MONASH University

The investigation of consensus design for protein stability, function and evolvability

Benjamin Thomas Porebski
B.S.c Hons.

A thesis submitted for the degree of Doctor of Philosophy at
Monash University in 2016

Department of Biochemistry and Molecular Biology
School of Biomedical Sciences

Copyright notice

Table of contents

Chapter 1 - Introduction **1**
1.1 Overview **3**
1.2 Protein folding and stability **4**
1.3 Engineering protein stability **6**
1.4 Thermodynamic stability versus kinetic stability **10**
1.5 Directed evolution **11**
1.5.1 Overview 11
1.5.2 Generation of diversity 12
1.5.3 Transformation of the library 13
1.5.4 Screening and selection 14
1.5.5 Screening and selection for binding 14
1.5.6 Screening and selection for stability 15
1.5.7 Screening and selection for activity 15
1.6 Semi-rational and sequence based design **16**
1.6.1 Ancestral protein reconstruction 17
1.6.2 Consensus design 17
1.7 Thesis aims **19**

Chapter 2 - Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain **21**
Summary **21**
Abstract **23**
Introduction **23**
Results **24**
Discussion **32**
SI (methods, figures, data) **35**

Chapter 3 – Circumventing the stability-function trade off in the FN3 domain **47**
Summary **47**
Abstract **49**
Introduction **50**
Results and discussion **52**
Conclusions **61**
Methods **61**
SI (figures, tables, data) **67**

Chapter 4 – Exploring the evolvability of FN3con with yeast surface display **71**
Summary **71**
Introduction **73**
Results and Discussion **76**
Conclusions **84**
Methods **84**

Chapter 5 – Smoothing a rugged protein folding landscape by sequence-based redesign 89
Summary 89
Abstract 91
Introduction 92
Results and discussion 95
Conclusions 115
SI (methods, tables, figures, data) 118

Chapter 6 – Discussion and conclusions 137
Summary 137
5.1 Overview of consensus design 138
6.1 Factors to consider during consensus design 139
6.1.1 Acquisition of homologous sequences 139
6.1.2 Homology 141
6.1.3 Bias 142
6.1.4 Sequence count 143
6.1.5 Quality of the sequence alignment 144
6.1.6 The fundamental limitations of protein folds 145
6.1.7 Approaches to the implementation of consensus design 146
6.1.8 Statistical enhancements to consensus design 147
6.2 Biophysical effects of consensus design 149
6.2.1 Thermodynamic stability 149
6.2.2 Protein evolvability 149
6.2.3 Protein folding and kinetic stability 152
6.2.4 Function 154
6.2.5 Immunogenicity 155
6.3 Concluding remarks 156
Table 6.1 159

References 167

Abstract

At the molecular level, protein molecules embody a remarkable relationship between structure and function. They are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. Most proteins are only marginally stable under physiological conditions, with an overall thermodynamic stability, or Gibbs free energy of folding (ΔG), in the range of -5 to -15 kcal mol^{-1}. This marginal stability complicates the design and application of industrial enzymes and therapeutic drugs, whilst also leaving wild-type proteins susceptible to pathologically destabilizing mutations. There are currently several approaches employed to enhance protein stability. The rational approach to stabilization is challenging, as it is difficult to predict the energetic and structural response to mutations in proteins, whilst *in vitro* evolutionary approaches are often expensive and time consuming. An alternative approach is to utilize statistical sequence analysis of an entire protein fold, motif or domain of interest. This is based on the hypothesis that at a given position in a multiple sequence alignment (MSA) of homologous proteins, the respective consensus amino acid contributes more than average to the stability of the protein than non-conserved amino acids. Conservation can be applied as an engineering approach, called consensus design. Here, either point mutations are made to a target protein, or a de novo sequence is calculated, which is known as “full sequence design”. Consensus design has produced many successful examples of stabilization, however little is understood about how and why the method works, nor the cause and effect of design variables.

This thesis explores the application of full sequence consensus design to two protein folds, the fibronectin type III (FN3) domain and the serine protease inhibitor (serpin). A thorough biophysical characterization of the two resulting proteins, FN3con and conserpin, reveals remarkable thermodynamic and kinetic stabilities, with melting temperatures above 100°C, reversible folding and improved aggregation resistance. These results are exceptional achievements of protein engineering, with both FN3con and conserpin being the most stable variants (engineered or wild-type) of their respective protein family, and conserpin being the first serpin with true refoldability. In turn, this has allowed for the direct application of FN3con as a binding scaffold through rational loop grafting and directed evolution, while maintaining its biophysical properties. Further, conserpin provides key insights into evolution, function and stability of the serpin superfamily, and a long sought-after model system for the elucidation of the serpin folding pathway. These results advance our understanding of consensus design, suggesting the capacity of full sequence design to smoothen out the protein energy landscape. This thesis therefore highlights the utility of the technique for engineering highly stable and robust proteins that may serve as model protein systems for future biophysical studies or the basis of industrial enzymes and therapeutic drugs.

Publications during enrolment

1. **Porebski, B.T.**, Keleher, S., Hollins, J.J., Nickson, A.A., *et al.* (2016). Smoothing a rugged protein folding landscape by sequence-based redesign. *Sci. Rep.* Accepted.
2. **Porebski, B.T.**, and Buckle, A.M. (2016). Consensus Protein Design. *Prot. Eng. Des. Sel.*, 29 245–251.
3. **Porebski, B.T.**, Conroy, P.J., Drinkwater, N., Schofield, P., Vazquez-Lombardi, R., Hunter, M.R., Hoke, D.E., Christ, D., McGowan, S., Buckle, A.M. (2016). Circumventing the stability-function trade-off in an engineered FN3 domain. *Prot. Eng. Des. Sel.*, *10.1093/protein/gzw046*
4. Campbell, E., Kaltenbach, M., Correy, G., Carr, P.D., **Porebski, B.T.**, *et al.* (2016). The role of protein flexibility in the laboratory evolution of new enzyme function. *Nat. Chem. Biol.* doi: 10.1038/nchembio.2175.
5. Li, C., **Porebski, B.T.**, McCoey, J., Webb, G.I., Buckle, M., *et al.* (2016) Structural Capacitance in Protein Evolution and Human Diseases. *Sci. Rep*. Accepted.
6. Riley, B., Ilyichova, O., Costa, M.G.S., **Porebski, B.T.**, *et al.* (2016) Direct and indirect mechanisms of KLK4 inhibition revealed by structure and dynamics. *Sci. Rep*. Accepted.
7. Le, S.N., **Porebski, B.T.**, McCoey, J., Fodor, J., Riley, B., Godlewska, M., *et al.* (2015). Modelling of Thyroid Peroxidase Reveals Insights into Its Enzyme Function and Autoantigenicity. *PLoS ONE 10(12): e0142615. doi:10.1371/journal.pone.0142615*
8. Li, C., Chang, C.C.H., Nagel, J., **Porebski, B.T.**, Hayashida, M., Akutsu, T., Song, J and Buckle, A.M. (2015). Critical evaluation of *in silico* methods for prediction of coiled-coil domains in proteins. *Briefings in Bioinformatics*, *doi: 10.1093/bib/bbv047*
9. Allen, M.D., Christie, M., Jones, P., **Porebski, B.T.**, Roome, B., Freund, S.M.V., Buckle, A.M., Bycroft, M., and Christ, D. (2015). Solution structure of a soluble fragment derived from a membrane protein by shotgun proteolysis. *Prot. Eng. Des. Sel.*, *doi:10.1093/protein/gzv021*
10. **Porebski, B.T.**, Nickson, A.A., Hoke, D.E., Hunter, M.R., Zhu, L., McGowan, S., Webb, G.I., and Buckle, A.M. (2015). Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. *Protein Eng. Des. Sel., 28, 67–78*.
11. Kass, I., Hoke, D.E., Costa, M.G.S., Reboul, C.F., **Porebski, B.T.**, Cowieson, N.P., Leh, H., Pennacchietti, E., McCoey, J., Kleifeld, O., et al. (2014). Cofactor-dependent conformational heterogeneity of GAD65 and its role in autoimmunity and neurotransmitter homeostasis. *Proc. Natl. Acad. Sci. U.S.A. 111*, E2524–E2529.
12. Godlewska, M., Góra, M., Buckle, A.M., **Porebski, B.T.**, Kemp, E.H., Sutton, B.J., Czarnocka, B., and Banga, J.P. (2014). A redundant role of human thyroid peroxidase propeptide for cellular, enzymatic, and immunological activity. Thyroid *24*, 371–382.

13. **Porebski, B. T.**, Ho, B., & Buckle, A. M. (2013) Interactive visualization solutions for the structural biologist. *Journal of Applied Crystallography*. 46 (5), 1518-1520.

Patents during enrolment

1. **Porebski, B.T.** & Buckle, A.M. (2014). Highly Stable Polypeptide Scaffolds. *AusPatent 2014905069, PCT/AU2015/050795*

Thesis including published works General Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 1 original paper published in a peer reviewed journal and 2 papers submitted for publication. The core theme of the thesis is “exploring the use of sequence homology for protein engineering”. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the department of biochemistry and molecular biology under the supervision of Associate Professor Ashley Buckle.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of chapters 2, 3 and 5, my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status	Nature and extent (%) of students contribution
2	Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain.	Published	Study design, majority of all experimental work, data analysis and writing of the manuscript, 95%
3	Circumventing the stability-function trade-off in an engineered FN3 domain.	Accepted	Study design, majority of all experimental work, data analysis and writing of the manuscript, 95%
5	Smoothing a rugged protein folding landscape by sequence-based redesign.	Accepted	Determination of crystal structures, computational experiments, data analysis and writing of the manuscript, 70%

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: [redacted] **Date:** 25.8.16

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student and co-authors’ contributions to this work.

Main Supervisor signature: [redacted] **Date:** 25.8.16

Declaration for Thesis Chapter 2

Declaration by candidate

In the case of Chapter 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I contributed to the design of the study, performed the protein design, expression and purification. I performed the CD thermal melt experiments, crystallography, molecular dynamics simulations and analysis. I generated figures and wrote the manuscript.	95

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Adrian Nickson	Performed the folding kinetics and equilibrium measurement experiments	
David Hoke	Assisted with crystallography and writing of the manuscript	
Morag Hunter	Provided contributions to the development of figures and the manuscript	
Liguang Zhu	Assisted in study design	1
Sheena McGowan	Assisted with crystallography and structure determination	
Geoffrey Webb	Study design	
Ashley Buckle	Study design and writing of the manuscript	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work.

Candidate's Signature	[redacted]	**Date** 25.8.16
Main Supervisor's Signature	[redacted]	**Date** 25.8.16

Declaration for Thesis Chapter 3

Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
I designed the study, performed all protein design, protein expression, purification, biophysical characterisation, crystallography, data collection, structural characterisation, biacore experiments and wrote the manuscript.	95

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Paul Conroy	Provided guidance with the biacore experiments	
Nyssa Drinkwater	Assisted with crystallography and data collection	
Peter Schofield	Performed Blitz affinity characterisation	
Morag Hunter	Assisted with protein expression and purification	
Rodrigo Vazquez-Lombardi	Assisted with data analysis and writing of the manuscript	
David Hoke	Assisted with data analysis and writing of the manuscript	
Daniel Christ	Assisted with data analysis and writing of the manuscript	
Sheena McGowan	Assisted with crystallography and data collection	
Ashley Buckle	Design of the study, data analysis, structural determination and writing of the manuscript	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work.

Candidate's Signature	[redacted]	**Date** 25.8.16
Main Supervisor' Signature	[redacted]	**Date** 25.8.16

X

Declaration for Thesis Chapter 5

Declaration by candidate

In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Design of the study, protein expression, purification, CD melts, crystallography, structural determination, molecular dynamics simulations, data analysis and writing of the manuscript.	70

The following co-authors contributed to the work.

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Shani Keleher	Protein expression, purification, biophysical experiments and crystallography	25
Jeffrey Hollins	Folding kinetics and equilibrium measurement experiments	
Adrian Nickson	Folding kinetics and equilibrium measurement experiments. Writing of the manuscript	
Emilia Marijanovic	Biophysical experiments	2
Natalie Borg	Crystallography	
Mauricio Costa	Frustration analysis	
Mary Pearce	Protein expression and purification	
Weiwen Dai	Protein expression and purification	
Liguang Zhu	Protein design	1
James Irving	Protein design	
David Hoke	Analysis of biophysical data	
Itamar Kass	Analysis of molecular dynamics simulations	
James Whisstock	Protein design	
Stephen Bottomley	Design of the study and protein design	
Geoffrey Webb	Protein design	
Sheena McGowan	Crystallography and writing of the manuscript	
Ashley Buckle	Study design and writing of the manuscript	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work.

Candidate's Signature	[redacted]	**Date** 25.8.16
Main Supervisor's Signature	[redacted]	**Date** 25.8.16

Acknowledgements

I would first like to thank my main supervisor Ashley Buckle. His focus, enthusiasm and mentorship throughout the past five years I have been in his group have undoubtedly shaped my personal and professional development for the better. He has provided me with a unique environment that enabled me to follow my own curiosities, whilst providing a rich degree of guidance. He has also shown me the importance of collaboration and communication with other scientists, for which I am incredibly thankful.

I would also like to thank my co-supervisor Geoff Webb for his enthusiasm, guidance and availability throughout my candidature. Discussions with Geoff were highly motivating and promoted the perception of biology from alternate perspectives. Together, my supervisors have enabled new scientific opportunities and fostered my capacity for independent and balanced thinking.

I would next like to thank David Hoke, all the current and past members of the Buckle lab, everyone on level 2 and all of my collaborators. Through extensive discussions and collaborative research, you have all made the past five years an enjoyable scientific and social experience.

Finally, I would like to thank my partner Morag, and all of my family for their continuous support and assistance.

Chapter 1 – Introduction

1.1 Overview

Proteins embody a remarkable relationship between structure and function at the molecular level. They are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. They function as catalysts; they transport and store other molecules; they transcribe; translate and replicate DNA; they provide physical support and immune protection; they generate movement; they transmit and amplify signals; and they control cell growth and differentiation. To do this, proteins are built as a linear polymer of amino acids that fold into complex three-dimensional structures dependent on their specific sequence. The use of 20 amino acids provides a wide range of functional groups including alcohols, thiols, thioethers, carboxylic acids, carboxamides and a variety of basic groups. The specific sequence and subsequent three-dimensional structure of a protein accounts for the broad spectrum of protein function.

Most proteins are only marginally stable under physiological conditions, with an overall thermodynamic stability, or Gibbs free energy between the folded and unfolded states (ΔG), in the range of -5 to -15 kcal mol^{-1} [1-3]. This marginal stability complicates the design and application of industrial enzymes and therapeutic drugs, whilst also leaving wild-type proteins susceptible to pathologically destabilizing mutations. To put this in perspective, the energy of a single hydrogen bond contributes approximately -1.3 kcal mol^{-1} [3,4], thus it becomes apparent how a single amino acid substitution is capable of tipping the balance and disrupting the native state structure. Our ability to predict the thermodynamic consequences of even single amino acid substitutions is still surprisingly limited, and established computational methods of predicting stability are inaccurate and slow, let alone impractical for sampling the vast sequence space available [5]. Protein engineering is the process of creating a protein with desired properties. The discipline has come of age since proteins were first manipulated using site directed mutagenesis by Fersht and Winter in 1985 [6]. However, despite the advances in rational, computational design and directed evolution since then, significant challenges in the accuracy of rational predictions, sampling size and throughput of functional screens in directed evolution studies still remain.

