From Peptide Oligomers to Single-Chain Proteins

Debbie Nicol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of PhD in the Faculty of Science.

School of Chemistry, February 2022

37,351 words

33,825 words excluding references

Abstract

If we truly understand the relationship between protein sequence and structure, we can solve the protein folding problem and, in the process, realise fully *de novo* proteins — that is, proteins that do not exist in nature. As fold determines function, *de novo* folds could yield novel functions, or simply offer the opportunity to probe known functions within new scaffolds.

The facile synthesis, hyper-thermostability, and tuneable pore size of type II α-helical coiled coils, also known as α-helical barrels (αHBs), predispose these quaternary structures to later-stage installation of function. However, the inherent symmetry of homomeric coiled coils may impact their utility. The Woolfson group have previously described the installation of rudimentary hydrolytic activity into a homomeric, seven-helix coiled coil barrel. Unfortunately, its efficiency pales in comparison to the natural hydrolase α-chymotrypsin. Computational transition state modelling suggests that the placement of the catalytic residues within the barrel core could be optimised by spreading them across adjacent helices. However, this cannot be achieved in the current homomeric aHB systems. As such, more sophisticated coiled coil scaffolds in which mutations can be made with pinpoint precision are sorely needed.

One possible route towards de-symmetrizing coiled coils is the construction of single-chain proteins with a coiled-coil core. We posit that controlled oligomerisation of polypeptide motifs with tertiary structure *e.g.* helix-loop-helix motifs will allow access to such *dark matter* proteins. The monomeric PPα miniprotein comprises a three-heptad α helix buttressed by an antiparallel polyproline type-II (PPII) helix, and provides a plausible template for single-chain parallel aHBs.

In chapter 3, we show that an optimised PPα miniprotein can be oligomerised in a controlled manner by embedding designed coiled-coil interfaces into the solvent-exposed face of the α helix of the miniprotein, resulting in thermostable helical bundles and barrels. An attempt is made to describe these PPα-based oligomers parametrically in chapter 4. Possible connecting loop sequences from protein structures deposited in the Protein Data Bank are explored, with a view to generating sequences for single-chain proteins.

With a solid footing in the design of PPα-based assemblies, we examine the recombinant expression of 'extended' PPα oligomers (specifically, oligomers of α-L-PPα motifs, where L is a flexible linker) and single-chain parallel bundles and barrels in chapter 5. Whilst only a modicum of success is achieved regarding the expression and characterisation of the oligomers, there is more success with the designed α-helical bundles. Two bundle designs express well, and biophysical characterisation confirms that they are thermostable and monomeric in solution, as designed. However, no expression is observed in the case of the barrels comprising five, six and seven α helices in a common expression strain of *E.coli*. However, preliminary experiments conducted with Lemo21 (DE3) *E. coli* show promise for the expression of single-chain α-helical barrels.

In summary, this thesis describes the successful oligomerisation of the optimised PPα miniprotein, and the utility of these PPα-based oligomers as templates for single-chain α-helical bundles and barrels. These *de novo* scaffolds possess great potential for catalysis and other biotechnological applications.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: 18/02/2022

Acknowledgements

Firstly, thank you to Dek, for the warm welcome to the group for a lab rotation, and for planting an enthusiasm for peptide design so strong that it brought me back for an entire PhD. Thank you for all the subsequent guidance, encouragement, and support that got me from start to finish.

Next, thank you to Emily, for starting this project and taking me under your wing in the early days. I owe a lot to you. I can't begin to describe how grateful I am to have met and worked with you. Thank you to Will, for your help with molecular biology and everything else, and for answering all the silly questions I asked you. You are a saint. Thanks again to you both, for proofreading this thesis. Thank you to Chris, for your help with ISAMBARD. Thanks to Jess, for all the joy you have brought to my time in the group. I couldn't ask for a better lab BFF. Thanks to all the members of the Woolfson lab, past and present, for making my time here so wonderful.

Many thanks to the Bristol Chemical Synthesis CDT for the opportunity to do a PhD, the EPSRC for funding, and to the 2016 cohort (especially Jenny, Beth, and Dan G.), for making it so much fun.

Thanks to my parents, for supporting me despite not having a clue what it is I do, and for always doing that. Thanks to Nyla, for being my best friend and biggest champion.

Finally, thanks to Rob, for absolutely everything.

Abbreviations

αHB	α-Helical barrel
Agg.	Aggregated
AMPAL	Atom, monomer, polymer, assembly, ligand
AUC	Analytical ultracentrifugation
BUDE	Bristol University docking engine
BUFF	BUDE forcefield
CC	Coiled coil
CCD	Cyclic coordinate descent
CD	Circular dichroism
CH_2Cl_2	Dichloromethane
Cl-HOBt	1-hydroxy-6-chloro-benzotriazole
DE	Differential evolution
DIC	*N,N*'-diisopropylcarbodiimide
DHB	2,5-dihydroxybenzoic acid
DMF	*N,N*-dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPH	1,6-Diphenyl-1,3,5-hexatriene
DSSP	Dictionary of secondary structure of proteins
E. coli	Escherichia coli
EtOH	Ethanol
Exp.	Expected
Fmoc	Fluorenylmethoxycarbonyl
GA	Genetic algorithm
h	Hydrophobic
H_2O	Water
H-bond	Hydrogen bond

HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HLH	Helix-loop-helix
HPLC	High performance liquid chromatography
HT	High tension
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ISAMBARD	Intelligent system for analysis, model building and rational design
K_d	Dissociation constant
KIC	Kinematic closure
KIH	Knobs-into-holes
LB	Lysogeny broth
MALDI-TOF	Matrix assisted laser desorption/ionisation-time of flight
MeCN	Acetonitrile
MRE	Mean residue ellipticity
MS	Mass spectrometry
NaCl	Sodium chloride
NCI	Non-covalent interaction
ND	Not determined
NMR	Nuclear magnetic resonance
Obs.	Observed
OD_{600}	Optical density at 600 nm
O.S.	Oligomeric state
p	Polar
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate-buffered saline
PDB	Protein Data Bank
pI	Isoelectric point
PMSF	Phenylmethane sulfonyl fluoride
POI	Protein of interest

PP	Pancreatic polypeptide
PPII	Polyproline-II
RMSD	Root mean square deviation
RPR	Rise per residue
RPT	Residues per turn
sc	Single-chain
SDS	Sodium dodecyl sulfate
SE	Sedimentation equilibrium
SEC	Size exclusion chromatography
SSE	Secondary structure element
SV	Sedimentation velocity
TBC	To be confirmed
TBS	Tris-buffered saline
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
TIM	Triose-phosphate isomerase
TIPS	Triisopropylsilane
T_M	Thermal midpoint temperature of unfolding
TPR	Tetratricopeptide repeat
UV	Ultraviolet
$\bar{v}$	Partial specific volume

Amino acids

Alanine (Ala, A), arginine (Arg, R), asparagine (Asn, N), aspartic acid (Asp, D), cysteine (Cys, C), glutamic acid (Glu, E), glutamine (Gln, Q), glycine (Gly, G), histidine (His, H), isoleucine (Ile, I), leucine (Leu, L), lysine (Lys, K), methionine (Met, M), phenylalanine (Phe, F), proline (Pro, P), serine (Ser, S), threonine (Thr, T), tryptophan (Trp, W), tyrosine (Tyr, Y), valine (Val, V).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Protein Folding and Protein Design Problems

1.1.1 The Protein Folding Problem

The central dogma in biology describes the process by which programmed genetic information is transformed into protein primary sequence. Anfinsen's dogma states that, under standard conditions, a protein's sequence dictates its structure,[1] and, furthermore, its physiological function. What sounds like a simple relationship, however, is in fact anything but. Despite recent advances in protein structure prediction algorithms,[2, 3] the so-called *protein-folding problem* — that is, the manner by which sequence dictates structure — remains unsolved.[4, 5]

The roots of the protein-folding problem date back to the early 20th century, when Kendrew and Perutz determined the first X-ray crystal structures of the globular proteins myoglobin and haemoglobin, respectively,[6-8] for which they were jointly awarded the 1962 Nobel Prize in Chemistry. The structures of these globular proteins were composed of regions of α-helical secondary structure, first proposed by Pauling in 1951.[9] However, the irregular

arrangement of the helices and the overall asymmetry of the protein structures obtained proved baffling: at the time, the structure of myoglobin was “more complicated than has been predicted by any theory of protein structure.”[6] This, in conjunction with Anfinsen’s theory that the native state of a protein must be the most thermodynamically favoured conformation, and the successful phasing of an ever-increasing volume of protein crystal X-ray data led to Cyrus Levinthal questioning the process by which proteins ultimately adopt their native conformation.[10] Levinthal was particularly interested in the short timescale of spontaneous protein folding, and noted that this appeared at odds with the astronomical number of degrees of freedom in a large protein,[11] a problem now referred to as Levinthal’s paradox. Levinthal concluded that a polypeptide chain could not arrive at its most thermodynamically stable state by trial and error alone. Instead, spontaneous protein folding must be directed somehow — through a protein folding pathway, for example.

A ‘new view’ of protein folding appeared in the literature in the 1990s.[12-14] Instead of envisaging protein folding as an open, flat field in which a lone, deep well is located (figure 1.1A), imagining a landscape with rolling hills, gullies and valleys (energy barriers, downward slopes towards negative free energies, kinetic traps, and the thermodynamic energy minimum) sits more comfortably with the realities of spontaneous protein folding (figure 1.1B). By this analogy, there are many possible routes to a protein’s optimal conformation, and it demonstrates that a given protein in different states of denaturation can fold to the same native state.[15] It is also easy to imagine how certain proteins could misfold given the opportunity, if the barrier between the native state’s energetic minimum and the next lowest energy minimum is shallow. However, not all proteins linger in their native state. Indeed, the serpins are an example of biologically active proteins that are kinetically trapped.[16]

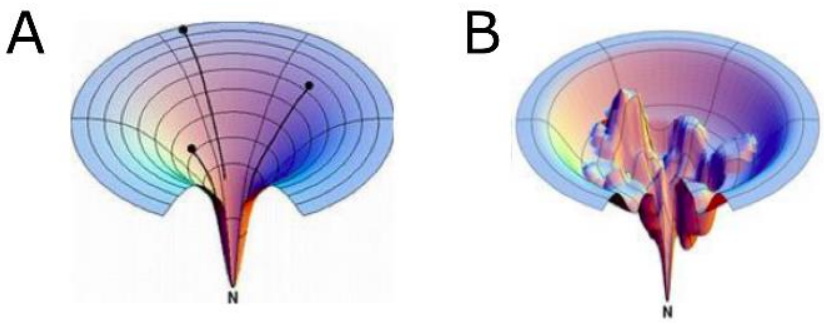


Figure 1.1: *Examples of protein energy landscapes.* ***A****, An idealised smooth funnel.* ***B****, A rugged funnel with kinetic traps. Adapted from an image by Ken Dill sourced on Wikimedia Commons,[17] covered by a Creative Commons Attribution 4.0 International license.[18] N indicates the position of the native state.*

However, the route to the native state is not quite as easy as "funnelling wine into a bottle", as Karplus states.[19] The simplistic nature of the funnel diagram, the two-dimensional precursor to the three-dimensional energy landscape, often allows misinterpretations regarding the difficulty of protein folding — the funnel itself does not aid or guide the process. The free energy (ΔG) of spontaneous protein folding is the sum of the potential energy (ΔH) and the configuration entropy (ΔS).[19] The former decreases on nearing the native state and favours folding; the latter also decreases upon approaching the native state, and is unfavourable in this context.

1.1.2 The Ramachandran Plot and Secondary Structure

The Ramachandran plot, first published by G. N. Ramachandran and colleagues in 1963,[20] is a map of energetically favourable conformations of the dihedral angles ϕ and ψ (figure 1.2A) of the polypeptide backbone. A further dihedral angle, ω, is largely ignored due the pseudo double bond character of the peptide bond; in this case, ω is restricted to 180°.

In order to achieve regular secondary structure, the same ϕ,ψ angles must be repeated across a given number of residues. The most populated regions of the Ramachandran map (figure 1.2B) depict the regions corresponding to secondary structure elements. The α and β regions (parameters defined in table 1.1) are the most densely populated, although β space is shallower and more widespread compared to α space. Smaller areas of the Ramachandran map are attributed to left-handed α, 3_{10}, π, and polyproline helices (parameters defined in table 1.1).

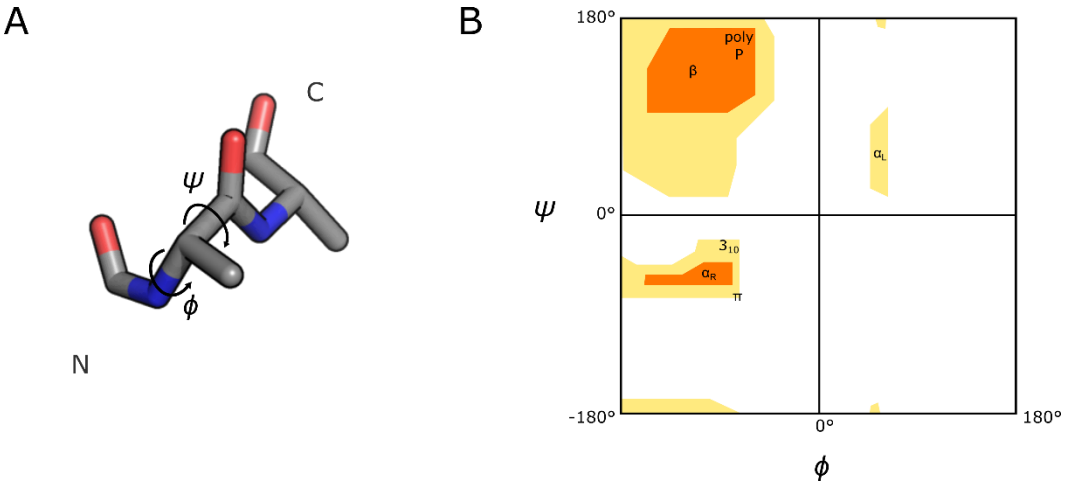


Figure 1.2: *Main chain dihedral angles and the Ramachandran plot.* ***A****, Example of a polypeptide backbone, showing the dihedral angles phi (ϕ) and psi (ψ).* ***B****, Simplified map of Ramachandran space, showing allowed main chain dihedral angles phi and psi for amino acids possessing a C_β atom i.e. all proteinogenic acids excluding Gly. Favoured regions are shown in dark orange, allowed regions in pale orange, and disfavoured regions in white. Regions are labelled according to the secondary structures which populate them. α_R = right-handed α; α_L = left-handed α; poly P = polyproline.*

Table 1.1: *Examples of helix-defining parameters for selected secondary structural elements.*

SSE	RPT	RPR (Å)	Radius (Å)
α helix	+ 3.6	1.5	2.3
3_{10} helix	+ 3.0	2.0	1.9
π helix	+ 4.3	1.1	2.8
β strand	- 2.3	3.3	1.0
PPII helix	- 3.0	3.1	1.6

SSE = secondary structure element, RPT = residues per turn, and RPR = rise per residue, PPII = polyproline II.

The majority of the SSEs outlined above are dictated by repeating patterns of intramolecular hydrogen bonds. For example, the backbone of the α helix is defined by *i→i+4* hydrogen bonds between a carbonyl oxygen (O_i) and an amide nitrogen (N_{i+4}) that is situated four residues down the polypeptide chain. The 3_{10} helix is so named as it has three residues per turn and ten atoms in its hydrogen bond pattern, between a carbonyl oxygen and an amide nitrogen three residues down the polypeptide chain (*i→i+3* hydrogen bonding). The majority of 3_{10} helices published in the literature contain non-proteinogenic α-alkylated amino acids, predominantly α-aminoisobutyric acid, which possess two methyl groups on the C_α atom.[21] The π helix is a rare secondary structure motif possessing an *i→i+5* hydrogen bonding pattern.[22]

Whilst β sheets are satisfied by intermolecular hydrogen bonds between individual β strands (at least two of which can adopt parallel or antiparallel orientations, each with their own distinctive hydrogen bond patterns), the β strand itself is not stabilised by local hydrogen bonds. Similarly, polyproline helices are not defined by local hydrogen bond patterns. Polyproline-I helices, composed of *cis*-Pro residues, are rarer than polyproline-II (PPII) helices, in which all peptide bonds adopt a *trans* conformation. The PPII

helix is typified by a ProXxxXxx residue repeat. However, the presence of Pro is not always necessary for its formation.[23, 24]

1.1.3 The Role of Non-Covalent Interactions

Folded proteins are stabilised by the interplay of multiple defined non-covalent interactions (NCIs).[25, 26] Individually, NCIs offer little in the way of energetic compensation for the entropically expensive process of protein folding. However, these weak interactions are both accumulative and cooperative, and their individual contribution to protein stability is often difficult to prise apart.

Hydrogen bonds (H-bonds) between backbone amide NH and carbonyl oxygen atoms, electrostatic interactions, van der Waals interactions, and the hydrophobic effect all contribute significantly to protein folding and stability.[25] H-bonds perhaps most so, as regular patterns of H-bonds are responsible for defining and stabilising the majority of secondary structure elements (SSEs). Electrostatic interactions between non-polar residues of opposing charges are capable of forming stabilising interactions such as salt bridges (≤ 4 Å).[27] However, intramolecular electrostatic interactions are predominantly restricted to the surface, where charged residues can be easily solvated. The main driver of protein folding, however, is widely attributed to van der Waals interactions. After the determination of the structure of myoglobin, Kendrew stated: "… it is clear that by far the most important contribution comes from the van der Waals forces between non-polar residues that make up the bulk of the interior of the molecule."[28] Over the years, the contribution of a number of additional, weaker NCIs — CH···O H-bonds,[29, 30] n→π* interactions,[31] C5 H-bonds,[32] π→π* interactions,[33-35] anion/cation-π interactions,[36, 37] and CH-π interactions,[38, 39] for example — to stabilising protein folds have been studied.

1.1.4 The Protein Design Problem

The inverse of the protein-folding problem — *i.e.*, finding an amino acid sequence that will adopt a pre-defined protein structure — can also be referred to as the *protein-design problem.* The aims of protein design are many but can be reduced to the following two basic desires: to better understand biology, and to improve upon biology. In brief, protein designers aim to enrich natural proteins with novel function *via* new sequences, or to provide novel function within folds not seen in nature in order to solve biological problems. The number of possible non-natural proteins (sometimes referred to as *never born proteins* or *fully de novo proteins*, which can be thought to populate the *dark matter of protein space*) is incomprehensibly vast, even if only working within the confines of proteinogenic amino acids.[40-42]

1.2 Approaches to Peptide and Protein Design

The peptide and protein design fields are much too vast to cover here in sufficient depth. Instead, the focus will primarily be on *rational* and *computational de novo* peptide and protein design. Both of these terms refer to the design of new sequences and/or polypeptide backbones *i.e.* novel protein folds, occasionally inspired by naturally-occurring peptide assemblies or protein structures. The boundary between these two approaches is blurry at best, and there is often significant overlap. In line with a recently published history of protein design,[43] we define rational and computational design as follows:

Rational protein design embodies the straightforward hydrophobic/polar or binary patterning of minimal protein design,[43-45] but also incorporates understood sequence-to-structure relationships that have been learned from nature. Bioinformatics has informed rational design approaches greatly, for example, in the extraction of amino acid propensities for certain

helix types from the PDB.[46-48] Rational protein design can also be considered to include consensus design, which makes use of information contained in the sequences of multiple homologous protein structures, which have been conserved over the course of evolution.[49]

Computational protein design, on the other hand, most often refers to the generation of full atomistic protein models (built either from protein fragments or parametrically) for which stabilising sequences are subsequently scored. This definition of computational protein design encompasses the essence of the protein-folding problem *i.e.* where 3D structures are predicted for a given sequence, as structure prediction often plays a role in the sequence optimisation that is part and parcel of modern computational design.

The field of rational protein design surfaced between the late 1980s and early 1990s with attempts to design four-helix bundles. Regan and DeGrado utilised a minimal design approach with a small palette of amino acids (with Leu at all hydrophobic sites, and Glu and Lys at all hydrophilic sites),[50] whereas Hecht *et al.* endeavoured for a less repetitive but still novel sequence that adopted the same fold.[51]

Extensive success in the *de novo* design of coiled coils, assemblies of two or more α helices, which are dictated by specific patterns of hydrophobic and polar residues, has been achieved using a rational approach. Early examples include efforts from DeGrado, who reported an antiparallel trimer,[52] and Alber, with a parallel, heterotrimeric coiled-coil design.[53] Later examples include the design of coiled-coil heterodimers,[54, 55] a pH-sensitive coiled-coil switch,[56] an antiparallel tetramer,[57] artificial metalloenzymes,[58, 59] predictable oligomerising coiled coils for use in synthetic biology applications,[60] and the design of peptide cages.[61] Coiled-coil design is covered in more detail in section 1.3.

Larger repeat proteins based on α-helical building blocks have been designed utilising a consensus design approach.[62] Kohl *et al.* created a library of Ankyrin repeat proteins (a 33-residue repeat sequence comprising a β turn, two antiparallel α helices and a loop) to inform their Ankyrin repeat protein design efforts.[63] The design rationale rested on the idea that all Ankyrin repeats belong to a canonical ensemble, the repeat unit of which can be described by the consensus sequence. Multiple Ankyrin repeat proteins possessing a varying number of repeats were expressed, and structures of two proteins composed of five repeats were obtained *via* X-ray crystallography. Tetratricopeptide repeat (TPR) proteins, defined by a 34-residue α-helical repeat) have also been designed by consensus. In nature, TPR proteins rarely contain fewer than three repeats. To determine whether three repeats is the minimum required to stabilise this fold, Main *et al.* utilised consensus design to achieve TPR proteins consisting of one, two or three repeats.[64] Whilst all of the resulting proteins behaved well *in vitro*, it was found that, in order to function in protein-protein interactions, a minimum of three repeats is necessary — not for stability, but for the construction of a peptide binding site.

Whilst the *de novo* design of α-helical structures was increasingly reported on, success in β-structure design advanced more sluggishly. In 1996, the Serrano group were amongst the first to tackle the challenge of β-hairpin design. The β hairpin is the simplest all-β-tertiary structure motif observed in natural proteins, and its formation is proposed to be the 'nucleation' event that precedes β-sheet formation.[65] The designed eight-residue hairpin (from *N* to *C* terminus, -B3, -B2, -B1, L1, L2, +B1, +B2, +B3, where B represents a strand residue and L a turn residue) was based on the most abundant hairpin structures observed in the protein structure database at the time.[66] Thr was placed at -B2 and +B2 positions, due to their high propensity for β sheet formation. Asn and Gly were placed at positions L1 and L2, respectively, as these residues were found to be the most favourable

combination for a type I' β turn during a search of the database. The sequence was completed based on statistical preferences, fixing Val at -B1, Ile at -B3, Lys at +B1, and Tyr at +B3, to yield peptide BH8, RGITVNGKTYGR. Arg residues were added to both termini to prevent lateral oligomerisation of the hairpin; Gly was added between Arg and the remainder of the BH8 sequence as a spacer. BH8 is an interesting design, as it eschews the strict alternating hydrophobic/polar patterning typically observed in β structure in an attempt to suppress β-sheet formation and stacking. Circular dichroism spectroscopy revealed BH8 possessed a minimum at 216 nm, indicative of β structure, and subsequent NMR experiments provided evidence of β hairpin structure.

Successful three-stranded antiparallel β sheet proteins were published in 1998 by the Serrano and Gellman groups. Kortemme *et al*. took an iterative, hierarchical approach, initially studying β-hairpin stability to inform their β sheet designs.[67] Whilst this initial design was monomeric in solution up to 3 mM, NMR studies only detailed structure in the region of the sequence corresponding to the second and third β strands. The subsequent two designs, differing from the first primarily in their inter-strand packing, behaved similarly, suggesting that the buried hydrophobic surface area between the first and second β strands was insufficient to drive β sheet formation. Examining different rotamers of aromatic residues in the first β strand and a non-β-branched residue at the corresponding position in the second strand (akin to the arrangements found in WW domains) seemed favourable. Incorporating Trp into the sequence of the initial sheet design resulted in Betanova, a stable *de novo* three-stranded β sheet. Schenck and Gellman explored β-sheet cooperativity with their three-stranded antiparallel β sheet design, which incorporated D-Pro-Gly loops to dictate the necessary β turns (as opposed to the AsnGly loops favoured by Kortemme *et al*).[68] The choice to include the D-Pro-Gly loops circumnavigated control issues posed by other β turn loops, as the

introduction of L-Pro at the same position should fatally disrupt the structure of the β sheet — a theory that held up to experimental scrutiny.

Notable efforts in the development of algorithm-aided redesign and *de novo* design appeared in the literature between 1980 and 2000. In 1987, Ponder and Richards used an algorithm to repack internal, fold-defining residues in natural proteins using rotamer libraries.[69] In 1995, Desjarlais and Handel combined custom rotamer libraries and a global optimisation algorithm to design the hydrophobic core of the phage 434 cro protein.[70] Dahiyat and Mayo produced a novel sequence for a ββα motif based on a natural zinc finger domain in 1997.[71]

Parametric protein design, in which protein structures are described mathematically *in silico*, is a useful tool for generating the fixed polypeptide backbones required to begin computational design or redesign efforts. In 1998, Harbury *et al.* utilised parametric design to produce a number of coiled coils with a right-handed super-helical structure.[72] Many other repetitive and highly α-helical proteins can readily be described using a small number of parameters. In addition to coiled-coil bundles, α-helical barrels, α-solenoids, and α/α-toroids can all be described parametrically, for example.[73-75]

In recent years, web-based, user-friendly coiled-coil modelling tools have been released by the Woolfson group. CCBuilder was released in 2014,[76] and updated in 2018.[77] The latest iteration is powered by ISAMBARD, parametric design software for building and optimising biomolecular models first released in 2017.[78] However, working only with a reduced set of structure-describing parameters effectively limits the scope of ‘designable’ polypeptide backbones to highly symmetric folds. To advance the field of protein design, a new approach was needed to target possible asymmetric protein backbones.

In 2003, the repurposing of protein structure prediction software to tackle the challenge of *de novo* protein design caused a seismic shift in the field. Researchers from the Baker group published the first account of a *de novo* designed globular protein with an architecture unlike any protein structure in the Protein Data Bank or the SCOP (Structural Classification of Proteins) database.[79] The initial fragment-based design was subjected to iterative rounds of sequence optimisation and structure prediction. The 93-residue mixed α/β protein, named Top7 (figure 1.3), was found to be highly stable in solution, and the experimentally determined structure matched the design model closely (1.17 Å RMSD between the design model and the crystal structure).

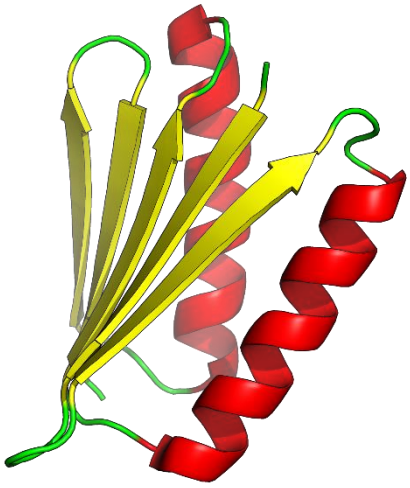


Figure 1.3: *Top7, the first de novo designed globular protein designed by Kuhlman et al.[79] β strands are shown in yellow, α helices in red, and connecting loops in green.*

Since its introduction with the design of Top7, the fragment-based approach remains popular in computational protein design, and the output from the

Rosetta suite of *in silico* modelling tools is particularly impressive. Baker, DeGrado and Kortemme have recently reviewed the successes in this area.[80-82] Over the years, computational methods have been improved by switching from working exclusively with fixed backbones to flexible backbones, scoring functions have been improved, and neural networks have been adopted for *de novo* protein design.[82-84] Protein designers are now focused on encapsulating function in the early stages of the design process as designing novel proteins with functionality in mind from the start allows for more flexibility.[85] Fragment-based design lends itself well in these ventures, as demonstrated by the Correira group's work on TopoBuilder,[86, 87] although some computational design efforts require multiple rounds of directed evolution post-design.

However, despite the advances in fragment-based design towards function, installing function into or re-fitting *de novo* designed proteins allows us to further investigate the dark matter of protein space, whilst working within an already well-understood scaffold. The predictability and versatility of α-helical coiled coils lend this super secondary structure to such exploration.

1.3 α-Helical Coiled Coils

1.3.1 Definition

Coiled coils (CCs) — assemblies of two or more supercoiling amphipathic α helices — are vastly abundant in nature. For every known genome, approximately 3 % of all residues in predicted protein sequences are present in coiled coil motifs,[88] and coiled-coil regions are present in approximately 10 % of all eukaryotic proteins.[89] These super secondary structures are important for facilitating and disrupting protein-protein interactions and also play structural roles at a cellular level.[90] For coiled coils, we possess perhaps the most well-studied — and arguably the best understood — sequence-to-structure relationship in the proteome. Thus,

these predictable and versatile structures are often targeted by protein designers.

Typically the sequences of the α helices in coiled coil show 7-residue or heptad repeats, although larger 11- and 18-residue repeats are known. These 7-residue repeats are usually denoted *abcdefg*, where *a* and *d* mark the hydrophobic (*h*) residues, and *b*, *c*, *e*, *f*, and *g* represent the polar (*p*) residues, giving an *hpphppp* repeat. The folding of α helices and the subsequent formation of coiled coils is, like the folding of most proteins, largely driven by the sequestration of hydrophobic side-chains away from the bulk solvent. The mismatch between the number of residues per turn (3.6) and the average spacing of hydrophobic residues (3.5 residues in the case of an *hpphppp* repeat) results in a hydrophobic seam that twists around the face of the folded α helix. This seam meanders in the opposite direction to the helix backbone (right-handed). The result is that, when two helices pack together to form the simplest coiled-coil dimer, the helices supercoil around one another with a left-handed twist (figure 1.4).

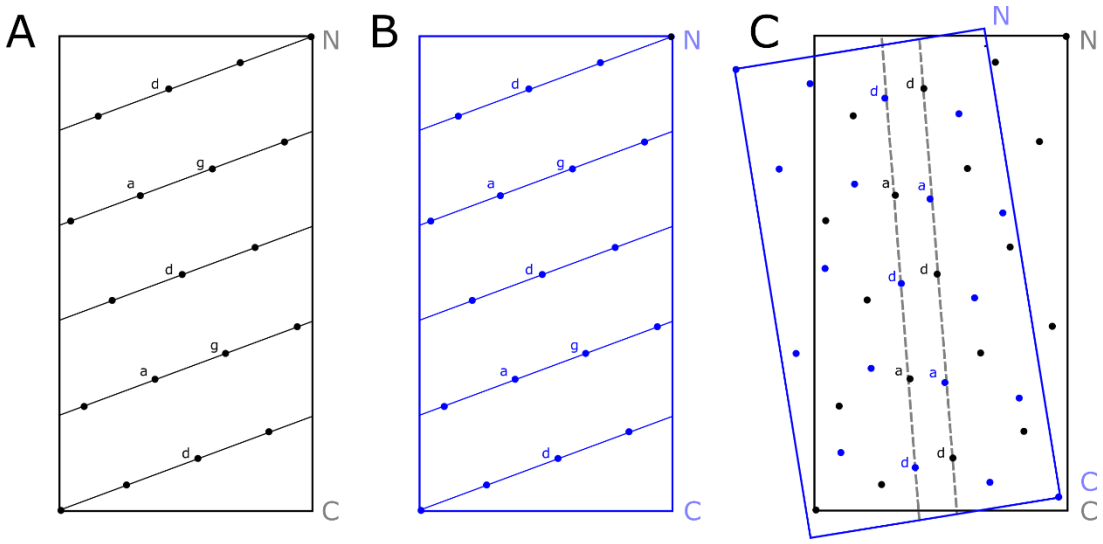


Figure 1.4: *Helical net representations.* ***A*** *and* ***B*** *show helical net representations of two α helices, drawn to scale using the parameters outlined in table 1.1.* ***C*** *shows an overlay of the two helical nets, with dotted lines denoting the interhelical hydrophobic seam formed.*

This *hpphppp* repeat is responsible for the simplest coiled coils (dimers) and is referred to as a type N coiled coil interface (figure 1.5A).[91] Incorporating additional hydrophobic residues into the heptad repeat results in more complex coiled coils with increasing oligomeric states i.e. a larger number of helices in the assembly. An *hpphhpp* repeat yields type I coiled coils (trimers/tetramers); an *hhphhpp* repeat gives type II coiled coils (tetramers to heptamers), and an *hhphphp* repeat gives type III coiled coils (octamers and above) (figure 1.5B, C, and D, respectively). The angle between two hydrophobic faces on one helix increases on moving from type I to type III coiled coils and is known as the offset angle.

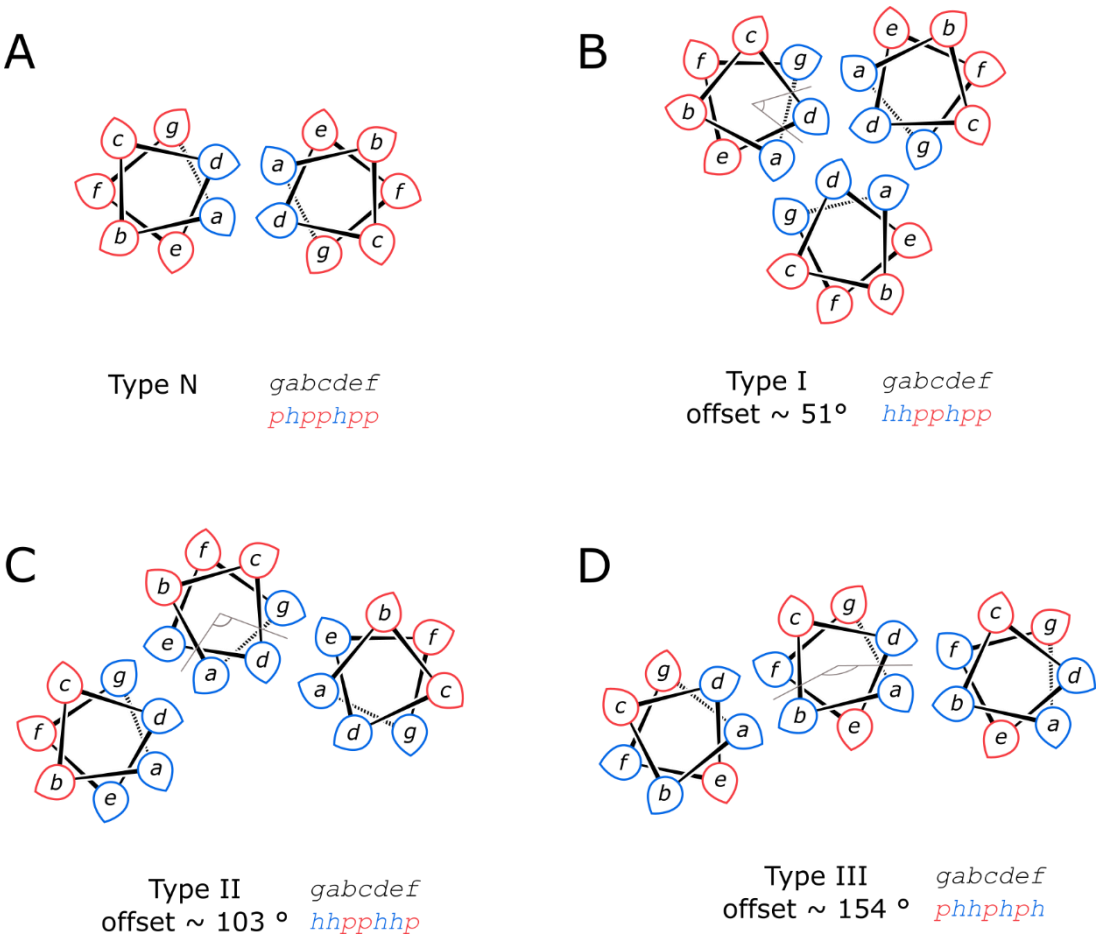


Figure 1.5: *Helical wheel representations of coiled coil interfaces.* ***A****, type N,* ***B****, type I,* ***C****, type II, and* ***D****, type III coiled coil interfaces. Positions of hydrophobic and polar residues are highlighted in blue and red, respectively. The offset angle between the two hydrophobic interfaces on each helix in type I, II and III coiled coils is outlined in grey. Type N and I coiled coils refer to α-helical bundles; type II and II coiled coils refer to α-helical barrels. For type II and III helical wheel representations, some helices from the α-helical barrels have been omitted for clarity. Figure adapted from reference.*[92]

1.3.2 Rules for *De Novo* Coiled Coil Design

In addition to the patterning of the hydrophobic and polar residues in a coiled-coil sequence, the identity of the hydrophobic residue directs oligomer state. Hydrophobic residues can be split into two subgroups: aliphatics and aromatics. Both naturally-occurring and designed coiled coils

tend to favour aliphatic residues at core *a* and *d* sites, although aromatic side-chains can occasionally be tolerated.[93] A seminal mutagenesis study by Harbury *et al*. on the 33-residue GCN4-p1 leucine zipper showed that *a* = Ile, *d* = Leu (GCN4-p-IL) specifies a dimer; *a* = *d* = Ile (GCN4-p-II) specifies a trimer, and *a* = Leu, *d* = Ile (GCN4-pLI) specifies a tetrameric coiled coil.[94, 95] Other mutants (p-VI, p-VL, p-LL and p-LV) were also prepared, but were found to populate multiple oligomerisation states. The helix-helix interactions that underpin the structures of the three discrete coiled coil assemblies — termed 'knobs into holes' (KIH) — were first postulated by Francis Crick in 1953.[96] In KIH packing, 'knob' residues at the *a* and *d* positions on one helix pack into 'holes' on an adjacent helix, the vertices of which are defined by *g*, *a*, *d* and *e* residues (figure 1.6). KIH packing is described in terms of the spatial relationship between the C_α-C_β vector of the knob side chain and the corresponding hole on an adjacent helix. KIH packing changes with regard to the coiled-coil interface type. In dimers, the C_α-C_β vector of the *a* knob is parallel to the hydrophobic interface on the adjacent helix, whereas the C_α-C_β vector of the *d* knob is perpendicular to the hole whose sides are defined by the C_α atoms of the *d* and *e* sidechains on the adjacent helix (figure 1.6A). In trimers, the knobs-into-holes packing is neither perpendicular or parallel, but falls somewhere in between, and is termed 'acute' packing (figure 1.6B). In tetramers, the KIH packing is opposite to that observed in dimers (figure 1.6C). Ultimately, Harbury's research shows the packing angle preferences for leucine and isoleucine (parallel and perpendicular, respectively, unless these angles are disallowed by nature of the assembly *e.g.* in trimeric coiled coils), and how KIH packing specifies oligomerisation state. Additionally, the work suggests that not all β-branched aliphatics are equal in helping to specify oligomeric state; isoleucine, for example, can dictate discrete oligomers, whereas valine, at the same positions, cannot. In addition to the knobs at *a* and *d* positions in simple coiled coil structures, the additional hydrophobic

residues incorporated into the hydrophobic interface at *g* and *e* can also act as knobs.

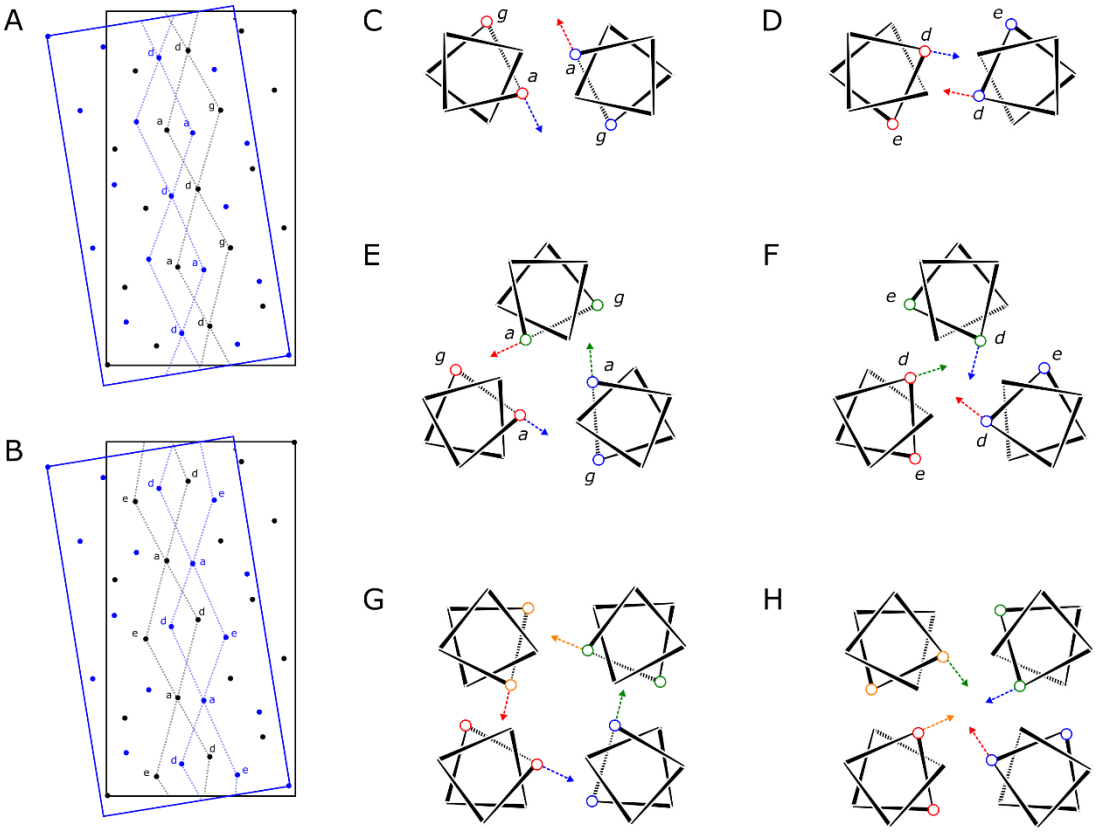


Figure 1.6: *KIH packing of coiled coils. Overlaid helical nets from figure 1.4, showing* ***A****,* a *knobs resting in the centre of diamond-shaped holes formed by* g*,* a*, and* d *residues, and* ***B****,* d *knobs resting in the centre of diamond-shaped holes formed by* a*,* d*, and* e *residues. Cartoon depictions of* ***C****, parallel packing in an* a *layer of a CC dimer,* ***D****, perpendicular packing in the* d *layer of a CC dimer,* ***E****, acute packing in an* a *layer of a CC trimer,* ***F****, acute packing in a* d *layer of a CC trimer,* ***G****, perpendicular packing in an* a *layer of a CC tetramer, and* ***H****, parallel packing in a* d *layer of a CC tetramer. Knobs are depicted by dashed lines with arrowheads showing their trajectory, and holes by open circles. Colours indicate the knob projected by one helix and the corresponding hole on an adjacent helix.*

In 2012, the Woolfson group reported a 'basis set' of *de novo* designed dimeric, trimeric and tetrameric coiled coils (figure 1.7).[60] The sequences

are informed by previous mutational studies of the GCN4-p1 leucine zipper and studying natural homo-oligomeric coiled coils. The *a* and *d* positions are populated by a combination of leucine and isoleucine or isoleucine on its own as necessary for oligomer specificity; the *b* and *c* positions by helix-promoting but otherwise unobtrusive Ala, and the *f* positions are populated by Lys and Gln (for solubility) and Trp (which provides a chromophore).

Whilst examples of α-helical barrels (αHBs), coiled coils which oligomerise in solution and possess a central channel, exist in nature, realising novel αHBs was only first achieved around 15 years ago, when further engineering of the GCN4-p1 leucine zipper yielded the first seven-helix coiled coil.[97] Following on from a previous study in which all *g* positions were mutated to Arg,[98] resulting in a switch of oligomer state and orientation of the helices to an antiparallel tetrameric assembly, researchers were keen to explore how the chemical nature of the side-chains flanking the *ad* hydrophobic core in coiled coils would affect the structure of the dimeric leucine zipper. Liu *et al.* mutated all the hydrophilic amino acids at *g* and *e* to Ala. The result was a heptameric assembly, with each helix z-shifted with respect to each other so that the first and last helices differed in position by one heptad. The core of the assembly is composed of alternating *a* and *d* layers, a phenomenon which had not been observed in natural coiled coils prior. Additionally, the researchers found that the Asn-17 residues (located at an *a* position) were involved in a network of buried hydrogen bonding which was essential to folding (mutation of Asn-17 residues to non-polar amino acids resulted in unfolded peptide).

The discovery of the first parallel hexameric coiled coil was also serendipitous.[99] Zaccai *et al.* synthesised and characterised a mutant of CC-Tet, swapping Lys at *e* for Ala. An X-ray crystal structure of the mutant revealed that the peptide adopted a parallel hexameric assembly. Further mutations at position 24 of the hexamer (at an *a* position) to His or Asp were made, resulting in an A_3B_3 hexameric assembly which was supported by

charge complementarity. Like GCN4-pAA, the channel of CC-Hex was found to be permeable to water, despite its hydrophobic nature.

