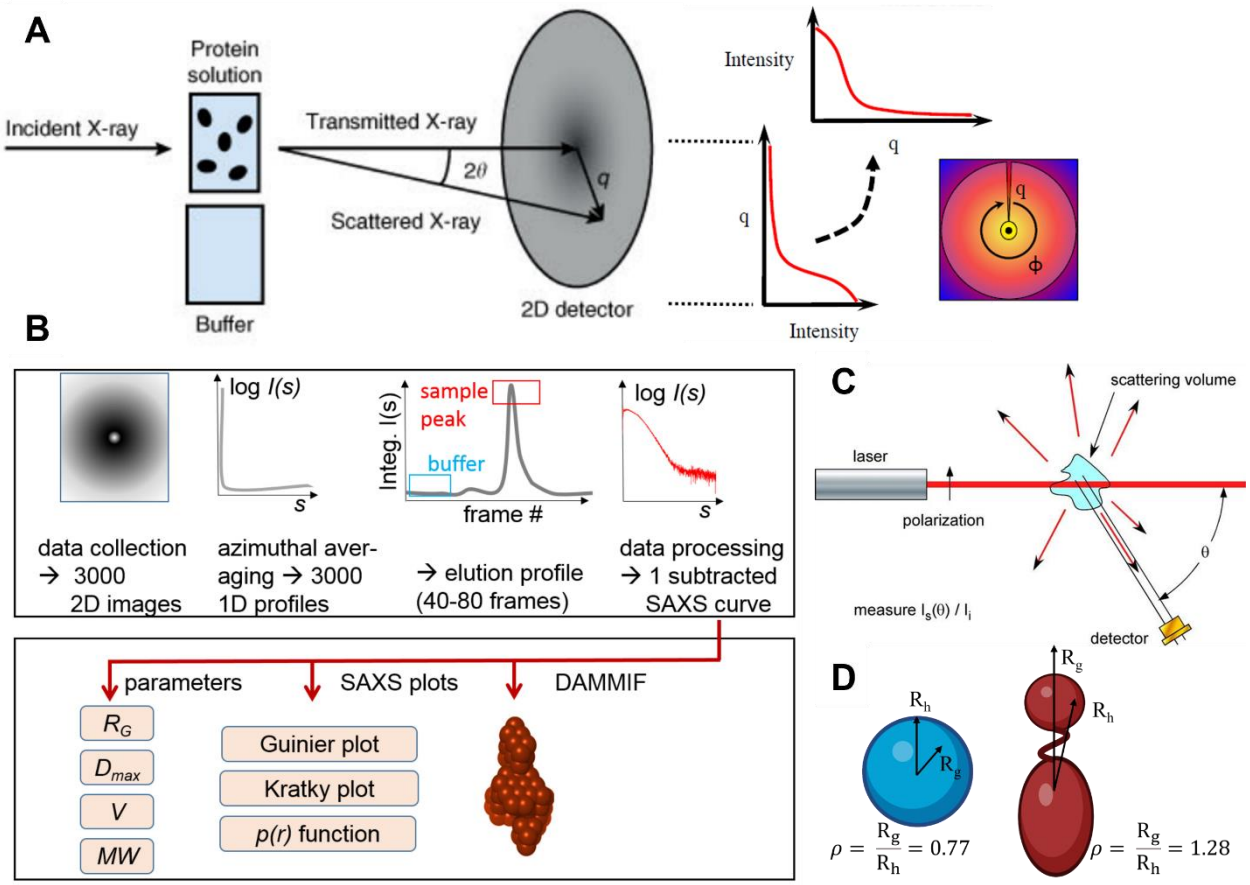

Figure 1.17 Diagrams of Protein SAXS, SEC-SAXS and MALS. **A)** A simplified diagram of protein SAXS data collection and initial processing. A protein sample is exposed to X-rays to produce a number of frames, each of which produce a scattering profile which directly relates the intensity of the scattering to the distance from the centre of the image, where intensity is radially averaged at every distance q of the 2D detected SAXS image[315]. These scattering profiles are then averaged and subtracted from the average scattering profile of the buffer alone. Figure adapted from Skou *et al.* 2014[308] **B)** An example of the SEC-SAXS data processing workflow, in which data is collected at the outlet of the SEC column, is averaged, and the integrated frames are displayed as a chromatograph. Optimal sample and buffer frames must be chosen to produce a scattering curve from which biophysical data of the protein of interest will be extracted. Figure adapted from Graewert *et al.* 2020[316]. **C)** Schematic of MALS, displaying the measurement of scattered light intensity (I_s) and its dependence on the angle of scattering and the intensity of the incident polarized laser light. Figure from Wyatt technologies (available at: https://www.wyatt.com/support/ask-the-expert-how-can-i-tell-if-i-have-fluorescing-or-absorbing-samples.html). **D)** Shape determinations can be calculated by dividing the R_g by the R_h, both of which can be measured through MALS[317]. Globular proteins will have a larger R_h than R_g, whereas proteins that are prolate and/or have extensions will typically have a larger R_g than R_h. Figure created with Biorender.com.

can be used to determine the size, shape, structure, oligomerization state, interactions, conformational heterogeneity, flexibility, and intrinsic disorder of biological macromolecules[316].

SAXS data is highly sensitive to sample quality and experimental technique; minimal amounts of protein aggregation, polydispersity of the sample, subtraction buffer mismatch, and radiation damage can greatly influence the output parameters[308,316]. These issues can be mitigated through the inclusion of size exclusion chromatography (SEC), where the sample will be eluted from an appropriate SEC column in-line with a flow cell where the eluate is exposed to the X-ray beam (**Figure 1.17B**). This allows particles in the sample to be separated by size prior to SAXS data collection, which aids in removing aggregates and separating polydisperse samples into individual species. As the eluent used will optimally be the subtraction buffer, highly accurate solvent frames can be collected prior to the void volume and after the total column volume to provide an optimal background subtraction[316].

1.10.3.6 Multi-Angle Light Scattering

Multi-angle light scattering (MALS) employs monochromatic incident light to cause Rayleigh scattering that is detected by angularly determined detectors (**Figure 1.17C**)[318]. When coupled with analytical SEC, it is a highly accurate technique for the determination of the molar mass of a molecule in solution, and can also determine the R_g of a molecule, the second virial coefficient A_2, and the translational diffusion coefficient which can be used to calculate the radius of hydration, R_h[318–320]. A_2 can provide a measure of the molecule's interactions with the solvent, and R_h provides the radius of the hypothetical sphere of aqueous solvent that diffuses at the same

rate as the macromolecule; when comparing R_g/R_h one can calculate a relative shape factor determination of the molecules (**Figure 1.17D**).

In a typical light scattering experiment, a collimated, single frequency, polarized light beam is used to illuminate a solution containing a suspension of macromolecules[320]. Measurements of intensity and angular dependence of the scattered light is performed by multiple detectors which are along the plane from which the polarized light beam is produced, as the electric field of the polarized light beam is produced perpendicular to the plane in which the light is incident on the sample. This is due to the refractive index increment which relates the concentration of molecules in solution to their ability to scatter light. As the detectors measure the intensity of light, the refractive index increment can be directly related to the square of the electric field of the scattered molecules. As two linked molecules would scatter light coherently unlike the incoherent scattering that would occur due to Brownian motion of two separate molecules in solution, the difference in coherent and incoherent scattering can be observed via intensity changes in detectors that are directly related to the mass of the molecule. In this way the change in light intensity carries information about the molar mass, while the change in the angular dependence of the scattering carries information about the molecular size and shape. Changes in angular dependence of scattering can be observed for particle sizes above 10 nm for incident light with a wavelength of 690 nm; effects due to intramolecular interference leads to a variation of the angle of the scattered light in the plane perpendicular to the plane of polarization[318,321]. The mathematical relationship describing the variation in intensity depends on the size of the particle, the wavelength of the light, and the observation angle. Therefore, size information can be retrieved from the angular dependence of the scattering intensity alone. Direct refractive index (dRI) detectors are typically coupled in-line with MALS detectors to measure the refractive indices of a solution in a flow cell

and the solvent in a reference cell to easily determine the refractive index increment produced from the particles of interest. Linking MALS with SEC allows for the sample to separated, where batch MALS experiments cannot determine the polydispersity of the sample[318–321]. SEC-MALS is analyzed similarly to a series of batch mode experiments but only if the peaks corresponding to the molecules of interest can be resolved, and therefore requires higher protein concentrations and purity.

1.10.3.7 Size Exclusion Chromatography linked to MALS and SAXS

SEC-MALS-SAXS is an equilibrium based SAXS method which links multiple biophysical techniques to accurately inform on particle shape and size[316]. Coupling SEC-SAXS with MALS and a dRI detector has become established as a robust method for acquiring highly accurate biophysical data[316]. MALS and dRI detectors provide a more accurate MW determination than SAXS and provide radius of hydration R_h values to couple with the highly accurate radius of gyration R_g (the radius obtained from the rotationally averaged volume of the macromolecule) values obtained from SAXS[313,316,322]. As well, shape factor reconstructions derived from the SAXS P(r) function can be confirmed with R_g/R_h values from MALS. For information regarding the parameters used for the validation of SAXS data, see **Section A.5.2**.

1.10.3.8 Advanced Neural Networks in Protein Structural Prediction Methods

Protein structural prediction has been developing alongside the advancement of computing power; neural networks were first applied to molecular biology in 1982 through “Perceptrons” used to distinguish translation initiation sites in DNA sequences, followed by another algorithm

which used a neural network for secondary structural prediction of proteins from their primary sequence in 1988[323–326]. Neural networks have advanced significantly since then, and a deep learning artificial intelligence (AI) algorithm for *de novo* structural prediction of proteins named AlphaFold (AF) provided a breakthrough in the accuracy of their output models[327,328]. AF and RoseTTAFold (RF) were published in 2020 and 2021, respectively, and use distinct novel deep learning attention-based algorithms to predict protein structures with high accuracy[327–329]. Most protein structural prediction methods rely heavily on information from structures deposited in the PDB that is integrated into protein structural prediction algorithms, however the method in which this wealth of structural information is used is essential for the reliability of the output model[324,327].

There are two main approaches that are used in structural prediction algorithms, template-based modelling (TBM) or template-free modelling (FM), also called *ab initio* or *de novo* modelling[324]. TBM algorithms copy and refine structural frameworks from related proteins to construct a model from a primary sequence in four steps: (1) identify a solved template that matches the query, (2) align the query and template, (3) construct the structural framework using the aligned region of the template structure, and (4) construct the unaligned regions based on secondary structure predictions and refine the global structure. Many modern TBM algorithms use both comparative modelling for models with close homology and threading to detect distant homologs using a Hidden Markov Model alignment. I-TASSER is an example of an accurate TBM which features some aspects of FM modelling[330]. FM modelling relies upon fragment assembly, which uses local structural fragments (1-20 aa for continuous lengths, or discrete 3 and 9 aa fragments) from unrelated proteins based on profile similarity to the query primary sequence[324,327]. These fragments are structured based on comparison of local structural features such as secondary structure, solvent accessibility, and torsion angles. The fragments are then assembled in a manner

that reduces the entropy of the conformational search space by replacing the backbone torsion angles for a region with either ideal bond angles or those from previously solved homologous structures; this comparison to experimental structures can help compensate for inaccuracies in the algorithm.

AF provided innovation in FM modelling by using a deep learning neural network which incorporates training procedures based on the evolutionary, physical and geometric constraints of known protein structures[328]. The architecture of the network is comprised of two main stages, firstly a network block that was termed an Evoformer which produces an array which compares number of sequences to number of residues of multiple sequence alignment results for the query to known structures, as well as an array that pairs number of residues (**Figure 1.18**)[328,331]. A number of attention-based and non-attention-based components are used in this evoformer block, allowing for a structural hypothesis to be built and continuously refined in this block in a two-track attention method in which information at the 1D sequence level and the 2D distance map level is iteratively transformed. Secondly, a structure module is used to introduce an explicit 3D structure by rotating and translating each residue of the protein in global rigid body frames. It also allows for chain structure breaks to optimize local refinement of all parts of the structure. While AF uses a two-track-attention method in the Evoformer, RF adapted a three-track attention (TTA) algorithm to have the 1D sequence comparisons, the 2D distance maps and the 3D coordinates training simultaneously and iteratively to more effectively extract sequence-structure relationships[329].

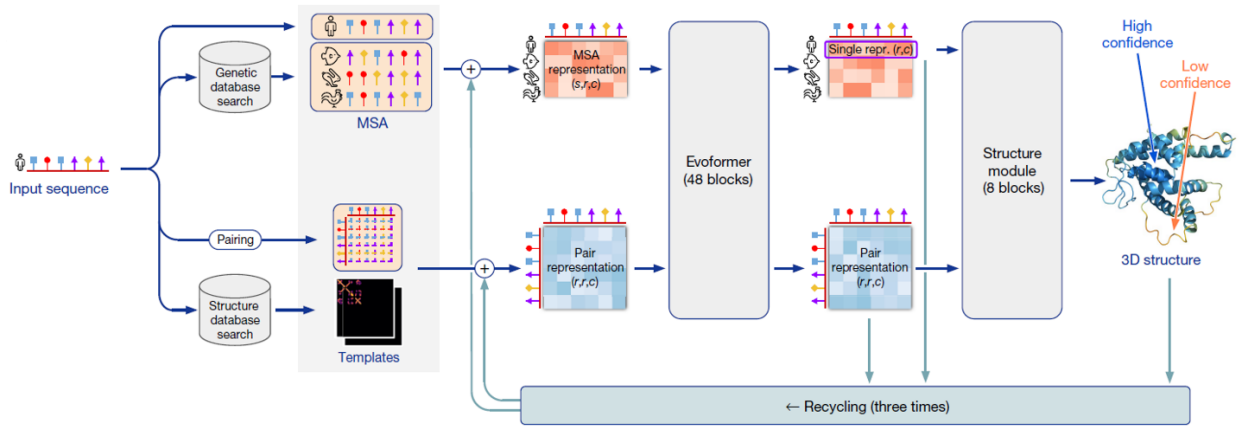

Figure 1.18 Architecture of the AlphaFold FM Modelling Algorithm. Double track attention is used to compare genetic multiple sequence alignment and structure pair representations in an iterative fashion to produce a structural model from an input primary amino acid sequence. Figure adapted from Jumper et al. 2021[328].

CHAPTER II: THE BIOPHYSICAL CHARACTERISATION OF THE CONJUGATIVE ENTRY EXCLUSION PROTEIN TRAG FROM F-LIKE PLASMIDS

2.1 Research Objectives

A high-resolution structure of TraG* has the potential to elucidate the mechanism by which TraG interacts with the cognate TraS of a recipient cell to affect Eex[7,67]. Determining methods to disrupt the Eex process and promote lethal zygosis through ceaseless conjugation, or to activate TraG and the Eex process and prohibit conjugation could allow for the development of novel classes of antibiotics[119,121]. This chapter describes the structural and biophysical study of TraG*, with the objective of creating a TraG* construct which facilitates protein crystallisation for the solution of a high-resolution structure of the protein. Initial studies of TraG* showed it was recalcitrant to proper crystallisation; the phase separating protein predictor (PSP)[246] software was first to identify a highly flexible N-terminal domain that was hypothesized to serve as a linker region between the two functional TraG domains[332]. N-terminal truncations of TraG* from both the F (TraG*$_{F}$) and the R100 (TraG*$_{R100}$) plasmids were designed based on these data.

Thermofluor assays, CD and CIU-MS were used to confirm the improved thermal, chemical, and conformational stability of the TraG*$_{R100}$ N-terminal truncation mutant relative to TraG*$_{R100}$. SEC-MALS-SAXS studies show the improved solubility of the truncated TraG*$_{R100}$ constructs based on a lowered propensity for aggregation resulting in improved data and a more reliable bead model reconstruction. However, diffraction-quality crystals of the mutant protein were not produced; SAXS data and other software such as predicted models from RF and AF confirmed the presence of the N-terminal flexible region but also indicated the presence of a dynamic C-terminal region[7,246,328,329,331]. Further CD and SEC-MALS-SAXS results corroborated the hypothesis that an additional truncation removing the C-terminal region yields a more stable

TraG* construct, however protein crystals suitable for X-ray diffraction remained elusive. The region responsible for binding TraS was hypothesized to be a dynamic loop and may be responsible for disrupting proper crystal formation in the N- and C-terminal truncation mutant. Glutathione S-Transferase (GST) pulldowns of $TraS_F$ peptides with all $TraG^*_F$ constructs were attempted to contrive a binding partner which could stabilize the proposed interacting loop region of $TraG^*_F$ and facilitate crystallisation of the complex. Additionally, other efforts were made to crystallize the different constructs of $TraG^*_{F\ \&\ R100}$, including the use of limited proteolysis through addition of trypsin and chymotrypsin to the crystallisation drop, and lysine methylation. Although an X-ray diffraction dataset suitable for solving the crystal structure of a TraG* construct was not achieved, these results provide a biophysical characterisation of $TraG^*_{F\ \&\ R100}$, yielding useful insight into the dynamic nature of the F-like conjugative T4SS.

2.2 Materials and Methods

2.2.1 Reagents and Equipment

All chemicals and antibiotics used in this study were purchased from Sigma-Aldrich, VWR, Thermo-Fisher, and BioBasic unless otherwise indicated. Restriction enzymes and cloning reagents were purchased from New England Biolabs (NEB); DNA ladders and protein markers were purchased from Thermo-Fisher. Plasmid vectors pET28a, pET28a:G^*_FHis, pBAD33:pelB_$TraG_{R100}$, and pGEX4T2 as well as *E.coli* BL21 (DE3) and DH5α were obtained from cryogenic lab stocks; competent cells were cultured and prepared as per **Appendix A.1**. Primers and the vectors pUCIDT:TraS were purchased from Integrated DNA Technologies (IDT), and pGEX4T1:$TraS_{P5}$ was purchased from BioBasic; all primers used in this study are listed in

Table S1. Qiagen, Hampton, Molecular Dimensions, and Anatrace crystallisation screens were used for initial vapour diffusion experiments.

Polymerase Chain Reaction (PCR) was performed using a TECHNE TC3000-G thermocycler. All sonication was carried out using a Sonic Dismembrator 5000 (Ficher Scientific). Liquid chromatography was performed utilizing an ÄKTA purifier 10S (Cytiva) using columns from Cytiva, Wyatt, and Phenomenex. Agarose and Sepharose resins for protein purification were from Fisher Scientific. Concentration and dialysis of proteins was performed using 30,000 Dalton (Da) MW cut-off (MWCO) centrifugal concentrators (Millipore, Corning). DNA quantification was estimated using a NanoDrop 2000 spectrophotometer (Thermo Scientific). Protein quantification assays were conducted using a Beckman Coulter DU-730 Spectrophotometer for UV_{280nm} readings and/or the Synergy H4 Hybrid microplate reader from BioTek for the Bicinchoninic Acid (BCA) assays (Pierce/Thermo Scientific). Native ESI-MS and CIU-MS was performed on a Waters Synapt G2 operating in positive ion mode. CD experiments were conducted on a J-815 CD Spectrometer from Jasco. Thermofluor experiments were performed on a RotorGene Q thermocycler from Qiagen and SYPRO Orange was used to detect protein unfolding (Life, ThermoFisher). Observations of crystallisation experiments were carried out using a Nikon SMZ1500-Fibre Lite MI 150 stereoscopic microscope, and in-house X-ray diffraction of protein crystals was performed on a Rigaku MicroMax-007 HF rotating anode X-ray generator with a Saturn944+ charge coupled device detector.

2.2.2 Analysis of $TraG_{F\ and\ R100}$ Using Software for the Detection of Disorder

The primary sequences of $TraG_{F\ \&\ R100}$, $TraS_F$, and a $TraG_F$-$TraS_F$ complex were input into AF (ColabFold v.1.5.2) and RF (Robetta Server) (the $TraG_F$-$TraS_F$ complex could not be entered into RF) for predictive modelling of the proteins, all using default settings[329,333]. Decisions in designing primers for cloning of TraG* mutants and TraS peptides were based on the results of these analyses. Later, $TraG_F$ was found in the PhaSePred database, allowing for a predictive sequence characterization based on a suite of different software[254]. Based on the catGRANULE results from PhaSePred, sequences of $TraG_{F\ \&\ R100}$ constructs were entered into the catGRANULE software for a more nuanced scoring[334].

2.2.3 Expression and Purification of Protein Constructs

2.2.3.1 Expression of TraG and TraS Variants and Lysis of Recombinant Cells*

All constructs were cloned as per **Appendix A.2**. A 10 mL feeder culture of BL21 (DE3) *E. coli* harbouring a pET28a plasmid containing the appropriate Hexahistidine (His)-tagged TraG* construct was grown in Luria-Bertani (LB) medium with 50µg/mL of Kanamycin (Kan) overnight at 37 °C. Information regarding the protein constructs used in this study are listed in **Table 2.1**. The feeder culture was used to inoculate 1 L of LB with 50 µg/mL of Kan and 1 mM glucose. After the 1 L culture had reached the mid-logarithmic growth phase (optical density at 600 nm (OD_{600nm}) = 0.5 – 0.8), ethanol was added to a final concentration of 1% (v/v) to induce a stress response and promote chaperone protein assembly[335]. Expression of the TraG* variant was then induced through addition of Isopropyl-β-D-thiogalactoside (IPTG) to a final concentration of 1 mM. An incubation was performed for 16 hours (hrs) at 18 °C with shaking at 200 rotations per

Table 2.1: A List of the TraG* Protein Constructs and TraS Peptides used in this study, as well as the plasmid they were expressed from and their respective abbreviations. Full protein sequence information is listed in **Appendix A3**.

Protein	Plasmid	Abbreviations	Residues
N-terminally His_6-tagged $TraG^*_F$	pET28a:$HisG^*_F$	$HisTraG^*_F$, $HisG^*_F$	Ala452-Glu938
N-terminally His_6-tagged $TraG^*_{R100}$	pET28a:$HisG^*_{R100}$	$HisTraG^*_{R100}$, $HisG^*_{R100}$	Ala452-Gln940
N-terminally His_6-tagged $TraG^*_F$ N-terminal truncation	pET28a:$HisG^*_{F\ 497-938}$	$HisTraG^*_{F\ 497-938}$, $HisG^*_{F\ 497-938}$	Thr497-Glu938
N-terminally His_6-tagged $TraG^*_{R100}$ N-terminal truncation	pET28a:$HisG^*_{R100\ 497-940}$	$HisTraG^*_{R100\ 497-940}$, $HisG^*_{497-940}$	Thr497-Gln940
N-terminally His_6-tagged $TraG^*_F$ N- and C-terminal truncation	pET28a:$HisG^*_{F\ 497-863}$	$HisTraG^*_{F\ 497-863}$, $HisG^*_{F\ 497-863}$	Thr497-Asn863
N-terminally His_6-tagged $TraG^*_{R100}$ N- and C-terminal truncation	pET28a:$HisG^*_{R100\ 497-863}$	$HisTraG^*_{497-863}$, $HisG^*_{497-863}$	Thr497-Asn863
N-terminally GST-tagged TraS peptide 1	pGEX4T2: $TraS_{P1}$	$GSTTraS_{P1}$	Met1-His17
N-terminally GST-tagged TraS peptide 2	pGEX4T2: $TraS_{P2}$	$GSTTraS_{P2}$	Cys45-Thr67
N-terminally GST-tagged TraS peptide 3	pGEX4T2: $TraS_{P3}$	$GSTTraS_{P3}$	Lys84-Val101
N-terminally GST-tagged TraS peptide 4	pGEX4T2: $TraS_{P4}$	$GSTTraS_{P4}$	Met114-Gly130
N-terminally GST-tagged TraS peptide 5	pGEX4T1: $TraS_{P5}$	$GSTTraS_{P5}$	Gln152-Lys173

minute (rpm), following which cells were then pelleted at 5,000 x g for 30 min at 4 °C. Cell pellets were resuspended with immobilized metal affinity chromatography (IMAC) running buffer (RB) (20 mM Tris-HCl pH 7.5, 200mM NaCl, 10% (v/v) glycerol). Protease inhibitors phenylmethane sulfonyl fluoride (PMSF), Benzamidine HCl and Aminocaproic acid were added to final concentrations of 1.25 mM, 5 mM and 5 mM, respectively. Lysis was performed through sonication at 25% amplitude with a 20 s on, 40 s off cycle for 10 min; the lysate was then separated from cell debris through centrifugation at 25,000 x g for 30 min at 4 °C.

GST-tagged TraS peptides were all expressed in a similar fashion to TraG* constructs with 50 µg/mL Ampicillin (Amp) as the selective antibiotic instead of Kan, however 1% (v/v) ethanol was not added prior to induction with IPTG to a final concentration of 1 mM. As well, the buffer used for lysis was a GST RB (50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA pH 7.5), and only 1.25 mM of PMSF was added to this buffer prior to resuspension of the cell pellet.

2.2.3.2 Purification of His-tagged TraG Constructs*

Ni^{2+} IMAC was performed for His-tagged proteins on a XK26 Nickel-nitrilotriacetic acid (Ni-NTA) agarose column. The lysate was loaded into the column at 1 mL/min and impurities were washed from the stationary phase with 100 mL of IMAC RB with 30mM Imidazole (Imid), and then 75 mL of IMAC RB with 50 mM Imid. Fractions of the eluted His-tagged protein were collected at 3 mL volumes using 75mL IMAC RB with 300 mM Imid. A column cleaning elution was performed with 2 CV of IMAC loading buffer with 1 M imidazole. Fractions from the affinity purifications with high quantities of TraG* relative to impurities based on the UV_{280nm} chromatogram and confirmed via SDS PAGE were pooled and dialyzed in 4 L of 20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol, pH 7.0; dialysis buffer was replaced three times over 24 hrs at 4 °C. In some instances, following dialysis and quantification the His-tag was cleaved using thrombin at a ratio of 0.5 U (0.16 mg) of thrombin to 1mg of protein, and the sample was left on a rotator overnight at 4°C. Un-tagged TraG* was collected in the flow-through after being added to the IMAC column with IMAC RB. When further secondary purification was required, SEC was performed using a HiPrep Sephacryl 16/60 S100 HR column (Cytiva) using 20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol pH 7.0 as the running buffer.

2.2.3.3 Purification of GST-tagged TraS Peptides

GST affinity purification of GST-tagged TraS peptides was performed on a XK26 GST Sepharose column. The lysate was loaded into the column at 1 mL/min and impurities were washed from the stationary phase with 80 mL of GST RB. Elution of purified GST-tagged peptides was performed with 40 mL of GST elution buffer (20 mM Tris-HCl, 150 mM NaCl, 10 mM reduced Glutathione (GSH), pH 8.0); 4 mL elution fractions were collected. Fractions from the affinity purifications with high quantities of GST-TraS relative to impurities based on the UV_{280nm} chromatogram and confirmed via SDS PAGE were pooled and dialyzed in 4 L of 20 mM Tris-HCl, 150 mM NaCl, pH 7.5; dialysis buffer was replaced three times over 24 hrs at 4 °C.

2.2.4 Visualization of Proteins using SDS PAGE

Samples of interest were aliquoted and appropriately diluted to a 15 µL volume, then mixed with 5 µL of 4 X SDS buffer (160 mM Tris-HCl pH 6.8, 4.0% (v/v) SDS, 20% (v/v) glycerol, 0.0012% (w/v) Bromophenol blue, 20% (v/v) β-mercaptoethanol). After boiling at 100 °C for 5 min, 10 µL of each sample was loaded onto a 12.5% SDS PAGE gel and electrophoresis was performed. The gel was stained with Coomassie solution (50% (v/v) methanol, 10% (v/v) glacial acetic acid, 0.1% (w/v) Coomassie Brilliant Blue R250) to visualize protein bands.

2.2.5 Quantification of Proteins

TraG* samples were quantified by the Edelhoch method and by BCA assay using the manufacturers protocols[336,337]. In the Edelhoch method, the molar absorption coefficient (ε) was

determined for TraG* constructs by denaturing protein in 6 M Guanidine HCl (GdHCl) in a ratio that would provide a concentration between 0.1-1 mg/mL protein yet retain a volumetric 10:1 GdHCl to protein ratio, and then measuring the A_{280nm} of the denatured protein. The ε_{280nm} of the denatured protein was predicted using Protparam[338], then divided by the A_{280nm} of the denatured protein and multiplied by the A_{280nm} of the native protein in buffer diluted to the same concentration to determine the ε_{280nm} of the native protein. The Beer-Lambert equation could then be used to determine concentration when the sample is measured spectroscopically at A_{280nm}. For other proteins, a BCA assay was used in which 12.5 µL volumes of various dilutions of the protein sample were mixed with 237.5 µL of the working reagent and loaded on a 96 well plate (Sarstedt). After incubation at 37 °C for 30 minutes (mins) the OD_{562nm} of each sample was determined using the Synergy H4 Hybrid plate reader. Protein samples were stored either at 4 °C for short term use or flash frozen in liquid N_2 and stored at -80 °C.

2.2.6 Analysis of TraG Structural Properties*

2.2.6.1 Thermofluor Assay

To quantify the beneficial properties of buffers and salts in aiding the stability of proteins, and to compare the thermal stability of N-terminally His-tagged TraG* (HisG*) and N-terminally His-tagged TraG*$_{497-940}$ (HisG*$_{497-940}$) from F and R100 plasmids, thermofluor assays were performed. The His tag was left intact for this experiment and many of the described biophysical experiments as further processing of HisG* often resulted in aggregation, therefore protein constructs were left tagged to allow for a proper comparison. Thermofluor involves the use of a fluorescent hydrophobic probe to determine optimal solvent-protein interactions using the melting

temperature (T_m) of the protein as an indicator of stability[255–258]. SYPRO Orange dye was used as the probe, and it was prepared through serial dilution from a 5,000X commercial stock to a final concentration of 10X in the buffers of interest, with 1 mg/mL as the final concentration of TraG* in a total volume of 50 µL. Fluorescence output was measured on the thermocycler with the excitation wavelength set to 470 nm and the emission wavelength detected at 610 nm, with a gain of 7; samples were heated from 25 °C to 99 °C at a ramp rate of 1 °C/min. The choice of buffers used for this assay was inspired by Seabrook and Newman, 2013[258]. A total of 14 buffers were tested at 50 mM concentration with either 50 mM NaCl or 200 mM NaCl, and each condition was performed in triplicate, with a lysozyme positive control in 1X Phosphate buffered saline (PBS) pH 7.4, dye-only negative control and a protein-only negative control in every experiment. This was repeated for both proteins to gather a total of six data sets for each buffer condition. The data was analysed using DSFworld33 which allowed for crude normalisation, averaging of the curves to produce more accurate T_m values and aided in creating graphs of the output thermal melting curves.

2.2.6.2 Circular Dichroism

The molar ellipticity (mdeg) of HisG*$_{R100}$ mutants were measured by CD spectroscopy at concentrations of 2 µM over wavelengths 190-260 nm. To normalize the CD spectra of the proteins, the spectrum of the solvent (10 mM MES pH 6.5, 10% (v/v) glycerol) was subtracted from the experimental spectra. All spectra were obtained through a continuous scan performed at 100 µm/min with mdeg measurements every 0.1 nm, with an accumulation factor of 8. Urea denaturation experiments were performed to provide an indication of protein chemical stability when comparing HisG* with and without the putative IDR, as denaturing studies are common for

determining the relative stability of protein mutants[340–342]. HisG* variants were added to solutions with different urea concentrations such that the final concentration of protein was 2 μM. Samples were incubated at 25 °C for 1 hour prior to CD measurements. Urea concentrations of 0.5, 1, 2, 3, 4, and 5 M to determine the approximate range for protein denaturation. Spectra measurements were performed with the same scanning protocol as stated above, and all results were normalized to their respective solvent conditions. Deconvolution of all CD data was performed using BeStSel,[264] an algorithm for protein fold recognition and secondary structural determination using input CD spectra, unique in its ability to distinguish parallel from antiparallel β-sheets and useful in converting raw data measurements of mdeg to Δε based on concentration, MW of the protein, and path length (0.1 cm).

2.2.6.3 CIU-MS

HisG* samples were prepared for MS analysis using a 5 kDa MWCO dialysis cassette to buffer exchange the protein at 4 °C into 100 mM ammonium acetate (MS grade), pH 6.6. Native MS analysis was performed with 5 μM HisG* or TraG*$_{497\text{-}940}$ flowing at a rate of 5 μL/min. In the CIU-MS experiments, data was collected from 5 V trap CE to 150 V in 5 V increments, with an acquired m/z range of 2,000-5,000 with no manual trapping. Capillary voltage was set to 3.0 kV, the sampling cone was 150.0, with source and desolvation temperatures at 120 °C and 250 °C respectively, cone gas flow 71.0 L/hr, nanoflow gas pressure at 2 bar, desolvation gas flow at 600 L/hr, the transfer CE was 10.0 V, the trap gas flow was 4.0 mL/min, the IMS wave delay was 1 ms with a wave height start of 10 V and a wave end height of 40 V.

2.2.6.4 SEC-MALS-SAXS

HisG*, HisG*$_{497\text{-}940}$, N-terminally His-tagged TraG*$_{497\text{-}863}$ from the R100 plasmid (HisG*$_{497\text{-}863}$), and HisG*$_{F\ 497\text{-}863}$ were purified as described (**Section 2.2.3.2**), however protein samples were dialyzed using a 25 kDa MWCO dialysis membrane into 20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol, pH 7.0, and further dialyzed into 30 kDa MWCO concentrator using a 10X dilution of a buffer stock (200 mM HEPES, 1 M NaCl, 50% (v/v) glycerol, pH 7.0, 0.2 µm membrane filtered). Five hundred milliliters of this buffer stock was chilled at -20 °C and was packaged into an insulated box along with 1,000 µL of 5 mg/mL HisG*, 1,000 µL of 5.6 mg/mL HisG*$_{497\text{-}863}$, and 1,000 µL 5.6 mg/mL HisG*$_{F\ 497\text{-}863}$, as quantified using the Edelhoch method. These were driven directly to the biophysics collaborative access team (BioCAT) facility at the Advanced Photon Source to prevent aggregation seen in previous experiments[7] and to improve data quality. The samples were centrifuged at 10,000 x g at 4 °C for 10 mins prior to experimentation. SAXS diffraction data was collected using the parameters seen in **Table 2.2**. Samples were injected onto a Superdex 200 10/300 Increase SEC column (Cytiva) using an Agilent Inifinity II HPLC, then sent through a Wyatt DAWN Heleos II MALS instrument and a Wyatt Optilab T-rEX dRI detector, and finally into the SAXS flow cell in the path of the synchrotron beamline. HisG*$_{497\text{-}940}$ data was collected at an earlier date with a sample at 5.2 mg/mL using a similar experimental method[7]. Data were analyzed using the RAW and ATSAS software packages[343,344]. *Ab-initio* bead models were generated from the output files of 20 cycles using the integrated DAMMIF tool with DAMAVER averaging and refinement, and clustering using DAMCLUST[345–347]. Bead models were aligned to RF and AF models using DAMCLUST and were otherwise generated with the same parameters[329,333]. Ensemble optimisation methods (EOM) was performed using ATSAS online (v2.1) to generate models which characterize the

conformational space of the flexible system[348,349]. No rigid bodies or distance constraints were defined as no regions have solved protein structures or were of high confidence in predictive models, RANCH was set to generate 10,000 models, FFMAKER was left with default settings and GAJOE was set to take models from RANCH and attempt 1,000 generations with 50 ensembles. The background constant was adjusted to improve the fit of the models to the data, curve repetition in the ensemble was maintained, and the final ensemble size was not fixed.

Table 2.2 SEC-MALS-SAXS Data Collection Parameters for HisG*, HisG*$_{497\text{-}940}$, HisG*$_{497\text{-}863}$, and HisG*$_{F\ 497\text{-}863}$, which were at concentrations of 5.0, 5.2, 5.6 and 5.6 mg/ml, respectively.

Data collection parameters	
Instrument	BioCAT (Sector 18, APS)
Detector	Eiger2 XE 9M
Wavelength (Å)	1.033
q-measurement range (1/Å)	0.0028 to 0.42
Exposure Time (s)	0.5
Size Exclusion Column	Superdex 200 10/300 Increase
Flow rate (mL/min)	0.6
Temperature (°C)	20
Loaded volume (µL)	300
Buffer	20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol, pH 7.0

*2.2.7 GST-TraS Pulldowns of TraG**

Purified HisG*$_F$ constructs and GST-TraS$_F$ were buffer exchanged in a 5 kDa MWCO micro-dialysis chamber to PBS RB (1X PBS pH 7.4, 1 mM PMSF, 5% (v/v) glycerol). GSH agarose beads were made into a slurry prior to aliquoting 25µL into a microcentrifuge tube, then prepared by washing three times with 750 µL GST loading buffer and removing the supernatant after centrifuging at 7,000 x g at 4 °C for 2 mins. GST-TraS$_F$ was added to the GST agarose beads

to a volume that would equate 100 µg based on the concentration; they were then rotated for 1 hr at 4 °C to promote binding. The samples were then spun at 700 x g at 4 °C for 2 mins (low speed to prevent protein aggregation), and the supernatant was decanted with an aliquot collected for SDS PAGE. The beads were washed once with 750 µL of GST loading buffer, then centrifuged, decanted and washed with 750 µL of PBS RB. HisG*$_F$ constructs were diluted to 7.5 µM and 100 µL were added to the GST-TraS$_F$-bound beads, then rotated at 4 °C for 2.5 hrs. After the pulldown incubation, the beads were centrifuged at 700 x g at 4 °C for 2 mins, the supernatant was decanted with an aliquot collected for SDS PAGE. The beads were then washed 6 times with 250 µL PBS RB with aliquots taken for each wash after centrifuging at 700 x g at 4 °C for 2 mins, following which the supernatant was decanted. After the final wash and removal of the supernatant, the beads were transferred to a new tube and 7.5 µL of 4 X SDS loading dye and 7.5 µL of double distilled Nanopure water (ddH_2O) was added to the beads for running on an SDS PAGE gel for confirmation of binding. This scheme was also reversed to perform a His-pulldown, with Ni-NTA agarose beads and HisG*$_F$ constructs as the bait and the GST-TraS$_F$ constructs as the potential binder.

*2.2.8 Crystallisation Trials of TraG**

2.2.8.1 Screening and Optimization of Crystallisation Conditions

HisG* and untagged TraG* protein constructs were loaded with concentrations of 1–40 mg/mL in 0.75 µL volumes to the sitting drop wells of a 96 well plate from Axygen. A 90 µL volume of a screening solution was added to the reservoir well, and 0.75 µL of that reservoir solution was mixed with the protein solution drop in the corresponding sitting well. This was

performed using the solutions from the MCSG -1, -2, -3, and -4 kits, as well as PUREPEG and TOP 96 from Anatrace, as well as the Morpheus I, II and III screens from Molecular Dimensions and the JCSG+ screen from Qiagen, and GRAS screens (1-8), the MembFac screen and the Natrix screen from Hampton. Following quantification of a purified protein batch a pre-crystallisation test (PCT) from Hampton was performed to determine whether the concentration of the protein was suitable for crystal trials, and concentration was adjusted based on the results. These plates were sealed to allow vapour diffusion to occur and were kept in a place that experiences little vibration at both 4 °C and 19 °C. Results were checked periodically under a stage microscope with a birefringent lens.

Conditions that were determined to produce TraG* crystals were optimized through altering precipitant concentration, protein concentration, pH of the buffer, salt concentration, temperature and drop sizes in 24 well hanging drop plates. Typically, the plate was set up with 1 mL of mother liquor in the reservoir, with 1 µL of this solution placed on a plastic coverslip. Then 1 µL of the protein solution was added to the reservoir solution droplet, the coverslip was placed onto the corresponding reservoir such that the drop was hanging above it, and the chamber was sealed using high vacuum grease.

In some 96 well plates, random matrix microseeding (rmms) would be performed in an attempt to improve likelihood of crystal nucleation using a 10^5 diluted lysozyme microseed stock, where 0.15 µL of seed was added to a sitting drop containing 0.75 µL of protein and 0.6 µL of reservoir solution[350]. All lysozyme seed stocks were created from looping crystals made in 24 well hanging drop plates which were made using 1 µL of 20 mg/mL lysozyme with 1µL of the reservoir solution (100 mM Sodium Acetate (NaAc) pH 4.6, 600 mM NaCl, 25% (w/v) polyethylene glycol (PEG) 3350). A single large crystal (~300−500 microns) was looped, moved to a 1 µL droplet of

the reservoir solution, then added to a microfuge tube containing the reservoir solution and a polytetrafluoroethylene (PTFE) seed bead (Hampton), and then vortexed for 1 min prior to making the appropriate dilution. 5-amino-2,4,6-triiodoisophthalic acid (I3C) was also used as a nucleating agent; if incorporated into the crystal structure properly, the tri-iodine electron density can be used to solve the phase problem through single wavelength anomalous dispersion[351]. I3C was prepared by dissolving it in 2 M LiOH to a concentration of 0.5 M; pH was adjusted to 7.5 with HCl. This was then added to the protein buffer stock, then diluted 1:1 with protein such that the final concentration of I3C was 5-30 mM. This solution was then used as the protein stock for crystal screening, in some instances with the addition of lysozyme seed for rmms.

2.2.8.2 Modifications to TraG Constructs to Aid in Crystallisation*

In situ proteolysis crystallisation experiments were performed following the protocol of Dong *et al.* (2007)[352] using HisG* and untagged TraG* constructs. Either trypsin or α-chymotrypsin was prepared at 2 mg/mL in 1 mM HCl, 2 mM $CaCl_2$ and added to the protein sample at a ratio of 1 µg protease:100 µg experimental protein prior to pipetting the protein droplet; otherwise, experiments were prepared for crystallisation as described in **Section 2.2.8.1** and monitored.

Lysine methylation of HisG* and untagged TraG* constructs was performed chemically as per Walter *et al.* (2006)[353]. Dimethylamine-borane complex (ABC) was freshly prepared in 4 °C ddH_2O to 1 M and 20 µL was added per 1mL of concentrated protein. Formaldehyde was prepared to 1 M and 40 µL was added to the protein sample, and the solution was mixed at 4 °C for 2 hrs. Another 20 µL of 1 M ABC and 40 µL of 1 M formaldehyde were added and another 4

°C, 2 hr incubation was performed. 10 µL of 1 M ABC was added and the protein was left to incubate at 4 °C overnight. Afterwards a 5,000 x g, 4 °C, 10 min incubation was performed to remove aggregates; the supernatant was isolated, quantified and qualified. If necessary, a SEC purification (as described in **Section 2.2.3.1**) was performed to isolate methylated proteins and remove further aggregates.

2.2.8.3 Crystal Diffraction Conditions

Potential TraG* crystals were looped with appropriately sized microloops (Mitegen), added to a cryogenic solution which consisted of conditions mimicking that of the reservoir but with a discrete addition of either glycerol, ethylene glycol, or an increased amount of the same precipitant used in the reservoir, typically to a final concentration of 5–30 % (v/v). The crystal was flash frozen in liquid N_2 and mounted on the system where diffraction screening was performed for 10 seconds of X-ray exposure with 4 images collected at 60.0mm crystal-to-detector distance with angles from 0°-2° and 70°-72°.

2.3 Results and Discussion

*2.3.1 Structural Prediction Software Indicates Dynamic Regions in TraG**

The PScore software[246] for predicting phase separating IDRs provided insight in determining which region of TraG was initially targeted for mutation or deletion to improve protein stability. The region from residues 447-498 were predicted to have phase separating qualities in $TraG_{F\ \&\ R100}$, with 27 of these residues surpassing the PSP algorithm's score threshold

for the designation of residues as intrinsically disordered (4.0)[246], thus providing $TraG_F$ and $TraG_{R100}$ with overall scores of 4.06 and 3.74, respectively (**Figure S2**). This region was a logical target for deletion in the TraG* construct based on this result; in the full-length protein it likely serves as a flexible linker that connects the membrane-bound portion of TraG and periplasmic TraG*, permitting the dynamic movements TraG* likely to be necessary in contacting both TraN and TraS. When the IM bound domain is removed, the highly flexible linker region remains on the N-terminus of TraG* and is the likely cause for destabilization of the protein, thus $TraG^*_{497\text{-}940}$ was designed. The lower P score of $TraG_{R100}$ compared to $TraG_F$ deemed it more favorable to undertake further experiments.

When AF was made available the predicted model of $TraG_F$ was found in the AF database (Code: AF-B1VCA9-F1) and indicated high-medium confidence in the prediction of the N-terminal membrane bound domain with medium-low confidence throughout the C-terminal periplasmic domain (**Figure 2.1A**)[328,331]. The overall confidence value of the models from RF was slightly higher for $TraG_{R100}$ at 0.55 vs. $TraG_F$ at 0.52. Notable regions which had high error in their predictions were the hypothesized flexible linker region from Ala452-Ala496 and the non-homologous region from Arg610-Asp673; both were predicted to be loop regions in both AF and RF models (**Figure 2.1A & B**). The putative linker region was also predicted to have a small 3-stranded β-sheet in both RF and AF. These algorithms also indicate that the final 100 residues of the protein are difficult to model in a defined space despite certainty in the secondary structure, and this is best illustrated by the per-residue angstrom error estimate plot from RF (**Figure 2.1D**). There were some distinctions in this region when modelling TraG* and $TraG^*_{497\text{-}940}$ truncation constructs; RF and AF predicted this region to be a large α-helix with no particular lowest entropy conformation regardless of *in silico* N-terminal deletions and sequence changes (**Figure S3**).

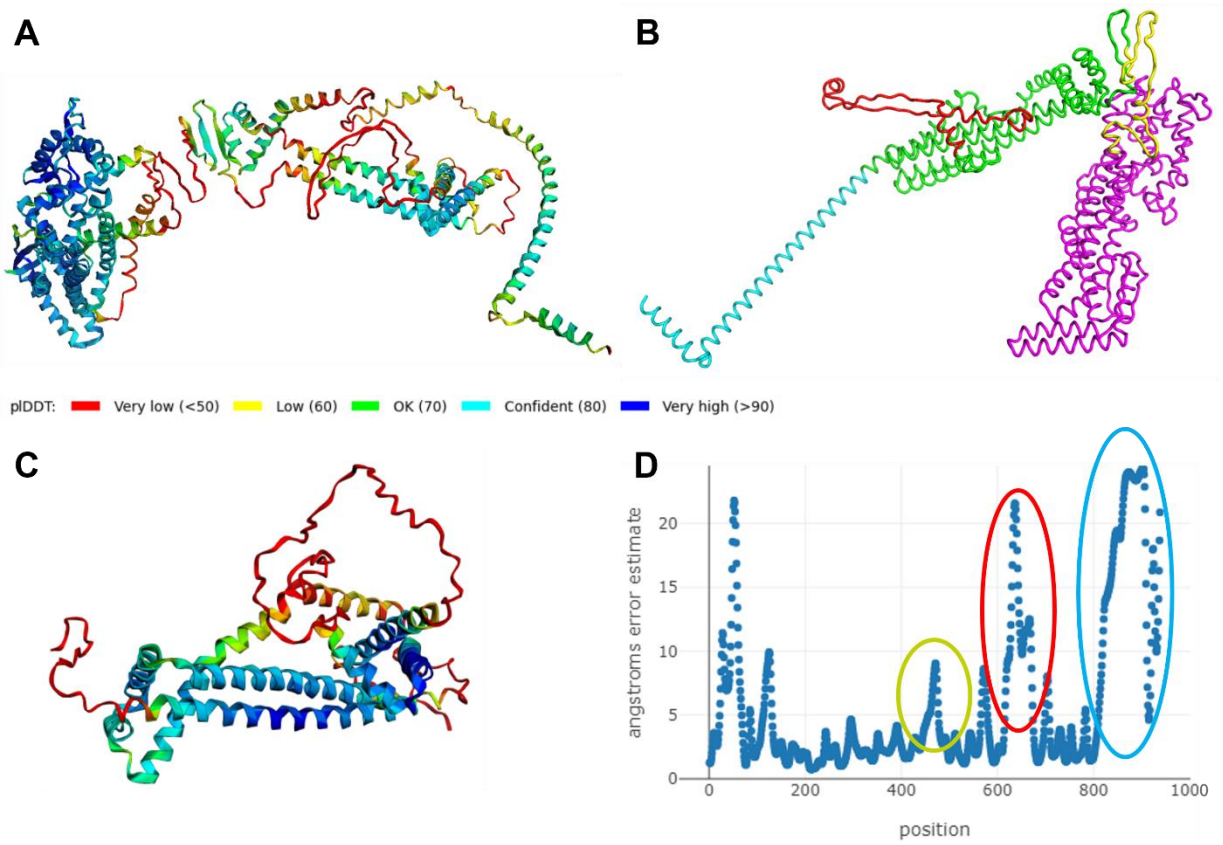

Figure 2.1 AlphaFold and RoseTTAFold Predicted Models of TraG$_{R100}$. Predicted models of TraG$_{R100}$ from **A**) AF and **B**) RF. **C**) The AF model of TraG* $_{R100\ 497\text{-}863}$; both AF models are coloured based on predicted local distance difference test (pLDDT) score (as per the bar), whereas the RF model is coloured based on domains and truncations. The N-terminal IM domain is in purple, TraG* is green, with the flexible linker region coloured in yellow, the region responsible for interacting with TraS is in red and the C-terminal truncation is in cyan. **D**) The per-residue angstrom error estimate graph for TraG$_{R100}$, with regions of interest circled in a similar colour scheme as described.

Based on this modelling data and other experimental data, TraG*$_{497\text{-}863}$ was designed and modelled (**Figure 2.1C**) prior to cloning and performing experiments with this truncated protein.

Based on the AF and RF models which predict the protein as mainly consisting of extended α-helices, it is visualized why TraG* aggregates commonly and is recalcitrant to forming diffraction-quality crystals; non-covalent interactions promoted by crystallisation conditions may induce the formation of non-specific coiled-coils[354]. AF and RF both predict the linker region of

TraG to contain a small β-sheet surrounded by loop regions; this moiety may fold cryptically in the context of the full T4SS, then uncoil like a spring to extend the periplasmic TraG* to its interacting partner TraS in the IM of the recipient cell. Another insight from the AI-predicted structures is the region of TraG* known to interact with TraS, as it is predicted to be an extended loop. This topology could be required for its functionality wherein this region may become structured only when interacting with its binding partner. Determining the crystal structure of a TraG* construct that includes this interacting motif will elucidate the role of this extended loop in TraG-TraS interactions. This was the reasoning behind performing GST-pulldowns with TraS peptides; if an interacting peptide could be found that stabilizes this domain it has the potential to form diffraction-quality crystals, as this region may also be disrupting proper crystal formation.

Phyre2 was used to identify TM and hydrophilic regions of $TraS_F$ that may be important for forming a binding interface for $TraG_F$ (**Figure S4**). $TraS_F$ and a $TraG_F$-$TraS_F$ complex was also modelled with AF and displayed the same scheme of 4 TM α-helices, with an α-helix at each terminus (**Figure 2.2**). The model of the complex corroborates the known interacting region of $TraG_F$[113], however as the AF algorithm does not identify regions as TM the predicted TraS interface consists mainly of the 2nd and 4th TM helices, and partially with the C-terminal α-helix. From these predictions, peptides of 16-20 residues in length were designed for GST pulldowns (**Table 2.1**).