The need to solve these obstacles is greater than ever. In biomedicine, highly specific therapeutics are required to target receptors or interfere with a particular biological process, and protein engineering can aid in the development of selective therapeutics. For example, adalimumab (trade names Humira or Exemptia) is a monoclonal antibody generated by phage display for the treatment of rheumatoid arthritis was the first blockbuster protein drug and functions by inhibition of the tumor necrosis factor-alpha (TNFα) resulting in an anti-inflammatory effect [7]. Five of the top ten selling drugs in 2014 were protein-based compositions, with many more now on the market and in research and development pipelines, reaching a total cumulative sales value of USD$140 billion in 2013 [8]. In industry, the ability to manufacture enzymes of minimal cost, which are stable, selective and highly active, is essential to market competition.

The engineering of protein stability under practical conditions is one of the most critical properties for downstream applications and is thus one of the most explored and well understood [9]. High thermostability is often accompanied by improved protein expression yields in large-scale production, good performance in unfavourable conditions (pH, temperature, pressure, presence of harsh solvents and salts) and longer shelf [10]. In medicinal proteins, improved thermal stability typically results in high serum retention times (biological half-lives) and may also resist proteolysis [11,12]. The engineering of protein function, such as modified selectivity and activity (catalysis or binding) typically results in a loss of protein stability. As such, thermostable proteins are able to tolerate a larger number of mutations than their mesostable counterparts, and thus are better suited as starting points for protein engineering studies [13-15]. However, thermostable proteins are often not available from the natural biodiversity, thus necessitating the need for the engineering of protein stability.

1.2 Protein folding and stability

Theories of protein folding and stability suggest that the native state is the structure with the lowest free energy, or the conformation that is predominantly occupied (Fig. 1.1A, B) [16]. Protein

stability is determined by a multitude of local and long-range interactions. In order to improve protein conformational stability, one usually aims to either increase the free energy difference between the folded and unfolded state, or decrease the rate of unfolding by increasing the free energy difference between the folded and any potential transition states of unfolding. Thus in turn, protein folding and stability are heavily influenced by polar contacts, which include hydrogen bonding, salt bridges, van der Waals forces, and the hydrophobic effect [4,17-22]. Protein stability comes in two different forms: thermodynamic stability, which is related to the equilibrium between the unfolded, partially unfolded and native states of a protein, and kinetic stability, which is related to a high free-energy barrier or transition state that separates the native state from unfolded states, and ultimately relates to how a protein folds and unfolds (Fig. 1.1A).

In terms of stability at temperature and under harsh chemical conditions, both thermodynamic and kinetic stability properties culminate to a particular and measurable result. A protein in its denatured or unfolded state has a considerable amount of conformational entropy and makes many non-covalent interactions with the surrounding solvent water. As the protein folds, it exchanges those interactions with others that it makes within itself. Hydrophobic side chains become packed together, thus increasing the entropy of surrounding water, which releases water molecules into the bulk solvent; this is the basis of the “hydrophobic effect” which drives protein folding [2,23,24]. Interaction energies tend to be small, and depending on the surrounding environment, a hydrogen bond contributes about -1.3±0.6 kcal mol^{-1} [3,25], a salt bridge is between -0.1 and -5.0 kcal mol^{-1} [26-29], van der Waals interactions are roughly -1 kcal mol^{-1} [3,30], hydrophobic residues contribute roughly -1.3±0.5 kcal mol^{-1} per buried CH_2 group [16,20] and disulfide bonds can contribute up to -4.0 kcal mol^{-1} [16,31]. Total interaction energies for proteins in the native and denatured states tend to be huge, in excess of a few thousand kilocalories per mole. However, proteins are only marginally stable, with free energy differences of -5 to -20 kcal mol^{-1} between the native and denatured states [2].

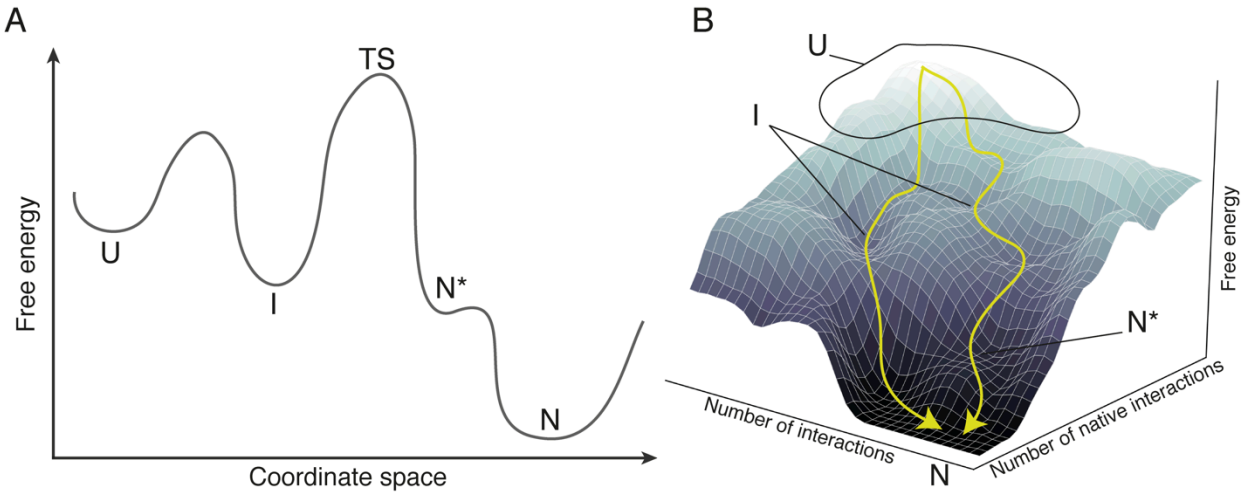


Figure 1.1. Protein folding is shown according to classical thermodynamic and kinetic principles **(A)** and a hypothetical free energy landscape **(B)**. The unfolded state (U) consists of an ensemble of unstructured conformations that readily collapses into the partially folded intermediate state (I) and then folds across a large free energy barrier (transition state – TS) before folding to the native state (N). One or more native-like states (N*), which may be partially unfolded, may be accessible along the possible folding pathway(s) (yellow lines). The free energy landscape for each protein is different, with potential for alternate low energy states that confound folding to the native state, or allow for aggregation to occur.

1.3 Engineering thermostability

A number of well established methods for engineering protein thermostability have been developed and have provided important insight into the structural features that govern stability. These principles arose from mutagenesis studies on small reversible folding proteins and have led to the identification of several general strategies for protein stabilization that are employed by an approach called rational design. For example, rigidifying a protein can be achieved by Gly -> Xaa or Xaa -> Pro mutations (with Xaa being any amino acid); the introduction of disulfides [31-34]; optimization of alpha helices [35-41]; the introduction of salt bridges [42-46]; and the introduction of clustered aromatic-aromatic interactions [17,20,47-49]. Subsequent analysis of proteins isolated from extremophiles and those engineered by directed evolution indicate that stability may be achieved through many alternate routes [50-55]. However, these studies have revealed substantial

trends, such as the significance of surface residues and surface electrostatics [54,56-60], the prevalence of electrostatic networks and higher oligomerization states [61-67].

The literature presents a continuous stream of articles highlighting the importance of specific stabilizing interactions, such as salt bridges or hydrophobic packing. Although optimizing these interactions has worked well in specific examples, a general strategy for improving protein stability has yet to emerge. Therefore, it has become evident that engineering protein stability is a highly context dependent process, typically requiring a detailed structural and functional understanding of the target protein, with no universal approach for success [68]. As such, knowledge-based or rational designs can be tedious, time consuming and expensive, with each mutation and combination of mutations requiring experimental validation. Thankfully, mutations that strongly affect thermostability often appear to cluster in particular regions of the protein, whereas similar substitutions introduced elsewhere may have a negligible effect on stability [69].

An approach for identifying target sites in the presence of crystallographic information is the B-fit method [70], in which residues with the highest crystallographic B-factors are selected for mutagenesis. The rationale is that highly flexible residues are more likely to unfold first and that substitutions in these positions are more likely to stabilize the folded state [70]. Besides B-factor guided mutagenesis, which may be influenced by the non-natural crystal conditions, various computational methods can be used to predict regions as targets for stabilization. Molecular dynamics (MD) simulations, which simulate Newtonian physics of atomic protein coordinates under a given forcefield has contributed greatly to the understanding of how proteins fold; unfold; interact with biological and non-biological molecules; undergo conformational change; and perform an exquisite degree of diverse actions [71,72]. Through the application of MD, it was possible to identify and redesign five flexible residues of a xylanase to improve its melting temperature (T_m) by 4°C, obtain a 30-fold longer half-life at 50°C [73], and improve the stability of a haloalkane dehalogenase by identifying key regions for the introduction a disulfide bond [74].

With increasing computational power and improvements to protein design algorithms, it has become possible to successfully predict stabilizing mutations and reduce the screening effort by orders of magnitude in comparison to random directed evolution approaches. This has typically been achieved in the context of an atomic protein structure and an energy forcefield, which calculates the ΔG of a provided conformation and can compare this with the effect of *in silico* designed mutations. The mutagenesis regime may be conducted either randomly or by knowledge based approaches using design algorithms that aim to improve features known to be important in protein stability or through manual guidance. Mutations then need to be scored, which require some degree of conformational sampling and energy minimization, which may occur locally to just that side chain, expansion to the surrounding region, or full atom MD simulations, since single point mutations are more than capable of eliciting conformational change. As greater accuracy is required, so too are the extent of the calculations and the fewer number of independent mutations that can be assessed.

Energy functions such as FoldX [75,76] and Rosetta [77] are able to assess hundreds, if not thousands of possible mutations in the timeframe of hours to days, whilst all atom MD simulations can take hours to weeks for the assessment of a single mutation. The high throughput nature of FoldX and Rosetta has been utilized extensively to brilliant success, however this still comes down to making a decision between which mutations to experimentally validate, which is not always trivial [78]. Another successful approach has been through the design of favorably charged networks on the protein surface (supercharging), which has shown to increase the T_m of an acylphosphophatase and GTPase by 9°C [79], introduce reversible folding after thermal denaturation to a GFP [80] and increase the heat survival of antibodies [81]; this stabilized antibody was also monomeric in solution and aggregation resistant, unlike the parent molecule. Finally, the application of a design algorithm to identify and optimise hydrophobic packing interactions between buried residues has seen much success when it was used to introduce three

mutations into a xylanase, which increased the T_m by 8°C [82], and improved the T_m of a methionine amino peptidase by 18°C [83].

Although there are many more computational design approaches throughout the literature, each with excellent cases of success, they frequently gloss over their limitations and do not report on the number of failed attempts before success was attained [78,84]. This is exemplified by Murphy et al., [85] who redesigned the core the four helix bundle CheA using four different approaches for backbone flexibility and core repacking. Two of the designs were incredibly successful with a T_m >140°C and a ΔG between 15 and 16 kcal mol^{-1}. However, one design did not express, whilst the other only exhibited wild type stability. Investigation into understanding the differences of stability between designs has unfortunately remained obscured. This and many other examples of design suggest there to not be a single magic bullet that can generally applied [78]. The reason for such difficulty in design is in part due to the highly context dependent nature of protein structure, whereby amino acids are not independent entities, but rather form complex interaction networks with one another, extending beyond 10 amino acids, that subsequently impose physiochemical restrictions on what substitutions can be tolerated at a given position [68,84].

Context dependence may be overcome to some extent by combinatorial knowledge based search approaches, however the addition of subsequent parameters greatly increases the sequence search space that rapidly exceeds what is feasibly assessed with todays computing resources. Therefore it is essential to find the right balance between the size of the problem and speed in which an accurate solution can be determined. Computational design is further challenged by the marginal stability between the native and denatured state and the need for a high degree of forcefield accuracy. Unfortunately, such accuracy is presently lacking, with FoldX and Rosetta having average unsigned errors in ΔΔG determination of 1.28±1.37 kcal mol^{-1} and 1.68±2.32 kcal mol^{-1} respectively for single point mutations [5]; meaning the average difference between experimental and predicted ΔΔG values. Further, prediction of whether a mutation is stabilizing or destabilizing is more challenging, with FoldX and Rosetta having reported accuracies

of 69.5% and 73.4%, respectively [5], whilst a recent comparison amongst 11 predictors reported accuracies of approximately 60% across the board [86]. Although one should not dismiss the utility of computational techniques, caution must be taken in the detailed and specific interpretation of their results.

1.4 Thermodynamic stability versus kinetic stability

So far, the techniques and examples discussed all affect stability of the native state or low energy configurations as determined by structural techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, stabilizing a low energy configuration does not necessarily consider the overall energy landscape and folding pathway of the particular polypeptide. It is well known that mutations which affect the folding pathway can significantly alter the resulting population of low energy configurations and affect rates of folding, resulting in loss of stability and increase in aggregation propensity [87]. This is particularly evident in misfolding and aggregation diseases, where mutations do not necessarily alter the thermodynamic stability of the native state, but rather alter the energy landscape, stabilizing folding intermediates, destabilizing transition states, and/or stabilizing off-pathway low energy states [23,87-89]. As such, misfolding diseases are not a simple case of being more or less thermodynamically stable (although this does occasionally happen), but rather, there is an effect on the kinetics of folding. However, understanding changes to the energy landscape and folding pathways of a protein is particularly difficult due to the present inability to observe intermediate ensembles, alternate low energy configurations and unfolded higher energy configurations at atomic resolution. This problem persists on an experimental level, with structural biology presently unable to provide a general solution, and also on a theoretical level, with computational modeling techniques failing to provide accurate models, thereby limiting the effectiveness of rational design for the engineering of non-native states [90].

1.5 Directed evolution

1.5.1 Overview

One solution to the challenges posed by engineering native and non-native states is to leverage on the highly parallel nature of evolution in a process known as directed evolution (DE). Evolution in its own right is a process in which the selection for specific traits is accomplished by diversification and application of an environmental pressure. In nature, genetic diversity is obtained by time consuming spontaneous mutations during DNA replication or recombination events. In the early 1990's, it became apparent that natural evolution could be harnessed and accelerated in the laboratory. The pioneering work of Stemmer [91] and Chen and Arnold [92] led to the general protocols for DE, which may be summarized in three steps: (section 1.5.2) generate a gene pool of diversity, (section 1.5.3) transform cells with the initial gene pool or selected variant and (sections 1.5.4-7) apply an evolutionary pressure through *in vivo* screening/selection (Fig. 1.2).

This process is repeated in an iterative fashion until the desired outcome is achieved or no further improvements can be made. In contrast to rational design, where an in depth knowledge of structure and function are essential, DE in the broadest sense does not require any such prior knowledge. However, unless one has a highly effective screen or selection system for the desired property, screening libraries can be laborious, costly, and sometimes not practical at all [93]. Further, it is impossible to cover the entire sequence space of a typical protein, where complete NNN codon randomization of a mere 10 amino acid peptide would yield 1.15×10^{18} unique oligonucleotide combinations, which well exceeds the library size of all known generation and selection methods. Thus, DE studies produce the best results with small libraries of high quality that sample a functionally rich portion of the fitness landscape, which requires some understanding of structure and function [94,95].