Further work published by the Woolfson group in 2014 sought to design further coiled coil 'largermers', aided by computation.[73] Initial designs married two type N interfaces on one helix (at *cdga* and *deab*, thought of as equivalent to the *gade* face in traditional type N interfaces), and typically placed hydrophobic residues at *g*, *a*, *d* and occasionally *e*. The resulting *hhpphhp* (*g* to *f*) type II repeats were expected to program αHBs of five to seven helices. Sequences were restricted to the use of Ala, Glu, Ile, Lys, Leu, Asn, Gln, Arg, Ser and Val, all abundant in known parallel coiled coils. Sequences were further screened to select for possible interhelical salt bridges at *b*→*c'*, *b*→*g'*, *e*→*c'* or *e*→*g'*, and sequence redundancy was reduced by choosing sequences containing Lys over Arg. A representative set of 22 sequences out of 76 which met the listed criteria were selected for further study. Each sequence was modelled in CCBuilder as a parallel oligomer with four to eight component helices. The predicted preferred oligomeric state corresponded to the lowest energy model, determined after model optimisation with a genetic algorithm (with floated parameters including the radius, pitch and rotational offset between helices). Synthesis and characterisation of the 22 peptides yielded a number of successful designs, the experimentally determined oligomer states of which matched the predicted preferred oligomer state: a pentamer (CC-Pent: `ILQKIE`; figure 1.7), two hexamers (CC-Hex2: `SLKEIA`, figure 1.7; CC-Hex3: `SIKEIA`), and a heptamer (CC-Hept: `ALKEIA`, figure 1.7). X-ray crystal structures of these peptides revealed solvent-accessible channels in the centre of the assemblies.

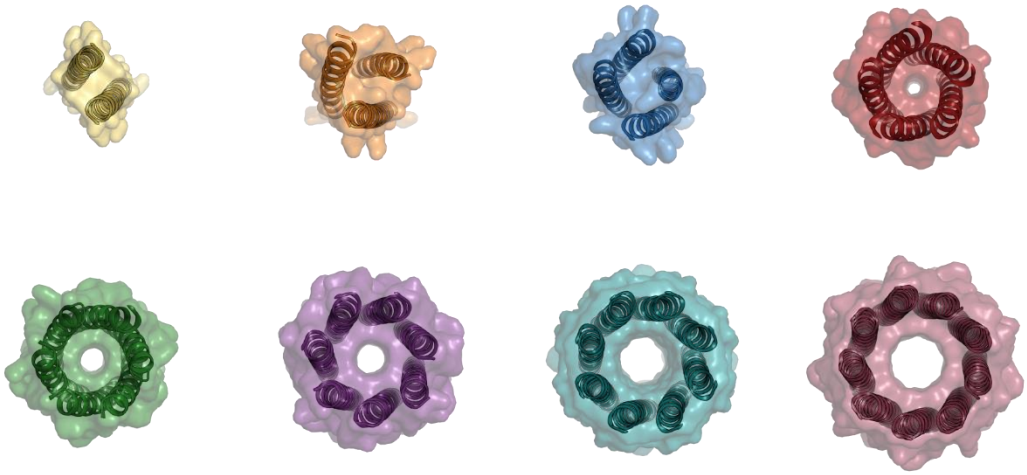


Figure 1.7: *End-on view of selected de novo designed coiled coils from the Woolfson group. Top left to right, CC-Di (yellow, 4DZM),[60] CC-Tri (orange, 4DZL),[60] CC-Tet (blue, 3R4A),[60] and CC-Pent (red, 4PN8).[73] Bottom left to right, CC-Hex2 (green, 4PN9),[73] CC-Hept (purple, 4PNA),[73] CC-Oct (teal, 6G67),[100] and CC-Non (pink, 7BIM).[101] Views are not scaled.*

Research by Dawson *et al.* published in 2021 showed that the oligomeric state of type II α-helical barrels could be increased by successive decreases of the side-chain steric bulk at *g*.[101] In the *g*LKEIA background, Thr at *g* dictates a pentamer (CC-Pent2); Ser at *g* specifies a hexamer (CC-Hex2); Ala at *g* a heptamer, and, somewhat surprisingly, Gly at *g* yields a nonamer in the solid state. However, CC-Non (GLKEIA) exists as a hexameric barrel in solution. Similarly, a close relative of CC-Hept, CC-Oct (AIKEIA) crystallises as an octamer but is hexameric in solution.[100]

In summary, coiled coils of four heptads in length can be reliably and robustly designed *de novo*, either by rational or computational means, in accordance with the sequence specifications set out in table 1.2.

Table 1.2: *A summary of design rules for de novo coiled coils.*

O.S.	Peptide	Position						PDB ID
		g	***a***	***b***	***c***	***d***	***e***	
2	CC-Di	Glu	Ile/Asn	Ala	Ala	Leu	Lys	4DZM
3	CC-Tri	Glu	Ile	Ala	Ala	Ile	Lys	4DZL
4	CC-Tet *	Gln	Leu	Lys	Glu	Ile	Gln	6XY1
5	CC-Pent2	Thr	Leu	Lys	Glu	Ile	Ala	7BAS
6	CC-Hex2	Ser	Leu	Lys	Glu	Ile	Ala	4PN9
7	CC-Hept	Ala	Leu	Lys	Glu	Ile	Ala	4PNA
8	CC-Oct	Ala	Ile	Lys	Glu	Ile	Ala	6G67
9	CC-Non	Gly	Leu	Lys	Glu	Ile	Ala	7BIM

O.S. = oligomer state. * This sequence is that that of 4-KE-4 peptide from the Class 4 homotetramers (collectively known as CC-Tet*), a family of homotetrameric coiled coils which are robust to truncation and charge inversion, unlike the original CC-Tet sequence, `ELAAIKX`.[102]

It should be noted that the entries in table 1.2 denote rules of thumb for *water-soluble*, *parallel* CCs; membrane-spanning parallel CCs can also be designed, using either buried hydrogen bond networks, or membrane-spanning or water-soluble peptide and protein sequences as inspiration.[103-105] The *de novo* design of *antiparallel* CCs remains in its infancy. Initial designs, however, stress the importance of complementary charge patterning at positions flanking the hydrophobic core.[57]

1.3.3 *De Novo* Coiled Coils and Function

Further to the successes of designing coiled coils *de novo*, progress has been made towards functionalising them. Symmetric metal-binding sites have been introduced into α-helical bundles, for example. Notable cases include the *de novo* due ferro (two-iron) helical bundles from Lombardi and

DeGrado,[59, 106] and lanthanide-binding coiled coils from the Peacock group with applications as imaging agents.[107] Designed CC bundles can facilitate and/or inhibit protein-protein interactions,[108] and in doing so can regulate transcription.[109, 110] Orthogonal *de novo* α-helical bundles have also been used as building blocks for nanomaterials, for example *via* coiled-coil protein origami.[111]

α-Helical barrels are innately functionalisable. Research from the Woolfson group published in 2018 found that α-helical barrels are capable of binding small hydrophobic molecules such as the environmentally-sensitive dyes in their central channel.[112] Additionally, barrels of differing pore sizes can selectively discriminate between lipophilic analytes of various sizes, and so can act as reporters in a displacement assay. A year prior, Burton *et al.* retrofitted a coiled-coil heptamer, CC-Hept, with a Cys-His-Glu catalytic triad, like those found in natural Cys/Ser hydrolases, producing a rudimentary hydrolase, CC-Hept-CHE (figure 1.8).[113] The Cys-His-Glu residues were placed at *d*-*a*-*d* positions in the third and fourth heptads, as opposed to exclusively at the *C* terminus, over fears of helix fraying. A colorimetric assay with *para*-nitrophenyl acetate was used to probe the activity of CC-Hept-CHE, and ester hydrolysis was observed: the catalytic efficiency (k_{cat}/k_M) of CC-Hept-CHE was determined to be 3.7 ± 0.6 M^{-1} s^{-1}. However, whilst CC-Hept-CHE's performance matched that of other redesigned proteins,[114, 115] it is rudimentary in comparison to the natural serine hydrolase α-chymotrypsin (figure 1.8) with the same substrate — approximately 10^3 times less efficient.[116, 117] Transition-state modelling in RosettaMatch suggests that spreading the catalytic triad across three adjacent helices — as opposed to being located on the same helix — could improve catalytic efficiency (figure 1.8C).[118]

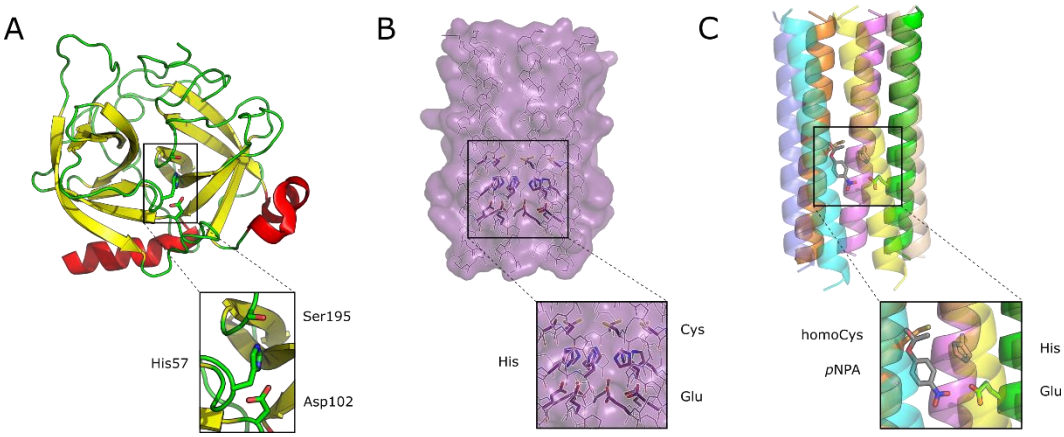


Figure 1.8: *Natural vs. designed hydrolases.* ***A****, The natural serine hydrolase α-chymotrypsin, with its single catalytic triad (Ser195-His57-Asp102) highlighted (4CHA).*[119] ***B****, The de novo designed homomeric cysteine hydrolase CC-Hept-CHE, based on CC-Hept, with all seven catalytic Cys-His-Glu triads highlighted (5EZC).*[113] ***C****, A transition state model of a heteromeric CC-Hept-CHE with a single catalytic homoCys-His-Glu triad highlighted, generated in RosettaMatch by Antony Burton.*[118] *The molecule in the central channel is para-nitrophenyl acetate (pNPA).*

1.3.4 Avenues to Coiled Coil De-Symmetrisation

In order to explore single-site mutations in α-helical barrels — to derive the optimum placement of each residue in the catalytic triad of CC-Hept-CHE, for example — the symmetry of the parent coiled coil must be broken. Herein, 'symmetry' refers to the highly repetitive nature of a homomeric coiled coil, in which the identity of a residue at any given site is the same across the whole assembly. Unfortunately, higher-order heteromeric αHBs have not yet been achieved, and so a new scaffold must be constructed for such experiments. There are three possible routes towards de-symmetrised, higher-order coiled coils. One, *templating a heteromeric αHB*s. Interactions between two complementary coiled-coil strands have previously been templated using DNA nanostructures.[120, 121] More complex DNA templates would allow for ever greater control over the spatial organisation of peptide pieces, without necessitating the design and

realisation of increasingly complex helical assemblies. Two, *designing a single-chain antiparallel αHB*. Spencer and Hochbaum first reported an antiparallel coiled-coil hexamer in 2016.[122] The assembly is driven by a specific Phe/Ile interaction in the CC core,[123] which renders the core inaccessible — not an ideal starting point for catalysis in the barrel channel. Finally, *designing a single-chain parallel αHB*. There is an advantage here that parallel systems are already well-understood, and the design will focus primarily on a scaffold to support and connect the α helices in the barrel as opposed to the design and connection of a new, antiparallel scaffold. Hexameric and octameric assemblies of helix-loop-helix (HLH) motifs in which six and eight α helices are arranged around a central axis have recently been reported, although to date, no single-chain parallel barrels have emerged.[124] This thesis concerns itself with this final approach.

1.4 Miniproteins as Inspiration for Dark Matter Proteins

Miniproteins are polypeptides composed of fewer than 40-50 amino acids that possess an overall well-defined topology comprising two or more SSEs.[125] Due to their small size, the number of possible stabilizing forces present in miniproteins are limited. As such, their compact nature necessitates an optimal arrangement of residues involved in stabilizing non-covalent interactions. Due to their reduced complexity, miniproteins are ideal candidates for the study of sequence-to-structure and sequence-to-stability relationships. Practically speaking, miniproteins allow the simplification of both the protein folding and protein design problems.

Whilst interesting structures in their own right, miniproteins can also act as inspiration for the construction of larger *de novo* proteins with a view to functionalisation. Many miniproteins — for example, βαβ motifs and pancreatic polypeptides — possess at least one helix-loop-helix motif which defines their tertiary structure. These HLH motifs are abundant in biology.

For example, they are present in TIM barrels, Rossman folds, and leucine rich repeats in the form of β-loop-α motifs, and α-loop-α motifs are observed in α-solenoids and α-helical bundles. Here, we propose that miniproteins (specifically, *designed* miniproteins) possessing HLH motifs go beyond inspiration for *de novo* proteins — they are peptide pieces capable of directly templating them.

1.4.1 ββα Motifs

The ββα motif was first identified as the repeating motif in transcription factor IIIA from *Xenopus* oocytes,[126] and a solution-phase structure was later determined for the Xfin domain.[127] The ββα fold (comprising a β-hairpin packed against an α helix) in nature is typically directed not by the sequestration of the conserved hydrophobic core, but by the binding of a metal ion, for example zinc in DNA-binding zinc finger domains, *via* multiple His and/or Cys sidechains.

A monomeric metal-free ββα miniprotein (BBA1) has designed by Struthers *et al. via* an iterative design process, starting from a consensus of five natural zinc-binding ββα sequences (figure 1.9A).[128] BBA1 is the result of the truncation and improvement of loops connecting the β-hairpin and the α helix, and the β-hairpin connecting loop, respectively, stabilisation of the SSEs present, and mutation of two residues to a D- or a non-natural amino acid to stabilise the β-hairpin turn and to incorporate reporter functionality. Dahiyat and Mayo validated their computational protein design algorithm experimentally with the characterisation of a 28-residue ββα miniprotein, FSD-1.[71] FSD-1 produced a far-UV CD spectrum akin to that of natural zinc finger domains, and displayed a broad but cooperative thermal denaturation and renaturation, with a T_M = 39 °C. BLAST searches of the FSD-1 sequence returned no natural zinc finger domains. Comparisons of the

sequences of FSD-1 and Zif268 (the protein backbone utilised in the design process) showed that the two sequences shared only six residues.

Headway has already been made in the oligomerisation of ββα miniproteins. A homo-tetrameric, 21-residue ββα miniprotein, BBAT2, has been designed *via* deletion of two residues of a loop region in the parent monomer (figure 1.9B).[129-132] In addition, computational modelling of mutations to surface and core residues of BBAT2 led to the design of two ββα hetero-tetramers, although it appears that hetero-specificity is achieved at the expense of stability.[133] The miniproteins oligomerise *via* their α helices, setting a precedent for miniprotein oligomerisation through this SSE.

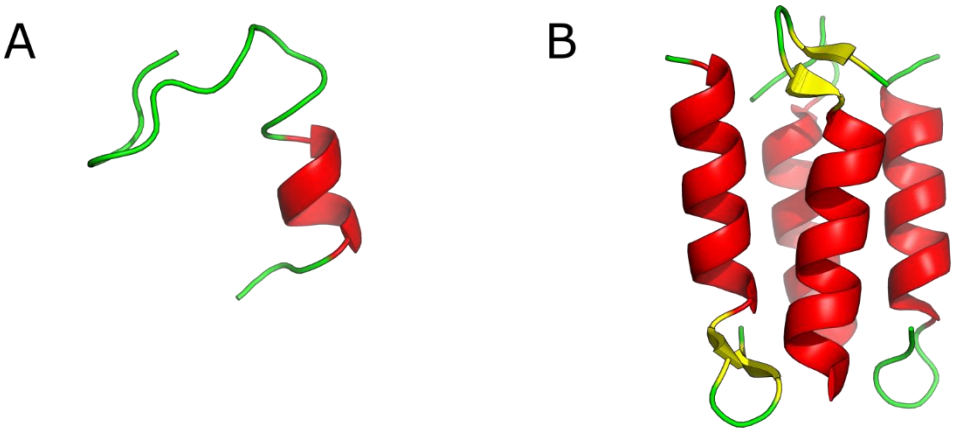


Figure 1.9: *Designed ββα miniproteins.* ***A****, The designed 23-residue, metal-free monomeric ββα miniprotein BBA1.*[128] ***B****, The homo-tetrameric ββα-based oligomer, BBAT2.*[129-132]

1.4.2 The Family of Pancreatic Polypeptides

The family of pancreatic polypeptides (PPs) are miniproteins that are less than 40 amino acids in length. The structure of the pancreatic polypeptides

comprises an α helix buttressed by a PPII helix. A key feature of the pancreatic polypeptide family structure is the interdigitation of proline residues from the PPII helix with aromatic residues (typically Tyr) presented by the α helix, forming the hydrophobic core of the miniprotein. Avian pancreatic polypeptide was the first of such peptides to be isolated in 1975 from chicken pancreas.[134] Similar proteins have since been identified in other mammals as well as in bacteria (table 1.3).

Table 1.3: *Sequences of the pancreatic polypeptide family of miniproteins.*

Peptide	Sequence	PDB ID
aPP	G**P**SQ**P**TY**P** **G**DD**A**P VEDLIRF YND**L**QQ**Y** LNVVTRH RY	1PPT [135]
bPP	A**P**LE**P**EY**P** **G**DN**A**T PEQMAQ**Y** AAE**L**RR**Y** INMLTRP RY	1BBA [136]
hPP	A**P**LE**P**VY**P** **G**DN**A**T PEQMAQ**Y** AAD**L**RR**Y** INMLTRP RY	-
pPP	A**P**LE**P**VY**P** **G**DD**A**T PEQMAQ**Y** AAE**L**RR**Y** INMLTRP RY	-
nPY	Y**P**SK**P**DN**P** **G**ED**A**P AEDLAR**Y** YSA**L**RH**Y** INLITRQ RY	1RON [137]
pYY	Y**P**IK**P**EA**P** **G**ED**A**S PEELNR**Y** YAS**L**RH**Y** LNLVTRQ RY	2DEZ [138]
AgI/II	P**P**VK**P**TA**P** / EAKLAK**Y** QAD**L**AK**Y** QKDLADY PV	3IOX [139]
GBPC	K**P**TP**P**EK**P** / PQKQEQ**Y** NKDFEK**Y** QSDVKEY EA	6CAM [140]

PP = pancreatic polypeptide; aPP = avian PP; bPP = bovine PP; hPP = human PP; pPP = porcine PP; nPY = neuropeptide Y; pYY = peptide tyrosine tyrosine; AgI/II = bacterial adhesin and antigen from Streptococcus mutans; GBPC = glucose-binding protein C from Streptococcus mutans. Residues conserved between the majority of sequences are shown in bold.

Many pancreatic polypeptide-like miniproteins do not exist as obligate monomers *in vitro*, but as dimers.[141] For example, avian pancreatic polypeptide (aPP) dimerises in solution at concentrations in the micromolar range,[141,142] and exists as an antiparallel dimer in the solid state.[135] Monomerised pancreatic polypeptide-type structures have proven to be useful scaffolds on to which functional residues can be grafted, but the field

is limited in terms of functionalisation of the naturally-occurring PP dimers.[143-145]

1.4.3 PPα Miniproteins

Partially inspired by the structure of the pancreatic polypeptides, the fragment-based design of the PPα miniprotein has previously been described.[146] The tertiary structure of the miniprotein borrows from the structures of a bacterial adhesin and antigen and bovine pancreatic polypeptide,[136, 139] and comprises an *N*-terminal polyproline (PPII) helix and an antiparallel amphipathic α helix connected by a five-residue loop (figure 1.10). PPα has a thermal midpoint temperature of unfolding of 39 °C and is monomeric in aqueous solution. Structural NMR studies have shown that PPα is stabilised by numerous CH-π interactions, primarily between Pro and Tyr residues, and that the strength of these NCIs and thus the stability of the miniprotein is tuneable based on the electronics of the *para*-substituted phenylalanine rings. Functional groups capable of donating electron density into the aromatic ring stabilise the miniprotein — tyrosine with its *para*-hydroxy group is the most stabilising, whereas *para*-nitrophenylalanine stabilises PPα the least of all the functional groups probed.

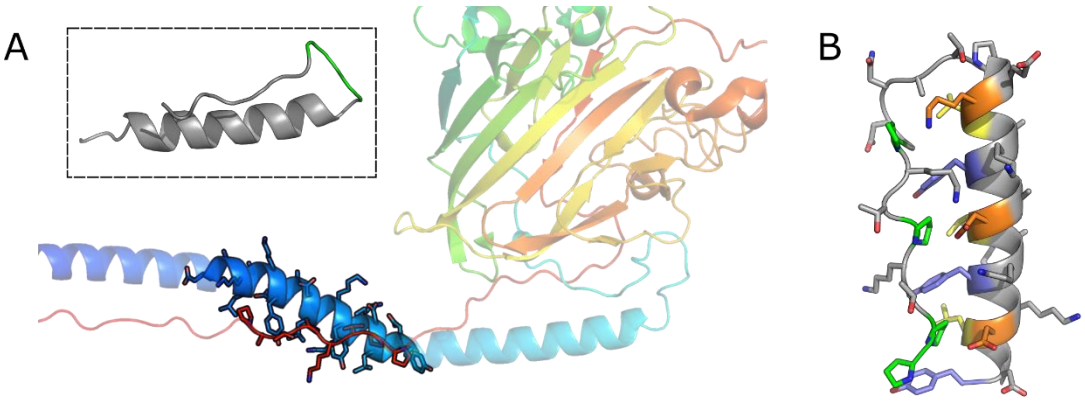


Figure 1.10: *Design of PPα and optimised oPPα.* ***A****, Inspiration for the fragment-based design of the PPα family of miniproteins. The structure of adhesin and antigen (AgI/II) from Streptococcus mutans (3IOX),[139] with α and PPII regions of interest highlighted. Inset: the NMR structure of bovine pancreatic polypeptide (1BBA),[136] the loop (highlighted in green) of which was used to connect the SSE sequences borrowed from AgI/II.* ***B****, An example structure from the NMR ensemble of the PPα miniprotein (5LO2).[146]*

The rational redesign of PPα-Tyr (10 mutations to the parent sequence) led to an optimised PPα miniprotein, oPPα (table 1.4, figure 1.11).[147] oPPα is more thermally stable than PPα, with T_M of 51 °C — a 12 °C increase compared to its parent. Whilst oPPα is also stabilised by CH-π interactions (with more of such interactions found in the 20-model NMR ensemble of oPPα compared to PPα), Lys is the predominant CH-π donor, compared to Pro in PPα.

Table 1.4: *Sequences of the PPα and optimised PPα miniproteins.*

Peptide	Sequence and Register *efgabcd efgabcd efgabcd*	T_M (°C)
PPα [146]	Ac-PPTKPTKP GDNAT PEKLAKY QADLAKY QKDLADY-NH_2	39
oPPα [147]	Ac-PPKKPKKP GDNAT PEKLAAY EKELAAY EKELAAY-NH_2	51

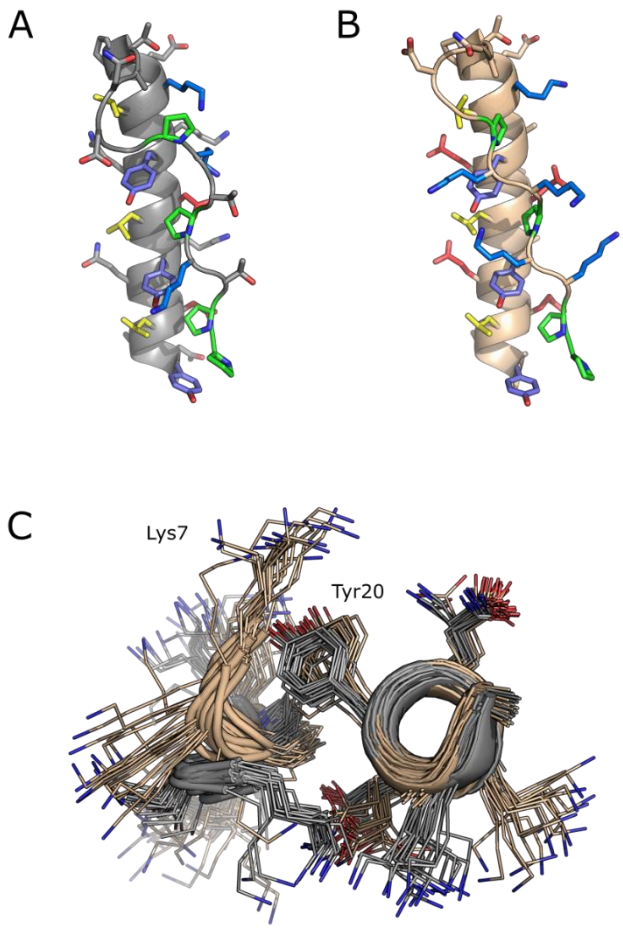


Figure 1.11: *Comparison of the sequences and structures of PPα and oPPα.* ***A****, Model 1 of from the NMR ensemble of PPα (5LO2).* ***B****, Model 1 from the NMR ensemble of oPPα (6GWX). Residues key to the PPII:α helix interface are highlighted. Colour key: Pro is shown in green, Lys in blue, Glu and Asp in red, Leu in yellow, and Tyr in purple.* ***C****, An overlay of all NMR ensemble models for PPα and oPPα, showing the differing positions of Lys7 and the PPII helix relation to Tyr20 and the α helix in both miniproteins.*

In oPPα, the *adg* hole presented by the α helix is a key component of the interhelical interface that stabilises the miniprotein's tertiary structure. Whilst Tyr at *d* is critical to maintaining this interface, little is known of the role played by the remaining vertices of the hole at *a* and *g*. A series of mutations probing the nature of the residue at *a* from hydrophobic, γ-

branched Leu to hydrophobic, β-branched Ile and Val, non-bulky Ala, and charged Lys and Glu were synthesised and characterised (table 1.5). Of these six mutants, oPPα (Leu at *a*) was the most stable. Substituting a β-branched Ile or Val at *a* resulted in an 11 °C and 17 °C decrease in T_M, respectively. Of the remaining mutants, only oPPα-K@*a* was folded. Mutating Glu at *g* to Ala, taking care to ensure that the interhelical charge patterning was maintained, resulted in loss of thermostability. Alternatively, mutating Glu at *g* to the hydrophobic Leu, yielding an all-hydrophobic 'hole', resulted in a broad thermal transition and concentration-dependent CD spectra, indicating aggregation of the monomer. The results of this series of mutations showed that Tyr at *d*, Leu at *a* and Glu at *e* indeed make for the most stable variant of the optimised PPα miniprotein.

Table 1.5: *Mutations made to oPPα to probe the stability of the* adg *hole presented by the α helix.*

Peptide	Sequence and Register	MRE_{222}	T_M
	efgabcd efgabcd efgabcd		
oPPα	PPKKPKKP GDNAT PEKLAAY EKELAAY EKELAAY	-18319	51
oPPα-I@*a*	PPKKPKKP GDNAT PEKIAAY EKEIAAY EKEIAAY	-16493	40
oPPα-V@*a*	PPKKPKKP GDNAT PEKVAAY EKEVAAY EKEVAAY	-14854	34
oPPα-A@*a*	PPKKPKKP GDNAT PEKAAAY EKEAAAY EKEAAAY	-8522	-
oPPα-E@*a*	PPKKPKKP GDNAT PEKEAAY EKEEAAY EKEEAAY	-1642	-
oPPα-K@*a*	PPKKPKKP GDNAT PEKKAAY EKEKAAY EKEKAAY	-12448	19

All peptides were *N*-terminally acetylated and amide-capped at the *C* terminus. MRE_{222} in units of deg cm^2 $dmol^{-1}$ res^{-1}. T_M in °C.

1.5 Scope of Thesis

Despite the vastness of space covered by known (natural and designed) proteins, a dizzying proportion of the dark matter of protein space remains unexplored. A number of design rationales can aid in the exploration of this unchartered territory, with particular promise in the sub-fields of rational and computational protein design. Herein, we adopt a strategy akin to fragment-based design in which tertiary structures — as opposed to individual SSEs in the case of fragment-based design — are organised to form a template for larger oligomeric and single-chain proteins. Our efforts are directed towards the combination of two known SSEs (PPII and α helices, inspired by the PPII:loop:α helix topology of the pancreatic polypeptides and PPα miniproteins and the structures of designed coiled coils) in a previously unseen topology, and we present initial *in vitro* biophysical characterisation for the resulting tuneable peptide oligomers. Attempts are made to construct parametric models of these novel assemblies, and to combine strands into larger oligomeric proteins and single-chain proteins.

Chapter 2

Materials and Methods

All Fmoc-protected amino acids and Cl-HOBt were purchased from Cambridge Reagents. Supporting resins (100-200 mesh) were purchased from Novabiochem®. DIC and oxyma was purchased from Carbosynth, morpholine from Merck Millipore, formic acid, TIPS, and TFA from Acros Organics, pyridine and CH_2Cl_2 from Fisher, acetic anhydride from BDH Laboratories, diethyl ether from Honeywell, and HPLC-grade MeCN from VWR Chemicals. All chemicals were used as supplied. Synthesis-grade ultra-pure DMF (Cambridge Reagents) was used exclusively during peptide synthesis.

2.1 Peptide Synthesis and Purification

2.1.1 Peptide Synthesis

Peptides were synthesised on a 0.1 mmol scale using a microwave-assisted Liberty Blue automated peptide synthesiser (CEM corporation) on a MBHA rink amide resin using standard Fmoc coupling chemistry. Fmoc deprotection was achieved with 20% morpholine in DMF. For sequences containing the GDNAT motif, the addition of 5% formic acid to the deprotection solution was used to suppress aspartimide formation. Peptides were synthesised from *C*- to *N*-terminus, with amino acid (0.2 M in DMF) couplings achieved using either Cl-HOBt (0.5 M in DMF) or oxyma (0.5 M in DMF) and DIC (1 M in DMF) at 70 °C. Single couplings under microwave conditions were used, except for polyproline-II helix sequences, which required double coupling.

N-terminal acetylation of the peptides was achieved with an excess of pyridine (0.5 mL) and acetic anhydride (0.25 mL) in DMF (5 mL) for 20 minutes. The resin was subsequently washed with DMF × 3 and CH_2Cl_2 × 3 prior to a three hour cleavage of the peptide from the resin under agitation, using H_2O:TIPS:TFA (5:5:90 vol%, 20 mL). The cleavage mixture was filtered from the resin and the volume reduced to < 5 mL under a flow of nitrogen. The peptide was precipitated in ice-cold diethyl ether (45 mL), and isolated by centrifugation and removal of the supernatant. The peptide pellet was dissolved in H_2O:MeCN (50:50 vol%, 10 mL) and lyophilised, yielding the crude peptide as a white powder.

2.1.2 Peptide Purification

Peptides were purified using a preparative JASCO HPLC system, with a reverse-phase Luna® C18 column (Phenomenex, 5 μm particle size, 100 Å

pore size, 150 × 10 mm). A linear gradient of 20-80% 0.1 vol% TFA in MeCN (Buffer B) vs. 0.1 vol% in H_2O (Buffer A) was typically used for initial purification. A second purification step was added as necessary. Peptide purities were determined using a JASCO analytical HPLC system, fitted with a reverse-phase Kinetex® C18 analytical column (Phenomenex, 5 μm particle size, 100 Å pore size, 100 × 4.6 mm); all peptides exceeded 97 % purity.

2.2 Recombinant Protein Production

2.2.1 Construct Design

Designed gene sequences were optimized and synthesised by GeneArt® (Thermo Fisher Scientific, Germany) or Twist Bioscience (USA). Genes were installed in pET-28a vectors, between NdeI and HindIII or NcoI and XhoI restriction sites to achieve *N*- or *C*-terminally His-tagged proteins, respectively.

2.2.2 Bacterial Transformation

1 μL plasmid (approximately 100 ng uL^{-1}) was transformed into ultracompetent *E. coli* Rosetta™ 2 (DE3) cells (Novagen) or Lemo21 (DE3) cells (New England BioLabs) *via* heat shock. Cells were incubated with plasmid on ice for 30 min prior to heat shock (42 °C for 30–60 s), followed by a 2 min incubation on ice. 1 mL LB media was added, and the cells were incubated for 1 hr 30 min at 37 °C, 180 rpm. Cells were pelleted by centrifugation at 13,000 rpm for 1 min and the majority of the supernatant was decanted. Cells were mixed with the remaining supernatant gently with pipetting and 50 μL of this mixture was plated on a 10 cm LB agar plate containing appropriate antibiotics and incubated overnight at 37 °C. Plates on which bacterial colonies were observed were stored at 5 °C for the preparation of overnight cultures or glycerol stocks.

2.2.3 Protein Expression

500 mL LB media was supplemented with kanamycin sulfate (50 ng μL^{-1} final concentration) and chloroamphenicol (35 ng μL^{-1} final concentration) and inoculated with 5 mL overnight culture and incubated at 37 °C, 180 rpm (MaxQ 4000, Thermo Scientific, USA). Growth was monitored and protein expression was induced when the absorbance at $OD_{600\ nm}$ reached 0.6–0.8. Protein expression was induced *via* addition of isopropyl-β-D-1-thiogalactopyranoside (IPTG, 0.5 mM final concentration), followed by overnight incubation at either 25 or 37 °C, 180 rpm.

2.2.4 Nickel Affinity Chromatography

Cells were pelleted by centrifugation at 4,500 × g for 10 min at 4 °C (Sorvall Lynx 4000, Thermo Scientific, USA). The supernatant was decanted and the pellet lysed with 20 mL lysis buffer per L of culture (20 mM HEPES, 5 mM imidazole, 500 mM NaCl at either pH 7 or 7.5, depending on the pI of the protein of interest). PMSF was added to the cell lysate to a final concentration of 2 mM. The cell lysate was incubated on ice prior to sonication at 40 % power, 70 % pulse (Model 3000 Ultrasonic Homogenizer with 12.7 mm solid tip, BioLogics, USA) for 5 min. For cell lysates containing overexpressed thermostable proteins, the lysate was subjected to heat shock treatment at 60 °C for 1 min. Cell debris was pelleted *via* centrifugation at 29,000 × g for 30 min at 4 °C. Affinity purification of the His-tagged proteins was undertaken on a column charged with Super Ni-NTA Affinity Resin (Generon, UK), equilibrated with wash buffer (20 mM HEPES, 10 mM imidazole, 500 mM NaCl, pH 7 or 7.5) The flow through was collected first, followed by wash and elution fractions (elution buffer: 20 mM HEPES, 250 mM imidazole, 100 mM NaCl, pH 7 or 7.5). Occasionally, Ni-NTA affinity chromatography was undertaken on an ÄKTAprime plus fitted

with a 5 mL HiFliQ Ni-NTA column (Protein Ark, UK). In this case, pellets were lysed with a HEPES buffer (50 mM HEPES, 300 mM NaCl, 1 mM PMSF, pH 7.5) and protein eluted from the column with a high concentration of imidazole (50 mM HEPES, 250 mM imidazole, 300 mM NaCl, pH 7.5).

2.2.5 Size Exclusion Chromatography

Further protein purification was undertaken using an ÄKTAprime plus fitted with a HiLoad™ 16/600 Superdex™ 75 pg column (GE Life Sciences, USA), during which the protein was exchanged into and appropriate buffer (PBS or 0.1 M phosphate buffer, 57.8 mM Na_2HPO_4, 42.2 mM NaH_2PO_4, pH 7) for biophysical characterisation.

2.2.6 SDS-PAGE

Selected fractions were visualized on a 12% or 13% SDS-PAGE gel stained with Coomassie Blue. Fractions of interest were combined and concentrated using either a 3 mL using an Amicon® Stirred Cell (50 mL capacity, Merck, Germany), or an Amicon® Stirred Cell Model 8003 (3 mL capacity, Merck, Germany) fitted with a 3 or 10 kDa cut-off membrane before determination of concentration and biophysical characterisation.

2.3 Biophysical Characterisation

2.3.1 Concentration Determination

Concentrations of peptides (in H_2O) and proteins (in an appropriate buffer) were determined using a Nanodrop 2000 spectrophotometer (Thermo Scientific) using the Beer-Lambert law (equation 2.1) and the known molar extinction coefficient for Tyr at 280 nm (ε_{280} = 1280 M^{-1} cm^{-1}).

$$A = \varepsilon . c . l \qquad \textit{Eq. 2.1}$$

Where A = absorbance, ε = molar extinction coefficient (M^{-1} cm^{-1}), c = concentration (M), and l = pathlength (cm).

2.3.2 Mass Spectrometry

The identities of the peptides were confirmed by MALDI-TOF mass spectrometry using a Bruker ultrafleXtreme II instrument in reflector mode. Peptides and matrix (a saturated solution of α-cyano-4-hydroxycinnamic acid in a 50:50 vol% solution of H_2O and MeCN) were spotted on a ground steel plate in 1:1 mixture. Protein masses were determined by ESI mass spectrometry using a Waters Synapt GS-2i instrument. Protein samples were prepared in a 50:50 vol% solution of H_2O and MeCN with 5 % formic acid using Millipore C18 ZipTip® pipette tips.

2.3.3 Circular Dichroism Spectroscopy

Peptide samples were prepared at 5, 10, 50 or 100 μM in phosphate-buffered saline (PBS, comprising 8.2 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 137 mM NaCl, and 2.7 mM KCl) at pH 7.4. Cysteine-containing peptide samples included a 10 × excess of TCEP relative to peptide concentration, to reduce the possibility of disulphide formation. CD spectra were recorded at 5 °C using a JASCO 810 spectropolarimeter fitted with a Peltier temperature controller, either a 1, 5, or 10 mm path length quartz cuvette (Starna; 1 mm cuvettes for 100 μM samples, 5 mm cuvettes for 5 and 10 μM samples, and 10 mm cuvettes for 50 and 100 μM near-UV CD experiments), a scanning speed of 100 nm min^{-1}, and a bandwidth of 1 nm. Spectra were recorded as the average of 8 scans and baseline corrected. Thermal denaturation profiles were obtained by monitoring peptide absorbance at 222 nm (unless otherwise stated) at 1 °C intervals from 5-95 °C, with a temperature ramp rate of 60 °C hr^{-1}, 1 nm bandwidth, 16 s delay and 16 s response times. When possible, thermal midpoint of unfolding (T_M) temperatures were

obtained from the first derivative of the thermal denaturation curve using a Savitsky-Golay algorithm. The absorption in mdeg (θ) was converted to mean residue ellipticity (MRE) using equation 2.2:

$$MRE = \left(\frac{\theta.100}{c.n.l}\right) \qquad \textit{Eq. 2.2}$$

Where *c* is the peptide concentration, *n* is the number of peptide bonds including the *N*-terminal acetyl and *l* is the path length.[148]

Peptide helicity was calculated using equations 2.3, 2.4 and 2.5:

$$Fraction\ helix\ (\%) = 100\left[\frac{(MRE_{222} - MRE_{coil})}{(MRE_{helix} - MRE_{coil})}\right] \qquad \textit{Eq. 2.3}$$

$$MRE_{coil} = 640 - 45.T \qquad \textit{Eq. 2.4}$$

$$MRE_{helix} = -42{,}000 \times \left(1 - (3/n)\right) \qquad \textit{Eq. 2.5}$$

Where *T* = 5 °C and *n* is the number of peptide bonds including the *N*-terminal acetyl.[148, 149]

All data shown is representative of one individual dataset, but percentage fraction helix and T_M values are the mean ± standard deviation from n = 3 datasets.

2.3.4 Analytical Ultracentrifugation

Sedimentation velocity analytical ultracentrifugation experiments were carried out on peptides that showed a reasonable degree of α-helical character at 20 °C. SV AUC experiments were carried out at 100 μM in PBS

(unless otherwise stated in section 8.4) at 20 °C in a Beckman XL-I or XL-A analytical ultracentrifuge using an An-50 Ti or An-60 rotor (Beckman Coulter). Peptide solution (310 μL) and buffer (320 μL) were placed in a sedimentation velocity cell with an Epon two channel centrepiece and quartz windows. After temperature equilibration and radial calibration at 3 krpm, samples were centrifuged at 40, 50 or 60 krpm, with absorbance scans taken across a radial range of 5.8 to 7.3 cm at 5 min intervals to a total of 120 scans. Peptide absorbance was monitored at 280 nm. Datasets were fitted to a continuous c(s) distribution model using SEDFIT (http://www.analyticalultracentrifugation.com/) at 95% confidence level.[150] The baseline, frictional coefficient (f/f_0), systematic time-invariant, radial-invariant noise and meniscus or bottom were fitted. The partial specific volume ($\bar{v}$) of each peptide and buffer density and viscosity at 20 °C (PBS, pH 7.4, 1.0054 g cm^{-3}, 0.0102 Pa) was calculated using SEDNTERP. Residuals are shown as a bitmap in which the grayscale shade indicates the difference between the fit and the raw data. The horizontal axis is the radial range over which the data was fitted, from the meniscus (left) to the cell bottom (right).

Sedimentation equilibrium analytical ultracentrifugation experiments were carried out on peptides that showed a reasonable degree of α-helical character at 20 °C. SE AUC experiments were carried out at 100 μM in PBS (unless otherwise stated in section 8.4) at 20 °C in a Beckman XL-A or XL-I analytical ultracentrifuge using an An-60 or An-50 Ti rotor (Beckman Coulter). Peptide solution and buffer (110 and 120 μL per channel, respectively) were placed in a sedimentation equilibrium cell with an Epon six-channel centrepiece and quartz windows. The samples were centrifuged at speeds in the range 18–42 krpm. Peptide absorbance was monitored at 280 nm. Equilibrium datasets were initially processed in SEDFIT, and then fitted initially to single, ideal species models using

SEDPHAT. If poor fits were obtained, the curves were fitted instead to an ideal, two-species model. 95 % confidence limits were obtained *via* Monte Carlo analysis of the fits (10000 total iterations). The reported molecular weight, Mw, is the average Mw from these fits.

2.3.5 Ligand Binding

Ligand binding experiments were prepared using an Eppendorf epMotion 5070 liquid handler. The experiments were carried out on a BMG Labtech Clariostar plate reader at 25 °C in quadruplicate, in PBS with 5% (v/v) DMSO (pH 7.4). Peptide concentrations typically varied from 0 – 500 μM, and DPH concentration remained constant at 1 μM. Peptide and ligand were equilibrated for a minimum of 30 minutes with shaking at room temperature prior to recording data. Fluorescence spectra were measured using an excitation wavelength of 350 nm and emission wavelengths in the range 380-602 nm. Data from at least three experiments were averaged and normalised prior to fitting to a single-site binding model (equation 2.6) using an in-house Python script.

$$y = \frac{B_{max}.x}{K_d + x} \qquad \textit{Eq. 2.6}$$

B_{max} is the fluorescence signal when all of the DPH is bound, and K_d is the dissociation constant.

2.3.6 Peptide Crystallisation Trials

Lyophilised peptide was resuspended in MilliQ H_2O to a concentration of 10 mg mL^{-1}; purified protein in PBS or TBS was concentrated and crystallisation trials conducted at concentrations in the range of 1–7 mg mL^{-1}. Sitting drop vapour diffusion crystallisation trials were set up using an Oryx8 Protein Crystallisation Robot (Douglas Instruments, UK),

mixing 0.3 µL peptide or protein and 0.3 µL of a selected commercial crystallisation screen (Structure Screen 1+2, JCSG Plus™, Morpheus®, ProPlex™ and PACT Premier™, Molecular Dimensions, UK) and incubated at 20 °C.

2.4 Parametric Modelling and Loop Database Searches

ISAMBARD (Intelligent System for Analysis, Model Building and Rational Design) is a modular open-source Python-based parametric protein design package.[78] The AMPAL (Atom, Monomer, Polymer, Assembly and Ligand) framework is a stand-alone module which can be used in conjunction with ISAMBARD for analysis of protein structures.

2.4.1 Extraction of Parameters

Parameters of mock PPα oligomer models constructed in PyMOL were extracted using functionality in AMPAL (https://isambard-uob.github.io/ampal/). Parameters for α helices and PPII helices were extracted separately.