PhaSePred was used to re-assess the structural properties of the TraG sequence through a variety of disorder prediction algorithms (**Figure 2.3)**. PhaSePred calculates a score from 0 to 1 based on the cumulative scores of each program executed, where 0 represents a predicted low propensity for the protein to undergo LLPS and 1 is the highest predicted certainty; $TraG_F$ had a score of 0.980, indicating a high propensity for LLPS[242]. PhaSePred uses PScore as part of the

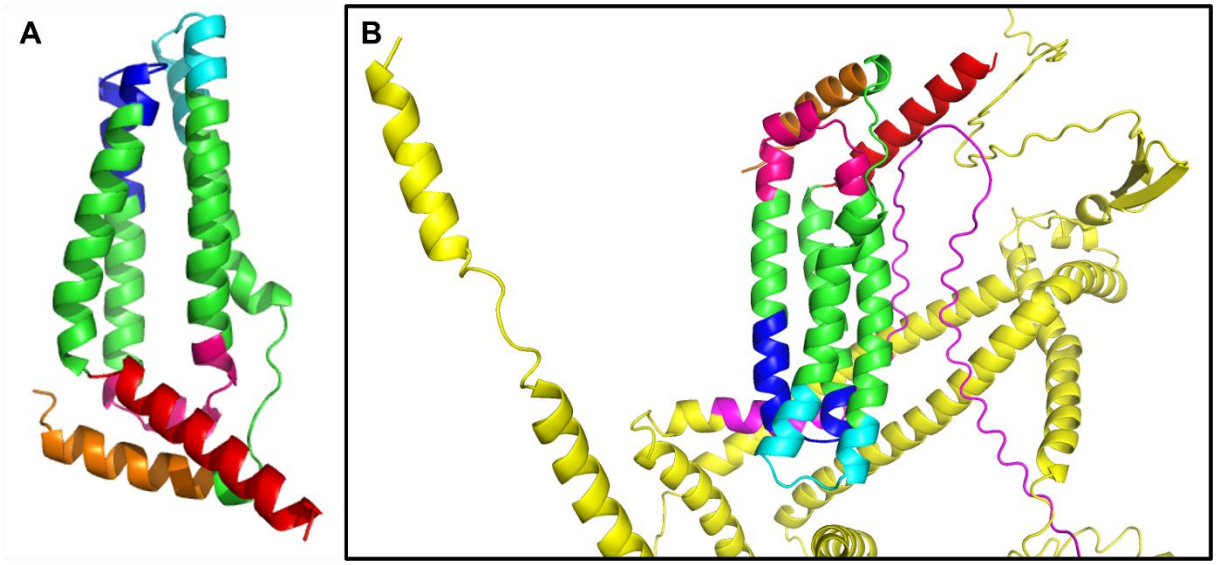

Figure 2.2 $TraS_F$ AlphaFold Models, where peptides designed for GST-pulldowns are coloured as follows: P1 (orange), P2 (cyan), P3 (pink), P4 (blue), P5 (red). **A**) The $TraS_F$ AF model, which was highly similar in structure to the RF prediction which had a score of 0.72. **B**) An AF model of the $TraG_F$-$TraS_F$ complex, where TraG is coloured in yellow with the interacting region from Arg610-Asp673 coloured in purple.

LLPS scoring analysis and recognized the linker region as being a PS IDR as expected. PhasePred's entry for $TraG_F$ showed other software also predicted the disorder present in the linker region, with the catGRANULE[334], PLAAC[355], and SEG[356] algorithms recognizing portions of this region as having the propensity to perform LLPS (Pro447-Val456, Glu468-Met486), having prion-like amino acid composition (Asn459-Thr495), and being of low complexity (Ser434-Ser454), respectively. Portions of the region responsible for interacting with TraS and the C-terminal helix were also predicted by catGRANULE[334] as having a high propensity for LLPS (Ser616-Ser668, Asp807-Asn905), and ESpritz[357] recognized a portion of the C-terminal helix as being disordered (Lys898-Glu938). The N-terminal domain was predicted to have high hydropathy by localCIDER[358], as expected for a TM protein, with some small regions exceeding the 0.5 score threshold in $TraG^*_F$. DeepCoil[359] did not recognize any regions as having the propensity for

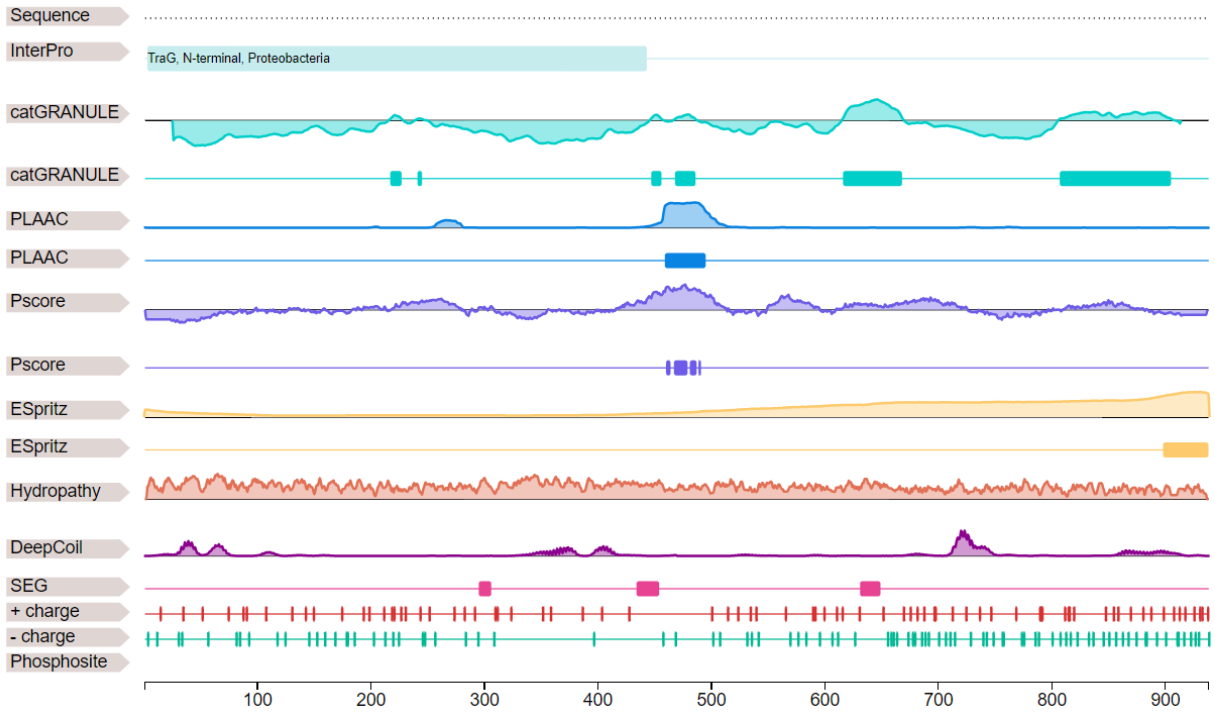

Figure 2.3 Predicted Propensity of the $TraG_F$ Primary Sequence for LLPS, as determined by PhaSePred[242].

forming coiled coils, which is not congruent with hypotheses that the many α-helices of TraG* may form coiled-coils to cause the experimentally observed aggregation.

Sequence analysis using catGRANULE showed three regions of TraG* were predicted to have high propensity for LLPS; these regions were similarly predicted to have the highest disorder in AF and RF models (**Figure 2.1**). All TraG constructs were analyzed by catGRANULE to better rank how truncations change the predicted LLPS propensity (**Table S1**). The scoring system for catGRANULE is from a cumulative distribution profile of the yeast proteome, from -3 to 3, where sequences scored above 1 are proteins highly likely to form phase separating stress granules, a form of LLPS, when expressed in yeast. Proteins from a variety of species with catGRANULE scores above 1 have been shown to form LLPS in their native cellular environments as well[360]. Contrary to the PScore algorithm, the catGRANULE scores for $TraG_{R100}$ constructs exceeds the

score for the $TraG_F$ constructs, however the difference is miniscule at ~0.040. The scores are below 1 for the full-length TraG, however they exceed 1 for TraG* and TraG*$_{497\text{-}940}$ from both F and R100 plasmids. The C-terminal deletion constructs TraG*$_{497\text{-}863}$ score below 1, lower than the full-length TraG sequences, indicating the predicted stability of this construct. The LLPS propensity plots for $TraG_{F\ \&\ R100}$ indicate that the residues in the flexible linker region have a propensity score from 0-2.5; the removal of the N-terminal TM domain is likely the cause of the increase in score for TraG*, as the TM domain scores low (**Figure S5**). The C-terminal residues from 807-905 score at ~5.0 and their partial removal is the reason for the improved score of TraG*$_{497\text{-}863}$. The highest scoring region for TraG is the interacting region, which reaches a propensity score of 10; the differences in score between F and R100 TraG constructs is caused by the increased score of the interacting loop region of $TraG_{R100}$ relative to $TraG_F$, as well as a minor difference in scores for residues 807-860.

2.3.2 Thermofluor Indicates Enhanced Thermal Stability of TraG$_{497\text{-}940}$*

Thermofluor experiments were performed to compare the thermal and chemical stability of HisG* and HisG*$_{497\text{-}940}$, as it was necessary to demonstrate that removal of the putative N-terminal linker improved stability of the protein. As mentioned, the His tag was left uncleaved for many of the described biophysical experiments as further processing of HisG* often resulted in aggregation, therefore protein constructs were left tagged to allow for a proper comparison. Thermofluor experiments demonstrated that HisG*$_{497\text{-}940}$ is more stable than HisG* based on lower average T_m values in all tested buffer conditions (**Figure 2.4**). The mean T_m of the proteins in all buffers (excluding 50mM citrate pH 3.5) was 50.6 ±0.13°C and 48.0 ±1.01°C for HisG*$_{497\text{-}940}$ and

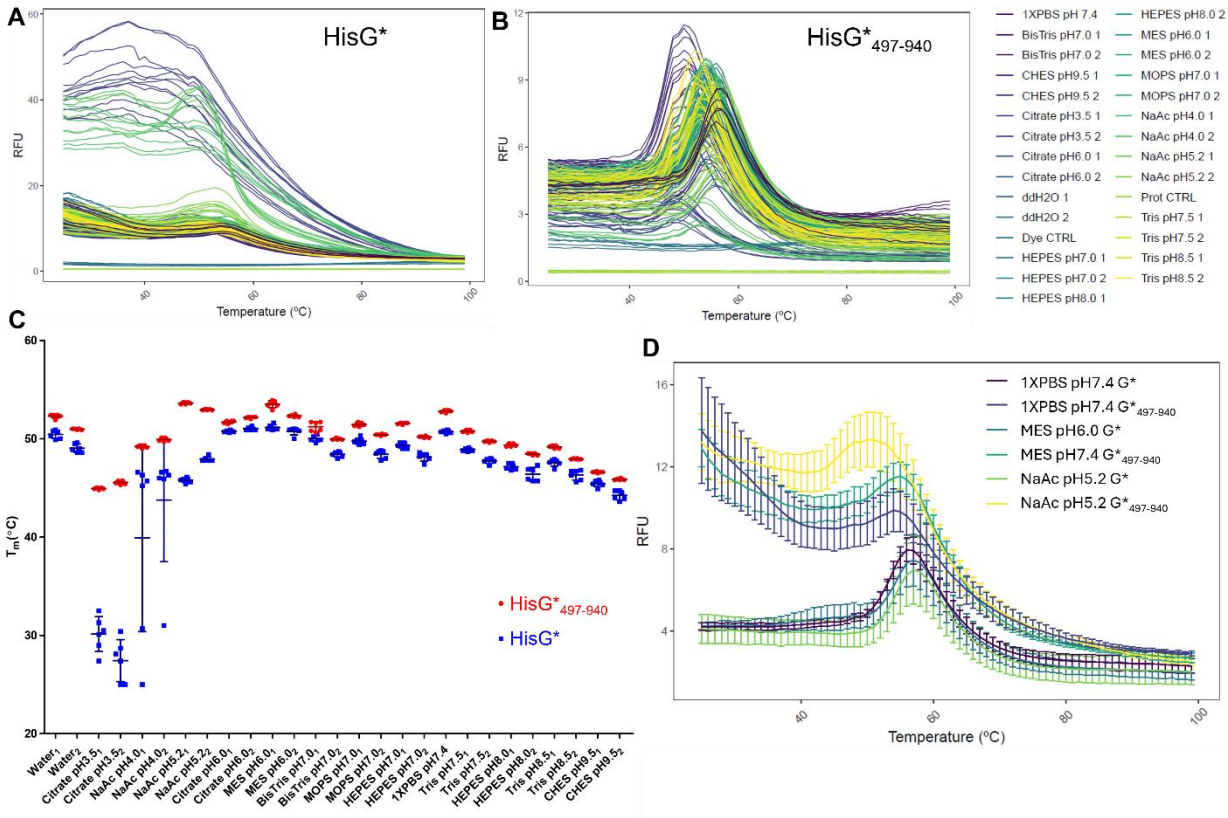

Figure 2.4 Melting Curve Plots from Thermofluor Assays, displaying the temperature dependent unfolding of **A**) HisG* and **B**) HisG*$_{497\text{-}940}$ in various buffer conditions (seen in the legend right of **B**)) as detected using SYPRO orange. Each buffer was prepared at a concentration of 50mM (except ddH_2O and 1 X PBS), where 1 represents the addition of 50mM NaCl in the buffer and 2 includes 200mM NaCl. Six (6) replicates of each experiment are shown. **C)** Mean melting temperatures with ± standard deviation (SD) of the TraG* variants in all tested buffer conditions. **D)** Average melting curve plots of the top three stabilizing buffer conditions for both HisG* and HisG*$_{497\text{-}940}$, as determined by the T_m of these average curves.

Table 2.3 Average Melting Temperatures from Thermofluor Assays. T_m values of HisG* and HisG*$_{497\text{-}940}$ were determined from the melting curves shown in **Figure 2.4A & B**; each buffer was prepared at a concentration of 50mM (except ddH_2O and 1 X PBS), where (1) represents a buffer with 50 mM NaCl and (2) represents a buffer with 200 mM NaCl.

	HisG*$_{497\text{-}940}$		HisG*			HisG*$_{497\text{-}940}$		HisG*	
Condition	T_m (°C)	±SD	T_m (°C)	±SD	Condition	T_m (°C)	±SD	T_m (°C)	±SD
Water (1)	52.30	0.20	50.43	0.43	MOPS pH 7.0 (1)	51.45	0.19	49.75	0.31
Water (2)	50.97	0.05	49.03	0.43	MOPS pH 7.0 (2)	50.4	0.06	48.42	0.42
Citrate pH 3.5 (1)	44.93	0.08	30.13	1.78	HEPES pH 7.0 (1)	51.55	0.08	49.33	0.28
Citrate pH 3.5 (2)	45.55	0.18	27.43	2.13	HEPES pH 7.0 (2)	50.18	0.08	48.15	0.42
NaAc pH 4.0 (1)	49.20	0.13	39.92	9.53	1xPBS pH 7.4	52.78	0.12	50.72	0.15
NaAc pH 4.0 (2)	49.92	0.17	43.75	6.25	Tris pH 7.5 (1)	50.77	0.12	48.88	0.15
NaAc pH 5.2 (1)	53.62	0.10	45.78	0.25	Tris pH 7.5 (2)	49.73	0.08	47.75	0.25
NaAc pH 5.2 (2)	52.95	0.05	47.92	0.24	HEPES pH 8.0 (1)	49.37	0.15	47.15	0.30
Citrate pH 6.0 (1)	51.68	0.13	50.75	0.1	HEPES pH 8.0 (2)	48.43	0.10	46.38	0.70
Citrate pH 6.0 (2)	52.15	0.08	51.02	0.16	Tris pH 8.5 (1)	49.17	0.14	47.57	0.36
MES pH 6.0 (1)	53.50	0.36	51.13	0.27	Tris pH 8.5 (2)	47.93	0.05	46.32	0.51
MES pH 6.0 (2)	52.32	0.12	50.75	0.37	CHES pH 9.5 (1)	46.6	0.06	45.43	0.32
BisTris pH 7.0 (1)	51.22	0.50	50.00	0.31	CHES pH 9.5 (2)	45.87	0.08	44.23	0.46
BisTris pH 7.0 (2)	49.97	0.08	48.42	0.27					

HisG*, respectively (**Table 2.3**). The difference in T_m between HisG* constructs is minimal in many cases, however the shapes of HisG* melting curves are poorer than HisG*$_{497\text{-}940}$, as the high fluorescence in the lower temperatures (25-45 °C) and the shallow peak indicates that there is a proportion of the HisG* molecules in the protein sample which have hydrophobic regions exposed immediately upon SYPRO addition (**Figure 2.4A**)[255]. A reduced quantity of HisG*$_{497\text{-}940}$

molecules are unfolded at the beginning of the assay relative to HisG*, and the peak RFU at the T_m is further pronounced in most of the conditions attempted.

The variability in the HisG* melting curves is higher than those of HisG*$_{497-940}$ as demonstrated by the larger error bars; this is likely due to differences in abundance of aggregation of the protein prior to the start of the experiment. The thermal stability of the truncation mutant is consistent between replicates as the error bars for all buffers tested is smaller, indicating HisG*$_{497-940}$ remains consistent between purified batches whereas the experimental replicates for HisG* have different thermal stability depending on the batch (**Figure 2.4C**)[339]. This provides evidence that the removal of the N-terminal tag increased the thermal and chemical stability of TraG*. Although HisG*$_{497-940}$ has higher thermal stability relative to HisG*, the mean T_m of the lysozyme control is 69.8 ±0.06°C in 1 X PBS pH 7.4, indicating that HisG*$_{497-940}$ is not as stable as a small, well-folded globular protein.

NaAc pH 5.2 and MES pH 6.0 were observed to be optimal buffers for HisG* variants based on this assay (**Figure 2.4D**). The increased T_m of conditions where buffer pH was approximate to the respective predicted isoelectric point (pI) of the proteins (from ProtParam[338], pI TraG*: 5.83 and TraG*$_{497-940}$: 5.95) is an interesting phenomenon. Maintaining protein charge at net neutral is predicted to aid in the formation of intermolecular interactions, which may be the reason for increased thermal stability seen in these conditions. Protein crystallisation is often suggested to be started with the pH of the mother liquor equal or proximal to the pI of the protein[361,362], suggesting optimal buffers of sodium acetate (pH 5.2) and MES (pH 6.0) be explored. As well, buffers with 50mM NaCl concentrations appeared to better maintain folding of both proteins as temperature was increased in comparison to buffers with 200mM NaCl. This could be caused by an increased heat capacity of the solvent when less ionic strength is present, or from

higher ionicity affecting intermolecular interactions by masking charges in TraG*, or that both phenomena contribute to the observed changes. As the thermocycler employed for the thermofluor assay measures the internal temperature of the instrument rather than the temperature of the buffers themselves it is possible the buffers with higher salt content absorbed more kinetic energy at the same temperatures and therefore caused the proteins to unfold at slightly lower temperatures. Additionally, increasing salt concentrations may create 'salting out' effects; as ionicity is increased, salt ions can mask the inter- and intra- molecular ionic interactions that prevent protein unfolding at increased temperatures[363].

2.3.3 Improved Stability of Truncated TraG as Observed by Circular Dichroism*

The CD spectra of all HisG* variants appear to be mainly α-helical based on the characteristic negative Δε peaks at 222 nm and 208 nm (**Figure 2.5**)[266]. Deconvolution in BeStSel indicates differences in the extent to which the proteins fold into canonical α helices[264]. The model providing the best fit for HisG* indicated that 72.3% of the protein was predicted to be α-helical, albeit 20.9% of helices are of a bent or imperfect topology. Turns were of low abundance in the overall structure of the protein at 7.4%, and the remaining 20.3% of the protein was deemed as 'other,' which include 3_{10} and π helices, but also loops that may not change the direction of polarized UV light and are therefore undetectable by CD[264,364,365]. The spectrum of HisG*$_{497\text{-}940}$ appears to have better defined peaks at 222 nm and 208 nm, and the protein is predicted to have 85.1% α-helices, with 67% of optimal topology contrasting with the 51.4% seen in HisG*. HisG*$_{497\text{-}940}$ has only 14.8% 'other' character and no strongly structured turns. The predicted change in α-helical topology of 15.6% is higher than expected for a loss of only 8.6% of the protein's residues. This strongly indicates that the 45 residues removed from the protein are

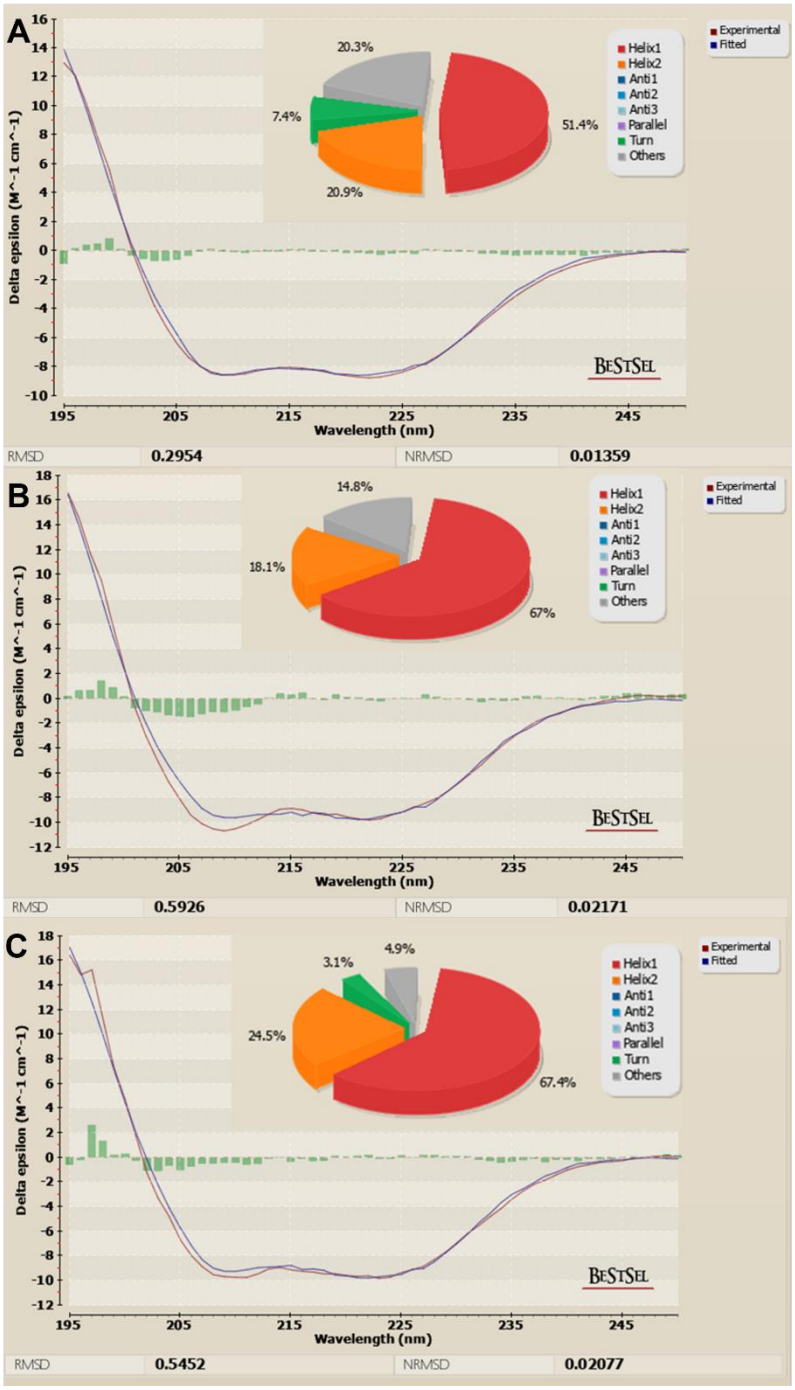

Figure 2.5 The CD spectra of TraG* Constructs. **A**) HisG* **B**) HisG*$_{497\text{-}940}$ and **C**) HisG*$_{497\text{-}863}$ as displayed by BeStSel[264]. Fitting spectra of the TraG* variants (experimental data shown in red) to those of proteins with known structure (shown in blue) allowed for the prediction of the secondary structure composition of each protein, as seen in the pie charts. Distorted helices represent regions in which the α helix geometry is imperfect resulting in helices more closely resembling a 3_{10} helix, while regions referred to as 'Other' include 3_{10} and π helices, but also loops that may not change the direction of polarized UV light and are therefore undetectable by CD[264,364,365].

disordered; the flexible N-terminal region in HisG*$_{497\text{-}940}$ could be causing distortions in the remaining protein fold based on the predicted high relative change in the percentage of assigned secondary structure. This provides a biophysical explanation for the observed aggregation propensity for HisG* relative to HisG*$_{497\text{-}940}$ and explains difficulties in attempts to crystallise HisG*. The high dynamicity of TraG* residues Ala452-Ala496 when not anchored to the N-terminal TM domain results in changes to the overall topology of TraG*.

In comparing the chromatograms of HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$, the lack of difference in these characteristic wavelengths does not indicate an increase in percentage of α-helical content. BeStSel predicts a miniscule increase in α-helices of 0.4% relative to HisG*$_{497\text{-}940}$, which may merely be within the bounds of error for the prediction algorithm. However, there is a notable decrease in the percentage of predicted loop regions for HisG*$_{497\text{-}863}$; a 9.9% decrease in loops is accompanied by a 6.4% increase in the percentage of distorted α-helices and a 3.1% increase in the percentage of detected turns. This indicates that the 77 C-terminal residues may have some disordered content, however the remaining protein has an equal percentage of ordered α-helical content, indicating some of the truncated content was α-helical. This reflects what is seen in AF and RF predictions (**Figure 2.1A & B**) where the C-terminal region is seen as a large, extended α-helix with several unstructured loop regions depending on the prediction method.

Urea denaturation revealed that HisG*$_{497\text{-}940}$ has higher chemical stability than HisG*, as observed by the loss of characteristic α-helical absorbance at lower urea concentrations (**Figure 2.6**). For HisG* at 1 M urea, the absorbance peaks of 222 nm and 208 nm were significantly less pronounced, almost decreasing by 50% of their values from the native spectra. The truncation mutant shows only a minor change in spectra from incubation with 1 M urea in the respective α-helical regions. Quantification via BeStSel predictions indicated a 19.1% α-helical character after

incubating HisG* in 1 M urea while HisG*$_{497\text{-}940}$ had 28.1%, both from an average of 3 replicates (**Table 2.4**).

Together with the thermofluor data demonstrating the lowered thermal stability of HisG*, the urea denaturation and native CD experiments demonstrate an increase in overall stability of the truncation mutant HisG*$_{497\text{-}940}$. It is evident that the presence of residues Ala452-Ala496 destabilizes TraG* as it displayed a higher susceptibility to unfolding compared to HisG*$_{497\text{-}940}$ based on the decrement of Δε at the characteristic α-helical wavelengths as the urea concentration was increased. The secondary structure of HisG*$_{497\text{-}940}$ remained relatively intact until incubation with 3.0 M urea, while the secondary structure of HisG* was partially unfolded at 1.0 M urea based on the topology analysis from 260-200 nm by BeStSel[264].

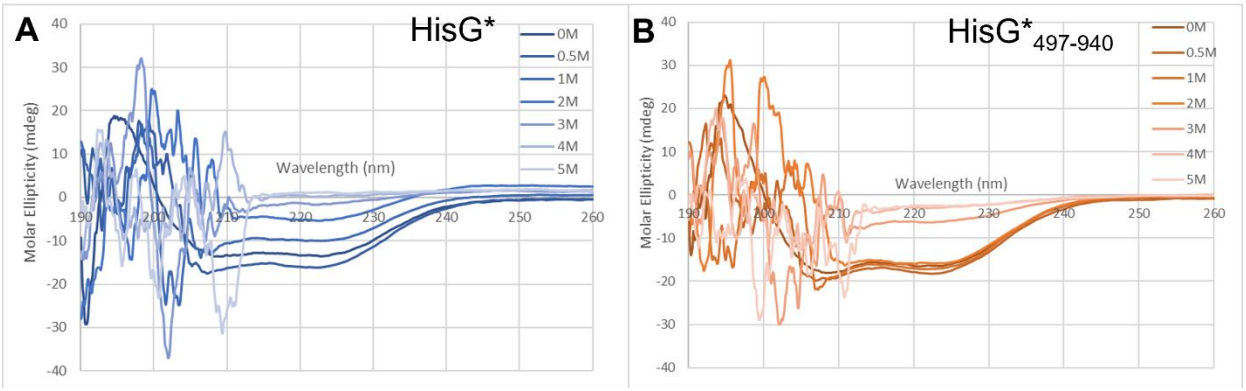

Figure 2.6 Urea Denaturation of TraG* Constructs Measured by CD Spectroscopy, where **A**) HisG* and **B**) HisG*$_{497\text{-}940}$. All experiments were performed with a final protein concentration of 2μM and were incubated in the respective concentrations of urea (shown in the figure legends) for 1 hr at 25 °C prior to measurement. The resultant data was normalized by subtracting using the blank spectra, wherein the buffer conditions in the same concentration of urea were replicated and CD data was collected. Molar ellipticity measurements were obtained every 0.1 nm from 260-190 nm. Data from 195-260 nm was entered into BeStSel for deconvolution[264], and the resultant analysis from triplicate runs is shown in **Table 2.4**; spectra shown is from single representative experiments.

Table 2.4 Secondary Structure of TraG* Constructs Under Urea Denaturation. Data presented from triplicate measurements and deconvoluted with BeStSel[264]. Representative CD spectra shown in **Figure 2.6**.

HisG*								
	α-Helix		Antiparallel β-Sheet					
[Urea] (M)	Regular	Distorted	Left-twisted	Relaxed	Right-twisted	Parallel β-sheet	Turn	Other
0	40.70	17.10	0.00	2.97	0.00	0.00	8.47	30.77
0.5	18.37	16.80	0.00	0.00	14.73	1.03	9.10	40.03
1	19.13	19.07	6.97	3.60	12.33	0.00	12.37	26.57
2	8.50	9.00	5.07	2.60	9.90	6.70	11.27	47.00
3	5.07	11.17	2.33	0.00	12.43	0.00	15.63	53.37
4	0.00	0.00	3.50	21.13	22.80	0.00	19.90	32.70
5	8.10	10.80	3.40	5.93	8.80	0.00	15.20	47.77
HisG*$_{497\text{-}940}$								
	α-Helix		Antiparallel β-Sheet					
[Urea] (M)	Regular	Distorted	Left-twisted	Relaxed	Right-twisted	Parallel β-sheet	Turn	Other
0	64.20	17.13	0.00	0.03	0.03	1.47	6.77	10.30
0.5	34.63	16.57	0.00	0.00	26.83	0.87	7.07	14.10
1	28.10	19.40	9.03	0.00	7.23	0.77	8.77	26.70
2	24.43	10.80	10.20	0.00	8.57	10.03	5.87	30.17
3	7.50	8.27	0.00	6.77	17.77	0.00	18.40	41.33
4	9.37	0.00	0.00	15.50	12.27	3.57	20.67	38.63
5	1.37	9.00	0.90	5.30	5.90	0.00	18.83	58.63

2.3.4 CIU-MS Demonstrates Differences in the Conformational Stability of HisG and HisG*$_{497\text{-}940}$*

The native mass spectra of HisG* and HisG*$_{497\text{-}940}$ were highly reproducible in biological replicates, where HisG* had highly abundant charge states of 18+ (3144 m/z), 17+ (3329 m/z), and 16+ (3537 m/z) while HisG*$_{497\text{-}940}$ had more abundant 17+ (3054 m/z), 16+ (3245 m/z), and 15+ (3462 m/z) charge states (**Figure 2.7**). The propensity for HisG* to occupy higher energy charge states more frequently indicates that the presence of the putative linker region results in

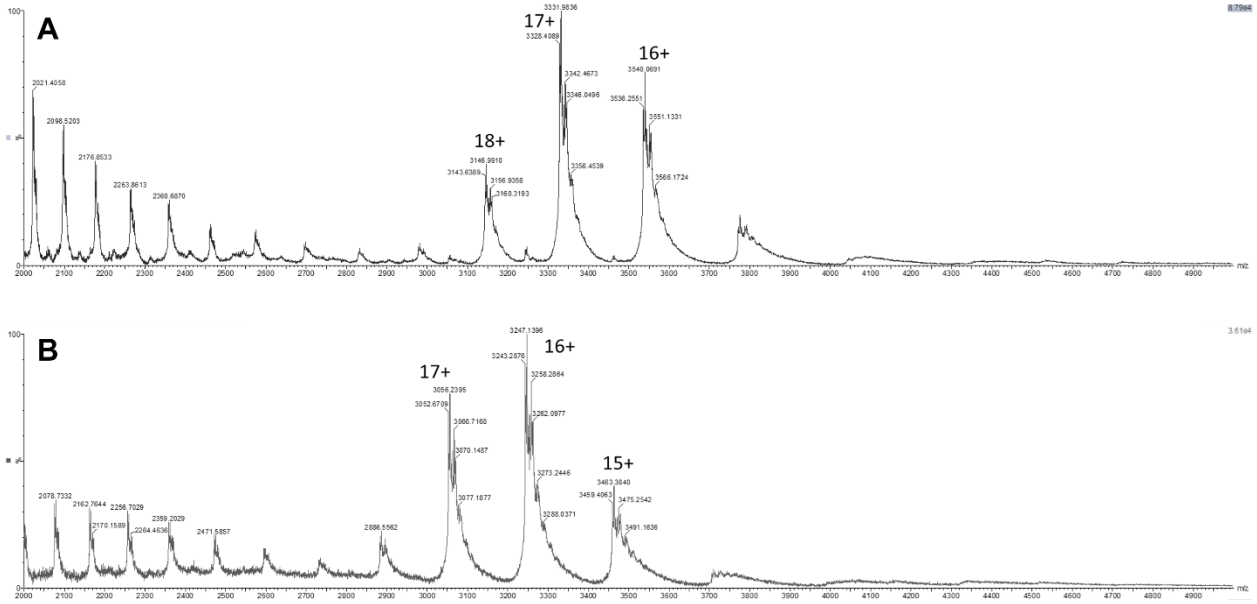

Figure 2.7 ESI-MS Analysis of Native Folds of TraG* Constructs, where **A**) HisG* and **B**) HisG*$_{497\text{-}940}$. ESI was performed on proteins in 100mM ammonium acetate pH 6.8 at concentrations of 5 μM at 5 μL/min. Monomers for both proteins are seen in the 3000-3900 m/z range with prominent charge states of +18 to +16 for HisG* and +17 to +15 for HisG*$_{497\text{-}940}$.

lower stability[366,367]. The mass of each protein as determined using ESIprot was 56,588 ±8.71 Da for HisG* and 51,906 ±7.54 Da for HisG*$_{497\text{-}940}$[368]. The changes relative to expectations based on sequence (HisG*: 56699.30 Da, HisG*$_{497\text{-}940}$: 52010.34 Da via ProtParam[338]) might be the result of salt adducts carried over from the purification. The spectra show that both HisG* and HisG*$_{497\text{-}940}$ samples have unfolded conformations as seen in the 2,000-2,700 m/z range, demonstrating the dynamic nature of the protein regardless of the N-terminal truncation. However, the peaks seen in the HisG* spectrum are more abundant indicating a higher propensity for unfolding relative to HisG*$_{497\text{-}940}$.

CIU data was analysed using charge states 17+ and 16+ for both proteins, as the 15+ and 18+ charge states were of insufficient abundance in some CEs to be compared. To visualize changes in drift time of these species as a function of CE, CIUSuite2[369] was used to generate CIU heat maps (**Figure 2.8**). The replicates of each plot have some differences in the intensity of the

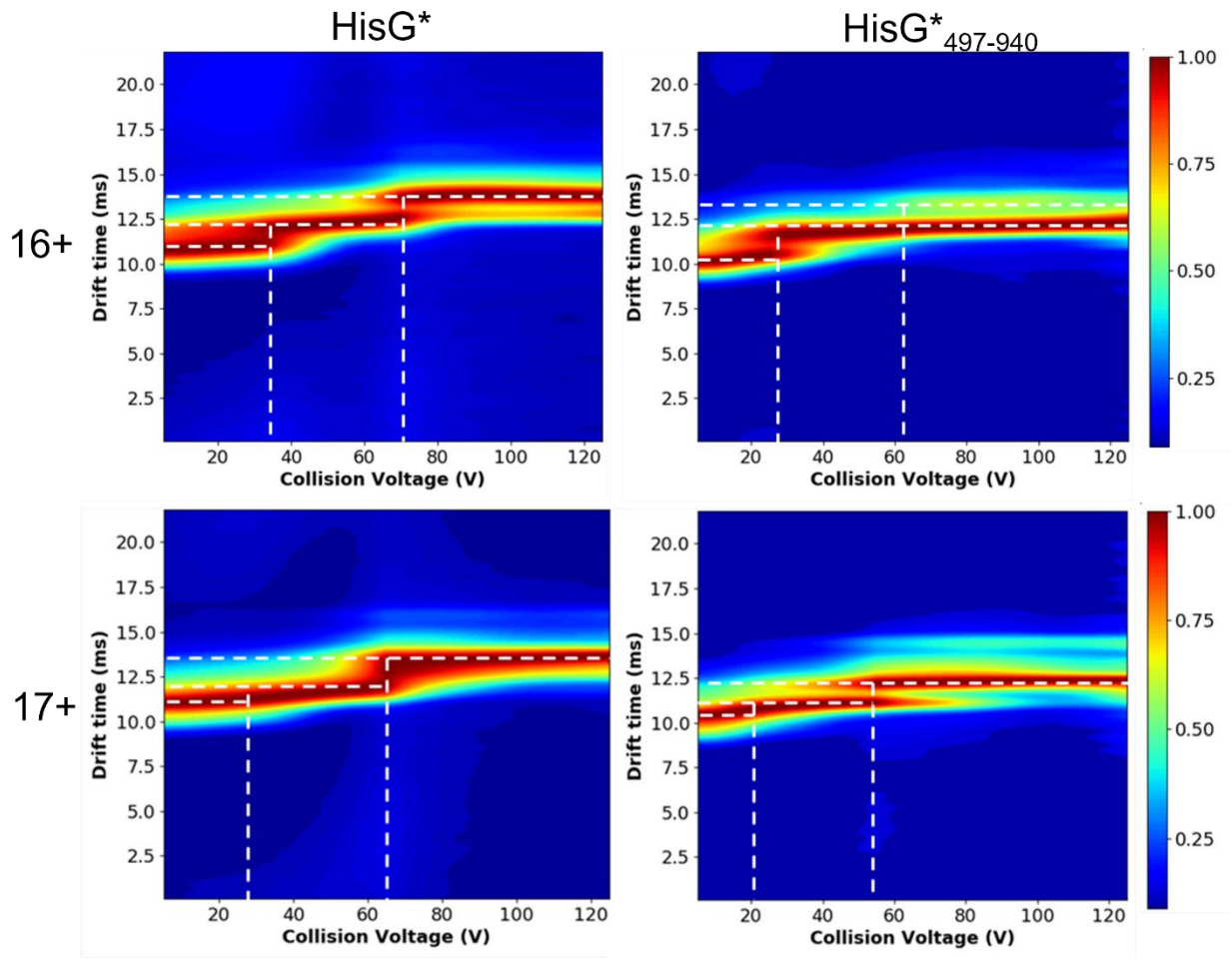

Figure 2.8 CIU-MS Heat Maps of 16+ and 17+ Charge States from TraG* Constructs, HisG* and HisG*$_{497\text{-}940}$ as trap CE is increased from 5 V to 125 V. These maps were plotted with CIUSuite2[369] and smoothed with default settings from the software. The m/z range set for plotting the IMS of the 17+ charge states were 3320-3355 for TraG* and 3040-3090 for HisG*$_{497\text{-}940}$, and the m/z range used for the 16+ charge states were 3525-3575 for HisG* and 3230-3280 for HisG*$_{497\text{-}940}$. A biological replicate for each experiment is shown in **Figure S6**.

transitions, however, there are common shifts in drift time for the same species (**Figure S6**). The 16+ charge state of HisG* displays 2 shifts, one at ~35 V CE and one at ~70 V CE. The 16+ charge state of HisG*$_{497\text{-}940}$ has a single shift at ~25 V CE, however, there appears to be a low abundance of an unfolded state from ~60 V CE onward. The third HisG* species with the longest drift time (~14 ms) is observed at nearly 100% abundance for both the 16+ and 17+ charge states above 70 V CE, whereas for HisG*$_{497\text{-}940}$ the third species (~13.5 ms) is only partially occupied above 70 V

CE and not well defined for the 17+ charge state. This indicates a third, highly unfolded conformer is more often induced in HisG* than HisG*$_{497\text{-}940}$, supporting the interpretation that the flexibility of the N-terminal region affects the conformational dynamics of HisG*. A similar trend is seen in the 17+ charge state as well. This can also be visualised by observing the trends in the IMS spectra, which aids in viewing the reproducibility and comparability between replicates (**Figure S7**). HisG*$_{497\text{-}940}$ and HisG* have 3 conformers each: approximately 10.5 ms, 12.5 ms, 13.5 ms and 11 ms, 13 ms, 14 ms, respectively, occupied in differing percentages depending on the CE applied. The differences in drift time between conformers of HisG* and HisG*$_{497\text{-}940}$ are appropriate based on a 45-residue truncation. The shift to the 14 ms conformer is well defined for HisG* in the 16+ charge state at 100 V CE, however the IMS spectrum of the 17+ charge state has a broad peak at 100 V CE that is centered at a smaller drift time (~13.75 ms) when comparing to the drift time for the most unfolded conformer of HisG*$_{497\text{-}940}$ (~14.5 ms). Although the most unfolded conformer from HisG*$_{497\text{-}940}$ has a longer drift time relative to that of HisG* at the same CE, it is of low abundance at ~40% signal, whereas the second HisG*$_{497\text{-}940}$ conformer (~12.5 ms) is at 100% abundance, and the most folded conformer is present at ~30% signal (~10.5 ms). Therefore, the interpretation that HisG* can access a more unfolded conformational state at a higher propensity than HisG*$_{497\text{-}940}$ is maintained. However, as each charge state was not isolated during data collection prior to the CIU analysis, charge stripping events may have caused some additional noise in the data284,285. Therefore, no quantitative claims can be made from this CIU-MS data regarding the degree of disorder in each protein construct.

2.3.5 SEC-MALS-SAXS for the Biophysical Characterisation of TraG Constructs*

2.3.5.1 SEC-SAXS Reveals the Aggregation Propensity of Larger TraG Constructs*

Previous SEC-MALS-SAXS data of HisG* showed mailing samples to the APS resulted in significant aggregation such that the signal from the monomeric protein could not be appropriately resolved[7]. Freshly purified protein was transported directly to the APS to limit aggregation and generate useful analyses of the HisG* monomer, however SEC-MALS-SAXS data of this sample indicates that some aggregation occurred despite preventative measures (**Figure 2.9A**). A highly intense peak from frames 1007-1410 results from high MW aggregates of the protein as it was eluted in the void volume of the SEC column. The low intensity of the monomeric protein peak at frames 1580-1767 is caused by the loss of monomeric HisG* to aggregate formation and the high sensitivity of the technique to aggregates, resulting in obscuration of monomer scattering. The MALS data indicates that approximately double the amount of monomeric protein was present compared to aggregate based on the UV_{280nm} chromatogram, which shows a large light scattering peak for the aggregate at 14.0-17.5 mL and the lower light scattering peak corresponding to the monomer at 23.5-26.0 mL (**Figure 2.10A**). This indicates that the measures taken to prevent aggregation greatly reduced their presence in the sample and therefore increased the SAXS data quality of the HisTraG* monomer. $HisG^*_{497\text{-}940}$ showed the presence of monomeric protein at frames 1355-1475, and $HisG^*_{497\text{-}863}$ showed monomeric protein at frames 1680-1895 (**Figure 2.9A**), with no observable aggregation seen for either sample in SAXS or MALS chromatograms (**Figure 2.10B & C**). This demonstrates the increased propensity of HisG* for aggregation as provided by the flexible linker at the N-terminus of the construct. The $HisG^*_{F\ 497\text{-}863}$ MALS data indicates there was more aggregate than monomeric protein present in this sample (**Figure 2.10D**). The sample was purified earlier than

the HisG*$_{R100}$ samples and was frozen, therefore the observed aggregation may be due to oxidation, the freeze-thaw process, or the interacting region which differs between TraG*$_{F\ \&\ R100}$ has a higher propensity for aggregation in HisG*$_{F\ 497\text{-}863}$.

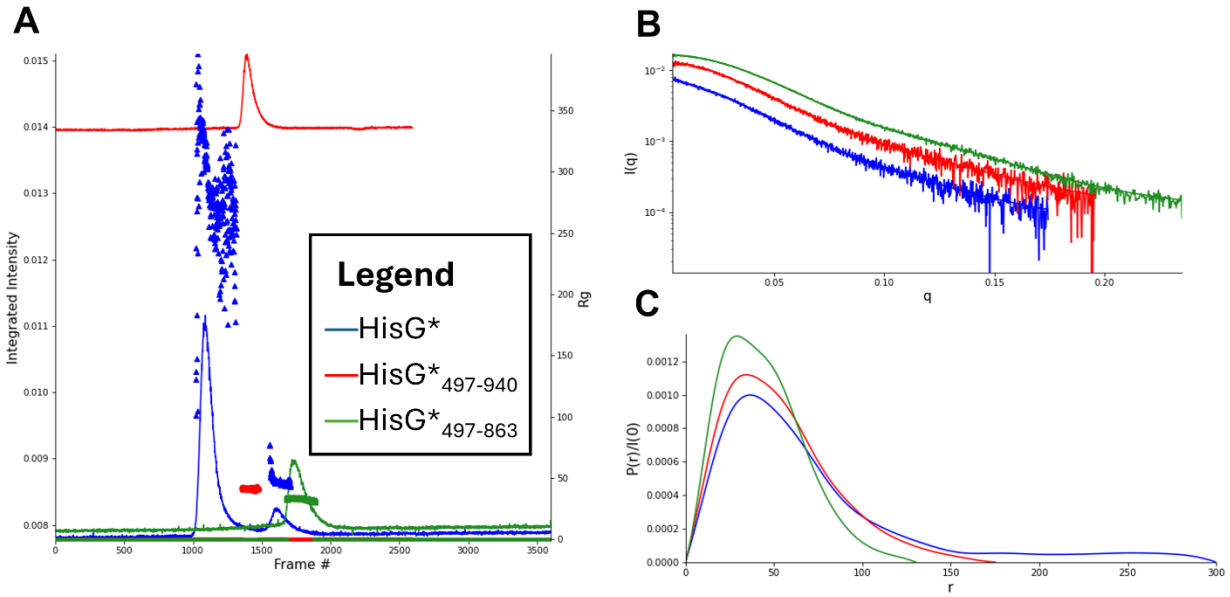

Figure 2.9 SEC-MALS-SAXS Data from TraG* Constructs. **A**) SEC-SAXS chromatograms, **B**) scattering profiles and **C**) P(r) function plots resulting from SEC-MALS-SAXS data collection and analysis of TraG* variants. HisG*$_{497\text{-}940}$ data was collected in a previous experiment, hence the baseline intensity of the SEC-SAXS chromatogram is higher. P(r) function plots were produced from taking the inverse Fourier transform of scattering profiles using the integrated GNOM tool in RAW[343,344]. The scattering profiles display the lines for the fit of the scattering data to the output P(r) functions, with HisG* and HisG*$_{497\text{-}940}$ data truncated for DAMMIF modelling. Biophysical parameters obtained from these figures can be seen in **Table 2.5**. All SAXS data images were produced using RAW[343].

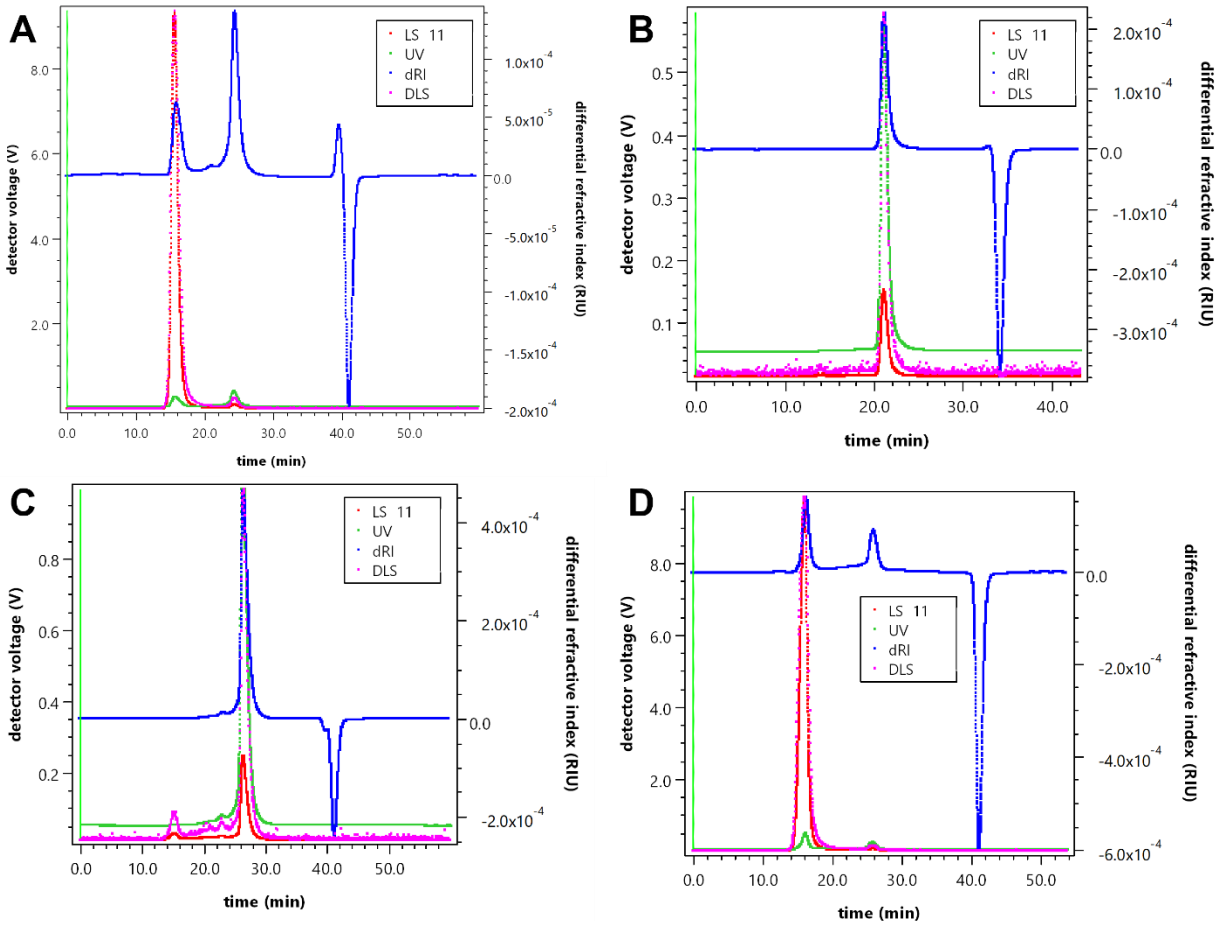

Figure 2.10 SEC MALS Chromatograms of TraG* Constructs, where **A**) HisG*, **B**) HisG*$_{497\text{-}940}$, **C**) HisG*$_{497\text{-}863}$, and **D**) HisG*$_{F\ 497\text{-}863}$, collected using a Superdex 200 10/300 Increase column (Cytiva) and a DAWN Heleos LS and T-rEX dRI instrument (Wyatt), presented using the ASTRA 7 software package. The monomeric protein peak of each sample is seen at 24-25 min, 20.5-21.5 min, 26-28 min and 25-26 min of elution, respectively.