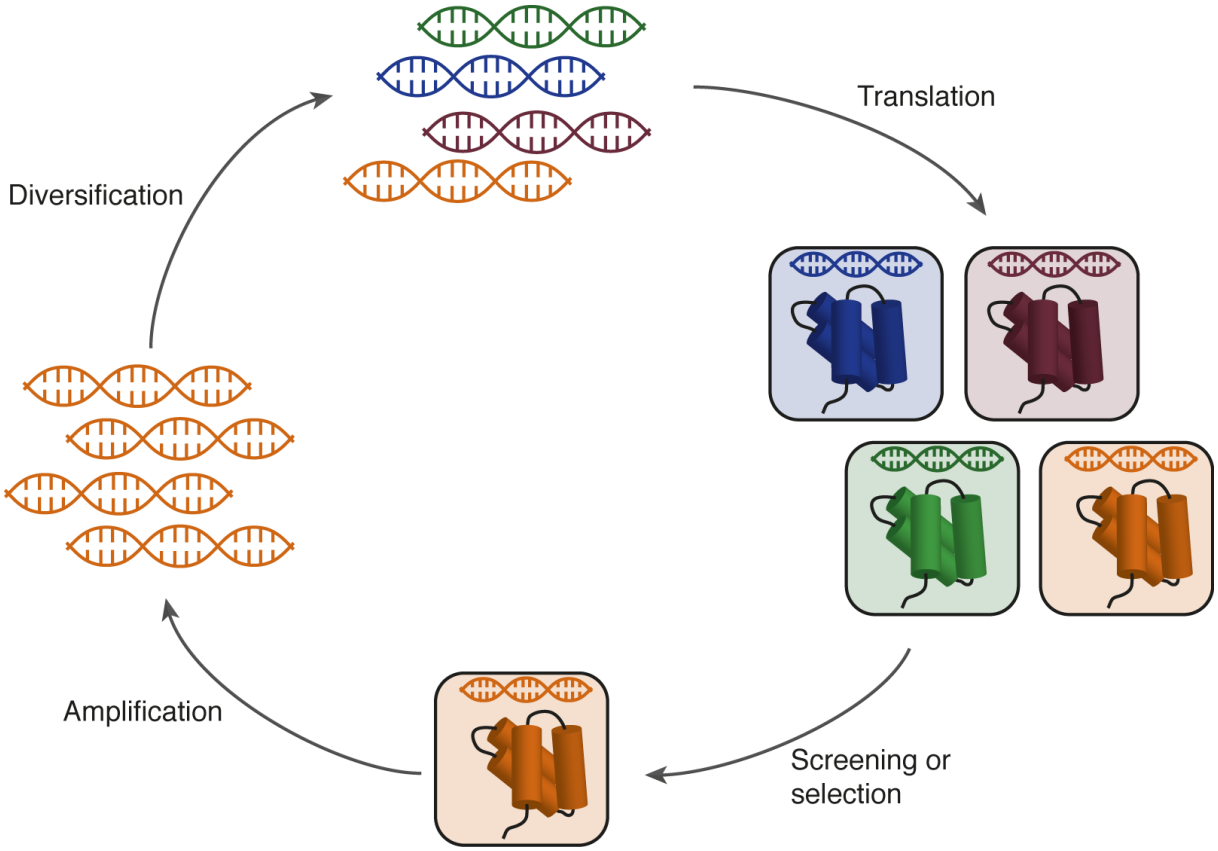


Figure 1.2. Key steps in the cycle of directed evolution. A diverse library of genes are translated into a library of gene products and screened or selected for functional variants using a system that links phenotype and genotype. These isolated genes are replicated and serve as the basis for subsequent diversification and screening.

1.5.2 *Generation of diversity*

There are numerous ways to generate genetic diversity in the lab and new methods are frequently developed (see ref [96] for a recent review). Briefly, three approaches are available. One such approach creates diversity by introducing point mutations (substitutions, insertions and deletions) at random. In a second approach, randomization is restricted to specific positions or regions within the gene – this approach is called focused mutagenesis and typically requires structural and functional understanding about the target protein. A third approach involves the use of recombination techniques applied to gene pools derived from one or both of the first two methods and also from nature. Recombination has been most successful in a method known as gene shuffling, where fragments of gene pool members are switched or recombined, leading to a very efficient sampling of sequence space.

The most common technique for diversity generation of a single gene is error prone polymerase chain reaction (PCR and epPCR) [97,98]. epPCR can be used for random whole gene mutagenesis with controllable error rates (10^{-4}~10^{-3} per replicated base) [97]. epPCR can also be used for focused mutagenesis, however, strong polymerase biases for A->G and T->C mutations exist [98], skewing the resulting population; although this problem is readily corrected by using uneven distributions of dNTPs or proprietary polymerase mixtures [96]. If a finer level of control is necessary, especially in the context of computational library design strategies, DNA oligonucleotide synthesis of degenerate or biased codon populations is an essential and incredibly powerful technique (see [96] for a recent review of methods). These oligonucleotide fragments are typically then assembled into a gene library with traditional restriction enzyme cloning, more contemporary cloning methods, such as leveraging on the homologous recombination machinery in yeast [99,100].

1.5.3 Transformation of the library

With an assembled library, the next challenge is to couple individual phenotypes with their corresponding genotype, as there is no point expressing a pool of diverse proteins if there is no simple way to retrieve the protein sequence or DNA that encodes it. There are effectively two ways to achieve this: (I) the expressed protein is covalently linked to the genetic material (phage display [101], mRNA display [102], ribosome display [103]) or (II) the expressed protein is compartmentalized with the gene either using a cell based system (internal compartmentalization, bacterial [104] and yeast surface display [105]) or an artificial compartment that typically uses an oil droplet emulsion (compartmentalized self replication (CSR) [106], compartmentalized partnered replication (CPR) [107]). Each approach has a scope of evolvable phenotypes that must be taken into consideration for the project at hand, along with their respective advantages and limitations, which is comprehensively discussed in ref [96].

1.5.4 *Screening and selection*

In order to isolate library members with function of interest, there are two general approaches. Screening, by definition, requires the inspection of individual phenotypes. The resulting data can be incredibly rich, providing information about desirable subpopulations and appropriate screening stringency in subsequent rounds of evolution. In comparison, selection bypasses the need to individually inspect each phenotype and instead links an activity of interest to physical separation of the encoding DNA or to the survival of the organism – producing only replicative or active library members. Development of an effective screening protocol can be a major undertaking that requires creativity and molecular intuition. A well-designed selection scheme can offer unparalleled throughput, although this comes at the expense of potentially rich screening data.

1.5.5 Screening and selection for binding

One of the simplest phenotypes to evolve for is binding. Binding studies are typically applicable to covalent and cell surface display technologies and may be used in both screening and selection. In a typical target-binding isolation scheme, library members with their encoding DNA sequencing are captured using an immobilized target, whilst non-binding members are washed away. Selection has been particularly effective with phage display for the development of therapeutic antibodies [7]. In contrast to selection, screening is applicable to bacterial and yeast surface display as they are both capable of using fluorescence activated cell sorting (FACS) for the differentiation of binders. Although FACS is experimentally not as straight forward as selection, it can be markedly more powerful due to the generation of real-time population data that allows for the interrogation of binding affinity (although accuracy limitations may persist). This information rapidly allows the user to adjust the screening stringency for optimal result. However, this advantage comes at a cost of throughput. State-of-the-art flow cytometry offers one of the highest capacities of any screening method, achieving up to 10^8 library members screened in roughly 24 hours [105], but this does not compare to selection studies which are limited by library size and transformation efficiency, which

can range from 10^8 to 10^{13} variants (10^{12} to 10^{13} for mRNA [102] and ribosome display [103] respectively).

1.5.6 Screening and selection for stability

Most strategies for evolving stability revolve around enzyme stability and subsequently rely on being able to measure residual activity after exposure to a denaturing challenge (temperature, chemical denaturant, organic solvent and pH). This is particularly difficult, as it requires prior separation of library members, such as picking colonies, growth and expression in a 96 well plate, inducing the denaturing challenge and assaying each well for residual activity. This process is incredibly laborious and is not high throughput by any means. Linkage of the enzyme of interest to survival of the organism, if possible, resolves this problem. Tamakoshi and colleagues [108] used the thermophile *Thermus thermophilus* directly as their selection system as the entire organism is tolerant to temperature. By engineering a leucine synthesis negative strain of *T. thermophilus,* they were able to evolve the leucine biosynthesis enzyme 3-isopropylmalate dehydrogenase from *Saccharomyces cervisiae* and improve its thermal stability five-fold. Other strategies have utilized phage display, where the target protein library is inserted between two domains of the protein g3p in the filamentous phage *fd* [109]. After denaturant challenge, phage library members that are not correctly folded are unable to propagate and are thus eliminated from the gene pool. However, this method is limited by the stability of the phage itself and general tolerance of the target protein to being inserted into this construct. Finally, as throughput may ultimately be a limiting factor, lessons from rational design of stability may be fruitful in designing smaller high quality libraries that better sample predicted hotspots of stability.

1.5.7 Screening and selection for activity

Evolving for activity is similar to that of stability, where activity is detected by a fluorescent reporter. However, without the need for denaturation, the use of compartmentalization systems is more readily available, removing the need for separating individual library members. In 2010,

Agresti *et al.,* [110] very elegantly used microfluidics to encapsulate individual cells from a yeast surface display library in a droplet of oil, allowing for compartmentalized enzyme activity and subsequent fluorescent droplet sorting. By doing so, the authors were able to evolve horseradish peroxidase (HRP) to exhibit catalytic rates that are 10-fold faster than their parent molecule, with a 1,000 fold increase in experimental speed and 1-million fold decrease in cost. Prior to the development of such techniques, individual library members had to be isolated and assayed in 96 well plates, a task usually performed by graduate students, or robots.

1.6 Semi-rational and sequence based design

Although directed evolution and the emerging wave of informatics-based rational design has fundamentally transformed the field of protein engineering, these methods in isolation are unable meet the current demands from research and industry, with practical and monetary limitations hindering the scale of a project. Rather than addressing this by brute force researchers are opting for the combination of techniques, such as mixing rational design with directed evolution schemes, to create small libraries of very high quality. This is often referred to as semi-rational or knowledge-based design. These approaches combine information from protein structure, function, sequence homology and predictive computational algorithms to preselect sites for focused mutagenesis with limited amino acid diversity. This focus translates into dramatically reduced library sizes with a major increase in functional content allowing for a more efficient sampling of sequence space.

A popular strategy in semi-rational design of stability is the use of evolutionary information encapsulated in homologous protein sequences. Multiple sequence alignments (MSAs) and phylogenetic analyses have become standard tools for exploring sequence conservation [111] and ancestral relationships [112] amongst protein homologues. Such sequences and alignments can be acquired from natural sequence databases [113,114], curated alignment databases [115-118] and neutral drift experiments [119,120], [112,121-123].

1.6.1 Ancestral protein reconstruction

Ancestral protein reconstruction involves the use of statistical phylogenetic analysis of protein homologues to predict or "reconstruct" a likely sequence of an ancestral protein [121]. The approach was originally developed to explore how present day sequences diverged from one another without the sequence of common ancestor [112,121-123]. However, the technique has been observed to consistently produce protein sequences with high thermodynamic stability and was subsequently adopted by protein engineers [124-127]. The exact reason behind improved thermodynamic stability is somewhat controversial, with debate over whether ancestral proteins were really thermphilic as a result of ancient conditions [125], or if improvements are artefacts resulting from biases [128] or correlations identified by the maximum likelihood framework used to infer the ancestral sequences [78,129].

1.6.2 Consensus design

Consensus design, like ancestral sequence reconstruction, utilizes evolutionary history; however, rather than inferring phylogenetic hierarchy, all sequences are aligned and the most frequently observed amino acid is identified at each position in the alignment (Fig. 1.3) [111]. The consensus design approach has been widely successful in improving the stabilities of functional and non-functional proteins, for example increasing melting temperatures by 10-32°C [78,130-137]. However, only ~50% of conserved residues are associated with improved stability, with ~10% being stability neutral, and ~40% being destabilizing, leading to challenges and trade-offs during implementation [111,130,136,138-142].

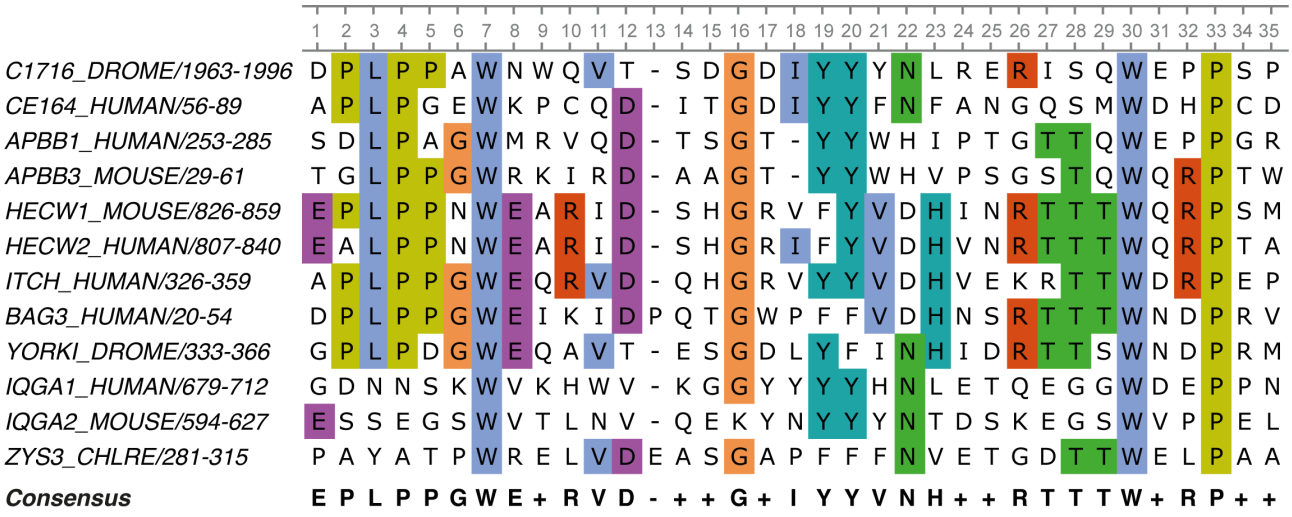


Figure 1.3. Sequence alignment of 12 WW domains across several species and parent proteins. In the consensus, a – is a gap, whilst a + is an ambiguous position that needs manual intervention. The most conserved residues are coloured.

Consensus design involves four steps; (1) identification of domain to be targeted (for example, boundaries within a larger sequence context); (2) Acquisition and pre-processing of homologous sequences; (3) Iterative assessment of several multiple sequence alignment (MSA) regimes and removal of disruptive sequences; (4) Calculation of sequence conservation. Application of sequence conservation is typically performed in one of three ways. First, single or multiple point mutations of the most conserved amino acid positions can be made to a target protein [138,139,141,143,144] and these mutations may further be filtered or weighted by other statistical or computational methods [141,145]. Second, full-length sequences can be created *de novo*, avoiding the problem of identifying residues that are truly stabilizing [134,137,140,146-150]. Third, conserved residues and positions can be spiked or targeted in directed evolution studies to increase sampling of functionally relevant sequence space [120,145,151,152]. The strategy of implementation is highly dependent on requirements and available resources; however, all approaches have seen impressive results, with an exhaustive catalogue of consensus-designed proteins shown in Chapter 6.