Example code:

```
import ampal
import numpy

protein=ampal.load_pdb("path_to_pdb.pdb")
ref_ax=ampal.analyse_protein.reference_axis_from_chains
       (protein)
helix=protein[helix_ID_number]

# cr_a are the Crick angles
cr_a=ampal.analyse_protein.crick_angles(helix, ref_ax)
```

```
# pca is the phi Cα angle i.e. the interface angle
# Z is 7 if helix is an α helix, or 3 if helix is a PPII helix
pca_list=[cr_a[x] for x in range(0,len(cr_a),Z) if cr_a[x] is
          not None]
pca_a1=numpy.mean(pca_list)
# r is the helix radius
r_list=ampal.analyse_protein.polymer_to_reference_axis_distances
       (a1, ref_ax)
r=numpy.mean(r_list)

# a is the alpha angle
a_list=ampal.analyse_protein.alpha_angles(helix, ref_ax)

pitch_list=[(2* numpy.pi * r)/numpy.tan(numpy.deg2rad(x)) for x
           in a_list if x is not None]
pitch=numpy.mean(pitch_list)
```

2.4.2 PPα Oligomer Specification Class

A specification class for building PPα oligomer models was constructed based on the pre-existing `CoiledCoil` specification class in ISAMBARD v2.0 (https://isambard-uob.github.io/isambard/).

```
import ampal
import isambard.specifications as specifications
import numpy

class PPaOligomer(specifications.CoiledCoil):
    def __init__(self, n, pp_aas, alpha_aas, pp_radius,
                 alpha_radius, pp_pitches, alpha_pitches,
                 pp_phi_c_a, alpha_phi_c_a, pp_zs,
                 pp_shr, auto_build=True):
        super().__init__(2*n, auto_build=False)
        self.aas = [pp_aas, alpha_aas] * n
        self.minor_helix_types = ['PPII', 'alpha'] * n
        self.major_radii = [pp_radius, alpha_radius] * n
        self.major_pitches = [pp_pitches, alpha_pitches] * n
        self.phi_c_alphas = [pp_phi_c_a, alpha_phi_c_a] * n
        self.z_shifts = [pp_zs, 0] * n
        zs_shr_adjust = [(z/p) * 360 for z, p in
                         zip(self.z_shifts, self.major_pitches)]
        shr_adjust = [pp_shr, 0] * n
```

```
        self.rotational_offsets = [x + y + z for x, y, z in
                                   zip(self.rotational_offsets,
                                   shr_adjust, zs_shr_adjust)]
        self_minor_repeats = [3.0, 3.5] * n
        self.orientations = [1, -1] * n
        if auto_build:
            self.build()
```

2.4.3 PPα Oligomer Model Optimisations

The following code describes the PPα oligomer model optimisations, using PPα-Tri as an example:

```
import matplotlib.pyplot as plt

from pprint import pprint

import ampal

import isambard.specifications as specifications

import isambard.modelling as modelling

import isambard.optimisation as ev_opts

from isambard.optimisation import Parameter


class PPCCX(specifications.CoiledCoil):

    def __init__(self, n, pp_aas, alpha_aas, pp_radius,
alpha_radius,

                 pp_pitches, alpha_pitches, pp_phi_c_a,
alpha_phi_c_a,

                 pp_zs, pp_shr, auto_build=True):

        super().__init__(2*n, auto_build=False)

        self.aas = [pp_aas, alpha_aas] * n

        self.minor_helix_types = ['PPII', 'alpha'] * n

        self.major_radii = [pp_radius, alpha_radius] * n

        self.major_pitches = [pp_pitches, alpha_pitches] * n

        self.phi_c_alphas = [pp_phi_c_a, alpha_phi_c_a] * n

        self.z_shifts = [pp_zs, 0] * n

        zs_shr_adjust = [(z/p) * 360 for z, p in
zip(self.z_shifts, self.major_pitches)]

        shr_adjust = [pp_shr, 0] * n
```

```
        self.rotational_offsets = [x + y + z for x, y, z in
zip(self.rotational_offsets, shr_adjust, zs_shr_adjust)]
        self_minor_repeats = [3.0, 3.5] * n
        self.orientations = [1, -1] * n
        if auto_build:
            self.build()
ppcctri = PPaOligomer(3, 8, 21, 15.0, 6.8, 131.4, 168.5, 82.3, -
86.8, 6, 40)


specification = PPaOligomer
sequences = ['PPKKPKKP','PIELIKYEIELIKYEIELIKY']*3
parameters = [
    Parameter.static('Oligomeric State', 3),
    Parameter.static('PPII helix length', 8),
    Parameter.static('Alpha helix Length', 21),
    Parameter.dynamic('PPII radius', 15.0, 1.0),
    Parameter.dynamic('Alpha radius', 6.8, 1.0),
    Parameter.dynamic('PPII pitch', 131.4, 100),
    Parameter.dynamic('Alpha pitch', 168.5, 100),
    Parameter.dynamic('PPII phi_ca', 82.3, 27),
    Parameter.dynamic('Alpha phi_ca', -86.8, 27),
    Parameter.dynamic('z-shift', 6.0, 1.0),
    Parameter.dynamic('SHR', 40, 10),
]


import budeff
def get_buff_total_energy(ampal_object):
    return budeff.get_internal_energy(ampal_object).total_energy
```

For optimisations employing a genetic algorithm:

```
opt_ga = ev_opts.GA(specification, sequences, parameters,
get_buff_total_energy)
opt_ga.run_opt(100, 5)
optimized_model = opt_ga.best_model
```

For optimisations employing a differential evolution algorithm:

```
opt_de = ev_opts.DE(specification, sequences, parameters,
get_buff_total_energy)

opt_de.run_opt(100, 5)

optimized_model = opt_de.best_model
```

2.4.4 Loop Database Construction

The PDB was culled using the PISCES server to compile a dataset of high-resolution (≤ 2 Å) X-ray crystal structures (18th July 2018).[151, 152] The sequences were further culled by entry at 40% maximum sequence identity; R-factor = 0.3; sequence length between 100 and 10000 residues. The loop database was constructed in a developer version of ISAMBARD using tools in the `modelling` module.

2.4.5 Loop Fitter

The newly created high resolution loop database was searched for suitable loops using the `loop fitter` suite of tools in the `modelling` module of the developer version of ISAMBARD. A loop database based on the top8000 database (the 2011 version of which was used to update the torsional distributions used in MolProbity[153]) was also searched. A list of loops meeting the following criteria were compiled for assessment:

- Maximum 12 residues in length (lower limit assigned based on calculated minimum number of residues to span the helix-helix distance);
- Loop between any two DSSP-defined SSEs;[154]
- Distance threshold = 1.0 Å, angle threshold = 10.0°, dihedral threshold = 10.0°;
- Loop fit score ≤ 2 Å, and
- Sequences containing no Cys residues.

Chapter 3

Controlled Oligomerisation of PPα-Based Miniproteins

3.1 Introduction

3.1.1 The Rational Redesign of the PPα Miniprotein

The monomeric PPα miniprotein — the result of a fragment-based design inspired by the structures of pancreatic polypeptides and bacterial adhesins — and its subsequent redesign have previously been described (figure 3.1).[136, 139, 146, 147] The optimised PPα (oPPα) miniprotein is the result of ten mutations to the sequence of PPα, which confer a 12 °C increase in thermostability over its parent (T_M = 51 °C for oPPα *vs*. T_M = 39 °C for PPα).

The PPα miniprotein family comprises an *N*-terminal polyproline-II (PPII) helix and a *C*-terminal α helix, connected by a five-residue loop. The side chains of these two miniproteins which are not involved in the PPII:α interface are solvent exposed, providing a scaffold on to which functional motifs can be grafted. The PPα miniprotein could therefore be thought of as a polypeptide building block, part of a synthetic biologist's 'toolbox' of *de novo* designed peptide components which can be used to interrupt or introduce selected protein-protein interactions and to manipulate other cellular processes.[108, 155] It is also an interesting structure in its own right,

with its helical hairpin topology, and so could act as a source of inspiration for new protein folds. With this view, efforts have previously been made to produce an inverted PPα structure, αPP, with an *N*-terminal α helix and a *C*-terminal PPII helix, and to combine these two topologies into extended, entirely *de novo* PPαPP-type structures.[156] However, we envisage that the PPα miniprotein has more to offer the field of synthetic biology than a mutable 'plug-and-play' tool, and that controlled oligomerisation of this miniprotein — and derivatives of these oligomers — could allow access to a number of functional proteins.

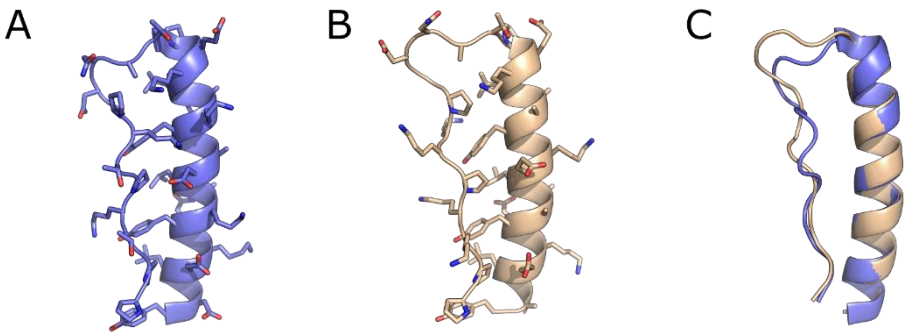


Figure 3.1: *The PPα and optimised PPα miniprotein. Representative structures from the NMR ensembles of* ***A****, PPα (model 12) and* ***B****, oPPα (model 12), and* ***C****, an overlay of the two structures.*

To date, no attempts have been made at oligomerising these PPα-like miniproteins with a view to further functionalisation. There has been some success, however, with the oligomerisation of mixed α/β miniproteins. In pared-down proteins, every residue present plays a vital role in determining tertiary structure, so mutations must be made with care, lest the fold be destabilised.[125] This is particularly poignant for oligomeric miniproteins, due to the positioning of residues which are critical for structure, and because the number of positions at which new interfaces can be introduced is limited.

We reasoned that the utility of pancreatic polypeptide-like structures could be expanded upon by controlled oligomerisation of one of the secondary structural elements (SSEs) present i.e. either the PPII helix or the α helix. Left-handed PPII helices can assemble into all-parallel, right-handed triple helices known as collagens.[157] Strands of the most characterised collagens typically follow a three-residue repeat pattern: Gly-Xxx-Yyy, where the Xxx and Yyy residues are frequently Pro or hydroxyPro. However, there is greater scope for homo-oligomerisation *via* the α helix of these miniproteins.

α-Helical coiled coils (CCs), assemblies of supercoiled amphipathic α helices, are abundant in nature. The sequence-to-structure relationship of these polypeptide assemblies is incredibly well-understood.[94, 158] Coiled-coil structures are easily parameterised, meaning that models can be generated with ease, and the most energetically favourable oligomeric state for any given sequence can be predicted quickly and cheaply in terms of computer power.[78,77] Depending on their sequence, *de novo* designed coiled coils are capable of forming oligomers from dimers through to nonamers, depending on conditions.[60,73,100, 101]

In this chapter, we deliver a number of thermostable and homo-oligomeric PPα miniproteins based on a toolkit of *de novo* designed coiled coils and examine their solution-phase assemblies.

3.2 Proof-of-Concept: A Trimeric PPα

3.2.1 Design

In order to successfully oligomerise the PPα miniprotein, it is imperative to maintain the basic structural features of both the miniprotein and the coiled-coil scaffold. Structural changes may be imparted upon the component α helices of the coiled coil upon addition to and mutation of the sequence to incorporate the PPII helix and loop, and maintain the PPII:α interhelical

interface. Such changes could result in poorer interhelical packing of hydrophobic residues, and, by proxy, a reduction in thermostability of the oligomer compared to the parent coiled coil. In an attempt to mitigate this possibility, the optimised PPα sequence was selected over that of PPα due to its increased thermostability, in the hope that this would mitigate any potential decrease to the stability of the oligomer. We decided to utilise three-heptad repeats, in keeping with the length of oPPα. Thermostable, three-heptad trimeric coiled coils have previously been characterised by the Woolfson group.[61, 102]

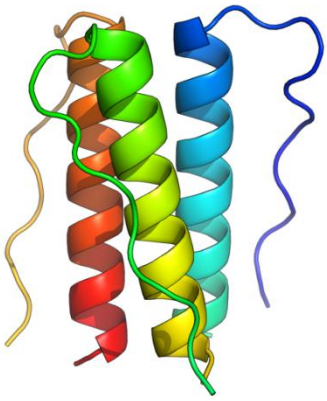


Figure 3.2: *A model of a trimeric PPα assembly.*

The simplest possible sequence combination for oPPα and a truncated trimer, CC-Tri3,[61] was explored. A PPII helix was appended to the *N*-terminus of CC-Tri3, connected *via* a `GDNAT` loop, and the *f* positions of the α helix were mutated to Tyr to give PPα-Tri-1 (table 3.1, figure 3.3). Ile at *a* and *d* dictate the oligomeric state of the coiled coil, and the flanking Glu and Lys residues can aid stabilisation of the assembly. The PPα-Tri-1 sequence resembled that of a mutant of oPPα, oPPα-A@*a* (table 3.1), more than oPPα itself. During the rational redesign of the PPα miniprotein, a number of mutations were made to the *g*, *a*, and *d* positions of the α helix in order

to determine which residues best stabilised the fold. In this monomeric system, it was found that Ala at *a* (corresponding to Ala at *c* in PPα-Tri-1) was destabilising.[147] It is possible that maintaining Ala at this position could likewise result in the destabilisation of the PPα trimer. As such, the next sequence combination incorporated additional residues from oPPα.

Of all the oPPα-X@*a* mutants (X = Leu, Ile, Val, Ala, Lys, and Glu) characterised, Leu at *a* yielded the most stable fold (T_M = 51 °C), followed by Ile, Val, and Lys (T_M = 40, 34 and 19 °C, respectively).[147] Ala at *a* displayed only partial folding at 5 °C, whereas the far-UV CD spectrum of Glu at *a* was characteristic of a random coil. In addition to the identity of the residue at the *a* position, salt studies conducted on oPPα indicated an electrostatic component to the stability of the miniprotein. However, the average distance of all Lys N_ζ/Glu C_δ pairs in the NMR ensemble of oPPα exceeded the accepted definition of a salt bridge (≤ 4 Å).[27] Long-range electrostatic interactions were posited to form between the Lys residues on the PPII helix and Glu residues at *e* and *g* on the α helix. Incorporating Glu at *b* (in addition to Glu at *g* in PPα-Tri-1) in the CC register would allow these long-range interhelical interactions to be mimicked in a PPα trimer. In addition to Glu at *b* and Leu at *c*, Pro was incorporated at the *N*-terminus of the α helix at position 14. Pro is highly conserved at this position in pancreatic-polypeptide like miniproteins, and helps to differentiate the start of the α helix from the preceding loop region.[159] Pro is commonly found between helical and non-helical regions in proteins, particularly at the *N*-terminus of α helices.[160] The result of these seven mutations yielded PPα-Tri-2 (table 3.1, figure 3.3).

PPα-Tri-2 showcases just one way that the sequences of oPPα and CC-Tri3 can be combined. It may be possible to shift the CC interface around the helix with respect to the PPα interface. This shift requires a swapping of the positions of the charged residues in the parent PPα-Tri-2, giving PPα-Tri-3 (table 3.1, figure 3.3).

Whilst CC-Tri is a robust trimer at both three and four heptads in length, the same is not true for other coiled coils. Pruning CC-Tet to three-heptads results in a trimeric assembly.[102] A more robust tetramer has since been designed. It was reasoned that an even more thermostable trimer could be produced from using the three-heptad CC-Tet (PPα-Tri-4, table 3.1, figure 3.3). The only difference between PPα-Tri-2 and PPα-Tri-4 is the identity of the residue at the *a* position in all heptads of the α helix: in PPα-Tri-4, this position is occupied by Leu, as opposed to its β-branched isomer Ile in PPα-Tri-2.

Table 3.1: *PPα, oPPα, CC-Tri3, and PPα-Tri sequences.*

Peptide	Sequence and Register
	efgabcd efgabcd efgabcd
PPα	PPTKPTKP GDNAT PEKLAKY QADLAKY QKDLADY
oPPα	PPKKPKKP GDNAT PEKLAAY EKELAAY EKELAAY
oPPα-A@a	PPKKPKKP GDNAT PEKAAAY EKEAAAY EKEAAAY
	gabcdef gabcdef gabcdef
CC-Tri3	G EIAAIKK EIAAIKQ EIAAIKQ GYG
PPα-Tri-1	PPKKPKKP GDNAT EIAAIKY EIAAIKY EIAAIKY
PPα-Tri-2	PPKKPKKP GDNAT PIELIKY EIELIKY EIELIKY
PPα-Tri-3	PPKKPKKP GDNAT PIELKIY EIELKIY EIELKIY
PPα-Tri-4	PPKKPKKP GDNAT PLELIKY ELELIKY ELELIKY

All peptides were *N*-terminally acetylated and amide-capped at the *C* terminus.

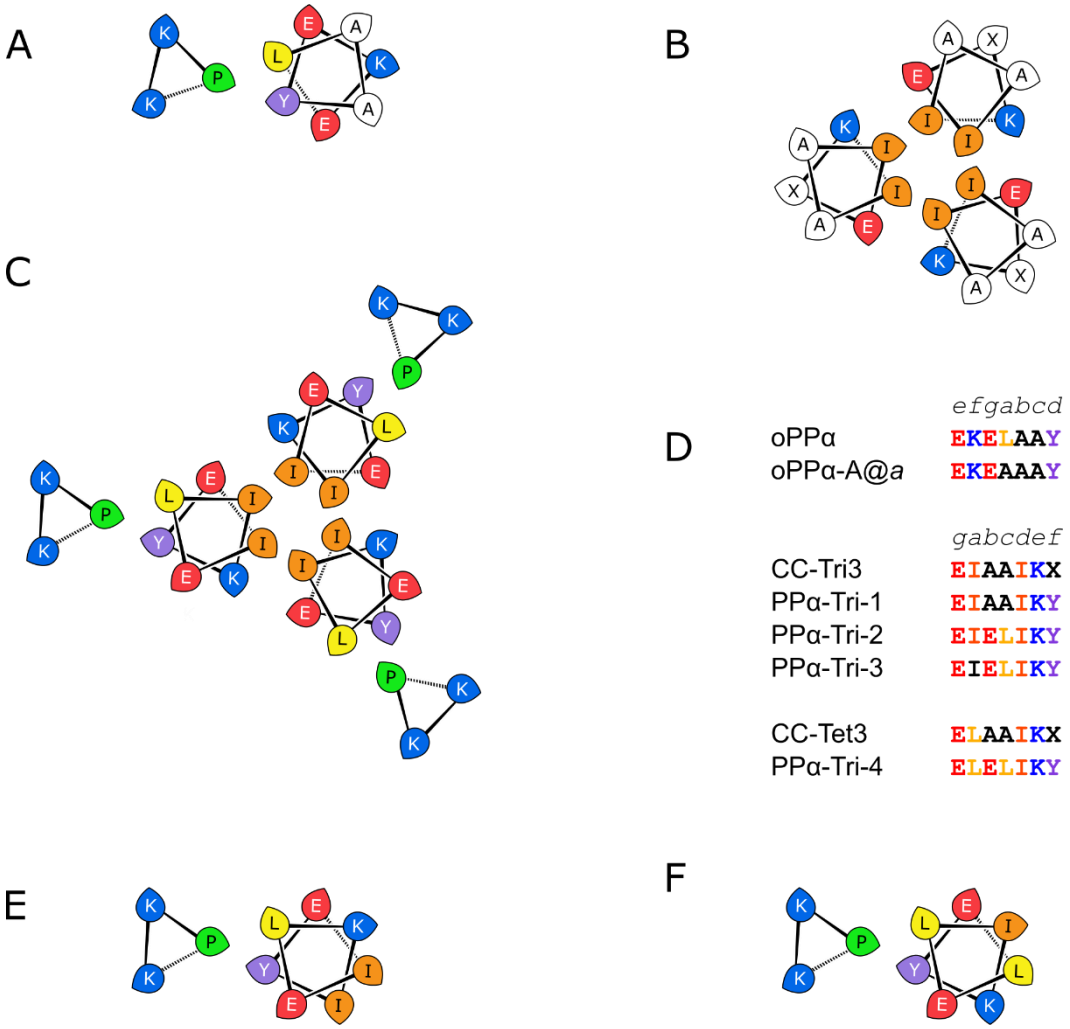


Figure 3.3: *Helical wheel representations for PPα-Tri peptides. Helical wheel representations of **A**, oPPα, **B**, CC-Tri3, and **C**, PPα-Tri-2. **D** Heptad sequences for the parent peptides and the resulting oligomers. Helical wheel representations of **E**, PPα-Tri-3 and **F**, PPα-Tri-4, with additional α and PPII helices omitted for clarity.*

3.2.2 Biophysical Characterisation of PPα-Trimers

PPα-Tri-1 was synthesised by solid-phase peptide synthesis, purified by reverse-phase HPLC, and its mass confirmed by MALDI-TOF mass spectrometry (section 7.4). Initial circular dichroism (CD) spectroscopy showed PPα-Tri-1 was folded with 46% helical character at 100 μM, similar

to oPPα, which was 50% helical under the same conditions (figure 3.4A). Both oPPα and CC-Tri3 unfold in a reversibly cooperative manner. However, the behaviour of PPα-Tri-1 in response to increasing temperature was quite different to that of its parents. Around 50 °C, the mean residue ellipticity (MRE) at 222 nm (a characteristic of α-helical secondary structure) reached a maximum, after which the peptide appeared to refold somewhat as the temperature was increased to 95 °C (figure 3.4B). The thermal midpoint of unfolding, T_M, of this shallow sigmoidal curve was determined to be 42 °C, lower than that of both oPPα and CC-Tri3 (T_M = 51 and 70 °C, respectively). After reducing the temperature to 5 °C, the typical α-helical CD spectrum with minima at 208 and 222 nm was not observed. Instead, only one minimum was observed at 216 nm (section 7.4), indicative of β structure. Far-UV CD spectra were subsequently recorded at 5 °C intervals from 5 to 95 °C to determine at what temperature the structural change occurs (figure 3.4C). At 35 °C, PPα-Tri-1 appears to be only partially folded; between 50 and 55 °C, the CD spectrum of PPα-Tri-1 morphs into one displaying a single minimum below 220 nm, correlating with the drop off in the signal at 222 nm at temperatures higher than 50 °C.

Both PPα and oPPα show characteristic CD signals in the near-UV region (340-250 nm), with minima around 275 nm (data for oPPα shown in figure 3.4D). These near-UV CD signals typically arise from chromophores that absorb around 280 nm i.e. Trp or Tyr which are embedded in tertiary structures.[161] In this case, the signal is due the Tyr residues presented by the α helix, which are surrounded by sidechains from both the α and PPII helices. There is no obvious near-UV signal from PPα-Tri-1 (figure 3.4D), suggesting the environment surrounding Tyr is different (and less structurally defined) from that of its parent. In PPα and oPPα, additional hydrophobic (Leu at *a*) and charged residues (Glu or Asp at *g*) occupy positions on the α helix in close proximity to Tyr. This is not the case in PPα-Tri-1, where Ala is present at both *e* and *g* positions (table 3.1). Thus, this

initial design may not contain enough information to robustly program the KIH-like packing of oPPα.

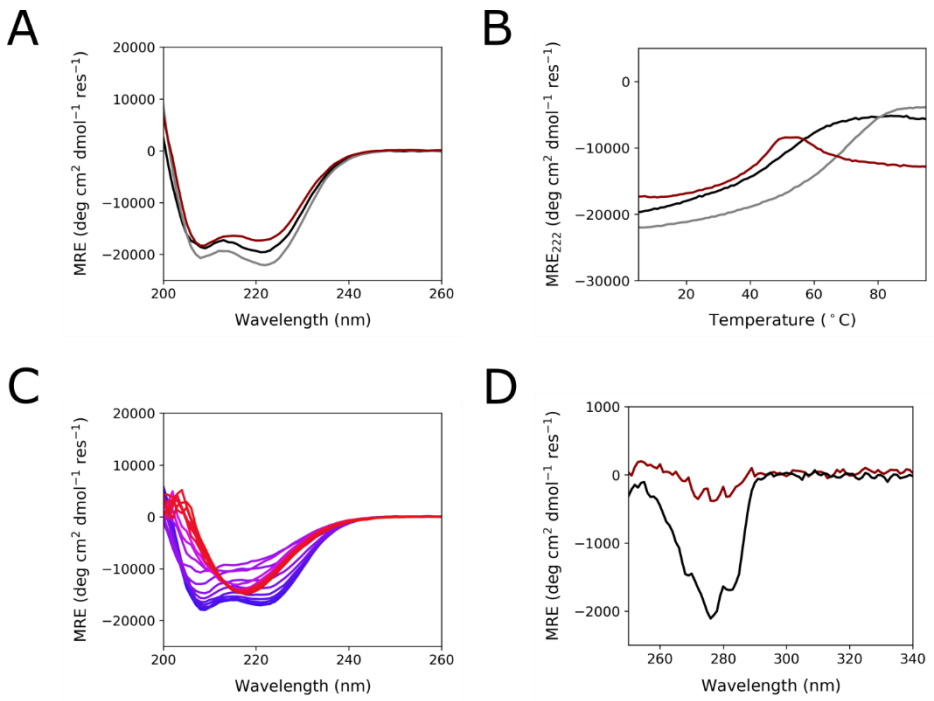


Figure 3.4: *Circular dichroism spectra for CC-Tri3, oPPα and PPα-Tri-1.* ***A****, Far-UV CD spectra recorded at 5 ºC and* ***B****, thermal denaturation profiles monitored at 222 nm for CC-Tri3 (grey), oPPα (black) and PPα-Tri-1 (maroon).* ***C****, Far-UV CD spectra recorded between 5 °C (blue) and 95 °C (red) at 5 °C intervals for PPα-Tri-1.* ***D****, Near-UV CD spectra recorded at 5 ºC for oPPα and PPα-Tri-1. Conditions: 100 μM peptide, PBS, pH 7.4. oPPα data were recorded by Kathryn Porter Goff, and previously published in Biochemistry.*[147]

Attempts were made to examine the oligomeric state of PPα-Tri-1 *via* sedimentation velocity analytical ultracentrifugation, but the peptide was aggregated, and the data could not be analysed.

With these results in hand, it was apparent that the sequence of PPα-Tri-1 was not ideal regards the aim of controlled oligomerisation of the optimised PPα miniprotein. In light of the low T_M and the lack of a near-UV CD signal,

it was obvious that further stabilisation of the fold was indeed necessary, as had been anticipated.

Initial characterisation of the second designed sequence, PPα-Tri-2, by CD spectroscopy showed that it was α-helical in solution, but slightly less well folded compared to PPα-Tri-1, with 43% fraction helix (figure 3.5A). Thermal denaturation of PPα-Tri-2, however, gave a T_M of 64 °C — closer to that of CC-Tri3 than oPPα (figure 3.5B). PPα-Tri-2 is not reversibly cooperative like its parents, and precipitated upon cooling from 95 to 5 °C. The heptad repeat of PPα-Tri-2 is more similar to what is expected of β structure, with alternating hydrophobic (*h*) and polar (*p*) residues (*hphphhp, a* to *g*), as opposed to the *hpphppp* (*a* to *g*) repeat typically observed in α-helical structures. However, the fact that this sequence adopts a helical conformation in solution at low temperature and is thermostable is encouraging.

Like oPPα, near-UV studies of PPα-Tri-2 showed a distinct signal with a minimum at 276 nm (figure 3.5C). Following the MRE at this wavelength also showed a sigmoidal unfolding curve, with a T_M of 77 °C, a 13 °C increase compared to the T_M measured by far-UV CD. This could indicate that the PPII helix unfolds prior to the α helix.

Analytical ultracentrifugation (AUC) experiments were undertaken to determine both the homogeneity and molecular weight of PPα-Tri-2 in solution. Sedimentation velocity experiments conducted at 100 μM revealed a single species present in solution, and the continuous c(s) fit to the data indicated that the peptide was trimeric (3.1 × monomer mass; figure 3.6A). This oligomeric state was subsequently confirmed by sedimentation equilibrium AUC. When fitted to an ideal, single species model, the data showed the molecular weight of the assembly to be 12.4 kDa i.e. a trimer, as designed (figure 3.6D).

The sequence of PPα-Tri-2 was thought to be too repetitive for solution structure determination by NMR, so sitting-drop crystallisation using numerous commercial screens was employed in an attempt to determine the 3D structure of PPα-Tri-2. However, these attempts were unsuccessful, and no crystals were obtained.

In the absence of a crystal structure, PPα-Tri-2 was subjected to a saturation binding assay with diphenylhexatriene (DPH), an environmentally sensitive dye which fluoresces only in a hydrophobic environment. DPH does not bind in a specific manner to CC-Tri, but binds specifically to coiled coils possessing hydrophobic channels with low micromolar K_d values.[112] As with CC-Tri, DPH did not show specific binding to PPα-Tri-2, indicating that the peptide assembly does not possess a hydrophobic pore (figure 3.7).

PPα-Tri-3, like its sister sequence PPα-Tri-2, was folded and thermostable (41% fraction helix, T_M = 65 °C), showed similar spectral features in the near-UV region, and was determined to be trimeric in solution (figures 3.5 and 3.6).

The last peptide in this series, PPα-Tri-4, was found to be more thermostable than all previous designs, with an increase in T_M of about 10 °C compared to PPα-Tri-2 and PPα-Tri-3 (PPα-Tri-4 T_M = 77 °C, figure 3.5). Like PPα-Tri-2 and -3, the peptide precipitated upon cooling from 95 to 5 °C. Sedimentation velocity and equilibrium experiments returned molecular weights of 11491 Da (2.8 × monomer mass) and 12493 Da (3.1 × monomer mass), respectively (figure 3.6C and 3.6F).

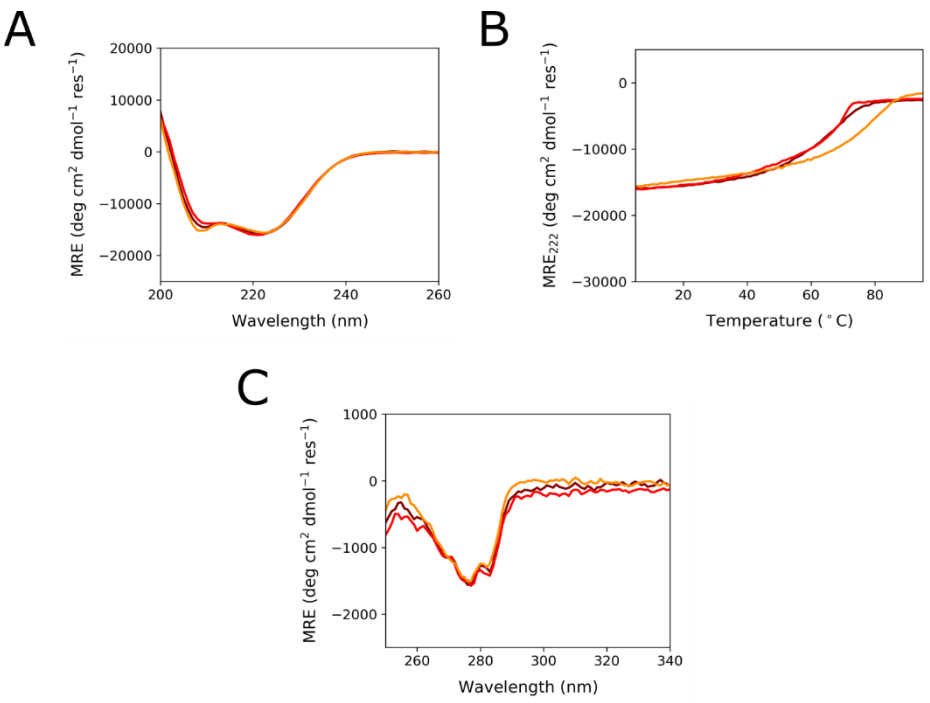


Figure 3.5: *Circular dichroism spectra for PPα-Tri-2, PPα-Tri-3 and PPα-Tri-4.* ***A****, Far-UV CD spectra recorded at 5 ºC,* ***B****, thermal denaturation profiles monitored at 222 nm, and* ***C****, near-UV CD spectra recorded at 5 ºC for PPα-Tri-2 (brown), PPα-Tri-3 (red) and PPα-Tri-4 (orange). Conditions: 100 μM peptide, PBS, pH 7.4.*

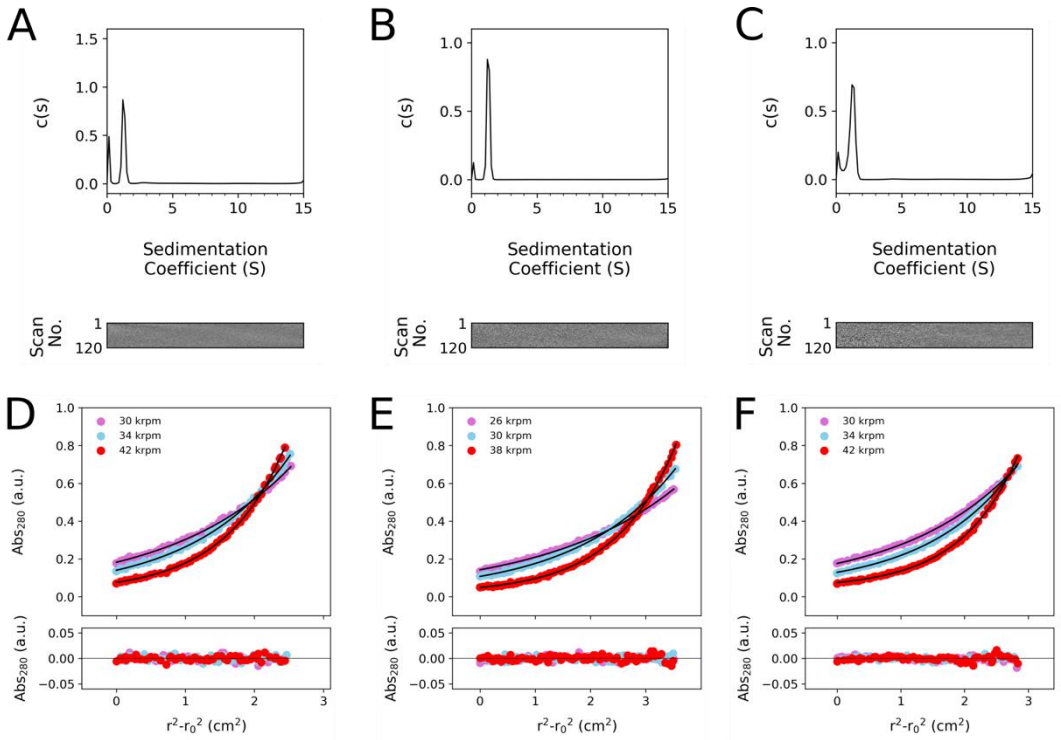


Figure 3.6: *Sedimentation velocity and equilibrium AUC data for PPα-Tri-2, PPα-Tri-3, and PPα-Tri-4 ($\bar{\nu}$ = 0.7726 cm^3 g^{-1}).* ***A****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Tri-2 at 60 krpm returning Mw = 12368 Da (3.1 × monomer mass), 95% confidence limits.* ***B****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Tri-3 at 60 krpm returning Mw = 13326 Da (3.3 × monomer mass), 95% confidence limits.* ***C****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Tri-4 at 50 krpm returning Mw = 11491 Da (2.8 × monomer mass), 95% confidence limits.* ***D****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 30, 34 and 42 krpm for PPα-Tri-2, returning Mw = 12604 Da (3.1 × monomer mass, 95% confidence limits: 12499 – 12713 Da), and residuals for the fits (bottom, dots).* ***E****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 26, 30 and 38 krpm for PPα-Tri-3, returning Mw = 12388 Da (3.1 × monomer mass, 95% confidence limits: 12283 – 12495 Da), and residuals for the fits (bottom, dots).* ***F****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 30, 34 and 42 krpm for PPα-Tri-4, returning Mw = 12493 Da (3.1 × monomer mass, 95% confidence limits: 12393 – 12591 Da), and residuals for the fits (bottom, dots). Conditions: 100 μM peptide for SV AUC and 100 or 110 μM peptide for SE AUC, PBS, pH 7.4.*

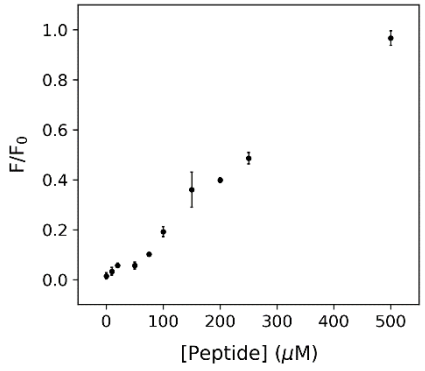


Figure 3.7: *Saturation binding curve for DPH with PPα-Tri-2. Conditions: 0-500 μM peptide, 1 μM DPH, PBS with 5 v/v% DMSO, pH 7.4.*

In summary, incorporating an increasing number of residues which stabilise a PP:α interface into three-heptad trimeric coiled coils is well tolerated, and, whilst small decreases in thermostability can be observed with respect to the parent coiled coil, the desired oligomeric state is maintained. There is also no one optimal positioning of the CC interface with respect to the PP:α interface — both options explored were equally thermostable. Now that this design strategy has been successfully realised, it can be applied towards constructing larger templates for single-chain α-helical barrels.

3.3 Combining PPα with *De Novo* Designed α-Helical Barrels

α-Helical barrels are coiled-coil assemblies consisting of five or more helices which possess a central hydrophobic pore, the diameter of which increases with increasing oligomer state. αHBs are accessible to small molecules, such as DPH, as well as biologically-relevant fatty acids and terpenoids.[73, 101, 112] In addition to the previously computationally designed pentamers, hexamers and heptamers, octamers and nonamers have also been determined.[100, 101] The formation of the larger αHBs is condition-dependent: heptamers form in tris- or HEPES-buffered saline, but not

phosphate-buffered saline, and the sequences which crystallise as assemblies of more than seven helices form hexamers in solution.

A heptamer is a particularly interesting target when considering functionalisation of and applications for PPα oligomers. CC-Hept, with its larger pore, is amenable to binding a larger variety of substrates compared with hexamers or pentamers. Additionally, CC-Hept has previously been transformed into a rudimentary hydrolase *via* the introduction of seven Cys-His-Glu triads to the core.[113] CC-Hept-CHE hydrolyses simple esters, albeit with low catalytic efficiency (1000 × less) compared to α-chymotrypsin, a natural serine hydrolase, against the same substrate.[116,117]

Computational transition state modelling in RosettaMatch suggests that catalytic activity towards *para*-nitrophenyl acetate could be boosted by spreading the triad across three adjacent helices, and reducing steric bulk in the pore as necessary for good substrate binding.[118] However, to date there has been no way to translate this computationally-supported hypothesis to *in vitro* experiments. A heteromeric seven-membered α-helical barrel would first need to be realised. The highest order, parallel hetero-oligomer in the CC+ database (≤ 50% redundancy, canonical repeat, > 11 residues; last updated on the 9th January 2020) is an A_3B_3 hexamer, in which the sequences of the A and B chains differ by one residue (residue 24 = His or Asp; PDB ID 3R48).[99, 162] There are no *ABCDEF* parallel, hetero-hexamers, nor any *ABCDE* parallel, hetero-pentamers listed in the database, meaning that a completely heteromeric seven-helix (*ABCDEFG*) parallel barrel would need to be designed from scratch. A second option is the construction of a seven-pronged template which maintains the parallel arrangement of the component helices. Alternatively, an antiparallel approach could be considered, but the *de novo* design of obligate antiparallel coiled coils remains in its infancy. Additionally, such an approach is more suited to barrels containing an even number of helices, to accommodate $(up\text{-}down)_n$ geometry. A heptameric PPα oligomer would

therefore provide a solid foothold towards achieving a single-chain, parallel, seven-helix barrel, making use of the antiparallel arrangement of PPII helices along the solvent-exposed face of the heptamer helices. Such a template would also offer a starting point from which to begin improving the catalytic efficiency of rudimentary αHB catalysts.

3.3.1 PPα-Barrel Design

Unlike for the trimeric designs described *vide supra*, a direct combination of the sequences of oPPα and CC-Hept to create a PPα heptamer (figure 3.8) is not as simple. This is due to the register of CC-Hept beginning at a *c* position as opposed to a *g* position like CC-Tri3. If maintaining the CC-Hept register at *c*, there is no combination of the two sequences which allows for all the desired residues to be carried through, as Lys at *b* in CC-Hept must be substituted for Tyr so that the PP:α interface can be maintained (PPα-Hept-1, figure 3.9A and 3.9B). If the register is changed to start at a *g* position, as for the trimeric designs, all charged residues can be carried through, at the expense of the Leu residue that best stabilises the PPII:loop:α fold. This results in Glu being placed at *c*. At the equivalent *a* position in oPPα, Glu completely destabilised the tertiary structure of the miniprotein; no α-helical secondary structure was observed by CD spectroscopy.[147] However, Lys was tolerated at this position, although with significantly reduced thermostability compared to oPPα-Leu@*a* (ΔT_M = 32 °C). Swapping the charged residues in the combined sequence would circumvent destabilisation of the local secondary structure by Glu. It was expected that the sequence would be tolerant of this Glu/Lys swap. The resulting sequence is PPα-Hept-2 (figure 3.9 and 3.9C). Pro was again included at the first *g* position of the α helix (residue 14).

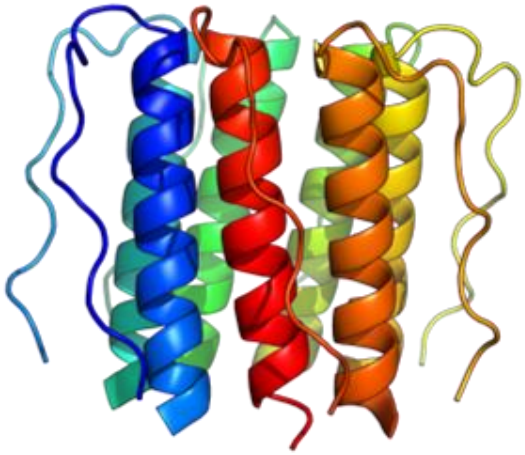


Figure 3.8: *A model of a heptameric PPα assembly.*

Whilst compromises would have to be made for both PPα-Hept-1 and PPα-Hept-2, efforts were focused on the characterisation of PPα-Hept-2 (henceforth renamed PPα-Hept, table 3.2), as it maintains both charged residues from the CC-Hept sequence. A decision was made to constrain the length of the α-helical segment to three heptads, as for previous designs. At this stage, uncertainty remained over whether a register rearrangement and truncation of the CC-Hept sequence (in isolation or in combination, as proposed here) would result in a switching of the oligomeric state. However, regards the aim of constructing single-chain parallel α-helical barrels, it was expected that the number of helices in the barrel (akin to the oligomeric state of the peptide assembly) could be determined by sequence length. If PPα-Hept resulted in a barrel assembly with an open conformation, the design could be considered a success.

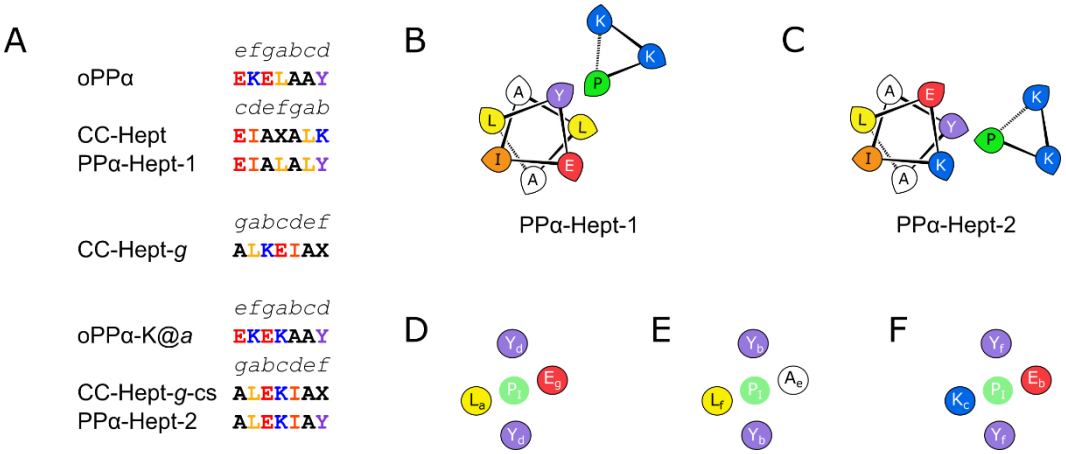


Figure 3.9: *Possible sequence combinations and helical wheel representations for PPα-Hept.* ***A****, Possible combinations of oPPα and CC-Hept sequences. Helical wheel representations for* ***B****, PPα-Hept-1 and* ***C****, PPα-Hept-2. Cartoon diamond-shaped holes representing the interhelical packing of* ***D****, oPPα,* ***E****, PPα-Hept-1, and* ***F****, oPPα-K@a and PPα-Hept-2.*

Table 3.2: *oPPα, oPPα-K@a, CC-Hept, and PPα-Hept sequences.*

Peptide	Sequence and Register
	efgabcd efgabcd efgabcd
oPPα	PPKKPKKP GDNAT PEK**L**AAY EKE**L**AAY EKE**L**AAY
oPPα-K@*a*	PPKKPKKP GDNAT PEK**K**AAY EKE**K**AAY EKE**K**AAY
	cdef gabcdef gabcdef gabcdef gab
CC-Hept	G EIAQ ALKEIAK ALKEIAW ALKEIAQ ALK G
PPα-Hept	PPKKPKKP GDNAT PLEKIAY ALEKIAY ALEKIAY

All peptides were *N*-terminally acetylated and amide-capped at the *C* terminus.