SEC-SAXS data was processed (**Section A.4**) to produce scattering profiles (**Figure 2.9B**). Guinier analysis of the scattering profiles showed that the HisG* monomer data is affected by the residual aggregate, as the scattering profile displayed an upward curvature at low q data resulting in a characteristic downwards arch-shaped residual plot (**Figure 2.11A**). HisG*$_{497\text{-}940}$ had minimal aggregate scattering contribution as can be seen by the slight curve in the residual plot, whereas the HisG*$_{497\text{-}863}$ plot is flat. This allowed for all q data to be used in further processing of the HisG*$_{497\text{-}863}$ data, while HisG* and HisG*$_{497\text{-}940}$ were truncated at both low and high q

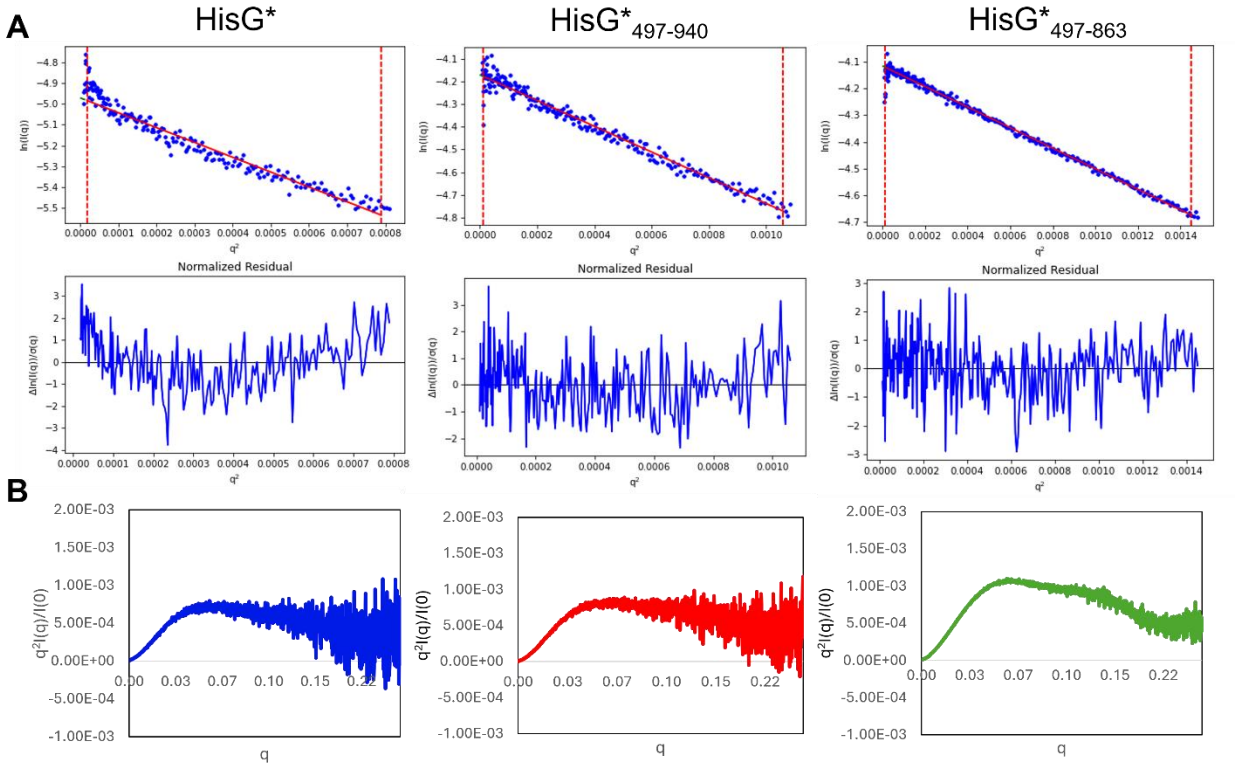

Figure 2.11 Guinier and Kratky Plots from TraG* SAXS Scattering Profiles. **A**) Guinier plots of scattering profiles from SAXS data of TraG* constructs with the normalized residual plots displayed immediately below. **B**) Normalized Kratky plots resulting from the scattering profile of HisG* (blue), HisG*$_{497\text{-}940}$ (red) and HisG*$_{497\text{-}863}$ (green). Images were produced using RAW[343].

(**Figure 2.9B**). A normalized Kratky analysis was performed on all scattering profiles, and the plot for HisG* has characteristics of a gaussian peak at low q, and at high q it appears to plateau, indicating an elongated or unfolded protein; this qualitative analysis is hindered by noise at high q, likely caused by the aggregate-obscured scattering of the monomeric protein (**Figure 2.11B**). The HisG*$_{497\text{-}940}$ Kratky plot has a similar bell shape to HisG*, returning to the x-axis at high q values (not pictured) indicating some disorder is present, however less noise is present at high q relative to the HisG* Kratky plot. The Kratky plot is further improved with HisG*$_{497\text{-}863}$, which shows minimal noise at high q and a refined Gaussian shape; the additional hill in the Kratky plot indicates the prolate shape of the protein[370,371]. Similar observations can be made from the P(r)

plot of HisG*$_{497\text{-}863}$, which has a left-leaning lobe[314]; these plots can also be used to illustrate the differences between the three constructs (**Figure 2.9C**). All proteins are prolate globules and the constructs with less truncations have unproportionally higher r values, indicating that these regions are unfolded. The significantly increased r value for HisG* is likely due to distortions from the presence of aggregate. Data trends for the Guinier analysis, Kratky plot, and P(r) function for HisG*$_{F\ 497\text{-}863}$ are similar to that of HisG*; in both, values are likely inflated due to the aggregate present in the sample (**Figure S8**). See **Section A.5.2** for more details regarding the validation of SAXS data.

2.3.5.2 Truncations of TraG Improve the Quality of SAXS and MALS Data*

SAXS and MALS data analysis allowed for the comparison of biophysical structural values between TraG* constructs (**Table 2.5**). The intensity at the zero scattering angle (*I(0)*) represents the abundance of the scattering molecule, which is ideal when it can be calculated accurately from both Guinier analysis and the P(r) function. The *I(0)* values for HisG* and HisG*$_{F\ 497\text{-}863}$ are low and are not accurate to each other due to the low abundance of monomer particles scattered in each dataset. This results in the difference between R_g calculations for HisG*, where the R_g from the Guinier analysis (46.2 Å) is akin to the expected size for the protein while the R_g from the P(r) function is overestimated (64.4 Å) resulting in an overestimated Dmax (300 Å). As described in **Section A.4**, Dmax was increased to this value to raise the χ^2 above 0.7, as overfitting of the IFT result occurred if the Dmax was lowered. The other constructs, HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ have accurate *I(0)*, R_g, and Dmax values with expected trends and appropriate fitting statistics; as the protein is truncated, the R_g and Dmax decrease and the *I(0)* increases because the protein has higher stability, resulting in less aggregation and higher scattering intensity from the monomeric protein.

Table 2.5 Structural Parameters and DAMMIF Modelling Results Acquired from TraG* SEC-MALS-SAXS Data[343,347]. Of the four methods in which MW is calculated from SAXS data, volume of correlation (V_c) was reported as it is considered to be the most accurate[372]. Shape factor ρ calculations were made using R_g values from Guinier and P(r) analyses and are thus reported as a range (Guinier-P(r)).

Structural Parameters				
	HisG*	**HisG*$_{497-940}$**	**HisG*$_{497-863}$**	**HisG*$_{F\ 497-863}$**
I(0) (cm^{-1}) [from Guinier]	6.93 ±0.031e^{-3}	12.6 ±0.054e^{-3}	18.9 ±0.026e^{-3}	5.38 ±0.023e^{-3}
R_g (Å) [from Guinier]	46.2 ±0.47	40.90 ±0.22	33.86 ±0.08	42.30 ±0.30
I(0) (cm^{-1}) [from P(r)]	7.42 ±0.065e^{-3}	12.6 ±0.040e^{-3}	18.9 ±0.026e^{-3}	5.32 ±0.024 e^{-3}
R_g (Å) [from P(r)]	64.36 ±1.53	42.27 ±0.81	34.95 ±0.07	42.67 ±0.28
χ^2 [from GNOM]	0.7696	0.8478	0.8779	0.9440
α [from GNOM]	2.137	10.94	43.85	19.06
Dmax (Å)	300	175	130	152
MW (kDa) [V_c from SAXS]	68.2	51.7	46.5	45.6
MW (kDa) [from MALS]	63.02 ±1.83	50.61 ±0.91	42.50 ±1.02	56.37 ±2.03
R_h (Å) (from MALS)	53.3 ±0.7	47.4 ±0.6	42.9 ±0.8	44.6 ±0.7
ρ (R_g/R_h)	0.87-1.21	0.86-0.89	0.79-0.83	0.95-0.96
DAMMIF Results				
AMBIMETER Score	0.00	2.029	2.720	1.398
Shape Categories	0	107	525	25
χ^2 [*ab initio*]	0.9519	0.8487	0.9230	0.9440
χ^2 [to RF]	0.9517	0.8465	0.9359	0.9631
χ^2 [to AF]	1.0290	0.8460	0.9586	0.9581

There is some overestimation of the R_g from the P(r) function for HisG*$_{497-940}$ which indicates some unfolding and/or aggregation may be occurring, thus inflating the Dmax (**Table 2.4**). The R_g values from the Guinier analysis and the P(r) function for HisG*$_{F\ 497-863}$ are accurate to each other but do not follow the observed trend as they are larger than the R_g of HisG*$_{497-940}$. The Dmax of HisG*$_{F\ 497-863}$ is smaller than that of HisG*$_{497-940}$, however it is not accurate as it should be similar to that of HisG*$_{497-863}$. The influence of aggregate scattering within the monomer peak is seen throughout the SAXS data for HisG* and HisG*$_{F\ 497-863}$.

In comparing the MW estimates from SAXS to the estimates acquired from MALS data, the values are proximal for constructs with no aggregate scattering. The HisG* SAXS and MALS MW calculations are overestimated as the expected weight from ProtParam is 56,699 Da, while the MW determination from native MS is the most accurate result with a value of 56,588 ±8.71 Da. The HisG*$_{497\text{-}940}$ MW from MALS (50.61 ±0.91 kDa) is lower than the expected MW from ProtParam (52,010 Da), where the MW from SAXS is proximal to the expected value at 51.7 kDa; again, the MW determined from MS is more accurate (51,906 ± 7.54 Da). This demonstrates the accuracy of MS in determining the MW of a protein despite the presence of aggregates in the sample. The MW estimate of HisG*$_{497\text{-}863}$ from MALS was within the margins of error for consensus with the expected weight from ProtParam (43,050 Da), indicating that this construct does not aggregate sufficiently to disrupt the proper calculation of MW from the MALS data. This is not true for HisG*$_{F\ 497\text{-}940}$, as the MW calculation from MALS is significantly overestimated whereas the MW from SAXS is closer to the expected MW from ProtParam (42,933 Da).

The R_h determination from MALS allows for a shape factor calculation based on the R_g/R_h ratio[316]. The results demonstrate that HisG* is an extended shape, and the experimental constructs become less prolate and more globular as regions are truncated; shape factors of ~0.75 represent compact, globular proteins and represent more prolate species as the shape factor increases[317]. To visualize the shapes of HisG* constructs, low resolution structural models were generated using DAMMIF[347]. AMBIMETER was used to quantify the ambiguity of the output DAMMIF bead models; the HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ models have reasonable outputs as scores below 3.0 are considered sufficiently unique, while the HisG* model is equivocal as it had an ambiguity score of 0 with 0 compatible shape categories[373]. Interestingly, the HisG*$_{F\ 497\text{-}863}$ data provided the best AMBIMETER scores relative to the other constructs, however the χ^2 of the output DAMMIF

model is worse than those for HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ indicating the HisG*$_{F\ 497\text{-}863}$ models may not be accurate. The χ^2 of the output *ab initio* model for HisG* is in an appropriate range despite an ambiguous model based on analysis by AMBIMETER; this is markedly improved from HisG* SAXS data previously collected, where the χ^2 of the DAMMIF model was 1.540 (SASBDB ID: SASDQH6). This demonstrates the improvement of the HisG* sample quality from previous mail-in experiments, suggesting kinetic energy due to harsh movements may cause HisG* to aggregate excessively.

The final refined HisG* DAMMIF *ab initio* bead model is an extended rod-like shape with two lobes on either terminus of the shape that bend in opposite directions, and a lobe extending directionally from the center of the model (**Figure 2.12**). The HisG*$_{497\text{-}940}$ bead model is a rod-like shape with one side showing a longer extension than the other, and the lobe in the middle of the HisG* model is present but has a defined hole within it, made visible due to the improved data quality (SASBDB ID: SASDQG6). The comparison of these models along with the significantly reduced R_g and Dmax of the protein indicates a large extension was removed through the N-terminal truncation of 45 residues. The reduction in size is significant compared to the number of residues removed, demonstrating the disordered nature of the proposed flexible linker region of TraG. In analyzing the DAMMIF model of HisG*$_{497\text{-}863}$ it is evident that the truncations performed have resulted in a more globular molecule, as the bead model is a small rod with an extending lobe in its center containing a hole. When comparing this model to that of HisG*$_{497\text{-}940}$ the C-terminal α-helix can be identified as the other extension seen in HisG*, and therefore this region is likely flexible as predicted. The lobe present in the center of all models may be wholly or partially from the interacting region; a region predicted to be disordered and to have a high propensity for LLPS by many predictive algorithms. DAMMIF bead models were aligned to AF and RF models of the

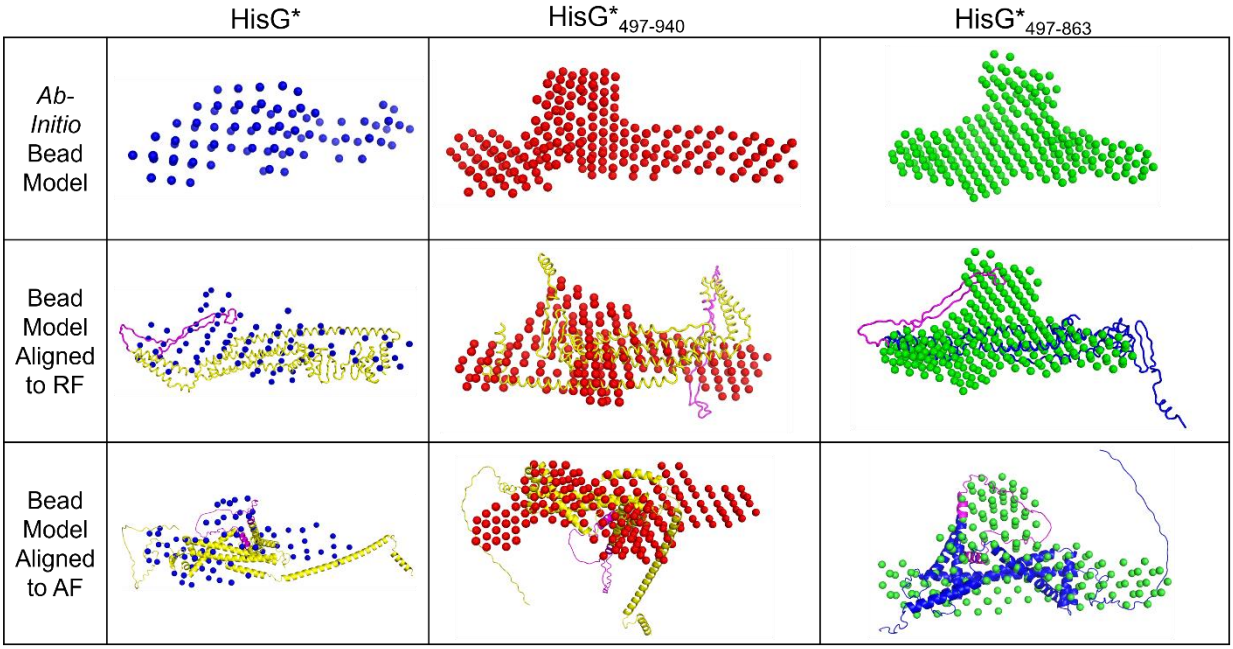

Figure 2.12 DAMMIF Bead Models of TraG* Constructs, generated *ab initio*, aligned to a RF model or an AF model[347]. Structural predictive models are coloured yellow (blue for HisG*$_{497-863}$) and show the interacting loop region in magenta.

His-tagged constructs to better visualize these hypotheses. In most cases, the χ^2 of the produced bead models aligned to the RF models are not significantly improved from the *ab initio* model, however, the bead models aligned to the AF models provide a less accurate fit (**Table 2.5**). The fit appears worst when the C-terminal residues in AF models are placed distant from the remaining rod-shaped globule, whereas the RF models fold this region such that it is transverse along the molecule. This indicates the structure may have the C-terminal region tucked into the globule rather than what has been seen in AF models. However, structural predictions for TraG* constructs made by AF and RF are generally medium-low confidence so these reconstructions may be inaccurate. As well, the HisG* and HisG*$_{F\ 497-863}$ models may not be comparable to the other constructs as the additional lobes seen may result from disruption by the scattering of the residual aggregate protein; this is evident in comparing the bead model from HisG*$_{F\ 497-863}$ to HisG*$_{497-863}$,

where the construct from the F plasmid has a significant extension not expected to be present based on the models from the accurately predicted constructs (**Figure S8F**). These HisG*$_{F\ 497\text{-}863}$ bead models show an improved shape when aligned to AF and RF models as they are similarly proportioned to the predictive models (**Figure S8G & H**).

2.3.5. EOM Provides Models for the Conformational Flexibility of TraG Constructs*

As TraG* was determined to have flexible segments, EOM was attempted to model the conformational space of the protein and the designed truncation mutants. The sequence character assignment which provided the optimal result for HisG* (χ^2=0.791) had residues 1-25 and 473-523 designated as disordered, with 26-472 designated as compact. The fit of the resultant models demonstrates that there is consensus with the assignment, as the conformational state of the protein is typically highly compact with models 1 and 2 representing half of the conformational states of the protein (**Figure 2.13A**). These models demonstrate that HisG* can behave as a globular protein with a small R_g compared to the R_g from the Guinier plot, however the other half of the conformational states show disordered extensions of which some make the molecule more prolate with R_g and Dmax values akin to those previously reported (models 3 and 4), with a quarter of the conformational states modeled as fully disordered strands with extremely high Dmax values (models 5 and 6) (**Table 2.6**). The averaged ensemble shows a smaller R_g and Dmax compared to the values from the P(r) function, indicating the conformational space typically sampled by HisG* is a moderately disordered conformation; somewhere between the states of model 4 and model 5. As positions of the residues in the models are predictions as a bead-on-a-string model can be oriented a number of ways and result in the same shape, it cannot be discerned whether the N- or C-terminus is more dynamic and further contributes to the prolate nature of the protein when

observing models 3 and 4, however as the assignment for the protein was optimal when only 25 residues of the N-terminal region and 50 C-terminal residues were assigned as dynamic, the C-terminal helix may be the significant contributor.

The distribution map displaying the R_g of conformational models indicates that the influence of the aggregate protein in the HisG* data may be distorting the models and providing the overly disordered conformations (**Figure 2.14A**). The R_{flex} of the ensembles is high at 78%, however they are only somewhat smaller than the R_{flex} of the pool at 84%, suggesting that the system is slightly less flexible than the pool. Mathematically, R_σ (SD of the ensemble/SD of the pool) should approach but not exceed 1.0 for fully flexible systems, however the R_σ is 1.98,

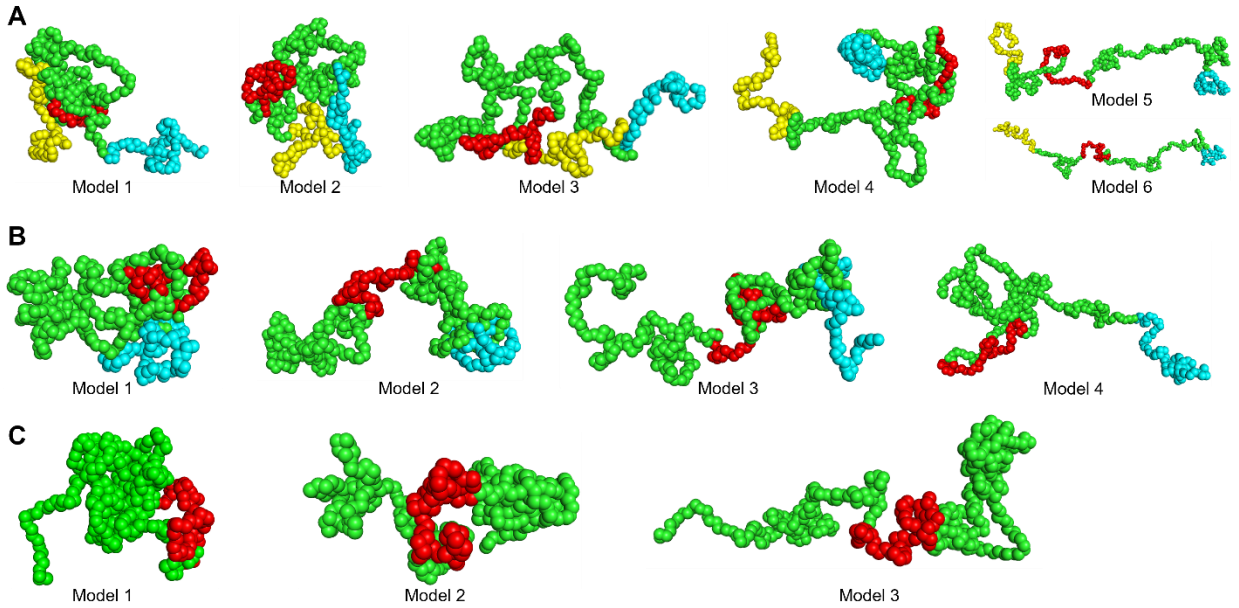

Figure 2.13 EOM Ensembles of TraG* Construct Models. The selected ensemble of models which best describe the conformations achieved by TraG* constructs **A**) HisG*, **B**) HisG*$_{497\text{-}940}$ and **C**) HisG*$_{497\text{-}863}$, as generated by EOM from SEC-MALS-SAXS data. Models are displayed with residues as spheres, coloured with the N-terminal linker in yellow, the interacting loop region in red, and the C-terminal helix in cyan with the remainder of the protein in green. Images were generated using PyMOL v2.5.0, with sphere radius set to 2.0. Statistics of the models are shown in **Table 2.6**.

Table 2.6 Structural Parameters of EOM Models Generated from SEC-MALS-SAXS Data. The fit to the data for the ensembles was measured by χ^2 values of 0.791, 0.918, 0.858, and 0.802 for HisG*, HisG*$_{497\text{-}940}$, HisG*$_{497\text{-}863}$, and HisG*$_{F\ 497\text{-}863}$, respectively.

	R_g (Å)	**Dmax (Å)**	**Fraction**
		HisG*	
Model 1	36.9	113.8	0.12 (1/8)
Model 2	36.8	140.8	0.38 (3/8)
Model 3	47.7	186.5	0.12 (1/8)
Model 4	51.6	159.7	0.12 (1/8)
Model 5	102.0	305.0	0.12 (1/8)
Model 6	119.4	403.0	0.12 (1/8)
Average Ensemble	58.50	198.80	
		HisG*$_{497\text{-}940}$	
Model 1	33.4	104.5	0.44 (4/9)
Model 2	46.6	149.8	0.33 (3/9)
Model 3	49.3	163.9	0.11 (1/9)
Model 4	65.3	205.1	0.11 (1/9)
Average Ensemble	42.94	137.38	
		HisG*$_{497\text{-}863}$	
Model 1	28.6	94.7	0.42 (5/12)
Model 2	37.9	114.3	0.50 (6/12)
Model 3	57.7	205.2	0.08 (1/12)
Average Ensemble	34.90	112.08	
		HisG*$_{F\ 497\text{-}863}$	
Model 1	33.2	106.7	0.22 (2/9)
Model 2	34.0	126.7	0.33 (3/9)
Model 3	36.5	126.9	0.11 (1/9)
Model 4	95.0	312.7	0.22 (2/9)
Model 5	90.3	317.2	0.11 (1/9)
Average Ensemble	53.91	184.78	

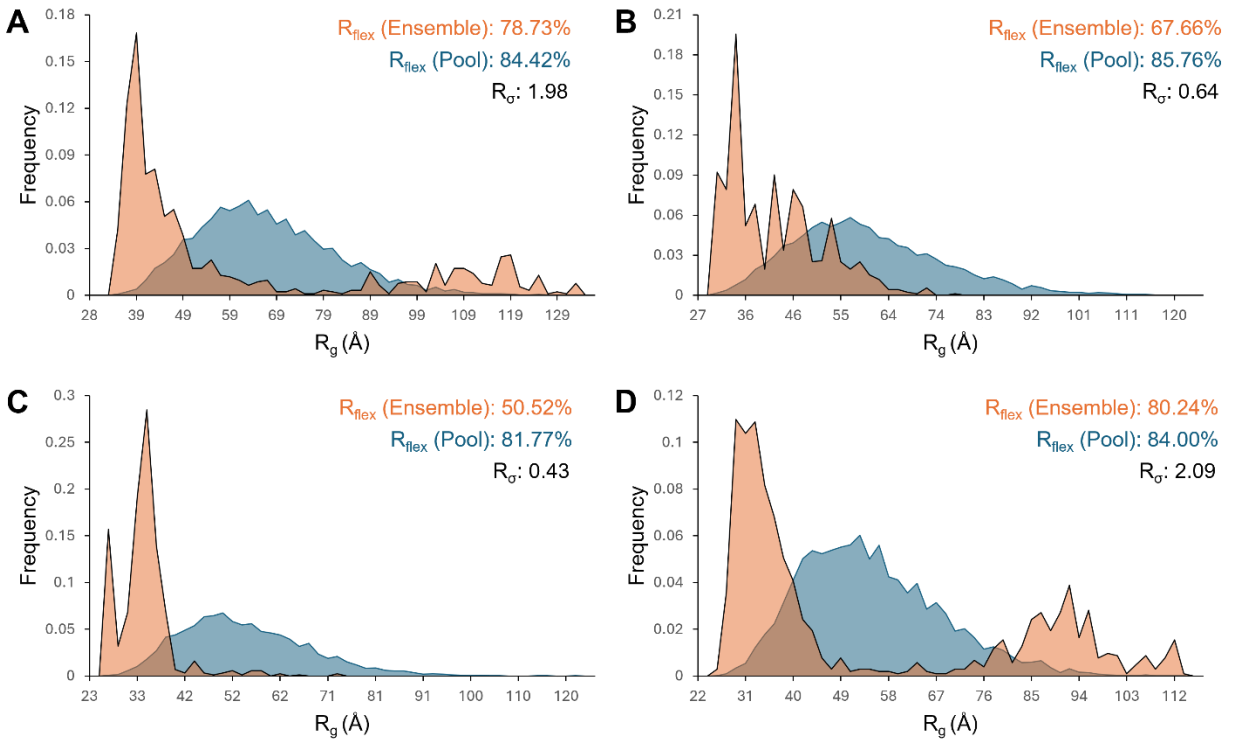

Figure 2.14 R_g Distribution of the EOM-generated TraG* Construct Models, where **A**) HisG*, **B**) HisG*$_{497\text{-}940}$, **C**) HisG*$_{497\text{-}863}$, and **D**) HisG*$_{F\ 497\text{-}863}$. Models with randomly generated conformations (shown in blue) are generated and the models from the pool which best fit the SAXS data are selected for the ensemble (shown in orange).

indicating an issue with the data[348]. The distribution map shows some selected ensemble models with larger R_g values than the pool, indicating they may be caused by the presence of aggregate. This is also seen with HisG*$_{F\ 497\text{-}863}$ which shows the presence of aggregates, whereas the other constructs which have no visible aggregate produced EOM models with distribution maps where selected ensembles are never larger than the largest models from the randomized pool (**Figure 2.14D**).

Despite issues with the EOM data from HisG* and HisG*$_{F\ 497\text{-}863}$, valuable structural information was gathered through comparing the models from HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$. The sequence assignments that provided the best χ^2 value for HisG*$_{497\text{-}940}$ (0.918) and HisG*$_{497\text{-}863}$

(0.858) involved designating the entirety of the sequences as compact (this was true for HisG*$_{F\ 497\text{-}863}$ as well, with a resultant χ^2 of 0.802). EOM models generated for these constructs therefore did not have any influence from these designations, allowing for an unbiased analysis of the conformational space and the regions predicted to be flexible. The models for HisG*$_{497\text{-}940}$ were mostly compact, with a highly globular model 1, and a moderately folded model 2, slightly larger in size when comparing to the DAMMIF model and the average of the ensemble (**Table 2.6**). These models represent 77% of the conformations sampled by the protein, and they indicate the dynamics of this construct may conformationally unfold mainly from the interacting region as it appears to be the most flexible region of the model (**Figure 2.13B**). The other two HisG*$_{497\text{-}940}$ EOM models are increasingly disordered, demonstrating additional disorder at both termini. The HisG*$_{497\text{-}863}$ EOM models validate the trend that this construct is the most globular of the truncation mutants tested; comparing the average ensemble statistics confirms this (**Table 2.6**). The compact HisG*$_{497\text{-}863}$ models follow a similar trend as the compact HisG*$_{497\text{-}940}$ models, as model 1 has a smaller R_g than the value from the P(r) function and model 2 is slightly larger, with the loop region displaying increased dynamicity (**Figure 2.13C**). The proportions with which these globular conformations are presented is increased at 50% for model 1 and 42% for model 2; an unfolded conformation similar to model 3 is achieved by only 8% of molecules. The distribution maps also show the flexible nature of HisG*$_{497\text{-}940}$, as the R_{flex} of the ensemble is above 50%, with an R_σ above 0.5 and less than 1 indicating a significantly flexible system (**Figure 2.14B**)[348]. These values are reduced compared to HisG*, indicating a less flexible construct was achieved by removing the N-terminal flexible region, and the trend continues for removal of the C-terminal helix as the R_{flex} of the ensemble for HisG*$_{497\text{-}863}$ is at 50% with an R_σ below 0.5, indicating the system is compact with moderate flexibility (**Figure 2.14C**). The EOM data for HisG*$_{F\ 497\text{-}863}$ does

not indicate the same compact nature likely due to the aggregate scattering present; the distribution of conformations is similar to HisG* that show a similar percentage of models representing an array of conformations from compact to fully disordered (**Figure S9**).

The conformational models for HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ provide evidence that the structure prediction software used in **Section 2.3.1** may be correct in their classification of the interacting region and the C-terminal region as a loop and a dynamic α-helix, respectively. In the highly globular model 1 of both constructs the interacting region is folded into the structure, however in model 2 when some unfolding occurs it is the residues of the interacting region which coincide with the least structured component. As the protein becomes less folded in models 3 and 4 of HisG*$_{497\text{-}940}$, the C-terminal region is predicted to unfold at a higher propensity. As mentioned, no sequence designations were used in EOM modeling of these proteins; no bias was introduced in their production, thus validating that the flexible regions shown in the models are likely to be accurate. The globular states of both TraG* constructs indicate that these regions are often folded tightly to the protein, however they can dynamically unfold into loops. It is not overlooked that there are other regions present in models from both TraG* constructs, specifically in the N-terminal regions, which also display higher dynamics despite removal of the 45 N-terminal residues. This may indicate that the N-terminal His-tag may be disordered, however, His-tags have been shown to reduce the dynamic movements of the conjugated protein[374,375]. Therefore, it is likely that there are more residues on the N-terminus of TraG*, and potentially the C-terminus, which are highly dynamic and could be truncated to produce a construct amenable to crystallisation.

2.3.6 Determination of the TraS$_F$ Interacting Region by GST-Pulldowns was Obscured by TraG$_F$ Aggregation*

In expressing GST-tagged TraS$_F$ peptides for their purification and subsequent use in GST-pulldowns, GST-tagged TraS$_F$ peptide 4 (GSTTraS$_{P4}$) formed inclusion bodies as the protein was seen most abundantly in the insoluble pellet following sonication and centrifugation, rather than in the soluble supernatant. As this peptide was from a region predicted by AF to have only a small hydrophilic loop with the remainder of the sequence being part of the TM helix (**Figure 2.2**), it is likely that this peptide is too hydrophobic to be properly expressed even when conjugated to GST. This property may be sufficient to justify that the region encompassed by the peptide sequence may not interact with TraG*, however GSTTraS$_{P4}$ was attempted to be purified as described in **Section 2.2.3.3** and was used for GST pulldowns to ensure that the region does not interact with TraG*.

The GST-pulldowns with HisG*$_F$ displayed HisG*$_F$ bound to the beads with high abundance after the final wash for all experiments, including the GST control. This is likely caused by the high aggregation propensity of HisG*$_F$, resulting in the association of the protein to the GSH-agarose beads leading to detection of the protein via SDS PAGE despite pipetting the beads to another tube prior to adding the SDS dye solution and boiling them, which was done to prevent any protein that may aggregate to the walls of the tube from appearing in the diagnostic sample. Some aggregation was seen when the experiment was performed with HisG*$_{F\ 497\text{-}938}$, as all experiments including the GST control showed the faint presence of a ~52 kDa band in the final sample (**Figure S10 A & B**). In pulldown experiments with HisG*$_F$ and HisG*$_{F\ 497\text{-}938}$ it was ensured that the sixth and final wash of the pulldown had no residual protein detectable by UV$_{280\ nm}$ and SDS PAGE gel. As the GST tagged peptides were all ~28.2kDa (except GSTTraS$_{P5}$, which

is 28.8 kDa), there is no detectable amount of GSTTraS$_{P4}$ bound to the beads as seen on the SDS PAGE gel, yet a band for HisG*$_{F\ 497\text{-}938}$ is faintly present indicating the unreliability of the experimental result with this construct.

To avoid issues caused by aggregation of TraG*$_F$ constructs, the GST pulldown experiment was reversed to a Ni^{2+} IMAC pulldown where HisG*$_F$ constructs were used as the bait and the GST-tagged TraS$_F$ peptides were introduced for the pulldown steps. It was theorized that in using this experimental configuration, the formation of HisG*$_F$ aggregate onto the beads would not alter the outcome of the experiment as they are expected to bind the beads, however only the GST-TraS peptide constructs which interact with HisG*$_F$ constructs would be visualized on the beads at the end of the experiment. Contrary to the hypothesis, a similar result to the GST-pulldown experiments was seen for both HisG*$_F$ and HisG*$_{F\ 497\text{-}938}$ Ni^{2+} IMAC pulldown experiments, with all GST-TraS constructs appearing in the beads, including the GST control (**Figure S10 C & D**). This may be caused by coaggregation of the HisG*$_F$ construct to the GSTTraS proteins; it is not uncommon for aggregated proteins to promote co-aggregation with other proteins and influence the result of pulldown experiments[376–379].

To accurately determine the TraS$_F$ peptide sequence which binds TraG*, the most globular form of the TraG*$_F$ protein was used for GST-pulldown, HisG*$_{F\ 497\text{-}938}$. Some aggregation is seen with this construct, as visualized via SDS PAGE by the presence of a band with a MW of 45 kDa in the final sample of the GST control (**Figure 2.15**). This construct has been shown to aggregate in SEC-MALS-SAXS experiments, therefore generating an accurate and reproducible GST-pulldown experiment for the interaction between TraG$_F$ and TraS$_F$ may be difficult. SEC-MALS of the purified and concentrated HisG*$_F$ constructs and GST-TraS$_F$ proteins was attempted after incubation at 4 °C to determine if the presence of a stationary phase which allows for affinity

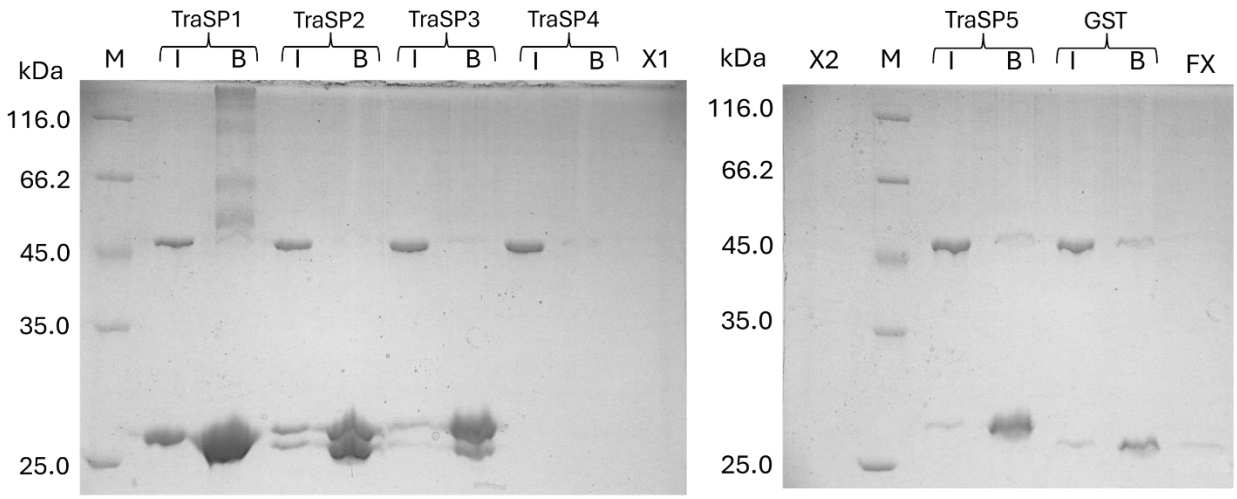

Figure 2.15 GST-TraS$_F$ Peptide Pulldown of HisG*$_{F\ 497\text{-}863}$. Samples were prepared as described in **Sections 2.2.7 & 2.2.4** and loaded on a 12.5% SDS PAGE gel before electrophoresis was performed at 180 V for 61 mins. M: unstained protein marker (ThermoFisher #26610), each GST-tagged TraS peptide that was bound to GSH-agarose beads is labeled above, as well as an un-conjugated GST control, where I: 2 µL of initial slurry after adding HisG*$_{F\ 497\text{-}863}$ to the GST-TraS-bound beads, prior to pulldown incubation, and B: 7 µL of pulldown bead sample after all wash steps were performed, X1: 20 µL supernatant after first wash to remove excess GST control during the binding step, X2: 20 µL supernatant after last wash to remove excess GST control during the binding step, and FX: 15 µL supernatant after last wash to remove excess HisG*$_{F\ 497\text{-}863}$ after the pulldown step.

interactions was disrupting the desired PPI, however no notable peaks representing interactions could be properly characterized due to the presence of aggregates obscuring the LS signal. However, some differences in binding to different GSTTraS$_F$ peptides were seen in this GST-pulldown experiment involving HisG*$_{F\ 497\text{-}938}$, yielding some valuable insight. The band associated with HisG*$_{F\ 497\text{-}938}$ was seen with higher intensity in the beads from the GSTTraS$_{P1}$ pulldown, and this lane featured additional high MW bands >50 kDa to >110 kDa despite the maintenance of quantitative and qualitative consistency between all experiments performed. Notably, the HisG*$_{F\ 497\text{-}938}$ bands seen in pulldown experiments involving GSTTraS$_{P2}$, GSTTraS$_{P3}$, and GSTTraS$_{P4}$, are virtually undetectable by SDS PAGE indicating that they do not interact with TraG*, while GST TraS$_{P5}$ and GST show similar abundance of bound HisG*$_{F\ 497\text{-}938}$. As the GST control was

purified from a blank pGEX4T2 vector and contains additional residues which are not present in the GST-TraS$_F$ constructs, there is a potential that the amino acid sequence present in the multiple cloning site of the vector causes unintentional binding to HisG*$_F$ constructs, resulting in the false positive interaction.

The residues which compose TraS$_{P1}$ and TraS$_{P5}$ have the potential to be an interacting region for HisG*$_F$ based on the results of the GST pulldown and the predicted topology of the TM structure; the region of TraS shown to bind TraG in the AF model was part of TraS$_{P5}$ (**Figure 2.2B**), however this model is likely inaccurate in the specifics of the binding interface. Based on the TraS$_F$ AF model (**Figure 2.2A**), which is predicted to decent accuracy (72% for RF as well), the hydrophilic regions more likely to form a binding interface include residues which make up TraS$_{P1}$, TraS$_{P3}$, and TraS$_{P5}$, with TraS$_{P1}$ and TraS$_{P5}$ forming a large portion of this interface as they are predicted to form laterally positioned α-helices. A proper interaction between TraS$_F$ and TraG$_F$ may require this entire TraS$_F$ interface.

2.3.7 Crystallisation Experiments on TraG Constructs Yielded Non-Diffracting Spherulites*

Previous vapour diffusion crystallisation experiments on HisG*$_F$ showed that in droplets where protein crystal nucleation occurred, the crystals formed were exclusively flat, 2D, disk-shaped spherulites which diffracted poorly[332]. Attempts at optimizing conditions and microseeding using these spherulites to improve the quality of the crystals proved unmeritorious as the droplets only produced HisG*$_F$ spherulites, some of which also resulted in protein aggregation that hindered harvesting for diffraction (**Figure 2.16A**). The spherulites were shown by SDS PAGE to be composed mainly of a truncated protein ~5 kDa smaller than the expected MW of the protein at

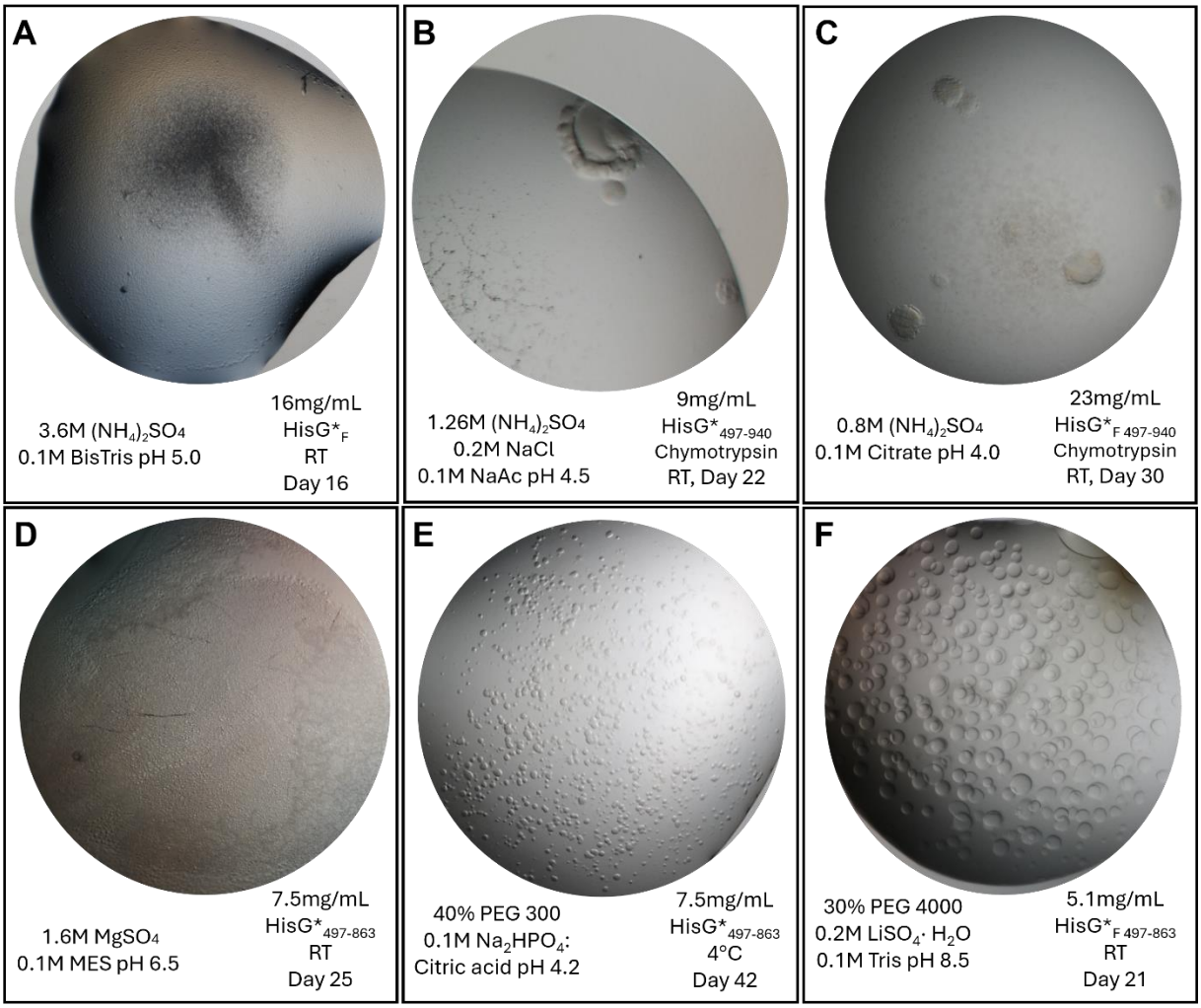

Figure 2.16 Crystallisation Experiments of TraG* Constructs. Reservoir conditions are shown in the bottom left of each panel while protein conditions, temperature, and the time elapsed since starting the experiment and capturing the picture are shown in the bottom right. Room temperature (RT) refers to the specific room in which these crystallisation experiments were incubated, which was maintained at 21 °C. Chymotrypsin refers to the addition of 1 : 100 µg ratio of α-chymotrypsin : TraG* in the protein solution immediately prior to starting the experiment. All protein stocks were in a buffer of or akin to 20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol, pH 7.0. **A**) is the result of a hanging drop vapour diffusion experiment, while **B**) to **F**) show results from sitting drop trials with reservoir solutions from commercial crystallisation screens. PEG 300 in **E**) was measured as a % (v/v) concentration, whereas PEG 4000 in **F**) was measured as a % (w/v) concentration.

56 kDa; this band was not present in the initial protein or mother liquor stocks used for the crystallisation experiment, providing the indication that a truncation mutant would be beneficial[332].

In crystallisation trials of HisG*$_{F\ and\ 100}$ truncation mutants, performing a PCT prior to performing random matrix screening with a sample ensured that the protein was sufficiently concentrated but not aggregated; in different purified batches of HisG*$_{497\text{-}863}$ a PCT produced a drop with the desired outcome of a light, granular precipitate when mixed with the high concentration precipitant (0.1 M Tris-HCl pH 8.5, 0.2 M Magnesium chloride hexahydrate, 30% (w/v) PEG 4,000) at both 7.5 mg/mL and 15 mg/mL. In all screening experiments, instances of crystal nucleation occurred more commonly in HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ than HisG* constructs from both plasmids, and the presence of visible aggregate in experimental droplets decreased significantly (**Figure 2.16**). The spherulites were also typically larger for HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ compared to those of HisG*, and some had different morphologies depending on the crystallisation conditions. This indicates that the removal of the N-terminal flexible linker region aids in crystal nucleation and growth and lowers the propensity of the protein for aggregation. As well, HisG*$_{497\text{-}863}$ is more soluble and could be brought to higher concentrations without aggregation, resulting in a higher propensity for crystal nucleation. However, all occurrences of HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ crystal nucleation manifested as microcrystals or spherulites, neither of which would change in quality during subsequent optimizations. The lack of crystal growth in the 3rd dimension but improvement of growth along the other 2 dimensions may lend to the theory that the lobe seen in DAMMIF bead models is preventing proper crystal formation, and the lobe is likely the dynamic interacting region. There was no perceivable difference in the quality or the quantity of crystals produced when comparing HisG*$_{497\text{-}940}$ and HisG*$_{497\text{-}863}$ crystallisation experiments, indicating that the truncation of the C-terminal helix had no significant effect on the

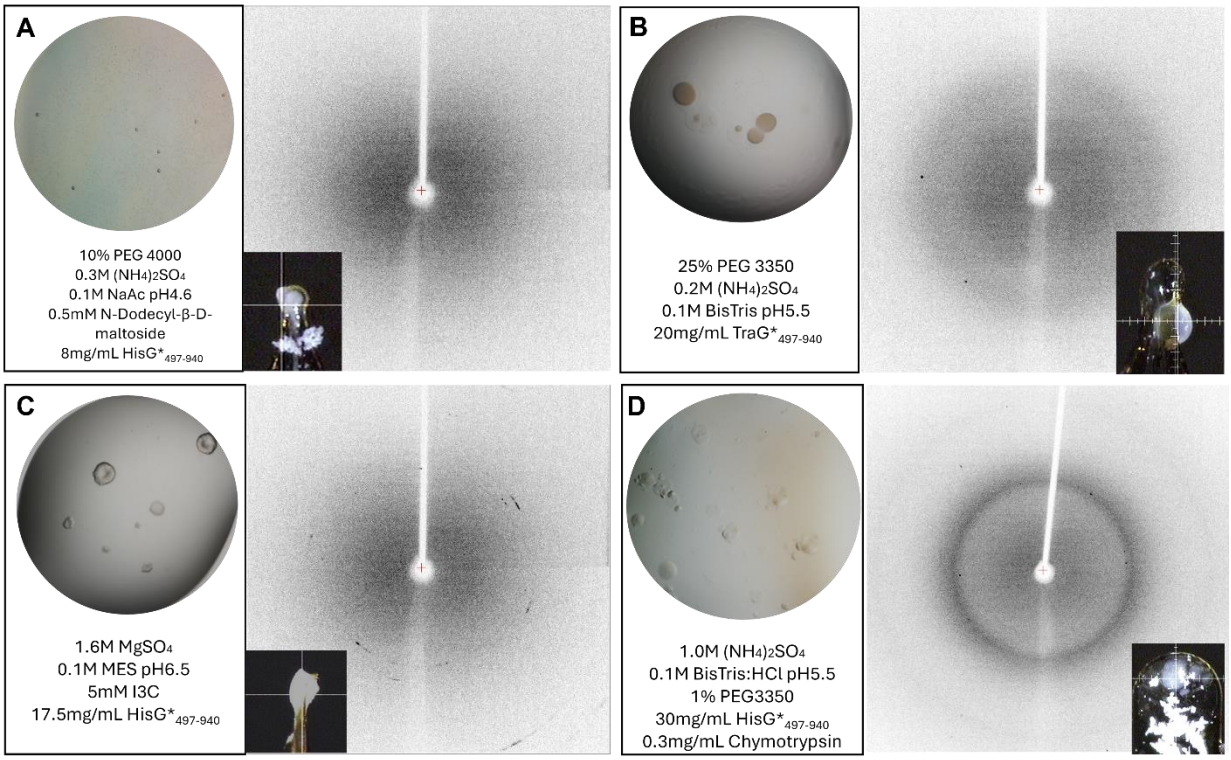

Figure 2.17 X-ray Diffraction of TraG* Construct Spherulites. Experimental conditions are shown in the bottom left of each panel with an image of the droplet above them, and the looped crystal within the picture of the resultant X-ray diffraction pattern. All conditions listed are the initial concentrations of solutions prior to droplet setup, and all protein stocks were in a buffer of or akin to 20 mM HEPES, 100 mM NaCl, 5% (v/v) glycerol, pH 7.0. Concentrations of PEG listed were measured as a % (w/v). The spherulites were grown in hanging drop vapour diffusion plates and were looped, flash frozen with liquid N_2, and diffracted while in a cryogenic stream. **A**) crystals were grown at 4 °C, imaged and harvested after 5 months of growth using a 50 µm microloop. **B**) the spherulites shown were grown at RT after 25 days of growth and were then harvested with a 150 µm microloop. **C**) this experiment was conducted for 3 months at 4 °C, the image was taken and spherulites were harvested using 200 µm microloop. **D**) thin spherulites resulted from this RT crystal trial which used a high concentration of HisG*$_{497\text{-}940}$; the droplet was imaged and spherulites were harvested with a 400/25 µm micromesh, to avoid cracking of the crystal, 23 days after commencing the experiment. All crystal images were taken using a Nikon stage microscope with a birefringent lens.

propensity for forming TraG* crystals. Crystallisation trials of HisG*$_{R100}$ constructs also appeared to result in a higher propensity of crystal formation than HisG*$_F$ constructs, providing further evidence that the interacting loop region disrupts crystal formation as this is the region with the

least sequence homology between the protein constructs from different plasmids. None of the produced spherulites diffracted sufficiently to produce a solved structure of a TraG* construct; all spherulites shown were diffracted in-house and at the Canadian Light Source (Canadian Macromolecular Crystallisation Facility, Insertion Device Beamline) and did not result in observable defined spots (**Figure 2.16 & 2.17**). Despite this, structural information of TraG* has been gleaned from the results of these crystal trials.

Conditions similar to those that were previously beneficial for the production of spherulites for HisG* and HisG*$_{497\text{-}940}$ also produced spherulites for HisG*$_{497\text{-}863}$[332]. These buffers were typically at a pH range of 5.0 to 6.5, which is near the pI of the protein (5.6), and the experimental conditions which most commonly resulted in crystallisation hits from matrix screening included NaAc, MES, or BisTris-HCl (**Figure 2.16 & 2.17**). These buffers were also seen to be those that best ameliorate the thermal stability of HisG* constructs in thermofluor assays (**Figure 2.4**)[7]. Many crystallisation trials were attempted with HisG* constructs after performing a buffer exchange to one of these optimal buffers, and **Figure 2.16D** shows small needle crystals that were produced. These crystals were difficult to harvest, were associated with low-quality diffraction, and were not able to be reproduced; there were no other differences in the outcome of crystallisation experiments when changing the buffer of the protein solution. As well, when performing crystallisation experiments at 4 °C, the pH range most optimal for crystallisation decreased to 4.0–5.5; as pH is known to increase at lower temperatures, the actual pH of the crystallisation conditions is likely higher than the reported pH at RT and may be akin to the optimal pH range previously reported for room temperature (RT) experiments (**Figure 2.16E**). Experiments performed at 4 °C took a longer period of time to grow spherulites, nucleation was

less common than for those grown at RT, and spherulites tended to grow larger at 4 °C, but they were otherwise no different from HisG* construct spherulites grown at RT.