The origin of consensus mutant stabilization is currently described as that at a given position in a MSA of homologous proteins, the respective consensus amino acid contributes more than average to the stability of the protein than non-consensus amino acids (Fig. 1.3) [111,130,136,138-142]. That is, a conserved residue is more likely to be stabilizing than a random mutation at that same position [13,141,153]. However, this does not explain why conserved residues are likely to be more stabilising. A possible explanation is that, as proteins evolved from a non-specialised but stable common ancestor, evolutionary drift allowed for the sampling of different stabilising mutations needed for adequate stability. Through the evolution of specialist function, many proteins now exist on a knife-edge of stability and function [13,154-156]; for this reason, stabilising residues tend to be conserved. Consensus design is therefore able to leverage on millions of years of evolution and identify stabilizing features from numerous protein homologues – amalgamating mostly additive mutations that no single protein has needed to amass. Regardless of mechanism, consensus design is a proven technique that allows for the rapid improvement of protein stability in the absence of structure or understanding of target protein function.

1.7 Thesis aims

With its relative ease of implementation, capacity to enhance other engineering methods and impressive variety of successful designs, consensus design is a powerful protein engineering approach that presently lacks a comprehensive understanding of implementation strategies, limitations, observed effects of design and the subsequent relationship to design parameters. This thesis will explore aspects of consensus design strategy and its effects on stability, structure, folding, dynamics, function and evolvability for two protein folds.

In order to explore the above themes, this thesis will discuss the following specific examples:

(a) Explore the effect of a large curated sequence alignment on the consensus design of a fibronectin type III (FN3) domain, in comparison to two previously designed variants and three naturally occurring FN3 domains of variable thermodynamic stability. The resulting design, *FN3con*, is the most stable FN3 domain reported to date, with key determinants of thermodynamic and kinetic stability thoroughly examined by structural and dynamics analysis (Chapter 2).

(b) Explore the mutational tolerance of FN3con as a binding scaffold. Rational loop grafting and iterative design was performed to circumvent the stability-function trade-off between FN3con and FNfn10 (Chapter 3).

(c) Explore the evolvability of FN3con by directed evolution with yeast surface display. Preliminary results indicate that FN3con may be able to tolerate and display a larger sequence space on three surface exposed loops than other alternative FN3 domains (Chapter 4).

(d) Explore the effect of consensus design on the serine protease inhibitor (serpin) family where stability and folding are directly linked to activity. This chapter provides significant insight into the effects of consensus design on complex energy landscapes (Chapter 5).

In summary, this thesis aims to advance the understanding of consensus design on several fronts, by exploring design parameters, resulting biophysical properties and applicability for downstream engineering studies.

Chapter 2

Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain

Summary

In this chapter, I explore the effect of a large curated sequence alignment on the full sequence consensus design of a fibronectin type III (FN3) domain. This resulted in a molecule of immense thermodynamic and kinetic stability, termed FN3con. Crystallographic structures and molecular dynamics simulations provide key insights into the determinants of stability, revealing optimization of the hydrophobic core, the accumulation of tyrosine corner residues and introduction of a surface exposed electrostatic mesh. This study further highlights the capacity for consensus design to not only stabilize the native state, but also destabilize non-native conformations.

This chapter resulted in the following publication and patent:

Porebski, B.T., Nickson, A.A., Hoke, D.E., Hunter, M.R., Zhu, L., McGowan, S., Webb, G.I., and Buckle, A.M. (2015). Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. *Protein Eng. Des. Sel. 28, 67–78*.

Porebski, B.T. & Buckle, A.M. (2014). Highly Stable Polypeptide Scaffolds. *AusPatent 2014905069, PCT/AU2015/050795.*

Protein Engineering, Design & Selection, 2015, vol. 28 no. 3, pp. 67–78
doi: 10.1093/protein/gzv002
Original Article

OXFORD

Original Article

Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain

Benjamin T. Porebski[1], Adrian A. Nickson[2], David E. Hoke[1], Morag R. Hunter[3], Liguang Zhu[4], Sheena McGowan[1], Geoffrey I. Webb[4], and Ashley M. Buckle[1,*]

[1]Department of Biochemistry and Molecular Biology, Faculty of Medicine, School of Biomedical Sciences, Monash University, Clayton, VIC 3800, Australia, [2]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK, [3]Centre for Brain Research and Department of Pharmacology and Clinical Pharmacology, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand, and [4]Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

*To whom correspondence should be addressed. E-mail: ashley.buckle@monash.edu
Edited by Lars Baltzer

Received 3 August 2014; Revised 8 January 2015; Accepted 15 January 2015

Abstract

Consensus protein design is a rapid and reliable technique for the improvement of protein stability, which relies on the use of homologous protein sequences. To enhance the stability of a fibronectin type III (FN3) domain, consensus design was employed using an alignment of 2123 sequences. The resulting FN3 domain, *FN3con*, has unprecedented stability, with a melting temperature >100°C, a ΔG_{D-N} of 15.5 kcal mol^{-1} and a greatly reduced unfolding rate compared with wild-type. To determine the underlying molecular basis for stability, an X-ray crystal structure of FN3con was determined to 2.0 Å and compared with other FN3 domains of varying stabilities. The structure of FN3con reveals significantly increased salt bridge interactions that are cooperatively networked, and a highly optimized hydrophobic core. Molecular dynamics simulations of FN3con and comparison structures show the cooperative power of electrostatic and hydrophobic networks in improving FN3con stability. Taken together, our data reveal that FN3con stability does not result from a single mechanism, but rather the combination of several features and the removal of non-conserved, unfavorable interactions. The large number of sequences employed in this study has most likely enhanced the robustness of the consensus design, which is now possible due to the increased sequence availability in the post-genomic era. These studies increase our knowledge of the molecular mechanisms that govern stability and demonstrate the rising potential for enhancing stability via the consensus method.

Key words: consensus design, fibronectin type III, FN3, molecular dynamics, stability

Introduction

There are currently several approaches employed to enhance protein stability. The rational approach to stabilization is challenging since it is difficult to predict the energetic and structural response to mutation in proteins, due to inaccuracies in predictive energy functions and the current inability to model the unfolded state (Magliery *et al.*, 2011). Much effort has been focused on stabilizing the native, folded state ('positive design' (Dantas *et al.*, 2003; Kuhlman, 2003;

Shah *et al.*, 2007)) and also destabilizing the non-native states ('negative design' (Richardson and Richardson, 2002; Jin *et al.*, 2003)) via rational design and structural comparison of thermophilic proteins with their mesophilic counterparts (Russell *et al.*, 1994; Russell and Taylor, 1995; Auerbach *et al.*, 1997; Davlieva and Shamoo, 2010; Nakamura *et al.*, 2010; Guelorget *et al.*, 2011; Sundaresan *et al.*, 2012). Although much insight has been gained from these studies, both approaches require structures of the target protein and or any thermophilic orthologs, which then needs to be followed up with extensive structural and functional analysis. These challenges are further complicated by the context dependence of stabilizing mutations and tend to be applicable to only a small subset of scaffolds.

An alternative approach is to utilize statistical analysis of the entire protein fold, motif or domain of interest. This is an attractive idea based on the hypothesis that at a given position in a multiple sequence alignment (MSA) of homologous proteins, the respective consensus amino acid contributes more than average to the stability of the protein than non-consensus amino acids (Steipe *et al.*, 1994; Lehmann and Wyss, 2001; Magliery *et al.*, 2011). However, the technique is not always simple to implement. In particular, generation of MSAs is challenging, especially in poorly conserved regions, which leads to a large amount of noise. As most sites across a protein family are not conserved, the most common amino acid tends to be no better than picking a residue at random (Dantas *et al.*, 2003; Kuhlman, 2003; Shah *et al.*, 2007; Magliery *et al.*, 2011). Regardless, the efficacy of consensus design in improving protein stability has been demonstrated numerous times; with examples including antibodies (Steipe *et al.*, 1994; Richardson and Richardson, 2002; Jin *et al.*, 2003), the GroEL minichaperone (Wang *et al.*, 1999), the Abp1p SH3 domain (Maxwell and Davidson, 1998), the p53 DNA-binding domain (Nikolova *et al.*, 1998), fluorescent proteins (Dai *et al.*, 2007), a fungal phytase (Lehmann *et al.*, 2002) and recently the FN3 domain (Jacobs *et al.*, 2012). The availability of a small number of homologs (10–50 sequences) has typically limited the technique to combining a relatively small number of the most conserved residues with rational engineering approaches, as opposed to complete sequence redesign. With the recent advances in high throughput sequencing, the number of available sequences is rapidly growing.

In this study, we investigated whether the availability of a greater number of protein sequences resulting from advances in genomics could enhance the consensus approach. We selected the fibronectin type III (FN3) domain, a small β-sheet sandwich of roughly 90–100 amino acids in length, due to its ubiquitous nature across phyla (Fraser *et al.*, 2006), and its popularity as a model for protein folding and engineering studies (Clarke *et al.*, 1997; Hamill *et al.*, 1998, 2000b; Cota and Clarke, 2000; Koide *et al.*, 2001, 2012; Bloom and Calabro, 2009; Jacobs *et al.*, 2012; Gilbreth *et al.*, 2014). This paper describes the structural and biophysical characterization of *FN3con*—a consensus-derived FN3 domain having increased stability.

Results

We constructed a consensus FN3 domain, which we call *FN3con*, using 2123 aligned FN3 sequences collected from the Prosite domain database (http://prosite.expasy.org/PDOC50853). The FN3 domains in this database are hand curated and sourced from numerous multidomain proteins spanning mostly higher order eukaryotic organisms. The full MSA can be found as a FASTA file in Supplementary Data S1. We generated the new protein sequence using the consensus method, which selects the most frequently observed residue at each column of the sequence alignment. His-tagged FN3con and FNfn8 were expressed as a soluble, monomeric domain in *Escherichia coli*. Purification by nickel affinity chromatography and size exclusion chromatography produced a homogenous, monomeric sample of the expected molecular weight (Supplementary Fig. S1) that was further characterized by biophysical and X-ray crystallographic methods. We subsequently selected a set of well-studied FN3 domains (FNfn10, FNfn8 and TNfn3) and the consensus FN3 domains produced by Jacobs et al. (2012) (Fibcon and Tencon) for comparative analysis (sequences in Supplementary Data S2). All of our comparison domains have extensive biophysical data and X-ray crystal structures available, and measured stabilities ranging from 57 to 90°C.

FN3con is the most stable FN3 domain reported

Thermal stability of FN3con and FNfn8 was measured by circular dichroism (CD) at a wavelength of 222 nm while heating from 20 to 110°C. FN3con gradually loses secondary structure signal until ~100°C, where a sharp unfolding transition starts but does not plateau before the thermal limit of the CD spectrophotometer is reached (110°C; Fig. 1A). We repeated the experiment in the presence of 2 M guanidine hydrochloride (GuHCl), which resulted in a complete unfolding transition and melting temperature (T_m) of 90.7°C (Fig. 1B). Furthermore, we found FN3con to be reversibly foldable (Supplementary Fig. S2), a common trait of the FN3 domain (Erickson, 1994), and for comparison, we measured the T_m of FNfn8 to be 58.0°C (Fig. 1B).

The unfolding and refolding equilibrium curves of FN3con show excellent agreement with one another, further indicating that the folding is reversible (Fig. 1C). The global fit to both datasets gives a denaturant activity midpoint, $[D']_{50}$, of 1.75 ± 0.01 M, an equilibrium *m*-value, m_{D-N}, of 8.80 ± 0.21 kcal mol^{-1} M^{-1} and hence a protein stability, ΔG_{D-N}, of 15.5 ± 0.4 kcal mol^{-1}. Note that these errors are those of the fit, and not the true errors of experimental replication. The *m*-value for FN3con (8.80 kcal mol^{-1} M^{-1}) is in the range expected for homologous FN3 domains (6.38 and 9.42 kcal mol^{-1} M^{-1} for FNfn10 and TNfn3 in guanidine isothiocyanate, respectively) (Cota and Clarke, 2000). However, it is clear that FN3con is far more stable than FNfn10 and TNfn3 (15.5 compared with 9.38 and 6.68 kcal mol^{-1}). The kinetic chevron (Fig. 1D) can be fitted extremely well using a modified equation to take into account both a refolding intermediate and a high-energy intermediate (see Supplementary Methods). The $[D']_{50}$ from the chevron is 1.80 ± 0.05 M, which is identical to the value obtained from the equilibrium studies and is strong evidence that both experiments are measuring the global unfolding of the protein domain and not a local effect. The fit gives a kinetic *m*-value of 10.2 ± 0.9 kcal mol^{-1} M^{-1}, and a stability in buffer of 18.6 ± 1.6 kcal mol^{-1} M^{-1}. Again, the errors are of the fit, and not the true errors of experimental replication. Taken together, the equilibrium and kinetic folding data indicate that, while FN3con is similar in structure to natural FN3 domains (based on the *m*-values) it is at least twice as stable. This increase in stability is predominantly due to a much slower unfolding rate, although the domain also has a slightly faster folding rate (see Table I).

FN3con structure reveals optimization of surface electrostatics and hydrophobic packing

In order to understand the structural basis for stability in FN3con, we determined its X-ray crystal structure to 2.0 Å resolution. Data processing and structure refinement statistics are shown in Table II. FN3con adopts the FN3 fold, consisting of seven anti-parallel

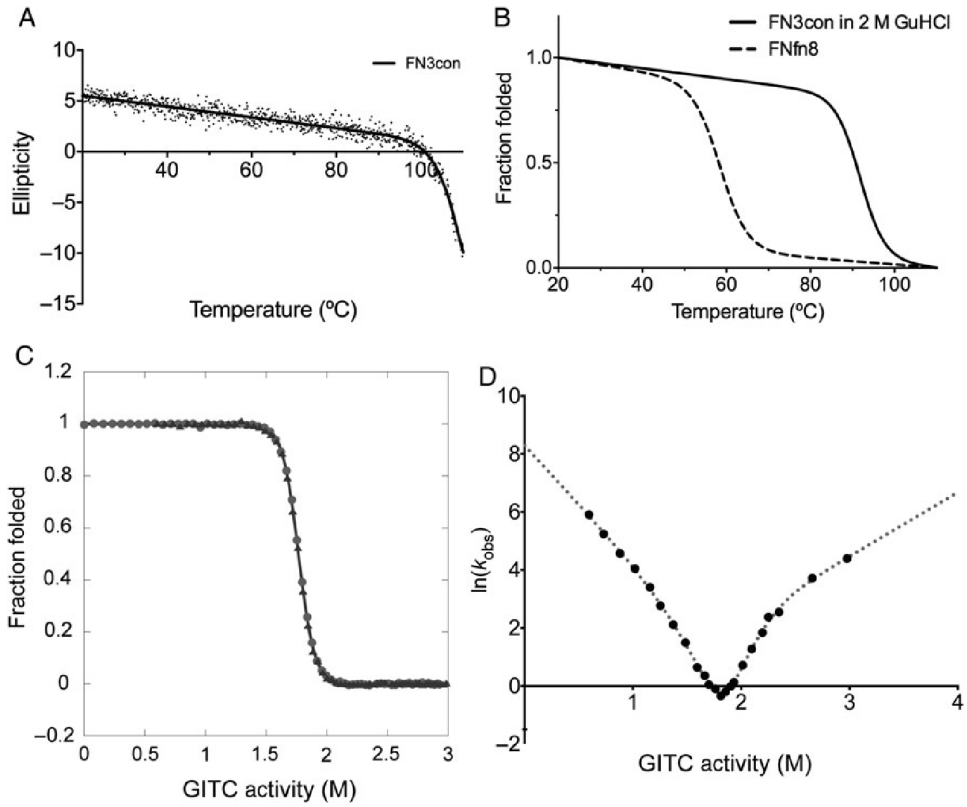


Fig. 1 Thermal stability, chemical stability and folding kinetics of FN3con. (**A**) Thermal unfolding monitored by CD at 222 nm with non-linear fit ($R^2 = 0.95$). (**B**) Thermal unfolding in 2 M GuHCl (solid line) and FNfn8 (dashed line) is represented as fraction folded with a non-linear fit ($R^2 = 0.98$ and $R^2 = 0.76$, respectively). (**C**) Equilibrium unfolding (circles) and refolding (triangles) curves. (D) Kinetic folding data showing curvature in both arms of the chevron plot.