3.3.2 Biophysical Characterisation of Higher Order PPα Oligomers

Characterisation of PPα-Hept by circular dichroism spectroscopy showed that the peptide was well-folded (49% fraction helix at 5 °C, figure 3.10A)

and subsequent monitoring of the peptide under increasing temperature revealed that PPα-Hept was hyper-thermostable, like CC-Hept, with a T_M greater than 95 °C (figure 3.10B). Examining the near-UV CD spectrum of PPα-Hept showed a characteristic signal from Tyr, which also showed no sigmoidal transition when monitored at increasing temperatures (figures 3.10C and D).

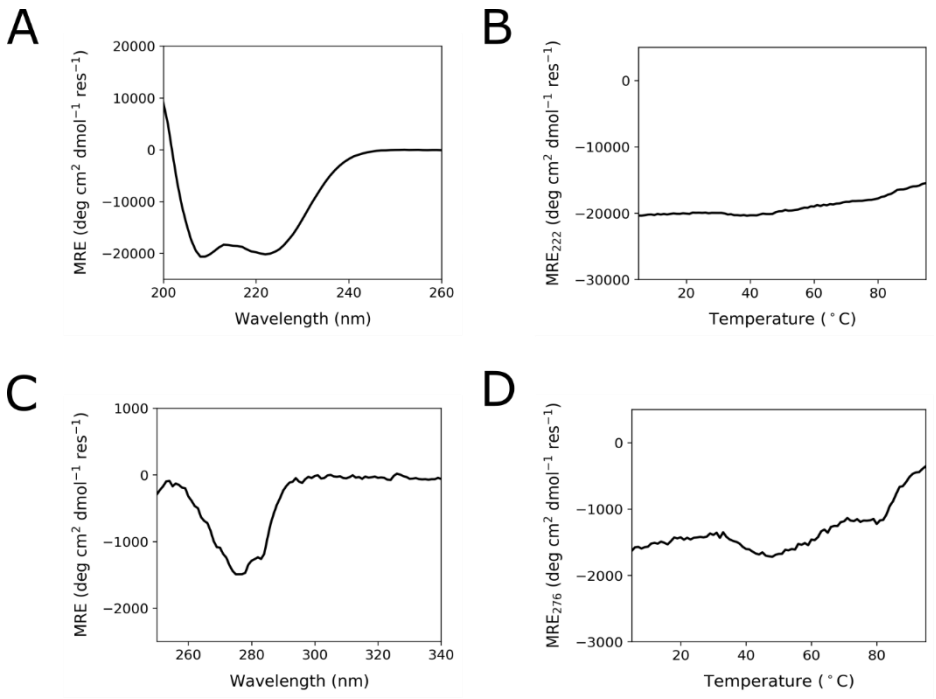


Figure 3.10: *Circular dichroism spectroscopy data for PPα-Hept.* ***A****, Far-UV CD spectra recorded at 5 ºC,* ***B****, profile monitored at 222 nm between 5 and 95 °C,* ***C****, near-UV CD spectra recorded at 5 ºC, and* ***D****, profile monitored at 276 nm.*

Encouraged by these results, analytical ultracentrifugation experiments were conducted. Continuous c(s) fits to sedimentation velocity data collected at 50 krpm in tris-buffered saline indicated that, whilst the sample was not homogenous, the expected molecular weight of the predominant species was 5.9 × monomer mass (figure 3.11A). A second, lower

concentration species was estimated to be 13.3 × monomer mass. Sedimentation equilibrium data recorded in phosphate-buffered saline did not fit well to a single species model. The residuals were non-random and displayed a positive arch, suggesting non-ideal behaviour. Fitting the data to a two-species model and floating the molecular weights obtained from SV AUC experiments resulted in a much better fit to the data and indicated that the major component was hexameric in solution (6.1 × monomer mass, figure 3.11B). The second species was estimated to be 17.6 × monomer mass.

Despite associating into a hexameric assembly as opposed to the desired heptamer, PPα-Hept exhibits the hallmarks of a PP:α interface and the thermostability of an αHB, as well as self-associating into a barrel-sized oligomer. It is possible that lengthening the sequence by one heptad may restore the desired oligomeric state. However, the fact that the peptide oligomer does not behave as was initially hoped is not a cause for concern, as it may be possible to force the formation of a seven membered single-chain αHB simply by encoding seven PPα units.

Despite attempts using commercial screens, sitting-drop crystallisation trials yielded no crystals of PPα-Hept, and thus its solid-state structure was unable to be determined.

Considering that PPα-Hept can self-associate into a hexamer, it was imperative to ascertain whether this assembly is accessible to small molecules. Preliminary data indicated that PPα-Hept was capable of binding 1,6-diphenylhextriene in a specific manner (figure 3.11C), indicating that it adopts an open conformation in solution, like many previous αHBs characterised by the Woolfson group

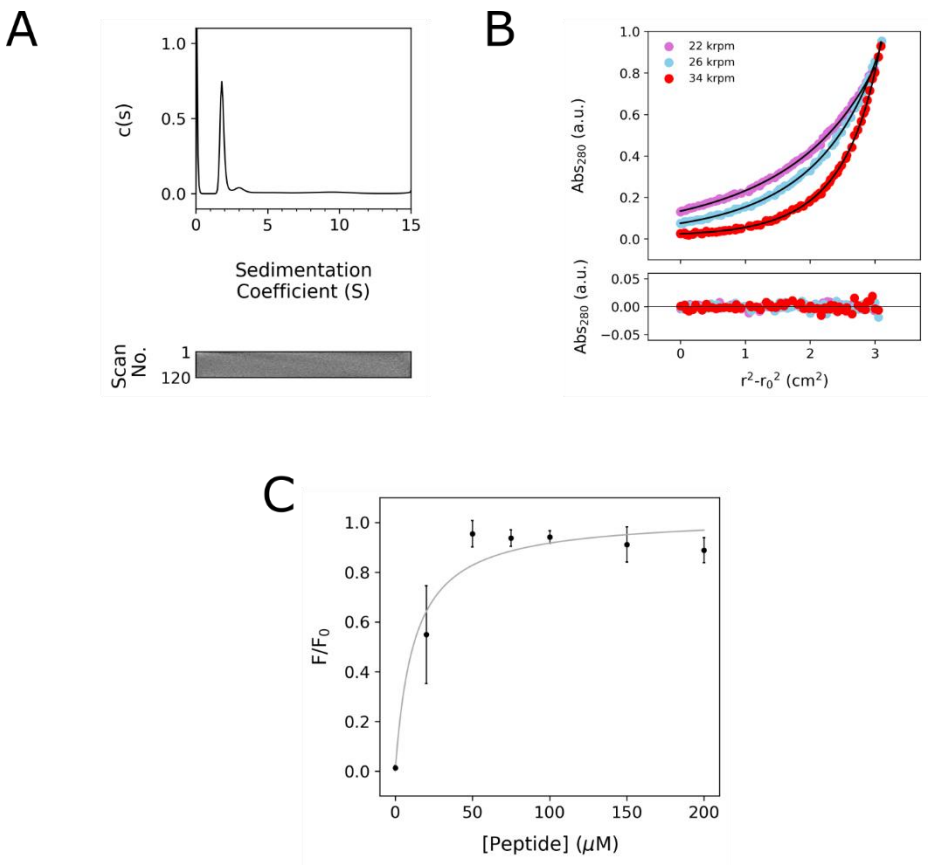


Figure 3.11: *AUC data and saturation binding curve for PPα-Hept ($\bar{\nu}$ = 0.7658 cm^3 g^{-1}).* ***A****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Hept at 50 krpm returning Mw = 22448 Da (5.9 × monomer mass), 95% confidence limits. Conditions: 100 μM peptide, TBS, pH 7.4.* ***B****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 22, 26 and 34 krpm for PPα-Hept, returning Mw = 23200 Da (6.1 × monomer mass, 95% confidence limits: 21718 – 24300 Da), and residuals for the fits (bottom, dots). Conditions: 110 μM peptide, PBS, pH 7.4.* ***C****, Saturation binding curve for DPH with PPα-Hept. Conditions: 0-200 μM peptide, 1 μM DPH, PBS with 5 v/v% DMSO, pH 7.4.*

As DPH binding indicated an accessible pore, we wondered whether the open channel of PPα-Hept could accommodate the same Cys-His-Glu triad that was introduced into CC-Hept.[113] Whilst CC-Hept was able to tolerate these polar mutations to its non-polar core (as confirmed by X-ray

crystallography; PDB ID 5EZC), the introduced triad had a negative effect on the thermostability of the peptide, reducing the T_M from > 95 °C to 57 °C.

Mutations to PPα-Hept were made in keeping with the *C*-terminal *d-a-d'* placement of the triad in CC-Hept-CHE (table 3.3). However, these polar mutations destabilised the fold significantly. The peptide was only partially folded at 5 °C at 100 μM (figure 3.12), and a T_M could not be determined as a result.

Table 3.3: *CC-Hept-CHE and PPα-Hept-CHE sequences.*

Peptide	Sequence and Register
	cdef gabcdef gabcdef gabcdef gab
CC-Hept-CHE	G EIAQ ALREIAK ALRE**C**AW A**H**RE**E**AQ ALR G
	gabcdef gabcdef gabcdef gab
PPα-Hept-CHE	PPKKPKKP GDNAT PLEKIAY ALEK**C**AY A**H**EK**E**AY ALK

All peptides were *N*-terminally acetylated and amide-capped at the *C* terminus.

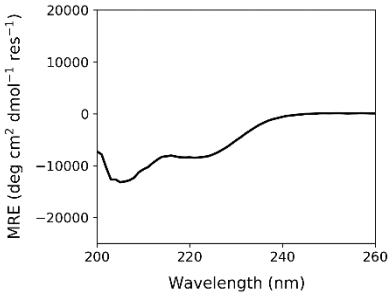


Figure 3.12: *Far-UV circular dichroism spectrum of PPα-Hept-CHE recorded at 5 °C. Conditions: 100 μM peptide in PBS, with 10 × excess TCEP relative to peptide concentration, pH 7.4.*

As mentioned previously, Ala at *g* specifies a heptamer in a four-heptad *g*LKEIAX (*gabcdef*) system, where X = Gln, Lys, or Trp. αHBs of decreasing oligomer state can be realised by increasing the steric bulk of the side chain at *g* i.e. by the successive addition of either carbon or oxygen.[101] Thr at *g* dictates a pentamer (CC-Pent2); Ser at *g* dictates a hexamer (CC-Hex2); and Gly at *g* gives a nonamer in the solid state (but a hexamer in solution).

Given the success of PPα-Hept, PPα-Pent2, -Hex2 and -Non were synthesised and characterised (table 3.4; figure 3.13).

Table 3.4: *PPα-Pent2, -Hex2, and -Non sequences.*

Peptide	Sequence and Register
	gabcdef gabcdef gabcdef
PPα-Pent2	PPKKPKKP GDNAT PLEKIAY **T**LEKIAY **T**LEKIAY
PPα-Hex2	PPKKPKKP GDNAT PLEKIAY **S**LEKIAY **S**LEKIAY
PPα-Non	PPKKPKKP GDNAT PLEKIAY **G**LEKIAY **G**LEKIAY

All peptides were *N*-terminally acetylated and amide-capped at the *C* terminus.

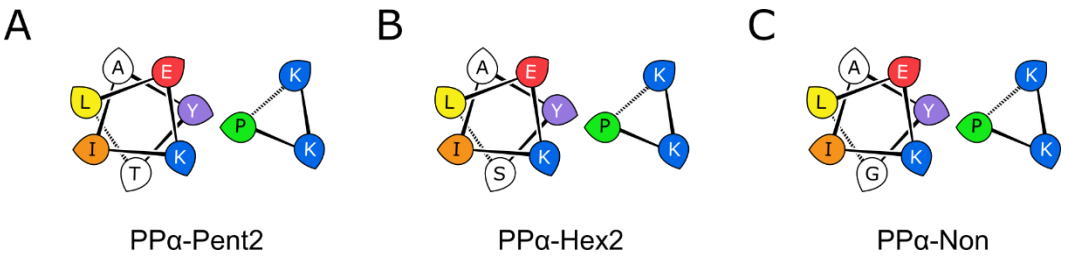


Figure 3.13: *Helical wheel representations of **A**, PPα-Pent2, **B**, PPα-Hex2 and **C**, PPα-Non.*

Because Pro is required at the first *g* position of the α-helical segment of PPα-like folds to discern the loop sequence and the beginning of the helix, only two mutations per helix (in the remaining *g* positions) were made. The peptides were all well-folded in solution (figure 3.14A) and showed near-UV CD signals characteristic of the PPα miniprotein family (figure 3.14C). PPα-Pent2 and -Hex2 are hyper-thermostable in solution, and did not unfold between 5 and 95 °C. PPα-Non unfolded reversibly with a T_M of 77 °C (figure 3.14B).

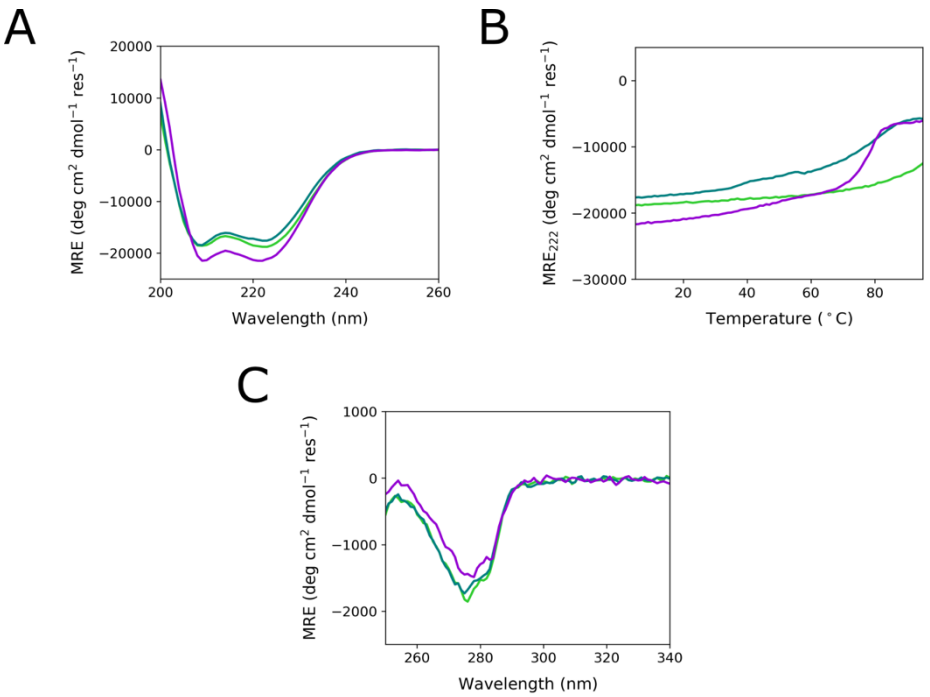


Figure 3.14: *Circular dichroism spectra for PPα-Pent2, PPα-Hex2 and PPα-Non. **A**, Far-UV CD spectra recorded at 5 ºC, **B**, thermal denaturation profiles monitored at 222 nm, and **C**, near-UV CD spectra recorded at 5 ºC for PPα-Pent2 (green), PPα-Hex2 (blue), and PPα-Non (purple). Conditions: 100 μM peptide, PBS, pH 7.4.*

However, sedimentation velocity AUC experiments showed that only PPα-Pent2 existed as a single species in solution (4.9 × monomer mass) at 100 μM (figure 3.15A). Sedimentation equilibrium data was obtained for

PPα-Hex2 at the same concentration. When the data was fitted to an ideal, single-species model, a molecular weight of 22178 Da, 5.7 × the monomeric mass, was returned (figure 3.15C).

Like PPα-Hept, PPα-Hex2 did not form a single species in solution, and the molecular weights of these two species corresponded roughly to a monomer and a dimer (5.3 and 11.6 × monomer mass, respectively), with the monomer being the most abundant species (figure 3.15B). Again, fits to a single, ideal species model indicated non-ideal behaviour. Fitting instead to an ideal, two-species model returned molecular weights of 24941 and 70760 Da (6.5 and 18.4 × monomer mass, respectively), suggesting that the three-heptad PPα-Hex2 behaves somewhat promiscuously in solution. Attempts were made to examine PPα-Non's solution-phase behaviour, but the peptide aggregated in the ultracentrifuge during temperature equilibration at 3 krpm, precluding molecular weight determination.

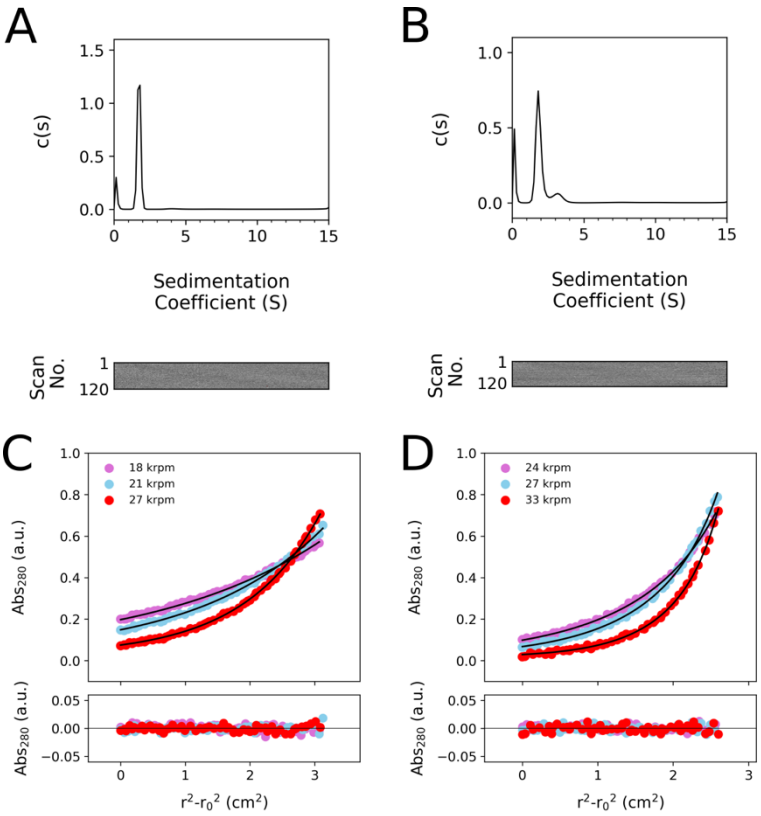


Figure 3.15: *Sedimentation velocity and equilibrium AUC data for PPα-Pent2 ($\bar{\nu}$ = 0.7633 cm^3 g^{-1}) and PPα-Hex2 ($\bar{\nu}$ = 0.7605 cm^3 g^{-1}).* ***A****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Pent2 at 50 krpm returning Mw = 18918 Da (4.9 × monomer mass), 95% confidence limits.* ***B****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for PPα-Hex2 at 50 krpm returning Mw = 20378 Da (5.3 × monomer mass), 95% confidence limits.* ***C****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 18, 21 and 27 krpm for PPα-Pent2, returning Mw = 22178 Da (5.7 × monomer mass, 95% confidence limits: 21886 – 22469 Da), and residuals for the fits (bottom, dots).* ***D****, SE data (top, dots) fitted to ideal, two-species model curves (top, black lines) at 24, 27 and 33 krpm for PPα-Hex2, returning Mw = 24941 Da (6.5 × monomer mass, 95% confidence limits: 23733 – 25730 Da) for the major species, and residuals for the fits (bottom, dots). Conditions: 100 μM peptide for SV and and SE AUC, PBS, pH 7.4.*

As with PPα-Hept, as PPα-Hex2 potentially formed a barrel-sized structure in solution, a saturation binding assay was undertaken in the presence of DPH to determine if the channel was open or closed to small molecules. This assembly also bound DPH in a specific manner, like PPα-Hept,

indicating it is adopted an open conformation in solution (figure 3.16). The binding of DPH by PPα-Pent2 is yet to be probed.

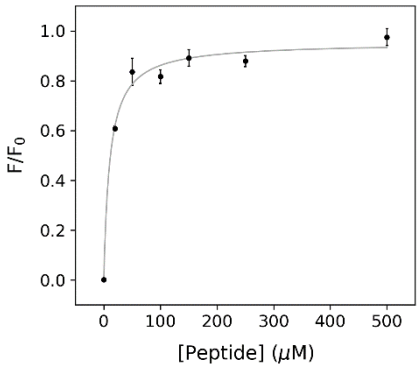


Figure 3.16: *Saturation binding curve for DPH with PPα-Hex2. Conditions: 0-200 μM peptide, 1 μM DPH, PBS with 5 v/v% DMSO, pH 7.4.*

Table 3.5: *A summary of solution-phase characterisation data for the PPα oligomers.*

Peptide	Sequence	Obs. Mass (Da)	Fraction Helix (%) [a]	T_M (°C) [b] Far- / Near-UV	O.S. [c] SV / SE
PPα-Tri-1	PPKKPKKP GDNAT EIAAIKYEIAAIKYEIAAIKY	3783.565	46 ± 3.6	42 ± 0.0 / -	ND / -
PPα-Tri-2	PPKKPKKP GDNAT PIELIKYEIELIKYEIELIKY	4051.362	43 ± 0.6	64 ± 1.2 / 77 ± 1.0	3.1 / 3.1
PPα-Tri-3	PPKKPKKP GDNAT PIELKIYEIELKIYEIELKIY	4051.319	41 ± 1.2	65 ± 1.2 / 83 ± 0.6	3.3 / 3.1
PPα-Tri-4	PPKKPKKP GDNAT PLELIKYELELIKYELELIKY	4051.877	43 ± 2.1	78 ± 1.0 / 79 ± 2.1	2.8 / 3.1
PPα-Hept	PPKKPKKP GDNAT PLEKIAYALEKIAYALEKIAY	3809.819	49 ± 3.5	≥ 95 / ≥ 95	5.9 / 6.1
PPα-Hept-CHE	PPKKPKKP GDNAT PLEKIAYALEKCAYAHEKEAYALK	4150.473	22	≤ 5 / -	- / -
PPα-Pent2	PPKKPKKP GDNAT PLEKIAYTLEKIAYTLEKIAY	3868.650	47 ± 1.5	≥ 95 / ≥ 95	4.9 / 5.7
PPα-Hex2	PPKKPKKP GDNAT PLEKIAYSLEKIAYSLEKIAY	3841.877	49 ± 2.6	≥ 95 / ≥ 95	5.3 / 6.5
PPα-Non	PPKKPKKP GDNAT PLEKIAYGLEKIAYGLEKIAY	3780.971	54 ± 2.9	77 ± 0.6 / 78 ± 3.5	Agg. / -

[a] Calculated using *Equation 2.3*. [b] Calculated from the 1st derivative of thermal denaturation profiles (mean ± std.). [c] O.S. = oligomeric state. Conditions: 100 or 110 μM peptide, PBS, pH 7.4. Obs. = observed. SV = sedimentation velocity AUC. SE = sedimentation equilibrium AUC. ND = not determined. Agg. = aggregated at 3 krpm.

3.4 Conclusions

In conclusion, simple selection and combination of residues which direct the folding of an entirely *de novo* miniprotein and a toolkit of designed coiled coils has resulted in numerous PPα-oligomers. Far-UV CD spectra of this limited palette of three-heptad oligomers show that these peptides adopt α-helical secondary structure in solution and are as folded as the optimised PPα miniprotein. Attempts to thermally denature the peptides has shown that these oligomers are highly thermostable. In this regard, they behave more like coiled coils than oPPα its derivatives, with only a slight reduction in T_M observed cf. the coiled coil parent. Near-UV circular dichroism spectroscopy indicates that the antiparallel arrangement of the PPII and α helices in optimised PPα is maintained in these oligomers. Analytical ultracentrifugation experiments have shown that several of the peptide sequences described herein self-associate predictably into discrete assemblies, although more complex (and promiscuous) behaviour is observed in the three-heptad oligomers based upon longer, more stable coiled-coil barrels. Additionally, preliminary saturation binding experiments with 1,6-diphenylhexatriene indicate that the oligomers inspired by the α-helical barrels are capable of binding small molecules. However, polar mutations to the core of a PPα-hexamer significantly destabilised the core, such that the peptide was largely unfolded. These data indicate that just 34 residues can encode an inverted TIM barrel-type structure. Unfortunately, attempts to crystallise the oligomers have so far been unsuccessful. Despite this, the biophysical characterisation reported herein indicates a solid first step towards templating single-chain, parallel α-helical barrels with peptide oligomers.

Chapter 4

Modelling PPα Oligomers

The work described in this chapter was conducted by the author of this thesis. Dr Christopher Wood wrote the Python script for building parametric models of PPα oligomers in ISAMBARD.

4.1 Introduction

There were two reasons for pursuing the single-chain, parallel α-helical bundles and barrels that are the subject of this thesis: Firstly, they present a protein design challenge. Secondly, the accessible hydrophobic cores of these proteins would be amenable to single-point mutations (a possibility excluded to the cores of coiled-coil oligomers due to their symmetric nature), meaning that catalysis or small-molecule binding in the core could be controlled more precisely. The successful oligomerisation of the optimised PPα miniprotein described in chapter 3 represents the completed first step of the process; *i.e.* the design of a homo-oligomeric template. The second step towards this goal is the concatenation of these template subunits into one polypeptide chain. This stepwise route is akin to the fusion of small, typically homo-oligomeric proteins over time leading to the

abundance of large, monomeric functional proteins observed in nature today.[163]

Unfortunately, all attempts at obtaining crystals of the oligomeric peptide templates for X-ray crystallography were unsuccessful. Additionally, their sequences were considered too repetitive for solution-phase structure determination by NMR spectroscopy. As these peptide oligomers ultimately form the basis of a template strategy towards single-chain, parallel α-helical bundles and barrels, having no means to visualise the template at atomic resolution complicates the remaining steps of the construction of these *de novo* proteins *i.e.*, modelling suitable connecting loop geometries and sequences. Therefore, this chapter explores a parametric route to modelling PPα oligomers, from which models of single-chain parallel α-helical bundles and barrels can be built.

4.1.1 Parametric Modelling

Parametric modelling is simply the *in silico* generation of biomolecular backbones that can be described mathematically by a set of parameters. Single α and polyproline-II (PPII) helices can both be described relatively simply by a small number of parameters, including: the number of residues per turn (3.6 and 3.0, respectively); the rise per residue (1.5 and 3.1 Å, respectively), and the radius of the helix (2.3 and 1.6 Å, respectively). Symmetric oligomers of these secondary structures, such as coiled coils (CCs) and collagens, trimeric assemblies of PPII helices,[157] can be modelled accurately using a small number of additional parameters.[78] For example, blunt-ended CCs can be encapsulated in just three parameters: the radius of the assembly (the distance in Ångstroms from the centre of one of the component helices to the centre of the assembly, figure 4.1); the pitch (the distance in Ångstroms over which a component helix turns 360 ° about the super-helical axis, figure 4.1), and the interface angle (the angle subtended by the C_α atom of a residue at a given position and the centre of

the assembly, figure 4.1).[164] Collagens can be modelled accurately using the same three parameters, as well as including a z-shift term to describe the leading and lagging strands, and a rotational offset for each helix.[78] Thus, models of PPα oligomers, comprising two parameterizable helix types, can also be built parametrically.

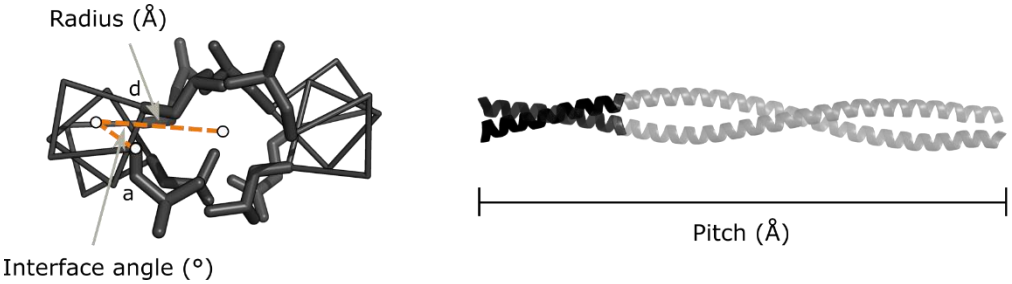


Figure 4.1: *Parameters required to describe a coiled coil. Left, the radius is the distance between the centre of the coiled-coil assembly and the centre of a constituent helix, and the interface angle is defined as the angle between the centre of the assembly, the centre of the helix and the position of a C_α carbon of one of the sidechains (e.g. at residue position* a*, as shown here). Right, the pitch is the distance it would take for a coiled coil (blocked out in black) to turn 360 ° about the super-helical axis (highlighted in grey). The backbone is shown as a ribbon, and sidechains are shown as sticks.*

4.1.2 Introduction to ISAMBARD

ISAMBARD (*Intelligent System for Analysis, Model Building and Rational Design*) is a Python-based package developed in the Woolfson lab to generate and optimise parametric models of proteins and other biomolecules.[78] In ISAMBARD, biomolecules are built following the hierarchical AMPAL (Atom, Monomer, Polymer, Assembly and Ligand) framework. The updated ISAMBARD package (v2.0.0) is split into various modules which can be used either in isolation (for structural analysis, for example) or in combination to generate and optimise models. The `specifications` module defines numerous biomolecule assembly

classes (including coiled coils, collagens, α-solenoid structures, and single- and double-stranded DNA) and the parameters required to build these assemblies. The `modelling` module contains functionality pertaining to the modelling of side chains with the external program SCWRL4, which utilises a backbone-dependent rotamer library to assign specific side-chain rotamers to a polypeptide backbone.[165] The `evaluation` module contains tools to aid the evaluation of generated models. The `optimisation` module introduces numerous evolutionary optimisation algorithms for use in optimisation of a generated model using a selected force field (BUDE, Bristol University Docking Engine, is ISAMBARD's default force field).[166, 167] Finally, the standalone `ampal` module (https://isambard-uob.github.io/ampal/) generates hierarchical AMPAL objects from PDB files, and contains a set of tools for analysing imported or generated biomolecular structures.

4.1.3 Loop Closure Methods

Whilst a powerful tool for modelling proteins with regular or super-secondary structure, parametric modelling is limited to the description of secondary structure elements (SSEs). Any polypeptide linkers between two parametrizable SSEs must be modelled separately and then concatenated to the main chain to create a protein model. However, there are many means by which loop regions can be added to and modelled on existing structures, meaning that, whilst it presents an additional step in the modelling process, this issue can largely be sidestepped.

The *loop closure problem* was first described by Go and Scheraga in 1970, when considering the closure of linear molecules and polypeptides.[168] The problem, however, is not limited to achieving cyclic molecules and polymers, but also extends to *de novo* structure prediction, homology modelling, and protein design. Over the years, numerous solutions of varying degrees of complexity and computational expense have been

proposed. In their 1970 paper, Go and Scheraga noted that the problem of ring closure is greatly simplified when considering only the dihedral angles, as opposed to the Cartesian coordinates of all atoms involved.[168] Building a loop from *N*- and *C*-terminal peptide fragments simultaneously to converge at a point between the two, as demonstrated by DePristo *et. al*, is simple in theory, but sampling intensive in practice.[169] Inverse kinematic algorithms, adapted from the field of robotics, take steps to simplify the problem even further. For example, in cyclic coordinate descent (CCD), variable dihedral angles are changed one at a time, and the process is iterated until the loop is closed, providing, of course, that the loop is of sufficient length to join the protein backbone segments.[170] CCD has been adapted to solve the loop closure problem for protein models containing only C_α atoms — this technique is known as 'full CCD'.[171] Kinematic closure (KIC) is an analytical inverse kinematic algorithm, which achieves loop closure by calculating exact solutions for certain degrees of freedom (like backbone torsion angles) which act as constraints, and allowing other degrees of freedom to be sampled, in effect limiting the number of conformations to be sampled.[172] Alternatively, backbone torsion angle sampling or constraining can be avoided entirely by searching for loops in a large database such as the RCSB Protein Data Bank (PDB) that plug the gap between fragments.[173] However, it is worth noting that this method by itself may not always be sufficient, particularly in the case of *de novo* models and when there is a large distance between the SSEs to be bridged.

4.2 Parameterising PPα Oligomers

4.2.1 Examining the Component Helices of the PPα Miniproteins

Parametric protein design packages such as ISAMBARD are capable of modelling ideal, repetitive structures such as coiled coils, but we have yet to explore parametric descriptions of PPα-like miniproteins. An idealised helical net (a two-dimensional projection of the positions of the C_α atoms of

a given helix) for an α helix and a PPII helix overlay nicely in terms of the knobs-into-holes-like packing of Pro into diamond-shaped holes defined by the *a*, *d* and *g* residues on the α helix (figure 4.2). However, experimentally determined PPα structures (oPPα, 6G6X; PPα, 5LO2; and two PPα structures containing *para*-substituted phenylalanine rings, PPα-ϕOCH$_3$, 5LO3, and PPα-ϕCH$_3$, 5LO4) have not yet been probed. If the average number of residues per turn and the average rise per residue for both SSEs align well with the parametric descriptions of these helix types, then it is feasible that the PPα topology could be described parametrically.

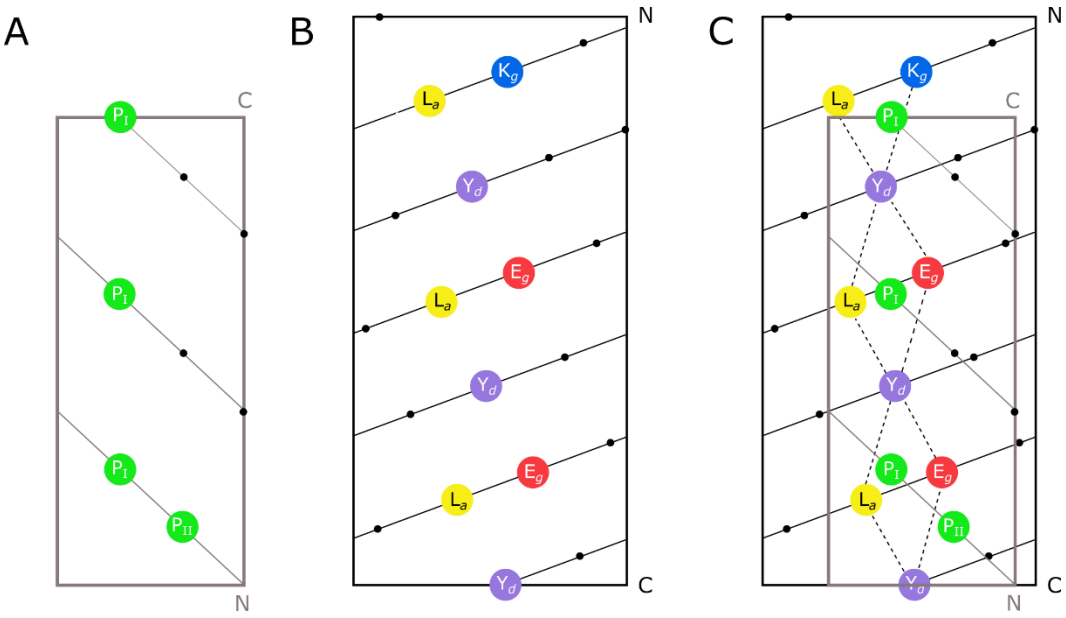


Figure 4.2: *Helical net representations of α and PPII helices. Helical net representations of* ***A****, the PPII helix and,* ***B****, the α helix of oPPα, and* ***C****, an overlay of the two helical nets, with Pro docking into* adg *diamond-shaped holes on the α helix.*

Representative models from the NMR ensembles of oPPα, PPα, PPα-ϕOCH$_3$, and PPα-ϕCH$_3$ (models 12, 12, 6 and 19, respectively) were determined using OLDERADO (https://www.ebi.ac.uk/pdbe/nmr/olderado/),

and each model probed using the protein analysis tools in `ampal`. The PPII and α helices were defined (residues 1 to 8 and residues 14 to 34, respectively), and the average number of residues per turn (RPT) and the average rise per residue (RPR) were calculated for each SSE.

For the PPII helices, the calculated RPT values for each miniprotein largely agreed with the ideal RPT value of 3.0 RPT (table 4.1). The exception, however, was PPα-ϕOCH_3, with an average RPT of 3.5 ± 1.6. Closer examination of the individual values calculated showed that the number of residues per turn were much larger at the termini than in the middle of the helix (6.0 RPT at the *N* terminus and 5.5 RPT at the *C* terminus). The average RPR of the PPII helices of PPα and PPα-ϕOCH_3 were lower than expected (table 4.1), but the RPR values reported for oPPα and PPα-ϕOCH_3 (3.0 and 3.1 Å, respectively) fall much closer to the ideal RPR for a PPII helix of 3.1 Å. With regards these two parameters, both oPPα and PPα-ϕOCH_3 can access PPII conformations that are comparable with an ideal model of a PPII helix.

There were minor deviations between the solution-phase structures and the parametric models in terms of the angle between the C_α atoms of Pro2, Pro5 and Pro8 ($\angle P_{258}$) of the PPII helix, which sit almost directly on top of each other and outline two consecutive ProXxxXxx repeats. The angle between these atoms in model 12 of the oPPα NMR ensemble was 161.7°; 167.8° in model 12 of oPPα; 151.0° in model 6 of PPα-ϕOCH_3, and 175.9° in model 19 of PPα-ϕCH_3. The same angle in an ideal, parametric model of a PPII helix (constructed using the `specifications.HelicalHelix` class in ISAMBARD, and defined only by its RPR and RPT values) was 178.0°. For PPα-ϕCH_3, there was no impact on the distance between the helix termini (table 4.1). For oPPα, the difference between these angles hardly impacted the distance between the PPII helix termini: the distance between the C_α atoms of Pro1 and Pro8 is 21.4 Å in oPPα and 21.8 Å in the parametric PPII helix model. However, there are more significant

differences in the termini distance in PPα and PPα-ϕOCH_3 — a difference of almost 2 Å in each case. These differences are important to consider as they may affect the interhelical termini distances in any *in silico* models generated parametrically, which could make adding loops of the required length (based on the sequences of the miniproteins and the PPα oligomer peptides) more difficult.

The α helices of the representative models of oPPα, PPα, PPα-ϕOCH_3, and PPα-ϕCH_3 were examined in the same manner. The average RPT and RPR values of each α helix was determined to be 3.6 and 1.5 Å (table 4.1), respectively, as expected for an ideal α helix. However, the angle between the C_α atoms of the three Tyr residues ($\angle Y_{20\,27\,34}$) in the sequence (all located at the end of a seven-residue repeat) differed between each miniprotein, as the distance between the α helix *N* and *C* termini (table 4.1). Out of the four miniprotein structures, the α helix of PPα-ϕCH_3 most closely matched that of the generated parametric α helix.

Overall, it is apparent that both the PPII helix and the α helix in oPPα (and other previously characterised PPα miniproteins) have the potential to be well-described parametrically (figure 4.3), and each SSE in the representative models could potentially be used as a base for a fragment-based model template for a PPα oligomer model.

Table 4.1: *RPT values, RPR values, distances and angles extracted from the PPα miniprotein family compared to parametric PPII and α helices.*

Peptide	**RPT**		**RPR (Å)**	
	PPII	α	PPII	α
oPPα [a]	2.9 ± 0.6	3.6 ± 0.2	3.0 ± 0.3	1.5 ± 0.2
PPα [a]	3.2 ± 0.7	3.6 ± 0.1	2.8 ± 0.5	1.5 ± 0.3
PPα-ϕOCH_3 [b]	3.5 ± 1.6	3.6 ± 0.2	2.7 ± 0.8	1.5 ± 0.3
PPα-ϕCH_3 [c]	2.9 ± 0.2	3.6 ± 0.1	3.1 ± 0.2	1.5 ± 0.2
Ideal SSE	3.0	3.6	3.1	1.5
Peptide	**∠P_{258} (°)**	**∠Y_{202734} (°)**	**Distance (Å) [d]**	
			PPII	α
oPPα [a]	161.7	164.4	21.4	29.0
PPα [a]	167.8	171.9	20.1	29.0
PPα-ϕOCH_3 [b]	151.0	164.7	19.4	28.8
PPα-ϕCH_3 [c]	175.9	177.9	21.8	29.7
Ideal SSE	178.0	176.5	21.8	30.7

[a] Model 12, [b] model 6, and [c] model 19 were determined to be most representative of the NMR ensemble for each peptide using Olderado. [d] Distances reported are measured between the C_α atoms of the *N*- and *C*-terminal residues. Where applicable, the values reported represent the mean ± the standard deviation.

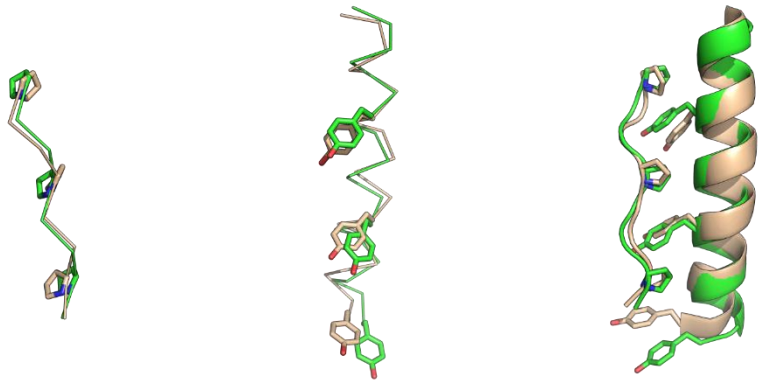


Figure 4.3: *Comparison of PPII (left) and α (right) helices from model 12 of the oPPα NMR ensemble (peach) and parametric models of the same SSEs (green).*

4.2.2 PyMOL Models of PPα Oligomers

A number of parameters are required to fully describe the PPα oligomers, such as the super-helical pitch and radius of the PPII helices in the PPα oligomers. These parameters unfortunately cannot not be extracted from the structures of the PPα miniproteins alone using ISAMBARD, as the protocol for calculating parameters first requires the generation of a reference axis; this reference axis can only be generated between two identical helices of equal length. However, sensible initial estimates of these parameters can be determined from models of the PPα oligomers. Such models were constructed in PyMOL based on the X-ray crystal structures of CC-Tri, CC-Pent, CC-Hex2 and CC-Hept (PDB IDs 4DZL, 4PN8, 4PN9, and 4PNA, respectively) and a representative model of the oPPα NMR ensemble.[60, 73, 147] The coiled-coil helices were truncated to three heptads in length so that the helical registers began at *g*, and residues at all *f* positions were mutated to Tyr to aid sequence alignment. A copy of the oPPα model was aligned to each coiled-coil helix *via* its α helix. The final models consisted of α helices from the coiled coil (each 21 residues in length) and the PPII helices (each eight residues in length) from the miniprotein.

4.2.3 Parameter Extraction

Parameters were extracted from the PyMOL models using tools from the `ampal` module (table 4.2). The radii of the assemblies of PPII helices and α helices were calculated by determining the average distance between points on each helix in the assembly and a central reference axis. The α or tilt angle is the angle of the component helix to the super-helical axis. A list of α angles was generated for each helix in the assembly. The pitch was then calculated using the following equation:

$$\text{pitch} = 2\pi r \,/\, \tan(x) \qquad \textit{Eq 4.1}$$

where x is the previously calculated α angle in radians.

The final assembly radius and pitch values were calculated as the means of these parameters for each CC helix. The interface angle, also known as the Crick angle, was measured between the C_α atom of a specific residue of a helix in the assembly, the centre of the helix, and the centre of the assembly. The final CC interface angle was taken to be the average of the interface angles generated for each helix, which in turn were calculated from the average of the Crick angles measured at positions 1, 8, and 15 of each α helix (the first residue of three consecutive heptad repeats). The same principles were applied for all PPII helices, with the exception that the Crick angles were measured at positions 1, 4 and 7 of each PPII helix, to complement the PPII helix's three-residue repeat pattern.

Table 4.2: *Average values of parameters extracted from PPα-Tri, -Pent, -Hex2 and -Hept.*

Peptide	Radius (Å)		Pitch (Å)		Interface Angle (°)	
	PPII	α	PPII	α	PPII	α
PPα-Tri	15.0	6.8	131.4	168.5	82.3	- 86.8
PPα-Pent	16.8	8.6	136.8	176.8	87.3	- 88.2
PPα-Hex2	17.6	9.4	125.9	158.9	84.7	- 84.1
PPα-Hept	18.0	9.8	196.3	464.4	85.9	- 91.1

4.3 Model Building

4.3.1 Construction of Model Class

The pre-existing `CoiledCoil` class in the `specifications` module of ISAMBARD v2.0.0 was expanded upon to create a new protein structure class for the PPα oligomers (see chapter 2, section 2.4.2). In the `CoiledCoil` class, the coiled coil is modelled as a discrete, symmetric unit. In order to create a PPα oligomer model, the PPII helices were treated as an additional but separate oligomer, constructed around and antiparallel to the coiled coil. In addition to the radius, pitch and interface angle, the exact placement of the PPII helices was described by a z-shift term (relative to the α helix), and a super-helical rotation term, introduced to control the position of the PPII helices around the central axis with respect to the α helices. These parameters could not be extracted from the PyMOL models, and instead were determined *via* iterative adjustments after comparisons to the original input models (table 4.3).