Many crystalline results were observed from performing rmms using either lysozyme crystals or TraG* spherulites as the seed, however it became evident that many of the potential hits were salt crystals (results not shown); the conditions that developed crystals did so within an hour of starting the experiment, and many contained a Ca^{2+} salt in the mother liquor. The microseed stabilizing solution contained either $(NH_4)_2SO_4$ or $MgSO_4$, thus the crystals were likely $CaSO_4$. The addition of I3C also increased the propensity for salt crystals, perhaps due to dilution of the protein or the presence of LiOH from the I3C buffer, however some spherulites were produced from these trials indicating that the presence of I3C does not disrupt HisG* construct nucleation (**Figure 2.17C**). The spherulites observed were reproducible and of a different morphology than others with a concave shape to the center of the disk, bulging edges, and a larger average size of ~300 microns, however the diffraction screening of these crystals resulted in diffraction images with poorly visible spots and some minor ice rings.

In situ proteolysis was performed on all TraG* constructs as structural predictions showed that there are regions of the proteins which may be disordered on either termini, especially residues 807-863 on the C-terminus of $HisG^*_{497\text{-}863}$ (**Section 2.3.1**). If the TraG* chain ends are highly disordered and disrupting proper crystal growth, this method allows for a slow degradation of the termini such that a globule which can pack appropriately may be achieved. Results from performing *in situ* proteolysis during crystallisation trials demonstrated that using α-chymotrypsin resulted in more crystal nucleation than using trypsin (**Figure 2.16 B & C, Figure 2.17 D**), which aligns with the results from Dong *et al.* 2007[352]. Spherulites were seen with a higher relative frequency in experiments where *in situ* proteolysis was performed; as $CaCl_2$ is present in the stock

solution for the protease, some of the observed crystals may be salt, however the concentration of calcium ions in the final droplet was minimal. The spherulites observed in **Figure 2.16 B & C, Figure 2.17 D** were fragile, thin, 2D disks with slight birefringence, indicating they are proteinaceous[380], however some intense and well-spaced diffraction spots provide evidence that these spherulites may be partially or completely composed of a small molecule.

When trials were performed with untagged TraG* constructs, no observable differences were seen in crystal nucleation occurrence or the quality of the output crystals (**Figure 2.17B**). This follows the trend seen in many proteins where the presence of a His-tag does not influence crystallisation results[381]. In comparing results from TraG* constructs from F and R100 plasmids, crystal trials of TraG*$_{F}$ constructs appeared to result in LLPS more often or with higher abundance of LLPS droplets compared to their TraG*$_{R100}$ counterparts. **Figure 2.16F** shows an example of an experiment for HisG*$_{F\ 497\text{-}863}$ with excessive LLPS, whereas the same conditions resulted in a drop with light precipitate when experimenting with HisG*$_{497\text{-}863}$. This indicates that there are significant structural differences in the proteins despite their overall high sequence homology[113]; as the region responsible for interaction with TraS shares the lowest homology it is likely causing the differences in outcome for vapour diffusion experiments. Therefore, determining high-resolution crystal structures of TraG* from different plasmids may be useful for identifying the Eex protein's mechanism of function by comparing the structures of these interacting regions.

As TraG has a TM domain, it was hypothesized that performing crystallisation experiments appropriate for membrane proteins with TraG* constructs may yield crystals with different morphologies; proteins that are soluble but associate with membranes have been crystallized in the lipidic cubic phase[382,383]. Crystal trials were attempted using the MembFac screen performed as per the manufacturer's protocol[384] with either 0.1% (v/v) Nonidet P40 substitute (NP40) or

0.25mM N-Dodecyl-β-D-maltoside (final concentration in experimental drop) as these concentrations are sufficiently above their respective critical micellar concentrations[385,386]. No crystals with defined edges were produced in either case, and the most substantial spherulites were small pebble-like shapes approximately 50µm in size observed when using HisTraG*$_{497\text{-}940}$ with N-Dodecyl-β-D-maltoside; these spherulites were difficult to reproduce and diffracted similarly to other spherulites despite their different morphology (**Figure 2.17A**).

The interacting loop region was suggested to be the reason for difficulties in crystallising the optimized TraG*$_{497\text{-}863}$ construct based on SEC-MALS-SAXS results. Based on the results of the pulldown experiments, co-crystallisation experiments of TraG* constructs were attempted with untagged TraS$_{P1}$ and TraS$_{P5}$, however without sufficient evidence for the interaction it was not explored to an extent appropriate for co-crystallisation experiments. These experiments did not result in any observable protein crystals.

2.4 Conclusions and Future Works

TraG is a multifunctional protein in T4SSs encoded on F-like plasmids, composed of a membrane-bound N-terminal domain required for pilus generation and a C-terminal periplasmic domain (TraG*) required for Eex through interaction with TraS in a trans fashion and Mps through interaction with TraN in a cis fashion[4,7,113]. TraG* is presumed to have some plasticity in its structure to facilitate these functions that are expected to require numerous PPIs. Based on the findings presented it can be stated that the protein has many dynamic regions: the N-terminal region of TraG* (residues Ala452-Ala496) is highly dynamic and its presence destabilizes the protein[7], the C-terminal region (Met864-Gln940) is an extended α-helix which increases the

prolate nature of the protein, and the interacting region (Arg610-Asp673) is highly dynamic and may prevent proper crystallisation. These findings have been reflected in an updated schematic of the F-like T4SS that represents the impact that the work described in this chapter has on this research field (**Figure 2.18**).

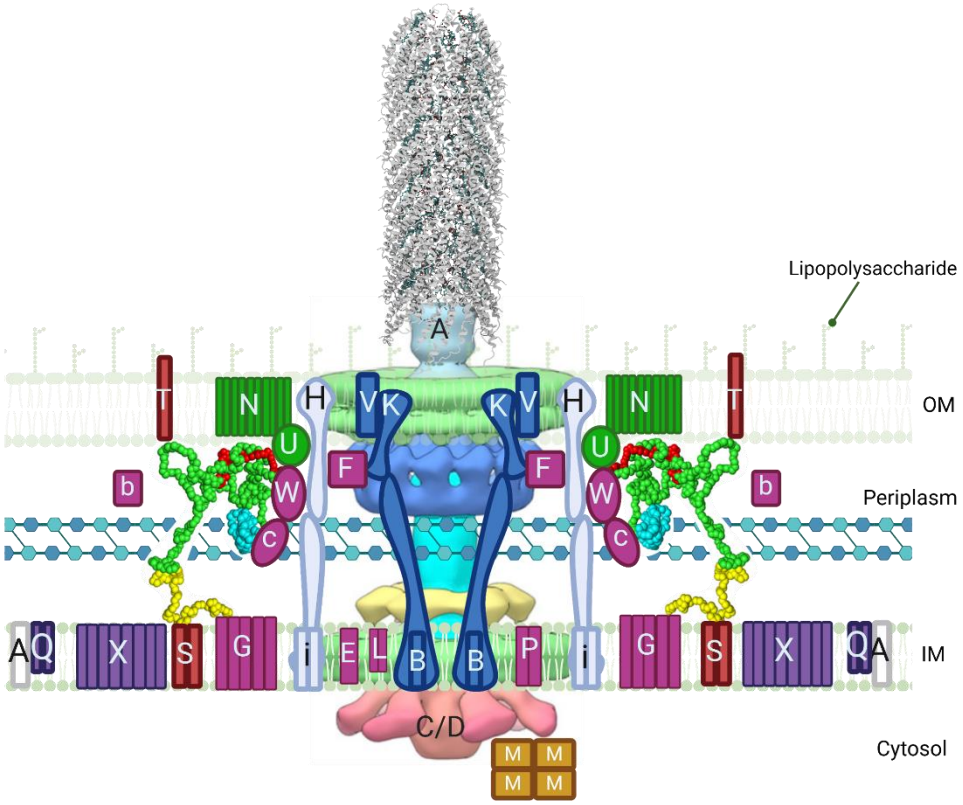

Figure 2.18 Updated Model of the F-like T4SS, based on available structural information including the findings on TraG described in this dissertation. The TraG* domain has been updated with the average EOM model of TraG* from **Figure 2.13A** with the same colour scheme representing the different characterized regions of the protein. The remainder of the schematic is the same as in **Figure 1.3B**. Figure created with BioRender.com[4].

The characteristics of higher thermal and chemical stability of the N-terminal truncation mutant observed from thermofluor and CD experiments are sufficient in stating that TraG*$_{497-940}$ is more stable than TraG*. In addition, the decrease in the presence of an unfolded population

observed in comparing TraG*$_{497\text{-}940}$ to TraG* via CIU-MS provides further evidence that the removal of the N-terminal region stabilizes the protein. TraG* also has a higher prevalence for aggregation as shown in SEC-MALS-SAXS and crystallisation experiments, and modelling via DAMMIF and EOM shows the region deleted in TraG*$_{497\text{-}940}$ is responsible for a higher proportion of unfolded conformations. It is proposed that when TraG* is expressed separately from the N-terminal TM domain, the linker region of TraG displays properties similar to an IDR. However, the data presented are not sufficient in identifying the deleted region as an IDR; IDRs shorter than 50 residues are not uncommon, however they typically serve a functional purpose in the native protein[387]. As this region has not been shown to be intrinsically disordered in TraG and its function is not yet confirmed, the region cannot be classified as an IDR. This study proposes the region be classified as a flexible linker, as this is congruent with known moieties in structural biology as the occurrence of unstructured regions approximately 50 residues in length are common in functional proteins[212,216,221]. Intrinsically disordered proteins are more common in eukaryotes and viruses due to a proclivity for complexity in their morphologies; it is suggested that there exists a link between intrinsic disorder and evolution[228]. However, there are many examples of prokaryotic proteins with IDRs (**Section 1.10.1**), and it is not uncommon in bacterial proteins for disordered regions to be important flexible linkers of folded domains[216,388,389]. Based on structural predictions (**Section 2.3.1**) and the known function of TraG*, it is reasonable to conclude the overall structure of the protein would require flexibility and therefore have some intrinsically disordered properties.

In comparing the produced DAMMIF bead models of TraG* constructs to the AF and RF predicted structures, it is highly likely that TraG* is an elongated rod-shaped structure with a distal lobe at its center, consisting entirely of extended α-helices and loops based on CD. The prolate shape of TraG is supported by the data shown in this study and is likely of significance for its

function in extending to contact its protein partners. It is difficult to assess whether the N- or C-terminal regions of TraG* truncated in the constructs are important in the PPIs of TraG as many of the interactions in T4SS proteins are highly transient and require a specific environmental context, and TraG* is non-functional when expressed on its own[113]. Confirming the functions of the N- and C-terminal regions are important in identifying whether they are non-essential for future high-resolution structural studies of TraG*.

The C-terminal region of TraG* was predicted to be an extended α-helix which may be responsible for increasing LLPS propensity of the protein; a C-terminally truncated construct was designed in an attempt to characterize the region and generate a protein which is sufficiently globular to properly crystallize. SEC-MALS-SAXS data demonstrated the globular nature of the TraG*$_{497\text{-}863}$ truncation mutant, indicating that the 77 C-terminal residues are an elongated structure that destabilizes TraG*$_{R100}$, however LLPS remains a frequent result in crystal trials for the protein. To further characterize this region, the suite of biophysical tests employed to compare TraG* with TraG*$_{497\text{-}940}$ could be performed on TraG*$_{497\text{-}863}$. A TraG* construct with only the C-terminal region truncated was not designed, however performing these biophysical tests to a TraG*$_{452\text{-}863}$ construct could reveal the nature of this region and further confirm the destabilizing nature of the N-terminal region as well.

Despite limited knowledge of their high-resolution structures, the interaction between TraN and TraG is predicted to involve an induced fit in which the C-terminal region of TraG conforms to interact with the well-structured TraN, perhaps with aid from other periplasmic T4SS components[4,112]. This could be confirmed through the use of conjugative mating assays similar to those performed for determining the interacting region of TraG with TraS[113,390]. These assays could also be used to determine the region of TraN which interacts with TraG; if this TraN region

could be expressed as a peptide perhaps it could be used for co-crystallisation of a TraG*-TraN peptide complex. Recent cryo-ET experiments by Kishida *et al.* (2024) examining the structure and function of the F-like T4SS after the systematic removal of various genes have elaborated on the role of TraN for pilus extension[65], however many advancements in the technique would be required to develop a high-resolution structure of the TraG-TraN interaction *in vivo* as the abundance of the complex within an *E. coli* minicell overexpressing a T4SS would be miniscule due to the transient nature of these Mps interactions.

TraG* has been demonstrated to be recalcitrant to proper crystallisation, and the results of this dissertation attribute the interacting loop region as the main hinderance to the production of diffraction-quality 3D crystals. Despite the production of an optimized TraG*$_{497\text{-}863}$ construct and a concerted effort to optimize crystallisation conditions based on results, the crystals produced were non-diffracting 2D spherulites; crystals likely do not grow in the 3rd dimension as the interacting region is predicted to be a highly dynamic loop which protrudes from the globular core of the protein, proposedly preventing crystal packing along the interface with the loop. Mutating this region may aid in crystallisation, however it is a highly important region for interacting with TraS and therefore a high-resolution structure of this region is desired. In future works, this region could be rationally altered in TraG* constructs to be made more easily crystallized; if successful the crystals may result in a high-resolution structure which can be used as a starting point for further mutational experiments to generate a construct with an intact interacting region that will crystallize[391,392]. As well, high-resolution biophysical techniques such as atomic force microscopy could be used on a growing spherulite of TraG* to confirm whether the interacting loop is the region preventing crystallisation in the 3rd dimension or if another region should be targeted for mutation[393].

Pulldown experiments using TraS peptides were attempted to contrive a binding partner to stabilize the interacting region of TraG*, however they failed to produce a valid result due to the aggregative nature of TraG* constructs. As the interaction between TraG and TraS is known to occur in a highly specific context, a successful pulldown experiment may not be achievable without expressing full-length TraG and TraS, and the interaction may require the context of other T4SS component and the lipid membrane. Conjugative mating assays in which TraS is mutated to discover the residues of TraS which interact with TraG may be difficult to assess as well due to the small size and hydrophobic nature of TraS, as mutating regions which are TM and not part of the interacting region may result in loss of Eex function merely due to misfolding of the protein in the OM[390].

As the only deletion mutants examined biophysically were from the described truncations, the sizes of the dynamic regions in TraG* have not been defined in full, however the findings of this study have several implications. Many structural predictive software packages such as Phyre2, IUPred2A, ANCHOR2A, RF and AF[328,329,394–396] improperly identified secondary structure properties of TraG*, overestimating the disordered content of the periplasmic protein to be approximately 50%. CD analysis of TraG* indicates the protein is not predominantly disordered, rather some regions are predicted to feature distorted helices at 20.3% of the protein (**Figure 2.5**). This indicates that these programs are incorrectly assigning regions of TraG* as disordered, likely due to an inability to identify TraG homologs with solved structures. Many software including AF and RF rely on solved protein structures in the PDB to train their machine learning AI algorithms, and the inaccurate prediction of TraG* by these programs indicates the lack of similar structures in the database, enforcing the postulation regarding the novelty of this protein structure. The regions predicted to be disordered may be folding cryptically, as identified by EOM models

(**Figure 2.13**) which show that TraG* is a flexible protein which can often sample a proportion of well-folded conformations, possibly afforded by these regions which may display secondary structure transiently depending on their environment. This cryptic folding may be a compounding factor to the difficulty in the accurate prediction of the structure of TraG by AF and RF, especially the interacting region which is highly variable between different plasmids.

TraG and TraS form the Eex system of the F-like T4SS and are heavily relied upon for preventing donor-donor plasmid exchange[111,137]. As excessive conjugation can lead to lethal zygosis it has been theorized that disruption of Eex systems would be detrimental to bacterial colony survival[67]. The development of molecules which can disrupt the TraG–TraS interactions and cause ceaseless conjugation or completely prohibit conjugation could permit the creation of a new class of antibiotics targeting gram-negative bacterial infections and prevent the conjugation-mediated spread of antibiotic-resistant plasmids in a colony. These antibiotics would be useful in tandem with other antibiotics as a drug cocktail for targeting bacteria with a characterized F-like plasmid containing an antibiotic resistance gene[7]. However, developing these antibiotics in a rational drug discovery effort is contingent on the high-resolution structural knowledge of the system's protein subunits. Other than the experiments suggested thus far, to solve the structure of TraG* it may be required to express the full-length protein, solubilizing the membrane protein into a lipid nanodisc, then purifying and crystallizing TraG in the lipidic cubic phase[285,382,383]. A MembFac screen was attempted with TraG* constructs[384], however the proper methodology for crystallising a membrane protein *in meso* may lead to a diffraction-quality crystal. Optimizing the expression of TraG into lipid nanodiscs may be challenging due to the dynamic nature of TraG*; this would also complicate attempting techniques such as cryo-EM to solve the structure of TraG to a high resolution in the nanodisc as the protein is smaller than 150 kDa[397,398]. The smallest

TraG* construct designed is 45kDa, which is likely too large for a high-resolution characterisation of the protein by nuclear magnetic resonance techniques[399]. The dynamic nature of TraG* renders the high-resolution structural solution of TraG from F-like T4SSs a challenging task. Continuing to approach the characterisation of these proteins in a stepwise manner using structural techniques and orthogonal biophysical techniques is the most logical manner of solving the structure of TraG and its interactions with TraS and TraN to elucidate the unique mechanism of this system.

CHAPTER 3 – THE 1.3Å STRUCTURE OF THE GROUP IA PILIN FROM PSEUDOMONAS AERUGINOSA STRAIN P1

3.1 Research Objectives

The ΔP1 pilin was initially expressed, purified, crystallized and diffracted in 2004 around the same time as the ΔK122-4 pilin[210]. However, the phase problem could not be solved for ΔP1 at the time (GF Audette, personal communication); purification, crystallization details are provided for completeness. In 2021 when AlphaFold became accessible, a predictive model of ΔP1 was generated; many have demonstrated the algorithm's ability to generate protein models suitable for use in MR[328,400–402]. Prior to this, a reliable MR search model for the ΔP1 pilin was not readily available, though there was the submission of the structure of the truncated 1244 (Δ1244) pilin to the PDB (ID 6BBK)[403]. This chapter describes how the phase problem was overcome using MR with a processed search model generated from the ColabFold algorithm[333]. The ΔP1 pilin structure was refined to 1.30 Å, allowing for a resolved view of side chain positioning and important solvent contacts. An analysis of Phaser results from testing different MR models shows using a processed AF model enhances the quality of initial maps compared to using a nearly identical structure from the Δ1244 pilin. SEC-MALS on ΔP1 after attempting *in vitro* oligomerization triggered by a hydrophobic catalyst provides evidence of differing behaviors between ΔP1 and ΔK122 and provides a model for dimerization of the ΔP1 pilin. Protein Interfaces, Surfaces and Assemblies (PISA), Procrustes Structural Matching Alignment and Restraints Tool (ProSMART), and Clustal Omega analyses were used to compare solved pilin structures from different pilin groups[404–406]. These results indicate structural homology may exist between pilins from different groupings, thus displaying the need for further structural analysis of *P. aeruginosa* pilins for a nuanced structural grouping to aid in anti-virulence drug and vaccine development[8].

3.2 Materials and Methods

3.2.1 Expression and Purification of ΔP1

Cloning, expression and purification of the ΔP1 pilin was performed in a similar fashion to that as previously reported for ΔK122[210]. Briefly, the P1 pilin gene was excised from a pRLD plasmid containing $pilA_{P1}$ by restriction digest using EcoRI and HindIII. DNA was isolated from a 1.2% agarose gel using standard procedures and ligated in-frame into the pMAL-p2 expression vector (NEB). ΔP1 [*pilA(Δ1-28)*] was expressed periplasmically in *E. coli* strain ER2507 as a maltose-binding protein (MBP) fusion protein and purified using an amylose column. The purified MBP-ΔP1 fusion protein was trypsinized to release ΔP1 pilin from MBP. The monomeric ΔP1 pilin contains four N-terminal residues (ISEF) from the expression construct followed by residues 28-148 of the P1 pilin (this numbering scheme excludes the 6-residue cleaved signal sequence, MKAQKG). Monomeric ΔP1 was then purified by cation-exchange chromatography and quantified using UV_{280nm} spectroscopy and the known molar absorptivity coefficient 0.694 ml mg^{-1} cm^{-1}.

3.2.2 Crystallization and Structural Determination of ΔP1

All crystallization experiments were performed using the hanging-drop vapour diffusion method at 295 K. Initial crystals of the ΔP1 pilin were grown in 1-2 days from 2 µl drops containing equal volumes of protein (20 mg/ml in 10 mM Tris-HCl pH 7.4, 100 mM sodium chloride) and reservoir solution (35% (w/v) PEG 4000, 100 mM sodium cacodylate pH 6.0, 100 mM monobasic potassium phosphate). These initial crystals were used as a microseed stock for further crystallization experiments. Crystals suitable for X-ray diffraction analysis (typical dimensions of

~ 180 x 60 x 30 microns) were grown from 3 µl drops containing equal volumes of protein, reservoir (30% (w/v) PEG 4000, 100 mM sodium cacodylate pH 6.0, 100 mM monobasic potassium phosphate), and microseed solution. Crystals were mounted in cryoloops (Hampton Research) and flash-cooled by direct immersion into liquid nitrogen prior to X-ray diffraction analysis.

X-ray diffraction data was collected on a single crystal of ΔP1 (dimensions of ~ 180 x 60 x 30 microns) on beamline 8.3.1 at the Advanced Light Source in 2003 on an ADSC Quantum 315 charge coupled device detector. Diffraction data was collected at 100 K in two passes. The first, lower resolution data set was collected with the following experimental parameters: 300 mm crystal-to-detector distance, and λ = 1.0236 Å, 10 seconds per exposure, angular range covering 0–90° in a total of 90 frames with 1.0° per image. The second, high resolution data set was collected with the following experimental parameters: 150 mm crystal-to-detector distance, and λ = 1.0332 Å, 30 seconds per exposure, angular range covering 0–146° in a total of 292 frames with 0.5° per image. The low- and high-resolution passes were indexed using MOSFLM[407], following which the indexed reflections were merged and scaled with SCALA from the CCP4 suite to 1.20 Å resolution. At the time, structure solution of the ΔP1 pilin was unsuccessful (either by MR or anomalous signal phasing). The images were archived, with only the indexed data files (one each for low and high-resolution pass) remaining available. Recently the merging and scaling of the MOSFLM indexed data was revisited employing AIMLESS from the CCP4 suite[408], identifying a more appropriate resolution cut-off as 1.3Å based on $CC_{1/2}$ and other recently deployed metrics. Based on these metrics, 3% of reflections for R_{free} were selected for comparison in monitoring the R_{work} value in refinement. Data collection statistics for ΔP1 are shown in **Table 3.1**.

The structure was solved using MR by Phaser[409] via the CCP4i graphical use interface[410] using an AF *ab initio* model generated using ColabFold (v.1.5.2) that was processed using the "process predicted model" function in CCP4i2, which generates a model separated into discrete regions based on the pLDDT score of the model[333,411]. Patterson map analysis would automatically set the translational non-crystallographic symmetry (tNCS) to 2 indicating a dimer in the unit cell, however based on a Matthew's coefficient analysis the MW to which the volume of the asymmetric unit applies was 22,423 Da, while a ΔP1 dimer would be 26,068 Da. Restricting the tNCS to 1 molecule per asymmetric unit allowed for the solution of the structure, which was solvated at 30.72%. The MR result was then input into ARP/wARP, resulting in a solution with R and R_{free} values of 0.1848 and 0.256, respectively, due to dummy glycine residues in the place of the N-terminal I25 to T67 and the C-terminal Ser147. The model was then examined in COOT, sidechains were manually assigned, and a number of residues showed occupancy in multiple conformations (Thr34 [0.5, 0.5], Thr45 [0.5, 0.5], Ser118 [0.5, 0.5], Lys120 [0.5, 0.5], Cys127 [0.8, 0.2], Lys128 [0.6, 0.4], Pro146 [0.5, 0.5])[412,413]. Solvent atoms were added with ARP/wARP and confirmed with visual inspection; solvent density corresponding to a cacodylic acid (dimethylarsinic acid) molecule was observed and modeled accordingly. The model was processed in several rounds of remodeling and refinement performed in COOT and with REFMAC5, respectively[413–416]. Restrained individual anisotropic B values were refined in the final cycle prior to validation using SfCHECK, ProCHECK and RAMPAGE[415,417–419]. The validity of the anisotropic B value refinement at this resolution was supported by the R_{free} statistic, which dropped from 0.1987 to 0.1738. The final model contains 124 residues, a single cacodylate ion, and 216 waters; refinement statistics are summarized in **Table 3.2**, and the final model and structure factor amplitudes submitted to the RCSB Protein Data Bank with accession code 8V7P[8].

Table 3.1 Data Collection and Processing Statistics of the N-terminally truncated P1 pilin. See **Section A.5.1** for more information on parameters for validating X-ray diffraction data.

Diffraction Statistics	
Diffraction source	ALS Beamline 8.3.1
Wavelength (Å)	1.0236
Temperature (K)	100
Detector	ADSC Quantum 315
Space group	$P2_1$
a, *b*, *c* (Å)	22.45, 54.85, 37.77
α, β, γ (°)	90, 96.15, 90
Mosaicity (°)	0.79
Resolution range (Å)	54.85-1.30
	(1.33-1.30)
No. of observations	89356 (3512)
No. of unique reflections	22344 (1309)
Completeness (%)	99.5 (99.2)
Redundancy	4.0 (2.7)
$\langle I/\sigma(I)\rangle$	14.2 (3.9)
$R_{p.i.m.}$	0.069 (0.355)
R_{merge}	0.141 (0.498)
$CC_{1/2}$	0.988 (0.639)
Overall *B* factor from Wilson plot ($Å^2$)	7.6

Values for the outer shell are given in parentheses.

Table 3.2 Refinement Statistics of the N-terminally truncated P1 pilin. See **Section A.5.1** for more information on parameters for validating X-ray diffraction data.

Refinement Statistics	
Resolution range (Å)	54.85–1.300 (1.334–1.300)
Completeness (%)	99.5
No. of reflections, working set	21642 (1603)
No. of reflections, test set	682 (39)
Final R_{cryst}	0.1306 (0.131)
Final R_{free}	0.1738 (0.176)
No. of non-H atoms	
Protein	957
Ligand	5
Water	216
R.M.S. deviations	
Bonds (Å)	0.0100
Angles (°)	1.9165
Average *B* factors (Å^2)	14.43
Protein	11.525
Ramachandran plot	
Most favoured (%)	97.5
Allowed (%)	2.5

Values for the outer shell are given in parentheses.

3.2.3 Sequence and Structure Comparisons

PISA (Protein Interfaces, Surfaces and Assemblies) analysis was performed on the ΔP1 structure using PDBePISA v1.52 with processing mode set to auto[404]. Interface results were used to search for other proteins with similar interfaces, and monomers and assemblies were analyzed

to view potential quaternary structures that can form based on the ΔP1 structure. Model *P. aeruginosa* pilin proteins with solved structures and structures with similar interfaces as ΔP1 from PISA results were compared using ProSMART (Procrustes Structural Matching Alignment and Restraints Tool)[405]. ProSMART align was used from the CCP4i interface with default settings to compare the structures in a conformation-independent manner. Multiple sequence alignment of the pilins was performed using Clustal Omega to compare to structural alignment values; the results were visualized on JalviewJS v2.12[420,421].

3.2.4 SEC-MALS

The ΔP1 pilin was purified as described above and concentrated to 23 mg/mL using a 5 kDa MWCO centrifugal concentrator (Millipore) at 3,500 x g, 4°C. The protein sample was injected at a concentration of 20 mg/mL at a volume of 100 µL onto a Yarra s3000 SEC column (separation range 5 kDa – 700 kDa; Phenomenex) using an ÄKTA Purifier 10S FPLC system (Cytiva) connected in-line to a Dawn Heleos II MALS system and Optilab T-rEX dRI detector (Wyatt Technology). The running buffer used was 20 mM Tris-HCl, 150 mM NaCl, pH 7.4 and was loaded at a flow rate of 0.5 mL/min.

Petrov et al. (2013)[171] characterized the oligomerization of ΔK122-4 into PNTs using a variety of hydrophobic trigger molecules and determined the optimal catalyst as 2-methyl-2,4-pentane-diol (MPD). A similar procedure was followed to oligomerize ΔP1 into a PNT through the addition of the trigger solution (10 mM Tris-HCl, 300 mM NaCl, 1 mM EDTA, 1 mM DTT, 1 M MPD, pH 7.4) to 23 mg/mL ΔP1 in a 10:1 (v/v) protein-to-hydrophobe ratio. The oligomerization reaction was incubated at RT with nutation for 96 hrs. SEC-MALS was performed

in the same manner as described above; analysis of SEC-MALS data was performed using the Astra software package (v6.0) and provided R_h, MW, polydispersity, and translational diffusion coefficient distributions as calculated from an autocorrelation function. Based on SEC-MALS data and previous experimental data showing a three-start helical assembly for the oligomerization of pilins, a model of the assembled P1 pilus was created using the truncated P1 monomeric structure[171,202].

3.3 Results and Discussion

3.3.1 Quality of the Final Model

The structure of the monoclinic crystal form of ΔP1 pilin was refined to 1.3 Å resolution, with a final R_{work} and R_{free} refined to 0.1306 and 0.1738, respectively (**Table 3.2**). In the final model all residues were observed and 216 water molecules and a cacodylic acid molecule could be modeled in the electron density. Ile25 and Thr66 had poor fit to the electron density, with real space R (RSR) Z-scores[422] of 4.4% and 2.6%, respectively. The high real-space-R value Z (RSRZ) score of Ile25 is common for the first residue of the protein chain as terminal residues typically have higher conformational disorder; also, Ile25 is a residue resulting from the N-terminal MBP tag and therefore is not relevant in discussing the native ΔP1 structure. The average B factors (main chain/side chain) are 10.02 $Å^2$/13.21 $Å^2$ and 27.21 $Å^2$ for the solvent molecules. Analysis of the stereochemical quality of the final model was performed using ProCHECK, SfCHECK and RAMPAGE[417–419,423]. Prototypical Ramachandran plots compare ϕ and ψ angles around C_α in every residue of a polypeptide to validate main chain geometry in a protein structure; the maps are typically derived from strict parameters of the most favourable and lowest-energy covalent bond

angles based on a residue's sidechain and the secondary structural element it resides in[424]. RAMPAGE provides a more nuanced validation by accounting for both the bond angles and the torsion angles together within the local combinations of residues, as well as using the C_β deviation (the observed C_α-C_β distance from the ideal one) as a single, simple measure of geometrical nonideality around the C_α[417]. These conditions allow for acceptability of some residues with atypical bond angles which have been seen in high resolution, low B-factor protein structures; these residues have minor strain but are valid conformations, for example in the active site serine of lipases where steric overlap is present in the C_β and the carbonyl[417,425]. The RAMPAGE Ramachandran plot features contours for favored (core), allowed, less favoured but allowed, and disallowed bond angles in Gly, Pro, pre-Pro (residues immediately preceding Prolines), and the general case of the 18 other amino acids[417]. The results of RAMPAGE analysis indicated that 97.5% of residues (120) in ΔP1 are in the core or allowed regions of the Ramachandran plot, and only 2.5% of residues (Asp65, Thr73, and Lys104) are in less favored regions (**Figure 3.1**). ProCHECK showed that the residues with less favoured bond angles have reasonable side chain RMSDs (0.0624 Å, 0.1090 Å, 0.0638 Å, respectively), indicating that they were properly modeled into the density and their conformations are not an anomaly. A single side chain planarity outlier was seen in Arg35, and a bond angle outlier was seen in Arg119; no strained stereochemical conformations were observed.

3.3.2 Features of the ΔP1 Structure

The ΔP1 crystal structure exhibits the characteristic type IVa pilin fold, with the truncated N-terminal α-helix (α1-C) packed onto the head domain consisting of a four-stranded antiparallel

β-sheet (**Figure 3.2A**). The α-β loop immediately follows the amphipathic α1-C helix, and the receptor binding D-region is present with the conserved disulfide bond covalently linking it to the last strand of the β-meander. While these features are common, minor differences in T4 pilin structures result in the observed functional diversity of the *P. aeruginosa* pili; as no other T4a group I pilin structures have been described in literature this chapter will detail the side chain interactions that result in the observed conformations of the α-helix, the β-sheet, the D-region, and the connecting loop regions.

3.3.2.1 The α1-C Helix Displays a Kink Stabilized by Solvent Interactions

The α1-C helix has some interesting side chain interactions that may be unique despite the region holding high sequence conservation in other T4 pilin proteins. Some residues on the helix contribute to stabilizing the conformation of proximal hypervariable loop regions, such as Gln32, which has two conformers, one which uses the carboxyl group of the side chain to hydrogen (H) bond to the backbone nitrogen of Gly107 (2.99 Å) and its amine H-bonds water molecule 211 (2.96 Å), while the other conformer H-bonds to water 2 (2.66 Å). Glu39 H-bonds strongly with Ser75 (2.48 Å), which connects the α-β loop to the helix. Thr34 also has two conformers, one of which performs H-bonding with backbone carbonyl groups of Arg30 and Thr31 (3.10 Å and 3.15 Å, respectively) presumably stabilizing the helix's conformation, and the other conformer H-bonds two water molecules, water 10 and water 105 (2.85 Å and 2.70 Å, respectively). There is a shallow bend in the helix starting at Ser41 which offsets canonical hydrogen bonding of the main chain (**Figure 3.3**). The carbonyl in the backbone of Ser41 H-bonds weakly with both backbone amides

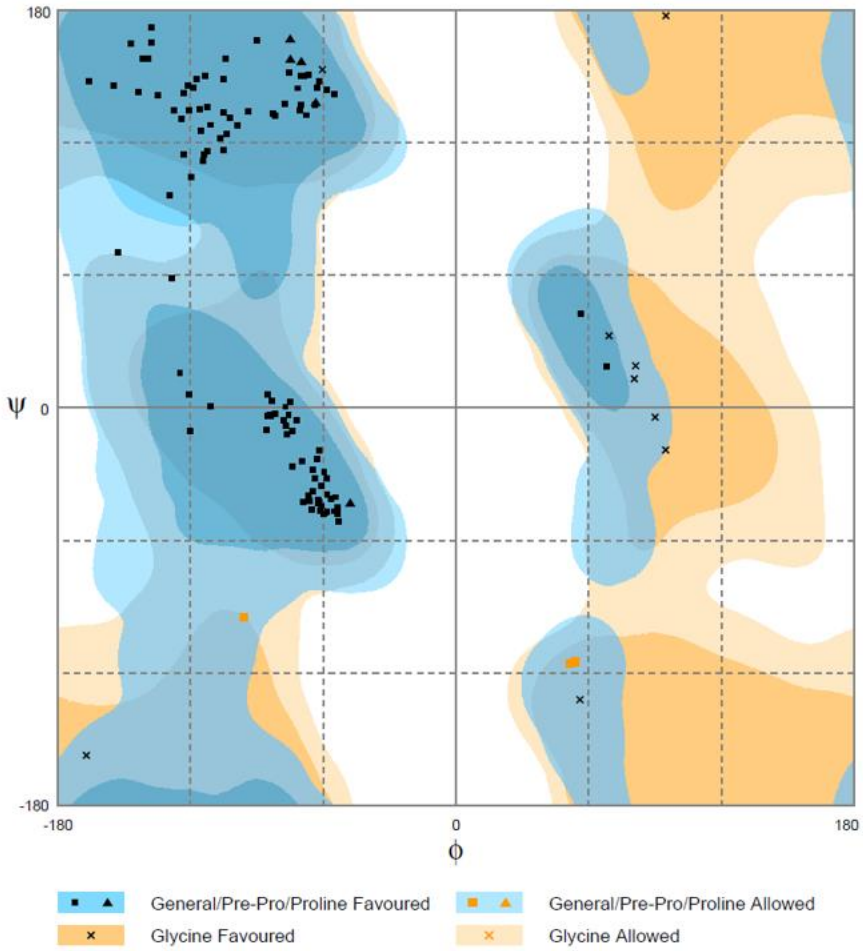

Figure 3.1 Ramachandran Plot of ΔP1 Residues, from RAMPAGE[417]. All pre-proline, proline, and glycine amino acids have favoured ϕ and ψ C_α bond angles; of the other amino acids, only Asp65, Thr73, and Lys104 are in a less favoured but allowed region.

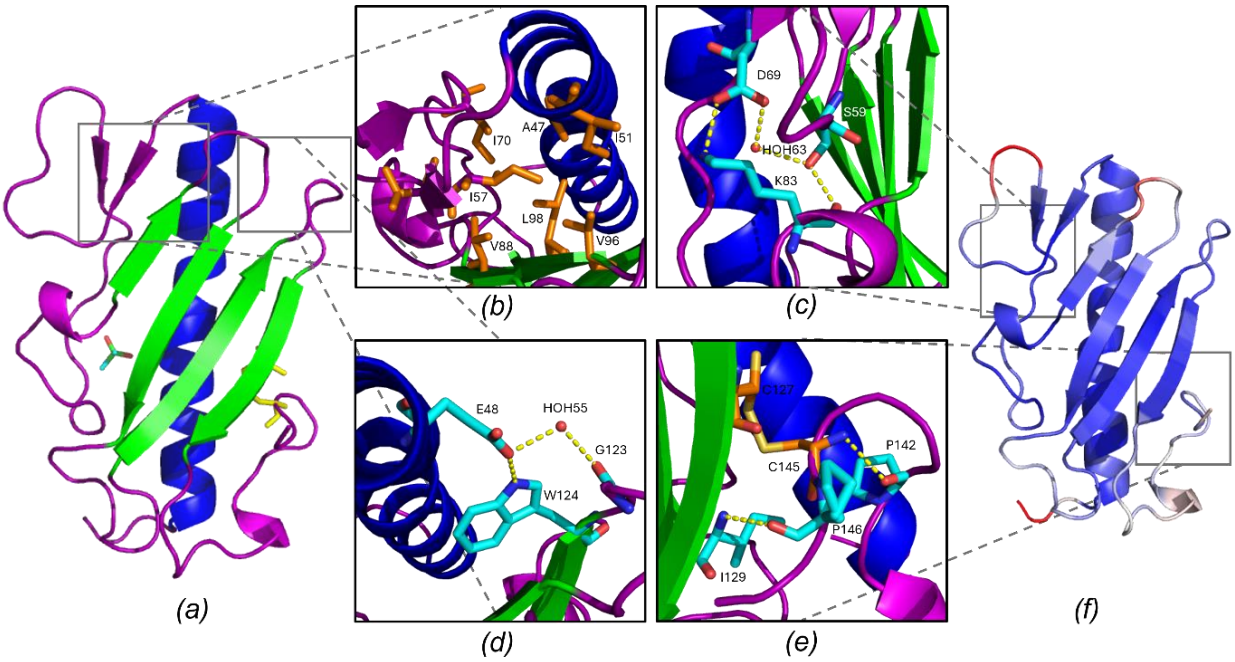

Figure 3.2 The Structure of the ΔP1 Pilin, with important residue interactions highlighted (PDB ID: 8V7P). **A**) The ΔP1 pilin monomer; the N-terminal α-helix (α1-C) is in blue, the β-sheet is in green, and the coil regions are in purple. The disulfide between Cys 127 and 145 is shown in yellow and the cacodylic acid molecule is shown as sticks with atomistic colouring. **B**) A portion of the hydrophobic interface seen between the α1-C helix, the β-meander and the α-β loop. Side chains of the hydrophobic residues (orange) are pointed towards the interior of the molecule to create a hydrophobic pocket. **C**) At 1.3 Å resolution numerous solvent contacts were seen in the ΔP1 structure, including the hydrogen bonding network (dotted yellow lines) containing a water molecule (in red) which aids in maintaining the conformation of the α-β loop. **D**) A view of the bend in the α1-C helix and the H-bonding between the conserved Trp125 and Glu48, as well as the backbone of Gly123 and water55 which stabilizes the kinked α-helix. **E**) The receptor binding D-loop; residues making H-bonds which support the disulfide bond. Cys127 is depicted with its reduced conformer, and Pro146 is shown with both endo and exo conformers. ΔP1 is coloured via B-factors in **F**), with low-high B-factors set from blue to red (minimum: 5 $Å^2$, maximum: 22 $Å^2$). The highest B-factors are in the α-β loop and hypervariable loop regions on the side opposite to the RBD, with some dynamics seen in the RBD. All images were generated using PyMOL v.2.5.0[426].

of Leu43 (3.08 Å) and Lys44 (3.11 Å), and this conformation is stabilized by both the backbone amide and the side chain of Ser41 performing an H-bond to the carbonyl backbone of Val37 (2.78 Å). This causes a shift in the next residue, where the carbonyl group of Ala42 H-bonds with Gln67 of the crystallographic symmetry mate (3.11 Å) instead of the backbone amide of Ala46; instead, the H-bond is bridged by water 20 (2.79 Å, 3.07 Å, 11.14 $Å^2$ B factor). Typical hydrogen bonding of the main chain resumes at Glu48 until the end of the helix at Glu53, resulting in a stretched and bent helix for one-and-a-half helical turns. Another important residue to note in the α1-C region is Arg35, the only planarity outlier. Using $N_{\eta 1}$ of the side chain, Arg35 strongly hydrogen bonds with Glu27 (2.40 Å), and $N_{\eta 2}$ is connected to the cacodylic acid through H-bonding with water 8 (2.93 Å, 17.52 $Å^2$ B factor), Glu39 (2.69 Å), and water 17 (2.75 Å, 20.59 $Å^2$ B factor). The uncommon planarity of the Arg35 side chain may be due to these interactions and its high B-factor (18.7 $Å^2$ B factor).

3.3.2.2 The α-β Loop Features a Network of Hydrogen Bonds to Maintain its Shape

The ΔP1 structure has a few important features in the α-β loop, namely the parallel β-sheet and the type III β-turn. The parallel β-sheet is not connected to the β-meander with backbone H-bonding and is therefore not considered to be a part of the head domain, however they are kept in place via side chain-main chain H-bonding between Asp90 of the first β-strand of the β-meander and the backbone amide of Ile57 (2.89 Å), as well as H-bonding between Asp90, water 35 (2.61 Å, B factor: 17.76 $Å^2$), and the backbone carbonyl of Lys55 (2.80 Å). Residues on the β-sheet proximal to the head domain Ile57 and Val58 are hydrophobic; Ile57 is buried inward to the space between the α-helix and the head domain to perform hydrophobic interactions with many β-sheet residues, including Val88, Val96 and Leu98, as well as α-helix and α-β loop residues Ala47, Ile51,

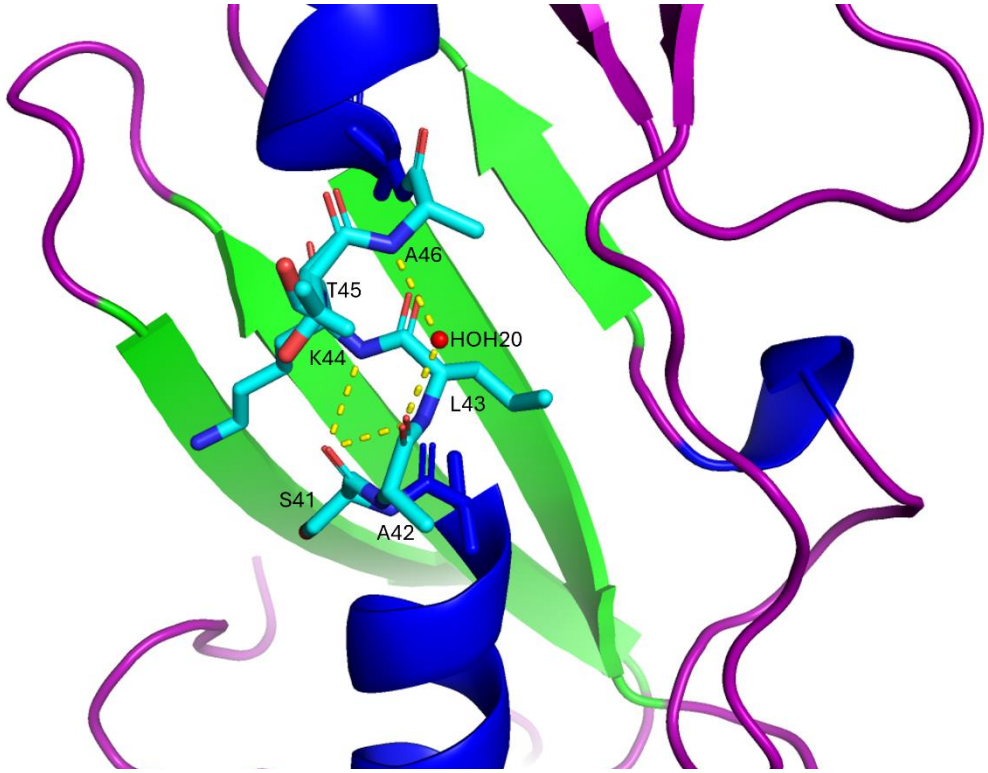

Figure 3.3 The Bend in the α1-C region of ΔP1. The N-terminal α-helix (α1-C) is in blue with relevant residues coloured in cyan and displayed as sticks, the β-sheet is in green, the coil regions are in purple, water molecules are red balls, and hydrogen bonding is shown as yellow dotted lines. Unusual main-chain hydrogen bonding occurs between Ser41 and Leu43, Lys44 of the helix, resulting in a kinked topology. Ala42 retains a stretched helical shape through interaction with water 20 (11.14 $Å^2$ B factor), which H-bonds Ala46. This image was generated using PyMOL v.2.5.0[426].

and Ile70 (**Figure 3.2B**). In the context of an assembled pilus, this parallel β-sheet may be shifted to align its hydrophobic residues with those of the head domain and promote a larger hydrophobic surface that contributes to the interior of the pilus.

The α-β loop makes several important solvent interactions that aid in the stabilization of the observed architecture, such as in the first and second parallel β-strands which are bridged by water 37 (7.88 $Å^2$ B factor) which H-bonds the side chain of Ser59 (2.85 Å) and the side chain of Asp69 (2.77 Å), resulting in the observed pitch between the strands. Ser59 and Asp69 also stabilize the type III β-turn, as the Asp69 side chain H-bonds with the side chain of Lys83 (2.83 Å), which

is H-bonded via its backbone amide to the side chain of Ser59 (2.71 Å), creating an impressive web of powerful non-covalent bonds between these structural moieties (**Figure 3.2C**). The cacodylic acid molecule also makes H-bonds which stabilize the α-β loop and the α1-C helix as its oxygen atoms H-bond with the side chain of Thr73 (2.46 Å), and water 17 (2.94 Å, 20.59 $Å^2$ B factor) which connects to the helix through an H-bond with the side chain of Glu39 (2.75 Å). Thr73 has non-favorable side chain angles (**Figure 3.1**), likely due to H-bonding with the cacodylic acid molecule which stabilizes the observed conformation suitable for crystallization; in the context of the monomeric pilin's typical environment, the bacterial periplasm, this region may behave differently. Another important residue in the type III β-turn is Gln85, which performs H-bonds to the backbone carbonyl of Leu78 (2.98 Å) using its side chain amine, and with the side chain carbonyl Gln85 H-bonds with the backbone amide of Ser75 (2.74 Å) to hold the β-turn and the proximal loop region in the observed conformation. As well, the carbonyl group of the Ser84 backbone connects the β-turn to the first β-sheet of the head domain through an H-bond with the side chain carbonyl of Gln87 (2.86 Å); the β-turn is a central point of contact for non-covalently fixing the α-β loop in its observed conformation and in proximity to the head domain.

3.3.2.3 The β-Meander and Hypervariable Loops Present Important Stabilizing Residues

The β-meander is amphipathic; half (13/26) of the residues are hydrophobic and mostly occupy the space between the head domain and the α-helix (**Figure 3.2B**), and the other half of the residues are polar, 5 of which are charged amino acids. These hydrophilic residues face away from the interior of the molecule and either make solvent contacts, contact side chains from the residues of neighbouring β-strands, or contact the symmetry mate. This is expected as they would face the solvated exterior of the assembled pilus.

The hypervariable loops in between the β-strands have some important residues which support their conformation, such as Ser106 which H-bonds the carbonyl of Gly103 with its backbone N (2.92 Å) to hold the β_2-β_3 loop in its conformation, and its side chain hydroxyl performs a H-bond to the main chain carbonyl of Leu77 (2.87 Å) to bind it with the neighbouring α-β loop. At the end of this loop, the side chain of Lys111 has higher dynamicity based on a side chain B-factor of 17.5 Å, providing flexibility prior to the third β-strand of the β-meander. Lys104 is part of the β_2-β_3 hypervariable loop and displays unfavorable Ramachandran angles (**Figure 3.1**) due to its H-bonds with the carbonyl backbone of Gly80 (2.85 Å) and the symmetry mate Ser148 on the terminal carboxyl (2.86 Å), which stabilizes the C-terminal residue of the molecule in the context of this crystal conformation. Arg119, at the beginning of the β_3-β_4 hypervariable loop is the only residue with a bond angle outlier, which is present in the N_ε-C_ζ-NH_2 covalent bonds of the guanidinium group. This is likely caused by multiple hydrogen bonds that stabilize the atypical conformation; N_ε H-bonds the side chain of Glu48 (2.88 Å), $N_{\eta 1}$ bonds the other oxygen of Glu48 (2.84 Å) and water 30 (2.90 Å, 21.03 $Å^2$ B factor), while $N_{\eta 2}$ H-bonds Gly94 (2.90 Å) and water 31 (2.89 Å, 24.31 $Å^2$ B factor). This sequestration provides Arg119 with a B-factor 10 $Å^2$ lower than other arginine residues in this protein (8.4 $Å^2$). Asp122 in the β_3-β_4 hypervariable loop also makes important contacts to maintain the conformation of this region, including an H-bond with its side chain to one of the conformers of Lys120 (2.90 Å) and H-bonds with the neighbouring backbone amides of Gly123 (3.10 Å) and Val124 (2.98 Å). Gly123 H-bonds with its backbone carbonyl to water 33 (2.72 Å, 15.91 $Å^2$ B factor), which along with the side chain of Trp125 H-bonds to the side chain of Glu48 (2.73 Å, 2.84 Å for Trp125) to bring the β_3-β_4 loop into contact with the α-helix (**Figure 3.2D**). Trp125 has been shown to be one of few highly conserved residues in the C-terminal half of *P. aeruginosa* pilins and may influence the bend of the α-helix, as Glu48

is the first residue to resume canonical α-helical backbone H-bonding geometry after the bend in the α1-C region.

3.3.2.4 The Structure of the D-loop and the B-factors of ΔP1 Support the Region's Role in Pilus Assembly and Binding

The D-loop is a conserved structural motif in *P. aeruginosa* pilins; however, the shape of the region differs due to changes in primary sequence. The conserved disulfide bond is formed by Cys127 in the fourth β-sheet and Cys145, however Cys127 is observed to have electron density representative of an unbonded cysteine (~20% occupancy), which is likely due to radiation-induced breaking of the disulfide bond[427]. There are some neighbouring residues that stabilize the main chain conformation near the disulfide; the backbone amide of Ile129 H-bonds with the carbonyl backbone of Pro146 (3.03 Å), and a proline from the loop immediately preceding the disulfide, Pro142, H-bonds to the main chain amide of Cys145 using its backbone carbonyl (3.02 Å) (**Figure 3.2E**). Prolines are expected to be rigid, hence the perceived conformational stabilization presented by these hydrogen bonds[428]. However, despite low side chain B-factors, Pro146 (10.1 $Å^2$) is observed in an endo-exo conformational flip at 50:50 ratios. Other important prolines in this region include Pro133 which induces the turn from the loop extending out of the final β-strand, and Pro138, which provides the kink in the type III β-turn observed in the D-loop region. The residues which make up this β-turn have the highest B-factors of this moiety: Lys137, Pro138 and Asn139 have main chain B-factors of 14.1 $Å^2$, 14.8 $Å^2$ and 14.8 $Å^2$, respectively. This region may be flexible to allow for broad-substrate recognition in the receptor binding process as it is the tip of the RBD, or it may simply be dynamic to allow for conformational changes during the oligomerization process.