Table I. Summary of stability, equilibrium and kinetic measurements for FN3 domains

Protein	T_m (°C)	ΔG (kcal mol$^{-1)}$	Folding rate (s^{-1})	Unfolding rate (s^{-1})	Source
FN3con	>100.0	15.5 ± 0.4	4020	3.09×10^{-8}	
Fibcon	89.6	11.4 ± 1.5	N/A	N/A	Jacobs *et al.* (2012)
FNfn10	82.5	9.38 ± 0.13	240	2.3×10^{-4}	Cota and Clarke (2000)
Tencon	78.0	10.6 ± 0.9	N/A	N/A	Jacobs *et al.* (2012)
FNfn8	58.0	N/A	N/A	N/A	
TNfn3	57.1	6.68 ± 0.18	6.2	4.8×10^{-4}	Hamill *et al.* (1998)

β-strands connected by surface exposed loops (Fig. 2). A structural alignment with our comparison domains shows very high similarity, with an average root mean square deviation (RMSD) of 1.2 Å across backbone Cα atoms in all structures (Fig. 2A).

To investigate the structural basis for increased stability in FN3con, we first calculated several physicochemical and structural parameters that are known to affect protein stability and folding, for the set of comparison domains (Table III). Analysis reveals FN3con to have the highest number of H-bonds (46) and salt bridges (48), with the smallest accessible surface area (ASA). Comparatively, the number of H-bonds is relatively equal across the assessed domains, with a mean count of 43.5. Salt bridge counts are highly varied across the comparison set. Although FN3con has the highest number of salt bridges (48), consistent with its high stability, TNfn3 (lowest stability) has the second highest count with 41 salt bridges. However, comparisons with the ratio of acidic:basic residues show large differences between FN3con and TNfn3. Specifically, FN3con has 48 salt bridges being formed from 10 positive and 7 negatively charged residues, while TNfn3 has 41 salt bridges being formed by 18 positive and 9 negatively charged residues. Interestingly, FN3con harbors a unique and extensive complementary charged electrostatic network that is distributed over β-sheet 2, spanning strands C′, C and F. This network results from the presence of four arginine residues and four glutamic acid residues (R45, R49, R81, R83, and E47, E57, E79, E90), which are not present in any of the other FN3 domains (Fig. 2B). Comparatively, TNfn3 reveals clustering of like-charged residues on the peripheral loops (Fig. 2B).

Calculations of ASA values correlate weakly to thermal stability, with FN3con and Fibcon having the smallest ASA values of 4545.5 and 4882.3 Å^2 and the highest thermal stability; however, this trend does not appear to be linear for the other domains. Similarly, the grand average hydropathicity (GRAVY) scores vary quite dramatically

Table II. Crystallographic data and refinement statistics

	FN3con (4U3H)
Data collection	
Temperature	100 K
X-ray source	Australian Synchrotron MX1
Detector	ADSC Quantum 210R
Wavelength (Å)	0.9537
Space group	$P4_132$
Unit cell axes (Å)	86.1, 86.1, 86.1
Angles (°)	90, 90, 90
Mol./ASU	1
Resolution (Å)[a]	35.15–1.98 (2.05–1.98)
Total reflections[a]	80 030 (95 144)
Unique reflections[a]	8078 (804)
Completeness (%)[a]	100.0 (100.0)
Multiplicity[a]	35.3 (33.5)
R^a_{pim}	0.018 (0.138)
$\langle I/\sigma I\rangle$[a]	28.5 (5.33)
CC1/2[a]	1.0 (0.935)
Structure refinement	
Resolution (Å)	35.15–1.98
Number of non-hydrogen atoms	801
Number of solvent molecules	61
R_{work} (%)	0.1970
R_{free} (%)[b]	0.2432
RMSD bond lengths (Å)	0.013
RMSD bond angles (°)	1.37
Ramachandran plot	
% favored (% outliers)	100.00 (0.0)
Clash score	0.7
MolProbity score	0.73 100th percentile[c] (n = 12 332, 1.980 ± 0.25 Å)

[a]Values for highest resolution shell are in parentheses.
[b]The free R factor was calculated with 5% of data omitted from refinement.
[c]100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.

across the set of comparison domains and are not related to thermal stability (Supplementary Fig. S3).

While salt bridge interactions are thought to make a relatively minor contribution to stability (Horovitz *et al.*, 1990; Serrano *et al.*, 1990), the presence of unfavorable clusters with like-charged residues is known to be destabilizing and may offer clues to the differences in stabilities of the assessed FN3 domains (Horovitz *et al.*, 1990; Loladze *et al.*, 1999; Koide *et al.*, 2001; Sanchez-Ruiz and Makhatadze, 2001). Indeed, such like-charged clusters are present in the metastable FN3 domains (FNfn10, Tencon, FNfn8 and TNfn3) but absent in the highly stable FN3con and Fibcon (Figs 2B and 3). This is clearly seen in FNfn10, which features both negatively (D7, E9 and D23) and positively (R30, R78 and D80) charged clusters (Fig. 3A). The destabilizing effect of the first cluster has been validated by mutagenesis, where mutation of D7 to asparagine or lysine increased thermal stability by ~10°C at pH 7.0 (Koide *et al.*, 2001). Similarly, potential destabilizing clusters are also present in Tencon (E67 and E87) (Fig. 3B), FNfn8 (D26 and D52) (Fig. 3C) and TNfn3 (E33 and D49; E28, D30 and D78; E9 and E8; D15 and D65; D40 and E67) (Fig. 3D). Unsurprisingly, there is a strong similarity in the distribution of charged residues among Tencon and TNfn3. However, Tencon appears to have reduced the presence of like-charged residue clusters, resulting in increased coordination of complementary charged residues (Fig. 2).

The hydrophobic effect is a major determinant of protein folding and stability (Fersht *et al.*, 1992; Serrano *et al.*, 1992; Buckle *et al.*, 1993; Fersht and Serrano, 1993; Axe *et al.*, 1996; Kellis *et al.*, 1988). We therefore assessed differences in hydrophobic packing among the comparison set of FN3 domains, focusing on a hydrophobic 'banding' pattern that is orthogonal to the direction of the β-strands (Fig. 4) (Lappalainen *et al.*, 2008). Strikingly, the degree of uniformity and alignment among hydrophobic residues in each band appears to be proportional to the stability of the domain. In general, we observe higher stability to be associated with uniform hydrophobic banding as well as greater burial and reduction of bulky hydrophobic residues, which is consistent with the current understanding of the hydrophobic effect and its role in stability (Fig. 4).

As packing density of the hydrophobic core is a known factor in protein stability (Karpusas *et al.*, 1989; Chothia and Finkelstein, 1990; DeDecker *et al.*, 1996; Levitt *et al.*, 1997; Ratnaparkhi and Varadarajan, 2000), we calculated the volumes of solvent inaccessible cavities and the mean occluded surface packing (OSP) value for each FN3 domain, as a measure of packing density (Table IV). The most striking observation from these calculations is the significantly reduced solvent inaccessible cavity volume of FN3con (60.8 $Å^3$) compared with the next most stable domain, Fibcon (171.0 $Å^3$) (Table IV and Supplementary Fig. S4). This value alone indicates superior packing of the hydrophobic core in FN3con and may contribute to its fast folding rate. Interestingly, FNfn8 has a cavity volume of 185.8 $Å^3$, suggesting that while cavity volume may be an indicator of stability, it is by no means absolute. A similar anomaly was also seen for a chimera of FNfn10 and TnFN3, which had a stability that was intermediate between the two proteins, despite having a core that was less well packed than either parent (Billings *et al.*, 2008).

We next investigated the structural context of aromatic residues, which are known to contribute greatly to the stability of immunoglobulin-like domains (Hamill *et al.*, 2000a; Nicaise *et al.*, 2003). All assessed FN3 domains contain the highly conserved tryptophan 22 (W22), while FN3con further contains a unique solvent-exposed tryptophan (W55) on β-sheet 2 (Fig. 4). W55 packs tightly against the side chains of E47, R49, E79 and R81; however, its effect on stability is not immediately apparent. Tyrosine residues are another highly conserved motif among the immunoglobulin fold and are thought to contribute to stability via the concept of a 'tyrosine corner', where the tyrosine residues are positioned near the beginning or end of an anti-parallel β-strand (Hamill *et al.*, 2000a; Nicaise *et al.*, 2003). Comparisons among the selected FN3 domains reveal two highly conserved tyrosine residues, one at the N-terminal end of strand C (Y48 in FN3con, Y36 or Y34 in others) and the other at the C-terminal end of strand F (Y78 in FN3con, Y68 or Y66 in others) (Figs 4 and 5). The relatively stable FN3con, Fibcon and TNfn10 contain a tyrosine residue at the C-terminal end of strand C (Y44 in FN3con and Y32 in Fibcon and FNfn10), potentially providing stabilizing interactions to both loop regions, which is absent in the less stable domains. Interestingly, FN3con, Tencon and TNfn3 share a unique tyrosine residue (Y67 and Y57, respectively) on β-sheet 1, which is absent in Fibcon, FNfn10 and FNfn8 (Figs 4 and 5).

Simulations reveal global and local differences in the dynamics of FN3 domains

Having performed a thermodynamic, kinetic and structural characterization of FN3con we next investigated its dynamic properties. We performed MD simulations of the FN3 domains listed in Table I in

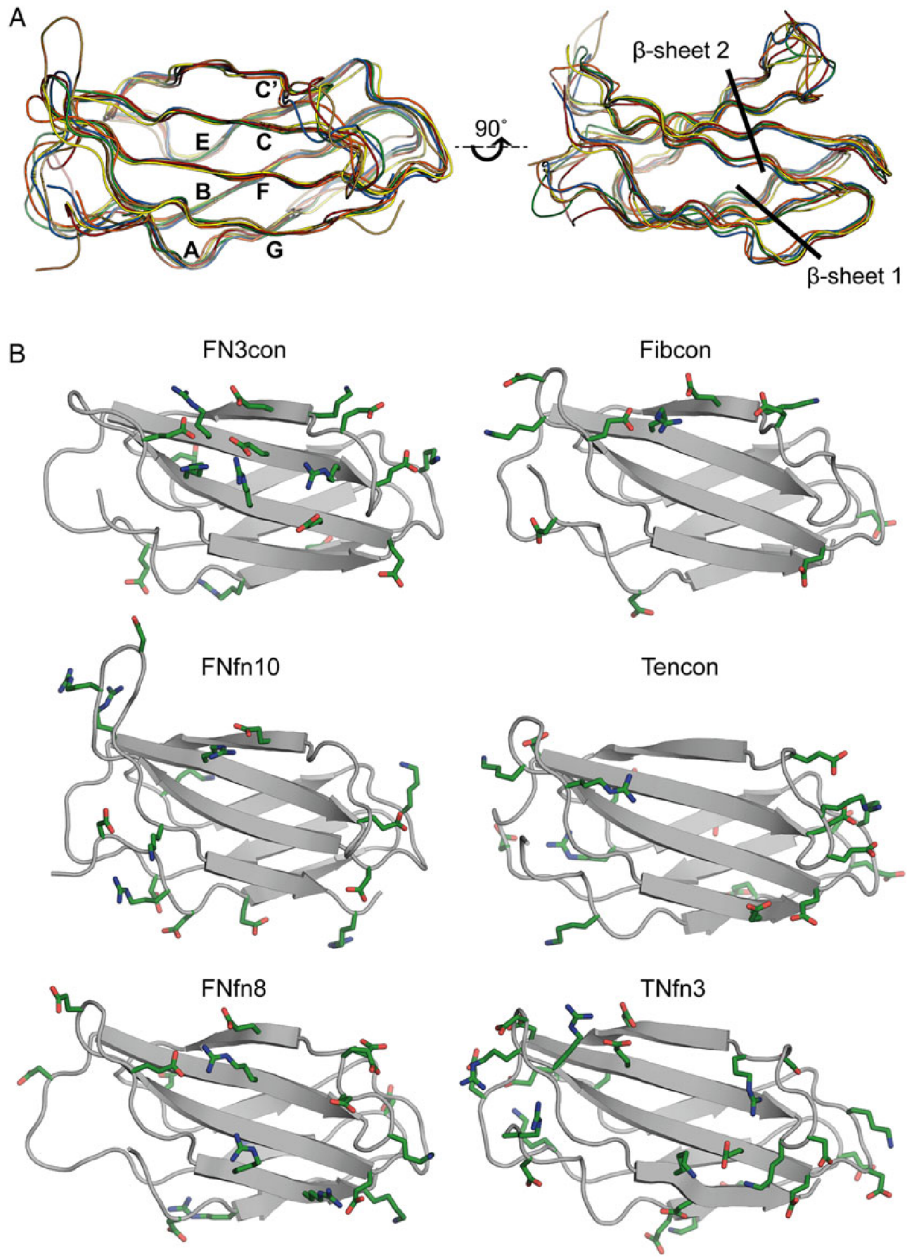


Fig. 2 Structural alignment and distribution of charged residues on β-sheet 2 (strands C, C′, F and G). (**A**) Structural alignments of FN3con (orange), Fibcon (blue), FNfn10 (brown), Tencon (red), FNfn8 (green) and TNfn3 (yellow). (**B**) Structural analysis showing charged residues (green) of FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3. Residue numbering is not included for clarity.

triplicate at 300 K (26.9°C) for 1 μs to investigate dynamics at room temperature, and at 368 K (94.9°C) for 2 μs to investigate structural response at high temperature. All domains display a similar dynamic behavior at 300 K, showing relatively low flexibility within the β-sheet and greater motion in the flexible loops, as expected (Fig. 6A). FN3con and Fibcon are both slightly more rigid than FNfn10, Tencon, FNfn8 and TNfn3 at 300 K; however, at 368 K dramatic differences are evident. At 368 K, FN3con, Fibcon and FNfn10 remain folded, with an average RMSD of 3.4, 4.1 and 4.1 Å, respectively. Comparatively, FNfn8 and TNfn3 start to unfold after 500 ns, with unfolding essentially complete by 1 μs, while Tencon shows signs of partial unfolding in some of the replicates at ~500 ns (Fig. 6B and Supplementary Movies S1, S2 and S3). Strikingly, the MD simulations faithfully support the experimentally derived stability hierarchy (Fig. 6B and Table I).