Table 4.3: *Additional parameters necessary to define selected parametric PPα oligomer models.*

Peptide	Z-shift (Å)	Super-helical Rotation (°)
PPα-Tri	6.0	40.0
PPα-Pent	6.0	-50.0
PPα-Hex2	6.0	-50.0
PPα-Hept	6.0	-30.0

Comparisons of the PyMOL models and the generated parametric models showed that, on the whole, the α helices were well described by the extracted parameters. However, there was more variation in the backbone of the PPII helices for all four PPα oligomer models (figure 4.4A, B, C, and D), resulting in sub-optimal placement of the final Pro8 residue of the PPII helix. In each case, Pro8 pointed away from the interhelical interface (figure 4.4E), the opposite of what is seen in the PPα miniproteins.

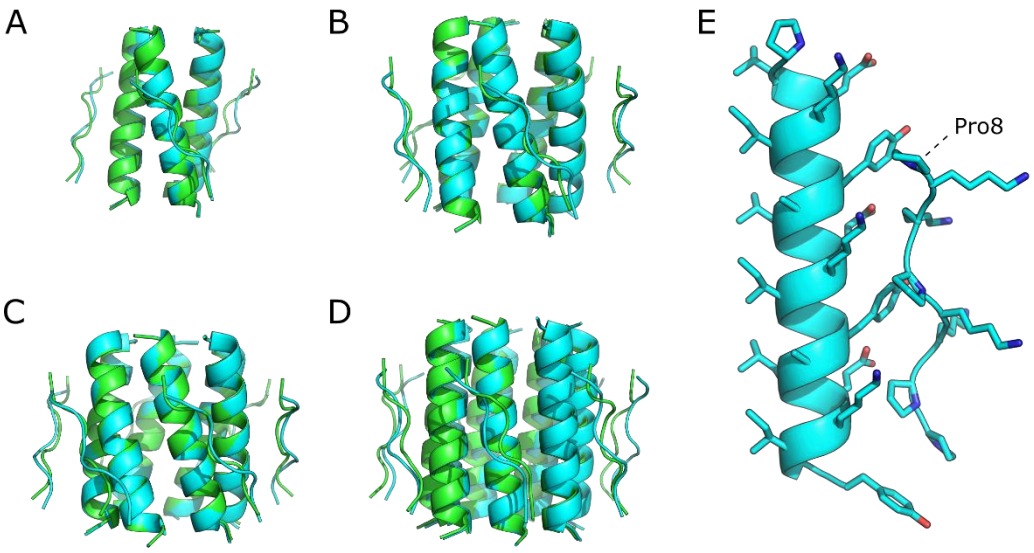


Figure 4.4: *Comparisons of the polypeptide backbones of PyMOL models of the PPα oligomers and models of PPα oligomers generated from extracted parameters. Models for* ***A****, PPα-Tri,* ***B****, PPα-Pent,* ***C****, PPα-Hex2, and* ***D****, PPα-Hept.* ***E****, A PPII helix and an α helix from PPα-Hept, with side chains shown as sticks. Colour key: PyMOL models are shown in green, and models generated from parameters are shown in cyan.*

4.3.2 Model Optimisation

To test the theory that model optimisation may improve the parametric models, the PPα-Tri model was subjected to optimisation using two different algorithms: a genetic algorithm (GA) and a differential evolution (DE) algorithm. The optimisations were conducted on a relatively small scale: over 100 generations, with a population of five in each generation (see chapter 2, section 2.4.3). The GA is good at quickly eliminating unfavourable regions of search space, whilst the DE algorithm is versatile, and helps avoid getting stuck exploring local minima. All parameters other than the oligomer state and the number of amino acids per helix type were subject to modification by the genetic algorithm (table 4.4).

Table 4.4: *A list of parameters floated during the optimisations of PPα-Tri and their allowed ranges.*

Parameter	Parameter Value ± Range
PPII radius	15.0 ± 1.0 Å
α radius	6.8 ± 1.0 Å
PPII pitch	131.4 ± 100 Å
α pitch	168.5 ± 100 Å
PPII interface angle	82.3 ± 27°
α interface angle	-86.8 ± 27°
Z-shift	6.0 ± 1.0 Å
Super-helical rotation	40.0 ± 10°

The optimisations using the GA ran to a minimum, improving on the BUFF (BUDE forcefield) internal energy score of the PPα-Tri model (figure 4.5A), whereas, two-thirds of the time, the optimisations testing the DE algorithm found best-scoring parameters earlier in the optimisation (figure 4.5C, table 4.5). Unfortunately, the issue of the undesirable positioning of the side chains of the PPII helix remained after optimisation (figure 4.5B and D). It is possible that these optimisations do not delve deep enough into the parameter space (*i.e.* the parameter ranges set for this optimisation are not wide enough, or that the optimisations need to occur over an increased number of generations), or that more parameters are needed to describe the desired structure. These problems, however, all require significantly more compute to solve.

Table 4.5: *Best parameters obtained from optimisations of the PPα-Tri model with genetic and differential evolution algorithms.*

Parameter	**GA**			**DE**		
	Opt. 1	**Opt. 2**	**Opt. 3**	**Opt. 1**	**Opt. 2**	**Opt. 3**
PPII radius (Å)	14.4	14.7	15.2	15.2	14.7	14.3
α radius (Å)	6.0	6.0	6.1	5.9	6.4	6.1
PPII pitch (Å)	154.8	120.3	116.1	183.5	129.0	156.2
α pitch (Å)	150.5	221.7	140.3	142.6	109.5	117.1
PPII ϕC_α (°)	81.0	85.1	84.5	87.9	91.2	86.9
α ϕC_α (°)	-74.0	-73.1	-80.4	-75.9	-76.6	-68.5
Z-shift (Å)	6.0	6.0	6.1	5.9	7.0	5.2
Super-helical rotation (°)	38.5	40.9	36.8	37.6	30.6	30.5
BUFF internal energy score	-1503.5	-1514.7	-1508.5	-1483.4	-1512.3	-1494.5
Average BUFF score	**GA:** -1508.9 ± 5.6			**DE:** -1496.7 ± 14.6		

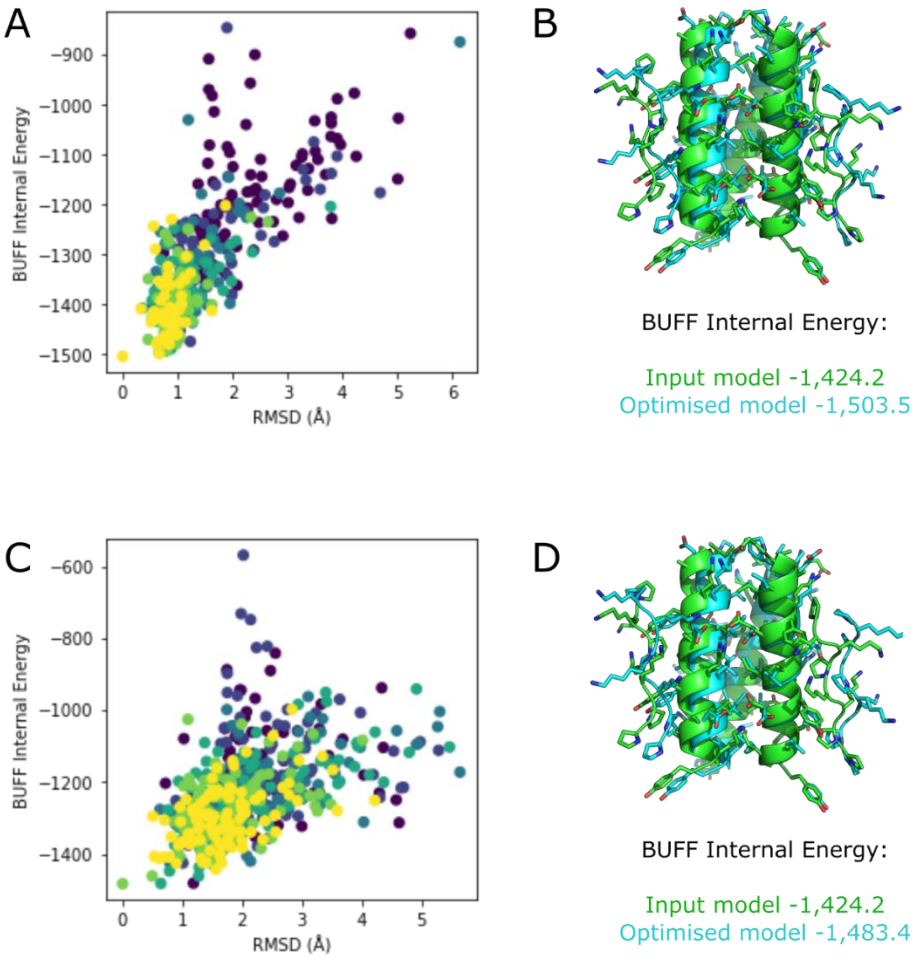


Figure 4.5: *Outputs from optimisation tests of a PPα-Tri model generated from extracted parameters.* ***A****, Graph of BUFF internal energy vs. RMSD for an optimisation with a genetic algorithm. Each circle represents a model generated by the algorithm during the optimisation process. The more negative the BUFF score, the more energetically favourable the model. Purple circles indicate the earliest generation, and yellow circles indicate the last generation. RMSD refers to the difference in RMSD of the best-scoring model and the next best-scoring model.* ***B****, PyMOL cartoon representations of the input PPα-Tri model (green) and the best-scoring model identified by the genetic algorithm (cyan) and their associated BUFF internal energies.* ***C****, Graph of BUFF internal energy vs. RMSD for an optimisation using a differential evolution algorithm. Colour scheme as before.* ***D****, PyMOL cartoon representations of the input PPα-Tri model (green) and the best-scoring model identified by the differential evolution algorithm (cyan) and their associated BUFF internal energies.*

As parametric modelling is primarily concerned with the polypeptide backbone, the modelling of specific rotamers adopted by residue side chains are a secondary consideration. SCWRL4 was used to model low-energy side-chain rotamers onto the newly generated model backbones. In all cases, one or more Tyr residues were modelled with a different rotamer from that observed in the pancreatic polypeptide-like miniproteins. Whilst this is a lesser issue compared to accurately representing the backbone of the desired structure mathematically, the lack of control over side-chain rotamers in ISAMBARD is problematic, considering that the topology of the PPα miniproteins is dictated by the relationship between selected side-chains (and specific side-chain rotamers), and that the design rationale for the PPα oligomers was based on the sequence repeats that facilitate these interhelical interactions.

Without definitive answers to these questions, we decided that it would have been unwise to take these models further. Instead, their precursors — the initial PyMOL PPα oligomer models — were used to estimate distance limits bewteen SSEs needed to be closed by loops to create single-chain proteins.

4.4 Closing the Template

A database loop search tool is implemented in the `modelling` module of a developer version of ISAMBARD. We used the tools within this module to create a loop database and search it for potential loop sequences to transform the peptide-based PPα oligomers into single-chain α-helical bundles and barrels.

4.4.1 Loop database construction

The construction of a suitable loop database requires input from a large number of high-resolution protein structures from the PDB. Searching the

whole PDB for loop segments that fit the desired geometric constraints, however, is not time-economical, and is likely to produce many redundant results, particularly as the PDB contains multiple homologous protein structures. As such, the number of loop structures (and sequences) in the database had to be reduced so that search results can be returned in a reasonable time. The PISCES server allows for culling of the PDB in accordance with parameters selected by the user, such as maximum percentage sequence identity, structure resolution and sequence length.

We used the PISCES server to cull high-resolution (≤ 2 Å) X-ray crystal structures in the PDB (18th July 2018).[151, 152] The sequences were culled by entry at 40% maximum sequence identity, with an R-factor = 0.3 and total sequence length between 100 and 10,000 residues, resulting in 12,668 PDB entries. The loop database was then constructed in ISAMBARD using these files. 251,051 loops were extracted; 4,951 loops failed to parse as no atoms were read before the TER record in the PDB file. The final database consisted of 246,100 loop sequences. It is worth noting that ISAMBARD utilises DSSP (the Dictionary of Secondary Structure of Proteins) to assess SSEs,[154, 174] and this is inherent in the creation of the loop database as DSSP is used to define the local structure of the four residues immediately prior to and after the loop region. DSSP uses hydrogen-bonding patterns to define SSEs, and consequently is unable to identify PPII helices, as these are not defined by hydrogen bonding. Several secondary structure assignment methods can identify PPII helices in protein structures, such as XTLSSTR, PROSS and SEGNO, but do so variably.[175-177] More recently, Mansiaux *et al*. showed that PPII structure can be assigned using the coil class of DSSP, and is in good agreement with other methods.[178] However, to date, only DSSP has been implemented in ISAMBARD.

4.4.2 Searching the loop database

Prior to searching the loop database, the distance between the termini of the two helices — specifically, the distance between the carboxy carbon of the final Tyr residue on the α helix and the nitrogen of the first Pro residue on an adjacent PPII helix — of the PyMOL PPα oligomer models were measured. Knowing these distances allows an estimation of the minimum number of residues required to close the gap, which in turn informs the database search (table 4.6). The minimum number of residues is defined as the result of the pre-determined distance (in angstroms) divided by 3.5 (the span in Ångstroms of a fully extended amino acid).

The database searches were conducted so that the loop placement would result in a right-handed single-chain α-helical barrel (αHB, figure 4.6A). This was envisaged to be the more likely arrangement over a left-handed single-chain αHB (figure 4.6B), as natural proteins are predominantly right-handed in nature. Loop searches were constrained to a maximum of 12 residues, with the minimum number of residues set at 7, 6, 6 and 4 for the PPα-Tri, PPα-Pent, PPα-Hex2 and PPα-Hept models, respectively. Loop searches were not restricted by termini SSE type, but were inherently limited by the SSEs defined by DSSP: α helices (H), extended strands that participate in β ladders (E), 3_{10} helices (G), π helices (I), isolated β bridge residues (B), hydrogen bonded turns (T), and bends (S). Whilst the SSEs that bookend the loops are not of particular interest, loops between *N*-terminal α helices and *C*-terminal β strands would be the most desirable of any results in order to closely mimic PPα-like structures, as PP structure falls within the β region of Ramachandran space.

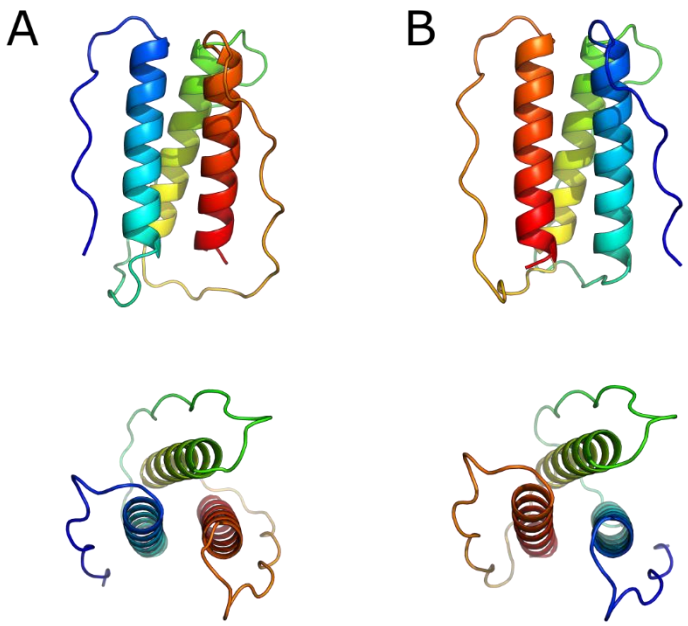


Figure 4.6: *Two possible ways to connect the SSEs in single-chain PPα-based protein models.* ***A*** *shows a right-handed single-chain three-helix bundle, and* ***B*** *a left-handed three-helix bundle. The N terminus is depicted in dark blue, and the Cterminus in red.*

In addition to searching for suitable loop sequences in the purpose-built database, searches were also conducted using the Top8000 database of 7,957 high resolution (< 2 Å) protein structures (≥ 38 residues in length), curated from PDB entries in 2011 for use in rotamer validation in the MolProbity program.[179]

After the database search results (table 4.6) were returned, the 'loop fitter' tool was employed in an attempt to better align each loop with the helix termini it would connect, and the best fit was retained. The lower the fit score, the better the fit. Only the results with a fit score < 2 were considered to limit the potentially high number of loop sequences to examine. In other circumstances, this same cut-off would be used as a measure of the quality

of the resulting models. This excluded one loop sequence for PPα-Pent, one loop sequence for PPα-Hex2, and 37 loop sequences for PPα-Hept.

Table 4.6: *Best-scoring loop sequences identified for PPα oligomer models*

O.S.	Sequence	Type	# Res	Span (Å)	PDB	Fit Score
Pent	GYAHGG†‡	GE	6	20.6	1UWC	1.11
	GYQHPSD†	GE	7	20.2	3UUE	1.61
			(6)	(20.8)		
Hex	GYAHGG†‡	GE	6	20.6	1UWC	1.37
			(6)	(19.4)		
Hept	SKDID†	GE	5	12.8	4RK2	1.30
	GQGAD†	GE	5	13.6	2GHA	1.31
	GIDTF†	GE	5	14.6	3AXB	1.47
	GQPPDN†	HE	6	12.6	1I9G	1.92
	SQIYPGIKV‡	HH	9	12.9	2ZD1	1.96
	WGLKP†	HE	5	14.8	2VK8	1.99
			(4)	(13.6)		

Values in brackets denote the distance between the carboxy carbon of Tyr_{α} and the nitrogen of an adjacent Pro_{PPII} in the corresponding PPα oligomer model. †Loop sequence identified in the created high resolution loop database. ‡Loop sequence identified in the Top8000 loop database.

All of the filtered loops were between two antiparallel (but not necessarily proximal and/or interacting) SSEs, predominantly connecting 3_{10} helices and β strands (figure 4.7, table 4.6), and found at solvent-exposed sites.

No results were returned from either database for PPα-Tri. This is unsurprising, given that the distance between the two termini in the model is quite large (22.0 Å). Loops spanning such distances are uncommon in small, globular protein structures. All resulting loops from the PPα-Pent, PPα-Hex2 and PPα-Hept database searches have been appended to the

corresponding model for visualisation purposes (figure 4.8). Due to the small number of loop sequences being considered, it was reasonable to assess each sequence individually, as opposed to developing and implementing an automated sequence-scoring function.

Two possible loops were returned for the PPα-Pent model: GYAHGG (figure 4.7A), and GYQHPSD (figure 4.7B), six and seven residues in length, respectively (table 4.5). The Gly residues at the termini of the GYAHGG loop sequence act as good breakers for secondary structure; however, the inclusion of a Tyr residue so close to the end of the α helix (already capped by a Tyr residue) could be problematic for maintaining the integrity of the desired PPII:α interface. For the GYQHPSD loop, the same issue with Tyr placement could be argued. In addition, the inclusion of Pro may trigger an early start to the PPII helix, particularly as Asp is favoured in PPII helices, and Ser is even more favoured than its β-branched cousin,[180] Thr, which is observed in several PPII helices in natural and designed pancreatic polypeptide-like miniproteins (including avian pancreatic polypeptide, the bacterial adhesin AgI/II, the glucan-binding protein GBPC, and PPα). The only result obtained for the PPα-Hex2 model is GYAHGG, as previously described for PPα-Pent.

For PPα-Hept, more filtered results were available from both databases: six new sequences in total. SKDID and GQGAD (figures 4.7C and D) are both fairly innocuous in terms of their sequences, each containing majority polar/charged groups and generally small hydrophobic side-chains. GIDTF (figure 4.7E), is a somewhat reasonable sequence, as Phe has low propensity for PPII helix formation, but it does add unwanted hydrophobic character to an ideally solvent-exposed loop. GQPPDN (figure 4.7F) may not be a suitable choice for similar reasons to GYQHPSD *i.e.* by possibly promoting early PPII helix formation. SQIYPGIKV (figure 4.7G), is perhaps the poorest sequence choice of the returned results, with its relatively high number of hydrophobic residues (five out of nine residues), and a less-than-

ideal positioning of some of these hydrophobic side-chains. The Ile and Tyr residues in the loop sequence would be situated three and four residues away from the Tyr residue that caps the α helix, which could potentially allow for helical read-through of the α helix, and possible distortion of the desired tertiary structure as a result. The `WGLKP` sequence (figure 4.7H) may not be a favourable choice, for two reasons: Trp, as it is positioned, may disrupt the envisioned PPII:α interface (Trp mutants of both PPα and oPPα are stable),[146, 147] and expression of protein sequences containing a triple Pro motif are often difficult to express, as these can cause ribosome stalling (although this stalling can be alleviated in bacteria if elongation factor P is present).[181]

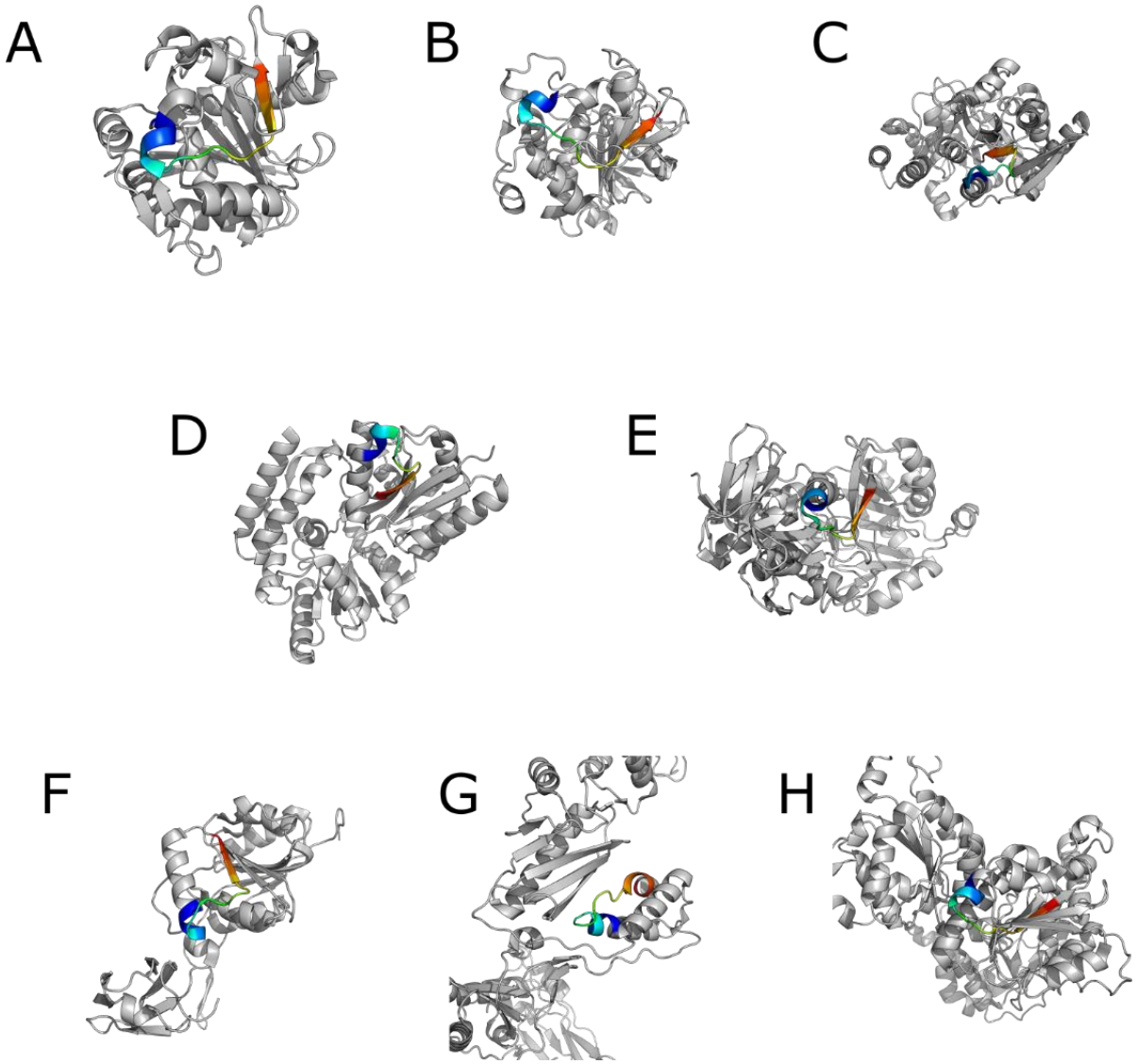


Figure 4.7: *Protein structures from which the loop sequences detailed in table 4.4 originate.* ***A****,* `GYAHGG`*, 1UWC,* ***B****,* `GYQHPSD`*, 3UUE,* ***C****,* `SKDID`*, 4RK2,* ***D****,* `GQGAD`*, 2GHA,* ***E****,* `GIDTF`*, 3AXB,* ***F****,* `GQPPDN`*, 1I9G,* ***G****,* `SQIYPGIKV`*, 2ZD1, and* ***H****,* `WGLKP`*, 2VK8. The preceding four residues of the entering SSE, the loop sequence, and the final four residues of the exiting SSE are highlighted from N- (blue) to C-terminus (red).*

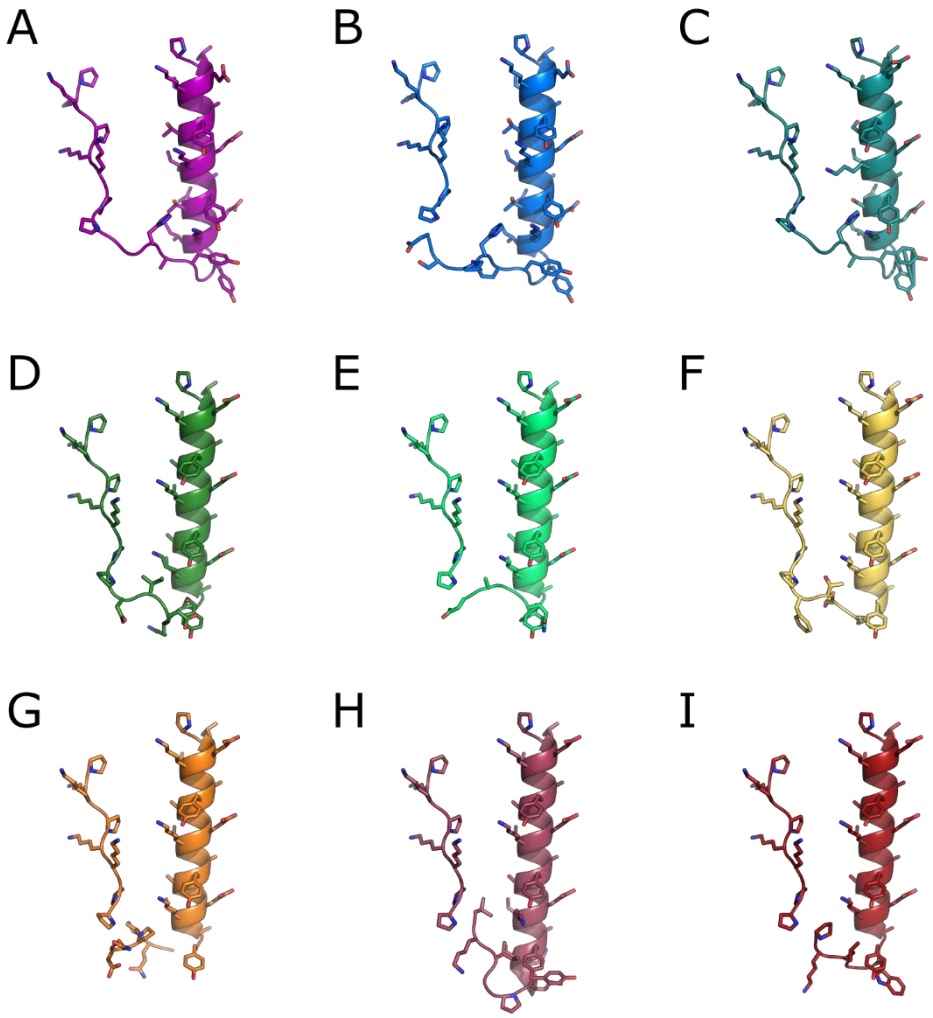


Figure 4.8: *All loop search results fitted to the PyMOL PPα oligomer models. The* ***A****,* `GYAHGG` *and* ***B****,* `GYQHPSD` *loops fitted to PPα-Pent.* ***C****, The* `GYAHGG` *loop fitted to PPα-Hex2. The* ***D****,* `SKDID`*,* ***E****,* `GQGAD`*,* ***F****,* `GIDTF`*,* ***G****,* `GQPPDN`*,* ***H****,* `SQIYPGIKV` *and* ***I****,* `WGLKP` *loops fitted to PPα-Hept. Only the loops and immediately preceding N-terminal α helices and C-terminal PPII helices are shown for clarity. All models are orientated with the N-terminal α helix on the right. The cause for the breaks in the polypeptide backbone as seen for the* `GYQHPSD`*,* `GQGAD`*,* `GQPPDN`*,* `SQIYPGIKV` *and* `WGLKP` *loops is unknown.*

4.5 Conclusions

In this chapter, we set out to determine whether or not PPα oligomers (and subsequently, single-chain, parallel α-helical bundles and barrels) can be described parametrically, and, if so, how to proceed with efforts on loop modelling and the identification of suitable connecting loop sequences.

Coiled coils are already known to be readily parameterizable super-secondary structures, but models of the PPα miniproteins have, to date, never been modelled *in silico* parametrically. Examination of the NMR structure of a number of PPα miniproteins shows that the structures of the component SSEs are very close to that of idealised PPII and α helices, and as such, the optimised PPα miniprotein and derived oligomers should also be amenable to parameterisation.

Our parameterisation attempt is based on that of the available `CoiledCoil` class in ISAMBARD, the Woolfson group's parametric protein design software.[78] Initial parameters were obtained from hybrid PyMOL models, built from a combination of shortened CC helices and PPII helices from a representative model of the oPPα NMR ensemble. However, it was not possible to obtain all the parameters needed to describe the PPα oligomers from the PyMOL models, such as the z-shift and super-helical rotation terms necessary to describe the exact positioning of the PPII helices, and so estimates of these parameters were determined manually.

Unfortunately, the parametric models generated from the extracted parameters did not fully encapsulate the topology of the PPα miniprotein family: the final Pro residue of the PPII helices points away from the interhelical interface, even after test backbone optimisations with two different algorithms. As such, further work is required to improve either the initial parameters, or widen the scope of the optimisation, as well as examining other optimisation algorithms and different forcefields. Efforts must also be made to fix the side-chain rotamers which determine the interhelical interactions in the PPα miniproteins.

Additionally, other routes to modelling the PPα oligomers could be taken. The Rosetta protein design software package also allows for the construction of symmetric models, for example.[84] Alternatively, structure prediction methods, like AlphaFold,[2] hold great potential for delivering models based on input sequences. However, it is worth noting that the methods employed in this chapter assume that the PPα fold is the only fold that can be adopted by the sequences described in chapter 3, and this may not necessarily align with the output of a structure prediction algorithm which lacks knowledge of the previously designed, and to-date unreported peptide sequences.

Despite the simplicity of parametric modelling, a major drawback is the difficulty of modelling the connecting loops (and identifying suitable loop sequences) required to produce a completely *in silico* model of a single-chain α-helical bundle or barrel. To this end, we have attempted to find suitable loop sequences in protein structures reported in the PDB that would span the distance between an α helix and an adjacent PPII helix from each PyMOL oligomer model. However, few results are returned from searching a pre-existing and newly created high-resolution databases, with no results from either database for PPα-Tri (the PPα oligomer possessing the longest distance between SSEs to be connected). Additionally, the sequences returned are mostly unsuitable due to possible helical read-through or the presence of a high percentage of hydrophobic residues for an ideally solvent-exposed loop. Additional loop modelling methods for parametric templates will need to be examined in the future. In the meantime, flexible linkers (not explored in this chapter) may provide a temporary relief for the problems outlined herein.

Chapter 5

Towards Extended PPα Oligomers and Single-Chain Proteins

The work described in this chapter was conducted by the author of this thesis, who expressed, purified and characterised the single-chain proteins, and Dr William Dawson, who expressed, purified and characterised the oligomeric proteins.

5.1 Introduction

In chapter 4, parametric modelling was explored as a means to generate models of PPα-based single-chain α-helical bundles and barrels. However, poor parameterisation of the PPII helix from the NMR structure of optimised PPα meant that the PPα topology could not be faithfully described *in silico*. Instead, PPα oligomer models constructed in PyMOL were used to estimate the distances between the helix termini and, thus, the minimum length of

the loops required to link the structures. A database search for loop sequences purported to close these gaps yielded little by way of suitable linker sequences for heavily solvent-exposed regions for any of the target designs.

As evidenced by the near-UV CD data for the peptide assemblies of chapter 3, the interface between the PPII and α helices in the optimised PPα miniprotein appears to be preserved in the PPα oligomers. As such, any structure programmed into the connecting loops is likely unnecessary to achieve stable single-chain α-helical bundles and barrels. In other words, flexible linkers could be used in place of loop sequences 'cut and pasted' from natural protein structures. Linker sequences in recombinant fusion proteins (typically connecting two protein domains as opposed to two secondary structure elements, SSEs) are often based on small, non-polar residues (*e.g.* Gly) for flexibility and polar residues (*e.g.* Ser and Thr) for hydrophilicity in the solvent-exposed region between domains.[182] Yu and Lutz review distances between native protein termini in the Protein Data Bank (PDB) and find that around 20 % of protein *N* and *C* termini are within 10 to 20 Å.[183] Similarly, the range found in the PyMOL models of the PPα oligomers was approximately 14 to 22 Å. Yu and Lutz note that, for the circular permutation of native proteins with termini > 10 Å apart, care must be taken with linker design as this can have a more significant impact on protein stability. However, the literature they reviewed is skewed towards Gly-rich linker sequences for such distances.

5.2 Recombinant Expression of PPα Oligomers

5.2.1 Design

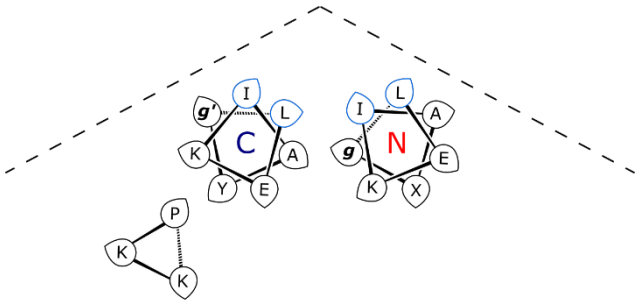
As the 34-residue oligomerising peptides described in chapter 3 are realised easily *via* solid-phase peptide synthesis, we focused here on the recombinant expression of longer α-L-PPα polypeptides, where α is an α

helix, L is a flexible linker sequence, and PPα is a polyproline-II helix connected to an α helix *via* the five-residue GDNAT loop. Should these recombinant oligomers be realised, the α-L-PPα scaffold represents a stepping stone between the peptide-based PPα oligomers and the single-chain αHBs, over which an ever-expanding level of sequence control can be exacted. Reducing the number of polypeptide chains *via* oligomerisation of larger subunits and ultimately concatenation of all polypeptide chains results in fewer and fewer mutations in the protein core.

Both α-helical bundles and barrels could be achieved *via* oligomerisation of α-L-PPα systems (figure 5.1) by the specific sequence patterning of the α helix, although the specific focus in this section is on achieving α-helical barrels as opposed to bundles containing four central α helices. As mentioned above, the three-heptad PPα-Pent2, PPα-Hex2 and PPα-Hept designs described in chapter 3 did not assemble quite as desired. By sedimentation equilibrium analytical ultracentrifugation, PPα-Pent2 and PPα-Hept formed hexamers (5.7 and 6.1 × monomer mass, respectively) instead of a pentamer and a heptamer, and PPα-Hex2 behaved somewhat promiscuously in solution (5.3 or 6.5 × monomer mass by sedimentation velocity and sedimentation equilibrium AUC, respectively). Ultimately, these three designs adopt hexameric assemblies in solution. As such, the fact that the α-L-PPα system is biased towards the formation of αHBs containing an even number of α helices is not of much consequence.

As Thr, Ser and Ala at the *g* position in the three-heptad *g*LKEIAY system all seemingly produce hexamers, it could be useful to explore whether a mixture of these different residues at the *g* and *g'* positions of the two adjacent α helices in the α-L-PPα scaffold can be tolerated — a notion that cannot be tested in the current family of homomeric αHBs. So as not to be burdened by a large library of constructs, mutations were confined to all *g* positions of one of the two helices in the α-L-PPα scaffold, resulting in nine different arrangements of Thr, Ser and Ala at *g* and *g'* positions of the α

helices in the α-L-PPα scaffold (table 5.1). These proteins were named systematically in the form *α2-gg'*, where *g* and *g'* are the identity of the residues at these positions.



g = A, S, T
g' = A, S, T

Figure 5.1: *A generic helical wheel for the α2-gg' protein series. The nine constructs encompass all possible combinations of Ala, Ser and Thr at the g and g' positions. N and C denote the N- and C-terminal helices of the monomer.*

Table 5.1: *Sequences, masses and pI values of the extended PPα oligomers.*

Protein	Sequence *gabcdefgabcdefgabcdef*	Mass (Da)	pI
α2-AA	MGSSHHHHHHSSGLVPR GSG ALEKIAQALEKIAWALEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY	8,987.7	8.0
α2-AS	MGSSHHHHHHSSGLVPR GSG ALEKIAQALEKIAWALEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY	9,019.7	8.0
α2-AT	MGSSHHHHHHSSGLVPR GSG ALEKIAQALEKIAWALEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY	9,047.7	8.0
α2-SA	MGSSHHHHHHSSGLVPR GSG SLEKIAQSLEKIAWSLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY	9,035.7	8.0
α2-SS	MGSSHHHHHHSSGLVPR GSG SLEKIAQSLEKIAWSLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY	9,067.7	8.0
α2-ST	MGSSHHHHHHSSGLVPR GSG SLEKIAQSLEKIAWSLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY	9,095.7	8.0
α2-TA	MGSSHHHHHHSSGLVPR GSG TLEKIAQTLEKIAWTLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY	9,077.7	8.0
α2-TS	MGSSHHHHHHSSGLVPR GSG TLEKIAQTLEKIAWTLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY	9,109.7	8.0
α2-TT	MGSSHHHHHHSSGLVPR GSG TLEKIAQTLEKIAWTLEKIAQ SEGSEGSEGPPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY	9,137.7	8.0

5.2.2 Expression Trials

DNA sequences for the designed proteins were optimised using a codon optimiser for *E. coli*, and purchased cloned directly into the pET-28a vector. This expression vector has a T7 promotor and a kanamycin resistance gene. It was chosen as this promotor allows facile overexpression of exogenous, soluble proteins, so long as the protein of interest is not toxic to the *E. coli* host. The synthetic proteins also had a hexa-histidine tag for facile protein purification. The histidine tag and the proteins were linked with a thrombin cleavage site (Leu-Val-Pro-Arg-/-Gly-Ser, where '/' indicates the cleavage site) to remove the histidine tag if needed.

The plasmids were transformed into Rosetta 2 (DE3) *E. coli* cells, a derivative of the BL21 *E. coli* strain considered to be the workhorse of protein expression. These Rosetta 2 cells are typically used in the expression of eukaryotic proteins in *E. coli*, as the strain has an additional plasmid with a resistance marker orthogonal to the kanamycin resistance gene in the pET-28a vector, and tRNAs for codons rarely used in bacteria. Rosetta 2 cells are also deficient in the protease OmpT, which reduces the possibility of proteolysis of the proteins of interest during cell lysis.

Of the nine constructs listed in table 5.5, the expression of five proteins has been attempted so far: α2-AA, α2-SA, α2-SS, α2-TA and α2-TT. Bacteria containing the pET-28a plasmids encoding the genes of interest were grown in LB media at 25 °C, and protein expression induced by the addition of IPTG when absorbance at $OD_{600\ nm}$ = 0.6–0.8. Soluble material from the cell lysate was harvested 18 hours after induction of protein expression and purified by nickel affinity chromatography. Subsequent size exclusion chromatography of the nickel column elution fraction provided a valuable first insight into whether or not any desired protein may be present (figure 5.2).

Disappointingly, as with sc-α5, sc-α6 and sc-α7, the SEC traces of α2-TT, α2-SS or α2-AA showed the elution of large aggregates early in the elution and no discernible monomer protein peaks. In addition, no protein bands with the expected masses were observed by SDS-PAGE. Whilst the SEC trace for α2-SA differed from these other traces (figure 5.2), SDS-PAGE of the SEC fractions has yet to be carried out to determine if any protein of interest is present. However, discrete peaks were observed in the SEC trace corresponding to α2-TA. This result was unexpected and interesting — it was not anticipated that simply mutating three residues (either in α2-TT or α2-AA to α2-TA) would yield such a distinct change in protein expression behaviour.

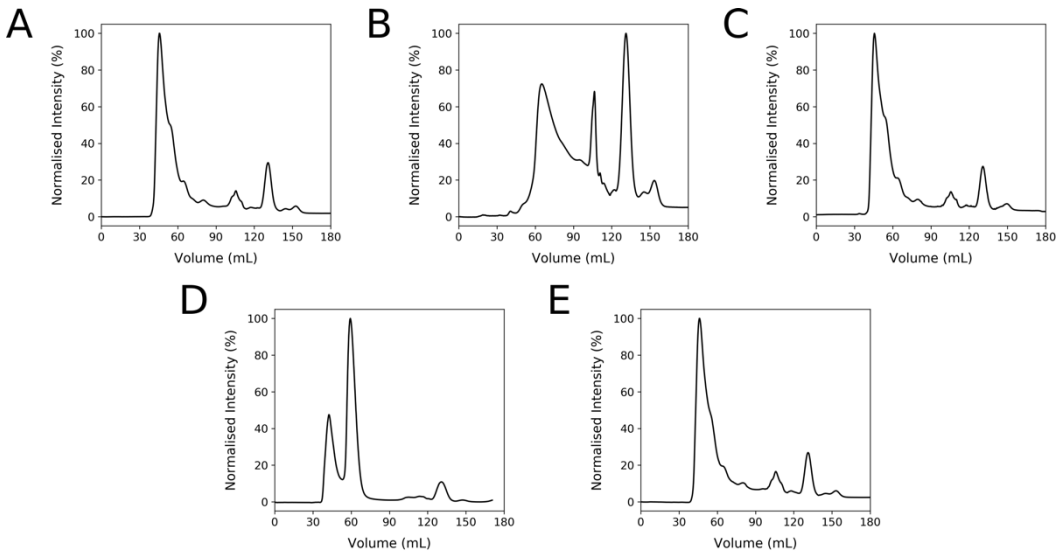


Figure 5.2: *SEC traces following nickel affinity chromatography for selected extended PPα oligomers.* ***A****, α2-AA,* ***B****, α2-SA,* ***C****, α2-SS,* ***D****, α2-TA, and* ***E****, α2-TT.*

Despite no apparent expression in Rosetta 2 (DE3) *E. coli*, soluble protein for α2-TT, α2-SS and α2-AA may result from expression in a different *E. coli* strain *e.g.* Lemo21 (DE3), as observed in the case of sc-α7. Further work must be carried out to determine if any of the remaining four proteins can be expressed under these same initial conditions.

5.2.3 Initial Characterisation of a Recombinantly-Expressed PPα-Based 'Oligomer'

The presence of monomeric protein in the SEC trace of α2-TA was confirmed by SDS-PAGE, although the protein band observed ran higher on the gel than expected (figure 5.3), and the mass of the monomer has yet to be confirmed by mass spectrometry. Initial characterisation of α2-TA was carried out by circular dichroism spectroscopy. The CD spectrum at 20 °C revealed that the protein was 36% helical at 10 μM and was

hyperthermostable with a T_M > 95 °C (figure 5.4, table 5.2). Near-UV CD studies of α2-TA have yet to be undertaken.