In looking at which regions have the highest B-factors, ΔP1 is observed to have more thermal movement in the loop regions on the side of the protein opposite to the receptor binding D-loop when compared to ΔK122 and Δ1244 (**Figure 3.2F**). The loop which connects the two parallel β-strands in the α-β loop region consisting of Pro63-Thr66 has the highest B-factors of the structural model (other than the N-terminal Ile25). Residue Asp65 has a main chain B-factor of 22.8 $Å^2$, and residue Thr66 has a poor RSRZ score of 2.6%. The side chain of Asp65 models in a non-favorable conformation, possibly due to the smeared electron density and the high B factors of the sidechain (**Figure 3.1**). Statistical outliers in both residues are likely due to the high thermal movement in this area as the region requires flexibility to be moved during pilus assembly. Another region that has high B-factors is the β_1-β_2 loop from Asp90-Thr95, which all have main chain B-factors above 10 $Å^2$. Residues Asn91 and Lys92 have particularly high B-factors for this region (17.5 $Å^2$ and 18.9 $Å^2$, respectively), which is especially interesting as the residue which differs between P1 and 1244 is Lys92 (Gln in 1244), indicating this flexible region may be of importance in the binding specificity and interbacterial recognition of *P. aeruginosa* using group Ia pili[163]. Despite the homology in structure between the T4P of *P. aeruginosa*, different residues play discrete roles in dictating how the α-helix orients with the β-meander of the head domain, and of the important conformations of the connecting loop regions that provide the observed functional diversity in pili from different strains.

3.3.3 Insights from Molecular Replacement Solutions

When solving the ΔP1 structure using Phaser, an optimal model was desired to produce an initial map with a high refined log-likelihood gain (LLG) score, a translational function Z (TFZ)

score above 8.0, and a R_{free} value below 0.50[409]. Initial attempts (circa 2003) at MR and experimental phasing for the ΔP1 structure were unsuccessful, and we were only *a posterori* aware of the structure of Δ1244 following refinement and deposition of the current ΔP1 structure. Therefore, AF and RF models were generated using ColabFold (v.1.5.2) and the Robetta server, respectively, and employed as search models in Phaser[329,333]. This allowed for the predictive modelling of the residues remaining from the cleaved tag (ISEF) and the differing residue between ΔP1 and Δ1244 (K92Q), as well as a comparison between algorithms for producing models suitable for MR by Phaser. Initial AF and RF models were inadequate for MR; Phaser could not find a valid solution using the RF model and the values for the solution resulting from MR using the AF model were poor (**Table 3.3**). Using the "process predicted models" tool in CCP4, suitable MR models were generated based on the regions determined to be in proximity based on the pLDDT score from the ColabFold output (**Figure 3.4A**). Both AF and RF models were separated into 5 domains to be used as separate search models, consisting of domain 1: Ser14-Ala18, domain 2: Ser2-Val13, Gly79-Val90, Ile105-Ser124, domain 3: Leu19-Phe48, domain 4: Thr49-Val64, Val75-Leu78, and domain 5: Thr65-Leu74, Thr92-Lys104, where Ile1 and Ile 91were removed from both search models and Ser2-Glu3 were removed only in the RF model. These processed search models provided solutions via Phaser with improved scores to their unprocessed counterparts, however the solution procured from the processed RF model was less reliable than the unprocessed AF model. Further, when the Δ1244 structure was used as a MR search model, Phaser provided a successful result with a R_{free} of 0.482, however the processed AF model was the optimal search model providing an output solution with a R_{free} of 0.453.

Comparing the refined ΔP1 structure to the MR search models, the trend in Phaser's output parameters is reflected in comparing the main chain RMSD of search models aligned with the ΔP1

structure; the RF model has a high RMSD of 7.041 Å, likely due to the hydrogen atoms present in the model that are not resolved in the electron density of the ΔP1 structure. The alignment of ΔP1 with AF and Δ1244 models are very similar as they have RMSD values of 0.513 Å and 0.616 Å, respectively (**Figure 3.4B**). Despite the higher RMSD of the Δ1244 model, there are some areas where the structure of ΔP1 more closely matches the structure of Δ1244, most notably in the hypervariable loop near the single non-homologous residue Lys92 (**Figure 3.4C**). This region has high B-factors in the ΔP1 structure, which may be why the AF predictive model wasn't successful in modelling the placement of the side chain. The similar positioning of the residues in ΔP1 and Δ1244 is interesting as there are no observable intermolecular hydrogen bonds holding this region of the loop in position, suggesting a role in pilus differentiation/recognition between different strains of *P. aeruginosa*. Notably, when the AF model was processed via the "process predicted model" functionality in CCP4i2, the solution was improved compared to using the full-length model for MR. The hypervariable loop region containing Lys92 may have needed to be processed to separate the search model into distance-related chain ensembles, allowing Phaser to more effectively use this local region as a search model, thus demonstrating the effectiveness and prudence of processing AI predicted models to produce superior MR search models.

Table 3.3 Phaser MR Statistics of Solutions for the ΔP1 Pilin Structure when using various search models. See **Section A.5.1** for more information on parameters for MR solutions.

	Refined LLG	R_{free}	TFZ	Main Chain RMSD (Å) ‡
AF Full	347.6	0.507	10.9	0.513
AF Processed†	1145.6	0.453	19.0	
RF Full	Failed			7.041
RF Processed	112.1	0.565	6.2	
Δ1224	593.7	0.482	16.2	0.616

†Processed models refer to the output from the process predicted model tool from CCP4i2[411].

‡Root mean square deviation distance between main chain atoms of the MR search model when aligned with the refined ΔP1 structure using the PyMOL v.3.0.0 align command[426].

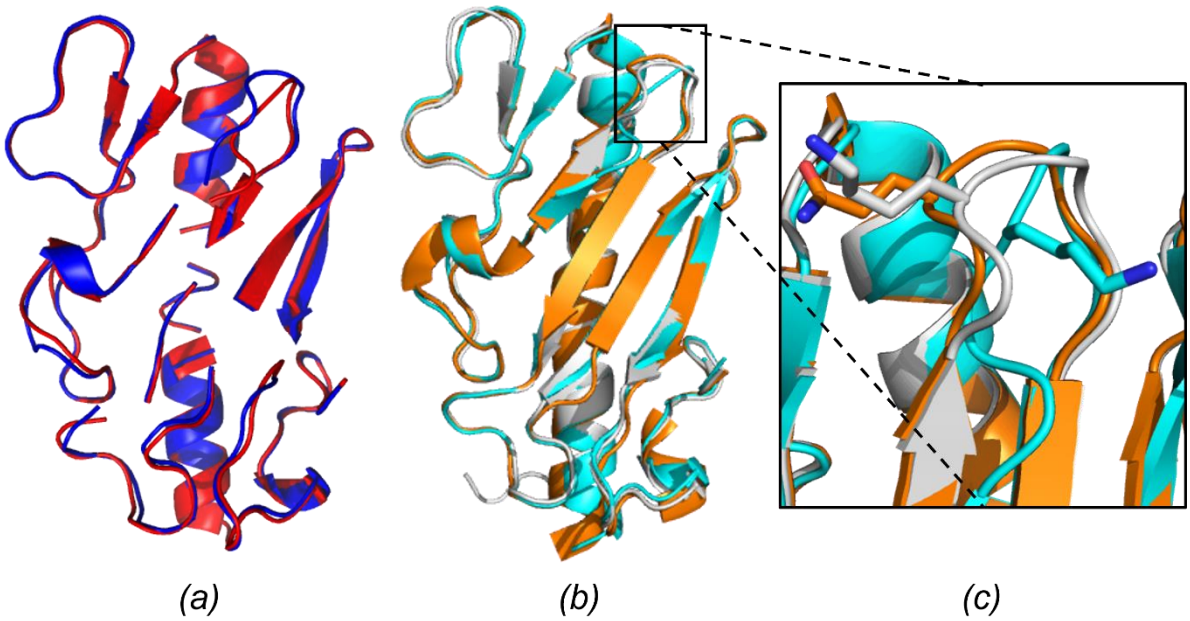

Figure 3.4 Processed AlphaFold and RoseTTAFold Models of ΔP1. **A**) AF and RF models of ΔP1 processed by the Process Predicted Model tool in CCP4i2, used as search models for Phaser molecular replacement. The processed AF model was generated using ColabFold (v.1.5.2) and is coloured red while the processed RF model was generated using the Robetta server and is coloured in blue[329,333]. **B**) Superimposition of the unprocessed AF ΔP1 model (cyan) and the Δ1244 structure (orange) (PDB ID: 6BBK) with the ΔP1 structure (white). The region with the highest RMSD between the models is highlighted in **C**); the differing side chain of residue 92 is shown and demonstrates the incorrect prediction of the positioning of Lys92 by AF. All images were generated using PyMOL v.2.5.0[426].

3.3.4 Binding Surface Analysis of ΔP1 Presented Similarities to Pilins from Different Species

PISA analysis of the ΔP1 structure predicted that all potential interfaces have no role in complex formation and are a result of crystal packing only. The main crystal packing interface consists of 77 atoms (8.4% of the protein atoms) from 29 residues (23.4% of total) of 112 surface accessible residues; this interface makes up 729 Å (10.9%) of the total solvent accessible area, 6681 Å (**Figure 3.5**). The residues contributing to the highest-scoring interface include those on the solvent-exposed side of the α-helix, parts of the α-β loop and the second hypervariable loop, the first β-strand of the β-meander, and the D-loop. There are 12 hydrogen bonds occurring directly between residues of crystallographic neighbors, and no salt bridges or intermolecular disulfides are present for this structure, however, as was mentioned water molecules are seen with low B-factors serving as a bridge for the other solvent exposed residues to make hydrogen bonds that maintain the interface.

Although there is no observed interface which contributes to dimerization of ΔP1 as was seen in other pilins such as ΔK122, PISA analysis is useful for finding T4P with structural homology based on similarities in their interfaces. The proteins with the highest Q-Scores (a structure alignment score which ranks structures from 0-1 based on the similarity of their interfaces, considering both the number of structurally aligned amino acid residues and RMSD of C_α atoms at best structure superposition[404]) were the crystallographic interfaces of ΔK122-4 triclinic and monoclinic crystal structures (Q-Scores: 0.436 and 0.428, respectively)188, the truncated minor pilin PilV pilin from *N. meningitidis* strain C8013 (Q-Score: 0.425) (PDB ID 5V0M)[181], the full length PAK pilus solved using cryo-EM (Q-Score: 0.379) (PDB ID 5VXY)[165], and the truncated pilus from *Shewanella oneidensis* strain MR-1 (Q-Score: 0.372) (PDB ID 4D40)[429]. These identified pilin proteins were used for comparative analyses.

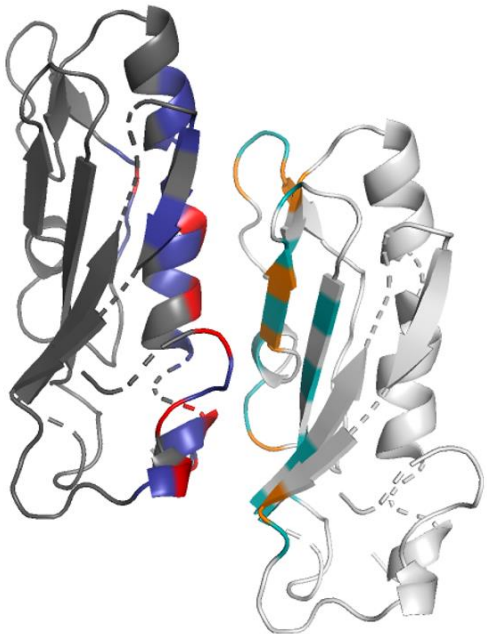

Figure 3.5 The Crystallographic Interface of ΔP1, specifically between two symmetry related molecules. The first monomer in grey shows interfacing residues in navy blue, with those making hydrogen bonds in red, while the second monomer is white with interfacing residues shown in cyan and those making hydrogen bonds in orange. Figure made in PyMOL v.2.5.0 from data obtained via PISA[404,426].

3.3.5 Sequence Alignment is not Indicative of Structural Alignment in Type IV Pilins

3.3.5.1 Clustal Omega Sequence Alignment Percentages Reveals Comparable T4aP Interspecies and Inter-strain Homology

Clustal Omega alignment of sequences of T4aP from *P. aeruginosa* (P1, K122, 1244, PAK, 110594), *N. gonorrhoeae* strain C30 (Ng_C30), *N. meningitidis* strain C8013 (Nm_C8013), *S. oneidensis* strain MR-1 (So_MR-1), and the T4bP from *V. cholerae* strain RT4236 (Vc_RT4236) was performed to compare amino acid sequences of pilins with similar structures and interfaces. The alignments start from their α1-N domain as the signal sequence is cleaved prior to pilus

assembly and is not present in any of the high-resolution structures of these T4P; an additional alignment was performed using the sequences of the head domain of each T4P. A sequence alignment matrix for comparing percentage of alignment between structurally related T4P demonstrates that the T4P from *P. aeruginosa* are unsurprisingly most closely related to each other, with the exception of K122-4 which is most closely related to Ng_C30 (**Table 3.4**). This is also seen when examining a phylogenetic tree created from the alignment results (**Figure 3.6**). In comparing the sequence alignments via the alignment matrix, it is evident that the α1-N helix is the region of highest sequence homology between the pilins, an established trend for the T4P from *P. aeruginosa* which interestingly appears to cross phylogenies. The trends in percentage of sequence alignment are approximate when comparing sequence alignments of full pilins to those without the α1-N helix; for brevity, only the alignments from sequences of the full pilins will be discussed. When comparing group Ia pilins to the other pilins it becomes evident that they share the highest sequence homology with group II pilins (PAK and K122), with the group V pilin PilA from P. aeruginosa strain 110594 (110594) having similar sequence alignment (34.1%) to T4aP from other species (Ng_C30 at 36.1 %)[195]. However, the difference in homology between the group II pilins in this study is lower than the homology between group Ia and II pilins, with an alignment of 34.6% between K122 and PAK, versus an alignment of 43.1% between PAK and P1. K122 also shows highest homology to Ng_C30 at 44.7% alignment versus any of the other *P. aeruginosa* pilins: the highest alignment score of the matrix. In comparing the group V pilin 110594 to the other *P. aeruginosa* pilins, K122 is the most homologous at 40.9% alignment while group Ia pilins share 34.1% alignment, which is slightly higher than the alignment of So_MR-1 at 30.3%. This demonstrates that although group I and group II pilins are of higher homology to each

other based on sequence alignment, the interspecies homology of T4aP is comparable to the inter-strain homology of T4aP from *P. aeruginosa*[430,431].

Table 3.4 Sequence Alignment Matrix of T4P, displaying primary sequence alignment of similar T4P as derived from Clustal Omega.[†‡]

	Group Ia	PAK	110594	K122	Ng_C30	Nm_C8013	So_MR-1	Vc_RT4236
Group Ia	-	43.1	34.1	38.7	36.1	25.5	32.0	14.4
PAK	28.7	-	37.3	34.6	31.8	23.8	29.0	15.8
110594	19.6	21.6	-	40.9	36.6	29.4	30.3	19.2
K122	24.3	26.4	29.3	-	44.7	27.3	28.6	14.7
Ng_C30	21.0	20.2	23.9	37.8	-	27.2	26.5	17.1
Nm_C8013	17.0	11.1	12.0	6.3	9.8	-	31.0	16.2
So_MR-1	14.6	2.3	13.0	15.4	14.3	18.8	-	21.6
Vc_RT4236	9.9	14.3	19.6	14.7	17.1	13.3	27.4	-

[†]Group Ia represents both ΔP1 and Δ1244, which are 99.3% identical; Ng_C30 is the major pilin PilE from *N. gonorrhoeae* strain C30; Nm_C8013 is the minor pilin PilV from *N. meningitidis* strain C8013; So_MR_1 is the major pilin PilD from *S. oneidensis* strain MR-1; Vc_RT4236 is the T4bP from *V. cholerae* strain RT4236.

[‡]Sequence similarities for alignments of full-length pilins (minus signal peptides), and solely the head domain are displayed in the upper right half and lower left half of the matrix, respectively.

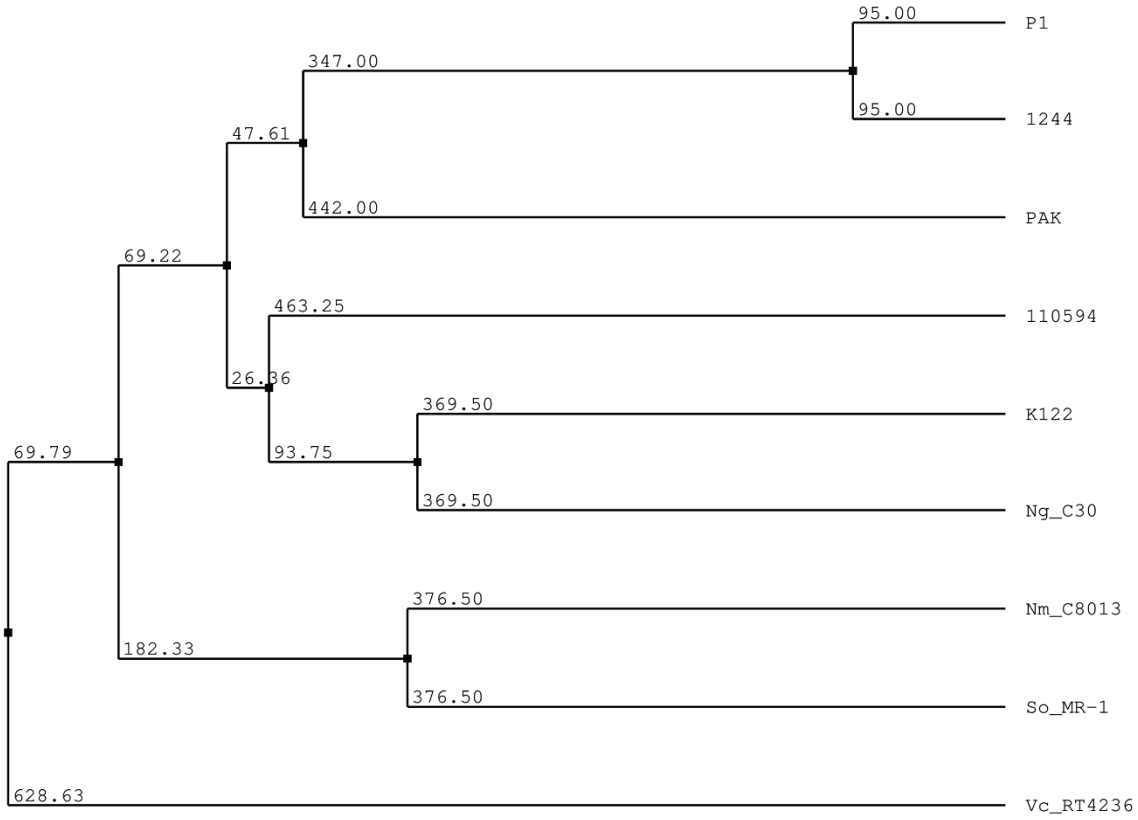

Figure 3.6 Phylogenetic Tree Displaying Relatedness of T4P, based on Clustal Omega alignment, scored via average distance using BLOSUM62. Sequence similarities shown were from alignments of full-length pilins (minus signal peptides). Ng_C30 is the major pilin PilE from *N. gonorrhoeae* strain C30; Nm_C8013 is the minor pilin PilV from *N. meningitidis* strain C8013; So_MR_1 is the major pilin PilD from *S. oneidensis* strain MR-1; Vc_RT4236 is the T4bP from *V. cholerae* strain RT4236. Image was created using Jalview[406,420,421].

3.3.5.2 Homologous Sidechain Interactions of T4aP as Determined by Sequence Alignment

The sequence alignment and conservation of the T4P determined with Clustal Omega shows expected results; the α1-N region has high sequence conservation, and the areas in which the α1-C sequence differs (such as Arg35 and Glu39 for P1 and 1244, Ser35 and Ser39 for PAK, Glu35 and Leu39 for the others) are simply residue swapped for the side chain interactions that

occur between the α-helix and the α-β loop (Ser75 for P1 and 1244, Thr84 in PAK, Ser81 in 110594, Lys80 in K122, and Glu80 in Ng_C30) such that an intermolecular H-bond holding these regions is maintained (**Figure 3.7**). Lys44 is a highly conserved residue in the α1-C region as it appears to be an important residue for maintaining head domain pilin-pilin interactions in the assembled pilus[165,432,433]. Other regions with moderate sequence alignment include the residues in the middle of the α-β loop, the last 3 strands of the β-sheet, the β_2-β_3 and β_3-β_4 hypervariable loops, and a few residues of the D-loop. The only areas with no sequence alignment for a length greater than 8 residues seen in P1 and 1244, compared to the other pilins, are the first part and last part of the α-β loop from residues Ile51-Ala61 and residues Asp79-Gln87. These regions include some important residues for maintaining the shape of the ΔP1 pilin highlighted in **Section 3.3.2**, including the hydrophobic interactions of the α-β loop (**Figure 3.2B**) and the hydrogen bonding network seen in (**Figure 3.2C**). Homologous residues in the β-sheets include Thr101 and Gly112, which in the case of ΔP1, creates a bend in the β-sheet between the second and third β-strands; Thr101's backbone carbonyl H-bonds to the backbone amide Asp79 (2.86 Å), connecting the bottom of the α-β loop to the second β-sheet and preventing a proper H-bond between Thr101 and Val114. This shift disrupts the standard H-bonding geometry of the β-sheet, and Gly112 H-bonds with its backbone amide to the backbone carbonyl of Leu102 at a non-planar angle (2.85 Å). Trp125 at the beginning of the 4th β-strand is also conserved and is shown H-bonding to residue Glu48 (**Figure 3.2D**). There is a conserved polar residue capable of H-bonding in other T4P, indicating that this interaction is likely important to maintain the proximity between the α-helix and the β-meander in these other pilins, and it may also influence the bend of the α-helix. The

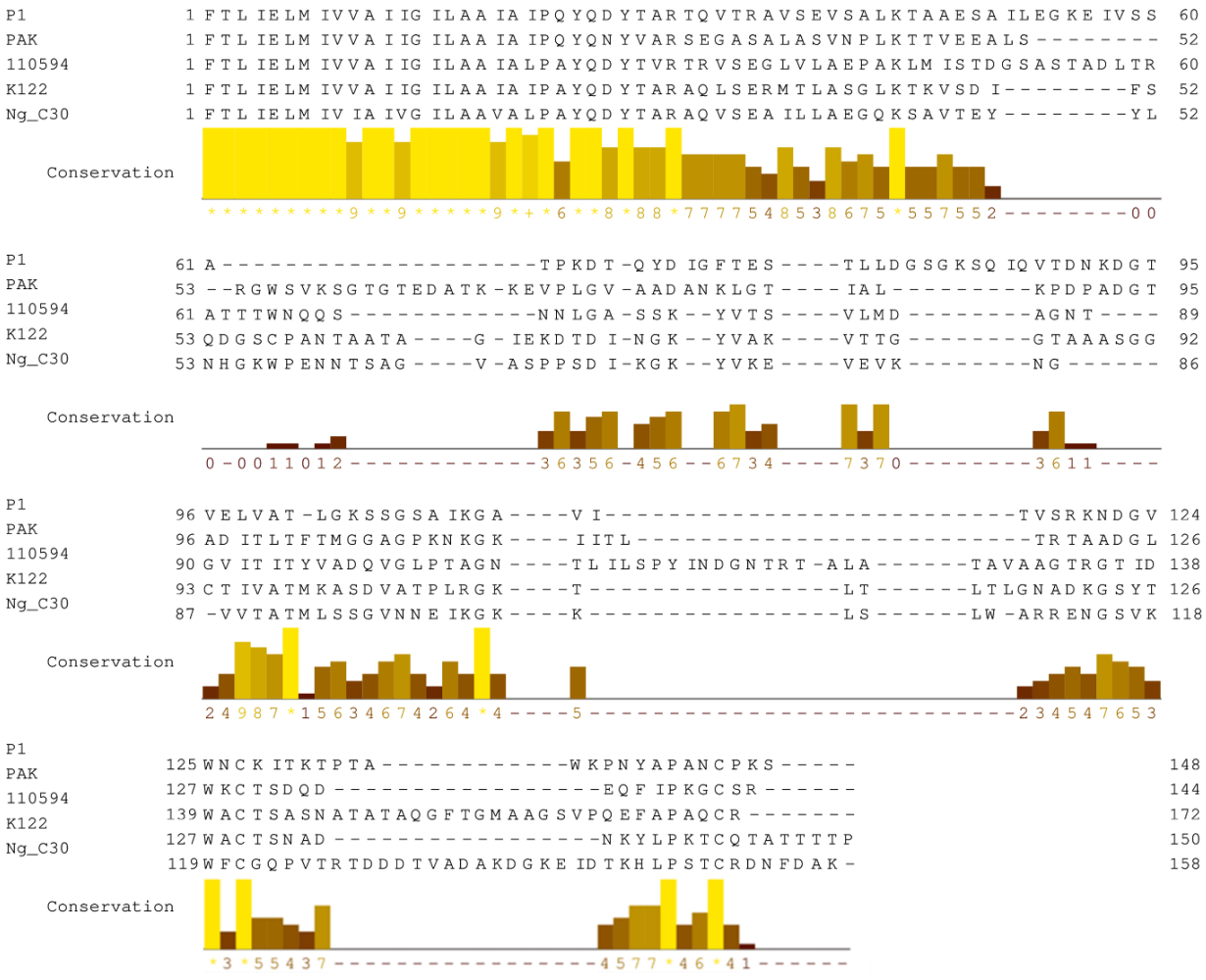

Figure 3.7 Clustal Omega Multiple Sequence Alignment of T4P, specifically with solved structures similar to ΔP1. All sequences have their N-terminal signal sequence removed. Per residue conservation scores based on homology are displayed with bar graphs underneath. Ng_C30 is the major pilin PilE from *N. gonorrhoeae* strain C30. Nm_C8013, So_MR-1, and Vc_RT4236 sequences were omitted from the figure to better emphasize the sequence conservation of certain regions of the sequence-related pilins. The image was generated with Jalview[420].

residues after the last β-strand that start the D-loop display low sequence conservation as the only homologous residues in the D-loop are Cys127, Pro142, and Cys145; they maintain the main chain conformation that is essential for T4P binding[190].

3.3.5.3 ProSMART Structural Alignment Statistics Suggest the Need for Structural Classification of P. aeruginosa T4aP

Analysis of structural alignment was done via ProSMART to ensure conformational flexibility was considered when comparing closely related T4P structures[405]. ProSMART uses the Procrustes algorithm to align local structural environments via backbone fragments of n-residue lengths from the compared high resolution structures[405,434]. The numerical results of the ProSMART alignments, which feature scores such as the average Procrustes score, the average Flexible score, the global RMSD, and the best fragment score, are presented in **Table 3.5**. Sequence identity values shown were from the ProSMART comparison of the PDB files so they contain features such as additional residues from protein tags, while sequence alignment values were derived from Clustal and consider the structures from the start of the α1-N helix.

In aligning the AF predictive model to the ΔP1 structure, the average Procrustes RMSD, the average Flexible score, the global RMSD and the Best Fragment score were all higher than when ΔP1 is aligned with the Δ1244 structure. Knowing that there are several well-refined pilin structures, likely employed in the training sets of AI model prediction suites including AF and RF, and that the best MR search model for Phaser was one processed prior to use, these data suggest that further improvements to prediction algorithms that incorporate such conformational flexibility could lead to better search model predictions.

Table 3.5 Structural (ProSMART) and Sequence (Clustal Omega) Alignment Values Comparing Type IV Pilins.

	AF ΔP1[†]	Δ1244	ΔK122	PAK	ΔPAK	Δ110594	Ng_C30	Nm_C8103	So_MR-1	Vc_RT4236
Residues Aligned (Total Residues)	124 (124)	124 (124)	116 (126)	116 (144)	116 (120)	119 (148)	115 (150)	98 (98)	89 (89)	118 (171)
PDB Sequence Alignment / Full-length (%)[‡]	100.0/ 100.0	96.0/ 99.3	10.3/ 38.7	22.4/ 43.1	25.47/ 43.1	10.1/ 34.1	13.9/ 36.1	10.2/ 25.5	7.87/ 32.0	8.47/ 14.4
Local Alignment RMSD (Å)	0.508	0.347	1.6	1.33	1.21	2.39	1.82	1.46	1.97	2.06
Average Flexible	0.292	0.2	1.24	1.03	1.01	1.62	1.25	1.14	1.48	1.60
Global RMSD (Å)	1.72	1.46	6.71	6.35	3.45	14.70	13.90	4.89	5.93	15.20
Best Fragment Score	0.0859	0.0725	0.302	0.146	0.183	0.252	0.230	0.197	0.157	0.203
PDB ID		6BBK	1QVE	1OQW	1DZO	3JZZ	2HI2	5V23	4D40	IODV
UniProt ID		P18774	P17838	P02973	P02973	Q8KQ36	P02974	A0A9K2KQ72	Q8EII5	P23024

†The UniProt ID and PDB ID of PilA from *P. aeruginosa* strain P1 is P17836 and 8V7P, respectively.

‡PDB sequence alignment values are derived from the PDB files and are compared using ProSMART. Full length sequence alignment values do not consider tags or mutations in the PDB files, only the residues in the UniProt sequence starting from after the signal sequence and are derived from Clustal Omega. AF ΔP1 is the Colabfold generated model of ΔP1.

As expected, those alignments with a low average Procrustes score often had proportionally low average Flexible, global RMSD and best fragment scores. There was an unusual relationship between Clustal Omega sequence alignment and structural alignment to ΔP1; there were many comparisons made in which the sequence alignment was lower than for another pilin but the average scores from ProSMART were relatively improved. This is most apparent when comparing the alignment of 43.1% for ΔK122 to the alignment of 25.5% from Nm_C8013, which has the lowest alignment to ΔP1 of the T4aP sequences compared, however the average Flexible score for Nm_C8013 is 1.14, superior to the scores of 1.24 and 1.62 from *P. aeruginosa* pili ΔK122 and Δ110594, respectively. Although most scores were slightly higher when aligning the PAK pilin to the ΔP1 structure versus the alignment of ΔPAK, as caused by the lack of the α1-N helix in the ΔP1 structure, the best fragment score went down for the full-length PAK pilin; as well, full-length PAK was recognized as having a similar interface to ΔP1 via PISA whereas ΔPAK had a lower Q-Score. This indicates that some regions alter when the hydrophobic α1-N is not present, demonstrating the importance of pilin structures with the α1-N helix. Other than Δ1244, the T4aP most closely resembling ΔP1 of the compared structures is ΔPAK, whereas the least homologous is Δ110594. Pilins from different species such as Ng_C30 and Nm_C8013 have lower scores from ProSMART, thus indicating the possibility for higher structural variability of T4aP between strains of *P. aeruginosa* than the variability between T4aP of different species. As well, the comparison of ΔP1 to the T4bP ΔVc_RT4236 provides further evidence that group V pilins are more similar in structure to T4bP than to group I and II pilins as sequence alignment scores and previous studies indicate[195], however structural alignment scores for ΔVc_RT4236 and Δ110594 to ΔP1 are also similar despite their large differences in sequence alignment to P1.

3.3.6 ProSMART Structural Comparisons Demonstrate that T4aP are Differentiated Mainly by their α-β Loop

3.3.6.1 Structural Alignment of Type Ia Pilins ΔP1 and Δ1244

The ProSMART alignment of ΔP1 with different T4Ps was performed and output aligned models were colored via the Flexible score (**Figure 3.8**); when colouring residues by this method, regions can be highlighted as scoring high if they are structurally distinct as the conformation is rigid, and if the backbone RMSD differences are merely due to dynamic regions causing conformational changes the score remains low[405]. In viewing the alignment between the ΔP1 and Δ1244 structures, the structural homology is extremely high, however the single residue difference between these pili affects the conformation of the α-β loop rather than influencing the main chain conformation of the differing residues in the first hypervariable loop where residue 92 resides (**Figure 3.8A**). Asp65-Thr73 feature residues with the highest side chain B-factors of ΔP1; they may merely be different due to differences in crystal contacts holding these regions in place, however the Flexible parameter in ProSMART indicates that this region has a discretely different conformation in ΔP1 compared to Δ1244.

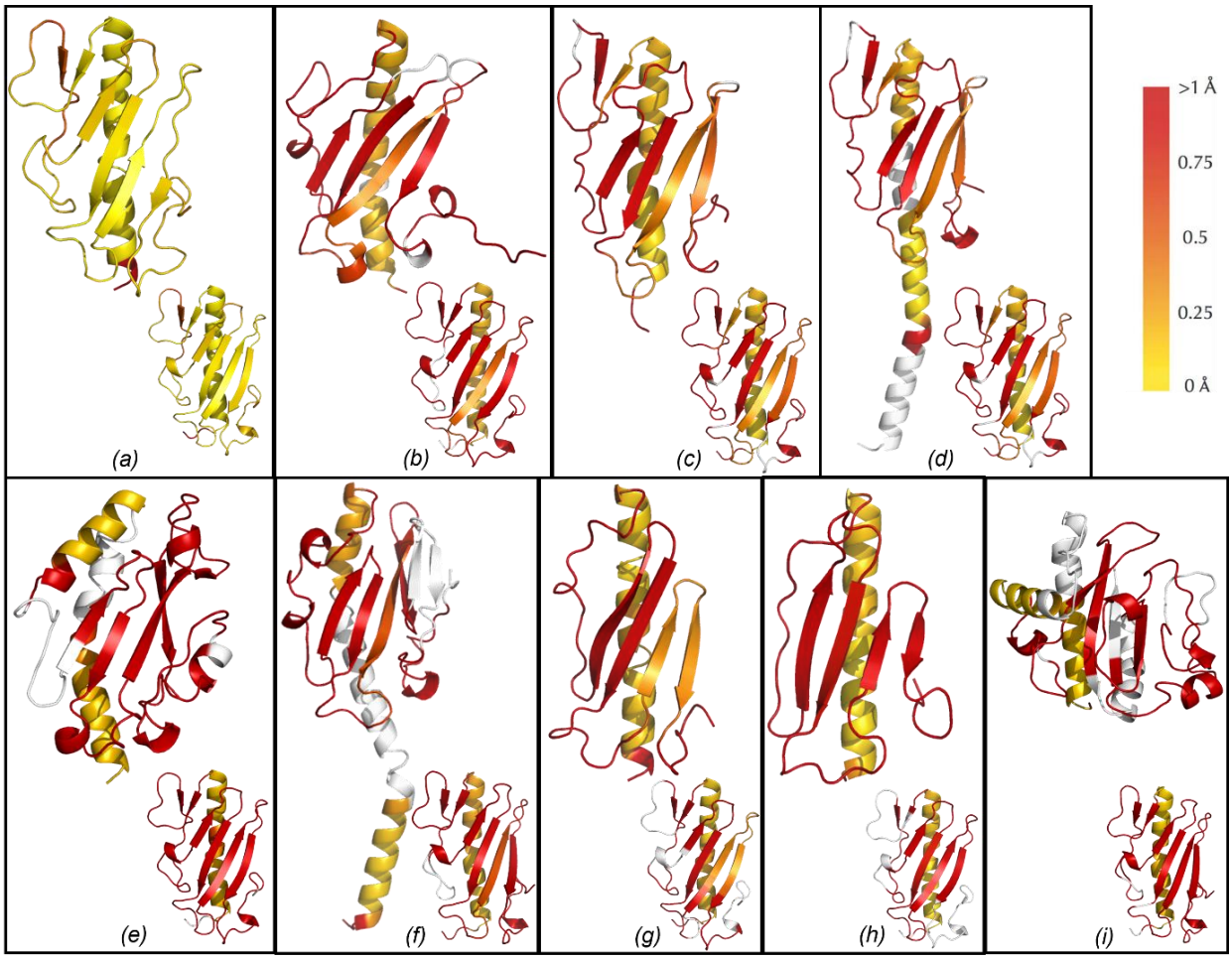

Figure 3.8 ProSMART Structural Alignment of ΔP1 with Similar T4P, coloured by the Average Flexible Score as per the legend in the top right of the figure based on the RMSD between backbone atoms of the aligned structures[405]. Regions coloured white are not present in the aligned protein and therefore are not comparable; the ΔP1 structure is to the bottom right of each compared pilin and is coloured by flexible scoring for the respective alignment. Structures aligned with ΔP1 are **A**) Δ1244 (PDB: 6BBK) **B**) ΔK122 (PDB: 1QVE) **C**) Δ110594 (PDB: 3JZZ) **D**) ΔPAK (PDB: 1DZO) **E**) PAK (PDB: 1OQW) **F**) Ng_C30 (PDB: 2HI2) **G**) So_MR-1 (PDB: 4D40) **H**) Nm_C8013 (PDB: 5V23) **I**) Vc_RT4236 (PDB: 1OQV).

3.3.6.2 Alignments Between ΔP1 and Group II Pilins

There are many structural distinctions observed in comparing ΔP1 and *P. aeruginosa* pilins from other groups. The α1-C helix is the only region that shares good alignment based on the Flexible score when aligning ΔP1 and group II pilins ΔK122 and ΔP1 (**Figure 3.8B**, **C**). In comparing ΔP1 and ΔK122, the third β-strands of the two pilins are moderately aligned from

Ile110-Arg118 in ΔP1, otherwise all regions have poor alignment. Some fragments in ΔK122 are recognized as not present in ΔP1, such as regions of the ΔK122 β_1-β_2 and β_3-β_4 hypervariable loops, Ala87-Ser90 and Ala120-Asp121. These regions have residues with some of the highest B-factors in ΔP1 and are known to be important for T4P differentiation; their mobility suggests their importance in intermolecular interactions. There are also fragments in ΔK122 that are not present in ΔP1; the receptor binding D-loop of ΔK122 features an extra type III β-turn from residues Lys136-Tyr137.

The alignment of ΔP1 and ΔPAK provided the best ProSMART scores for any pilin other than the two group Ia pilins; structural features look highly similar, however the Flexible scoring shows that there are major differences in the conformation of the pilins starting after the first β-strand of the α-β loop region to the second hypervariable loop from Ser59-Gly106. The α-β loops are also distinguished between the two pilins via residues in ΔPAK not present in ΔP1, Asp65-Thr67, and a residue in ΔP1 not present in ΔPAK that forms the type III β-turn, Asn85 (**Figure 3.8C**). The structures are moderately similar in the third and fourth β-strands, however the D-loop regions of the proteins are dissimilar, owing to a region in ΔP1 from Thr132-Trp136 not present in ΔPAK.

Comparing full length PAK to ΔP1, the head domain aligns in an identical fashion as ΔPAK, however the α-helix no longer aligns with the same residues; there is a gap in the α1-C helix from residues Ala34-Pro42 in PAK that is not structurally homologous in ΔP1 despite being recognized as highly aligned in ΔPAK (**Figure 3.8D**). Instead, a region in the α1-N helix of PAK from Ile21-Ser31 is predicted to be able to conformationally align with the α1-C helix of ΔP1. This provides evidence that structures of pilins with the α1-N region may provide more accurate comparisons of the α-helical regions of pilins, and as mentioned in **Section 3.3.5**, PISA may

recognize different aspects of the interfaces present in the structures with α1-N helices, however they do not often make a difference in the structure of the rigid head domain as shown via these alignments and likely are not essential for relating *in vivo* functions from structures of T4aPs.

3.3.6.3 Comparing Structures of ΔP1 and a Group V Pilin

Group Ia and group V pilins are structurally distinct, however when aligning ΔP1 and Δ110594 there are no regions of ΔP1 that are not recognized as fragments in Δ110594 based on Flexible scoring, the group V pilin merely has extra regions which differentiate it (**Figure 3.8E**). There is an additional region in the α1-C helix of the Δ110594 from Glu41-Ala53 not present in ΔP1, however the α-helix that extends to the α-β loop of Δ110594 from Ser54-Thr63 corresponds well to the α-helix of ΔP1 from Thr45-Glu53. Unlike the alignments to the other *P. aeruginosa* pilins, ΔP1 does not align well to Δ110594 throughout the whole α1-C helix, the extended helical turn which causes the kink in ΔP1 is not well aligned from Ser41-Lys44. All other regions in ΔP1 align poorly with Δ110594, and regions that are not present in ΔP1 that are specific to Δ110594 include Ser69-Ser81 which constitutes the α-β loop and part of the first β-strand, and a few residues in the D-region from Thr150-Gln152. Despite the complex secondary structural regions in the hypervariable loops and the differences in the β-sheet topology of Δ110594, ProSMART recognizes that these regions hold some structural similarity despite displaying poor sequence and structural alignment overall.

3.3.6.4 Structural Alignments of ΔP1 and Pilins from Other Species

ProSMART alignment of ΔP1 with T4aP from other species shows similarities in alignment to the group II pilins viewed. Of the three observed, Ng_C30 has the worst structural alignment scores despite having optimal sequence alignment to ΔP1, likely due to the additional regions in the head domain of Ng_C30 and the smaller size of the head domain of the other pilins ΔNm_C8013 and ΔSo_MR-1. The structure of Ng_C30 aligns with ΔP1 in a similar fashion as PAK, where the α1-N region of Ng_C30 from Glu5-Leu16 that isn't present in the ΔP1 structure is instead aligned with the initial α1-C residues, however residues in the α1-C of Ng_C30 from Val19–Gly42 have no counterpart in ΔP1 (**Figure 3.8F**), once again demonstrating the utility of structures with the α1-N region. The next part of the α1-C from Gln43-Asn53 is only partially aligned however, and the only other region with moderate structural alignment is the third β-strand from Ile101-Ala110, otherwise all regions have poor alignment to ΔP1. There is one other motif that makes up auxiliary secondary structural elements in the D-loop of Ng_C30 from Val125-Asp138 which appears to be unique to this T4aP and is not present in ΔP1.

ΔNm_C8013 and ΔSo_MR-1 both have similar crystallographic interfaces based on PISA, however ΔNm_C8013 shares higher structural similarity despite the higher sequence alignment of ΔSo_MR-1. In aligning with ΔP1, the α1-C of ΔNm_C8013 and ΔSo_MR-1 are highly homologous up until the residues that cause the kink in α-helix from residues Ser41-Glu53; ΔP1 is somewhat conformationally distinct from the other two pilins in this region, which appear to have canonically straight α1-C helices (**Figure 3.8G & H**). Some regions in ΔP1 stand out as not being present in ΔNm_C8013 and ΔSo_MR-1; large portions of the α-β loop, from Ser59-Asp65 and Gly82-Gln87 for ΔNm_C8013, from Ile57-Thr66 and Gly80-Ser84 for ΔSo_MR-1, and the D-loop region from Asn139-Ser148 for ΔNm_C8013 and the entire D-loop from Lys131-Ser148

for ΔSo_MR-1. The difference in size of these absent regions when aligning to ΔP1 likely contributes to the lower structural alignment to ΔSo_MR-1 than to ΔNm_C8013 despite the opposite being true for their respective sequence alignments. The third and fourth β-strands of ΔNm_C8013 align moderately well with ΔP1 from Thr116-Lys128, however ΔSo_MR-1 has poor structural alignment with ΔP1 for the remainder of the protein. Structural similarity of the third and fourth β-strands appears to be a trend of pilins that have improved scores, as global RMSD of the alignment between ΔP1 and ΔNm_C8013 is third lowest of those compared; similar alignments occur with ΔK122, PAK, and Ng_CR30 (**Figure 3.8B, D, & F**). Interestingly, ΔNm_C8013 and ΔSo_MR-1 do not have a receptor-binding D-loop; the ΔSo_MR-1 pilin is thought to be important to the species for extracellular electron transfer pathways and is not primarily used for cellular adhesion, however the minor pilin ΔNm_C8013 has not been well characterized; based on these structural comparisons and information on homologous minor pili it likely has alternative functions as well.

The structural comparison of ΔP1 to the T4bP ΔVc_RT4236 aligns similarly to the manner that ΔP1 aligns with Δ110594, again further supporting that group V pilins, and by extension group III pilins, share structural similarity to T4bP (**Figure 3.8I**)[195]. ΔP1 varies significantly from ΔVc_RT4236, with only the α1-C helix of ΔP1 scored as aligning well with half of the α1-C of ΔVc_RT4236 from Ser30-Gln44 and a part of the helix in the α-β loop from Ala66-Gly77; the region in between is considered not to be present in ΔP1. The rest of the T4bP is poorly aligned, with a lack of structural homology to ΔP1 in many of the hypervariable loop residues and most of the fourth β-strand. There are only a few residues in ΔP1 that are recognized as not being aligned with fragments from ΔVc_RT4236, and they included the α-β loop residues Leu78-Gly80 and Gly103 on the β_2-β_3 loop, which coincidentally perform intermolecular interactions; although these

are just a few residues of many that are not sequence aligned between ΔP1 and ΔVc_RT4236, the divergent evolution which causes the deletion of interacting residues such as these are an example of how the observed structural diversity in T4Ps arose and continues to evolve.

3.3.6.5 Comparisons of ProSMART Alignments Elucidates the Structurally Unique Regions of ΔP1

A kink in the α1-C helix was described in the structure of ΔP1 (**Figure 3.2A, Figure 3.3**), however in comparing the structures of pilins via ProSMART this bend was observed to vary in pitch in different pilins (**Figure 3.8**). This bend in the α1-C helix is common in T4P yet only the severe kink in the α1-N region is often discussed in literature; it is seen in cryo-EM reconstructions of the PAK pilus and the *N. gonorrhoeae* strain MS11 pilus and therefore is not caused by crystallographic packing interaction[165]. We propose that it allows for cushioning between the two moieties of the head domain as there are bulky hydrophobic side chains on the α1-C helix and the β-meander that populate the interior space between the two domains. This region excludes water, which contributes to the rigid rod-like behaviour of the assembled T4P[32]. The kink is of a shallower bend in group Ia compared to group II pilins, potentially due to the presence of a hydrophilic residue in the group Ia pilins, Glu48, which is one of few non-conserved residues in the α1-C. The group V pilin alignment showed no fragments aligned with the bent region from Ser41-Lys44 in ΔP1, while the minor pilin ΔNm_C8013 and the extracellular electron transfer pathway pilin ΔSo_MR-1 have a straight helix α1-C region and have different roles than the other T4P shown. This kink may serve a role similar to the α1-N helix break conserved in many T4P from residues 15-23, which 'melts' to accommodate the assembly of the pilus by aiding movement of the helix in the hydrophobic interior[165,433]. However, as it has a different kink depending on the pilin and it is not present in the pilins with alternative roles, the bend in the α1-C may provide a mechanical

signal for pilus attachment and/or depolymerization, straightening upon either event as the mechanism for function[163,435].

The architecture of α-β loops appeared to best distinguish pilins from each other; pilins of lower homology appeared to have additional or missing α-β loop regions as was confirmed by Clustal Omega alignment results (**Figure 3.7**), and those close in sequence homology in many instances were seen to have higher Flexible scoring and therefore are conformationally distinct despite appearing similar on first glance. As mentioned, this region in ΔP1 contains residues with the highest B-factors of the protein, however ProSMART recognizes it as being sufficiently unique to not align well with the other α-β loops of pilins. The D-loops of the pilins were also conformationally distinct despite previous studies demonstrating main chain homology, indicating that there is enough rigidity in the regions for each to be distinct and therefore differentiated from each other. Interestingly, the group Ia pilins studied are the only pilins compared which have prolines bordering both cysteines in the D-loop, and Pro146 is performing endo-exo conformational flipping; this may provide the rigidity required for the observed diversity in the main chain conformation but allows for some rotamer movement to accommodate the dynamics required for this region's function (**Figure 3.2E**). This provides further evidence for the D-loop region in aiding the mechanism of receptor binding specificity and recognition, but also adds the α-β loop as an additional distinguishing factor for T4aPs.

3.3.7 Differences in Self-Assembly of Type IV Pilins

Self-assembly of the ΔP1 pilus was assessed in a manner previously performed with the ΔK122-4 pilin[105,171]. Characterizing the oligomerisation of pilins is desirable as the mechanism for

the *in-vitro* oligomerisation of T4Ps remains unknown; understanding pilus self-assembly would aid in elucidating their *in-vivo* assembly to develop vaccines/drugs against *P. aeruginosa* and for the development of pilin-based PNT drug delivery vehicles[170]. The crystallographic interface of ΔP1 is most similar to that of ΔK122 despite the additional dimeric interface of ΔK122-4, both of which are not in the same configuration as the ΔP1 crystallographic interface[105,190]. The crystallographic interface of the ΔK122-4 pilin stacks together the β-meander head groups of the two symmetry-related molecules, while the dimeric interface is from the α-β loop to the bottom of the α1-C helix. Nevertheless, the ΔK122-4 self-oligomerizes into PNTs[105,171,201]. Therefore, an analysis of the self-assembly of ΔP1 was performed via SEC-MALS to separate any oligomers which may form by size, accurately determine their MW, and gain and understanding of the general shape of the eluted species. MALS is necessary as many pilins do not absorb UV_{280nm} light well due to low abundancy of aromatic residues and their small size; this also prevents good staining and visualization of pilins in PAGE gels[171].

When analyzed via SEC-MALS, ΔP1 was observed to be monomeric (**Figure 3.9A**). This is unlike ΔK122-4 which displays a monomer-dimer equilibrium when in the same buffer[105,171]. As well, when oligomerization was attempted in conditions demonstrated to be optimal for ΔK122-4, including the hydrophobic catalyst MPD, ΔP1 was observed to only partially dimerize and does not indicate the presence of any oligomers (**Figure 3.9B**). In overlapping the UV_{280nm} chromatograms the slight increase in dimer peak becomes visible (**Figure 3.9C**). Some useful biophysical parameters were obtained from MALS analysis; the MW of the monomer peak from the unoligomerized ΔP1 sample is calculated at 14.81 kDa (±0.909%) (expected MW 13.03 kDa), and the small shoulder corresponding to the dimer in the oligomerized sample has a MW of 21.93 kDa (±3.738%) (expected MW 26.06 kDa). The discrepancy in size for the monomer can be

attributed to solvent hydration, however the MW for the dimer is likely due to its small proportion of the LS signal causing inaccuracy in the MW estimation (9.1% of mass fraction). The small light scattering (LS) peaks corresponding to the UV_{280nm} signals of the monomer and dimer in both

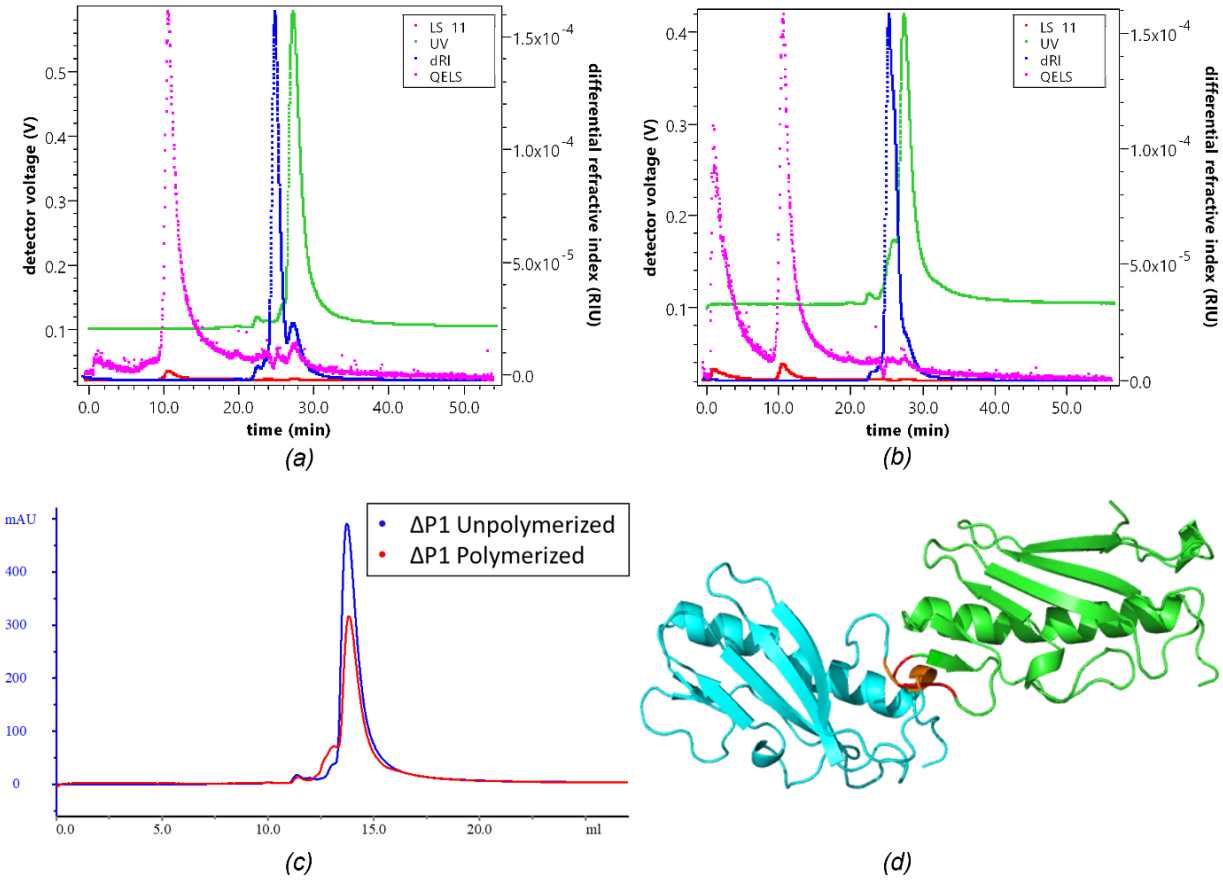

Figure 3.9 Oligomerization Analysis of ΔP1 Using SEC-MALS, and a three-start helical model. MALS chromatograms of **A**) unoligomerized and **B**) oligomerized ΔP1 samples after performing in-line SEC with a Yarra s3000 column (separation range 5 kDa – 700 kDa; Phenomenex) at a flow rate of 0.5 mL/min using an ÄKTA Purifier 10S FPLC system (Cytiva). Images were created using the Astra software suite v.6.0 (Wyatt Technology). **C**) Overlaid UV_{280nm} chromatograms from both experimental runs. **D**) A model for the assembly direction of the pilus using the ΔP1 pilin structure based on a three-start helical assembly[171,202]. Monomers are shown in green with clashing residues in red and cyan with clashing residues in orange; the direction of pilin oligomerization would be from the cyan molecule to the green molecule.

experiments prevented accurate calculations of other biophysical parameters like R_h and R_g. The LS peak with a high quasi-elastic light scattering (QELS) signal seen in both experiments has no UV_{280nm} signal and is considered an anomaly as it lies in the void volume of the SEC column.