Strand swapping may play a role in thermostability and unfolding

Analysis of the simulation trajectories at 368 K reveals that, with the exception of Fibcon and TNfn3, all domains reveal some degree of strand swapping from one sheet to the other at either the N- or C-terminus. Specifically, in FN3con, FNfn10 and FNfn8, we observe strand G to swap from β-sheet 2 to β-sheet 1 (Fig. 7A and Supplementary Movie S1). Intriguingly, this is reversed in Tencon, with strand A

Table III. Global analysis of molecular contacts

Protein	PDB code	T_m (°C)	H-bonds[a]	Salt bridge interactions <7 Å[a]	ASA (Å^2)[b]	GRAVY score[c]	Negatively charged residues[d]	Positively charged residues[d]
FN3con	4U3H	>100	46	48	4545.5	–0.536	10	7
Fibcon	3TEU	89.6	42	14	4882.3	–0.190	8	3
FNfn10	1FNF	82.5	41	9	5470.8	–0.115	8	8
Tencon	3TES	78.0	44	28	5093.3	–0.307	12	7
FNfn8	1FNF	58.0	43	28	5239.0	–0.377	11	7
TNfn3	1TEN	57.1	45	41	5163.5	–0.577	18	9

[a]Hydrogen bonds and salt bridges were calculated using the WHATIF server (Hekkelman *et al.*, 2010).
[b]Calculated using the ASA tool from ccp4 (Winn *et al.*, 2011).
[c]Calculated using the ProtParam tool provided by ExPASy and uses the Kyte and Doolittle hydropathy value for each amino acid (Kyte and Doolittle, 1982).
[d]Negatively charged amino acids are counted as Asp and Glu, while positively charged amino acids are counted as Arg and Lys.

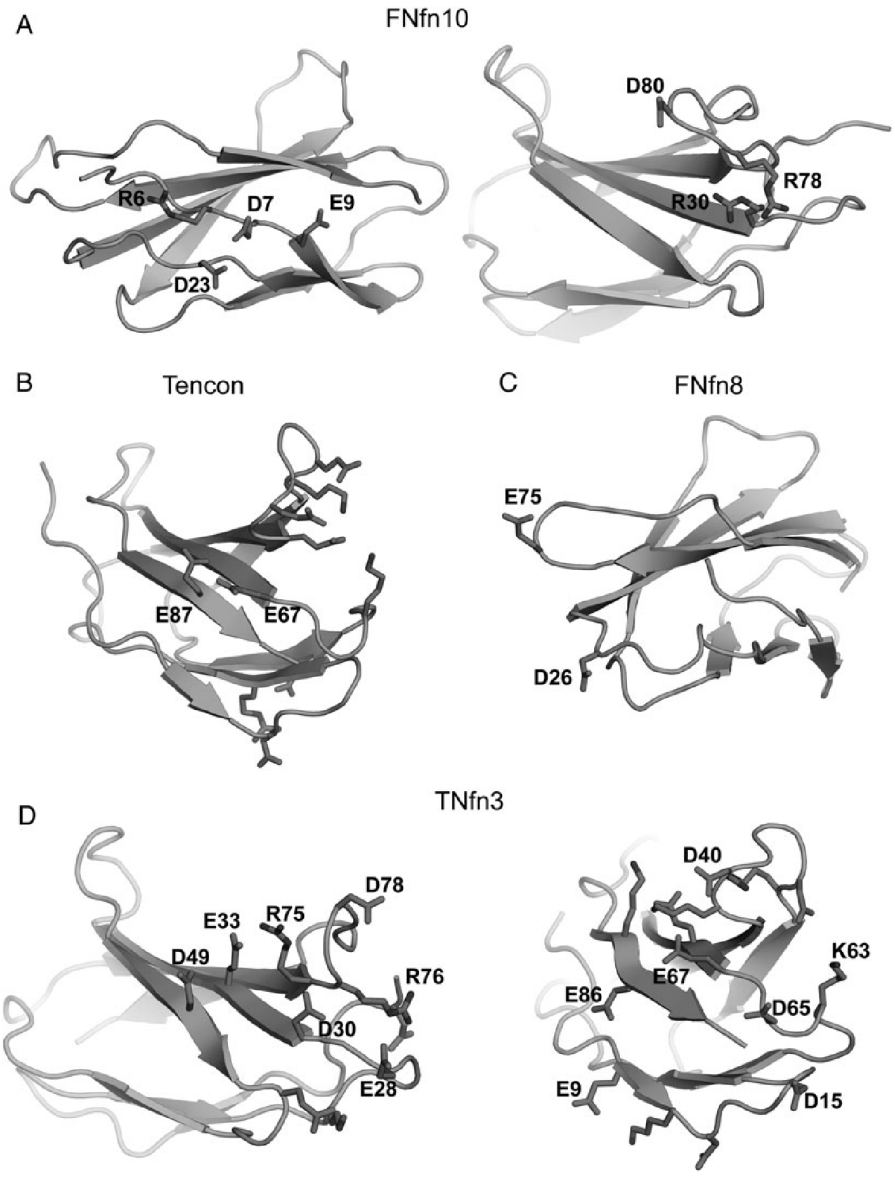


Fig. 3 Potentially destabilizing like-charged residue clusters. (**A**) FNfn10 showing two separate clusters. (**B**) Tencon, showing E67 and E78, which are surrounded by two complementary charged clusters. (**C**) FNfn8, showing D26 and E75. (**D**) TNfn3 showing like-charged residue clusters on each set of loop hairpins. The left panel shows the N-terminal loop region, highlighting potential destabilizing interactions between E33 in strand C and D49 in strand C′, as well as a potential long range repulsion from E28, D30 and D78. The right panel shows the C-terminal loop regions, highlighting potential destabilizing interactions between E9 in strand A and E86 in strand G, D15 in the A–B loop and D65 in the E–F loop, and D40 in the C–C′ loop and E67 in strand F.

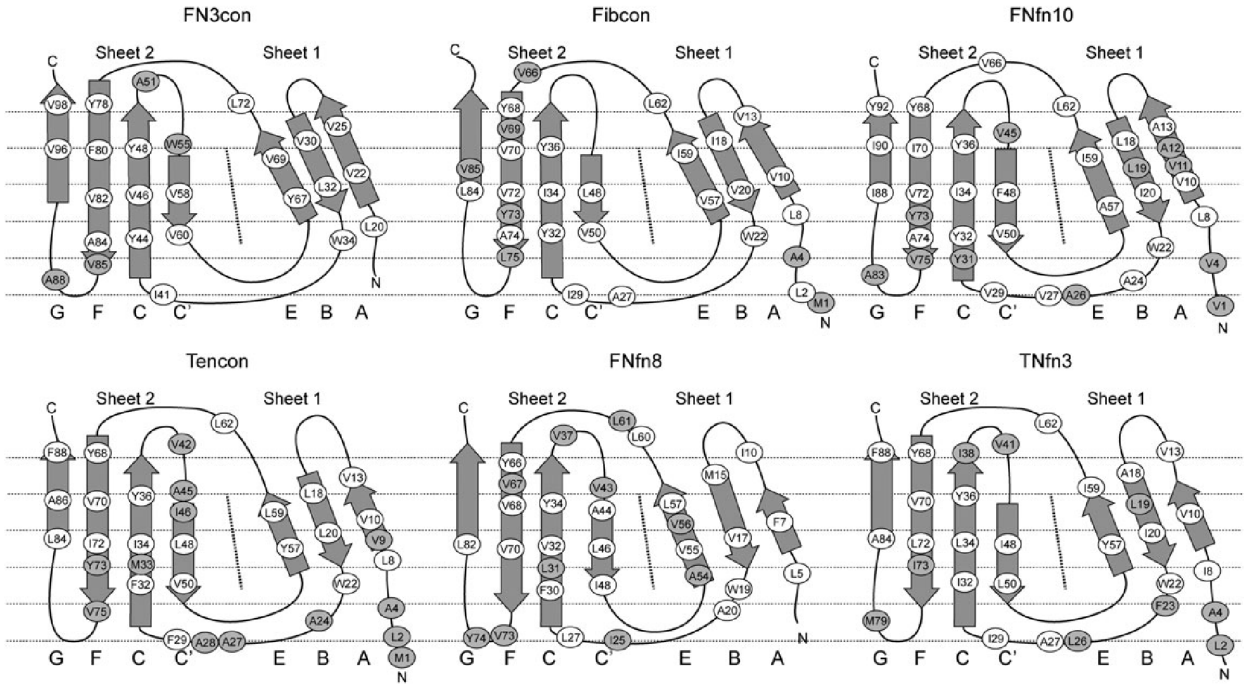


Fig. 4 Analysis of hydrophobic residue positions in FN3 domains. A schematic unfolded model of each FN3 domain is shown, indicating positions of the hydrophobic residues as ovals. White ovals indicate the residue as contributing to the hydrophobic core and, for the most part, not solvent exposed. Shaded ovals indicate exposure to solvent and lack of contribution to the hydrophobic core.

Table IV. Packing densities of FN3 domains

Protein	T_m (°C)	Total cavity volume (Å^3)[a]	Mean protein packing value (OSP)[b]
FN3con	>100	60.8	0.354
Fibcon	89.6	171.0	0.350
FNfn10	82.5	243.9	0.344
Tencon	78.0	260.7	0.343
FNfn8	58.0	185.8	0.356
TNfn3	57.1	334.1	0.335

[a]Calculated using the CASTp web server (Dundas *et al.*, 2006) with a 1.4 Å probe radius.

[b]Calculated using the OS software (Fleming and Richards, 2000); a higher value indicates better packing.

swapping from β-sheet 1 to β-sheet 2, forming a five-stranded β-sheet (Fig. 7B and Supplementary Movie S1). The effect of strand swapping on stability is not immediately obvious from the simulations. Strand swapping at 300 K is not observed during 1 μs, which may be due to a lack of conformational sampling. Both Tencon and FNfn8 exhibit partial to full unfolding after strand swapping, suggesting that strand swapping precedes or initiates the unfolding pathway by compromising the hydrophobic core. Although no strand swapping is seen in the Fibcon simulations, we instead observe the N-terminal strand to undergo large structural rearrangements that may expose the hydrophobic core to solvent and lead to eventual unfolding (Fig. 7C and Supplementary Movie S1). In the case of TNfn3, we do not observe any strand swapping, but rather, strands A and G of TNfn3 pull closer together in concert, followed by rapid unfolding. This motion does not appear to directly initiate unfolding, which is rapid and cooperative in nature; however, it is difficult to ascertain if this is due to the simulation temperature being significantly higher than the measured melting temperature. Given the prevalent like-charged residue clusters in TNfn3, unfolding may instead be initiated by electrostatic repulsion at both peripheral loops (Fig. 3D and Supplementary Movie S2).

The role of electrostatics in FN3 domain dynamics

Structural comparisons of FN3 domains revealed contrasting electrostatic interactions likely to induce positive or negative effects on stability (Figs 2 and 3). We therefore investigated whether electrostatics also play a role in the dynamics of FN3 domains. The complementary electrostatic mesh on β-sheet 2 of FN3con (Fig. 2) is stable throughout the simulations (at 300 and 368 K) indicating that the mesh is a stabilizing factor during the stress of high temperature, possibly by lowering the unfolding rate (Fig. 7A and Supplementary Movie S2). In contrast, one of the few surface electrostatic interactions in Fibcon (involving E47, E80 and R33) is short-lived during the simulation at 300 and 368 K and is unlikely to make a large contribution to stability (Supplementary Movie S2). During the simulations of FNfn10 at 368 K, the negatively charged cluster of D7, E9 and D23 is highly mobile, with charge repulsion causing the N-terminus to peel away into solvent, exposing the hydrophobic core. In addition, the neighboring positively charged residues R30 and R78 on strands C and F in FNfn10 rapidly rearrange throughout the simulation, with R30 burying itself into the hydrophobic core (Supplementary Movie S2). Our structural analysis of Tencon predicted charge repulsion of E67 and E87. The resulting dynamics simulations suggest that this may have some impact on the dynamics of the local area, with strands C and F regularly peeling away from one another at the E/F and C/C′ loop peripheries (Supplementary Movie S2). Finally, in FNfn8, the region surrounding residues D26 and E75 show pronounced motion prior to unfolding, suggesting a negative contribution to stability (Supplementary Movie S2).

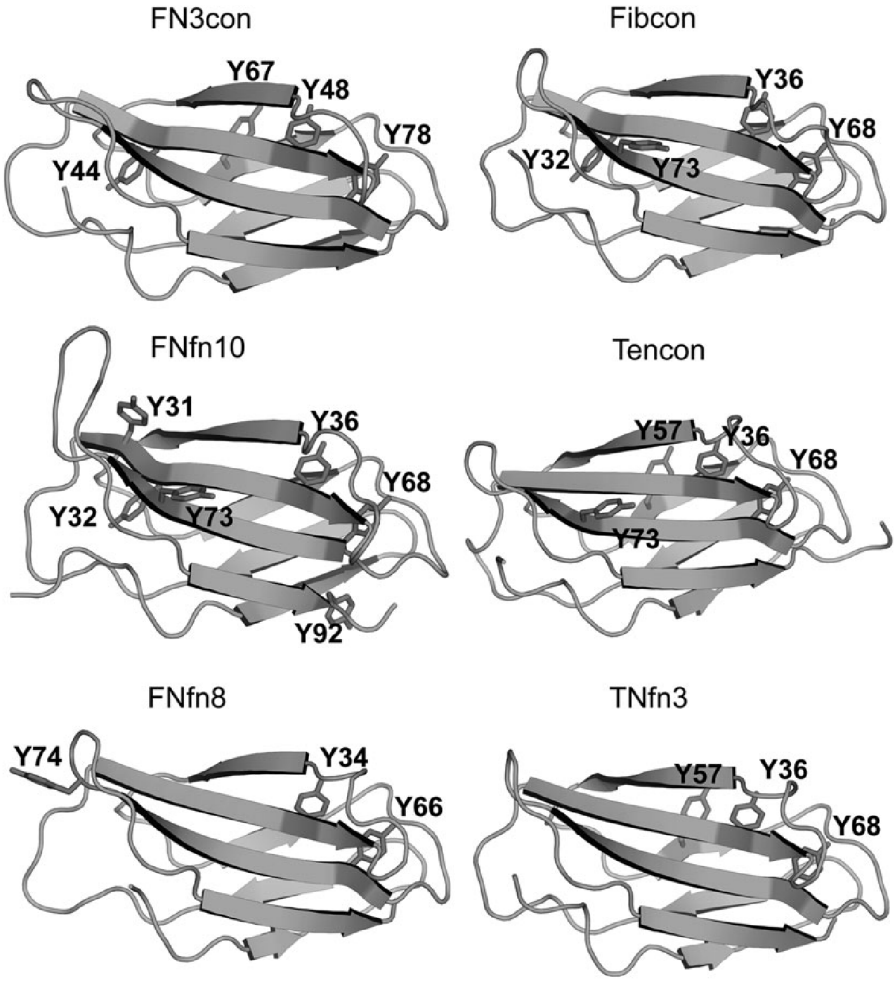


Fig. 5 Position of tyrosine residues in FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3.

Rigidity of the uniform hydrophobic core of FN3 domains may contribute to their stability and folding

The hydrophobic core of FN3con is highly regular, exhibiting uniform banding of hydrophobic residues (Fig. 4). Strikingly, this uniformity is retained throughout the high-temperature simulations and after strand swapping, a phenomenon that also occurs in FNfn10 and Tencon (Supplementary Movie S1 and Fig. 7A). In particular, the uniformity of FN3con is due to residues V96 and V98 realigning with L20 and V22 in strand A as strand G swaps from β-sheet 2 to β-sheet 1 (Supplementary Movie S1).