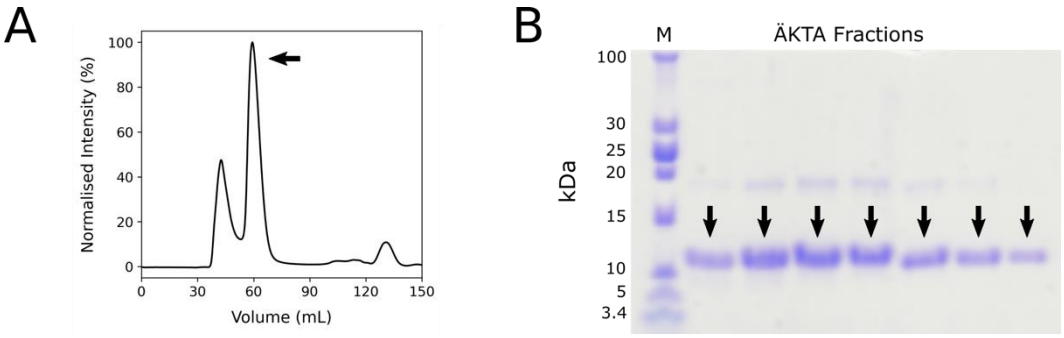


Figure 5.3: *SEC trace and SDS-PAGE gel for α2-TA.* ***A****, SEC trace, and* ***B****, 13% SDS-PAGE gel following SEC for α2-TA. M = marker (PageRuler low range unstained protein ladder); FT = flow through; W = wash fraction; E = elution fraction, and P = insoluble pellet. Arrows indicate bands and peaks of interest. The expected mass of α2-TA is 9.1 kDa (including N-terminal Met and His-tag).*

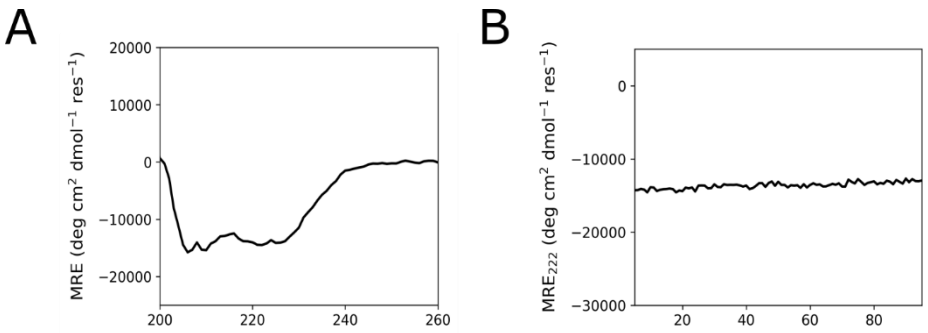


Figure 5.4: *Circular dichroism spectra for α2-TA.* ***A,*** *Far-UV CD spectra recorded at 20 ºC, and* ***B,*** *thermal denaturation profiles monitored at 222 nm for α2-TA. Conditions: 10 μM peptide, 100 mM $NaPO_4$, 100 mM NaCl, pH 5.8.*

Table 5.2: *Mean residue ellipticity at 222 nm, fraction helix (%) and T_M values for α2-TA.*

Protein	MRE_{222}	Fraction Helix (%)	T_M (°C)
α2-TA	-14,504	36	> 95

Unfortunately, no information on the oligomeric state of α2-TA could be determined as the protein aggregated in the ultracentrifuge at 10 μM at a low rotor speed (3 krpm). Further biophysical characterisation must be undertaken to probe the nature of this protein in solution, and additional work is necessary to understand why aggregation occurs, and whether it is a result of inherent 'oligomer frustration' or some other feature of the protein.

5.3 Single-Chain Parallel α-Helical Bundles

5.3.1 Design and Expression of Soluble *De Novo* Helical Bundles

From the work described in chapter 3, there are three viable templates for a single-chain, parallel, three-helix bundle: PPα-Tri-2, PPα-Tri3, and PPα-Tri-4 (table 5.3). All adopt the desired oligomer state in solution, and these are thermostable with T_M values between 64 °C and 78 °C. The sequences of PPα-Tri-2 and PPα-Tri-3 are very similar, as are their T_Ms (table 5.3). It was expected that linking the individual polypeptide chains of the assembly into a single-chain protein would lead to an increase in T_M. With this in mind, we selected a template with a somewhat reduced thermostability, so that this change might be measured explicitly. PPα-Tri-2 was ultimately chosen as the template sequence.

Table 5.3: *Sequences for PPα-Tri-2, PPα-Tri-3 and PPα-Tri-4.*

Peptide	Sequence and Register	T_M (°C)†	O.S.‡
	gabcdef gabcdef gabcdef		
PPα-Tri-2	PPKKPKKP GDNAT PIELIKY EIELIKY EIELIKY	64 ± 1.2	3.1
PPα-Tri-3	PPKKPKKP GDNAT PIELKIY EIELKIY EIELKIY	65 ± 1.2	3.1
PPα-Tri-4	PPKKPKKP GDNAT PLELIKY ELELIKY ELELIKY	78 ± 1.0	3.1

All peptides were N-terminally acetylated and amide-capped at the *C*-terminus. †Calculated from first derivative of signal monitored at 222 nm. ‡O.S. = oligomeric state (× monomer mass) determined from SE AUC experiments.

In chapter 4, approximate estimates of the loop lengths necessary to connect two assembly subunits for α-helical bundles and barrels were derived from PyMOL models of the PPα oligomers. For PPα-Tri, this distance was estimated to be 22 Å, which would require a minimum of seven residues to cross, assuming each residue in the loop adopts an extended conformation. To avoid strain, a loop length of ten residues was chosen. In the loop database search, no suitable loops of this length were found that could link two subunits of the PPα-Tri model. In the absence of any structural precedent, we looked to flexible linkers. Specifically, a linker based on a Gly-Ser-Glu repeat. Whilst the *C*-terminal Tyr of the preceding helical region can act as a good capping residue for an α helix, Gly is a better α helix cap.[160] Both Ser and Glu have low propensities for the α-helical *C*-cap position, if any propensity at all.[160] Therefore, the final loop sequence was `GSEGSEGSEG` to give the sc-α3-1 sequence (table 5.2).

Whilst the design of sc-α3-1 (with a PPα-L-PPα-L-PPα SSE pattern, where PPα is the sequence of PPα-Tri-2, and L a loop sequence) was directly drawn from PPα-Tri-2, the *N*-terminal PPII helix is not necessary for trimer formation or for subunit concatenation — it would merely be an heirloom from the parent peptide sequence (see figure 3.3, page 54). As such, a second single-chain three-helix bundle sequence without the *N*-terminal PPII helix was designed (sc-α3-2, table 5.4). If this protein is confirmed to be stable, this PPII helix need not be carried over into future designs.

Table 5.4: *Sequences, masses and pI values of sc-a3-1 and sc-a3-2.*

Protein	Sequence *gabcdefgabcdefgabcdef*	Mass (Da)	pI
sc-α3-1	MGSSHHHHHHSSGLVPRGSHM PPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEG PPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEG PPKKPKKPGDNATPIELIKYEIELIKYEIELIKYG	16,097.5	6.4
sc-α3-2	MGSSHHHHHHSSGLVPRGSHM G EIAAIKQEIAAIKKEIAAIKQGSEGSEGSEG PPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEG PPKKPKKPGDNATPIELIKYEIELIKYEIELIKYG	14,442.6	6.4

Bacteria containing the pET-28a plasmids encoding the genes of interest were grown in LB media at 37 °C, and induction of protein expression and purification were carried out as before. Bands corresponding to the masses of sc-α3-1 and sc-α3-2 were observed on denaturing SDS-PAGE gels stained with Coomassie Blue (figure 5.5A and 5.5D, respectively). Size exclusion chromatography (SEC, figure 5.5B and 5.5E) and further SDS-PAGE confirmed fractions containing monomeric protein for characterisation.

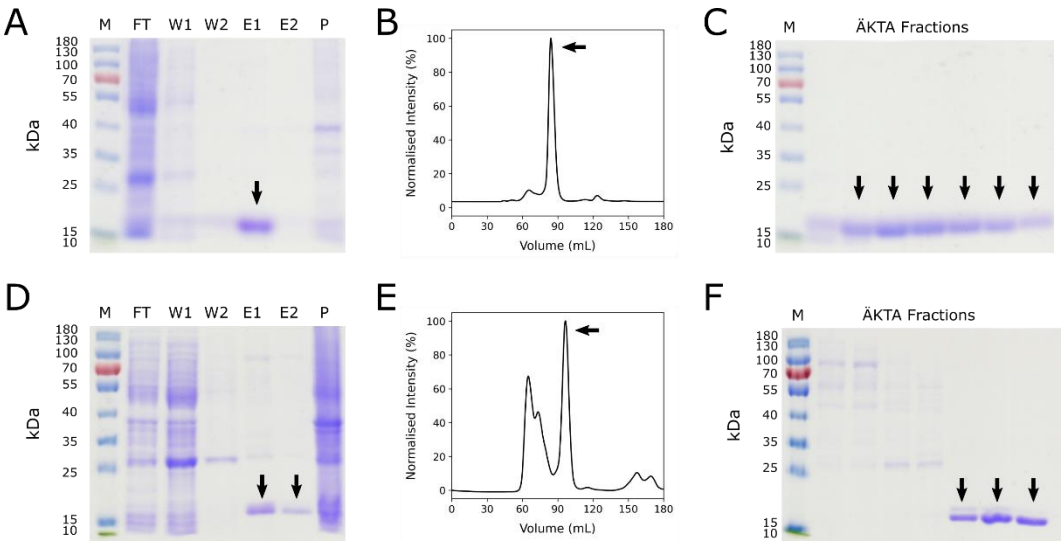


Figure 5.5: *SDS-PAGE gels and SEC traces for sc-α3-1 and sc-α3-2.* ***A****, 12% SDS-PAGE gel following nickel affinity chromatography,* ***B****, SEC trace, and* ***C****, 12% SDS-PAGE gel following SEC for sc-α3-1.* ***D****, 12% SDS-PAGE gel following nickel affinity chromatography,* ***E****, SEC trace, and* ***F****, 12% SDS-PAGE gel following SEC for sc-α3-2. M = marker (PageRuler pre-stained protein ladder); FT = flow through; W = wash fraction; E = elution fraction, and P = insoluble pellet. Arrows indicate bands and peaks of interest. The expected masses of sc-α3-1 and sc-α3-2 are 16.1 and 14.4 kDa, respectively (including N-terminal Met and His-tag).*

5.3.2 Monomeric, Stable, Single-Chain Parallel Bundles

As SDS-PAGE gels indicate only approximate protein masses, the masses of sc-α3-1 and sc-α3-2 were confirmed by mass spectrometry, using nanospray ionisation (figure 5.6). Despite bands for sc-α3-2 in the SDS-PAGE gels running slightly higher than the expected mass (just over 15 kDa instead of the expected 14.4 kDa), the observed mass matched the calculated mass well.

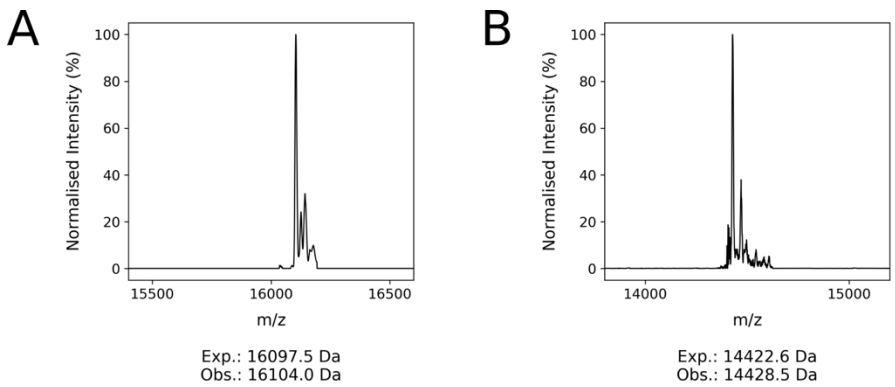


Figure 5.6: *ESI mass spectra for sc-α3-1 (left) and sc-α3-2 (right). Exp. = expected i.e. calculated mass; Obs. = observed mass.*

Both proteins were examined by CD spectroscopy to probe their secondary and tertiary structures. In comparison to PPα-Tri-2, their peptide precursor, sc-α3-1 was slightly less helical, and sc-α3-2 was slightly more helical (figure 5.7A, table 5.5), despite the presence of extra polypeptide fragments (*i.e.* His-tags and loops) which were not expected to be α-helical. This second three-helix bundle design, sc-α3-2, did not have an additional PPII helix, most likely resulting in its slightly improved helicity over the other design. A major difference between the peptide oligomer and the single-chain proteins is the thermostability — both single-chain proteins displayed sharper thermal unfolding transitions and were more stable than PPα-Tri-2 (figure 5.7B, table 5.5). However, and as with the peptides, the proteins did not refold post-denaturation, but precipitated (post-denaturation spectra are given in chapter 7.4). Both sc-α3-1 and sc-α3-2 show bands in the near-UV region, indicative of the PPα fold; indeed, these were slightly stronger signals than for the peptide precursor (figure 5.7C). The T_M values measured at 276 nm were similar to those measured at 222 nm for sc-α3-1 and sc-α3-2 (figure 5.7D, table 5.5), and higher than that for PPα-Tri-2; though there was a 13 °C difference between the far- and near-UV measurements for the peptide.

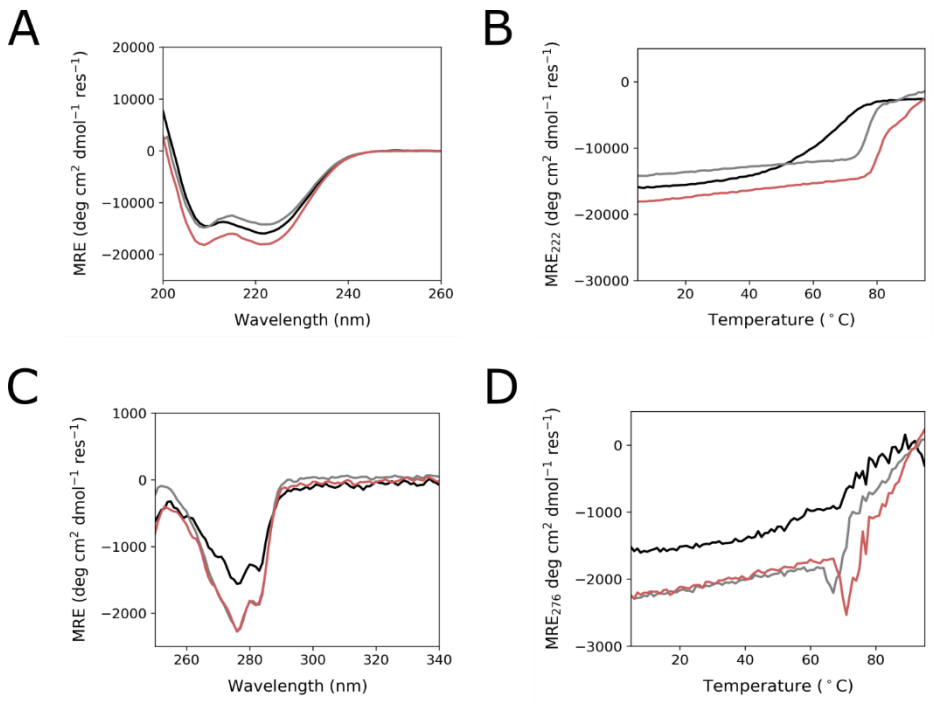


Figure 5.7: *Circular dichroism spectra for PPα-Tri-2, sc-α3-1, and sc-α3-2.* ***A,*** *Far-UV CD spectra recorded at 5 ºC, and* ***B,*** *thermal denaturation profiles monitored at 222 nm for PPα-Tri-2 (black), sc-α3-1 (grey), and sc-α3-2 (pink).* ***C,*** *Near-UV CD spectra recorded at 5 ºC and* ***D,*** *thermal denaturation profiles monitored at 276 nm for PPα-Tri-2 (black), sc-α3-1 (grey) and sc-α3-2 (pink). Conditions: 5 μM protein (sc-α3-1 and sc-α3-2) or 100 μM peptide (PPα-Tri-2), PBS, pH 7.4.*

Table 5.5: *Comparison of the mean residue ellipticity at 222 nm, fraction helix (%) and T_M values for PPα-Tri-2, sc-α3-1 and sc-α3-2.*

Peptide / Protein	MRE_{222}	Fraction Helix (%)	T_M (°C)
PPα-Tri-2	-16,202 ± 223	43 ± 0.6	64 ± 1.2 (77 ± 1.0)
sc-α3-1	-14,214	37	78 (78)
sc-α3-2	-18,034	47	84 (82)

Conditions are as listed in the caption for figure 5.7. Values in brackets denote those derived from near-UV CD experiments.

Both proteins were subjected to sedimentation velocity (SV) and sedimentation equilibrium (SE) analytical ultracentrifugation (AUC) experiments to determine their homogeneity and molecular weights (and, therefore, oligomeric states) in solution. Continuous c(s) fits to SV data collected for sc-α3-1 at 50 krpm showed a single species with a molecular weight of 15.9 kDa (1.0 × monomer mass; figure 5.8A). Fits to SE AUC data returned a molecular weight of 18.7 kDa for this protein when fitted to an ideal, single-species model (1.2 × monomer mass; figure 5.8B). Similar AUC experiments for sc-α3-2 indicated that it was also a homogeneous monomer as designed (figures 5.8C and D).

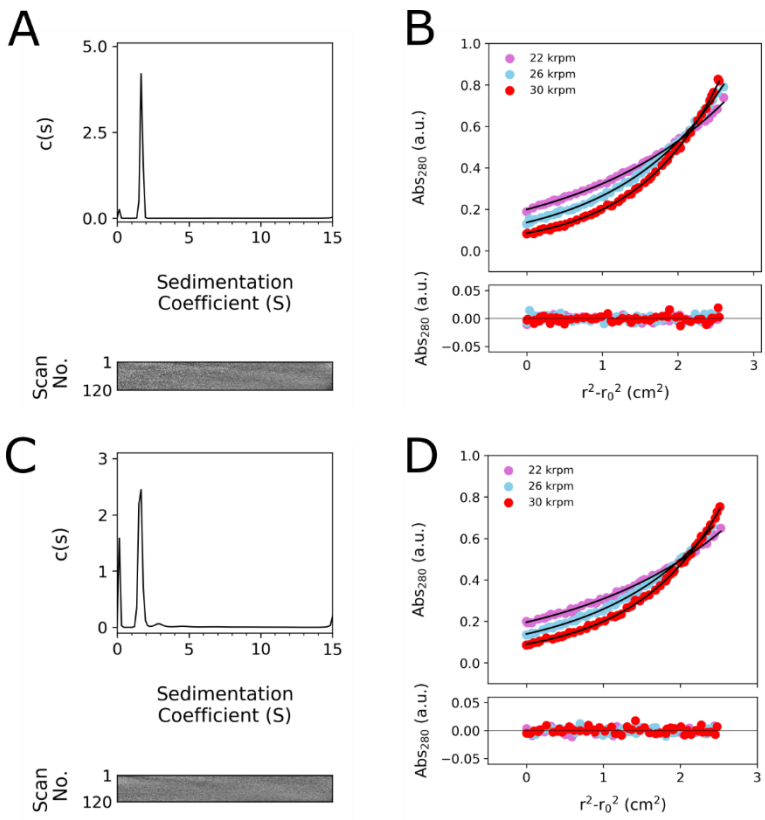


Figure 5.8: *Sedimentation velocity and equilibrium AUC data and fits for sc-α3-1 ($\bar{v}$ = 0.7466 cm^3 g^{-1}) and sc-α3-2 ($\bar{v}$ = 0.7446 cm^3 g^{-1}).* ***A****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for sc-α3-1 at 50 krpm returning Mw = 15921 Da (1.0 × monomer mass), 95% confidence limits. Conditions: 70 μM protein, PBS, pH 7.4.* ***B****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 22, 26 and 30 krpm for sc-α3-1, returning Mw = 18694 Da (1.2 × monomer mass, 95% confidence limits: 18385 – 19013 Da), and residuals for the fits (bottom, dots). Conditions: 35 μM protein, PBS, pH 7.4.* ***C****, Continuous c(s) fit to SV data (top, black line) and residuals bitmap (bottom) for sc-α3-2 at 50 krpm, returning Mw = 16559 Da (1.1 × monomer mass), 95% confidence limits. Conditions: 110 μM protein, PBS, pH 7.4.* ***D****, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 22, 26 and 30 krpm for sc-α3-2, returning Mw = 17880 Da (1.2 × monomer mass, 95% confidence limits: 17504 – 18270 Da), and residuals for the fits (bottom, dots). Conditions: 55 μM protein, PBS, pH 7.4.*

As with the peptide assemblies described in chapter 3, the sequences of sc-α3-1 and sc-α3-2 were deemed too repetitive for structure determination by NMR. Unfortunately, and despite several attempts, crystals needed for

X-ray crystallography could not be obtained from sitting-drop crystallisation using commercial screens and high protein concentrations, between 1 and 7 mg/mL.

5.4 Towards Single-Chain Parallel α-Helical Barrels

Encouraged by the success of the single-chain three-helix bundles, efforts turned towards single-chain, parallel α-helical barrels (αHBs). Three single-chain αHB sequences were targeted: sc-α5, sc-α6, and sc-α7, with five, six and seven α helices, respectively, table 5.6. These were based on the corresponding peptide designs PPα-Pent2, PPα-Hex2, and PPα-Hept and incorporating flexible linkers as for the three-helix bundles. Both PPα-Pent2 and PPα-Hex2 required a minimum of six residues to span the distance between the helix termini whereas PPα-Hept requires a minimum of 4 residues. As with the single-chain α-helical bundles, additional loop residues were added to ensure that the loops did not become a source of strain to the desired topology. Ultimately, the PPα subunits of sc-α5 and sc-α6 were joined with an eight-residue `GSE`-based loop, and the subunits of sc-α7 were joined by a seven-residue `GSE` loop.

Analytical ultracentrifugation experiments conducted on the PPα barrels showed that they did not quite assemble in solution as desired — the protein assemblies were determined to be hexamers, instead of the expected range between pentamer and heptamer. However, with recombinant protein expression, in principle at least, the size of the protein is determined by the length of the gene encoding it; for example, a five-helix bundle could be achieved by encoding five linked PPα subunits.

Table 5.6: *Sequences, masses and pI values of sc-a5, sc-α6 and sc-a7.*

Protein	Sequence *gabcdefgabcdefgabcdef*	Mass (Da)	pI
sc-α5	MGSSHHHHHHSSGLVPRGSHM [PPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAYGSEGSEGS]×4 PPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAYG	24,179.7	9.2
sc-α6	MGSSHHHHHHSSGLVPRGSHM [PPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAYGSEGSEGS]×5 PPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAYG	28,511.9	9.2
sc-α7	MG [PPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGSEGSEG]×6 PPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGGKLAAA LEHHHHHH	31,708.9	9.1

Genes encoding sc-α5, sc-α6 and sc-α7 were transformed into Rosetta 2 (DE3) *E. coli*., and the same protein expression protocol was followed as for the single-chain bundles. However, no bands corresponding to the masses of any of the three proteins were observed in the soluble and insoluble cell material by SDS-PAGE gels after nickel column chromatography.

There are numerous strategies that can be used to improve exogenous protein expression in bacteria. If the protein of interest is sequestered in insoluble inclusion bodies, it can be extracted from purified inclusion bodies and refolded with a decreasing concentration of denaturant, typically *via* dialysis. Alternatively, fusing solubility tags such as SUMO, small ubiquitin-like modifier, or MBP, maltose binding protein to the protein of interest (POI) can enable its solubilisation. Adjusting the temperature at which protein expression is induced can affect the concentration of the protein shepherded to inclusion bodies. However, these methods are only useful if protein expression occurs. Unfortunately, in the case of sc-α5, sc-α6 and sc-α7, no bands corresponding to these proteins were identified even in the

insoluble cell material — a difficult place from which to optimise protein expression.

Whilst the T7 expression system is a popular choice for the overexpression of exogenous proteins, a disadvantage of this system is that low levels of recombinant protein can be expressed prior to induction with a chemical agent. In other words, the T7 expression system is not tightly regulated, and is often referred to as a 'leaky' expression system. If the POI is toxic to the expression host, cell growth can be limited. Consistent with this, Rosetta 2 (DE3) *E. coli* containing plasmids encoding sc-α5, sc-α6 and sc-α7 grow more slowly than 'empty' cells without the expression vectors.

Despite seemingly no/little production of sc-α5, sc-α6 and sc-α7 in Rosetta 2 (DE3) *E. coli*, recombinant protein may be obtained from other expression strains. Numerous variants of the popular BL21 (DE3) *E. coli* are used for protein expression. Some of these derivative strains include an additional plasmid encoding T7 lysozyme, which inhibits basal expression of T7 RNA polymerase, thus inhibiting expression of any toxic proteins encoded in a T7 expression vector during the cell growth phase. Examples of such strains include BL21 (DE3) pLysS and Lemo21 (DE3). Lemo21 (DE3) *E. coli* have been used for the expression of proteins that are toxic, membrane proteins which are difficult to express, and other proteins which are known to be difficult to solubilise.[184, 185] The Lemo21 (DE3) strain has an advantage over the BL21 (DE3) pLysS strain: in Lemo21 (DE3) *E. coli*, the *pLEMO* plasmid encodes T7 lysozyme under the control of the titratable rhaBAD promoter, allowing for fine control over its expression. For this reason, the Lemo21 (DE3) strain was selected for further expression trials.

Focus was placed on achieving expression of sc-α7 over other single-chain αHBs, as its design forms the basis of this thesis. After transforming the pET plasmid containing the gene for sc-α7 into Lemo21 (DE3) cells and following the same induction protocol as before, a protein band corresponding to the mass of sc-α7 (approximately 31 kDa) was observed

by SDS-PAGE after nickel affinity purification of the soluble cell lysate (figure 5.9), although the yield of protein was very low. As a result, full characterisation of this protein was not possible. This initial work with the Lemo21 (DE3) *E. coli* strain gives hope that the remaining single-chain αHB proteins can be expressed, and further work will need to be carried out to improve the yield of pure sc-α7 protein.

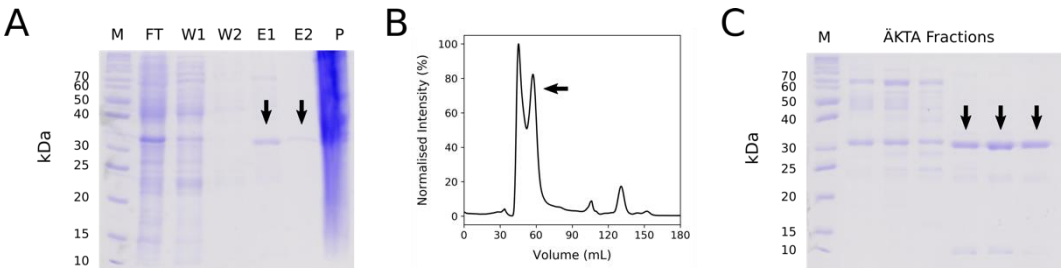


Figure 5.9: *SDS-PAGE gels and SEC traces for sc-α7.* ***A****, 12% SDS-PAGE gel following nickel affinity chromatography,* ***B****, SEC trace, and* ***C****, 12% SDS-PAGE gel following SEC for sc-α7. M = marker (PeqGOLD unstained protein marker); FT = flow through; W = wash fraction; E = elution fraction, and P = insoluble pellet. Arrows indicate bands and peaks of interest. The expected mass of sc-α7 is 31.7 kDa (including N-terminal Met and His-tag).*

Table 5.7: *A summary of conditions for successful expression of selected proteins of interest.*

Protein	Expression Conditions	Obs. Mass (Da)	Fraction Helix (%) [a]	T_M (°C) [b] Far- / Near-UV	O.S. SV / SE
sc-α3-1	Rosetta 2 (DE3) *E. coli*, 0.5 mM IPTG, 18 h, 37 °C	16104.0	37	78 / 78 [c]	1.0 / 1.2 [e]
sc-α3-2	Rosetta 2 (DE3) *E. coli*, 0.5 mM IPTG, 18 h, 37 °C	14428.5	47	84 / 82 [c]	1.1 / 1.2 [f]
sc-α7	Lemo21 (DE3) *E. coli*, 0.5 mM IPTG, 18 h, 37 °C	TBC	TBC	TBC / TBC	TBC / TBC
α2-TA	Rosetta 2 (DE3) *E. coli*, 0.5 mM IPTG, 18 h, 25 °C	TBC	36	≥ 95 / TBC [d]	Agg. [d]

[a] Calculated using *Equation 2.3*. [b] Calculated from the 1st derivative of thermal denaturation profiles. [c] Conditions: 5 μM protein (far-UV CD) or 50 μM protein (near-UV CD), PBS, pH 7.4. [d] Conditions: 10 μM protein, 100 μM $NaPO_4$, 100 μM NaCl, pH 5.8. [e] Conditions: 70 μM protein (SV AUC) or 35 μM protein (SE AUC), PBS, pH 7.4. [f] Conditions: 110 μM protein (SV AUC) or 55 μM protein (SE AUC), PBS, pH 7.4. Obs. = observed. TBC = to be confirmed. Agg. = aggregated at 3 krpm.

5.5 Conclusions

The aims of the work described in this chapter were as follows: to recombinantly express oligomers based on those PPα barrels synthesised by solid-phase peptide synthesis, and to achieve single-chain parallel α-helical bundles and barrels. To these ends, headway has been made, despite the path being peppered with unsuccessful protein expression attempts.

Simple constructs based on an α-L-PPα scaffold were designed to test expression of PPα-based oligomers in *E. coli*., as well as exploring whether or not αHBs comprised of these oligomerising α-L-PPα units would tolerate different residues at the *g* and *g'* positions in adjacent α helices. Of the five oligomers examined in this chapter, sequences defined by Thr, Ser and Ala at the *g* and *g'* positions of the α helix were unable to be expressed in Rosetta 2 (DE3) *E. coli*. Surprisingly, an oligomer sequence with Thr at *g* and Ala at *g'* expressed well, and preliminary biophysical characterisation showed it to be α-helical in nature. No information could be garnered on the oligomeric state of the protein in solution, however, as it aggregated at low rotor speeds in the analytical ultracentrifuge.

The successful design of two parallel single-chain helical bundles, sc-α3-1 and sc-α3-2, comprising subunits of PPII and α helices linked by flexible loop sequences is also reported. These monomeric single-chain bundles are folded in solution, exhibit signals of the tertiary structure inherent to their parent peptide assemblies, and are more thermostable than their precursors, nearing hyperthermostability in the case of sc-α3-2. Despite attempts, high-resolution structural data have yet to be obtained for these proteins. From here, focus segued into single-chain α-helical barrels comprising five, six and seven central α helices. However, attempts to express these proteins in Rosetta 2 (DE3) *E. coli* cells proved fruitless: no protein bands corresponding to the expected masses were observed in the soluble or insoluble fractions of the cell lysates. However, changing expression host to Lemo21 (DE3) *E. coli* yielded protein material of the

expected mass (as determined by SDS-PAGE) for the seven-helix (sc-α7) design. Further confirmation of the identity of sc-α7 is necessary, alongside full biophysical characterisation of the protein. It is hoped that this alternative *E. coli* strain can aid in the expression of the remaining single-chain αHB proteins and PPα oligomers.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Through the work described in this thesis, a set of peptide oligomers based on the redesigned PPα miniprotein and *de novo* coiled coils have been designed and characterised. These PPα oligomer designs form the basis of a strategy towards the de-symmetrisation of coiled coils and act as a template for the first single-chain, parallel α-helical bundles. Designs of single-chain, parallel α-helical barrels lie on the cusp of realisation.

Firstly, a number of oligomers have been designed by embedding residues dictating *de novo* coiled-coil interfaces in the solvent-exposed face of the optimised PPα miniprotein. These designs are inspired by the antiparallel helix-loop-helix topology of the miniprotein; such a motif is integral for the development of de-symmetrised coiled coils which maintain the parallel orientation of their component α helices. Initial efforts focused on the design of a three-heptad trimeric PPα assembly bearing the hallmarks of a *de novo* designed coiled-coil trimer. The residues dictating both coiled-coil and PPα interfaces merge seamlessly, and three trimeric, thermostable assemblies displaying near-UV CD signals characteristic of the PPα fold result. The success of this proof-of-concept work led to more complicated designs based on type II α-helical barrels comprising five, six, and seven helices. The designed sequences are folded and hyperthermostable, but not all adopt the oligomer state of their parent coiled coil — all are hexameric in nature. No peptide crystals for 3D structure determination were obtained for any of the designs. However, saturation binding experiments revealed that the assemblies possess accessible channels. Overall, just 34 residues are

sufficient to encode a *de novo* peptide assembly akin to an inverted TIM barrel structure.

In the absence of structural data, efforts were made to generate *in silico* renditions of the PPα oligomers on which loops for later single-chain designs could be modelled. Only a small number of parameters are required to faithfully describe oligomers of α and polyproline-II helices (*e.g.* coiled coils and collagens), and so parameterisation of the PPα oligomers was deemed to be feasible.[78] Estimates for initial parameters were extracted from PyMOL models and funnelled into ISAMBARD for reconstruction and optimisation. However, some parameters (such as the z-shift and the super-helical rotation of the helix types relative to each other) could not be extracted from these models, and the parameterisation was not faithful to key tenets of the PPα fold even after test optimisations with two different algorithms. Instead, the PyMOL models were used as a starting point from which to search the Protein Data Bank for suitable loop sequences for single-chain designs. No results were found for the trimeric PPα, and the filtered hits returned for the PPα barrels often contained a high percentage of hydrophobic residues or were at risk of causing deviations from the intended structure *via* helical read-through.

With no informative leads from *in silico* models, flexible linkers based on `SEG` repeats were utilised in recombinant protein designs. 'Extended' PPα oligomers, based on a α-L-PPα scaffold where L is a flexible linker, represent a stepping-stone between peptide assemblies and single-chain proteins in terms of the level of sequence control offered, and allow nuances in α-helical barrel interfaces to be probed for the first time. Nine constructs with different combinations of Thr, Ser, and Ala at every *g* position of the two α helices were designed to probe possible frustration in the oligomers. However, out of the five designs subjected to expression trials so far, only one oligomer (*g* = Thr, *g'* = Ala) expresses well. The protein has been partially characterised: whilst folded and thermostable, the protein aggregates at low speeds in the ultracentrifuge, precluding determination of

its assembly size. Flexible linkers were also utilised in single-chain α-helical bundle and barrel designs. The first examples of single-chain, parallel α-helical bundles are presented. These proteins are more thermostable and display sharper thermal unfolding transitions than their parent peptide assembly, and are monomers, as designed. The expression of single-chain α-helical barrels, however, proved more difficult, with no apparent expression of the designs in a common bacterial expression strain. Pleasingly, preliminary expression trials with Lemo21 (DE3) *E. coli* cells holds promise for the overexpression of soluble single-chain α-helical barrels: a band corresponding to the mass of a protein barrel comprising seven α helices is observed by gel electrophoresis.

6.2 Future Work

As demonstrated herein with the PPα miniprotein, oligomerising polypeptide motifs possessing tertiary structure offers the opportunity to investigate previously unexplored regions of protein space. These oligomers may prove themselves to be useful and informative in their own right; alternatively, as was intended with this work, the oligomer can act as a building block for ever more complex structures. Ideally, future efforts on the PPα oligomers and the single-chain parallel α-helical bundles and barrels will lead to the determination of their 3D structures, which may give insights into the optimisation of these designs and aid parameterisation or other modelling efforts.

Our initial driving force for realising single-chain parallel barrels remains: to investigate and improve catalysis within the lumen of α-helical barrels. These single-chain systems will undoubtedly prove to be useful tools to test our knowledge on the design of active sites for catalysis within coiled coils. It is hoped that efforts in this area will yield a *de novo* designed hydrolase with a catalytic efficiency comparable to its natural counterparts, aided by directed evolution if necessary. Indeed, a multitude of functionalities could

be introduced into such a scaffold. Alternatively, single-chain α-helical barrels could find use in sensors. α-Helical barrels are known to be accessible to biologically relevant small molecules and are capable of discriminating between differently sized ligands.[112] As the cores of the homomeric barrels are predominantly hydrophobic, single-point mutations of barrel receptors could help improve binding of more polar analytes, for example.

Additionally, alternatives to single-chain parallel α-helical barrels, such as single-chain antiparallel barrels or heteromeric barrels (DNA-templated or otherwise) could be explored in the future. Such investigations would further inform coiled coil design and allow protein designers access to an even more vast expanse of protein space.

Chapter 7

Appendix

7.1 Designed Peptide and Protein Sequences

Table 7.1: *Designed peptide and protein sequences.*

Peptide/Protein	Sequence
PPα-Tri-1	Ac–PPKKPKKP GDNAT EIAAIKYEIAAIKYEIAAIKY–NH_2
PPα-Tri-2	Ac–PPKKPKKP GDNAT PIELIKYEIELIKYEIELIKY–NH_2
PPα-Tri-3	Ac–PPKKPKKP GDNAT PIELKIYEIELKIYEIELKIY–NH_2
PPα-Tri-4	Ac–PPKKPKKP GDNAT PLELIKYELELIKYELELIKY–NH_2
PPα-Hept	Ac–PPKKPKKP GDNAT PLEKIAYALEKIAYALEKIAY–NH_2
PPα-Hept-CHE	Ac–PPKKPKKP GDNAT PLEKIAYALEKCAYAHEKEAYALK–NH_2
PPα-Pent2	Ac–PPKKPKKP GDNAT PLEKIAYTLEKIAYTLEKIAY–NH_2
PPα-Hex2	Ac–PPKKPKKP GDNAT PLEKIAYSLEKIAYSLEKIAY–NH_2
PPα-Non	Ac–PPKKPKKP GDNAT PLEKIAYGLEKIAYGLEKIAY–NH_2
α2-AA	MGSSHHHHHHSSGLVPRGSGALEKIAQALEKIAWALEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY
α2-AS	MGSSHHHHHHSSGLVPRGSGALEKIAQALEKIAWALEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY
α2-AT	MGSSHHHHHHSSGLVPRGSGALEKIAQALEKIAWALEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY
α2-SA	MGSSHHHHHHSSGLVPRGSGSLEKIAQSLEKIAWSLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY
α2-SS	MGSSHHHHHHSSGLVPRGSGSLEKIAQSLEKIAWSLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY

α2-ST	MGSSHHHHHHSSGLVPRGSGSLEKIAQSLEKIAWSLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY
α2-TA	MGSSHHHHHHSSGLVPRGSGTLEKIAQTLEKIAWTLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY
α2-TS	MGSSHHHHHHSSGLVPRGSGTLEKIAQTLEKIAWTLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAY
α2-TT	MGSSHHHHHHSSGLVPRGSGTLEKIAQTLEKIAWTLEKIAQSEGSEGSEG PPKKPKKPGDNATPLEKIAYTLEKIAYTLEKIAY
sc-α3-1	MGSSHHHHHHSSGLVPRGSHMMPPKKPKKPGDNATPIELIKYEIELIKYE IELIKYGSEGSEGSEGPPKKPKKPGDNATPIELIKYEIELIKYEIELIKY GSEGSEGSEGPPKKPKKPGDNATPIELIKYEIELIKYEIELIKYG
sc-α3-2	MGSSHHHHHHSSGLVPRGSHMGEIAAIKQEIAAIKKEIAAIKQGSEGSEG SEGPPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEGPPK KPKKPGDNATPIELIKYEIELIKYEIELIKYG
sc-α5	MGSSHHHHHHSSGLVPRGSHMPPKKPKKPGDNATPLQKIEYILQKIEYIL QKIEYGSEGSEGSPPKKPKKPGDNATPLQKIEYILQKIEYILQKIEYGSE GSEGSPPKKPKKPGDNATPLQKIEYILQKIEYILQKIEYGSEGSEGSPPK KPKKPGDNATPLQKIEYILQKIEYILQKIEYGSEGSEGSPPKKPKKPGDN ATPLQKIEYILQKIEYILQKIEYG
sc-α6	MGSSHHHHHHSSGLVPRGSHMPPKKPKKPGDNATPLEKIAYSLEKIAYSL EKIAYGSEGSEGSPPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAYGSE GSEGSPPKKPKKPGDNATPLEKIAYSLEKIAYSLEKIAYGSEGSEGSPPK KPKKPGDNATPLEKIAYSLEKIAYSLEKIAYGSEGSEGSPPKKPKKPGDN ATPLEKIAYSLEKIAYSLEKIAYGSEGSEGSPPKKPKKPGDNATPLEKIA YSLEKIAYSLEKIAYG
sc-α7	MGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGSEGSEGPPKKPKK PGDNATPLEKIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDNATPLE KIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDNATPLEKIAYALEKI AYALEKIAYGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY GSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGSEGSEGPP KKPKKPGDNATPLEKIAYALEKIAYALEKIAYGGKLAAALEHHHHHH

7.2 DNA Sequences and Vector Maps

7.2.1 DNA Sequences

α2-AA

ATGGGCTCAAGCCATCATCATCATCACCATTCCTCGGGATTAGTACCTCGTGGCTCGGGAG
CACTGGAAAAGATTGCCCAAGCTCTGGAAAAGATTGCTTGGGCGTTAGAGAAGATCGCGCA
AAGCGAGGGATCTGAAGGAAGCGAGGGACCGCCGAAGAAGCCAAAGAAACCCGGCGATAAT
GCAACCCCGTTAGAGAAGATCGCGTACGCGCTCGAAAAGATCGCCTACGCTTTGGAAAAGA
TTGCATAT

α2-AS

ATGGGCTCTTCGCATCACCACCATCACCACTCATCTGGTCTCGTTCCTCGCGGCAGCGGAG
CACTGGAGAAGATCGCACAAGCCTTGGAGAAGATTGCGTGGGCACTCGAAAAGATCGCCCA
ATCGGAAGGCAGTGAAGGGTCGGAAGGTCCCCCGAAGAAGCCCAAGAAGCCAGGAGATAAT
GCCACGCCGTTAGAGAAGATCGCTTACAGCTTGGAGAAGATCGCCTATTCGCTCGAAAAGA
TCGCGTAC

α2-AT

ATGGGCAGCTCTCATCACCACCATCATCACAGCTCAGGTTTGGTCCCGCGTGGTTCAGGGG
CGCTTGAAAAGATTGCGCAAGCGCTGGAGAAAATCGCATGGGCCTTAGAGAAGATCGCCCA
AAGCGAAGGAAGCGAAGGATCAGAGGGGCCACCCAAGAAGCCAAAGAAACCAGGTGATAAT
GCAACCCCTCTGGAAAAGATTGCTTATACGCTTGAGAAGATCGCATACACGCTTGAAAAGA
TTGCATAC

α2-SA

ATGGGCAGCTCTCATCACCACCATCATCACAGCTCAGGTTTGGTCCCGCGTGGTTCAGGGG
CGCTTGAAAAGATTGCGCAAGCGCTGGAGAAAATCGCATGGGCCTTAGAGAAGATCGCCCA
AAGCGAAGGAAGCGAAGGATCAGAGGGGCCACCCAAGAAGCCAAAGAAACCAGGTGATAAT
GCAACCCCTCTGGAAAAGATTGCTTATACGCTTGAGAAGATCGCATACACGCTTGAAAAGA
TTGCATAC

α2-SS

ATGGGCTCGTCCCACCACCATCACCATCACTCTTCGGGCCTTGTGCCCCGTGGCTCCGGCA
GTCTTGAAAAGATTGCTCAATCCCTGGAGAAGATTGCCTGGTCCTTAGAGAAAATCGCTCA
GTCGGAAGGGAGCGAGGGCAGTGAGGGGCCGCCCAAGAAACCGAAGAAGCCCGGCGATAAC
GCGACCCCGCTTGAGAAGATCGCTTATTCTCTGGAGAAAATCGCGTATTCCCTGGAGAAGA
TCGCGTAC

α2-ST

ATGGGCAGCTCTCATCACCACCATCATCACAGCTCAGGTTTGGTCCCGCGTGGTTCAGGGG
CGCTTGAAAAGATTGCGCAAGCGCTGGAGAAAATCGCATGGGCCTTAGAGAAGATCGCCCA
AAGCGAAGGAAGCGAAGGATCAGAGGGGCCACCCAAGAAGCCAAAGAAACCAGGTGATAAT
GCAACCCCTCTGGAAAAGATTGCTTATACGCTTGAGAAGATCGCATACACGCTTGAAAAGA
TTGCATAC

α2-TA

ATGGGCAGCTCGCATCACCACCACCATCATTCATCTGGTTTAGTTCCTCGTGGCAGTGGTA
CGCTGGAGAAAATTGCTCAAACGCTGGAGAAAATCGCGTGGACGCTCGAAAAGATCGCGCA
AAGTGAGGGTTCTGAGGGAAGCGAGGGACCGCCCAAGAAGCCTAAGAAGCCTGGCGACAAC
GCGACCCCACTGGAAAAGATTGCGTACGCGTTGGAGAAGATCGCCTACGCGTTAGAGAAGA
TCGCCTAC

α2-TS

ATGGGCAGCTCTCATCACCACCATCATCACAGCTCAGGTTTGGTCCCGCGTGGTTCAGGGG
CGCTTGAAAAGATTGCGCAAGCGCTGGAGAAAATCGCATGGGCCTTAGAGAAGATCGCCCA
AAGCGAAGGAAGCGAAGGATCAGAGGGGCCACCCAAGAAGCCAAAGAAACCAGGTGATAAT
GCAACCCCTCTGGAAAAGATTGCTTATACGCTTGAGAAGATCGCATACACGCTTGAAAAGA
TTGCATAC