To conceptualize why ΔP1 does not oligomerize *in vitro* in the same manner as ΔK122-4, the structure of the ΔP1 pilin was modeled in a three-start helical assembly, the configuration that best describes the way in which pilin monomers assemble into a polymer as originally characterized by cryo-EM experiments on the assembled Ng_C30 pilus (**Figure 3.9D**)[171,202]. Regarding the positioning of the monomers, there are clashes that occur between the α-β loop residues Pro63-Thr66 of one monomer and the D-loop residues Pro138-Ala141 of the other monomer; these regions are not pronounced in ΔK122-4. PISA provided further evidence into why the ΔP1 pilin does not oligomerize in the same fashion as ΔK122-4; an interface for oligomerization cannot be found. However, the region in the α-β loop that is clashing has the highest B-factors for the ΔP1 protein, and one could imagine that it might change conformation in forming oligomers.

3.4 Conclusions and Future Work

The high-resolution structure of ΔP1 allowed for a thorough structural analysis of the protein; the first such characterization of a group I pilin from *P. aeruginosa*. The structure was solved twenty years after diffraction data collection largely due to advancements in structural prediction software and complimentary software which allowed for the creation of an optimal search model for MR. However, AF might not have created such a reliable model if the truncated structure of the truncated PilA from strain 1244 was not uploaded to the PDB[403], reinforcing the

importance of having data uploaded to repositories allowing for more robust training sets for predictive modelling. The predicted ΔP1 model from AF was superior to that of RF; it is suggested that the Robetta server could include an option to provide an output suitable for MR, as even when hydrogen atoms were deleted from the model it would not align well with the solved ΔP1 structure (**Table 3.3**), and although the PyMOL command to remove hydrogens is simple there are minor differences which prevents the output from being a suitable MR search model. Alternatively, the "process predicted model" tool in the CCP4i2 GUI could provide the option to remove hydrogens from a search model.

Using sequence and structural analyses it was found that ΔP1 displays higher homology to some distantly related pili than with ΔK122, which is further supported by the differences in their propensity for dimerization in solution and their ability to detectably oligomerize in the presence of a hydrophobic trigger molecule. PISA analysis results indicate that there is potential for ΔP1 to dimerize which supports the SEC-MALS results, however the dimerization and crystallographic interfaces are likely different, and the ΔP1 dimerization interface may be different from the ΔK122 dimerization interface seen in mass spectrometry experiments[105] (**Figure 3.4, Figure 3.9D**). The hydrophobic trigger molecule MPD allowed for some dimerization of ΔP1, but not oligomerization; it is proposed that the difference in function between the two proteins is due to differences in the structures of their α-β loop and the D-region, especially considering the three-start helical model of ΔP1 shows these two regions clashing. Minor pilins and/or other chaperones may be required to induce the proper fold in both regions for ΔP1 oligomerization; there are several other factors that could influence the *in vitro* oligomerization of ΔP1 such as the presence of particular buffer conditions, hydrophobic trigger molecule, or TfpO glycosylation of the C-terminal serine[189]. These unexplored variables will be investigated in future experiments.

MS experiments on ΔP1 akin to the HDX MS experiments on ΔK122 would be useful for exploring the dynamics of the regions with high B-factors and would allow for a comparison in the conformations achieved by the two pilins[105]. This may elucidate the cause of unsuccessful *in vitro* oligomerization seen for ΔP1, and comparison to ΔK122 could reveal the mechanism behind the phenomenon. HDX MS revealed that the α-helix and the β_2-β_3 loop of ΔK122 displayed lowered dynamics upon dimerization, while a small part of the α-β loop and portions of β strands 1 and 2 were seen to be more conformationally dynamic[105]. As PISA did not identify a dimerization interface for ΔP1 and a dimer was in very low abundance after *in vitro* oligomerization, it is unlikely that sufficient quantities of the ΔP1 dimer could be isolated for these experiments. Instead, HDX MS results for monomeric ΔP1 would be compared with the monomeric ΔK122 data; the relative differences in dynamics of various regions may reveal some insights and confirm ProSMART comparisons. Additionally, performing MALS-SAXS in batch mode with an appropriate stop-flow setup, such as those used in studying fibrin network formation[436,437], has significant potential as a technique for studying the mechanism of the *in vitro* oligomerization reaction of Type IV pili such as ΔK122. Using this method fibrils of ΔK122 can be characterized via MALS and SAXS as they form via addition of MPD, and through aligning the crystal structure of ΔK122 with the low resolution SAXS information a model of T4P oligomerization can be achieved.

Comparing the structure of ΔP1 to a variety of T4Ps provided context for the distinctions between group Ia pilins and the other groups of *P. aeruginosa* pilins, as well as pilins from different species. Sequence and structural comparisons between ΔP1 and ΔPAK identified the proteins as similar in structure, while ΔK122 has more homology to ΔNg_C30 than to either of these *P. aeruginosa* pilins. These findings suggest that developing a nuanced structural grouping of *P.*

aeruginosa pilins to update the current genetic grouping of pilins could be beneficial. The results of the structural comparisons demonstrated that the α-β loop may be of functional importance as it differs dramatically between closely related pilins; the structures of ΔP1 and Δ1244 only differ by one residue and were most structurally distinct in this region (**Figure 3.8A**). All other comparisons identified the α-β loop and the D-loop as having residues that were not present or greatly differed in conformation, while the α-helix was highly conserved in every structure. Targeting the D-loop likely remains the best approach for both vaccine and anti-virulence drug development against T4P from *P. aeruginosa,* however care should be taken to prevent T4P from other bacteria to not be targeted as well to prevent destruction of the lung microbiome[430,431]. High resolution structural information on *P. aeruginosa* T4aP from all groups will help ascertain the structure-function relationship of the diverse pilins to better understand how they bind receptors, how they relay binding signals, and how they self-assemble for formation of PNTs in the development of drug delivery vehicles. The ΔP1 structure is one piece in a puzzle to dismantle the diverse weapons of the human pathogen *P. aeruginosa*.

CHAPTER IV – SUMMARY AND CONCLUSIONS

Bacterial secretion systems are sophisticated macromolecular machines comprised of many components, each of which play an integral role in the function of the complex. Secretion systems are used by bacterial species to compete for survival in their microenvironments, making them intriguing targets for the study of bacterial antagonism, commensalism, evolution, and pathogenesis (**Section 1.1**). This dissertation described the structural study of different components of a T4SS and a T4P, two secretion systems with different roles in promoting bacterial survival. The conjugative Eex protein from F-like T4SS, TraG, was characterized through a variety of biophysical methods which revealed that the C-terminal periplasmic domain TraG* has a number of putative disordered regions, many of which are likely essential for performing the many PPIs required for its function (**Chapter II**). The structure of the N-terminally truncated type IV pilin monomer from *P. aeruginosa* strain P1 was solved using a modified MR model from an AF prediction (**Chapter III**). The structure reported in this dissertation is the first group Ia pilin to be structurally characterized, and through comparing its structure to other pilins it was revealed that classifying type IV pilins based on structural homology rather than sequence homology may be beneficial for future drug development efforts.

The dynamics of the proteins in the secretion systems studied are essential to their function. In studying TraG* (**Chapter II**) and other T4SS proteins (TraS, TraF, TrbB[438] and TraY) throughout the course of this thesis, it became apparent that the T4SS is a highly dynamic complex which features disordered regions in many of its protein components, of which many can perform LLPS[4]. This likely serves a purpose in the function of the T4SS as LLPS occurs in a controlled fashion in many organisms to form membraneless organelles, such as the aggresome in bacteria and nucleoli in eukaryotes[230,439,440]. It is theorized here that similar to how nucleoli are formed to

stabilize single-stranded nucleic acid-protein complexes, LLPS in the bacterial T4SS may be essential in stabilizing conjugative ssDNA as it crosses the mating junction. As well, the mechanism behind the formation of the mating junction is unknown, however recent studies on OmpA have revealed the protein is responsible for phase separation in the LPS layer of the bacterial cell wall, allowing for some permeability in allowing endo/exocytosis to an otherwise high-integrity barrier[229]. OmpA has been shown to be an important interacting partner with many T4SS components on the recipient cell wall and has been proposed as a binding site for conjugative pili[65,108,441]. In a scenario where a conjugative donor cell extends its pilus and binds OmpA on a recipient cell, the pilus would be retracted such that OmpA is proximal to the T4SS complex of the donor cell; perhaps the proximity of many proteins which perform LLPS instigates the formation of a lipid-excluding granule, allowing for the temporary fusion of cellular membranes thus forming the mating junction through the diffusion of phospholipids via phase separation. As well, it is proposed here that the IDRs in prokaryotic secretion systems have evolved such that these proteins can fold transiently into different active conformations to interact with multiple proteins and have multiple functions to further increase the efficiency of a small genome.

In studying ΔP1 and comparing it to other type IV pilins, the dynamics of the assembled T4P are evidently dependent on the structure of the pilin monomer (**Chapter III**). As self-assembly of ΔP1 is different from ΔK122 despite structural homology in the α-helix and β-meander, the structural dynamics of variable connecting loop regions are responsible for the observed functional differences. These differences in structure are important in many mechanisms, including binding to a receptor on host mucosal cells for pathogenicity, providing twitching motility across surfaces, aiding in microcolony and biofilm formation, cell signaling, DNA uptake, and regulating virulence gene expression[105,162–164]. The pilin monomer must accommodate all

described functions through different conformations, and it is proposed here that conformational changes in the α-β loop of the pilin in the assembled pilus may be involved in the mechanism for recognition of different strains by *P. aeruginosa*, as attachment of T4P can proposedly transduce signals through the assembled pilus to regulate virulence gene expression. Kin recognition has been observed to be initiated by T4P in *V. cholerae*[435], and it is likely that minor differences in the structure and dynamics of the PilA major pilin and minor pilins is responsible for the specificity of these interactions.

The 2024 Nobel Prize in Chemistry was awarded to members of the research teams who designed AlphaFold and RoseTTAFold; these algorithms were amazing achievements in predictive modelling of protein structures that were deemed significant contributions to the advancement of science, and certainly contributed greatly to the success of the research presented in this dissertation. As shown with ΔP1, their predictions are highly accurate when there are homologous structures in the PDB, as the main chain of the solved ΔP1 crystal structure was nearly identical to the AF model used for MR. It must be emphasized that these algorithms and the predictive models they produce are excellent tools but are not replacements for structural experiments as they do not provide accurate information regarding sidechain placement or interactions with solvent. This information is critical for rational drug discovery as biomolecules predominantly perform their function and interactions in aqueous environments. The solvent interactions seen in the ΔP1 structure were useful for characterising how regions with no secondary structure maintained their fold; these regions were determined to be unique to group Ia T4P from *P. aeruginosa* using ProSMART. Predictive models also do not provide accurate information on dynamics as was acquired for TraG* constructs through CIU-MS and SEC-MALS-SAXS, and was visualized in ΔP1 through anisotropic B-factors of the crystal structure. As proteins function

through their conformational movements, characterising these dynamics are essential in understanding their mechanism of action. As well, the predictions made by AF and RF are less accurate when there are no homologous structures in the PDB, and they can have difficulty in properly identifying disordered sequences[442], as was seen for TraG. Therefore, structural studies such as the research described in this dissertation are essential to confirm protein structure predictions made by algorithms and to elaborate on their dynamic movements.

The study of component proteins of bacterial secretion systems provides valuable information on how they interact with other proteins in their complex, which aids in elucidating the functional mechanism of said system. These macromolecular apparatuses are important to study as increasing numbers of MDR bacterial pathogens have been observed globally as attributed to poor nosocomial hygiene, excessive use of antibiotics in animal agriculture, over-prescription of broad spectrum antibiotics, and furthered by the COVID-19 pandemic due to de-prioritization of bacterial infection prevention and control, over-sanitization with biocides, shortage of personal protective equipment and crowding of infected hospital patients[57,443–446]. The structural studies of TraG* from the F-like plasmid and the ΔP1 pilin from *P. aeruginosa* described in this dissertation aid in the understanding of bacterial function and evolution, and could improve future efforts for rational drug design to target these systems in bacterial pathogens and in adapting secretion system components for bionanotechnology applications.

CHAPTER V – REFERENCES

1. Green, E. R. & Mecsas, J. Bacterial Secretion Systems: An Overview. *Microbiol Spectr* **4**, (2016).
2. Grossman, A. S. *et al.* A Surface Exposed, Two-Domain Lipoprotein Cargo of a Type XI Secretion System Promotes Colonization of Host Intestinal Epithelia Expressing Glycans. *Front Microbiol* **13**, (2022).
3. Maphosa, S., Moleleki, L. N. & Motaung, T. E. Bacterial secretion system functions: evidence of interactions and downstream implications. *Microbiology (N Y)* **169**, (2023).
4. Bragagnolo, N. *et al.* Protein dynamics in f-like bacterial conjugation. *Biomedicines* **8**, (2020).
5. Baron, C. Antivirulence drugs to target bacterial secretion systems. *Curr Opin Microbiol* **13**, 100–105 (2010).
6. Costa, T. R. D. *et al.* Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat Rev Microbiol* **13**, 343–359 (2015).
7. Bragagnolo, N. & Audette, G. F. Solution characterization of the dynamic conjugative entry exclusion protein TraG. *Structural Dynamics* **9**, (2022).
8. Bragagnolo, N. & Audette, G. F. The 1.3 Å resolution structure of the truncated group Ia type IV pilin from *Pseudomonas aeruginosa* strain P1. *Acta Crystallogr D Struct Biol* **80**, 834–849 (2024).
9. Filloux, A. Bacterial protein secretion systems: Game of types. *Microbiology (N Y)* **168**, (2022).
10. Grossman, A. S., Mauer, T. J., Forest, K. T. & Goodrich-Blair, H. A Widespread Bacterial Secretion System with Diverse Substrates. *mBio* **12**, 16009 (2021).
11. Guerrero-Mandujano, A., Hernández-Cortez, C., Ibarra, J. A. & Castro-Escarpulli, G. The outer membrane vesicles: Secretion system type zero. *Traffic* **18**, 425–432 (2017).
12. Spitz, O. *et al.* Type I Secretion Systems—One Mechanism for All? *Microbiol Spectr* **7**, (2019).
13. Bocian-Ostrzycka, K. M., Grzeszczuk, M. J., Banaś, A. M. & Jagusztyn-Krynicka, E. K. Bacterial thiol oxidoreductases — from basic research to new antibacterial strategies. *Appl Microbiol Biotechnol* **101**, 3977–3989 (2017).
14. Ates, L. S., Houben, E. N. G. & Bitter, W. Type VII Secretion: A Highly Versatile Secretion System. *Microbiol Spectr* **4**, (2016).
15. Rivera-Calzada, A., Famelis, N., Llorca, O. & Geibel, S. Type VII secretion systems: structure, functions and transport models. *Nat Rev Microbiol* **19**, 567–584 (2021).

16. Anlauf, M. T. *et al.* Type 1 secretion necessitates a tight interplay between all domains of the ABC transporter. *Sci Rep* **14**, (2024).

17. Korotkov, K. V., Sandkvist, M. & Hol, W. G. J. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat Rev Microbiol* **10**, 336–351 (2012).

18. Naskar, S., Hohl, M., Tassinari, M. & Low, H. H. The structure and mechanism of the bacterial type II secretion system. *Mol Microbiol* **115**, 412–424 (2021).

19. Deng, W. *et al.* Assembly, structure, function and regulation of type III secretion systems. *Nat Rev Microbiol* **15**, 323–337 (2017).

20. Coburn, B., Sekirov, I. & Finlay, B. B. Type III Secretion Systems and Disease. *Clin Microbiol Rev* **20**, 535–549 (2007).

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

22. Fan, E., Chauhan, N., Udatha, D. B. R. K. G., Leo, J. C. & Linke, D. Type V Secretion Systems in Bacteria. *Microbiol Spectr* **4**, (2016).

23. Santos, M. N. M. *et al.* Redundancy and Specificity of Type VI Secretion vgrG Loci in Antibacterial Activity of Agrobacterium tumefaciens 1D1609 Strain. *Front Microbiol* **10**, (2020).

24. Russell, A. B., Peterson, S. B. & Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. *Nat Rev Microbiol* **12**, 137–148 (2014).

25. Bunduc, C. M. *et al.* Reconstitution of a minimal ESX-5 type VII secretion system suggests a role for PPE proteins in the outer membrane transport of proteins. *mSphere* **8**, (2023).

26. Bhoite, S., van Gerven, N., Chapman, M. R. & Remaut, H. Curli Biogenesis: Bacterial Amyloid Assembly by the Type VIII Secretion Pathway. *EcoSal Plus* **8**, (2019).

27. Paillat, M., Lunar Silva, I., Cascales, E. & Doan, T. A journey with type IX secretion system effectors: selection, transport, processing and activities. *Microbiology (N Y)* **169**, (2023).

28. Lauber, F. *et al.* Structural insights into the mechanism of protein transport by the Type 9 Secretion System translocon. *Nat Microbiol* **9**, 1089–1102 (2024).

29. Palmer, T., Finney, A. J., Saha, C. K., Atkinson, G. C. & Sargent, F. A holin/peptidoglycan hydrolase-dependent protein secretion system. *Mol Microbiol* **115**, 345–355 (2021).

30. Craig, L., Forest, K. T. & Maier, B. Type IV pili: dynamics, biophysics and functional consequences. *Nat Rev Microbiol* **17**, 429–440 (2019).

31. Giltner, C. L., Nguyen, Y. & Burrows, L. L. Type IV Pilin Proteins: Versatile Molecular Modules. *Microbiology and Molecular Biology Reviews* **76**, 740–772 (2012).

32. McCallum, M., Burrows, L. L. & Howell, P. L. The Dynamic Structures of the Type IV Pilus. *Microbiol Spectr* **7**, 1–12 (2019).

33. Vogan, A. A. & Higgs, P. G. The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biol Direct* **6**, 1 (2011).

34. Christie, P. J. The Mosaic Type IV Secretion Systems. *EcoSal Plus* **7**, 1–34 (2016).

35. Shala-Lawrence, A., Bragagnolo, N., Nowroozi-Dayeni, R., Kheyson, S. & Audette, G. F. The interaction of TraW and TrbC is required to facilitate conjugation in F-like plasmids. *Biochem Biophys Res Commun* **503**, 2386–2392 (2018).

36. Arutyunov, D. & Frost, L. S. F conjugation: Back to the beginning. *Plasmid* **70**, 18–32 (2013).

37. Hu, B., Khara, P. & Christie, P. J. Structural bases for F plasmid conjugation and F pilus biogenesis in *Escherichia coli*. *Proceedings of the National Academy of Sciences* 201904428 (2019) doi:10.1073/pnas.1904428116.

38. Christie, P. J. Type IV secretion: The Agrobacterium VirB/D4 and related conjugation systems. *Biochim Biophys Acta Mol Cell Res* **1694**, 219–234 (2004).

39. Christie, P. J., Whitaker, N. & González-Rivera, C. Mechanism and structure of the bacterial type IV secretion systems. *Biochim Biophys Acta Mol Cell Res* **1843**, 1578–1591 (2014).

40. Mariano, G. *et al.* A family of Type VI secretion system effector proteins that form ion-selective pores. *Nat Commun* **10**, (2019).

41. Gallique, M., Bouteiller, M. & Merieau, A. The Type VI Secretion System: A Dynamic System for Bacterial Communication? *Front Microbiol* **8**, (2017).

42. Lasica, A. M., Ksiazek, M., Madej, M. & Potempa, J. The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function. *Front Cell Infect Microbiol* **7**, (2017).

43. Duchêne, S. *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb Genom* **2**, e000094 (2016).

44. Souza, V. & Eguiarte, L. E. Bacteria gone native vs. bacteria gone awry?: Plasmidic transfer and bacterial evolution. *Proceedings of the National Academy of Sciences* **94**, 5501–5503 (1997).

45. Allen, H. K. *et al.* Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* **8**, 251–9 (2010).

46. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711–721 (2005).

47. Arber, W. Horizontal gene transfer among bacteria and its role in biological evolution. *Life* **4**, 217–224 (2014).

48. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679–687 (2005).

49. Gyles, C. & Boerlin, P. Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. *Vet Pathol* **51**, 328–340 (2014).

50. Grohmann, E., Muth, G. & Espinosa, M. Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev* **67**, 277–301, table of contents (2003).

51. Lerminiaux, N. A. & Cameron, A. D. S. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can J Microbiol* **65**, 34–44 (2018).

52. Jain, A. & Srivastava, P. Broad host range plasmids. *FEMS Microbiol Lett* **348**, 87–96 (2013).

53. Virolle, C., Goldlust, K., Djermoun, S., Bigot, S. & Lesterlin, C. Plasmid transfer by conjugation in gram-negative bacteria: From the cellular to the community level. *Genes (Basel)* **11**, 1–33 (2020).

54. Lujan, S. A., Guogas, L. M., Ragonese, H., Matson, S. W. & Redinbo, M. R. Disrupting antibiotic resistance propagation by inhibiting the conjugative DNA relaxase. *Proceedings of the National Academy of Sciences* **104**, 12282–12287 (2007).

55. Cabezón, E., de la Cruz, F. & Arechaga, I. Conjugation inhibitors and their potential use to prevent dissemination of antibiotic resistance genes in bacteria. *Front Microbiol* **8**, 1–7 (2017).

56. Waksman, G. From conjugation to T4S systems in Gram-negative bacteria: a mechanistic biology perspective. *EMBO Rep* **20**, e47012 (2019).

57. Thomson, J. M. & Bonomo, R. A. The threat of antibiotic resistance in Gram-negative pathogenic bacteria: β-lactams in peril! *Curr Opin Microbiol* **8**, 518–524 (2005).

58. Sultan, I. *et al.* Antibiotics, resistome and resistance mechanisms: A bacterial perspective. *Front Microbiol* **9**, (2018).

59. Bennett, P. M. Plasmid encoded antibiotic resistance: Acquisition and transfer of antibiotic resistance genes in bacteria. *Br J Pharmacol* **153**, 347–357 (2008).

60. Williams, J. J. & Hergenrother, P. J. Exposing plasmids as the Achilles' heel of drug-resistant bacteria. *Curr Opin Chem Biol* **12**, 389–399 (2008).

61. Baron, C. A novel strategy to target bacterial virulence. *Future Microbiol* **8**, 1–3 (2013).

62. Novick, R. P. Plasmid incompatibility. *Microbiol Rev* **51**, 381–95 (1987).

63. Fernandez-Lopez, R., de Toro, M., Moncalian, G., Garcillan-Barcia, M. P. & de la Cruz, F. Comparative Genomics of the Conjugation Region of F-like Plasmids: Five Shades of F. *Front Mol Biosci* **3**, 71 (2016).

64. Lawley, T. D., Klimke, W. a, Gubbins, M. J. & Frost, L. S. F factor conjugation is a true type IV secretion system. *FEMS Microbiol Lett* **224**, 1–15 (2003).

65. Kishida, K. *et al.* Contributions of F-specific subunits to the F plasmid-encoded type IV secretion system and F pilus. *Mol Microbiol* **117**, 1275–1290 (2022).

66. Oluwadare, M. *et al.* The Role of the Salmonella spvB IncF Plasmid and Its Resident Entry Exclusion Gene traS on Plasmid Exclusion. *Front Microbiol* **11**, 1–13 (2020).

67. Garcillán-Barcia, M. P. & de la Cruz, F. Why is entry exclusion an essential feature of conjugative plasmids? *Plasmid* **60**, 1–18 (2008).

68. Kamruzzaman, M., Mathers, A. J. & Iredell, J. R. A Novel Plasmid Entry Exclusion System in pKPC_UVA01, a Promiscuous Conjugative Plasmid Carrying the *bla* $_{KPC}$ Carbapenemase Gene. *Antimicrob Agents Chemother* **66**, (2022).

69. Wong, J. J. W., Lu, J. & Glover, J. N. M. Relaxosome function and conjugation regulation in F-like plasmids - a structural biology perspective. *Mol Microbiol* **85**, 602–617 (2012).

70. Fekete, R. A. & Frost, L. S. Mobilization of chimeric oriT plasmids by F and R100-1: Role of relaxosome formation in defining plasmid specificity. *J Bacteriol* **182**, 4022–4027 (2000).

71. Wawrzyniak, P., Plucienniczak, G. & Bartosik, D. The different faces of rolling-circle replication and its multifunctional initiator proteins. *Front Microbiol* **8**, 1–13 (2017).

72. Koraimann, G. & Wagner, M. A. Social behavior and decision making in bacterial conjugation. *Front Cell Infect Microbiol* **4**, 1–7 (2014).

73. Zahrl, D., Wagner, A., Tscherner, M. & Koraimann, G. GroEL plays a central role in stress-induced negative regulation of bacterial conjugation by promoting proteolytic degradation of the activator protein TraJ. *J Bacteriol* **189**, 5885–5894 (2007).

74. Costa, T. R. D. *et al.* Structure of the Bacterial Sex F Pilus Reveals an Assembly of a Stoichiometric Protein-Phospholipid Complex. *Cell* **166**, 1436-1444.e10 (2016).

75. Costa, T. R. D., Patkowski, J. B., Macé, K., Christie, P. J. & Waksman, G. Structural and functional diversity of type IV secretion systems. *Nat Rev Microbiol* **22**, 170–185 (2024).

76. Kohler, V., Keller, W. & Grohmann, E. Regulation of gram-positive conjugation. *Front Microbiol* **10**, (2019).

77. Koraimann, G., Teferle, K., Markolin, G., Woger, W. & Högenauer, G. The FinOP repressor system of plasmid R1: Analysis of the antisense RNA control of traJ expression and conjugative DNA transfer. *Mol Microbiol* **21**, 811–821 (1996).

78. Ghetu, a F., Gubbins, M. J., Frost, L. S. & Glover, J. N. Crystal structure of the bacterial conjugation repressor finO. *Nat Struct Biol* **7**, 565–569 (2000).

79. Lee, S. H., Frost, L. S. & Paranchych, W. FinOP repression of the F plasmid involves extension of the half-life of FinP antisense RNA by FinO. *MGG Molecular & General Genetics* **235**, 131–139 (1992).

80. van Biesen, T., Soderbom, F., Wagner, E. G. H. & Frost, L. S. Structural and functional analyses of the FinP antisense RNA regulatory system of the F conjugative piasmid. *Mol Microbiol* **10**, 35–43 (1993).

81. Arthur, D. C. *et al.* Mapping interactions between the RNA chaperone FinO and its RNA targets. *Nucleic Acids Res* **39**, 4450–4463 (2011).

82. Arthur, D. C. *et al.* FinO is an RNA chaperone that facilitates sense-antisense RNA interactions. *EMBO Journal* **22**, 6346–6355 (2003).

83. Glover, J. N. M. *et al.* The FinO family of bacterial RNA chaperones. *Plasmid* **78**, 79–87 (2014).

84. Jerome, L. J., van Biesen, T. & Frost, L. S. Degradation of FinP antisense RNA from F-like plasmids: the RNA-binding protein, FinO, protects FinP from ribonuclease E. *J Mol Biol* **285**, 1457–1473 (1999).

85. Schröder, G. & Lanka, E. The mating pair formation system of conjugative plasmids - A versatile secretion machinery for transfer of proteins and DNA. *Plasmid* **54**, 1–25 (2005).

86. Manchak, J., Anthony, K. G. & Frost, L. S. Mutational analysis of F-pilin reveals domains for pilus assembly, phage infection and DNA transfer. *Mol Microbiol* **43**, 195–205 (2002).

87. Majdalani, N. & Ippen-Ihler, K. Membrane insertion of the F-pilin subunit is Sec independent but requires leader peptidase B and the proton motive force. *J Bacteriol* **178**, 3742–3747 (1996).

88. Zechner, E. L., Lang, S. & Schildbach, J. F. Assembly and mechanisms of bacterial type IV secretion machines. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 1073–1087 (2012).

89. Frost, L. S., Ippen-Ihler, K. & Skurray, R. A. Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiol Rev* **58**, 162–210 (1994).

90. Deltoro, D. *et al.* Walker-A Motif Acts to Coordinate ATP Hydrolysis with Motor Output in Viral DNA Packaging. *J Mol Biol* **428**, 2709–2729 (2016).

91. Shirasu, K. & Kado, C. I. Membrane location of the Ti plasmid VirB proteins involved in the biosynthesis of a pilin-like conjugative structure on Agrobacterium tumefaciens. **111**, 287–293 (1993).

92. Anthony, K. G., Kathir, P., Moore, D., Ippen-Ihler, K. & Frost, L. S. Analysis of the traLEKBP sequence and the TraP protein from three F-like plasmids: F, R100-1 and ColB2. *J Bacteriol* **178**, 3194–200 (1996).

93. Willetts, N. S. Characterization of the F transfer cistron, traL. *Genet. Res. Camb.* **21**, 205–213 (1973).

94. Silverman, P. M. & Clarke, M. B. New insights into F-pilus structure, dynamics, and function. *Integr. Biol.* **2**, 25–31 (2010).

95. Bergeron, J. R. C. & Sgourakis, N. G. Type IV Pilus: One Architectural Problem, Many Structural Solutions. *Structure* **23**, 253–255 (2015).

96. Wang, Y. a., Yu, X., Silverman, P. M., Harris, R. L. & Egelman, E. H. The Structure of F-Pili. *J Mol Biol* **385**, 22–29 (2009).

97. Folkhard, W. *et al.* X-ray diffraction and electron microscope studies on the structure of bacterial F pili. *J Mol Biol* **130**, 145–160 (1979).

98. Lawley, T. D., Gilmour, M. W., Gunton, J. E., Tracz, D. M. & Taylor, D. E. Functional and mutational analysis of conjugative transfer region 2 (Tra2) from the IncHI1 plasmid R27. *J Bacteriol* **185**, 581–591 (2003).

99. Costa, T. R. D. *et al.* Type IV secretion systems: Advances in structure, function, and activation. *Mol Microbiol* **115**, 436–452 (2021).

100. Amin, H., Ilangovan, A. & Costa, T. R. D. Architecture of the outer-membrane core complex from a conjugative type IV secretion system. *Nat Commun* **12**, (2021).

101. Liu, X., Khara, P., Baker, M. L., Christie, P. J. & Hu, B. Structure of a type IV secretion system core complex encoded by multi-drug resistance F plasmids. *Nat Commun* **13**, (2022).

102. Clarke, M., Maddera, L., Harris, R. L. & Silverman, P. M. F-pili dynamics by live-cell imaging. **105**, 17978–17981 (2008).

103. Skerker, J. M., Berg, H. C., JM, S. & HC, B. Direct observation of extension and retraction of type IV pili. *Proc Natl Acad Sci U S A* **98**, 6901 (2001).

104. Arutyunov, D., Arenson, B., Manchak, J. & Frost, L. S. F Plasmid TraF and TraH Are Components of an Outer Membrane Complex Involved in Conjugation. *J Bacteriol* **192**, 1730–1734 (2010).

105. Lento, C., Wilson, D. J. & Audette, G. F. Dimerization of the type IV pilin from Pseudomonas aeruginosa strain K122-4 results in increased helix stability as measured by time-resolved hydrogen-deuterium exchange. *Structural Dynamics* **3**, 012001 (2016).

106. Harris, R. L. & Silverman, P. M. Tra proteins characteristic of F-like type IV secretion systems constitute an interaction group by yeast two-hybrid analysis. *J Bacteriol* **186**, 5480–5485 (2004).

107. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of Plasmids. *Microbiology and Molecular Biology Reviews* **74**, 434–452 (2010).

108. Klimke, W. A. *et al.* The mating pair stabilization protein, TraN, of the F plasmid is an outer-membrane protein with two regions that are important for its function in conjugation. *Microbiology (N Y)* **151**, 3527–3540 (2005).

109. Horiuchi, T. & Komano, T. Mutational analysis of plasmid R64 thin pilus prepilin: The entire prepilin sequence is required for processing by type IV prepilin peptidase. *J Bacteriol* **180**, 4613–4620 (1998).

110. Dunny, G. M. & Leonard, B. A. Cell-cell communication in gram-positive bacteria. *Annu Rev Microbiol* **51**, 527–64 (1997).

111. Achtman, M., Kennedy, N. & Skurray, R. Cell-cell interactions in conjugating Escherichia coli: role of traT protein in surface exclusion. *Proc Natl Acad Sci U S A* **74**, 5104–5108 (1977).

112. Klimke, W. A. & Frost, L. S. Genetic analysis of the role of the transfer gene, traN, of the F and R100-1 plasmids in mating pair stabilization during conjugation. *J Bacteriol* **180**, 4036–4043 (1998).

113. Audette, G. F., Manchak, J., Beatty, P., Klimke, W. A. & Frost, L. S. Entry exclusion in F-like plasmids requires intact TraG in the door that recognizes its cognate TraS in the recipient. *Microbiology (N Y)* **153**, 442–451 (2007).

114. Lu, J. *et al.* Structural basis of specific TraD-TraM recognition during F plasmid-mediated bacterial conjugation. *Mol Microbiol* **70**, 89–99 (2008).

115. Peng, Y. *et al.* Mechanistic basis of plasmid-specific DNA binding of the F plasmid regulatory protein, TraM. *J Mol Biol* **426**, 3783–3795 (2014).

116. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res* **42**, 5715–5727 (2014).

117. Lederberg, J., Cavalli, L. L. & Lederberg, R. M. Sex compatibility in Eschericha coli. *Genetics* 720–730 (1952).

118. Watanabe, T. Infective heredity of multiple drug resistance in bacteria. *Bacteriol Rev* **27**, 87–115 (1963).

119. Ou, J. T. Role of surface exclusion genes in lethal zygosis in Escherichia coli K12 mating. *MGG Molecular & General Genetics* **178**, 573–581 (1980).

120. Viljanen, P. Polycations which disorganize the outer membrane inhibit conjugation in Escherichia coli. *J Antibiot (Tokyo)* **40**, 882–6 (1987).

121. Skurray, R. A. & Reeves, P. F factor mediated immunity to lethal zygosis in Escherichia coli K 12. *J Bacteriol* **117**, 100–106 (1974).

122. Edwards, S. & Meynell, G. G. General method for isolating de-repressed bacterial sex factors. *Nature* **219**, 869–870 (1968).

123. Meynell, E. & Ewins, A. Effect on exclusion of alterations to the sex pilus. *J Bacteriol* **113**, 71–75 (1973).

124. Watanabe, T., Sakaizumi, S. & Furuse, C. Superinfection with R factors by transduction in Escherichia coli and Salmonella typhimurium. *J Bacteriol* **96**, 1796–1802 (1968).

125. Dubnau, E. & Maas, W. K. Inhibition of replication of an F'lac episome in Hfr cells of Escherichia coli. *J Bacteriol* **95**, 531–539 (1968).

126. Novick, R. P. Extrachromosomal inheritance in bacteria. *Bacteriol Rev* **33**, 210–63 (1969).

127. Couturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. Identification and classification of bacterial plasmids. *Microbiol Rev* **52**, 375–95 (1988).

128. Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P. & Bradbury, A. R. M. Plasmid incompatibility: More compatible than previously thought? *Protein Engineering, Design and Selection* **20**, 309–313 (2007).

129. Carattoli, A. *et al.* Identification of plasmids by PCR-based replicon typing. *J Microbiol Methods* **63**, 219–228 (2005).

130. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).

131. Clewell, D. B. Tales of conjugation and sex pheromones. *Mob Genet Elements* **1**, 38–54 (2011).

132. Marrero, J. & Waldor, M. K. Interactions between inner membrane proteins in donor and recipient cells limit conjugal DNA transfer. *Dev Cell* **8**, 963–970 (2005).

133. Yamada, Y., Yamada, M. & Nakazawa, A. A ColE1-encoded gene directs entry exclusion of the plasmid. *J Bacteriol* **177**, 6064–6068 (1995).

134. Furuya, N. & Komano, T. Surface Exclusion Gene of IncI1 Plasmid R64: Nucleotide Sequence and Analysis of Deletion Mutants. *Plasmid* **32**, 80–84 (1994).

135. Gunton, J. E. *et al.* Entry exclusion in the IncHI1 plasmid R27 is mediated by EexA and EexB. *Plasmid* **59**, 86–101 (2008).

136. Humbert, M., Huguet, K. T., Coulombe, F. & Burrus, V. Entry Exclusion of Conjugative Plasmids of the IncA, IncC, and Related Untyped Incompatibility Groups. *J Bacteriol* **201**, (2019).

137. Achtman, M., Manning, P. A., Kusecek, B., Schwuchow, S. & Neil, W. Genetic analysis of F sex factor cistrons needed for surface exclusion in Escherichia coli. *J Mol Biol* **138**, 779–795 (1980).

138. Willetts, N. & Achtman, M. Genetic analysis of transfer by the Escherichia coli sex factor F, using P1 transductional complementation. *J Bacteriol* **110**, 843–851 (1972).

139. Firth, N. & Skurray, R. Characterization of the F plasmid bifunctional conjugation gene, traG. *Mol Gen Genet* **232**, 145–53 (1992).

140. Moore, D., Sowa, B. a & Ippen-Ihler, K. The effect of tra mutations on the synthesis of the F-pilin membrane polypeptide. *Mol Gen Genet* **184**, 260–4 (1981).

141. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb Perspect Biol* **2**, a000414 (2010).

142. Mai-Prochnow, A., Clauson, M., Hong, J. & Murphy, A. B. Gram positive and Gram negative bacteria differ in their sensitivity to cold plasma. *Sci Rep* **6**, (2016).

143. Sun, J., Rutherford, S. T., Silhavy, T. J. & Huang, K. C. Physical properties of the bacterial outer membrane. *Nature Reviews Microbiology* vol. 20 236–248 Preprint at https://doi.org/10.1038/s41579-021-00638-0 (2022).

144. Miller, S. I. & Salama, N. R. The gram-negative bacterial periplasm: Size matters. *PLoS Biol* **16**, e2004935 (2018).

145. Low, W. W. *et al.* Mating pair stabilization mediates bacterial conjugation species specificity. *Nat Microbiol* **7**, 1016–1027 (2022).

146. Valvekens, D., Van Montagu, M. & Van Lijsebettens, M. Agrobacterium tumefaciens-mediated transformation of Arabidopsis thaliana root explants by using kanamycin selection. *Proc Natl Acad Sci U S A* **85**, 5536–40 (1988).

147. Bailey, S., Ward, D., Middleton, R., Grossmann, J. G. & Zambryski, P. C. Agrobacterium tumefaciens VirB8 structure reveals potential protein-protein interaction sites. *Proceedings of the National Academy of Sciences* **103**, 2582–2587 (2006).

148. Smith, M. A. *et al.* Identification of the binding site of Brucella VirB8 interaction inhibitors. *Chem Biol* **19**, 1041–1048 (2012).

149. Gillespie, J. J. *et al.* Structural Insight into How Bacteria Prevent Interference between Multiple Divergent Type IV Secretion Systems. *mBio* **6**, 1–17 (2015).

150. Fercher, C. *et al.* VirB8-like protein TraH is crucial for DNA transfer in Enterococcus faecalis. *Sci Rep* **6**, 1–13 (2016).

151. Casu, B. *et al.* VirB8 homolog TraE from plasmid pKM101 forms a hexameric ring structure and interacts with the VirB6 homolog TraD. *Proc Natl Acad Sci U S A* **115**, 5950–5955 (2018).

152. Kuroda, T. *et al.* Molecular and structural analysis of Legionella DotI gives insights into an inner membrane complex essential for type IV secretion. *Sci Rep* **5**, (2015).

153. Behzadi, P., Baráth, Z. & Gajdács, M. It's not easy being green: A narrative review on the microbiology, virulence and therapeutic prospects of multidrug-resistant pseudomonas aeruginosa. *Antibiotics* **10**, 1–29 (2021).

154. Haenni, M. *et al.* Resistance of animal strains of Pseudomonas aeruginosa to carbapenems. *Front Microbiol* **8**, 1–10 (2017).

155. Curran, C. S., Bolig, T. & Torabi-Parizi, P. Mechanisms and targeted therapies for pseudomonas aeruginosa lung infection. *Am J Respir Crit Care Med* **197**, 708–727 (2018).

156. Williams, B. J., Dehnbostel, J. & Blackwell, T. S. Pseudomonas aeruginosa: Host defence in lung diseases. *Respirology* **15**, 1037–1056 (2010).

157. Jurado-Martín, I., Sainz-Mejías, M. & McClean, S. Pseudomonas aeruginosa: An audacious pathogen with an adaptable arsenal of virulence factors. *Int J Mol Sci* **22**, 1–37 (2021).

158. Bongiovanni, M. & Barda, B. Pseudomonas aeruginosa Bloodstream Infections in SARS-CoV-2 Infected Patients: A Systematic Review. *J Clin Med* **12**, (2023).

159. Ng, Q. X. *et al.* Trends in Pseudomonas aeruginosa (P. aeruginosa) Bacteremia during the COVID-19 Pandemic: A Systematic Review. *Antibiotics* **12**, 1–11 (2023).

160. Ramírez-Estrada, S., Borgatta, B. & Rello, J. Pseudomonas aeruginosa ventilator-associated pneumonia management. *Infect Drug Resist* **9**, 7–18 (2016).

161. Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. Pseudomonas genomes: Diverse and adaptable. *FEMS Microbiol Rev* **35**, 652–680 (2011).

162. Van Schaik, E. J. *et al.* DNA binding: A novel function of Pseudomonas aeruginosa type IV pili. *J Bacteriol* **187**, 1455–1464 (2005).

163. Persat, A., Inclan, Y. F., Engel, J. N., Stone, H. A. & Gitai, Z. Type IV pili mechanochemically regulate virulence factors in Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A* **112**, 7563–7568 (2015).

164. Klausen, M. *et al.* Biofilm formation by Pseudomonas aeruginosa wild type, flagella and type IV pili mutants. *Mol Microbiol* **48**, 1511–1524 (2003).

165. Wang, F. *et al.* Cryoelectron Microscopy Reconstructions of the Pseudomonas aeruginosa and Neisseria gonorrhoeae Type IV Pili at Sub-nanometer Resolution. *Structure* **25**, 1423-1435.e4 (2017).

166. Roux, N., Spagnolo, J. & De Bentzmann, S. Neglected but amazingly diverse type IVb pili. *Res Microbiol* **163**, 659–673 (2012).

167. Campbell, A. P. *et al.* Interaction of the receptor binding domains of Pseudomonas aeruginosa pili strains PAK, PAO, KB7 and P1 to a cross-reactive antibody and receptor analog: implications for synthetic vaccine design. *J Mol Biol* **267**, 382–402 (1997).

168. Hertle, R., Mrsny, R. & Fitzgerald, D. J. Dual-function vaccine for Pseudomonas aeruginosa: Characterization of chimeric exotoxin A-pilin protein. *Infect Immun* **69**, 6962–6969 (2001).

169. Horzempa, J. *et al.* Immunization with a Pseudomonas aeruginosa 1244 pilin provides O-antigen-specific protection. *Clinical and Vaccine Immunology* **15**, 590–597 (2008).

170. Audette, G. F., Yaseen, A., Bragagnolo, N. & Bawa, R. Protein Nanotubes: From Bionanotech towards Medical Applications. *Biomedicines* **7**, 46 (2019).

171. Petrov, A., Lombardo, S. & Audette, G. F. Fibril-mediated oligomerization of pilin-derived protein nanotubes. *J Nanobiotechnology* **11**, 1 (2013).

172. Giltner, C. L. *et al.* The Pseudomonas aeruginosa type IV pilin receptor binding domain functions as an adhesin for both biotic and abiotic surfaces. *Mol Microbiol* **59**, 1083–1096 (2006).

173. Gold, V. A., Salzer, R., Averhoff, B. & Kühlbrandt, W. Structure of a type IV pilus machinery in the open and closed state. *Elife* **4**, (2015).

174. Misic, A. M., Satyshur, K. A. & Forest, K. T. P. aeruginosa PilT Structures with and without Nucleotide Reveal a Dynamic Type IV Pilus Retraction Motor. *J Mol Biol* **400**, 1011–1021 (2010).

175. Chiang, P. *et al.* Functional role of conserved residues in the characteristics secretion NTPase motifs of the Pseudomonas aeruginosa type IV pilus motor proteins PilB, PilT and PilU. *Microbiology (N Y)* **154**, 114–126 (2008).

176. Guzzo, C. R., Dunger, G., Salinas, R. K. & Farah, C. S. Structure of the PilZ-FimXEAL-c-di-GMP complex responsible for the regulation of bacterial type IV pilus biogenesis. *J Mol Biol* **425**, 2174–2197 (2013).

177. Tammam, S. *et al.* Characterization of the PilN, PilO and PilP type IVa pilus subcomplex. *Mol Microbiol* **82**, 1496–1514 (2011).

178. Koo, J., Lamers, R. P., Rubinstein, J. L., Burrows, L. L. & Howell, P. L. Structure of the Pseudomonas aeruginosa Type IVa Pilus Secretin at 7.4 Å. *Structure* **24**, 1778–1787 (2016).

179. Treuner-Lange, A. *et al.* PilY1 and minor pilins form a complex priming the type IVa pilus in Myxococcus xanthus. *Nat Commun* **11**, (2020).

180. Marko, V. A., Kilmury, S. L. N., MacNeil, L. T. & Burrows, L. L. Pseudomonas aeruginosa type IV minor pilins and PilY1 regulate virulence by modulating FimS-AlgR activity. *PLoS Pathog* **14**, 1–26 (2018).

181. Kolappan, S. & Craig, L. SeMet Crystal Structure of Neisseria meningitidis Non-core Minor Pilin PilV in the Monoclinic Form. PDB ID 5V0M. (2018).

182. Korotkov, K. V. & Hol, W. G. J. Structure of the GspK-GspI-GspJ complex from the enterotoxigenic Escherichia coli type 2 secretion system. *Nat Struct Mol Biol* **15**, 462–468 (2008).

183. Giltner, C. L., Habash, M. & Burrows, L. L. Pseudomonas aeruginosa minor pilins are incorporated into type IV Pili. *J Mol Biol* **398**, 444–461 (2010).

184. Webster, S. S. *et al.* Force-Induced Changes of PilY1 Drive Surface Sensing by Pseudomonas aeruginosa. *mBio* **13**, (2022).

185. Pelicic, V. Mechanism of assembly of type 4 filaments: everything you always wanted to know (but were afraid to ask). *Microbiology (N Y)* **169**, (2023).

186. Lu, C., Korotkov, K. V. & Hol, W. G. J. Crystal structure of the full-length ATPase GspE from the Vibrio vulnificus type II secretion system in complex with the cytoplasmic domain of GspL. *J Struct Biol* **187**, 223–235 (2014).

187. Patrick, M., Gray, M. D., Sandkvist, M. & Johnson, T. L. Type II Secretion in Escherichia coli . *EcoSal Plus* **4**, (2010).

188. Hazes, B., Sastry, P. A., Hayakawa, K., Read, R. J. & Irvin, R. T. Crystal structure of Pseudomonas aeruginosa PAK pilin suggests a main-chain-dominated mode of receptor binding. *J Mol Biol* **299**, 1005–1017 (2000).

189. Kus, J. V., Tullis, E., Cvitkovitch, D. G. & Burrows, L. L. Significant differences in type IV pilin allele distribution among Pseudomonas aeruginosa isolates from cystic fibrosis (CF) versus non-CF patients. *Microbiology (N Y)* **150**, 1315–1326 (2004).

190. Audette, G. F., Irvin, R. T. & Hazes, B. Crystallographic analysis of the Pseudomonas aeruginosa strain K122-4 monomeric pilin reveals a conserved receptor-binding architecture. *Biochemistry* **43**, 11427–11435 (2004).

191. Craig, L. *et al.* Type IV pilin structure and assembly: X-ray and EM analyses of Vibrio cholerae toxin-coregulated pilus and Pseudomonas aeruginosa PAK pilin. *Mol Cell* **11**, 1139–50 (2003).

192. Asikyan, M. L., Kus, J. V. & Burrows, L. L. Novel proteins that modulate type IV pilus retraction dynamics in Pseudomonas aeruginosa. *J Bacteriol* **190**, 7022–7034 (2008).

193. Kus, J. V. *et al.* Modification of Pseudomonas aeruginosa Pa5196 type IV pilins at multiple sites with D-Araf by a novel GT-C family arabinosyltransferase, TfpW. *J Bacteriol* **190**, 7464–7478 (2008).

194. Villela, A. D., Harvey, H., Graham, K. & Burrows, L. L. Pseudomonas aeruginosa TfpW is a multifunctional D-Araf glycosyltransferase and oligosaccharyltransferase. *bioRxiv* **2020**, (2020).

195. Nguyen, Y., Jackson, S. G., Aidoo, F., Junop, M. & Burrows, L. L. Structural Characterization of Novel Pseudomonas aeruginosa Type IV Pilins. *J Mol Biol* **395**, 491–503 (2010).

196. Ma, W. *et al.* The biological applications of DNA nanomaterials: current challenges and future directions. *Signal Transduct Target Ther* **6**, 351 (2021).

197. Mashaghi, S., Jadidi, T., Koenderink, G. & Mashaghi, A. Lipid Nanotechnology. *Int J Mol Sci* **14**, 4242–4282 (2013).

198. Li, T., Lu, X. M., Zhang, M. R., Hu, K. & Li, Z. Peptide-based nanomaterials: Self-assembly, properties and applications. *Bioactive Materials* vol. 11 268–282 Preprint at https://doi.org/10.1016/j.bioactmat.2021.09.029 (2022).

199. Ma, X., Luan, Z. & Li, J. Inorganic Nanoparticles-Based Systems in Biomedical Applications of Stem Cells: Opportunities and Challenges. *Int J Nanomedicine* **Volume 18**, 143–182 (2023).

200. Holmannova, D., Borsky, P., Svadlakova, T., Borska, L. & Fiala, Z. Carbon Nanoparticles and Their Biomedical Applications. *Applied Sciences* **12**, 7865 (2022).