Dynamic recruitment of tyrosine corner residues

All of the assessed FN3 domains contain the highly conserved tyrosine residue, Y78 in FN3con (Y68/Y66 in the other domains). During the high-temperature simulations of all domains, Y78 is capable of dynamic rearrangement during strand swapping and thermal warping. Specifically, Y78 is recruited from the C′/E solvent interface to mediate solvent interactions when strand F becomes slightly separated from strand C (Supplementary Movie S3). Furthermore, the relatively stable domains of FN3con, Fibcon and FNfn10 contain a conserved tyrosine corner (residues Y44, Y32 and Y32, respectively) (Figs 4 and 5). This residue is not present in the less stable domains of Tencon, FNfn8 and TNfn3. In the simulations of FN3con, Fibcon and FNfn10, the side chains of Y44/Y32 are relatively rigid, suggesting a specialist role in stability that is consistent with other findings (Cota *et al.*, 2000). In FNfn8, a tyrosine residue is not present in this position, and as such, high-temperature simulations show that the solvent-exposed Y74 in the G/F loop is recruited to fulfill this role. However, given its position in the structure, such recruitment appears to have a destabilizing effect in the local area (Supplementary Movie S3). In Tencon, although Y73 is nearby, it is not positioned in the G/F loop, but rather at the C terminus of strand F; this positioning restricts dynamic motion and thus does not appear to play a role in stability. In TNfn3, there are no nearby tyrosine residues available to fulfill this role. FN3con contains an additional tyrosine corner motif (Y67) (Figs 4 and 5), whose interactions are almost identical to the equivalently positioned Y57 of Tencon and TNfn3, but absent in all other domains. In a previous MD simulation of TNfn3, Y36 makes several potentially stabilizing, non-crystallographic interactions (H-bonds and VdW) with Y57 and I20 (Paci *et al.*, 2003), which may indicate that the equivalent Y67 of FN3con makes a similar contribution to stability. Our simulations of FN3con show long-lived conformations of Y67 and Y48, suggesting that they play a role in stabilizing the C/E strand solvent interface (Supplementary Movie S3).

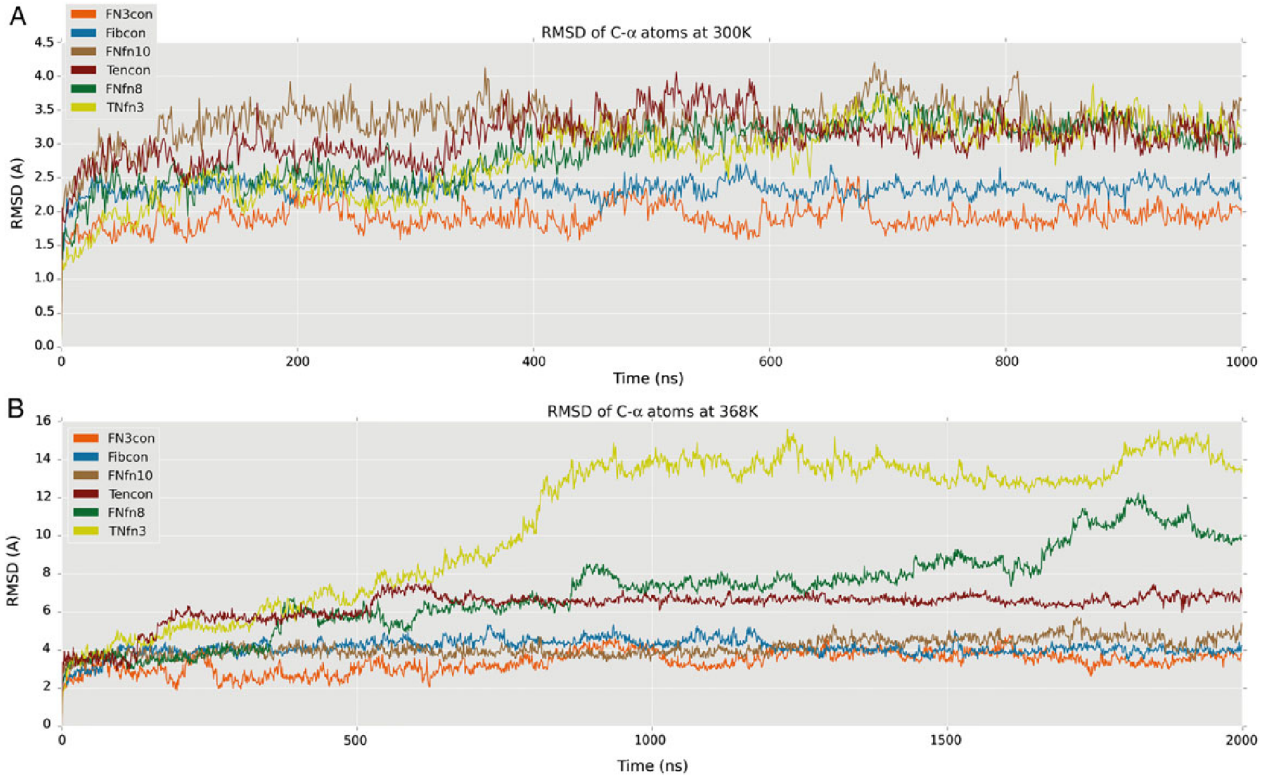


Fig. 6 RMSD plots at 300 and 368 K. All plots represent the mean RMSD across replicate simulations ($n = 3$), for Cα atoms. (**A**) RMSD plot of FN3con (orange), Fibcon (blue), FNfn10 (brown), Tencon (red), FNfn8 (green), TNfn3 (yellow) at 300 K. (**B**) RMSD plot of FN3con (orange), Fibcon (blue), FNfn10 (brown), Tencon (red), FNfn8 (green), TNfn3 (yellow) at 368 K.

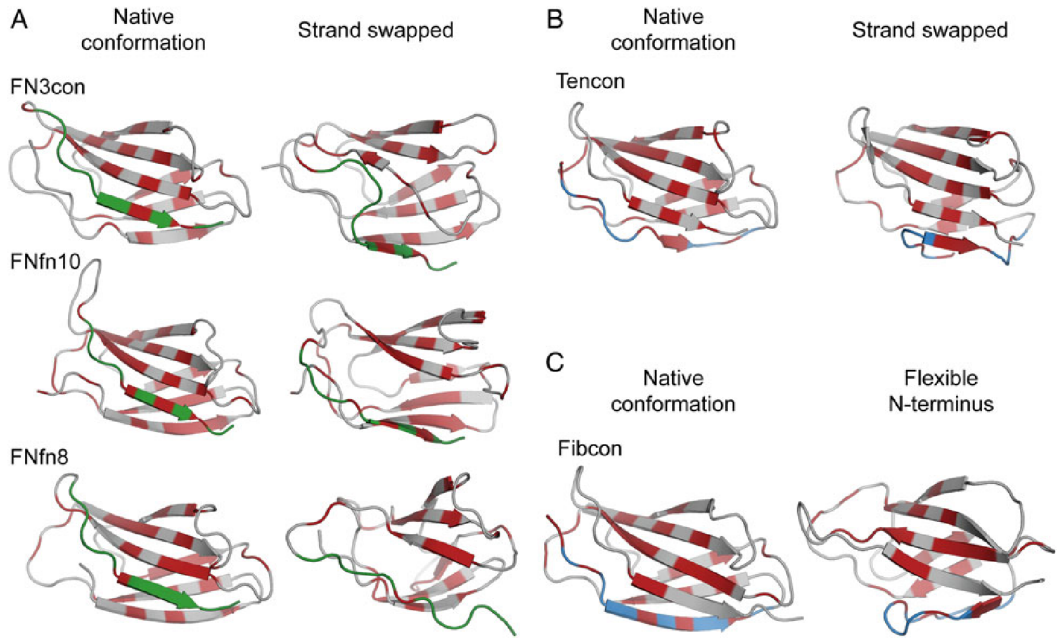


Fig. 7 Dynamics of C- and N-terminal strands at 368 K. (**A**) Cartoon representations of FN3con, FNfn10 and FNfn8 (gray) in their native (left) and strand swapped configurations (right), showing strand G (green) and hydrophobic residues (red). (**B**) Swapping of strand A (blue) in Tencon, showing the native conformation (left) and the five-stranded β-sheet (right) and hydrophobic residues (red). (**C**) The flexible N-terminus of Fibcon, showing strand A in blue and hydrophobic residues (red).

Discussion

In this study, we have described the consensus design and subsequent biophysical, structural and dynamical characterization of a novel FN3 domain, FN3con. Design of FN3con was carried out by complete sequence generation using 2123 homologous FN3 domain sequences. In a 2012 study (Jacobs *et al.*, 2012), two consensus designed FN3 domains were described (Fibcon and Tencon). This prior work was based on using 15 FN3 repeats from human Fibronectin to generate Fibcon and 15 repeats from human Tenascin to generate Tencon, as opposed to our large non-redundant selection from the Prosite database. Their resulting consensus domains had stabilities ranging from 78°C (Tencon) to 89°C (Fibcon), and have been further stabilized to 92.7°C using alanine scanning (Jacobs *et al.*, 2012). The authors of this study suggested that the quantity of sequences influences the outcome of consensus design to a greater extent than does quality. Specifically, they showed that using 15 sequences proved superior to using 7 of the most thermostable sequences. Overall our work supports this hypothesis, in that the use of 2123 sequences during the design of FN3con has made a significant contribution to its biophysical properties. Overall, FN3con is the most stable FN3 domain reported to date, having a T_m in excess of 100°C and a ΔG_{D-N} of 15.5 kcal mol^{-1}. It folds reversibly via two-state kinetics, with relatively fast folding and very slow unfolding rates (Fig. 1, Table I and Supplementary Fig. S2). As the FN3 superfamily is moderately conserved, with 18–41% sequence identity among the members (Fraser *et al.*, 2006), we therefore hypothesize that consensus design of a large, diverse family greatly benefits from the use of many sequences. We subsequently attribute the use of a large set of sequences to enhanced filtering of noise and a more authentic selection of conserved residues over the evolutionary landscape. In an effort to determine the molecular basis of stability in FN3con, we determined its X-ray crystal structure, which allowed structural and dynamics analyses and comparisons with Fibcon, FNfn10, Tencon, FNfn8 and TNfn3. Our results reveal that the superior stability of FN3con originates from highly specific and optimized electrostatic and hydrophobic interactions, as well as dynamic adaptability of the hydrophobic core at high temperature.

Calculations of physiochemical properties from the crystal structures revealed no relationship between the number of hydrogen bonds or salt bridges to stability. However, in a structural context, there are significant differences in the positioning of salt bridges and the ratio of positive and negatively charged residues, resulting in potential charge mismatches. The crystal structure of FN3con reveals a unique and extensive complementary charged electrostatic network that is distributed over β-sheet 2. This network consists of four arginine and four glutamic acid residues, and is not present in any of the other FN3 domains (Fig. 2B). Comparatively, TNfn3 contains a cluster of like-charged residues on the peripheral loops, which are likely to be destabilizing (Fig. 2B). The remaining FN3 domains show no sign of a linear correlation between salt bridge count and stability. This implies that stability is related to the structural context of salt bridge interactions rather than a numerical metric of potential interactions. The role of electrostatic interactions and their relation to thermal stability has been studied extensively. Surface electrostatic interactions typically make small contributions (~0.5 kcal mol^{-1}) to the overall stability, and tend to be context dependent and non-additive in nature (Serrano *et al.*, 1990). The energetic contribution provided by the electrostatic mesh in FN3con would be challenging to predict, given that each charged residue influences each other over long distances (2–7 Å) (Serrano *et al.*, 1990, 1993; Vaughan *et al.*, 2002). Although surface charged residues are unlikely to play a major role in thermodynamic stability, they may influence kinetic stability via effects on folding and unfolding rates (Cavagnero *et al.*, 1998; Karshikoff and Ladenstein, 2001; Sanchez-Ruiz, 2010). Accordingly, we hypothesize that the complementary electrostatic network seen in FN3con contributes to the dramatic reduction in unfolding rate, which has been reported for some thermophilic proteins.

Comparative analysis of hydrophobic packing in the set of FN3 domains reveals the presence of a banding pattern that is orthogonal to the direction of the β-strands (Fig. 4). This banding pattern is well known and important in formation of the folding nucleus (Lappalainen *et al.*, 2008). Strikingly, the degree of uniformity and alignment among hydrophobic residues in each band appears to be proportional to the stability of the domain. In general, we observe higher stability to be associated with uniform hydrophobic banding as well as greater burial and reduction of bulky hydrophobic residues, which is consistent with the established role of hydrophobic packing in protein stability (Fig. 4 and Table IV) (Fersht *et al.*, 1992; Serrano *et al.*, 1992; Buckle *et al.*, 1993; Fersht and Serrano, 1993; Axe *et al.*, 1996; Kellis *et al.*, 1988; Billings *et al.*, 2008). One of the most striking observations from our physiochemical properties was the dramatic decrease of solvent inaccessible cavity volume in FN3con, which is 2.8× smaller than the next best structure, Fibcon (Table IV and Supplementary Fig. S4). As packing density of the hydrophobic core is a known factor in protein stability, we suspect this attribute plays a significant role in the observed fast folding rate of FN3con (Karpusas *et al.*, 1989; Chothia and Finkelstein, 1990; DeDecker *et al.*, 1996; Levitt *et al.*, 1997; Ratnaparkhi and Varadarajan, 2000; Billings *et al.*, 2008).

Structural analysis of FN3con revealed the introduction of a cooperative electrostatic network, optimization of the hydrophobic core packing and accumulation of tyrosine corner residues in a positional pattern that is not seen in any of the other FN3 domains assessed. Given the complexity of interactions, we employed MD simulations to provide insight into the dynamics at ambient (300 K) and high temperature (368 K). Strikingly, the MD simulations at 368 K faithfully coincide with the experimentally derived stability hierarchy (Fig. 6B and Table I). Overall, the simulation trajectories reveal partial unfolding of Tencon and loss of native structure in FNfn8 and TNfn3 around 500 ns, which we attribute to the start of an unfolding pathway (Supplementary Movie S1). On closer inspection of the simulation trajectories at high temperatures, FN3con, FNfn10 and FNfn8 show C-terminal strand (strand G) swapping from β-sheet 2 to β-sheet 1 (Fig. 7A). Interestingly, as FN3con and FNfn10 strand swap the hydrophobic residues in strand G align perfectly to those in strand A (Fig. 7A and Supplementary Movie S1). This is in contrast to FNfn8, where the hydrophobic residues on strand G do not successfully align with those in strand A (Fig. 7A and Supplementary Movie S1), suggesting that the ability to realign the hydrophobic residues after strand swapping has an effect on stability. The simulations of Tencon also reveal strand swapping; however, there are dramatic differences compared with the other FN3 domains, with its N-terminal strand (strand A) swapping from β-sheet 1 to β-sheet 2 (Fig. 7B), forming a five-stranded β-sheet. Interestingly, mutations in the F/G loop of Tencon have been shown to promote strand swapping of the C-terminal strand (strand G), as well as influencing the resulting aggregation properties (Teplyakov *et al.*, 2014). However, it is unclear how this relates to the dynamics observed at 368 K, especially since strand G remains stable throughout MD of Tencon. Although there exists only one example of strand swapping within the current FN3 literature, folding studies, including Phi-value analysis, of FN3-like domains indicate folding occurs through a common-core ring involving strands B, C, E and F, leaving strands

A and G to pack last (Hamill *et al.*, 1998, 2000b; Cota and Clarke, 2000). This suggests a lack of constraints on strands A and G and is consistent with the strand swapping events we observe during the high-temperature simulations (Fig. 7 and Supplementary Movie S1). We therefore hypothesize that strand swapping is an event on the unfolding pathway.