α2-TT

ATGGGCAGCTCTCACCACCACCACCATCACTCGTCCGGTTTAGTACCACGCGGGAGTGGCA
CCCTTGAGAAGATCGCGCAGACGTTAGAAAAGATCGCCTGGACCCTTGAGAAGATCGCACA
ATCAGAAGGGAGCGAGGGCTCAGAAGGCCCGCCGAAGAAACCAAAGAAGCCGGGCGACAAC
GCTACCCCGTTGGAAAAGATTGCGTACACATTAGAGAAGATCGCCTACACGTTGGAGAAGA
TCGCCTAC

sc-α3-1

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATA
TGATGCCTCCGAAAAAACCTAAAAAACCGGGTGATAATGCAACCCCGATTGAGCTGATCAA
ATATGAAATCGAACTGATCAAGTACGAGATAGAGCTGATTAAGTATGGTAGCGAAGGTTCA
GAAGGCAGCGAAGGTCCGCCTAAAAAGCCGAAAAAGCCTGGCGATAACGCCACACCTATTG
AACTGATTAAATACGAAATTGAGTTAATAAAGTATGAGATCGAGTTAATCAAATACGGCAG
TGAAGGTAGCGAGGGTAGTGAAGGTCCACCGAAGAAACCAAAGAAACCTGGCGACAATGCG
ACGCCGATTGAACTTATTAAGTATGAGATTGAGCTTATAAAGTACGAAATAGAGCTTATCA
AATATGGC

sc-α3-2

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATA
TGATGGGTGAAATTGCAGCCATCAAACAAGAAATTGCCGCAATCAAGAAAGAGATCGCAGC
AATTAAACAGGGTAGCGAAGGTTCAGAAGGCAGCGAAGGTCCTCCGAAAAAACCTAAAAAA
CCGGGTGATAATGCAACCCCGATTGAGCTGATCAAATATGAAATCGAACTGATCAAGTACG
AGATAGAGCTGATTAAGTATGGTAGTGAAGGTAGCGAGGGCTCTGAAGGTCCGCCTAAAAA
GCCGAAGAAACCTGGCGATAACGCCACACCTATTGAACTGATTAAATACGAAATTGAGTTA
ATAAAGTATGAGATCGAGTTAATCAAATACGGC

sc-α5

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATA
TGCCTCCGAAAAAACCTAAAAAACCGGGTGATAATGCAACACCGCTGCAGAAAATTGAATA
CATCCTGCAGAAGATCGAGTATATTCTGCAAAAAATCGAATATGGTAGCGAAGGCAGCGAA
GGTAGCCCACCGAAAAAGCCGAAGAAACCTGGCGATAACGCGACCCCGTTACAAAAAATTG
AGTATATCTTACAGAAAATAGAGTACATACTTCAAAAGATTGAATACGGCAGTGAAGGTTC
AGAAGGTAGTCCGCCTAAGAAACCAAAAAAGCCTGGCGACAATGCTACCCCTCTTCAGAAG
ATAGAATATATCCTTCAGAAAATTGAGTACATTTTACAAAAAATCGAATACGGCTCCGAAG
GTTCTGAAGGCAGCCCTCCAAAAAAACCGAAAAAGCCAGGGGACAACGCAACGCCATTACA
AAAGATAGAATACATTTTACAGAAGATTGAATACATTCTTCAAAAAATAGAGTATGGCTCA
GAAGGTAGCGAAGGTTCACCGCCAAAAAAACCAAAGAAACCCGGTGACAACGCCACACCTT
TGCAAAAGATTGAGTACATATTGCAAAAGATAGAGTATATCCTGCAGAAGATAGAATACGG
C

sc-α6

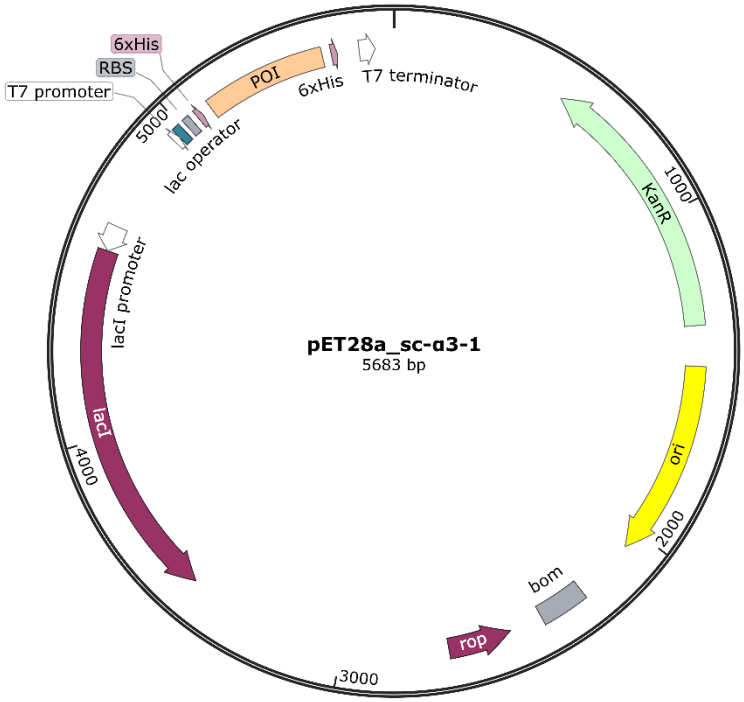
ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATA
TGCCTCCGAAAAAACCTAAAAAACCGGGTGATAATGCAACACCGCTGGAAAAAATTGCATA
CAGCTTAGAGAAAATCGCCTACAGCCTGGAAAAGATCGCCTATGGTAGCGAAGGTTCAGAA
GGTAGCCCACCGAAAAAGCCGAAGAAACCTGGCGATAACGCGACCCCGTTAGAAAAAATAG
CCTATTCGCTTGAAAAGATTGCGTATAGCCTTGAGAAGATCGCATACGGCAGTGAAGGCAG
CGAAGGTAGTCCGCCTAAGAAACCAAAAAAGCCTGGCGACAATGCTACCCCTCTTGAAAAA
ATCGCTTACTCCCTTGAGAAAATAGCGTACTCATTGGAGAAGATTGCTTATGGCTCTGAAG
GTAGTGAAGGCTCACCGCCAAAAAAACCCAAAAAGCCAGGGGATAACGCCACGCCTTTAGA
GAAGATAGCATATAGCTTAGAAAAGATAGCTTACTCACTGGAAAAAATCGCGTACGGTTCA
GAGGGTAGTGAGGGCAGCCCTCCAAAGAAACCGAAGAAACCAGGGGACAATGCGACGCCTC
TGGAAAAGATAGCGTACAGTTTGGAGAAGATAGCCTATTCATTAGAGAAAATTGCGTACGG
TAGCGAGGGTAGCGAAGGCTCTCCTCCTAAAAAGCCAAAAAAACCAGGTGACAACGCAACT
CCATTAGAAAAAATCGCATATAGTTTGGAAAAAATAGCGTATAGCCTGGAAAAAATTGCCT
ACGGC

sc-α7

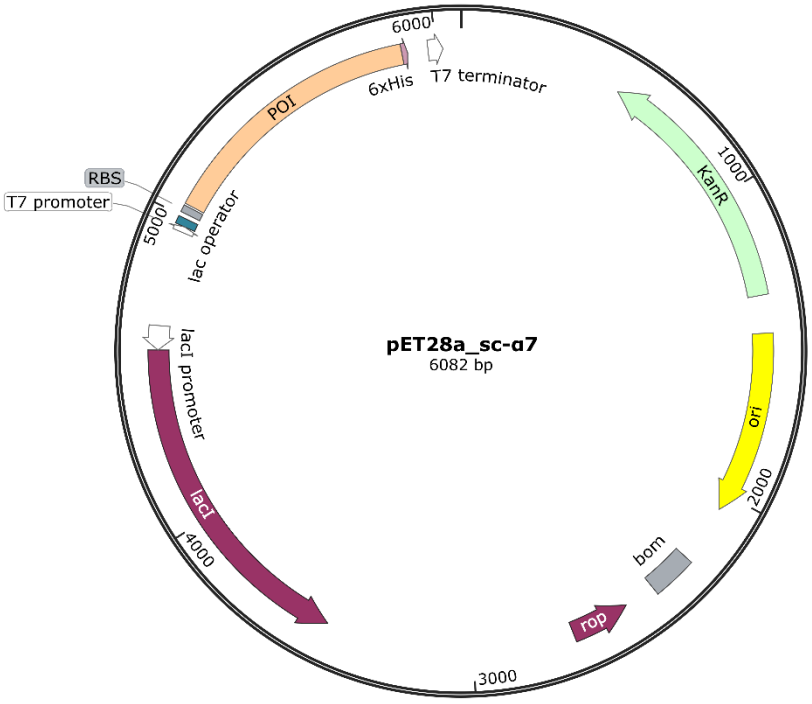
ATGGGGCCTCCGAAAAAACCTAAAAAACCGGGTGATAATGCAACACCGCTGGAAAAAATTG
CATATGCCTTAGAGAAAATCGCCTACGCACTGGAAAAGATCGCCTATGGTAGCGAAGGTAG
TGAAGGTCCGCCTAAAAAGCCGAAAAAGCCTGGCGATAATGCGACCCCTTTAGAAAAAATA
GCGTACGCCCTTGAGAAGATTGCGTATGCATTAGAAAAGATTGCTTATGGCTCAGAAGGCA
GCGAAGGTCCACCGAAGAAACCAAAGAAACCTGGCGACAACGCAACCCCACTTGAGAAAAT
AGCTTATGCGTTGGAGAAGATAGCATATGCCCTGGAAAAAATCGCGTATGGGTCTGAAGGT
TCAGAAGGCCCACCAAAAAAGCCAAAAAAACCAGGGGACAATGCCACGCCGTTAGAAAAGA
TCGCATACGCGTTAGAAAAAATTGCGTATGCGCTGGAAAAAATAGCCTACGGTTCAGAAGG
TTCTGAGGGTCCGCCAAAAAAACCGAAGAAACCCGGTGATAACGCGACTCCGTTGGAGAAA
ATTGCTTACGCTCTTGAAAAAATCGCTTACGCATTAGAGAAGATTGCATACGGTAGCGAAG
GCTCAGAAGGTCCTCCAAAGAAACCTAAAAAGCCAGGCGATAACGCCACACCACTGGAAAA
GATTGCGTACGCCTTGGAAAAGATAGCCTATGCACTTGAGAAAATTGCCTATGGCAGTGAA
GGATCTGAAGGACCTCCTAAAAAGCCCAAGAAACCGGGTGACAACGCTACACCTCTTGAGA
AGATCGCCTATGCGCTTGAAAAGATAGCGTATGCGTTAGAGAAAATAGCTTACGGCCTCGA
GCACCACCACCACCACCAC

7.2.2 Vector Maps

N-Terminally His-Tagged Proteins



C-Terminally His-Tagged Proteins



7.3 Peptide and Protein Characterisation Summary

Table 7.2: *A summary of peptide and protein characterisation.*

Peptide	Calc. Mass (Da)	Obs. Mass (Da)	Fraction Helix (%) [a]	T_M (°C) [b]	O.S. [c]
PPα-Tri-1	3783.123	3783.565	46 ± 3.6	42 ± 0.0 -	ND
PPα-Tri-2	4051.291	4051.362	43 ± 0.6	64 ± 1.2 (77 ± 1.0)	3.1 (3.1)
PPα-Tri-3	4051.291	4051.319	41 ± 1.2	65 ± 1.2 (83 ± 0.6)	3.1 (3.3)
PPα-Tri-4	4051.291	4051.877	43 ± 2.1	78 ± 1.0 (79 ± 2.1)	3.1 (2.8)
PPα-Hept	3809.139	3809.819	49 ± 3.5	≥ 95 (≥ 95)	6.1 (6.7)
PPα-Hept-CHE	4151.214	4150.473	22	≤ 5 -	-
PPα-Pent2	3869.160	3868.650	47 ± 1.5	≥ 95 (≥ 95)	5.7 (4.9)
PPα-Hex2	3841.129	3841.877	49 ± 2.6	≥ 95 (≥ 95)	6.5 (5.3)
PPα-Non	3781.108	3780.971	54 ± 2.9	77 ± 0.6 (78 ± 3.5)	Agg.
α2-TA	9077.715	ND	36	> 95	Agg.
sc-α3-1	16097.5	16104.0	37.4 [d]	78 [d] (78) [e]	1.2 (1.0)
sc-α3-2	14422.6	14428.5	47.1 [d]	84 [d] (82) [e]	1.2 (1.1)

Exp. = expected. Obs. = observed. [a] Calculated using *Equation 3*. Mean of three measurements ± standard deviation. Conditions: 100 μM peptide, PBS, pH 7.4. [b] 1st derivative of thermal denaturation monitored at 222 nm. Values in brackets are the 1st derivative of thermal denaturation monitored in the near-UV region. Mean of three measurements ± standard deviation. Conditions: 100 μM peptide, PBS, pH 7.4. [c] O.S.

= oligomeric state (× monomer mass), as determined by sedimentation equilibrium analytical ultracentrifugation experiments. Values in brackets are the oligomeric state as determined by sedimentation velocity experiments. ND = not determined. Agg. = aggregated at 3 krpm. See section 7.4 for conditions. [d] 5 µM peptide, PBS, pH 7.4. [e] 50 µM peptide, PBS, pH 7.4.

7.4 Characterisation of Designed Peptides and Proteins

PPα-Tri-1

Ac–PPKKPKKP GDNAT EIAAIKYEIAAIKYEIAAIKY–NH_2

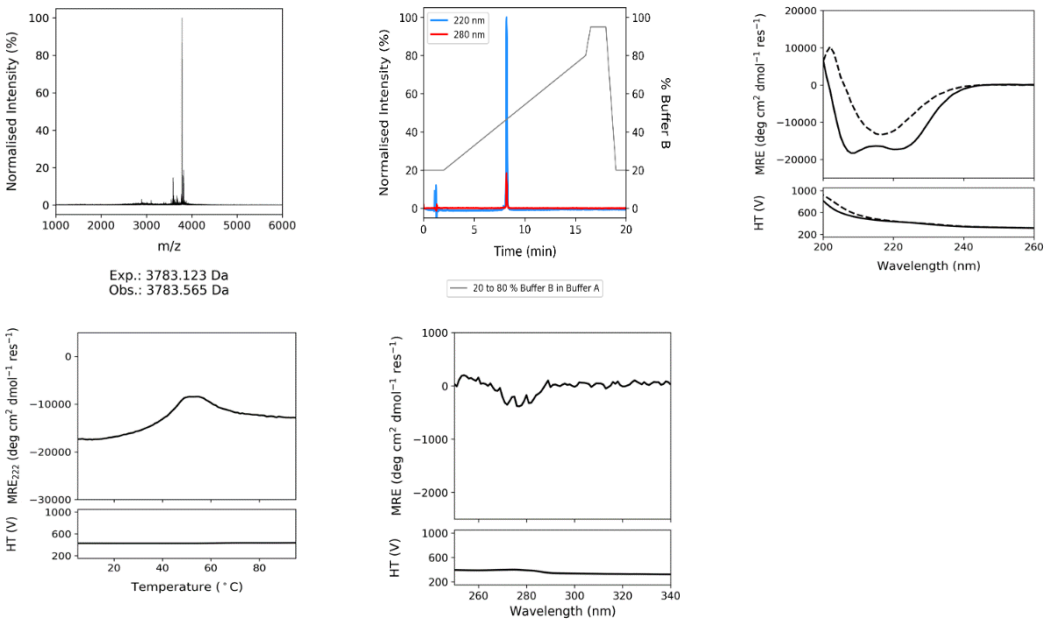


Figure 7.1: *Characterisation of PPα-Tri-1. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (3783.123 Da). Obs. refers to the observed experimentally determined mass (3783.565 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top left, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Bottom left, thermal denaturation profile monitored at 222 nm. Bottom centre, near-UV CD spectrum recorded at 5 ℃. Conditions: 100 μM peptide, PBS, pH 7.4.*

PPα-Tri-2

Ac-PPKKPKKP GDNAT PIELIKYEIELIKYEIELIKY-NH_2

$\bar{v}$ = 0.7726 $cm^3\ g^{-1}$

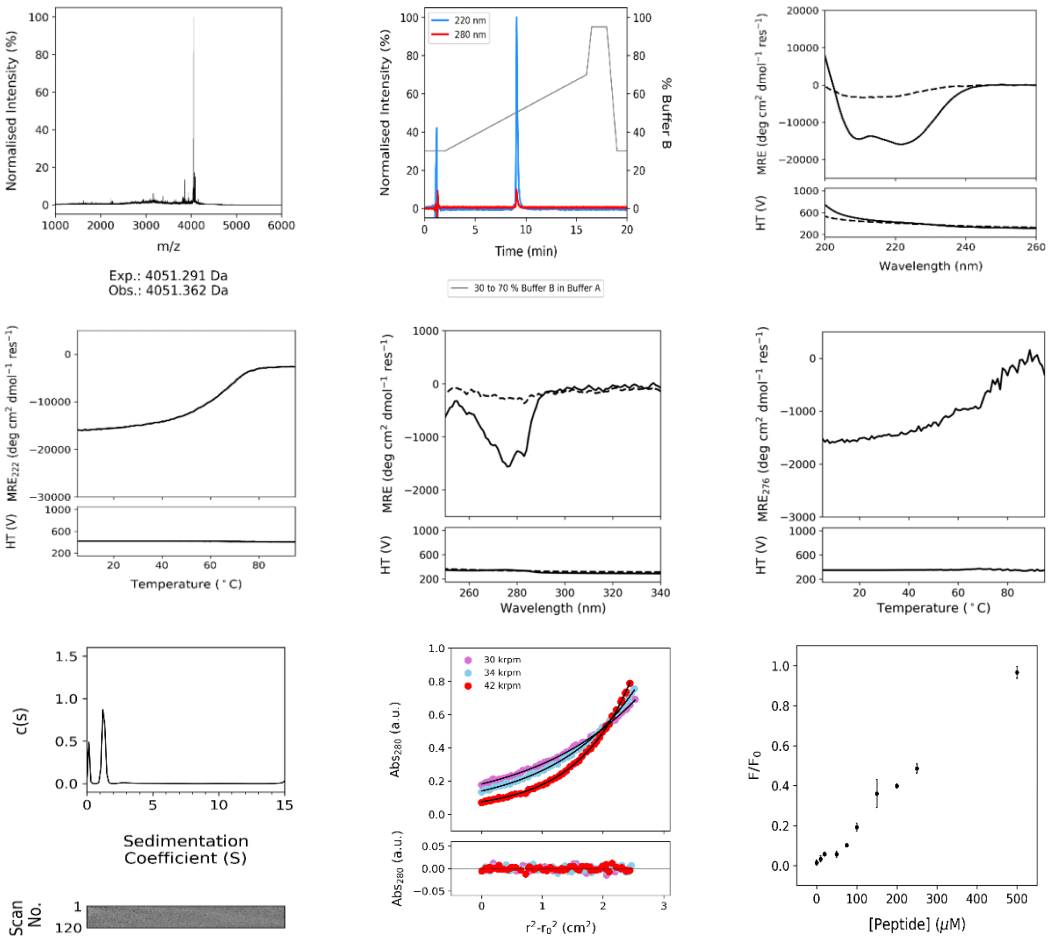


Figure 7.2: *Characterisation of PPα-Tri-2. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (4051.291 Da). Obs. refers to the observed experimentally determined mass (4051.362 Da). Top middle, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 276 nm. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 60 krpm, returning s = 1.275 S, $s_{20,w}$ = 1.331 S, f/f_0 = 1.199, and Mw =*

12368 Da (3.1 × monomer mass), 95% confidence limits. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom centre, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 30, 34 and 42 krpm, returning Mw = 12604 Da (3.1 × monomer mass, 95% confidence limits: 12499 – 12713 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 110 μM peptide, PBS, pH 7.4. Bottom right, 0 – 500 μM peptide, 1 μM DPH, PBS, 5 % (v/v) DMSO, pH 7.4.

PPα-Tri-3

Ac-PPKKPKKP GDNAT PIELKIYEIELKIYEIELKIY-NH_2

$\bar{v}$ = 0.7726 $cm^3\ g^{-1}$

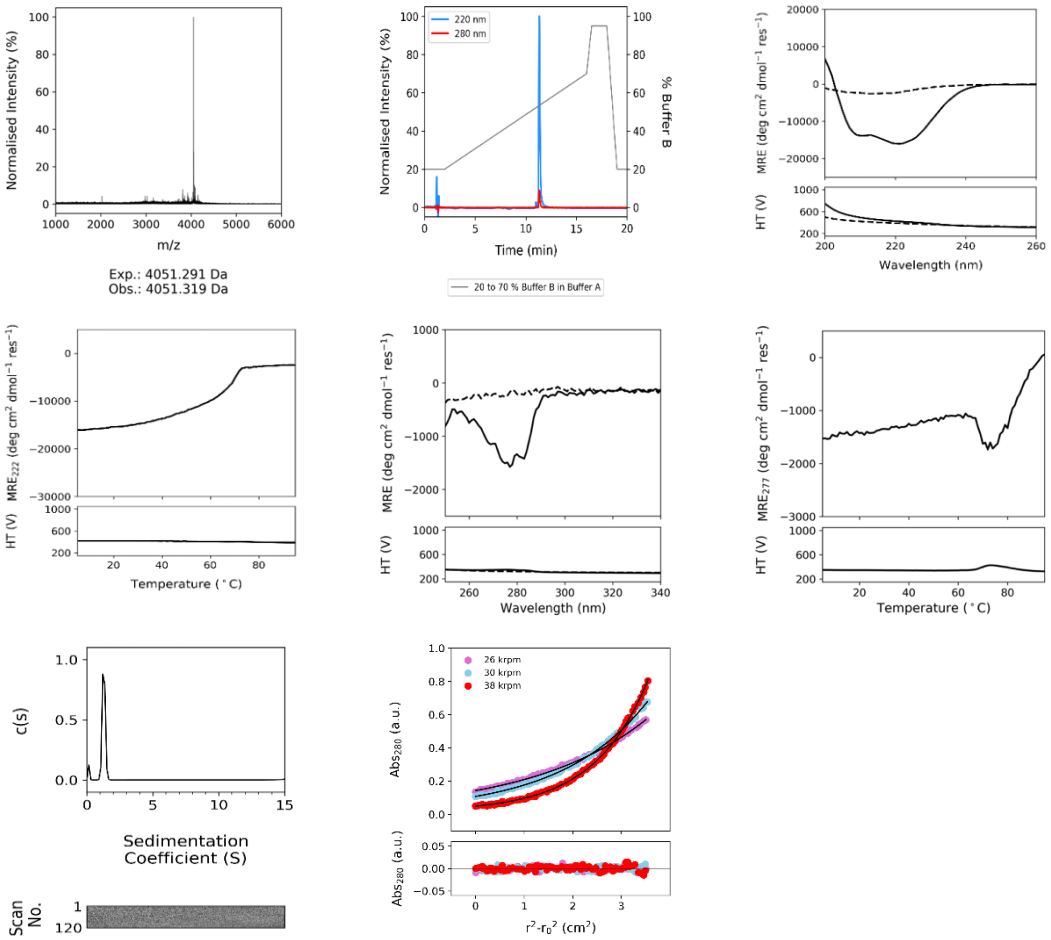


Figure 7.3: *Characterisation of PPα-Tri-3. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (4051.291 Da). Obs. refers to the observed experimentally determined mass (4051.319 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 °C (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 °C (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 277 nm. Conditions: 100 µM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 60 krpm, returning s = 1.285 S, $s_{20,w}$ = 1.340 S, f/f_0 = 1.251, and Mw =*

13326 Da (3.3 × monomer mass), 95% confidence limits. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom centre, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 26, 30 and 38 krpm, returning Mw = 12388 Da (3.1 × monomer mass, 95% confidence limits: 12283 – 12495 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 100 μM peptide, PBS, pH 7.4.

PPα-Tri-4

Ac-PPKKPKKP GDNAT PLELIKYELELIKYELELIKY-NH_2

$\bar{v}$ = 0.7726 $cm^3\ g^{-1}$

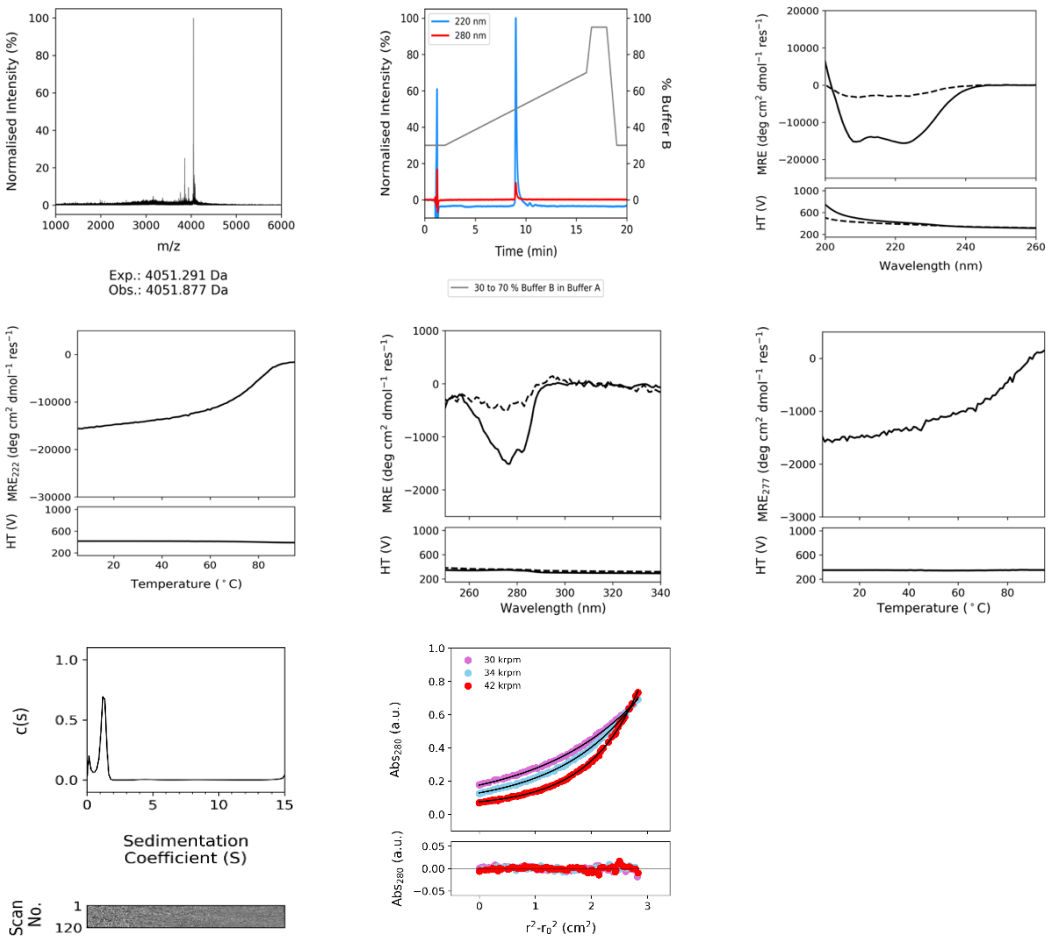


Figure 7.4: *Characterisation of PPα-Tri-4. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (4051.291 Da). Obs. refers to the observed experimentally determined mass (4051.877 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 277 nm. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.213 S, $s_{20,w}$ = 1.266 S, f/f_0 = 1.200, and*

Mw = 11491 Da (2.8 × monomer mass), 95% confidence limits. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom centre, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 30, 34 and 42 krpm, returning Mw = 12493 Da (3.1 × monomer mass, 95% confidence limits: 12393 – 12591 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 100 μM peptide, PBS, pH 7.4.

PPα-Hept

Ac-PPKKPKKP GDNAT PLEKIAYALEKIAYALEKIAY-NH_2

$\bar{v}$ = 0.7658 $cm^3\ g^{-1}$

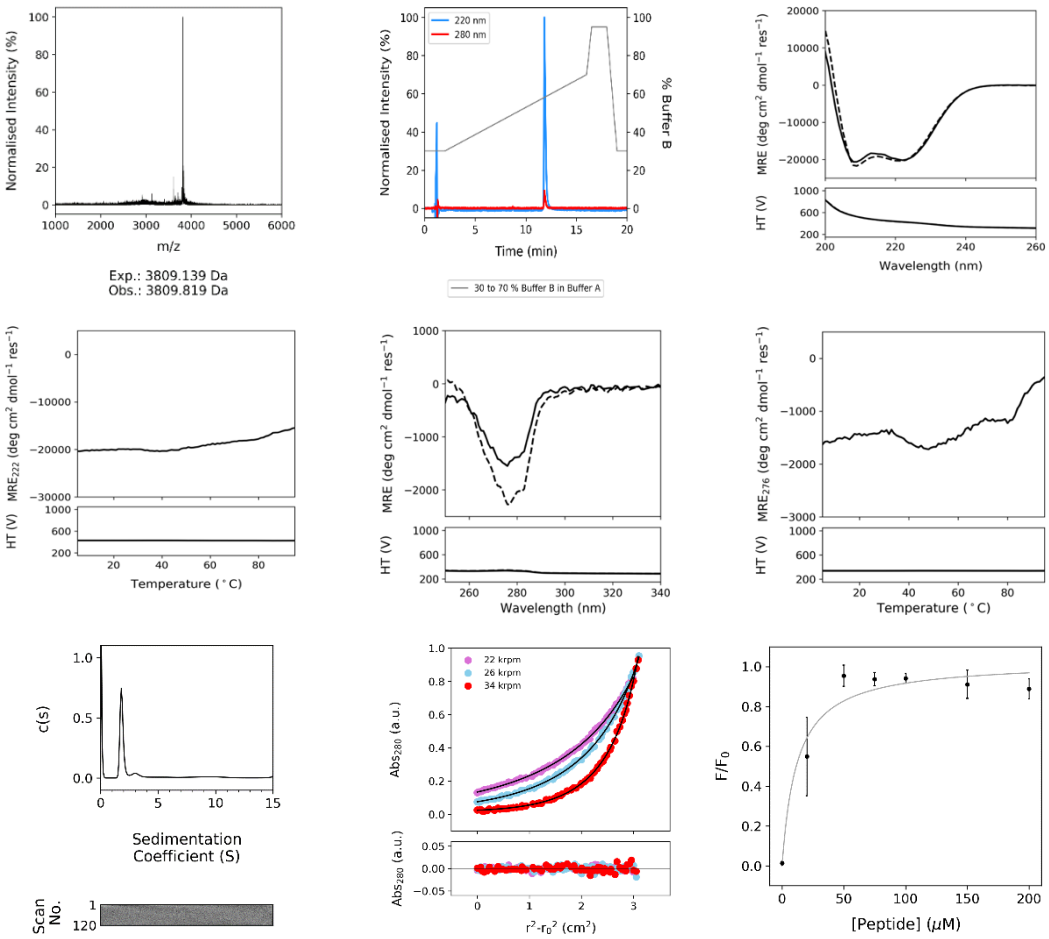


Figure 7.5: *Characterisation of PPα-Hept. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (3809.139 Da). Obs. refers to the observed experimentally determined mass (3809.819 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 276 nm. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.854 S, $s_{20,w}$ = 1.915 S, f/f_0 = 1.280, and Mw =*

22448 Da (5.9 × monomer mass) for peak 1, and s = 3.193 S, $s_{20,w}$ = 3.298 S, f/f_0 = 1.280, and Mw = 50726 Da (13.3 × monomer mass) for peak 2, 95% confidence limits. Conditions: 100 µM peptide, TBS, pH 7.4. Bottom centre, SE data (top, dots) fitted to ideal, two-species model curves (top, black lines) at 22, 26 and 34 krpm, returning Mw = 23200 Da (6.1 × monomer mass, 95% confidence limits: 21718 – 24300 Da) for species 1, Mw = 67121 Da (17.6 × monomer mass, 95 % confidence limits: 53017 – 84140 Da) for species 2, and residuals for the fits (bottom, dots, same colour scheme). Conditions: 110 µM peptide, PBS, pH 7.4. Bottom right, saturation binding curve for binding of PPα-Hept to DPH, fitted to a single-site binding model, returning K_d = 12.1 ± 5.4 µM. 0 – 200 µM peptide, 1 µM DPH, PBS, 5 % (v/v) DMSO, pH 7.4. The fit to the data is shown in grey.

PPα-Hept-CHE

Ac-PPKKPKKP GDNAT PLEKIAYALEKCAYAHEKEAYALK-NH_2

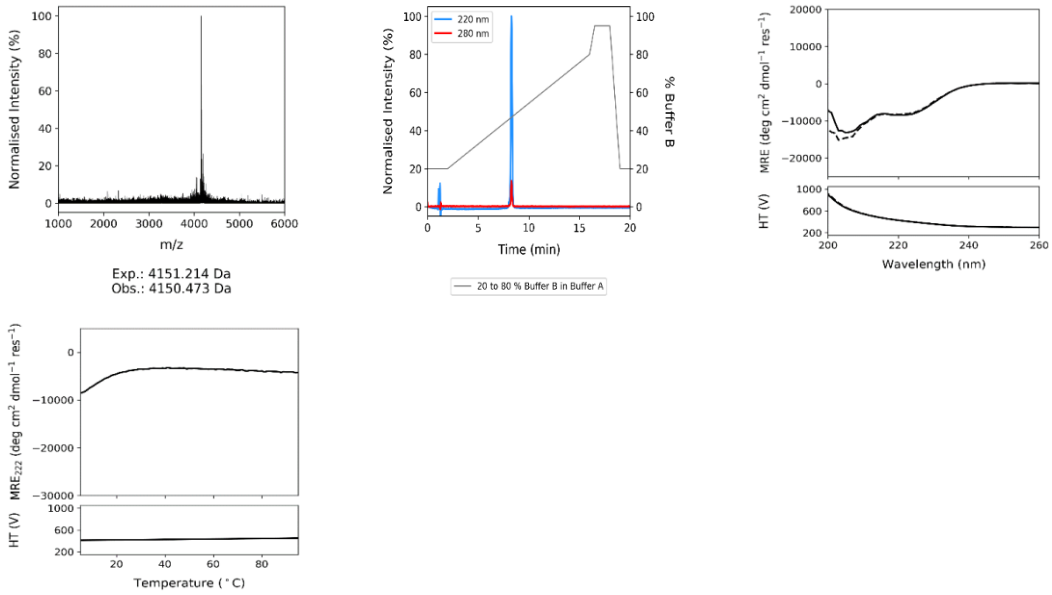


Figure 7.6: *Characterisation of PPα-Hept-CHE. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (4151.214 Da). Obs. refers to the observed experimentally determined mass (4150.473 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Bottom left, thermal denaturation profile monitored at 222 nm. Conditions: 100 μM peptide, 1000 μM TCEP, PBS, pH 7.4.*

PPα-Pent2

Ac-PPKKPKKP GDNAT PLEKIAYTLEKIAYTLEKIAY-NH_2

$\bar{v}$ = 0.7633 cm^3 g^{-1}

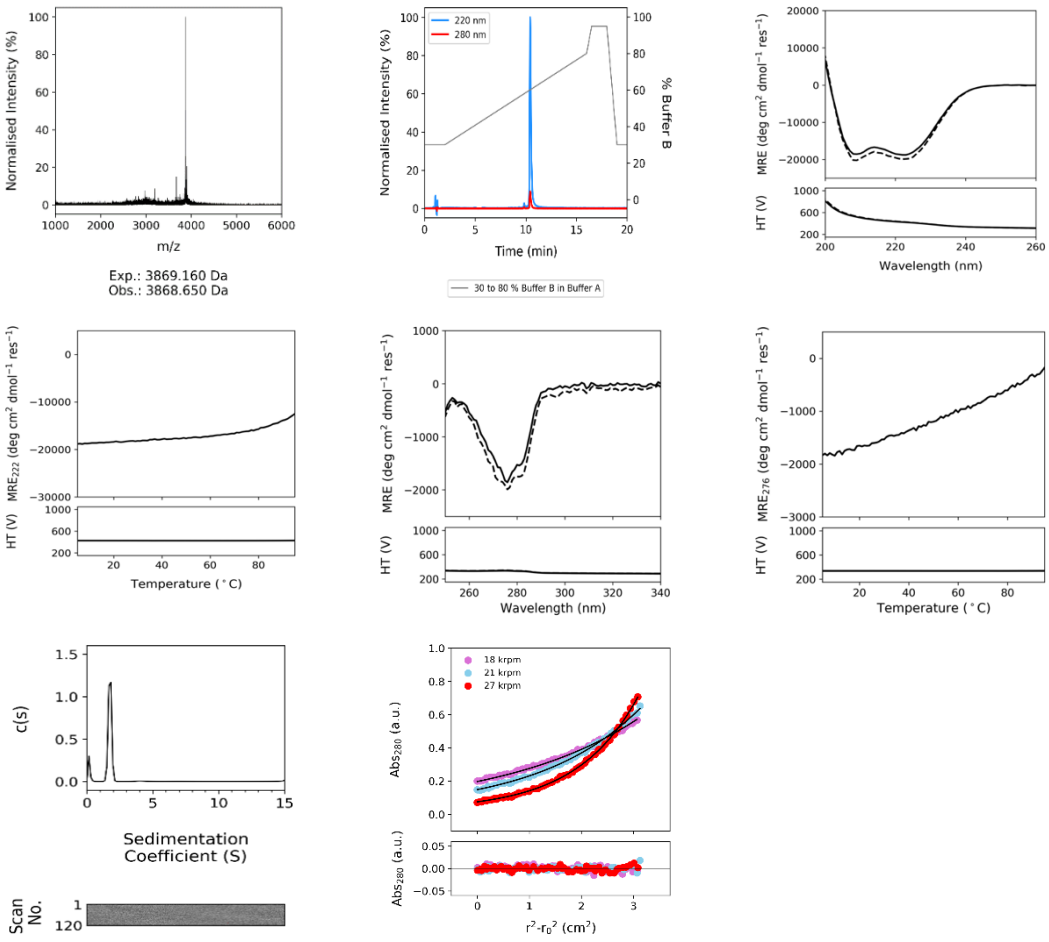


Figure 7.7: *Characterisation of PPα-Pent2. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (3869.160 Da). Obs. refers to the observed experimentally determined mass (3868.650 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 276 nm. Conditions: 100 µM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.747 S, $s_{20,w}$ = 1.820 S, f/f_0 = 1.216, and Mw =*

18918 Da (4.9 × monomer mass), 95% confidence limits. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom centre, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 18, 21 and 27 krpm, returning Mw = 22178 Da (5.7 × monomer mass, 95% confidence limits: 21886 – 22469 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 100 μM peptide, PBS, pH 7.4.

PPα-Hex2

Ac-PPKKPKKP GDNAT PLEKIAYSLEKIAYSLEKIAY-NH_2

$\bar{v}$ = 0.7605 $cm^3\ g^{-1}$

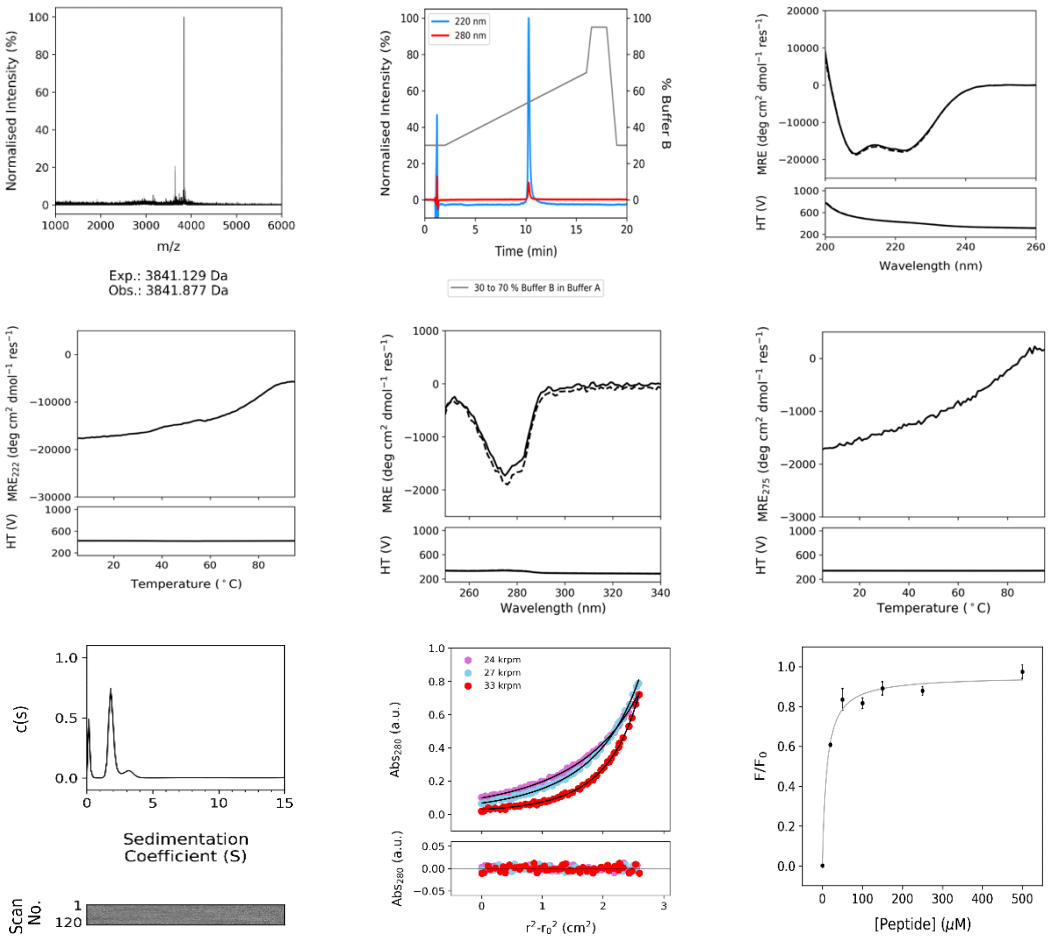


Figure 7.8: *Characterisation of PPα-Hex2. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (3841.129 Da). Obs. refers to the observed experimentally determined mass (3841.877 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle left, thermal denaturation profile monitored at 222 nm. Middle centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Middle right, thermal denaturation profile monitored at 275 nm. Conditions: 100 μM peptide, PBS, pH 7.4. Bottom left, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.873 S, $s_{20,w}$ = 1.935 S, f/f_0 = 1.217, and Mw =*

20378 Da (5.3 × monomer mass) for peak 1, and s = 3.210 S, $s_{20,w}$ = 3.318 S, f/f_0 = 1.217, and Mw = 45736 Da (11.9 × monomer mass) for peak 2, 95% confidence limits. Conditions: 100 µM peptide, PBS, pH 7.4. Bottom middle, SE data (top, dots) fitted to ideal, two-species model curves (top, black lines) at 24, 27 and 33 krpm, returning Mw = 24941 Da (6.5 × monomer mass, 95% confidence limits: 23733 – 25730) for species 1, Mw = 70760 Da (18.4 × monomer mass, 95 % confidence limits: 58215 – 83478 Da) for species 2, and residuals for the fits (bottom, dots, same colour scheme). Conditions: 100 µM peptide, PBS, pH 7.4. Bottom right, saturation binding curve for binding of PPα-Hex2 to DPH, fitted to a single-site binding model, returning K_d = 10.7 ± 2.2 µM. 0 – 500 µM peptide, 1 µM DPH, PBS, 5 % (v/v) DMSO, pH 7.4. The fit to the data is shown in grey.