201. Audette, G. F., Van Schaik, E. J., Hazes, B. & Irvin, R. T. DNA-binding protein nanotubes: Learning from nature's nanotech examples. *Nano Lett* **4**, 1897–1902 (2004).

202. Craig, L. *et al.* Type IV Pilus Structure by Cryo-Electron Microscopy and Crystallography: Implications for Pilus Assembly and Functions. *Mol Cell* **23**, 651–662 (2006).

203. Yu, B. *et al.* A Novel Biometallic Interface: High Affinity Tip-AssociatedBinding by Pilin-Derived Protein Nanotubes. *Journal of Bionanoscience* **1**, 73–83 (2008).

204. Reguera, G. *et al.* Extracellular electron transfer via microbial nanowires. *Nature* **435**, 1098–1101 (2005).

205. Reguera, G. Harnessing the power of microbial nanowires. *Microb Biotechnol* **11**, 979–994 (2018).

206. Malvankar, N. S. *et al.* Tunable metallic-like conductivity in microbial nanowire networks. *Nat Nanotechnol* **6**, 573–579 (2011).

207. Speert, D. P., Eftekhar, F. & Puterman, M. L. Nonopsonic phagocytosis of strains of Pseudomonas aeruginosa from cystic fibrosis patients. *Infect Immun* **43**, 1006–1011 (1984).

208. Pasloske, B. L., Sastry, P. A., Finlay, B. B. & Paranchych, W. Two unusual pilin sequences from different isolates of Pseudomonas aeruginosa. *J Bacteriol* **170**, 3738–3741 (1988).

209. Castric, P. A., Sidberry, H. F. & Sadoff, J. C. Cloning and sequencing of the Pseudomonas aeruginosa 1244 pilin structural gene. *Mol Gen Genet* **216**, 75–80 (1989).

210. Audette, G. F., Irvin, R. T. & Hazes, B. Purification, crystallization and preliminary diffraction studies of the Pseudomonas aeruginosa strain K122-4 monomeric pilin. *Acta Crystallogr D Biol Crystallogr* **59**, 1665–1667 (2003).

211. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–82 (2002).

212. Uversky, V. N. A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Science* **22**, 693–724 (2013).

213. Monod, J., Wyman, J. & Changeux, J. P. On the nature of allosteric transitions: A plausible model. *J Mol Biol* **12**, 88–118 (1965).

214. Nam, K. & Wolf-Watz, M. Protein dynamics: The future is bright and complicated! *Structural Dynamics* **10**, (2023).

215. Du, X. *et al.* Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int J Mol Sci* **17**, 144 (2016).

216. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197–208 (2005).

217. Reddy Chichili, V. P., Kumar, V. & Sivaraman, J. Linkers in the structural biology of protein-protein interactions. *Protein Science* **22**, 153–167 (2013).

218. Mittag, T., Kay, L. E. & Forman-Kaya, J. D. Protein dynamics and conformational disorder in molecular recognition. *Journal of Molecular Recognition* **23**, 105–116 (2010).

219. Uversky, V. N. The protein disorder cycle. *Biophysical Reviews* vol. 13 1155–1162 Preprint at https://doi.org/10.1007/s12551-021-00853-2 (2021).

220. Louros, N., Schymkowitz, J. & Rousseau, F. Mechanisms and pathology of protein misfolding and aggregation. *Nat Rev Mol Cell Biol* **24**, 912–933 (2023).

221. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739–56 (2002).

222. Xu, Z., Horwich, A. L. & Sigler, P. B. The crystal structure of the asymmetric GroEL–GroES–(ADP)7 chaperonin complex. *Nature* **388**, 741–750 (1997).

223. Lu, J. X. *et al.* XMolecular structure of β-amyloid fibrils in alzheimer's disease brain tissue. *Cell* **154**, 1257 (2013).

224. Chen, L. *et al.* Structural instability tuning as a regulatory mechanism in protein-protein interactions. *Mol Cell* **44**, 734–744 (2011).

225. Breg, J. N. *et al.* Solution Structure of Porcine Pancreatic Procolipase as Determined from 1H Homonuclear Two-Dimensional and Three-Dimensional NMR. *Eur J Biochem* **227**, 663–672 (1995).

226. Mulvenna, J. P., Foley, F. M. & Craik, D. J. Discovery, structural determination, and putative processing of the precursor protein that produces the cyclic trypsin inhibitor sunflower trypsin inhibitor 1. *Journal of Biological Chemistry* **280**, 32245–32253 (2005).

227. Havukainen, H., Underhaug, J., Wolschin, F., Amdam, G. & Halskau, Ø. A vitellogenin polyserine cleavage site: Highly disordered conformation protected from proteolysis by phosphorylation. *Journal of Experimental Biology* **215**, 1837–1846 (2012).

228. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**, 137–149 (2012).

229. Benn, G. *et al.* Phase separation in the outer membrane of Escherichia coli. *Proceedings of the National Academy of Sciences* **118**, (2021).

230. Feric, M. *et al.* Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell* **165**, 1686–1697 (2016).

231. Gao, C. *et al.* Intrinsic disorder in protein domains contributes to both organism complexity and clade-specific functions. *Sci Rep* **11**, (2021).

232. Pavlovic-Lazetic, G. M. *et al.* Bioinformatics analysis of disordered proteins in prokaryotes. *BMC Bioinformatics* **12**, (2011).

233. Barker, C. S., Meshcheryakova, I. V., Kostyukova, A. S., Freddolino, P. L. & Samatey, F. A. An intrinsically disordered linker controlling the formation and the stability of the bacterial flagellar hook. *BMC Biol* **15**, 1–14 (2017).

234. Wootton, J. C. & Drummond, M. H. The Q-linker: a class of interdomain sequences found in bacterial multidomain regulatory proteins. *Protein Eng* **2**, 535–43 (1989).

235. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).

236. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol* **28**, 420–435 (2018).

237. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019).

238. Mitrea, D. M. & Kriwacki, R. W. Phase separation in biology; Functional organization of a higher order Short linear motifs - The unexplored frontier of the eukaryotic proteome. *Cell Communication and Signaling* **14**, 1–20 (2016).

239. Su, X. *et al.* Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science* **352**, 595–9 (2016).

240. Sfakianos, A. P., Whitmarsh, A. J. & Ashe, M. P. Ribonucleoprotein bodies are phased in. *Biochem Soc Trans* **44**, 1411–1416 (2016).

241. Taylor, J. P., Brown, R. H. & Cleveland, D. W. Decoding ALS: From genes to mechanism. *Nature* **539**, 197–206 (2016).

242. Hou, C. *et al.* PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res* **51**, D460–D465 (2023).

243. Lin, Y. H., Forman-Kay, J. D. & Chan, H. S. Sequence-Specific Polyampholyte Phase Separation in Membraneless Organelles. *Phys Rev Lett* **117**, 1–6 (2016).

244. Kim, S. *et al.* Complexation and coacervation of like-charged polyelectrolytes inspired by mussels. *Proceedings of the National Academy of Sciences* **113**, E847–E853 (2016).

245. Yeo, G. C., Keeley, F. W. & Weiss, A. S. Coacervation of tropoelastin. *Adv Colloid Interface Sci* **167**, 94–103 (2011).

246. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, 1–48 (2018).

247. Dannenhoffer-Lafage, T. & Best, R. B. A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. *Journal of Physical Chemistry B* **125**, 4046–4056 (2021).

248. Efimov, A. V. & Brazhnikov, E. V. Relationship between intramolecular hydrogen bonding and solvent accessibility of side-chain donors and acceptors in proteins. *FEBS Lett* **554**, 389–393 (2003).

249. Kertesz, M. Pancake Bonding: An Unusual Pi-Stacking Interaction. *Chemistry - A European Journal* **25**, 400–416 (2019).

250. Haugh, E. F. & Hirschfelder, J. O. Pi-electron forces between conjugated double bond molecules. *J Chem Phys* **23**, 1778–1796 (1955).

251. McGaughey, G. B., Gagné, M. & Rappé, A. K. π-Stacking Interactions. *Journal of Biological Chemistry* **273**, 15458–15463 (1998).

252. Riley, K. & Tran, K.-A. Strength and Character of R–X···π Interactions Involving Aromatic Amino Acid Sidechains in Protein-Ligand Complexes Derived from Crystal Structures in the Protein Data Bank. *Crystals (Basel)* **7**, 273 (2017).

253. Liang, Q. *et al.* MolPhase, an advanced prediction algorithm for protein phase separation. *EMBO Journal* **43**, 1898–1918 (2024).

254. Chen, Z. *et al.* Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proceedings of the National Academy of Sciences* **119**, (2022).

255. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem* **357**, 289–298 (2006).

256. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. High-throughput thermal scanning: A general, rapid dye-binding thermal shift screen for protein engineering. *J Am Chem Soc* **131**, 3794–3795 (2009).

257. Phillips, K. & de la Peña, A. H. The combined use of the Thermofluor assay and ThermoQ analytical software for the determination of protein stability and buffer optimization as an aid in protein crystallization. *Curr Protoc Mol Biol* **2011**, 1–15 (2011).

258. Seabrook, S. A. & Newman, J. High-throughput thermal scanning for protein stability: Making a good technique more robust. *ACS Comb Sci* **15**, 387–392 (2013).

259. Bai, N., Roder, H., Dickson, A. & Karanicolas, J. Isothermal Analysis of ThermoFluor Data can readily provide Quantitative Binding Affinities. *Sci Rep* **9**, 2650 (2019).

260. Boivin, S., Kozak, S. & Meijers, R. Optimization of protein purification and characterization using Thermofluor screens. *Protein Expr Purif* **91**, 192–206 (2013).

261. Dupeux, F., Röwer, M., Seroul, G., Blot, D. & Márquez, J. A. A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallogr D Biol Crystallogr* **67**, 915–919 (2011).

262. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc* **2**, 2212–2221 (2007).

263. Provencher, S. W. & Glöckner, J. Estimation of Globular Protein Secondary Structure from Circular Dichroism. *Biochemistry* **20**, 33–37 (1981).

264. Micsonai, A. *et al.* Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences* **112**, E3095–E3103 (2015).

265. Kelly, S. M., Jess, T. J. & Price, N. C. How to study proteins by circular dichroism. *Biochim Biophys Acta Proteins Proteom* **1751**, 119–139 (2005).

266. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**, 2876–2890 (2007).

267. Sarem, M. & Lüdeke, S. Circular dichroism: A powerful tool for studying biomineralization promoter proteins. *MRS Bull* **40**, 490–498 (2015).

268. Benjwal, S., Verma, S., Röhm, K.-H. & Gursky, O. Monitoring protein aggregation during thermal unfolding in circular dichroism experiments. *Protein Sci* **15**, 635–9 (2006).

269. Usha, R. & Ramasami, T. The effects of urea and n-propanol on collagen denaturation: using DSC, circular dicroism and viscosity. *Thermochim Acta* **409**, 201–206 (2004).

270. Shuken, S. R. An Introduction to Mass Spectrometry-Based Proteomics. *J Proteome Res* **22**, 2151–2171 (2023).

271. Olshina, M. A. & Sharon, M. Mass spectrometry: a technique of many faces. *Q Rev Biophys* **49**, e18 (2016).

272. Maciel, E. V. S., Pereira dos Santos, N. G., Vargas Medina, D. A. & Lanças, F. M. Electron ionization mass spectrometry: Quo vadis? *Electrophoresis* **43**, 1587–1600 (2022).

273. Katta, V. & Chait, B. T. Conformational changes in proteins probed by hydrogen-exchange. *Rapid Commun Mass Spectrom* **5**, 214–217 (1991).

274. Oganesyan, I., Lento, C. & Wilson, D. J. Contemporary hydrogen deuterium exchange mass spectrometry. *Methods* **144**, 27–42 (2018).

275. Commisso, M., Anesi, A., Dal Santo, S. & Guzzo, F. Performance comparison of electrospray ionization and atmospheric pressure chemical ionization in untargeted and targeted liquid chromatography/mass spectrometry based metabolomics analysis of grapeberry metabolites. *Rapid Communications in Mass Spectrometry* **31**, 292–300 (2017).

276. Zaima, N., Hayasaka, T., Goto-Inoue, N. & Setou, M. Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry. *Int J Mol Sci* **11**, 5040–5055 (2010).

277. Lehmann, E., Zenobi, R. & Vetter, S. Matrix-assisted laser desorption/ionization mass spectra reflect solution-phase zinc finger peptide complexation. *J Am Soc Mass Spectrom* **10**, 27–34 (1999).

278. Ben-Nissan, G. & Sharon, M. The application of ion-mobility mass spectrometry for structure/function investigation of protein complexes. *Curr Opin Chem Biol* **42**, 25–33 (2018).

279. Ridgeway, M. E., Bleiholder, C., Mann, M. & Park, M. A. Trends in trapped ion mobility – Mass spectrometry instrumentation. *TrAC Trends in Analytical Chemistry* **116**, 324–331 (2019).

280. Dixit, S. M., Polasky, D. A. & Ruotolo, B. T. Collision induced unfolding of isolated proteins in the gas phase: past, present, and future. *Curr Opin Chem Biol* **42**, 93–100 (2018).

281. Vallejo, D. D. *et al.* A modified drift tube ion mobility-mass spectrometer for charge-multiplexed collision-induced unfolding. *Anal Chem* **91**, 8137–8146 (2019).

282. Luft, J. R., Wolfley, J. R. & Snell, E. H. What's in a drop? Correlating observations and outcomes to guide macromolecular crystallization experiments. *Cryst Growth Des* **11**, 651–663 (2011).

283. Smyth, M. S. x Ray crystallography. *Molecular Pathology* **53**, 8–14 (2000).

284. Dauter, Z. & Wlodawer, A. Progress in protein crystallography. *Protein Pept Lett* **23**, 201–210 (2016).

285. Kermani, A. A. A guide to membrane protein X-ray crystallography. *FEBS J* **288**, 5788–5804 (2021).

286. Kam, Z., Shore, H. B. & Feher, G. On the crystallization of proteins. *J Mol Biol* **123**, 539–555 (1978).

287. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr F Struct Biol Commun* **70**, 2–20 (2014).

288. Fomina, N. S., Masyukevich, S. V., Gall, L. N. & Gall, N. R. ESI current oscillations for a high-resistivity fluid with the simultaneous visualization of the Taylor cone. *Journal of Analytical Chemistry* **71**, 1260–1265 (2016).

289. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int J Anal Chem* **2012**, 1–40 (2012).

290. Mitchell Wells, J. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **402**, 148–185 (2005).

291. Morris, C. B., Poland, J. C., May, J. C. & McLean, J. A. Fundamentals of Ion Mobility-Mass Spectrometry for the Analysis of Biomolecules. in *Methods in Molecular Biology* vol. 2084 1–31 (Humana Press Inc., 2020).

292. Calabrese, A. N. & Radford, S. E. Mass spectrometry-enabled structural biology of membrane proteins. *Methods* **147**, 187–205 (2018).

293. Aroyo, M. I., Kirov, A., Capillas, C., Perez-Mato, J. M. & Wondratschek, H. Bilbao Crystallographic Server. II. Representations of crystallographic point groups and space groups. in *Acta Crystallographica Section A: Foundations of Crystallography* vol. 62 115–128 (2006).

294. Grosse-Kunstleve, R. W. & Adams, P. D. Algorithms for deriving crystallographic space-group information. II. Treatment of special positions. *Acta Crystallogr A* **58**, 60–65 (2002).

295. Ranatunga, W. K., Su, D. & Romero, M. F. Obtaining Crystals. in *Molecular Life Sciences* 1–9 (Springer New York, New York, NY, 2014). doi:10.1007/978-1-4614-6436-5_28-2.

296. Banayan, N. E. *et al.* Systematic enhancement of protein crystallization efficiency by bulk lysine-to-arginine (KR) substitution. *Protein Science* **33**, (2024).

297. Ameh, E. S. A review of basic crystallography and x-ray diffraction applications. *International Journal of Advanced Manufacturing Technology* **105**, 3289–3302 (2019).

298. Dimaio, F. *et al.* Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473**, 540–543 (2011).

299. Wukovitz, S. W. & Yeates, T. O. Why protein crystals favour some space-groups over others. *Nat Struct Mol Biol* **2**, 1062–1067 (1995).

300. Kemp, T. J. & Alcock, N. W. 100 Years of X-ray Crystallography. *Sci Prog* **100**, 25–44 (2017).

301. Sayre, D. *X-Ray Crystallography: The Past and Present of the Phase Problem*. *Structural Chemistry* vol. 13 (2002).

302. Pannu, N. S., Mccoy, A. J. & Read, R. J. Biological Crystallography Application of the complex multivariate normal distribution to crystallographic methods with insights into multiple isomorphous replacement phasing. *Acta Cryst* **59**, 1801–1808 (2003).

303. Foos, N., Rizk, M. & Nanao, M. H. Single-support serial isomorphous replacement phasing. *Acta Crystallogr D Struct Biol* **78**, 716–724 (2022).

304. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J* **9**, 1665–1672 (1990).

305. Son, S. K., Chapman, H. N. & Santra, R. Multiwavelength anomalous diffraction at high X-ray intensity. *Phys Rev Lett* **107**, (2011).

306. Luo, Z. Selenourea: A convenient phasing vehicle for macromolecular X-ray crystal structures. *Sci Rep* **6**, (2016).

307. Scapin, G. Molecular replacement then and now. in *Acta Crystallographica Section D: Biological Crystallography* vol. 69 2266–2275 (2013).

308. Skou, S., Gillilan, R. E. & Ando, N. Synchrotron-based small-angle X-ray scattering of proteins in solution. *Nat Protoc* **9**, 1727–39 (2014).

309. Jacques, D. A. & Trewhella, J. Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls. *Protein Science* **19**, 642–657 (2010).

310. Petoukhov, M. V. & Svergun, D. I. Applications of small-angle X-ray scattering to biomacromolecular solutions. *International Journal of Biochemistry and Cell Biology* **45**, 429–437 (2013).

311. Burger, V. M., Arenas, D. J. & Stultz, C. M. A Structure-free Method for Quantifying Conformational Flexibility in proteins. *Sci Rep* **6**, 1–9 (2016).

312. Tondnevis, F., Gillilan, R. E., Bloom, L. B. & McKenna, R. Solution study of the Escherichia coli DNA polymerase III clamp loader reveals the location of the dynamic ψχ heterodimer. *Structural Dynamics* **2**, (2015).

313. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* **589**, 2570–2577 (2015).

314. Liu, H. & Zwart, P. H. Determining pair distance distribution function from SAXS data using parametric functionals. *J Struct Biol* **180**, 226–234 (2012).

315. Jay Breidt, F., Erciulescu, A. & Woerd, M. van der. Autocovariance structures for radial averages in small-angle X-ray scattering experiments. *J Time Ser Anal* **33**, 704–717 (2012).

316. Graewert, M. A. *et al.* Adding Size Exclusion Chromatography (SEC) and Light Scattering (LS) Devices to Obtain High-Quality Small Angle X-Ray Scattering (SAXS) Data. *Crystals (Basel)* **10**, 975 (2020).

317. Rambaldi, D. C. *et al.* An Analytical Method for Size and Shape Characterization of Blood Lipoproteins. *Clin Chem* **53**, 2026–2029 (2007).

318. Li, Y., Weiss, W. F. & Roberts, C. J. Characterization of High-Molecular-Weight Nonnative Aggregates and Aggregation Kinetics by Size Exclusion Chromatography With Inline Multi-Angle Laser Light Scattering. *J Pharm Sci* **98**, 3997–4016 (2009).

319. Velours, C. *et al.* Determination of the Absolute Molar Mass of [Fe-S]-Containing Proteins Using Size Exclusion Chromatography-Multi-Angle Light Scattering (SEC-MALS). *Biomolecules* **12**, (2022).

320. Some, D., Amartely, H., Tsadok, A. & Lebendiker, M. Characterization of proteins by size-exclusion chromatography coupled to multi-angle light scattering (Sec-mals). *Journal of Visualized Experiments* **2019**, (2019).

321. Meyer, M. & Morgenstern, B. Characterization of gelatine and acid soluble collagen by size exclusion chromatography coupled with multi angle light scattering (SEC-MALS). *Biomacromolecules* **4**, 1727–1732 (2003).

322. Stetefeld, J., McKenna, S. A. & Patel, T. R. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys Rev* **8**, 409–427 (2016).

323. Qian, N. & Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* **202**, 865–884 (1988).

324. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**, 681–697 (2019).

325. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* **10**, 2997–3011 (1982).

326. DeLisi, C. Computers in Molecular Biology: Current Applications and Emerging Trends. *Science (1979)* **240**, 47–52 (1988).

327. Pearce, R. & Zhang, Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr Opin Struct Biol* **68**, 194–207 (2021).

328. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

329. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

330. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7–8 (2015).

331. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439–D444 (2022).

332. Bragagnolo, N. Characterization of a Highly Dynamic Region in TraG from the F Plasmid. (York University, 2019).

333. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).

334. Bolognesi, B. *et al.* A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep* **16**, 222–231 (2016).

335. Chhetri, G., Kalita, P. & Tripathi, T. An efficient protocol to enhance recombinant protein expression using ethanol in Escherichia coli. *MethodsX* **2**, 385–391 (2015).

336. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Science* **4**, 2411–2423 (1995).

337. Cortés-Ríos, J. *et al.* Protein quantification by bicinchoninic acid (BCA) assay follows complex kinetics and can be performed at short incubation times. *Anal Biochem* **608**, (2020).

338. Garg, V. K. *et al.* MFPPI – Multi FASTA ProtParam Interface. *Bioinformation* **12**, 74–77 (2016).

339. Wu, T., Gale-Day, Z. J. & Gestwicki, J. E. <scp>DSFworld</scp> : A flexible and precise tool to analyze differential scanning fluorimetry data. *Protein Science* **33**, (2024).

340. Griko, Y., Sreerama, N., Osumi-Davis, P., Woody, R. W. & Woody, A. Y. Thermal and urea-induced unfolding in T7 RNA polymerase: calorimetry, circular dichroism and fluorescence study. *Protein Sci* **10**, 845–53 (2001).

341. Glover, K., Mei, Y. & Sinha, S. C. Identifying intrinsically disordered protein regions likely to undergo binding-induced helical transitions. *Biochim Biophys Acta Proteins Proteom* **1864**, 1455–1463 (2016).

342. Matsuo, K., Sakurada, Y., Yonehara, R., Kataoka, M. & Gekko, K. Secondary-structure analysis of denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy. *Biophys J* **92**, 4088–4096 (2007).

343. Hopkins, J. B., Gillilan, R. E. & Skou, S. BioXTAS RAW: Improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J Appl Crystallogr* **50**, 1545–1553 (2017).

344. Manalastas-Cantos, K. *et al.* ATSAS 3.0 : expanded functionality and new tools for small-angle scattering data analysis . *J Appl Crystallogr* **54**, 343–355 (2021).

345. Volkov, V. V. & Svergun, D. I. Uniqueness of ab initio shape determination in small-angle scattering. in *Journal of Applied Crystallography* vol. 36 860–864 (2003).

346. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys J* **76**, 2879–2886 (1999).

347. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J Appl Crystallogr* **42**, 342–346 (2009).

348. Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2**, 207–217 (2015).

349. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* **129**, 5656–5664 (2007).

350. Till, M. *et al.* Improving the success rate of protein crystallization by random microseed matrix screening. *J Vis Exp* 1–6 (2013) doi:10.3791/50548.

351. Truong, J. Q., Nguyen, S., Bruning, J. B. & Shearwin, K. E. Derivatization of protein crystals with I3C using random microseed matrix screening. *Journal of Visualized Experiments* **2021**, 1–19 (2021).

352. Dong, A. *et al.* In situ proteolysis for protein crystallization and structure determination. *Nat Methods* **4**, 1019–1021 (2007).

353. Walter, T. S. *et al.* Lysine Methylation as a Routine Rescue Strategy for Protein Crystallization. *Structure* **14**, 1617–1622 (2006).

354. Fiumara, F., Fioriti, L., Kandel, E. R. & Hendrickson, W. A. Essential Role of Coiled Coils for Aggregation and Activity of Q/N-Rich Prions and PolyQ Proteins. *Cell* **143**, 1121–1135 (2010).

355. Lancaster, A. K., Nutter-Upham, A., Lindquist, S. & King, O. D. PLAAC: A web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**, 2501–2502 (2014).

356. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* **17**, 149–163 (1993).

357. Walsh, I., Martin, A. J. M., Di domenico, T. & Tosatto, S. C. E. Espritz: Accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).

358. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**, 16–21 (2017).

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

360. Cid-Samper, F. *et al.* An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Rep* **25**, 3422-3434.e7 (2018).

361. Kantardjieff, K. A. & Rupp, B. Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics* **20**, 2162–2168 (2004).

362. Kirkwood, J., Hargreaves, D., O'Keefe, S. & Wilson, J. Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics* **31**, 1444–1451 (2015).

363. Dumetz, A. C., Snellinger-O'Brien, A. M., Kaler, E. W. & Lenhoff, A. M. Patterns of protein-protein interactions in salt solutions and implications for protein crystallization. *Protein Science* **16**, 1867–1877 (2007).

364. Wilman, H. R., Shi, J. & Deane, C. M. Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins: Structure, Function and Bioinformatics* **82**, 1960–1970 (2014).

365. Geetha, V. Distortions in protein helices. *Int J Biol Macromol* **19**, 81–89 (1996).

366. Tamara, S. *et al.* Symmetry of Charge Partitioning in Collisional and UV Photon-Induced Dissociation of Protein Assemblies. *J Am Chem Soc* **138**, 10860–10868 (2016).

367. Watt, S. J. *et al.* Effect of Protein Stabilization on Charge State Distribution in Positive- and Negative-Ion Electrospray Ionization Mass Spectra. *J Am Soc Mass Spectrom* **18**, 1605–1611 (2007).

368. Winkler, R. ESIprot: A universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. *Rapid Communications in Mass Spectrometry* **24**, 285–294 (2010).

369. Polasky, D. A., Dixit, S. M., Fantin, S. M. & Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal Chem* **91**, 3147–3155 (2019).

370. Sethi, R. & Ylänne, J. Small-angle X-ray scattering reveals compact domain-domain interactions in the n-terminal region of filamin C. *PLoS One* **9**, (2014).

371. Franke, D., Jeffries, C. M. & Svergun, D. I. Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions. *Biophys J* **114**, 2485–2492 (2018).

372. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).

373. Petoukhov, M. V & Svergun, D. I. Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr D Biol Crystallogr* **71**, 1051–1058 (2015).

374. Booth, W. T. *et al.* Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* **3**, 760–768 (2018).

375. Thielges, M. C., Chung, J. K., Axup, J. Y. & Fayer, M. D. Influence of histidine tag attachment on picosecond protein dynamics. *Biochemistry* **50**, 5799–5805 (2011).

376. Prajapati, K. P. *et al.* Rapid Coaggregation of Proteins Without Sequence Similarity: Possible Role of Conformational Complementarity. *Biochemistry* (2024) doi:10.1021/acs.biochem.4c00282.

377. Kim, H. J. *et al.* Transglutaminase 2 crosslinks the glutathione S-transferase tag, impeding protein–protein interactions of the fused protein. *Exp Mol Med* **53**, 115–124 (2021).

378. Bai, Y. *et al.* Quantitative interrogation of protein co-aggregation using multi-color fluorogenic protein aggregation sensors. *Chem Sci* **12**, 8468–8476 (2021).

379. Wissmueller, S. *et al.* Protein-protein interactions: Analysis of a false positive GST pulldown result. *Proteins: Structure, Function and Bioinformatics* **79**, 2365–2371 (2011).

380. Echalier, A., Glazer, R. L., Fülöp, V. & Geday, M. A. Assessing crystallization droplets using birefringence. *Acta Crystallogr D Biol Crystallogr* **60**, 696–702 (2004).

381. Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. His-tag impact on structure. *Acta Crystallogr D Biol Crystallogr* **63**, 295–301 (2007).

382. Robert, X. *et al.* X-ray diffraction reveals the intrinsic difference in the physical properties of membrane and soluble proteins. *Sci Rep* **7**, (2017).

383. Caffrey, M. A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. *Acta Crystallographica Section F:Structural Biology Communications* **71**, 3–18 (2015).

384. Ostermeier, C. & Michel, H. Crystallization of membrane proteins. *Curr Opin Struct Biol* **7**, 697–701 (1997).

385. Zhang, L., Somasundaran, P. & Maltesh, C. Electrolyte effects on the surface tension and micellization of n-dodecyl β-D-maltoside solutions. *Langmuir* **12**, 2371–2373 (1996).

386. Calvo, E., Bravo, R., Amigo, A. & Gracia-Fadrique, J. Dynamic surface tension, critical micelle concentration, and activity coefficients of aqueous solutions of nonyl phenol ethoxylates. *Fluid Phase Equilib* **282**, 14–19 (2009).

387. Van Der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589–6631 (2014).

388. Estrozi, L. F., Boehringer, D., Shan, S. O., Ban, N. & Schaffitzel, C. Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. *Nat Struct Mol Biol* **18**, 88–90 (2011).

389. Focia, P. J., Shepotinovskaya, I. V., Seidler, J. A. & Freymann, D. M. Heterodimeric GTPase core of the SRP targeting complex. *Science* **303**, 373–7 (2004).

390. Erdogan, F. *et al.* Conjugative mating assays for sequence-specific analysis of transfer proteins involved in bacterial conjugation. *Journal of Visualized Experiments* **2017**, 1–10 (2017).

391. Wang, H. *et al.* Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* **6**, 21383 (2016).

392. Goldschmidt, L., Cooper, D. R., Derewenda, Z. S. & Eisenberg, D. Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Science* **16**, 1569–1576 (2007).

393. Rotondi, S. M. C. *et al.* Morphological Investigation of Protein Crystals by Atomic Force Microscopy. *Crystals (Basel)* **13**, 1149 (2023).

394. Mészáros, B., Erdös, G. & Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329–W337 (2018).

395. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).

396. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845–858 (2015).

397. Bayburt, T. H. & Sligar, S. G. Membrane protein assembly into Nanodiscs. *FEBS Lett* **584**, 1721–7 (2010).

398. Orioli, S., Hansen, C. G. H. & Arleth, L. Ab initio determination of the shape of membrane proteins in a nanodisc. *Acta Crystallogr D Struct Biol* **77**, 176–193 (2021).

399. Frueh, D. P., Goodrich, A. C., Mishra, S. H. & Nichols, S. R. NMR methods for structural studies of large monomeric and multimeric proteins. *Curr Opin Struct Biol* **23**, 734–739 (2013).

400. Oeffner, R. D. *et al.* Putting AlphaFold models to work with phenix.process_predicted_model and ISOLDE. *Acta Crystallogr D Struct Biol* **78**, 1303–1314 (2022).

401. McCoy, A. J., Sammito, M. D. & Read, R. J. Implications of *AlphaFold* 2 for crystallographic phasing by molecular replacement. *Acta Crystallogr D Struct Biol* **78**, 1–13 (2022).

402. Barbarin-Bocahu, I. & Graille, M. The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. *Acta Crystallogr D Struct Biol* **78**, 517–531 (2022).

403. Shen, Y., Nguyen, Y., Guarne, A. & Burrows, L. L. N-terminally truncated group I PilA from Pseudomonas aeruginosa strain 1244. PDB ID 6BBK. (2018).

404. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **372**, 774–797 (2007).

405. Nicholls, R. A., Fischer, M., Mcnicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr D Biol Crystallogr* **70**, 2487–2499 (2014).

406. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–80 (1994).

407. Leslie, A. G. W. The integration of macromolecular diffraction data. in *Acta Crystallographica Section D: Biological Crystallography* vol. 62 48–57 (2006).

408. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204–1214 (2013).

409. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674 (2007).

410. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the *CCP* 4 program suite. *Acta Crystallogr D Biol Crystallogr* **59**, 1131–1137 (2003).

411. Agirre, J. *et al.* The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **79**, 449–461 (2023).

412. Emsley, P. & Cowtan, K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126–2132 (2004).

413. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

414. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **3**, 1171–1179 (2008).

415. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355–367 (2011).

416. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP and Automatic Interpretation of Protein Electron Density Maps. in 229–244 (2003). doi:10.1016/S0076-6879(03)74011-7.

417. Lovell, S. C. *et al.* Structure validation by Cα geometry: φ,ψ and Cβ deviation. *Proteins: Structure, Function and Genetics* **50**, 437–450 (2003).

418. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291 (1993).

419. Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr D Biol Crystallogr* **55**, 191–205 (1999).

420. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

421. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27**, 135–145 (2018).

422. Gore, S. *et al.* Validation of Structures in the Protein Data Bank. *Structure* **25**, 1916–1927 (2017).

423. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21 (2010).

424. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–99 (1963).

425. Uppenberg, J., Hansen, M. T., Patkar, S. & Jones, T. A. The sequence, crystal structure determination and refinement of two crystal forms of lipase B from Candida antarctica. *Structure* **2**, 293–308 (1994).

426. Schrödinger, L. The Pymol Molecular Graphics System. Preprint at (2020).

427. Stachowski, T. R., Snell, M. E. & Snell, E. H. SAXS studies of X-ray induced disulfide bond damage: Engineering high-resolution insight from a low-resolution technique. *PLoS One* **15**, (2020).

428. Ge, M. & Pan, X. M. The contribution of proline residues to protein stability is associated with isomerization equilibrium in both unfolded and folded states. *Extremophiles* **13**, 481–489 (2009).

429. Gorgel, M. *et al.* High-resolution structure of a type IV pilin from the metal-reducing bacterium Shewanella oneidensis. *BMC Struct Biol* **15**, (2015).

430. Natalini, J. G., Singh, S. & Segal, L. N. The dynamic lung microbiome in health and disease. *Nat Rev Microbiol* **21**, 222–235 (2023).

431. Beck, J. M., Young, V. B. & Huffnagle, G. B. The microbiome of the lung. *Translational Research* **160**, 258–266 (2012).

432. Neuhaus, A. *et al.* Cryo-electron microscopy reveals two distinct type IV pili assembled by the same bacterium. *Nat Commun* **11**, (2020).

433. Ochner, H. *et al.* Structure of the Pseudomonas aeruginosa PAO1 Type IV pilus. *bioRxiv* **2024**, (2024).

434. Gower, J. C. & Dijksterhuis, G. B. *Procrustes Problems*. (Oxford University Press, 2004).

435. Adams, D. W., Stutzmann, S., Stoudmann, C. & Blokesch, M. DNA-uptake pili of Vibrio cholerae are required for chitin colonization and capable of kin recognition via sequence-specific self-interaction. *Nat Microbiol* **4**, 1545–1557 (2019).

436. Vos, B. E., Martinez-Torres, C., Burla, F., Weisel, J. W. & Koenderink, G. H. Revealing the molecular origins of fibrin's elastomeric properties by in situ X-ray scattering. *Acta Biomater* **104**, 39–52 (2020).

437. Rocco, M. *et al.* A comprehensive mechanism of fibrin network formation involving early branching and delayed single- to double-strand transition from coupled time-resolved X-ray/Light-scattering detection. *J Am Chem Soc* **136**, 5376–5384 (2014).

438. Apostol, A. J., Bragagnolo, N. J., Rodriguez, C. S. & Audette, G. F. Structural insights into the disulfide isomerase and chaperone activity of TrbB of the F plasmid type IV secretion system. *Curr Res Struct Biol* **8**, (2024).

439. Jin, X. *et al.* Membraneless organelles formed by liquid-liquid phase separation increase bacterial fitness. *Sci Adv* **7**, eabh2929 (2021).

440. Lafontaine, D. L. J., Riback, J. A., Bascetin, R. & Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nat Rev Mol Cell Biol* **22**, 165–182 (2021).

441. Riede, I. & Eschbach, M. L. Evidence that TraT interacts with OmpA of Escherichia coli. *FEBS Lett* **205**, 241–245 (1986).

442. Zhao, B., Ghadermarzi, S. & Kurgan, L. Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. *Comput Struct Biotechnol J* **21**, 3248–3258 (2023).

443. Serapide, F. *et al.* Is the Pendulum of Antimicrobial Drug Resistance Swinging Back after COVID-19? *Microorganisms* **10**, 957 (2022).

444. Seneghini, M. *et al.* It is complicated: Potential short- and long-term impact of coronavirus disease 2019 (COVID-19) on antimicrobial resistance—An expert review. *Antimicrobial Stewardship & Healthcare Epidemiology* **2**, 1–8 (2022).

445. Chang, Q., Wang, W., Regev-Yochay, G., Lipsitch, M. & Hanage, W. P. Antibiotics in agriculture and the risk to human health: How worried should we be? *Evol Appl* **8**, 240–247 (2015).

446. Pelat, C. *et al.* Hand hygiene, cohorting, or antibiotic restriction to control outbreaks of multidrug-resistant enterobacteriaceae. *Infect Control Hosp Epidemiol* **37**, 272–280 (2015).

447. Dauter, Z. Collection of X-Ray Diffraction Data from Macromolecular Crystals. in *Methods in Molecular Biology* vol. 1607 165–184 (Humana Press Inc., 2017).

448. Cruickshank, D. W. J., Helliwell, J. R. & Moffat, K. Multiplicity distribution of reflections in Laue diffraction. *Acta Crystallogr A* **43**, 656–674 (1987).

449. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Structural Biology* vol. 34 60–68 Preprint at https://doi.org/10.1016/j.sbi.2015.07.003 (2015).

450. Karplus, P. A. & Diederichs, K. Linking Crystallographic Model and Data Quality. *Science (1979)* **336**, 1030–1033 (2012).

451. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

452. Yang, H. *et al.* DCC: A Swiss army knife for structure factor analysis and validation. *J Appl Crystallogr* **49**, 1081–1084 (2016).

453. Read, R. J. *et al.* A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **19**, 1395–1412 (2011).

454. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A* **47**, 110–119 (1991).

455. Brooks-Bartlett, J. C. *et al.* Development of tools to automate quantitative analysis of radiation damage in SAXS experiments. *J Synchrotron Radiat* **24**, 63–72 (2017).

456. Jeffries, C. M., Graewert, M. A., Svergun, D. I. & Blanchet, C. E. Limiting radiation damage for high-brilliance biological solution scattering: practical experience at the EMBL P12 beamline PETRAIII. *J Synchrotron Radiat* **22**, 273–279 (2015).

457. Grant, T. D. *et al.* The accurate assessment of small-angle X-ray scattering data. *Acta Crystallogr D Biol Crystallogr* **71**, 45–56 (2015).

458. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J Appl Crystallogr* **25**, 495–503 (1992).

459. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* **589**, 2570–2577 (2015).

460. Roberts, A. P. *et al.* Revised nomenclature for transposable genetic elements. *Plasmid* **60**, 167–173 (2008).

461. Villa, L., García-Fernández, A., Fortini, D. & Carattoli, A. Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *Journal of Antimicrobial Chemotherapy* **65**, 2518–2529 (2010).

462. Manwaring, N. P., Skurray, R. A. & Firth, N. Nucleotide sequence of the F plasmid leading region. *Plasmid* **41**, 219–225 (1999).

463. Matson, S. W. & Ragonese, H. The F-plasmid TraI protein contains three functional domains required for conjugative DNA strand transfer. *J Bacteriol* **187**, 697–706 (2005).

464. Firth, N., Ippen-ihler, K. & Skurray, R. A. Structure and Function of the F Factor and Mechanism of Conjugation. *Escherichia coli and Salmonella: Cellular and Molecular Biology, Second Edition* 2377–2401 (1996).

CHAPTER VI – APPENDICES

A.1 Generation of Competent *E. coli* Cells

Overnight cultures of DH5α cells were prepared in LB medium (37 °C, >150 rpm). The overnight culture was used to inoculate sterile LB in a 1:100 ratio. Incubation was performed at 37 °C, >150 rpm until OD_{600nm} reached the beginning of the logarithmic growth phase (0.4-0.6 absorbance units). The cells were incubated on ice for 10 mins before centrifugation was performed for 10 mins at 2,700 x g, 4 °C. LB was decanted from the cell pellet and the pellet was re-suspended with 1.6 mL ice cold, filtered and autoclaved 100 mM $CaCl_2$. Cells were then incubated on ice for 30 mins before being centrifuged for 10 mins at 2,700 x g, 4 °C. Again, the supernatant was completely removed, and the pellet was re-suspended with 1.6mL 100 mM $CaCl_2$. The cells were incubated on ice for 20 mins to equilibrate. 0.25 mL of ice-cold autoclaved 80% (v/v) glycerol was slowly added to the 1.6 mL of cells and mixed gently until homogeneous. The competent cells were then stored at -80 °C for up to 6 months. The same procedure was followed for the generation of competent BL21 (DE3) cells.

A.2 Cloning of TraG* and TraS Constructs

A.2.1 Plasmid Extraction of Vectors

Overnight cultures of DH5α *E. coli* with plasmids were prepared in 10 mL LB with 50 μg/mL Kan for pET28a vectors or Amp for pGEX4T2, pBAD33:pelB_$TraG_{R100}$ and pUCIDT:TraS, at 37 °C with shaking at 200 rpm. Plasmid extraction was performed using the GeneJET plasmid miniprep kit from Thermo-Fisher following the manufacturer's protocol. The

extracted plasmids were quantified using a NanoDrop 2000 Spectrometer from Thermo Scientific and qualified using a 1.2% (w/v) agarose gel.

Table S1: Primers Used in the Creation of Plasmid Constructs for TraG and TraS. The uppercase portion of the primer sequences show their respective restriction enzyme cut sites.

Construct	Restriction Enzyme	T_m (°C)	Primer Sequence (5'–3')
pET28a:HisG*$_{R100}$_Fwd	BamHI	57	catGGATCCgcaggtagtgtggttgacg
pET28a:HisG*$_{R100}$_Rev[†]	HindIII	52	gccgcgAAGCTTttattgcttatacttatcctgatattcc
pET28a:HisG*$_{F\ \&\ R100\ 497\text{-}940}$_Fwd	BamHI	62	cgaGGATCCacacagacccgtgacggtaatatg
pET28a:HisG*$_{F\ 497\text{-}938}$_Rev[‡]	HindIII	56	cgcgAAGCTTttattctttatgctggtaactctttgc
pET28a:HisG*$_{F\ \&\ R100\ 497\text{-}863}$_Rev[+]	HindIII	58	ggcAAGCTTttaattatcaacctgatgttttacatcattacgg
pGEX4T2:TraS$_{P1}$_Fwd	BamHI	56	ggccGGATCCatgattacacaacaaattatcagtagtgag
pGEX4T2:TraS$_{P1}$_Rev	EcoRI	57	ggccGAATTCttaatgtttctttaaaacttccaactcactac
pGEX4T2:TraS$_{P2}$_Fwd	BamHI	58	ggccGGATCCtgtcccttgctaatgtcatttctaataacac
pGEX4T2:TraS$_{P2}$_Rev	EcoRI	59	ggccGAATTCttaagtgagtgtttcacttgtcttttgtg
pGEX4T2:TraS$_{P3}$_Fwd	BamHI	56	tgcgGGATTCttatatttctctctgccggaagaattcc
pGEX4T2:TraS$_{P3}$_Rev	EcoRI	57	ggccGAATTCttaaaccctcattacggatgatgtctttagg
pGEX4T2:TraS$_{P4}$_Fwd	BamHI	58	ggccGGATCCatgggagttatatttttatggtctttttagg
pGEX4T2:TraS$_{P4}$_Rev	EcoRI	61	ggccGAATTCttaaccgtatataatccctccacct

[†]This reverse primer was used for cloning pET28a:HisG*$_{R100\ 497\text{-}940}$ as well.
[‡]This primer was previously used for cloning pET28a:HisG*$_{F}$.
[+]pET28a:HisG*$_{F\ \&\ R100\ 497\text{-}863}$ were cloned with this primer and the pET28a:HisG*$_{F\ \&\ R100\ 497\text{-}940}$_Fwd primer.

A.2.2 PCR Amplification of Genes

Primers were designed based on guidelines from IDT and analyzed using the IDT OligoAnalyzer®. Melting temperatures were maintained within 5 °C between forward and reverse primers, GC content was between 35-65%, and the free energy to generate self-dimers, hairpins or heterodimers was kept above -9.0 kcal/mol. Q5 DNA polymerase was used to amplify the genes

from 200 ng of plasmid. The PCR protocol was performed as follows: initial denaturation at 98 °C for 30 seconds, then 30 cycles of denaturation, annealing and extension at 98 °C for 10 seconds, the desired annealing temperature for 20 seconds, and 72 °C for an extension time dependent on the length of the amplicon (1 min/kb). Then a final extension was performed at 72 °C for 10 mins. PCR products were purified using the GeneJET PCR Purification kit from Thermo-Fisher following the manufacturer's protocol. Amplicons were then quantified by NanoDrop 2000 and qualified using a 1.2% (w/v) agarose gel.

A.2.3 RE Digestion, Ligation and Transformation of Amplicons into Vectors

The vector and purified amplicons were double digested using the appropriate restriction enzymes in 1X NEB CutSmart Buffer. 1 μg of DNA was digested with 20,000 units of each restriction enzyme at 37 °C for 30 mins. After this incubation, 1 unit of Calf Intestinal Alkaline Phosphatase (CIP) was added to the double digested vector samples to de-phosphorylate the 5' and 3' ends of the linearized plasmid to prevent self-ligation. These samples were then incubated for an additional 30 mins at 37 °C. DD plasmids and amplicons were then isolated through separation on a 1.2% (w/v) agarose gel by performing electrophoresis at 100 V for ~40 mins, followed by gel extraction using the GeneJET Gel Extraction kit from Thermo-Fisher using the manufacturer's protocol. The DNA samples were then quantified by NanoDrop 2000.

NEBs ligation calculator was then used to design ligation experiments in 1:5 and 1:7 proportions of plasmid to amplicon, and a plasmid self-ligation control. Ligation was performed overnight at 4 °C using T4 DNA ligase in proportions described by the manufacturer's protocol. Heat inactivation was performed at 65 °C for 10 mins, and after cooling at 4 °C, 5 μL of the ligation

reaction was added to 90 µL of DH5α competent cells. The cells were incubated at 4 °C for 20 mins, heat-shock transformed at 42 °C for 90 seconds, cooled at 4 °C for 5 mins and then incubated at 37 °C for 1 hr following the addition of 900 µL of LB to equilibrate the transformed cells. The cells were spun down at 5000 x g for 10 mins at RT, then resuspended with 100 µL of LB. 30 µL of each cell resuspension was spread plated onto a LB agar plate with the appropriate antibiotic (either Kan^+ (50µg/mL) or Amp^+ (100µg/mL)), as well as control agar plates with another antibiotic and LB agar with no antibiotic to check for cell viability. Plates were kept at 37°C for 16-24 hrs.

Colonies grown from experimental LB agar plates were aseptically picked and used to inoculate 10 mL of LB broth with the appropriate selective antibiotic which was then incubated at 37 °C overnight. Plasmid extraction was performed as described in **Section A.2.1** and restriction enzyme digestion was performed as described above to check for insertion of the proper gene through performing agarose gel electrophoresis. The plasmids with the correct insert size were sent to BioBasic (Markham, ON) or The Center for Applied Genomics (Toronto, ON) for dideoxy sequencing of the multiple cloning site using the T7 Promoter and Terminator or pGEX5′ and pGEX3′ universal sequencing primers. Plasmids determined to have the correct gene inserted in the proper frame were used to transform BL21 (DE3) competent cells as in the above heat shock protocol. The plasmids pET28a:HisG*$_F$ and pET28a:HisG*$_{F\ 497\text{-}938}$ were cloned previously as described[332].

A.3 Primary Sequences of Proteins used in Chapter II

A.3.1 TraG$_F$ and R100

The below sequences are colour coded in the same scheme as **Figure S3**, where the N-terminal hydrophobic domain is coloured in purple, the region that is deleted for N-terminal truncations of TraG* is yellow, the region that is C-terminally truncated in TraG* constructs is blue, and the region present in all TraG* constructs is in green with the region responsible for binding TraS is in red.

TraG$_F$

```
        10         20         30         40         50         60
VNEVYVIAGG EWLRNNLNAI AAFMGTWTWD SIEKIALTLS VLAVAVMWVQ RHNVMDLLGW

        70         80         90        100        110        120
VAVFVLISLL VNVRTSVQII DNSDLVKVHR VDNVPVGLAM PLSLTTRIGH AMVASYEMIF

       130        140        150        160        170        180
TQPDSVTYSK TGMLFGANLI VKSTDFLSRN PEIINLFQDY VQNCVLGDIY LNHKYTLEDL

       190        200        210        220        230        240
MASADPYTLI FSRPSPLRGV YDNNNNFITC KDASVTLKDR LNLDTKTGGK TWHYYVQQIF

       250        260        270        280        290        300
GGRPDPDLLF RQLVSDSYSY FYGSSQSASQ IMRQNVTMNA LKEGITSNAA RNGDTASLVS

       310        320        330        340        350        360
LATTSSMEKQ RLAHVSIGHV TMRNLPMVQT ILTGIAIGIF PLLILAAVFN KLTLSVLKGY

       370        380        390        400        410        420
VFALMWLQTW PLLYAILNSA MTFYAKQNGA PVVLSELSQI QLKYSNLAST AGYLSAMIPP

       430        440        450        460        470        480
LSWMMVKGLG AGFSSVYSHF ASSSISPTAS AAGSVVDGNY SYGNMQTENV NGFSWSTNST

       490        500        510        520        530        540
TSFGQMMYQT GSGATATQTR DGNMVMDASG AMSRLPVGIN ATRQIAAAQQ EMAREASNRA

       550        560        570        580        590        600
ESALHGFSSS IASAWNTLSQ FGSNRGSSDS VTGGADSTMS AQDSMMASRM RSAVESYAKA

       610        620        630        640        650        660
HNISNEQATR ELASRSTNAS LGLYGDAYAK GHLGISVLGN GGGVGLQAGA KASIDGSDLD
```

```
       670        680        690        700        710        720
SHEASSGSRA SHDARHDIDA RATQDFKEAS DYFTSRKVSE SGSHTDNNAD SRVDQLSAAL

       730        740        750        760        770        780
NSAKQSYDQY TTNMTRSHEY AEMASRTESM SGQMSEDLSQ QFAQYVMKNA PQDVEAILTN

       790        800        810        820        830        840
TSSPEIAERR RAMAWSFVQE QVQPGVDNTW RESRRDIGKG MESVPSGGGS QDIIADHQGH

       850        860        870        880        890        900
QAIIEQRTQD SNIRNDVKHQ VDNMVTEYRG NIGDTQNSIR GEENIVKGQY SELQNHHKTE

       910        920        930
ALTQNNKYNE EKLAQERIPG ADSPKELLEK AKSYQHKE
```

TraG$_{R100}$

```
        10         20         30         40         50         60
VNEVYVIAGG EWLRNNLNAI AAFMGTRTWD SIEKIALTLS VLAVAVMWVQ RHNVMDLLGW

         70         80         90        100        110        120
VAVFVLISLL VNVRTSVQII DNSDLVQVHR VDNVPVGLAM PLSLTTRIGH AMVASYEMIF

       130        140        150        160        170        180
TQPDSATYSK TGMLFGANLI VKSTDFLSRN PEIINLFQDY VQNCVLGDIY LNHKYTLEDL

       190        200        210        220        230        240
MASADPYTLI FSRPSPLRGV YDSNNNFITC KDASVTLKDR LNLDTKTGGK TWHYYVQQIF

       250        260        270        280        290        300
GGRPDPDLLF RQLVSDSYSY FYGSSQSASH IMRQNVTMNA LKEGITSNAA RNGDTASLVS

       310        320        330        340        350        360
LATTSSMEKQ RLAHVSIGHV LMRNLPMVQT ILVGITIGIF PLLVLAAVFN RLTLSVLRGY

       370        380        390        400        410        420
VFALMWFQTW PLLYAILNSA MTFYAKQNGA PVVLSELSQI QLKYSDLAST AGYISVMIPP

       430        440        450        460        470        480
LSWMMVKGLG AGFSSVYSHF ASSAISPTAS AAGSVVDGNY SYGNMQTENV NGFSWSTNST

       490        500        510        520        530        540
TSFGQMMYQT GSGATATQTR DGNMVMDASG AMSRLPVGIN ATRQIAAAQQ EMAREASNRA

       550        560        570        580        590        600
ESALHGFSSS IASAWNSLSQ FGSNRGSSDS VTGGADSTMS AQDSMMASRM RSAVESYAKA
```

```
        610         620         630         640         650         660
HNISNEQATQ ELASRSTRAS AGMYGDAHAE WGVKPKILGV GGGLGVRGGG RAGIDWEDND

        670         680         690         700         710         720
AHTASSNTQS SHNARHDIDA RATQDFKEAS DYFTSRKVSE SGSHTDNNAD SRVDQLSAAL

        730         740         750         760         770         780
NSAKQSYDQY TTNMTRSHEY AEMASRTESM SGQMSEDLSQ QFAQYVMKHA PQDAEAILTN

        790         800         810         820         830         840
TSSPEIAERR RAMAWSFVQE QVQPGVDNAW RESRGDIGKG MESVPSGGGS QDIIADHQGH

        850         860         870         880         890         900
QAIIDQRTQD SNIRNDVKHQ VDNMVTEYKG NIGDTQNSVR GEENIVRGQY SELQNHHKTE

        910         920         930         940
ALSQNNKYNE EKSAQERMPG ADSPQELMKR AKEYQDKYKQ
```

A.3.2 TraS

The TraS primary sequence is coloured in the similar manner as in **Figure 2.2**, with peptide 1 coloured orange, peptide 2 in cyan, peptide 3 in rose, peptide 4 in blue and peptide 5 in red, with all other regions left uncoloured.

```
         10          20          30          40          50          60
MITQQIISSE LEVLKKHIDS GDIRIPSLWQ GLKPGLIIMG WMIFCPLLMS FLITQKTSET

         70          80          90         100         110         120
LTAVLAGGWL GLIILFIVAR IRMLYFSLPE EFLKTSSVMR VISSKLKVYF IVYMGVIFLW

        130         140         150         160         170
SFLGGGIIYG FGAILVTVIM AFLIQLDIGR YQFVGVIDAI NSYVKNKKLS RVK
```

A.3.3 Hexa-Histidine Tag

This sequence is at the N-terminus of the TraG* constructs and includes the linker with the thrombin cleavage site.

```
       10          20          30
MGSSHHHHHH SSGLVPRGSH MASMTGGQQM GRGS
```

A.3.4 Glutathione-S-Transferase Tag

This sequence is present at the N-terminus of the TraS constructs and includes the linker with the thrombin cleavage site.

```
        10         20         30         40         50         60
MSPILGYWKI KGLVQPTRLL LEYLEEKYEE HLYERDEGDK WRNKKFELGL EFPNLPYYID

        70         80         90        100        110        120
GDVKLTQSMA IIRYIADKHN MLGGCPKERA EISMLEGAVL DIRYGVSRIA YSKDFETLKV

       130        140        150        160        170        180
DFLSKLPEML KMFEDRLCHK TYLNGDHVTH PDFMLYDALD VVLYMDPMCL DAFPKLVCFK

       190        200        210        220
KRIEAIPQID KYLKSSKYIA WPLQGWQATF GGGDHPPKSD LVPRGS
```

A.4 Data Processing for SEC-MALS-SAXS

The HisG* SAXS data was buffer subtracted using frames 150-250 and 3250-3350, with the frames 1639-1672 chosen as the sample region (**Figure 2.9A**). Evolving Factor Analysis (EFA) was attempted on the aggregate and monomer peak to remove noise from aggregate scattering that disrupted the monomer signal, however the data quality was optimal when EFA was not performed and instead a sample region that did not include frames at the leading edge of the peak was used. The buffer regions used for subtraction were frames 500-600 and 2000-2100 for HisG*$_{497\text{-}940}$ with frames 1380-1418 used for the sample, and for HisG*$_{497\text{-}863}$ frames 80-145 were used as the buffer region, and frames 1715-1740 were used for the sample. HisG*$_{F\ 497\text{-}863}$ was buffer subtracted using frames 575-622 and 3150-3205; EFA was attempted, however an optimal result was produced by selecting the sample region from 1672-1689 instead of using the EFA result.

Guinier analysis was performed for each TraG* construct; in the analysis of HisG*, n_{min} was set to 13 and n_{max} was set to 209, for HisG*$_{497\text{-}940}$ the bounds of n were set as 5-233, whereas

for HisG*$_{497\text{-}863}$ and HisG*$_{F\ 497\text{-}863}$ they were set as 0-208 and 10-209, respectively. In performing an IFT with GNOM, the same values for n were set as above, in all calculations the Dmax was increased until it smoothly approached the x-axis and P(r) was then forced to equal 0 at Dmax, and all profiles but the data from HisG*$_{497\text{-}863}$ were truncated for DAMMIF/N. All IFTs were performed altering Dmax until χ^2 was optimal, and α values were reported (**Table 2.5**). For the HisG* *ab initio* model from DAMAVER, mean normalized spatial discrepancy (NSD) was 0.98 ±0.29, R_g was 47.0 Å, and MW was 119.6 kDa. For the HisG*$_{497\text{-}940}$, HisG*$_{497\text{-}863}$ and HisG*$_{F\ 497\text{-}863}$ *ab initio* models, NSD was 1.19 ±0.44, 1.43 ±0.34, and 1.15 ±0.18, R_g was 43.5 Å, 35.2 Å, and 42.8 Å, and MW was 92.3 kDa, 60.9 kDa, and 57.1 kDa, respectively. See **Section A.5.2** for more information regarding suitable values for these parameters.

A.5 Parameters for the Validation of Structural Data

A.5.1 Parameters Used for the Validation of X-ray Crystallography Data

In the collection of X-ray diffraction data, high resolution is desirable as it provides refined perceptible details within resultant electron density maps[283,447]. Resolution is dependent on the crystal's packing as it directly relates to the 'd' spacing between lattice planes. In a diffraction pattern, the highest resolution achievable is dependent on the smallest distance between crystal lattice planes that can be resolved, and therefore smaller numeric values represent better resolution[283,297,447]. Data is considered poor resolution when it is resolved to >4.0 Å resolution as chains of a protein are difficult to interpret, from 4.0–2.8 Å electron density of secondary structures can be seen, therefore the data is considered low resolution. At 2.7–1.7 Å resolution sidechains are interpretable and the resolution is considered good, while data of resolution <1.7 Å are considered

high resolution as most atoms can be clearly seen in the electron density. Higher resolution data can be collected by moving the detector closer to the crystal upon which the X-rays are being impinged. However, to generate a sufficient signal from high resolution spots, higher intensity or longer X-ray exposures are required; these can degrade crystals and induce radiation damage, which is why typically a lower resolution dataset is acquired before attempting to acquire a high resolution dataset.

Another parameter important in X-ray diffraction is completeness, which represents the number of expected diffraction spots in every frame as based on the known data resolution and space group[297,407,447]. The percentage of recorded reflections versus expected reflections should be above 95% in the overall dataset and above 80% in the highest resolution shell. Completeness can be linked to signal-to-noise, $\langle I/\sigma(I)\rangle$, which represents the intensity of the diffraction spots above the uncertainty in their measurement. Signal-to-noise will always decrease at higher resolution; benchmarks for proper data are an $\langle I/\sigma(I)\rangle$ above 10 for the overall dataset and >2 for the highest resolution shell.

In reducing data from multiple image frames to a single file representing the structure factor in reciprocal space, R_{merge} is an indication of the agreement between reflection intensity measurements in all frames that are merged for the processing of an indexed dataset. This value will increase depending on the symmetry of the space group; higher symmetry will always result in higher multiplicity in reflections. Multiplicity of the dataset should be >1 for a triclinic system and will increase for more symmetrical systems[448]. $R_{p.i.m.}$ provides an indication of the precision of the merged data and should be lower than the R_{merge} value. A high R_{merge} value is acceptable if the $CC_{1/2}$ (correlation coefficient at half peak height) is sufficiently high[449,450]. $CC_{1/2}$ is a modern metric which uses a standard statistical test to estimate the correlation of a dataset with the true

signal at half peak intensity; it is therefore related to the signal-to-noise. In this way the $CC_{1/2}$ value provides an indication of the proper resolution cut-off for the data; if the value is <0.5, the high resolution shells have poor intensity and the data should be cut-off at a lower resolution, and the $CC_{1/2}$ for the overall dataset should be as close to 1.0 as possible.

When overcoming the phase problem using MR, three parameters are used in evaluating the success of the phasing function in generating a unique solution: the refined LLG score, the TFZ score, and the R_{free} value[409]. The refined LLG score is dependent on the quality and completeness of the output model, as dictated by the fraction of scatterings the model explains, which is typically aided by the total number of reflections present and the resolution of the data. The TFZ score provides an indication of signal-to-noise as computed by comparing the LLG values from the provided rotation-translation search to those LLG values that would be acquired from a random set of rotation-translation search functions. A good solution from MR produces an initial map with a high refined LLG score where the minimum score for acceptability would be 120, a TFZ score above 8.0 is desirable, and a R_{free} value below 0.50 provides a good initial model.

In generating a proper real-space model from reciprocal space data, R_{work} (often abbreviated to 'R') and R_{free} are used to measure the similarity between the observed structure factor amplitudes and those calculated from the model[451,452]. Lower R values correspond to a model which is more representative of the diffraction data as it correlates with the electron density derived from the structure factor amplitudes. R_{free} is derived from a fraction of reflections (typically 5% of the total number of reflections, with lower numbers representing higher stringency) that are not used in refinement as a bias-free calculation for monitoring the fit of model to the data. As refinement of the model is performed, R and R_{free} should decrease proportionally, and if a situation arises where R increases yet R_{free} decreases it is an indication of overfitting. R and R_{free} will

decrease with better resolution, and for structures of ~1.5 Å benchmark values for R and R_{free} are 0.17 and 0.21, respectively.

In determining the quality of the final model, metrics are used to determine conformity of the model to expected physical and atomic properties[423,453]. Ramachandran outliers are determined through comparing ϕ and ψ angles around C_α in every residue of a polypeptide to validate main chain geometry in a protein structure; a Ramachandran map is typically derived from strict parameters of the most favourable and lowest-energy covalent bond angles based on a residue's side chain and the secondary structural element it resides in[424]. Sidechain outliers are monitored in a similar fashion, with improper torsion angles resulting in atoms which are unusually close to each other (referred to as clashes). Another metric called the RSR (real-space R-value) is a measure of the quality of fit between a modeled residue and the electron density[454]. To normalise the RSR-score to the residue type, the RSR Z-score (RSRZ) is calculated, and a residue is considered an outlier if its RSRZ value is >2. The B-factor (also known as the Debye-Waller factor or temperature factor) indicates the dispersity of atoms in the refined model[453]. It provides an indication of the structural rigidity or dynamicity of a region; if a residue has high B-factors (>80 $Å^2$) it is likely disordered.

A.5.2 Parameters Used for the Validation of SAXS Data

In batch-mode SAXS data collection, output diffraction frames must be monitored to ensure radiation damage does not affect the resultant scattering pattern[455]. In the collection of SEC-SAXS data, radiation damage is typically not an issue as sample is continuously flowed, resulting in limited exposure times of individual particles to radiation[316]. SEC-SAXS frames are inspected

by plotting the integrated intensity and corresponding expected R_g of the diffraction image on the y-axis and the volume of elution on the x-axis and checking for anomalies; sudden changes in intensity indicate a SEC-SAXS profile may have radiation-induced damage on some frames. Corresponding the SAXS scattering peaks to MALS data in a SEC-MALS-SAXS experimental configuration further simplifies qualitative data analysis.

Observing SAXS data profiles at small scattering angles, also called a Guinier approximation, is the standard method for protein SAXS data analysis and can provide indications of sample aggregation, radiation damage, interparticle interactions, and buffer mismatch[456,457]. All of these can be detected as deviations from linearity in the Guinier region, as can be observed from the normalized residuals of the Guinier fit[309]. The fit should extend to the lowest q point, if many q points (>5) are excluded in order to provide a linear fit then there may be an issue with data. If the minimum q of the fit multiplied by the R_g is >0.65, or the maximum q of the fit multiplied by the R_g is >1.3 then there may be issues with the data as these metrics define a globular protein's scattering; deviations <1.3 to 1.0 for the maximum q of the fit multiplied by the R_g indicate the scattering particle is an extended shape. A Kratky plot can also be used to validate the presence of unfolding in a scattering plot, as described in **Section 1.10.3.5**. As the information from the Guinier plot is used in subsequent calculations of MW and shape factor it is an essential qualitative analysis for SAXS data.

In performing an IFT to generate the P(r) function from the scattering profile using GNOM, quality of the calculation's fit is monitored by a χ^2 value which should be close to 1 and >0.7. The regularization parameter α would optimally be an appropriately large value depending on the dataset, as an α value close to zero produces an unstable solution and an α value approaching infinity represents and overfit calculation[343,458]. Optimal P(r) functions fall to 0 at Dmax, are

always positive, and the R_g and *I(0)* from the P(r) function match those from the Guinier analysis. Dmax values do not have an associated error as they are values that can be altered to better fit the P(r) function to the scattering data. In modeling the shape factor from the P(r) function, AMBIMETER can be used to quantify the ambiguity of the output DAMMIF bead models[373]. Ambiguity score and shape categories should be proportional; scores <3.0 should have <1000 compatible shape factors. An ambiguity score <1.5 indicates an *ab initio* shape reconstruction would be unique, scores in the range of 1.5–2.5 indicate that some care should be taken in analysing the bead model, and scores exceeding 2.5 indicate the reconstruction would be ambiguous. This ambiguity could be caused by high symmetry/anisometry of the scattered particle and is not indicative of a fault in the data. An ambiguity score of 0 with 0 compatible shape categories indicates that ambiguity could not be calculated using the algorithm due to an issue with the data. In validating output bead models from DAMAVER, the χ^2 should be close to 1.0, the normalized spatial discrepancy value should be <1.0, the model R_g and Dmax should be close to the values from the P(r) function, and the MW should be close to the expected MW[347]. Deviations from these metrics do not invalidate an *ab initio* model as structural heterogeneity arising from dynamic or disordered proteins can alter them; instead, caution should be taken in their analysis, as described by Franke *et al.* 2009[347] and Kikhney and Svergun 2015[459].

In validating EOM models derived from SAXS data, the average R_g and Dmax from the ensembles should be close to the values from the P(r) function and the χ^2 should be between 0.7–1.0[348]. The R_g and Dmax values of the models produced from EOM do not have error associated with these measurements as each model is itself an estimate dictated by the fit to the data as represented by the χ^2. The R_σ (SD of the ensemble/SD of the pool) should approach but not exceed 1.0 for fully flexible systems and the R_{flex} of the system should be appropriate based on the R_σ,

where an R_{flex} of 100% represents a fully flexible system and an R_{flex} of 0 represents a completely rigid system. If the R_{flex} value of the ensemble models is significantly smaller than that of the pool, but R_{σ} is >1.0, this may indicate a problem with the IFT data used for EOM.

A.6 Supplementary Figures and Tables

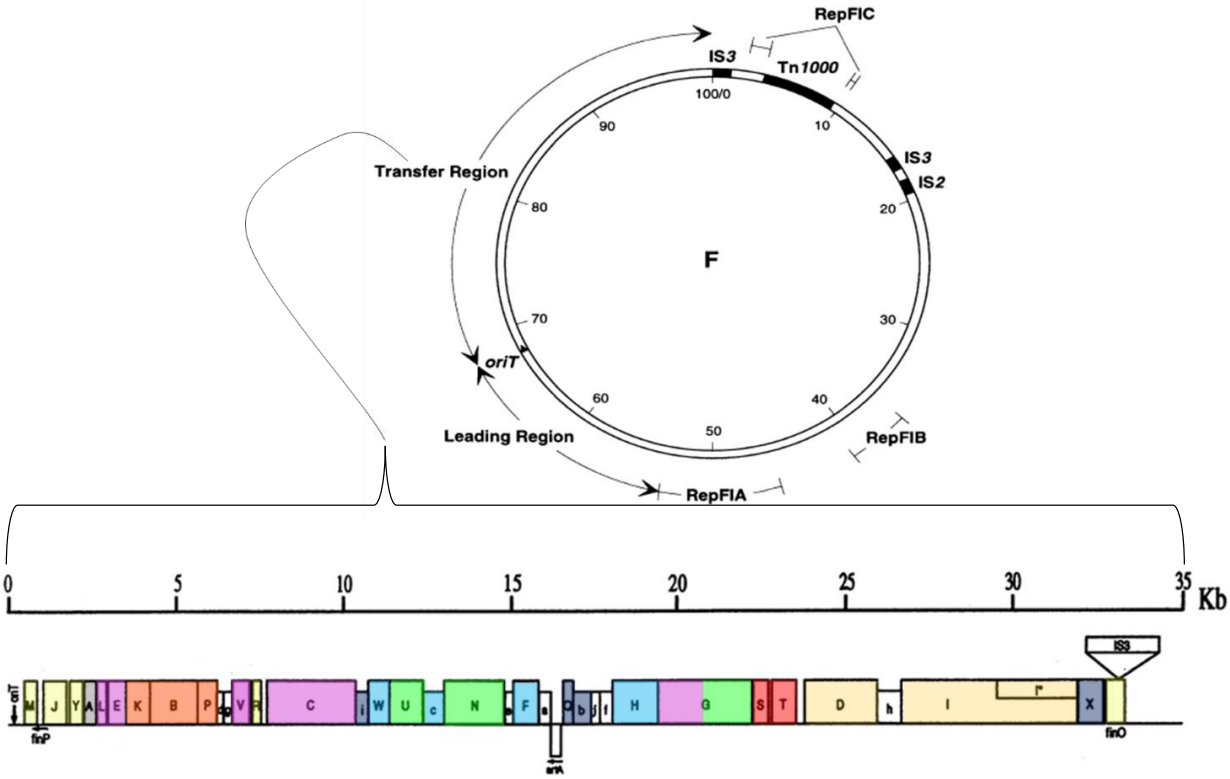

Figure S1 The Anatomy of the F plasmid, approximating positions of important genetic markers in a 100 kb map. The origin of transfer (*oriT*) represents the position where TraI nicks the plasmid, the small black triangle represents the direction of ssDNA transfer. Insertion sites 2 and 3 (IS*2* & *3*) and Tn*1000* represented by black boxes are simple and composite transposable elements, respectively[460]. RepFIA and B are replicons which function as origins of replication, while the FIC replicon is non-functional[461]. The leading region contains many genes involved in the regulation of the *tra* genes and their gene products; for more information see[462]. The transfer region is the 34 kb *tra* operon; all *tra* and *trb* genes are shown, as well as antisense genes *artA* and *finP*[89]. I* represents a gene product predicted to be expressed within the *traI* gene, but later this gene region was determined to be a separate domain of the TraI protein responsible for DNA transfer activity[463]. The *tra* genes are labelled with capital letters and *trb* proteins are shown with lower case letters. Genes are coloured based on translated protein function; the pilin gene *traA* is white, pilin processing genes are purple, genes responsible for pilus assembly/extension are coloured in fuchsia, the core complex is coloured in dark blue, while pilus retraction genes are coloured in light blue, green coloured genes are responsible for mating pair stabilization, the red genes are responsible for exclusion, and yellow genes are those involved in conjugation signalling, origin nicking, relaxase activity, and T-DNA transport. White genes are those with no known function to date. Figure adapted from Firth *et al.* (1996) and Frost *et al.* (1994)[89,464].

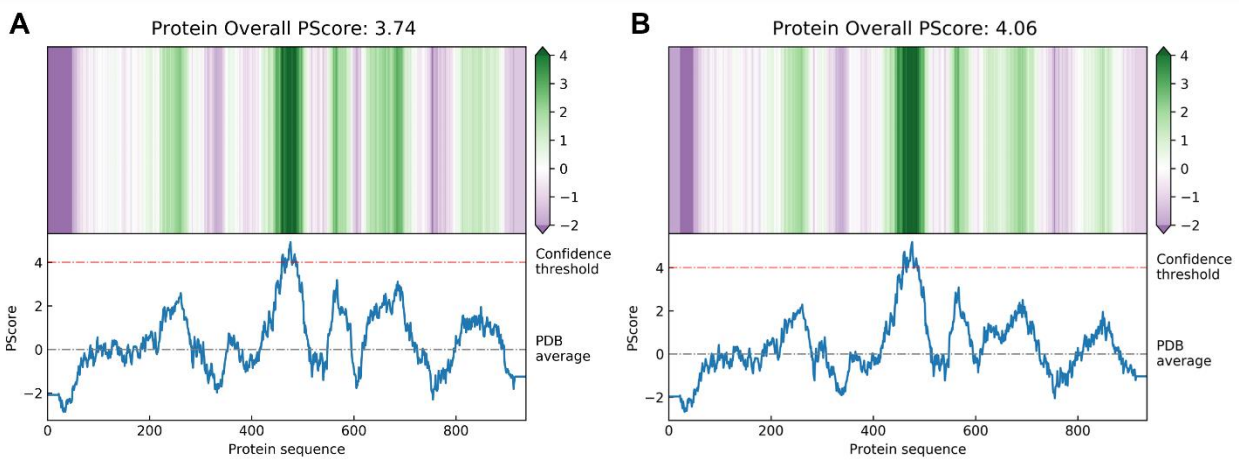

Figure S2 PScore Results from $TraG_{R100}$ and $TraG_F$ Sequences. The resultant 2D heat maps and graphs produced by entering the **A**) $TraG_{R100}$ and **B**) $TraG_F$ sequence into the PScore algorithm[246]. The only region predicted to be a phase separating IDR is the putative linker from residues 447-498, as it is the only region that surpasses the designated confidence threshold.

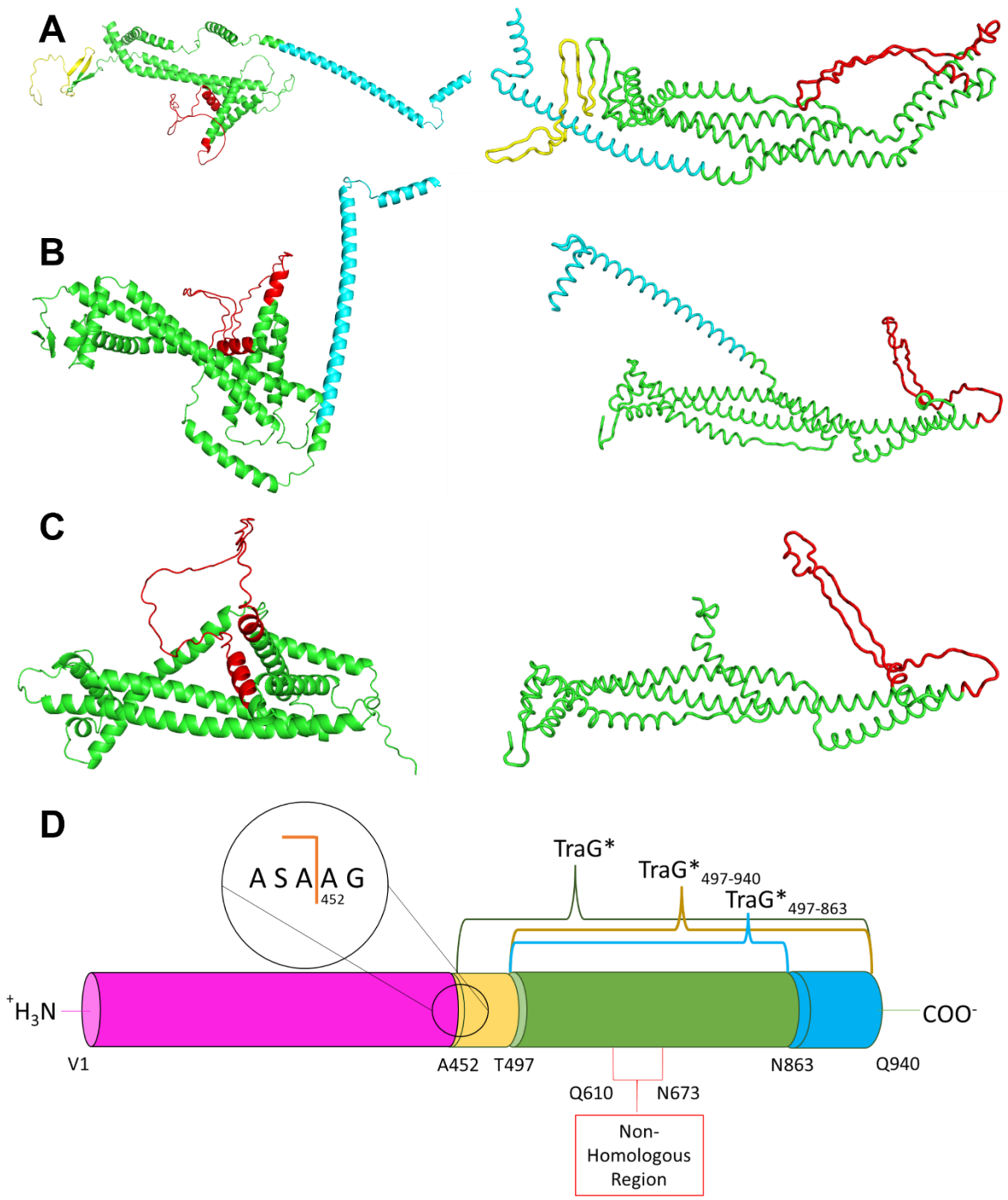

Figure S3 AF and RF models of TraG* Constructs & TraG*$_{R100}$ Schematic **A-C**) AF (left) and RF (right) models of TraG*$_{R100}$ constructs, coloured with TraG* in green, the flexible linker region in yellow, the region responsible for interacting with TraS is in red and the C-terminal truncation is in cyan. **A**) displays TraG*, **B**) is TraG*$_{497-940}$ and **C**) is TraG*$_{497-863}$; the RF prediction scores are 0.51, 0.55, and 0.66, respectively. Images were made in PyMOL v2.5.0[426]. **D**) A schematic of TraG$_{R100}$ with all TraG* constructs depicted and truncated regions of TraG* coloured as described.

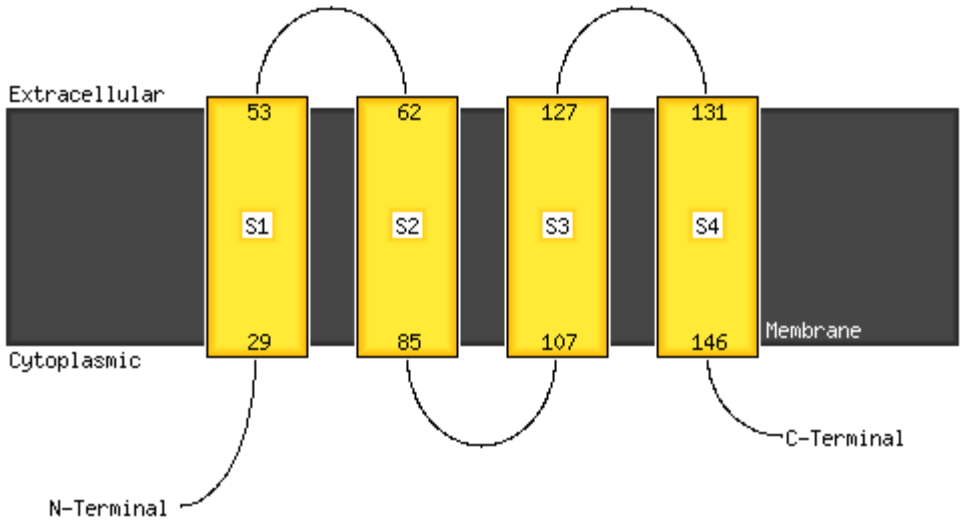

Figure S4 Phyre2 TM Protein Domain Prediction of $TraS_F$[396]. The five TraS peptide constructs were designed based on the loop regions predicted to be soluble by this algorithm, with aid from AlphaFold and RoseTTAFold models[328,329].

Table S2 LLPS Propensity for TraG Constructs from catGRANULE. The LLPS propensity for each TraG construct is determined by catGRANULE proteome distribution scores, where scores above 1 indicate the protein is predicted to form LLPS granules when expressed in yeast[334].

	F	**R100**
TraG	0.899	0.940
TraG*	1.240	1.280
TraG*$_{497\text{-}940}$	1.090	1.140
TraG*$_{497\text{-}863}$	0.861	0.900

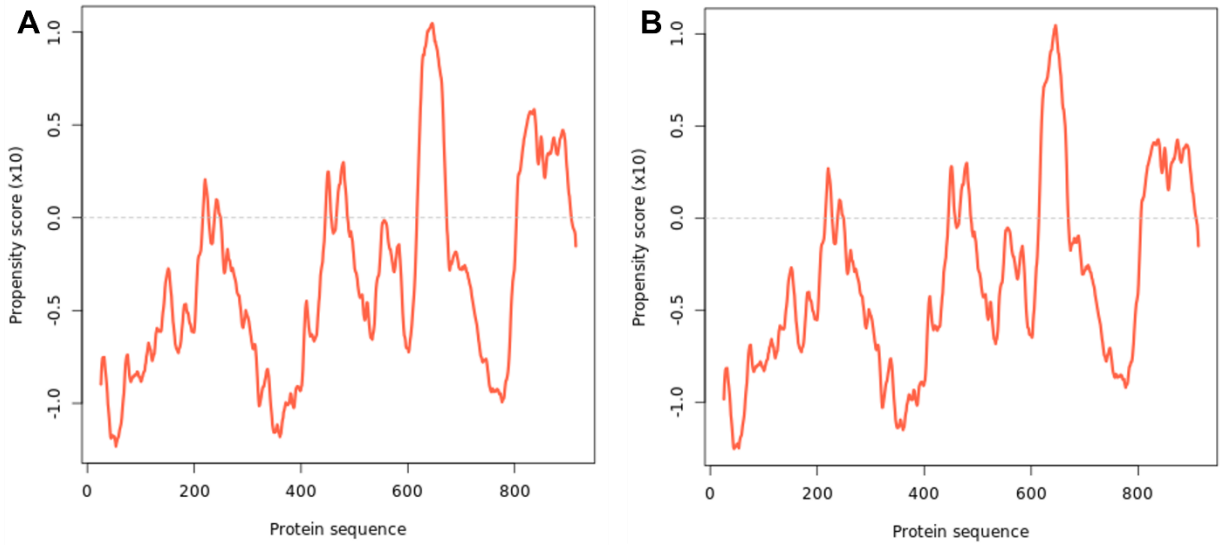

Figure S5 LLPS Propensity Profiles for TraG Constructs from catGRANULE for **A**) $TraG_{R100}$ and **B**) $TraG_F^{334}$. Overall scores are reported in **Table S2**.

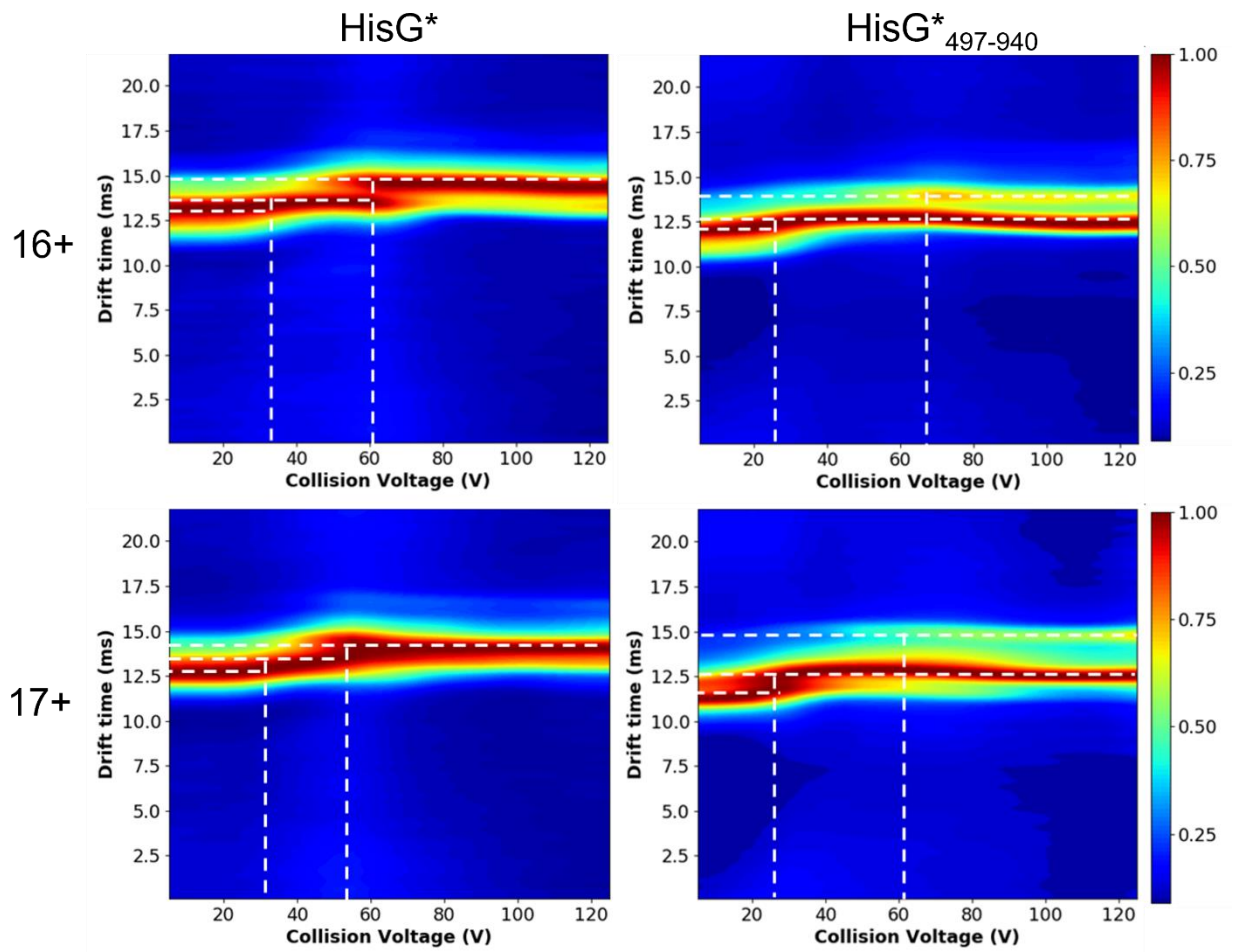

Figure S6 Replicate of CIU-MS Heat Maps of 16+ and 17+ Charge States from TraG* Constructs, HisG* and HisG*$_{497\text{-}940}$ as trap CE is increased from 5 V to 125 V. These maps were produced from biological replicates of the CIU-MS experiments seen in **Figure 2.8** and were plotted with CIUSuite2[369] with smoothing performed using default settings. The m/z range set for plotting the IMS of the 17+ charge states were 3320-3355 for HisG* and 3040-3090 for HisG*$_{497\text{-}940}$, and the m/z range used for the 16+ charge states were 3525-3575 for HisG* and 3230-3280 for HisG*$_{497\text{-}940}$.

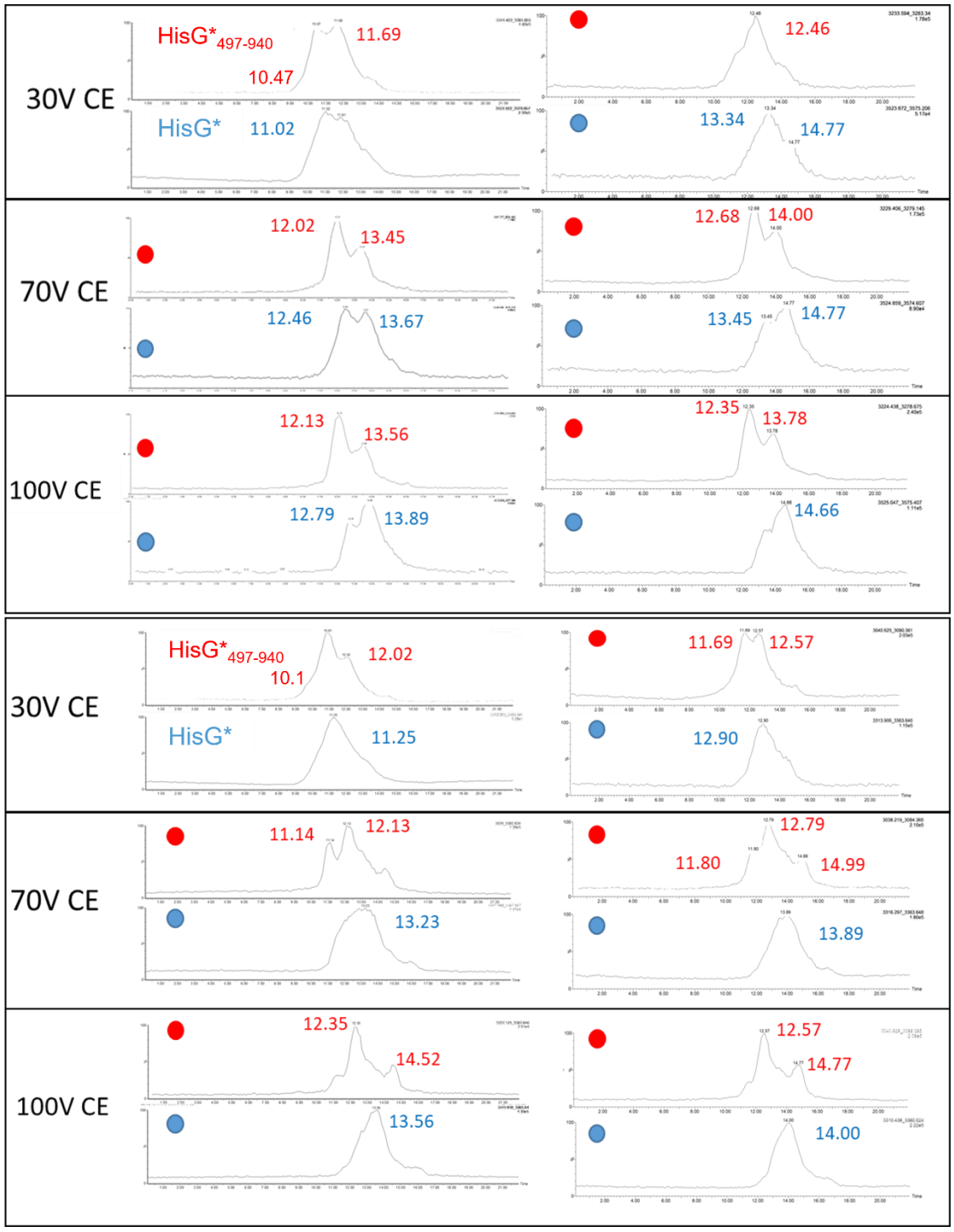
16+
30V CE
HisG*497-940
11.69
10.47
HisG*
11.02
12.46
13.34
14.77
70V CE
12.02
13.45
12.46
13.67
12.68
14.00
13.45
14.77
100V CE
12.13
13.56
12.79
13.89
12.35
13.78
14.66
17+
30V CE
HisG*497-940
12.02
10.1
HisG*
11.25
11.69
12.57
12.90
70V CE
11.14
12.13
13.23
12.79
11.80
14.99
13.89
100V CE
12.35
14.52
13.56
12.57
14.77
14.00

Figure S7 IMS spectra of **A**) 16+ and **B**) 17+ Charge States from TraG* Constructs, HisG* and HisG*$_{497\text{-}940}$, at trap collision energies of 30 V, 70 V and 100 V, gathered on the Synapt G2. Each IMS graph shows drift time on the x-axis with drift times for species with peak abundance labeled, and the biological replicate of each experiment immediately to the right. The 16+ IMS maps were obtained from 3525-3575 m/z for HisG* and 3230-3280 m/z for HisG*$_{497\text{-}940}$. The m/z range set for plotting the IMS of the 17+ charge states were 3320-3355 m/z for HisG* and 3040-3090 m/z for HisG*$_{497\text{-}940}$.

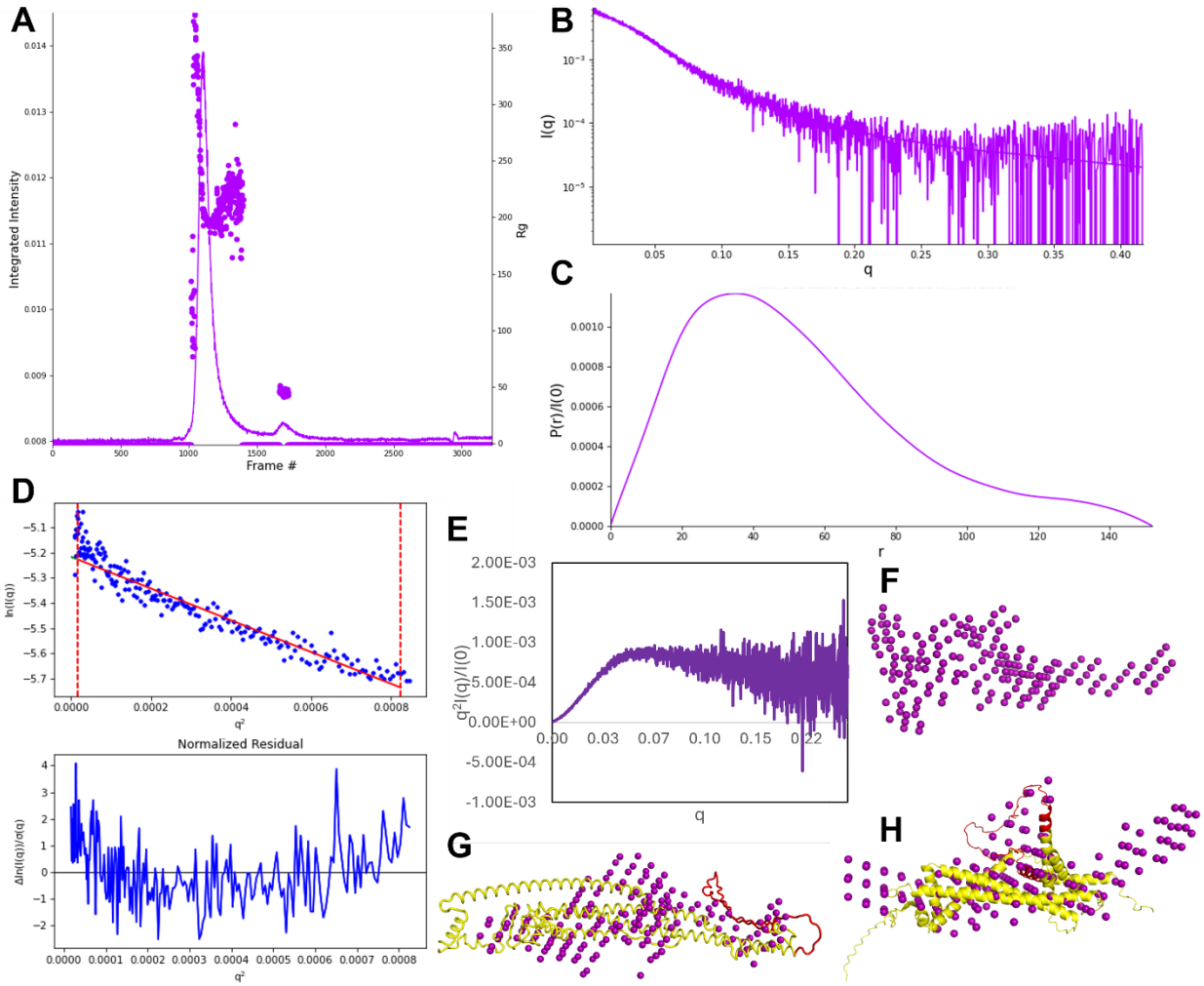

Figure S8 SEC-MALS-SAXS Data of HisG*$_{F\ 497\text{-}863}$. **A**) The SEC-SAXS chromatogram shows aggregation from frames 1010-1390 and the monomer from frames 1660-1740. **B**) The scattering profile and **C**) the P(r) function plot was processed from the data as described in **Section A.4** and **Section 2.3.5**, respectively. **D**) The Guinier plot with the normalized residual plot below it both indicate the HisG*$_{F\ 497\text{-}863}$ aggregate is affecting the monomer data[343]. **E**) The Kratky plot of HisG*$_{F\ 497\text{-}863}$ has an initial gaussian shape which plateaus at high q; the protein is prolate with some unfolding, and noise present in the plot indicates the data is obscured by aggregate. **F**) DAMMIF *ab initio* bead model, **G**) DAMMIF bead model to RF-predicted structure and **H**) DAMMIF bead model to AF-predicted structure, where bead models are in purple and predicted structures aligned to the bead models are yellow with the interacting loop regions coloured red. Bead model images were made with PyMOL v2.5.0[426]. See **Table 2.5** for statistics for the accuracy of the bead models.

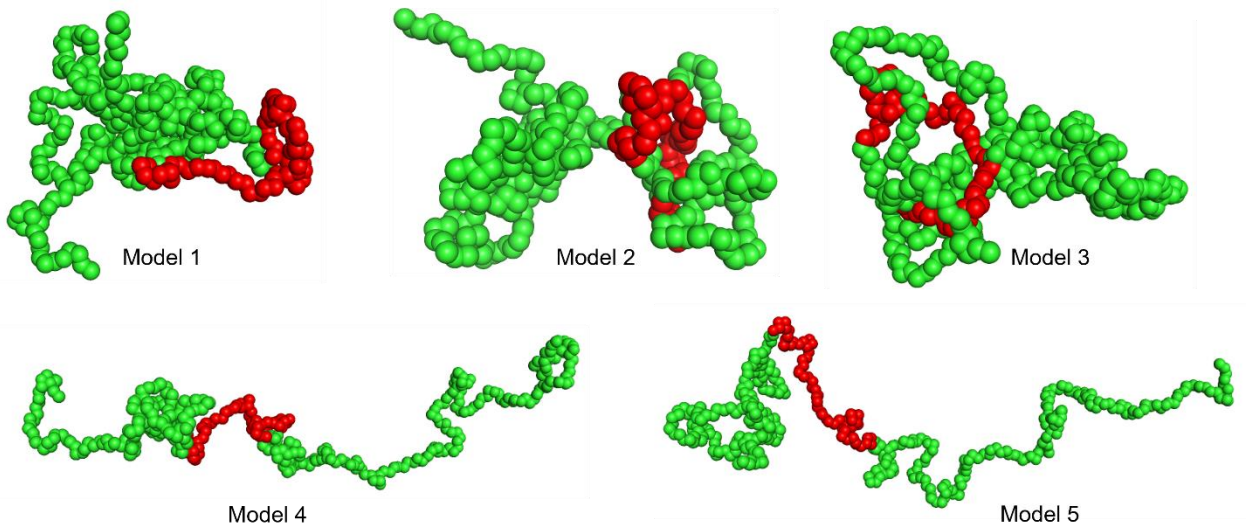

Figure S9 EOM Ensembles of HisG*$_{F\ 497\text{-}863}$ Models. The selected ensemble of models which best describe the conformations achieved by HisG*$_{F\ 497\text{-}863}$, as generated by EOM from SEC-MALS-SAXS data[348]. Models are displayed with residues as spheres, coloured with the interacting loop region in red and the remainder of the protein in green. Images were generated using PyMOL v2.5.0[426], with sphere radius set to 2.0. Parameters for the EOM models are shown in **Table 2.6**.

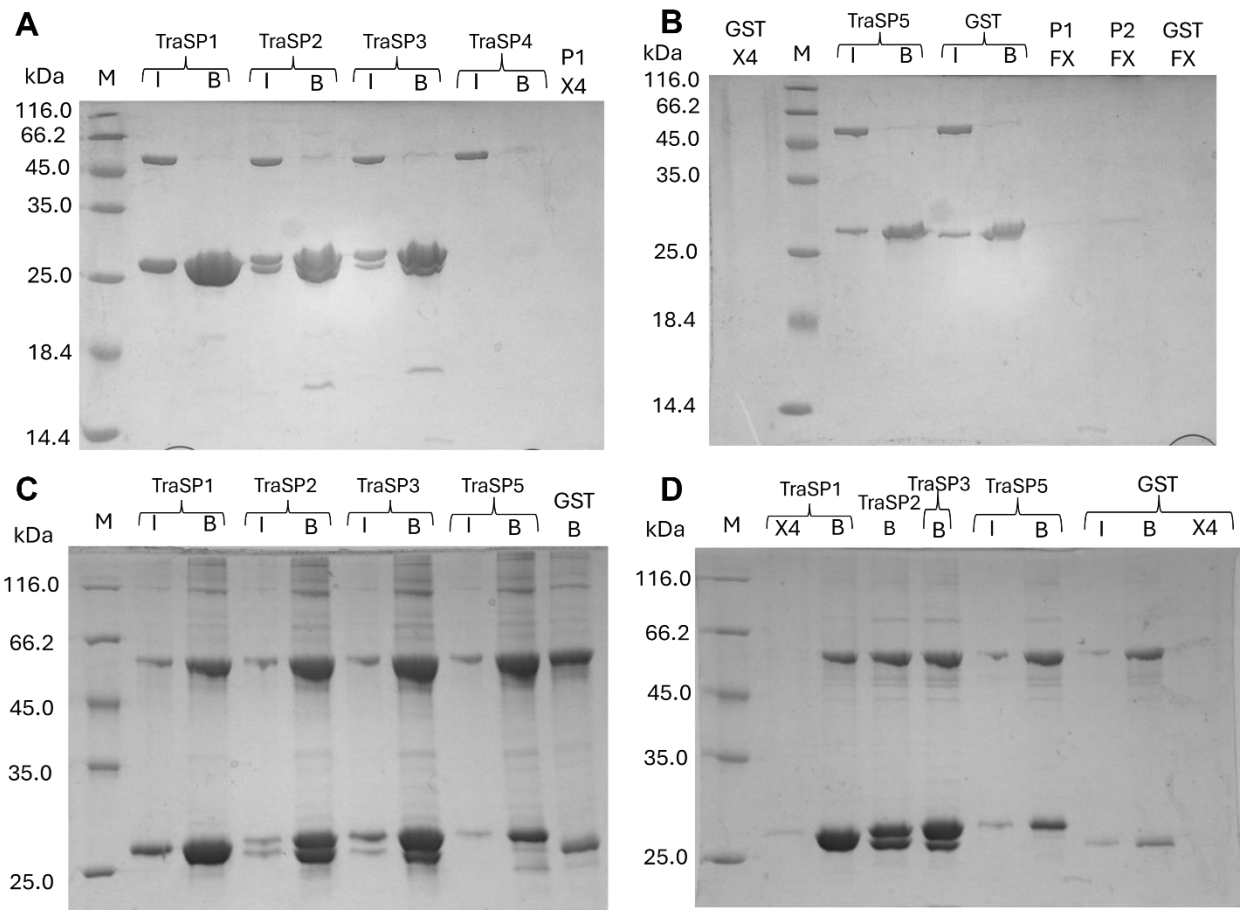

Figure S10 Pulldown Experiments Involving GST-TraS$_F$ Peptides and HisG*$_F$ Constructs. **A**) and **B**) are from GST pulldowns of HisG*$_{F\ 497\text{-}938}$, where samples were prepared as described in **Sections 2.2.7 & 2.2.4** and loaded on a 20% SDS PAGE gel before electrophoresis was performed at 180 V for 72 mins. M: unstained protein marker (ThermoFisher #26610), each GST-tagged TraS peptide that was bound to GSH-agarose beads is labeled above, as well as an un-conjugated GST control, where I: 2 µL of initial slurry after adding HisG*$_{F\ 497\text{-}863}$ to the GST-TraS-bound beads prior to pulldown incubation, and B: 7 µL of pulldown bead sample after all wash steps were performed, P1 X4, GST X4: 20 µL supernatant after fourth wash to remove excess GSTTraS$_{P1}$ and GST control, respectively, during the binding step. P1 FX, P2 FX, GST FX: 20 µL supernatant after last wash to remove excess HisG*$_{F\ 497\text{-}938}$ after the pulldown step for GSTTraS$_{P1}$, GSTTraS$_{P2}$, and GST control, respectively. Nickel-IMAC pulldowns of GST-TraS$_F$ peptides, where the bait was **C**) HisG*$_F$ and **D**) HisG*$_{F\ 497\text{-}938}$ was performed, and samples were loaded on 15% SDS PAGE gels; electrophoresis was performed at 180 V for 57 mins. Gel labelling is the same as described above, except the His-tagged TraG proteins were bound to Ni-NTA agarose beads and the labelling therefore refers to the GST-TraS or control GST added as part of the pulldown step. I: 2 µL of initial slurry after adding each GST-tagged TraS peptide and the GST control to the HisG*-bound beads, prior to pulldown incubation, and B: 10 µL of pulldown bead sample after all wash steps were performed.