PPα-Non

Ac-PPKKPKKP GDNAT PLEKIAYGLEKIAYGLEKIAY-NH_2

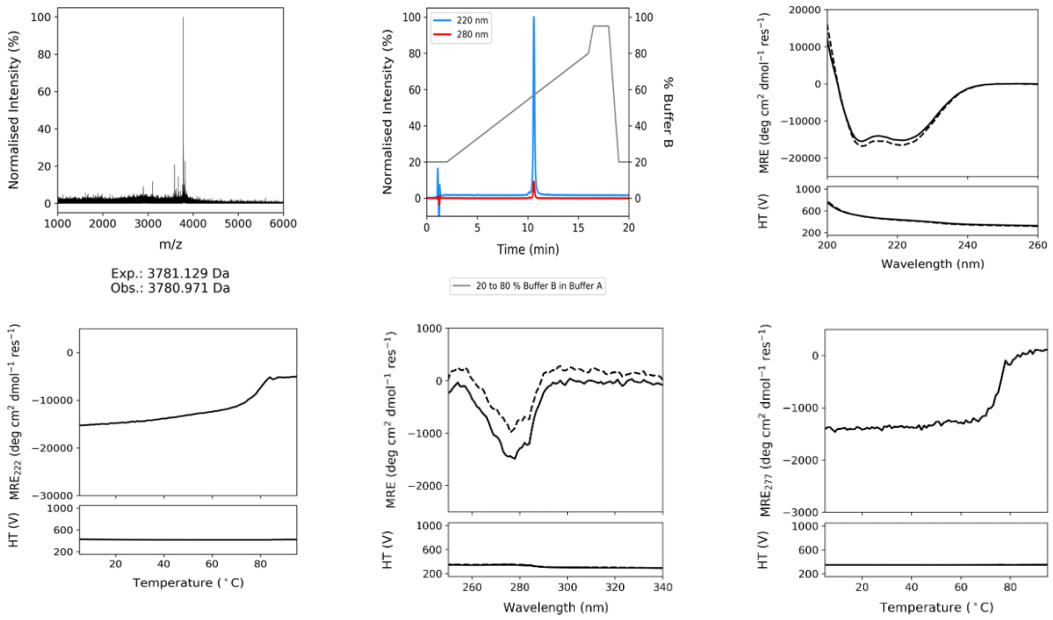


Figure 7.9: *Characterisation of PPα-Non. Top left, MALDI-TOF MS data. Exp. denotes the expected or calculated monomeric mass (3781.129 Da). Obs. refers to the observed experimentally determined mass (3780.971 Da). Top centre, analytical HPLC traces (220 nm, blue, and 280 nm, red). Top right, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Bottom left, thermal denaturation profile monitored at 222 nm. Bottom centre, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Bottom right, thermal denaturation profile monitored at 277 nm. Conditions: 100 µM peptide, PBS, pH 7.4.*

α2-TA

MGSSHHHHHHSSGLVPRGSGTLEKIAQTLEKIAWTLEKIAQSEGSEGSEGPPKKPKKPG
DNATPLEKIAYALEKIAYALEKIAY

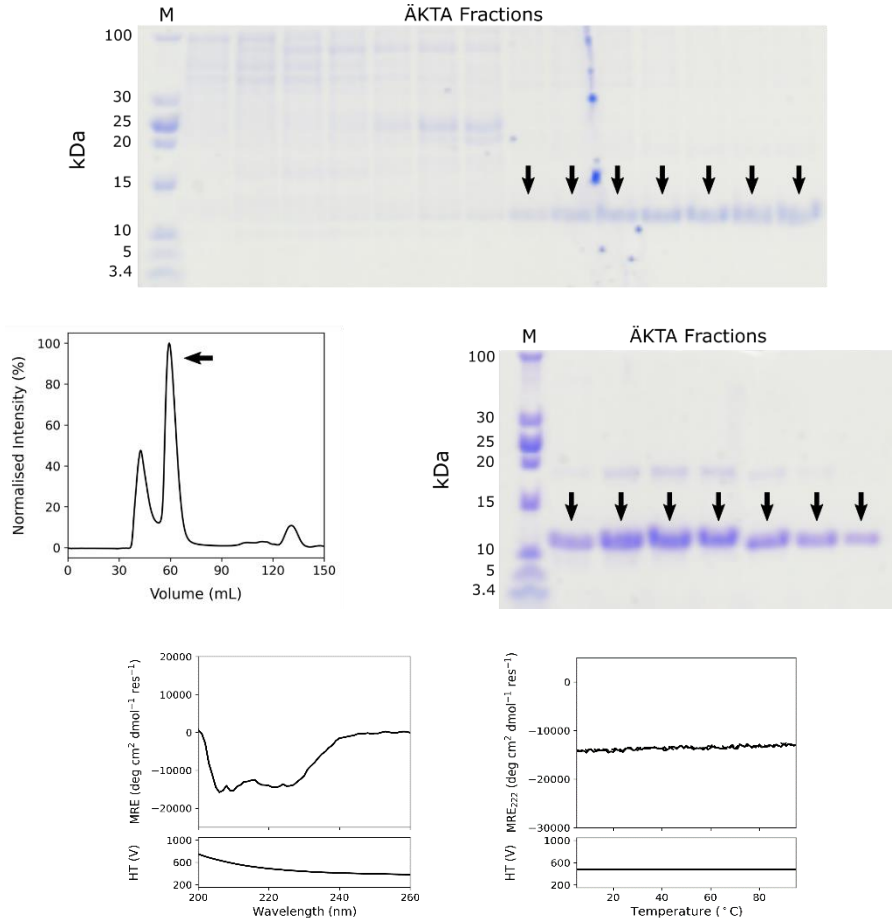
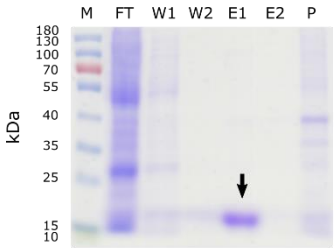


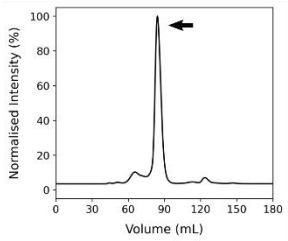
Figure 7.10: *Characterisation of α2-TA. Top, 13% SDS-PAGE of fractions from Ni affinity chromatography. Middle left, SEC trace monitored at 280 nm. Middle right, 13% SDS-PAGE of fractions from SEC. Bottom left, far-UV CD spectra recorded at 20 °C. Bottom right, thermal denaturation profile monitored at 222 nm. Conditions: 10 μM protein, 100 mM $NaPO_4$, 100 mM NaCl, pH 5.8.*

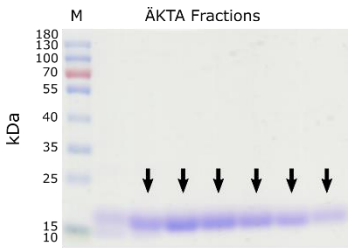
sc-α3-1

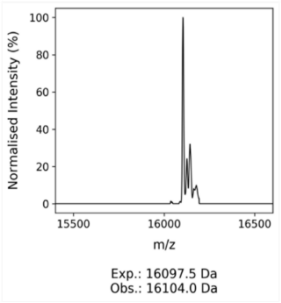
MGSSHHHHHHSSGLVPRGSHMMPPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSE
GSEGSEGPPKKPKKPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEGPPKKPKKP
GDNATPIELIKYEIELIKYEIELIKYG

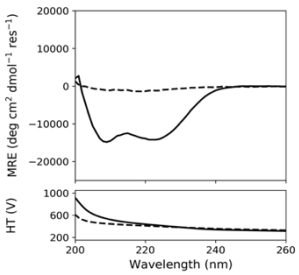
$\bar{v}$ = 0.7466 $cm^3\ g^{-1}$

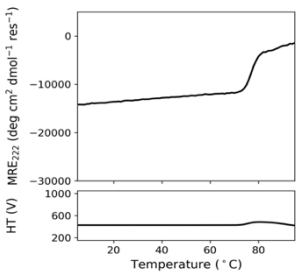


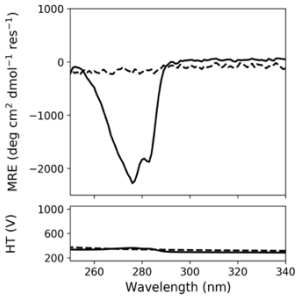


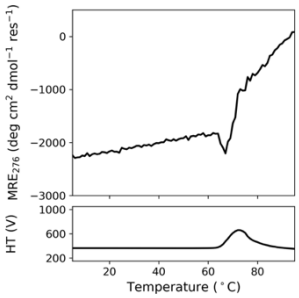


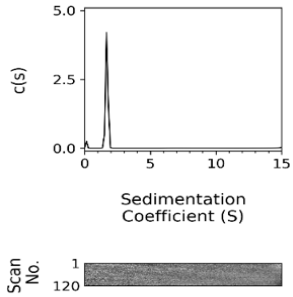












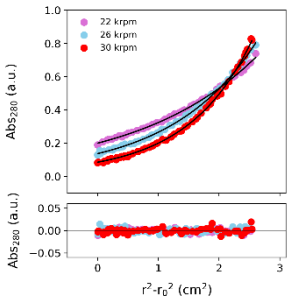
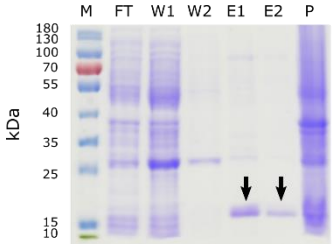


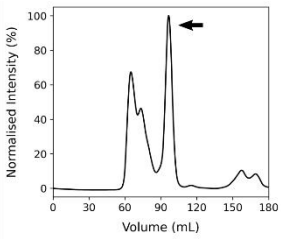
Figure 7.11: *Characterisation of sc-α3-1. Top left, 12% SDS-PAGE of fractions from Ni affinity chromatography. Top centre, SEC trace monitored at 280 nm. Top right, 12% SDS-PAGE of fractions from SEC. Upper middle left, ESI MS data. Exp. denotes the expected or calculated monomeric mass (16097.5 Da). Obs. refers to the observed experimentally determined mass (16104.0 Da). Upper middle centre, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Upper middle right, thermal denaturation profile monitored at 222 nm. Conditions: 5 μM protein, PBS, pH 7.4. Lower middle left, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Lower middle centre, thermal denaturation profile monitored at 276 nm. Conditions: 50 μM protein, PBS, pH 7.4. Lower middle right, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.692 S, $s_{20,w}$ = 1.760 S, f/f_0 = 1.209, and Mw = 15921 Da (1.0 × monomer mass), 95% confidence limits. Conditions: 70 μM protein, PBS, pH 7.4. Bottom left, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 22, 26 and 30 krpm, returning Mw = 18694 Da (1.2 × monomer mass, 95% confidence limits: 18385 – 19013 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 35 μM protein, PBS, pH 7.4.*

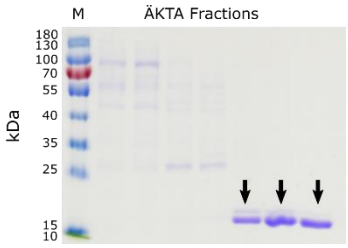
sc-α3-2

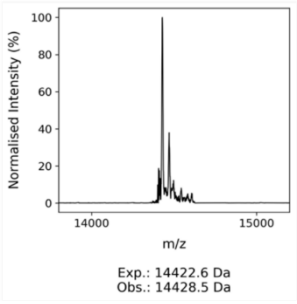
MGSSHHHHHHSSGLVPRGSHMGEIAAIKQEIAAIKKEIAAIKQGSEGSEGSEGPPKKPK
KPGDNATPIELIKYEIELIKYEIELIKYGSEGSEGSEGPPKKPKKPGDNATPIELIKYE
IELIKYEIELIKYG

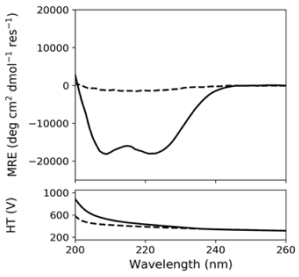
$\bar{v}$ = 0.7446 $cm^3\ g^{-1}$

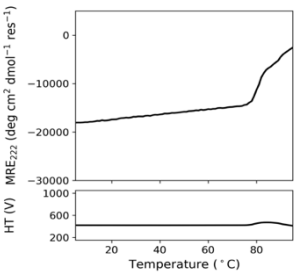


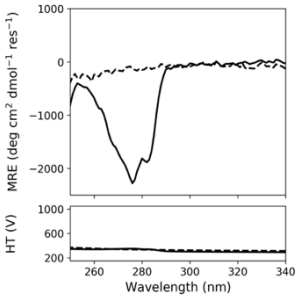


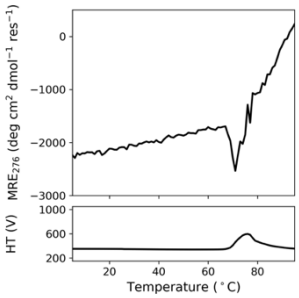


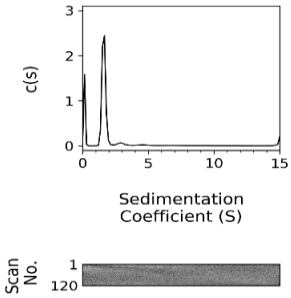












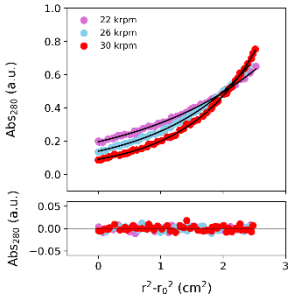


Figure 7.12: *Characterisation of sc-α3-2. Top left, 12% SDS-PAGE of fractions from Ni affinity chromatography. Top centre, SEC trace monitored at 280 nm. Top right, 12% SDS-PAGE of fractions from SEC. Upper middle left, ESI MS data. Exp. denotes the expected or calculated monomeric mass (14422.6 Da). Obs. refers to the observed experimentally determined mass (14428.5 Da). Upper middle centre, far-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Upper middle right, thermal denaturation profile monitored at 222 nm. Conditions: 5 μM protein, PBS, pH 7.4. Lower middle left, near-UV CD spectra recorded at 5 ℃ (dashed line indicates the post-thermal denaturation spectrum). Lower middle centre, thermal denaturation profile monitored at 276 nm. Conditions: 50 μM protein, PBS, pH 7.4. Lower middle right, SV data (top, dots), continuous c(s) fits (top, black lines), and residuals bitmap (bottom) at 50 krpm, returning s = 1.621 S, $s_{20,w}$ = 1.685 S, f/f_0 = 1.307, and Mw = 16559 Da (1.1 × monomer mass), 95% confidence limits. Conditions: 110 μM protein, PBS, pH 7.4. Bottom left, SE data (top, dots) fitted to single ideal species model curves (top, black lines) at 22, 26 and 30 krpm, returning Mw = 17880 Da (1.2 × monomer mass, 95% confidence limits: 17504 – 18270 Da), and residuals for the fits (bottom, dots, same colour scheme). Conditions: 55 μM protein, PBS, pH 7.4.*

sc-α7

MGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDNATPLE
KIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAY
GSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDN
ATPLEKIAYALEKIAYALEKIAYGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYAL
EKIAYGSEGSEGPPKKPKKPGDNATPLEKIAYALEKIAYALEKIAYGLEHHHHHH

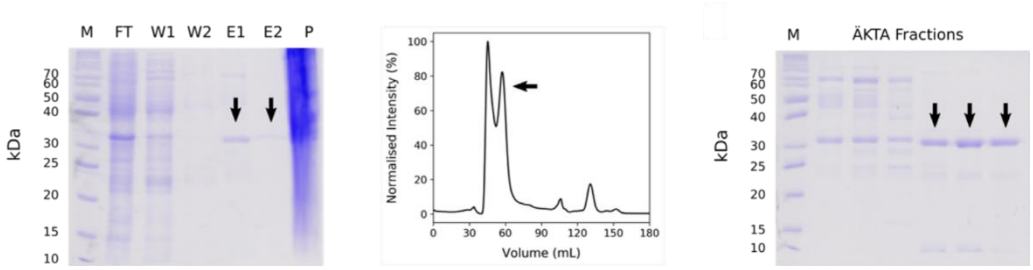


Figure 7.13: *Characterisation of sc-α7. Left, 12% SDS-PAGE of fractions from Ni affinity chromatography. Centre, SEC trace monitored at 280 nm. Right, 12% SDS-PAGE of fractions from SEC.*

References

1. Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H., Jr., *Proc. Natl. Acad. Sci. U.S.A.,* **1961**, *47*, 1309-1314.
2. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., *Nature,* **2020**, *577*, 706-710.
3. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; C., M.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pererira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D., *Science,* **2021**, *373*, 871-876.
4. Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R., *Annu. Rev. Biophys.,* **2008**, *37*, 289-316.
5. Dill, K. A.; MacCallum, J. L., *Science,* **2012**, *338*, 1042-1046.
6. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., *Nature,* **1958**, *181*, 662-666.
7. Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C., *Nature,* **1960**, *185*, 422-427.
8. Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C., *Nature,* **1960**, *185*, 416-22.
9. Pauling, L.; Corey, R. B.; Branson, H. R., *Proc. Natl. Acad. Sci. U.S.A.,* **1951**, *37*, 205-211.
10. Anfinsen, C. B., *Science,* **1973**, *181*, 223-230.
11. Levinthal, C. In *How to Fold Graciously*, Mossbauer Spectroscopy in Biological Systems, 67: 1969; pp 22-24.
12. Baldwin, R. L., *Nature,* **1994**, *369*, 183-184.
13. Baldwin, R. L., *J. Biomol. NMR,* **1995**, *5*, 103-109.
14. Dill, K. A.; Chan, H. S., *Nat. Struct. Biol.,* **1997**, *4*, 10-19.
15. Radford, S. E.; Dobson, C. M.; Evans, P. A., *Nature,* **1992**, *358*, 302-307.
16. Wang, Z.; Mottonen, J.; Goldsmith, E. J., *Biochemistry,* **1996**, *35*, 16443-16448.

17. Dill, K. A. https://commons.wikimedia.org/wiki/File:Funnel-shaped_energy_landscape.png (accessed 19/07/2021).
18. https://creativecommons.org/licenses/by/4.0/deed.en (accessed 19/07/2021).
19. Karplus, M., *Nat. Chem. Biol.,* **2011**, *7*, 401-404.
20. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V., *J. Mol. Biol.,* **1963**, *7*, 95-99.
21. Gessmann, R.; Bruckner, H.; Petratos, K., *J. Pept. Sci.,* **2003**, *9*, 753-762.
22. Fodje, M. N.; Al-Karadaghi, S., *Protein Eng.,* **2002**, *15*, 353-358.
23. Adzhubei, A. A.; Sternberg, M. J.; Makarov, A. A., *J. Mol. Biol.,* **2013**, *425*, 2100-2132.
24. Kumar, P.; Bansal, M., *J. Struct. Biol.,* **2016**, *196*, 414-425.
25. Dill, K. A., *Biochemistry,* **1990**, *29*, 7133-7155.
26. Pace, C. N.; Scholtz, J. M.; Grimsley, G. R., *FEBS Lett.,* **2014**, *588*, 2177-2184.
27. Barlow, D. J.; Thornton, J. M., *J. Mol. Biol.,* **1983**, *168*, 867-885.
28. Kendrew, J. C., *Science,* **1963**, *139*, 1259-1266.
29. Yohannan, S.; Faham, S.; Yang, D.; Grosfeld, D.; Chamberlain, A. K.; Bowie, J. U., *J. Am. Chem. Soc.,* **2004**, *126*, 2284-2285.
30. Arbely, E.; Arkin, I. T., *J. Am. Chem. Soc.,* **2004**, *126*, 5362-5363.
31. Newberry, R. W.; VanVeller, B.; Guzei, I. A.; Raines, R. T., *J. Am. Chem. Soc.,* **2013**, *135*, 7843-7846.
32. Newberry, R. W.; Raines, R. T., *Nat. Chem. Biol.,* **2016**, *12*, 1084-1088.
33. Tatko, C. D.; Waters, M. L., *J. Am. Chem. Soc.,* **2002**, *124*, 9372-9373.
34. Butterfield, S. M.; Patel, P. R.; Waters, M. L., *J. Am. Chem. Soc.,* **2002**, *124*, 9751-9755.
35. Serrano, L.; Bycroft, M.; Fersht, A. R., *J. Mol. Biol.,* **1991**, *218*, 465-475.
36. Shi, Z.; Olson, C. A.; Bell, A. J., Jr.; Kallenbach, N. R., *Biophys. Chem.,* **2002**, *101-102*, 267-279.
37. Meyer, E. A.; Castellano, R. K.; Diederich, F., *Angew. Chem. Int. Ed.,* **2003**, *42*, 1210-1250.
38. Brandl, M.; Weiss, M. S.; Jabs, A.; Suhnel, J.; Hilgenfeld, R., *J. Mol. Biol.,* **2001**, *307*, 357-377.
39. Nishio, M.; Umezawa, Y.; Fantini, J.; Weiss, M. S.; Chakrabarti, P., *Phys. Chem. Chem. Phys.,* **2014**, *16*, 12648-12683.
40. Luisi, P. L.; Chiarabelli, C.; Stano, P., *Orig. Life Evol. Biosph.,* **2006**, *36*, 605-616.
41. Woolfson, D. N.; Bartlett, G. J.; Burton, A. J.; Heal, J. W.; Niitsu, A.; Thomson, A. R.; Wood, C. W., *Curr. Opin. Struct. Biol.,* **2015**, *33*, 16-26.

42. Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, J. T., Jonassen, I., *Structure,* **2009**, *17*, 1244-1252.
43. Woolfson, D. N., *J. Mol. Biol.,* **2021**, 167160.
44. DeGrado, W. F.; Wasserman, Z. R.; Lear, J. D., *Science,* **1989**, *243*, 622-628.
45. Curnow, P., *Biochem. Soc. Trans.,* **2019**, *47*, 1233-1245.
46. Jiang, B.; Guo, T.; Peng, L.-W.; Sun, Z.-R., *Biopolymers,* **1997**, *45*, 35-49.
47. Costantini, S.; Colonna, G.; Facchiano, A. M., *Biochem. Biophys. Res. Commun.,* **2006**, *342*, 441-451.
48. Fujiwara, K.; Toda, H.; Ikeguchi, M., *BMC Struct. Biol.,* **2012**, *12*, 18.
49. Porebski, B. T.; Buckle, A. M., *Protein Eng. Des. Sel.,* **2016**, *29*, 245-251.
50. Regan, L.; DeGrado, W. F., *Science,* **1988**, *241*, 976-978.
51. Hecht, M. H.; Richardson, J. S.; Richardson, D. C.; Ogden, R. C., *Science,* **1990**, *249*, 884-891.
52. Lovejoy, B.; Choe, S.; Cascio, D.; McRorie, D. K.; DeGrado, W. F.; Eisenberg, D., *Science,* **1993**, *259*, 1288-1293.
53. Nautiyal, S.; Woolfson, D. N.; King, D. S.; Alber, T., *Biochemistry,* **1995**, *34*, 11645-11651.
54. Lindhout, D. A.; Litowski, J. R.; Mercier, P.; Hodges, R. S.; Sykes, B. D., *Biopolymers,* **2004**, *75*, 367-375.
55. O'Shea, E. K.; Lumb, K. J.; Kim, P. S., *Curr. Biol.,* **1993**, *3*, 658-667.
56. Zimenkov, Y.; Dublin, S. N.; Ni, R.; Tu, R. S.; Breedveld, V.; Apkarian, R. P.; Conticello, V. P., *J. Am. Chem. Soc.,* **2006**, *128*, 6770-6771.
57. Rhys, G. G.; Wood, C. W.; Beesley, J. L.; Zaccai, N. R.; Burton, A. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., *J. Am. Chem. Soc.,* **2019**, *141*, 8787-8797.
58. Yu, F.; Cangelosi, V. M.; Zastrow, M. L.; Tegoni, M.; Plegaria, J. S.; Tebo, A. G.; Mocny, C. S.; Ruckthong, L.; Qayyum, H.; Pecoraro, V. L., *Chem. Rev.,* **2014**, *114*, 3495-3578.
59. Lombardi, A.; Pirro, F.; Maglio, O.; Chino, M.; DeGrado, W. F., *Acc. Chem. Res.,* **2019**, *52*, 1148-1159.
60. Fletcher, J. M.; Boyle, A. L.; Bruning, M.; Bartlett, G. J.; Vincent, T. L.; Zaccai, N. R.; Armstrong, C. T.; Bromley, E. H.; Booth, P. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., *ACS Synth. Biol.,* **2012**, *1*, 240-250.
61. Fletcher, J. M.; Harniman, R. L.; Barnes, F. R.; Boyle, A. L.; Collins, A.; Mantell, J.; Sharp, T. H.; Antognozzi, M.; Booth, P. J.; Linden, N.; Miles, M. J.; Sessions, R. B.; Verkade, P.; Woolfson, D. N., *Science,* **2013**, *340*, 595-599.

62. Main, E. R.; Jackson, S. E.; Regan, L., *Curr. Opin. Struct. Biol.,* **2003**, *13*, 482-489.
63. Kohl, A.; Binz, H. K.; Forrer, P.; Stumpp, M. T.; Pluckthun, A.; Grutter, M. G., *Proc. Natl. Acad. Sci. U.S.A.,* **2003**, *100*, 1700-1705.
64. Main, E. R.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L., *Structure,* **2003**, *11*, 497-508.
65. Ptitsyn, O., *FEBS Lett.,* **1981**, *131*, 197-202.
66. Ramirez-Alvarado, M.; Blanco, F. J.; Serrano, L., *Nat. Struct. Biol.,* **1996**, *3*, 604-612.
67. Kortemme, T.; Ramirez-Alvarado, M.; Serrano, L., *Science,* **1998**, *281*, 253-256.
68. Schenck, H. L., Gellman, S. H., *J. Am. Chem. Soc.,* **1998**, *120*, 4869-4870.
69. Ponder, J. W.; Richards, F. M., *J. Mol. Biol.,* **1987**, *193*, 775-791.
70. Desjarlais, J. R.; Handel, T. M., *Protein Sci.,* **1995**, *4*, 2006-2018.
71. Dahiyat, B. I.; Mayo, S. L., *Science,* **1997**, *278*, 82-87.
72. Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S., *Science,* **1998**, *282*, 1462-1467.
73. Thomson, A. R.; Wood, C. W.; Burton, A. J.; Bartlett, G. J.; Sessions, R. B.; Brady, R. L.; Woolfson, D. N., *Science,* **2014**, *346*, 485-488.
74. Brunette, T. J.; Parmeggiani, F.; Huang, P. S.; Bhabha, G.; Ekiert, D. C.; Tsutakawa, S. E.; Hura, G. L.; Tainer, J. A.; Baker, D., *Nature,* **2015**, *528*, 580-584.
75. Doyle, L.; Hallinan, J.; Bolduc, J.; Parmeggiani, F.; Baker, D.; Stoddard, B. L.; Bradley, P., *Nature,* **2015**, *528*, 585-588.
76. Wood, C. W.; Bruning, M.; Ibarra, A. A.; Bartlett, G. J.; Thomson, A. R.; Sessions, R. B.; Brady, R. L.; Woolfson, D. N., *Bioinformatics,* **2014**, *30*, 3029-3035.
77. Wood, C. W.; Woolfson, D. N., *Protein Sci.,* **2018**, *27*, 103-111.
78. Wood, C. W.; Heal, J. W.; Thomson, A. R.; Bartlett, G. J.; Ibarra, A. A.; Brady, R. L.; Sessions, R. B.; Woolfson, D. N., *Bioinformatics,* **2017**, *33*, 3043-3050.
79. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., *Science,* **2003**, *302*, 1364-1368.
80. Huang, P. S.; Boyken, S. E.; Baker, D., *Nature,* **2016**, *537*, 320-327.
81. Korendovych, I. V.; DeGrado, W. F., *Q. Rev. Biophys,* **2020**, *53*, e3, 1-33.
82. Pan, X.; Kortemme, T., *J. Biol. Chem.,* **2021**, *296*, 100558.
83. Anishchenko, I.; Pellock, S. J.; Chidyausiku, T. M.; Ramelot, T. A.; Ovchinnikov, S.; Hao, J.; Bafna, K.; Norn, C.; Kang, A.; Bera, A. K.; DiMaio, F.; Carter, L.; Chow, C. M.; Montelione, G. T.; Baker, D., *Nature,* **2021**, *600*, 547-552.

84. Huang, P. S.; Ban, Y. E.; Richter, F.; Andre, I.; Vernon, R.; Schief, W. R.; Baker, D., *PLoS One,* **2011**, *6*, e24109.
85. Dawson, W. M.; Rhys, G. G.; Woolfson, D. N., *Curr. Opin. Chem. Biol.,* **2019**, *52*, 102-111.
86. Sesterhenn, F.; Yang, C.; Bonet, J.; Cramer, J. T.; Wen, X.; Wang, Y.; Chiang, C. I.; Abriata, L. A.; Kucharska, I.; Castoro, G.; Vollers, S. S.; Galloux, M.; Dheilly, E.; Rosset, S.; Corthesy, P.; Georgeon, S.; Villard, M.; Richard, C. A.; Descamps, D.; Delgado, T.; Oricchio, E.; Rameix-Welti, M. A.; Mas, V.; Ervin, S.; Eleouet, J. F.; Riffault, S.; Bates, J. T.; Julien, J. P.; Li, Y.; Jardetzky, T.; Krey, T.; Correia, B. E., *Science,* **2020**, *368*, 730-739.
87. Yang, C.; Sesterhenn, F.; Bonet, J.; van Aalen, E. A.; Scheller, L.; Abriata, L. A.; Cramer, J. T.; Wen, X.; Rosset, S.; Georgeon, S.; Jardetzky, T.; Krey, T.; Fussenegger, M.; Merkx, M.; Correia, B. E., *Nat. Chem. Biol.,* **2021**, *17*, 492-500.
88. Rackham, O. J.; Madera, M.; Armstrong, C. T.; Vincent, T. L.; Woolfson, D. N.; Gough, J., *J. Mol. Biol.,* **2010**, *403*, 480-493.
89. Liu, J.; Rost, B., *Protein Sci.,* **2001**, *10*, 1970-1979.
90. Hartmann, M. D., Functional and Structural Roles of Coiled Coils. In *Fibrous Proteins: Structures and Mechanisms*, Springer International Publishing: 2017.
91. Walshaw, J.; Woolfson, D. N., *J. Struct. Biol.,* **2003**, *144*, 349-361.
92. Woolfson, D. N.; Bartlett, G. J.; Bruning, M.; Thomson, A. R., *Curr. Opin. Struct. Biol.,* **2012**, *22*, 432-441.
93. Rhys, G. G.; Dawson, W. M.; Beesley, J. L.; Martin, F. J. O.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., *Biomacromolecules,* **2021**, *22*, 2010-2109.
94. Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T., *Science,* **1993**, *262*, 1401-1407.
95. Harbury, P. B.; Kim, P. S.; Alber, T., *Nature,* **1994**, *371*, 80-83.
96. Crick, F. H. C., *Acta Crystallographica,* **1953**, *6*, 689-697.
97. Liu, J.; Zheng, Q.; Deng, Y.; Cheng, C. S.; Kallenbach, N. R.; Lu, M., *Proc. Natl. Acad. Sci. U.S.A.,* **2006**, *103*, 15457-15462.
98. Deng, Y.; Liu, J.; Zheng, Q.; Eliezer, D.; Kallenbach, N. R.; Lu, M., *Structure,* **2006**, *14*, 247-255.
99. Zaccai, N. R.; Chi, B.; Thomson, A. R.; Boyle, A. L.; Bartlett, G. J.; Bruning, M.; Linden, N.; Sessions, R. B.; Booth, P. J.; Brady, R. L.; Woolfson, D. N., *Nat. Chem. Biol.,* **2011**, *7*, 935-941.
100. Rhys, G. G.; Wood, C. W.; Lang, E. J. M.; Mulholland, A. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N., *Nat. Commun.,* **2018**, *9*, 4132.
101. Dawson, W. M.; Martin, F. J.; Rhys, G. R.; Shelley, K. L.; Brady, R. L.; Woolfson, D. N., *Chem. Sci.,* **2021**, *12*, 6923-6928.

102. Lu, P.; Min, D.; DiMaio, F.; Wei, K. Y.; Vahey, M. D.; Boyken, S. E.; Chen, Z.; Fallas, J. A.; Ueda, G.; Sheffler, W.; Mulligan, V. K.; Xu, W.; Bowie, J. U.; Baker, D., *Science,* **2018**, *359*, 1042-1046.
103. Mravic, M.; Thomaston, J. L.; Tucker, M.; Solomon, P. E.; Liu, L.; DeGrado, W. F., *Science,* **2019**, *363*, 1418-1423.
104. Scott, A. J.; Niitsu, A.; Kratochvil, H. T.; Lang, E. J. M.; Sengel, J. T.; Dawson, W. M.; Mahendran, K. R.; Mravic, M.; Thomson, A. R.; Brady, R. L.; Liu, L.; Mulholland, A. J.; Bayley, H.; DeGrado, W. F.; Wallace, M. I.; Woolfson, D. N., *Nat. Chem.,* **2021**, *643*, 643-650.
105. Lombardi, A.; Summa, C. M.; Geremia, S.; Randaccio, L.; Pavone, V.; DeGrado, W. F., *Proc. Natl. Acad. Sci. U.S.A.,* **2000**, *97*, 6298-6305.
106. Webster, A. M.; Peacock, A. F. A., *Chem. Commun.,* **2021**, *57*, 6851-6862.
107. Fletcher, J. M.; Horner, K. A.; Bartlett, G. J.; Rhys, G. G.; Wilson, A. J.; Woolfson, D. N., *Chem. Sci.,* **2018**, *9*, 7656-7665.
108. Lebar, T.; Lainscek, D.; Merljak, E.; Aupic, J.; Jerala, R., *Nat. Chem. Biol.,* **2020**, *16*, 513-519.
109. Edgell, C. L.; Smith, A. J.; Beesley, J. L.; Savery, N. J.; Woolfson, D. N., *ACS Synth. Biol.,* **2020**, *9*, 427-436.
110. Lapenta, F.; Aupic, J.; Strmsek, Z.; Jerala, R., *Chem. Soc. Rev.,* **2018**, *47*, 3530-3542.
111. Thomas, F.; Dawson, W. M.; Lang, E. J. M.; Burton, A. J.; Bartlett, G. J.; Rhys, G. G.; Mulholland, A. J.; Woolfson, D. N., *ACS Synth. Biol.,* **2018**, *7*, 1808-1816.
112. Burton, A. J.; Thomson, A. R.; Dawson, W. M.; Brady, R. L.; Woolfson, D. N., *Nat. Chem.,* **2016**, *8*, 837-844.
113. Bolon, D. N.; Mayo, S. L., *Proc. Natl. Acad. Sci. U.S.A.,* **2001**, *98*, 14274-14279.
114. Moroz, Y. S.; Dunston, T. T.; Makhlynets, O. V.; Moroz, O. V.; Wu, Y.; Yoon, J. H.; Olsen, A. B.; McLaughlin, J. M.; Mack, K. L.; Gosavi, P. M.; van Nuland, N. A.; Korendovych, I. V., *J. Am. Chem. Soc.,* **2015**, *137*, 14905-14911.
115. Kezdy, F. J.; Bender, M. L., *Biochemistry,* **1962**, *1*, 1097-1106.
116. Faller, L.; Sturtevant, J. M., *J. Biol. Chem.,* **1966**, *241*, 4825-4834.
117. Burton, A. J. Design, Characterisation and Functional Evaluation of α-Helical Barrels. PhD thesis, University of Bristol, Bristol, 2015.
118. Tsukada, H.; Blow, D. M., *J. Mol. Biol.,* **1985**, *184*, 703-711.
119. Jin, J.; Baker, E. G.; Wood, C. W.; Bath, J.; Woolfson, D. N.; Turberfield, A. J., *ACS Nano,* **2019**, *13*, 9927-9935.
120. Buchberger, A.; Simmons, C. R.; Fahmi, N. E.; Freeman, R.; Stephanopoulos, N., *J. Am. Chem. Soc.,* **2020**, *142*, 1406-1416.

121. Spencer, R. K.; Hochbaum, A. I., *Biochemistry,* **2016**, *55*, 3214-3223.
122. Spencer, R. K.; Hochbaum, A. I., *Biochemistry,* **2017**, *56*, 5300-5308.
123. Xu, C.; Lu, P.; Gamal El-Din, T. M.; Pei, X. Y.; Johnson, M. C.; Uyeda, A.; Bick, M. J.; Xu, Q.; Jiang, D.; Bai, H.; Reggiano, G.; Hsia, Y.; Brunette, T. J.; Dou, J.; Ma, D.; Lynch, E. M.; Boyken, S. E.; Huang, P. S.; Stewart, L.; DiMaio, F.; Kollman, J. M.; Luisi, B. F.; Matsuura, T.; Catterall, W. A.; Baker, D., *Nature,* **2020**, *585*, 129-134.
124. Baker, E. G.; Bartlett, G. J.; Porter Goff, K. L.; Woolfson, D. N., *Acc. Chem. Res.,* **2017**, *50*, 2085-2092.
125. Miller, J.; McLachlan, A. D.; Klug, A., *EMBO J,* **1985**, *4*, 1609-1614.
126. Lee, M. S.; Gippert, G. P.; Soman, K. V.; Case, D. A.; Wright, P. E., *Science,* **1989**, *245*, 635-637.
127. Struthers, M. D.; Cheng, R. P.; Imperiali, B., *Science,* **1996**, *271*, 342-345.
128. Struthers, M.; Ottesen, J. J.; Imperiali, B., *Fold. Des.,* **1998**, *3*, 95-103.
129. Mezo, A. R.; Ottesen, J. J.; Imperiali, B., *J. Am. Chem. Soc.,* **2001**, *123*, 1002-1003.
130. Mezo, A. R.; Cheng, R. P.; Imperiali, B., *J. Am. Chem. Soc.,* **2001**, *123*, 3885-3891.
131. Ali, M. H.; Peisach, E.; Allen, K. N.; Imperiali, B., *Proc. Natl. Acad. Sci. U.S.A.,* **2004**, *101*, 12183-12188.
132. Ali, M. H.; Taylor, C. M.; Grigoryan, G.; Allen, K. N.; Imperiali, B.; Keating, A. E., *Structure,* **2005**, *13*, 225-234.
133. Kimmel, J. R.; Hayden, J.; Pollock, H. G., *J. Biol. Chem.,* **1975**, *250*, 9369-9376.
134. Blundell, T. L.; Pitts, J. E.; Tickle, I. J.; Wood, S. P.; Wu, C. W., *Proc. Natl. Acad. Sci. U.S.A.,* **1981**, *78*, 4175-4179.
135. Li, X. A.; Sutcliffe, M. J.; Schwartz, T. W.; Dobson, C. M., *Biochemistry,* **1992**, *31*, 1245-1253.
136. Monks, S. A.; Karagianis, G.; Howlett, G. J.; Norton, R. S., *J. Biomol. NMR,* **1996**, *8*, 379-390.
137. Nygaard, R.; Nielbo, S.; Schwartz, T. W.; Poulsen, F. M., *Biochemistry,* **2006**, *45*, 8350-8357.
138. Larson, M. R.; Rajashankar, K. R.; Patel, M. H.; Robinette, R. A.; Crowley, P. J.; Michalek, S.; Brady, L. J.; Deivanayagam, C., *Proc. Natl. Acad. Sci. U.S.A.,* **2010**, *107*, 5983-5988.
139. Mieher, J. L.; Larson, M. R.; Schormann, N.; Purushotham, S.; Wu, R.; Rajashankar, K. R.; Wu, H.; Deivanayagam, C., *Infect. Immun.,* **2018**, *86*, e00146-18.

140. Noelken, M. E.; Chang, P. J.; Kimmel, J. R., *Biochemistry,* **1980**, *19*, 1838-1843.
141. Chang, P. J.; Noelken, M. E.; Kimmel, J. R., *Biochemistry,* **1980**, *19*, 1844-1849.
142. Zondlo, N. J.; Schepartz, A., *J. Am. Chem. Soc.,* **1999**, *121*, 6938-6939.
143. Smith, B. A.; Daniels, D. S.; Coplin, A. E.; Jordan, G. E.; McGregor, L. M.; Schepartz, A., *J. Am. Chem. Soc.,* **2008**, *130*, 2948-2949.
144. Nicoll, A. J.; Allemann, R. K., *Org. Biomol. Chem.,* **2004**, *2*, 2175-2180.
145. Baker, E. G.; Williams, C.; Hudson, K. L.; Bartlett, G. J.; Heal, J. W.; Porter Goff, K. L.; Sessions, R. B.; Crump, M. P.; Woolfson, D. N., *Nat. Chem. Biol.,* **2017**, *13*, 764-770.
146. Porter Goff, K. L.; Nicol, D.; Williams, C.; Crump, M. P.; Zieleniewski, F.; Samphire, J. L.; Baker, E. G.; Woolfson, D. N., *Biochemistry,* **2019**, *58*, 3060-3064.
147. Myers, J. K.; Pace, C. N.; Scholtz, J. M., *Biochemistry,* **1997**, *36*, 10923-10929.
148. Scholtz, J. M.; Qian, H.; York, E. J.; Stewart, J. M.; Baldwin, R. L., *Biopolymers,* **1991**, *31*, 1463-1470.
149. Schuck, P., *Biophys. J.,* **2000**, *78*, 1606-1619.
150. Wang, G.; Dunbrack, R. L., Jr., *Bioinformatics,* **2003**, *19*, 1589-1591.
151. Wang, G.; Dunbrack, R. L., Jr., *Nucleic Acids Res.,* **2005**, *33*, W94-98.
152. Chen, V. B.; Arendall, W. B., 3rd; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., *Acta Crystallogr. D Biol. Crystallogr.,* **2010**, *66*, 12-21.
153. Kabsch, W.; Sander, C., *Biopolymers,* **1983**, *22*, 2577-2637.
154. Golemi-Kotra, D.; Mahaffy, R.; Footer, M. J.; Holtzman, J. H.; Pollard, T. D.; Theriot, J. A.; Schepartz, A., *J. Am. Chem. Soc.,* **2004**, *126*, 4-5.
155. Porter Goff, K. L. Design, synthesis and characterisation of new protein folds. PhD thesis, University of Bristol, Bristol, 2019.
156. Shoulders, M. D.; Raines, R. T., *Annu. Rev. Biochem.,* **2009**, *78*, 929-958.
157. Woolfson, D. N., Coiled-Coil Design: Updated and Upgraded. In *Fibrous Proteins: Structures and Mechanisms*, Springer International Publishing: 2017.
158. Edgell, C. L.; Savery, N. J.; Woolfson, D. N., *Biochemistry,* **2020**, *59*, 1087-1092.

159. Hodges, A. M.; Schepartz, A., *J. Am. Chem. Soc.,* **2007**, *129*, 11024-11025.
160. Kumar, S.; Bansal, M., *Proteins,* **1998**, *31*, 460-476.
161. Li, Z.; Hirst, J. D., *Chem. Sci.,* **2017**, *8*, 4318-4333.
162. Testa, O. D.; Moutevelis, E.; Woolfson, D. N., *Nucleic Acids Res.,* **2009**, *37*, D315-D322.
163. Ali, M. H.; Imperiali, B., *Bioorg. Med. Chem.,* **2005**, *13*, 5013-5020.
164. Lupas, A. N.; Bassler, J.; Dunin-Horkawicz, S., *Subcell. Biochem.,* **2017**, *82*, 95-129.
165. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr., *Proteins,* **2009**, *77*, 778-795.
166. McIntosh-Smith, S.; Wilson, T.; Ibarra, A. Á.; Crisp, J.; Sessions, R. B., *Comput. J.,* **2012**, *55*, 192-205.
167. McIntosh-Smith, S.; Price, J.; Sessions, R. B.; Ibarra, A. Á., *Int. J. High Perform. Comput. Appl.,* **2015**, *29*, 119-134.
168. Go, N., Scheraga, H. A., *Macromolecules,* **1970**, *3*, 178-187.
169. DePristo, M. A.; de Bakker, P. I.; Lovell, S. C.; Blundell, T. L., *Proteins,* **2003**, *51*, 41-55.
170. Canutescu, A. A.; Dunbrack, R. L., Jr., *Protein Sci.,* **2003**, *12*, 963-972.
171. Boomsma, W.; Hamelryck, T., *BMC Bioinformatics,* **2005**, *6*, 159.
172. Coutsias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A., *J. Comput. Chem.,* **2004**, *25*, 510-528.
173. van Vlijmen, H. W.; Karplus, M., *J. Mol. Biol.,* **1997**, *267*, 975-1001.
174. Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G., *Nucleic Acids Res.,* **2015**, *43*, D364-D368.
175. King, S. M.; Johnson, W. C., *Proteins,* **1999**, *35*, 313-320.
176. Srinivasan, R.; Rose, G. D., *Proc. Natl. Acad. Sci. U.S.A.,* **1999**, *96*, 14258-14263.
177. Cubellis, M. V.; Cailliez, F.; Lovell, S. C., *BMC Bioinformatics,* **2005**, *6*, S8.
178. Mansiaux, Y.; Joseph, A. P.; Gelly, J. C.; de Brevern, A. G., *PLoS One,* **2011**, *6*, e18401.
179. Hintze, B. J.; Lewis, S. M.; Richardson, J. S.; Richardson, D. C., *Proteins,* **2016**, *84*, 1177-1189.
180. Brown, A. M.; Zondlo, N. J., *Biochemistry,* **2012**, *51*, 5041-5051.
181. Doerfel, L. K.; Wohlgemuth, I.; Kothe, C.; Peske, F.; Urlaub, H.; Rodnina, M. V., *Science,* **2013**, *339*, 85-88.
182. Chen, X.; Zaro, J. L.; Shen, W. C., *Adv. Drug Deliv. Rev.,* **2013**, *65*, 1357-1369.
183. Yu, Y.; Lutz, S., *Trends Biotechnol.,* **2011**, *29*, 18-25.
184. Wagner, S.; Klepsch, M. M.; Schlegel, S.; Appel, A.; Draheim, R.; Tarry, M.; Hogbom, M.; van Wijk, K. J.; Slotboom, D. J.; Persson,

J. O.; de Gier, J. W., *Proc. Natl. Acad. Sci. U.S.A.,* **2008**, *105*, 14371-14376.

185. Schlegel, S.; Lofblom, J.; Lee, C.; Hjelm, A.; Klepsch, M.; Strous, M.; Drew, D.; Slotboom, D. J.; de Gier, J. W., *J. Mol. Biol.,* **2012**, *423*, 648-659.