MD simulations at 368 K reveal flexibility of loop regions in all structures, providing cavities for solvent to enter and potentially destabilize the hydrophobic core. Tyrosine corners feature tyrosine residues positioned near the beginning or end of an anti-parallel β-strand. This feature is highly conserved, ubiquitous and exclusive to Greek key proteins (Hemmingsen *et al.*, 1994; Hamill *et al.*, 2000a; Nicaise *et al.*, 2003). Tyrosine corners in the FN3 superfamily are involved in early structure formation and are important for stability of the structure, with tyrosine to phenylalanine mutations costing 1.5–3 kcal mol^{-1} in stability (Hamill *et al.*, 2000a). Our analysis of tyrosine residues showed a striking trend in that the most stable FN3 domains (FN3con, Fibcon and FNfn10) all contain tyrosine corners evenly spread throughout their structures and accessible to both peripheral loop regions. Specifically, FN3con, Fibcon and FNfn10 make use of a unique tyrosine residue (Y44, Y32 and Y32, respectively) at the C-terminal end of strand C; a trait not observed in Tencon, FNfn8 and TNfn3 (Figs 4 and 5). Intriguingly, FN3con, Tencon and TNfn3 share a unique tyrosine residue (Y67, Y57 and Y57, respectively) at the C-terminal end of strand E in sheet 1 (Figs 4 and 5). It has been suggested that Y57 makes a small contribution to stability in TNfn3 by forming H-bond and Van der Waals interactions with Y36 (Paci *et al.*, 2003). It is therefore likely that this residue makes a similar contribution to the stability of FN3con, given the close similarities in its environment and the rigidity of both residues in our MD analysis. In addition, simulations at 368 K reveal the capacity for rearrangement and recruitment of tyrosine residues at high temperature. One of the most striking differences is the lack of Y44/Y32 in Tencon, FNfn8 and TNfn3. Although FNfn8 attempts to recruit the solvent-exposed Y74, which is similarly positioned to Y44/Y32, it appears to destabilize the local area (Fig. 6 and Supplementary Movie S3). Furthermore, Tencon and TNfn3 lack the ability to reposition a tyrosine residue to this region. As such, we hypothesize that the presence of a unique distribution of tyrosine corners in FN3con provides stabilizing features and may contribute to the observed slow unfolding rate.

In conclusion, we have successfully generated an FN3 domain, FN3con, which has unprecedented stability, with experimental data highlighting a T_m in excess of 100°C, a ΔG_{D-N} of 15.5 kcal mol^{-1}, reversible folding via two-state kinetics, with the fastest folding and slowest unfolding rates reported to date. Structural and dynamical analysis reveals that FN3con stability does not result from a single mechanism, but rather the combination of several features and a strong tendency to remove non-conserved unfavorable interactions. These features include the introduction of a previously unseen complementary charged residue mesh on β-sheet 2, which we propose to contribute to the slow unfolding rate. FN3con includes the optimization of alignment within the hydrophobic core, resulting in superior packing, followed by removal of solvent-exposed hydrophobic residues and widespread adoption of tyrosine residues. Dynamics simulations reinforce the stability hierarchy determined by experiment and shed light on behavior of the FN3 domain at high temperature. Furthermore, we are the first to suggest that the flexibility and swapping of the N- and C-terminal strands of the FN3 domain are implicated in its unfolding pathway at high temperature, thus playing a role in its stability by allowing optimization of hydrophobic packing during conformational change. As such, FN3con features near perfect realignment of the hydrophobic core and recruitment of tyrosine residues. By exploiting the increased availability of genomic sequence data, this study further supports consensus design to be a rapid and effective method for the engineering of protein stability.

Methods

See Supplementary data.

Supplementary data

Supplementary data are available at *PEDS* online.

Author contributions

B.T.P., G.I.W., L.Z. and A.M.B. designed the study. B.T.P. performed the protein design, expression and purification, CD thermal melt experiments, crystallography, molecular dynamics simulations and analysis. A.A.N. performed the folding kinetics and equilibrium measurement experiments. D.E.H. assisted with crystallization of FN3con. S.M. assisted with crystallography and structure determination. B.T.P. generated figures and movies with assistance from M. R.H. B.T.P., D.E.H. and A.M.B. wrote the manuscript.

Acknowledgements

We would like to thank Itamar Kass, Andrew M. Ellisdon, Grischa R. Meyer and Bosco K. Ho for their helpful discussions and advice during the research and writing of this manuscript. We thank the Australian Synchrotron for beamtime and technical assistance.

Funding

A.A.N. is supported by the Wellcome Trust (grant number WT 095195). S.M. is an Australian Research Council Future Fellow (FT100100960). G.I.W. is an Australian Research Council Discovery Outstanding Researcher Award Fellow (DP140100087). A.M.B. is a National Health and Medical Research Senior Research Fellow (1022688). Funding to pay the Open Access publication charges for this article was provided by the Australian Research Council (grant number DP150101371).

References

Auerbach,G., Huber,R., Grättinger,M., Zaiss,K., Schurig,H., Jaenicke,R. and Jacob,U. (1997) *Structure*, **5**, 1475–1483.

Axe,D.D., Foster,N.W. and Fersht,A.R. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 5590–5594.

Billings,K.S., Best,R.B., Rutherford,T.J. and Clarke,J. (2008) *J. Mol. Biol.*, **375**, 560–571.

Bloom,L. and Calabro,V. (2009) *Drug Discov. Today*, **14**, 949–955.

Buckle,A.M., Henrick,K. and Fersht,A.R. (1993) *J. Mol. Biol.*, **234**, 847–860.

Cavagnero,S., Debe,D.A., Zhou,Z.H., Adams,M.W. and Chan,S.I. (1998) *Biochemistry*, **37**, 3369–3376.

Chothia,C. and Finkelstein,A.V. (1990) *Annu. Rev. Biochem.*, **59**, 1007–1039.

Clarke,J., Hamill,S.J. and Johnson,C.M. (1997) *J. Mol. Biol.*, **270**, 771–778.

Cota,E. and Clarke,J. (2000) *Protein Sci.*, **9**, 112–120.

Cota,E., Hamill,S.J., Fowler,S.B. and Clarke,J. (2000) *J. Mol. Biol.*, **302**, 713–725.

Dai,M., Fisher,H.E., Temirov,J., Kiss,C., Phipps,M.E., Pavlik,P., Werner,J.H. and Bradbury,A.R.M. (2007) *Protein Eng. Des. Sel.*, **20**, 69–79.

Dantas,G., Kuhlman,B., Callender,D., Wong,M. and Baker,D. (2003) *J. Mol. Biol.*, **332**, 449–460.

Davlieva,M. and Shamoo,Y. (2010) *Proteins*, **78**, 357–364.
DeDecker,B.S., O'Brien,R., Fleming,P.J., Geiger,J.H., Jackson,S.P. and Sigler,P. B. (1996) *J. Mol. Biol.*, **264**, 1072–1084.
Dundas,J., Ouyang,Z., Tseng,J., Binkowski,A., Turpaz,Y. and Liang,J. (2006) *Nucleic Acids Res.*, **34**, W116–W118.
Erickson,H.P. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 10114–10118.
Fersht,A.R. and Serrano,L. (1993) *Curr. Opin. Struct. Biol.*, **3**, 75–83.
Fersht,A.R., Matouschek,A. and Serrano,L. (1992) *J. Mol. Biol.*, **224**, 771–782.
Fleming,P.J. and Richards,F.M. (2000) *J. Mol. Biol.*, **299**, 487–498.
Fraser,J.S., Yu,Z., Maxwell,K.L. and Davidson,A.R. (2006) *J. Mol. Biol.*, **359**, 496–507.
Gilbreth,R.N., Chacko,B.M., Grinberg,L., Swers,J.S. and Baca,M. (2014) *Protein Eng. Des. Sel.*, **27**, 411–418.
Guelorget,A., Barraud,P., Tisné,C. and Golinelli-Pimpaneau,B. (2011) *BMC Struct. Biol.*, **11**, 48.
Hamill,S.J., Meekhof,A.E. and Clarke,J. (1998) *Biochemistry*, **37**, 8071–8079.
Hamill,S.J., Cota,E., Chothia,C. and Clarke,J. (2000a) *J. Mol. Biol.*, **295**, 641–649.
Hamill,S.J., Steward,A. and Clarke,J. (2000b) *J. Mol. Biol.*, **297**, 165–178.
Hekkelman,M.L., te Beek,T.A.H., Pettifer,S.R., Thorne,D., Attwood,T.K. and Vriend,G. (2010) *Nucleic Acids Res.*, **38**, W719–W723.
Hemmingsen,J.M., Gernert,K.M., Richardson,J.S. and Richardson,D.C. (1994) *Protein Sci.*, **3**, 1927–1937.
Horovitz,A., Serrano,L., Avron,B., Bycroft,M. and Fersht,A.R. (1990) *J. Mol. Biol.*, **216**, 1031–1044.
Jacobs,S.A., Diem,M.D., Luo,J., Teplyakov,A., Obmolova,G., Malia,T., Gilliland,G.L. and O'Neil,K.T. (2012) *Protein Eng. Des. Sel.*, **25**, 107–117.
Jin,W., Kambara,O., Sasakawa,H., Tamura,A. and Takada,S. (2003) *Structure*, **11**, 581–590.
Karpusas,M., Baase,W.A., Matsumura,M. and Matthews,B.W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8237–8241.
Karshikoff,A. and Ladenstein,R. (2001) *Trends Biochem. Sci.*, **26**, 550–556.
Kellis,J.T., Nyberg,K., Sali,D. and Fersht,A.R. (1988) *Nature*, **333**, 784–786.
Koide,A., Jordan,M.R., Horner,S.R., Batori,V. and Koide,S. (2001) *Biochemistry*, **40**, 10326–10333.
Koide,A., Wojcik,J., Gilbreth,R.N., Hoey,R.J. and Koide,S. (2012) *J. Mol. Biol.*, **415**, 393–405.
Kuhlman,B. (2003) *Science*, **302**, 1364–1368.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lappalainen,I., Hurley,M.G. and Clarke,J. (2008) *J. Mol. Biol.*, **375**, 547–559.
Lehmann,M. and Wyss,M. (2001) *Curr. Opin. Biotechnol.*, **12**, 371–375.
Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P.G.M. and Wyss,M. (2002) *Protein Eng.*, **15**, 403–411.
Levitt,M., Gerstein,M., Huang,E., Subbiah,S. and Tsai,J. (1997) *Annu. Rev. Biochem.*, **66**, 549–579.
Loladze,V.V., Ibarra-Molero,B., Sanchez-Ruiz,J.M. and Makhatadze,G.I. (1999) *Biochemistry*, **38**, 16419–16423.
Magliery,T.J., Lavinder,J.J. and Sullivan,B.J. (2011) *Curr. Opin. Chem. Biol.*, **15**, 443–451.
Maxwell,K.L. and Davidson,A.R. (1998) *Biochemistry*, **37**, 16172–16182.
Nakamura,A., Takumi,K. and Miki,K. (2010) *J. Mol. Biol.*, **396**, 1000–1011.
Nicaise,M., Valerio-Lepiniec,M., Izadi-Pruneyre,N., Adjadj,E., Minard,P. and Desmadril,M. (2003) *Protein Eng.*, **16**, 733–738.
Nikolova,P.V., Henckel,J., Lane,D.P. and Fersht,A.R. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 14675–14680.
Paci,E., Clarke,J., Steward,A., Vendruscolo,M. and Karplus,M. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 394–399.
Ratnaparkhi,G.S. and Varadarajan,R. (2000) *Biochemistry*, **39**, 12365–12374.
Richardson,J.S. and Richardson,D.C. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 2754–2759.
Russell,R.J., Hough,D.W., Danson,M.J. and Taylor,G.L. (1994) *Structure*, **2**, 1157–1167.
Russell,R.J. and Taylor,G.L. (1995) *Curr. Opin. Biotechnol.*, **6**, 370–374.
Sanchez-Ruiz,J.M. (2010) *Biophys. Chem.*, **148**, 1–15.
Sanchez-Ruiz,J.M. and Makhatadze,G.I. (2001) *Trends Biotechnol.*, **19**, 132–135.
Serrano,L., Horovitz,A., Avron,B., Bycroft,M. and Fersht,A.R. (1990) *Biochemistry*, **29**, 9343–9352.
Serrano,L., Kellis,J.T., Cann,P., Matouschek,A. and Fersht,A.R. (1992) *J. Mol. Biol.*, **224**, 783–804.
Serrano,L., Day,A.G. and Fersht,A.R. (1993) *J. Mol. Biol.*, **233**, 305–312.
Shah,P.S., Hom,G.K., Ross,S.A., Lassila,J.K., Crowhurst,K.A. and Mayo,S.L. (2007) *J. Mol. Biol.*, **372**, 1–6.
Steipe,B., Schiller,B., Plückthun,A. and Steinbacher,S. (1994) *J. Mol. Biol.*, **240**, 188–192.
Sundaresan,R., Ragunathan,P., Kuramitsu,S., Yokoyama,S., Kumarevel,T. and Ponnuraj,K. (2012) *Biochem. Biophys. Res. Commun.*, **420**, 692–697.
Teplyakov,A., Obmolova,G., Malia,T.J., *et al.* (2014) *Proteins*, **82**, 1359–1369.
Vaughan,C.K., Harryson,P., Buckle,A.M. and Fersht,A.R. (2002) *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 591–600.
Wang,Q., Buckle,A.M., Foster,N.W., Johnson,C.M. and Fersht,A.R. (1999) *Protein Sci.*, **8**, 2186–2193.
Winn,M.D., Ballard,C.C., Cowtan,K.D., *et al.* (2011) *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 235–242.

Supplementary information

Methods

Multiple Sequence Alignments

The sequence alignment for homologous fibronectin type III domains was sourced from the Prosite database (http://prosite.expasy.org/PDOC50853). The Prosite sequence alignments are hand curated with no definitive method. The sequences were retrieved in FASTA format in February 2012, returning 2,619 entries. In order to reduce bias from the sequence alignment, the CD-HIT web server [157] was used to remove redundant sequences above 95% similarity, reducing the sequence count to 2,123 entries. Sequences were re-aligned using ClustalW [158]. Application of the consensus algorithm was applied over all 2,123 sequences, resulting in a single consensus sequence for FN3con.

Protein Expression and Purification

Genes encoding FN3con and FnFN8, containing an N-terminal 6x HIS tag, followed by a thrombin cleavage site (LVPRGS), were chemically synthesized and provided in a pJexpress 404 plasmid by DNA2.0. The resulting plasmids were transformed into competent C41 *E. coli* cells for expression. A single colony from each transformation was picked and grown overnight at 37°C in 250 ml of 2xYT (16.0 g/L tryptone, 10.0 g/L yeast extract, 5.0 g/L NaCl) media containing 100 µg/ml of ampicillin. These cultures were then used to seed 2 L of 2xYT media for the FN3con and FNfn8. Cultures were induced at an OD_{600} of 0.9 with IPTG (0.5 mM final concentration), and grown for a further 5 hours at 37°C. The cultures were harvested and cell pellets resuspended in TBS (50 mM Tris, 150 mM NaCl, pH 7.4; EDTA free protease inhibitors, ThermoFisher), lysed via sonication and cellular debris removed by centrifugation (5,000x G). Recombinant protein was isolated from the whole cell lysate by metal affinity chromatography using loose NiNTA resin (Sigma). Protein eluted from NiNTA resin was filtered (0. 22 µm) then subjected to size exclusion chromatography using a Superdex 75 16/60 column (GE Healthcare) equilibrated in either PBS (140 mM NaCl, 2.7 mM KCl, 10 mM PO_4^{3-}, pH 7.4) for biophysical characterization or low salt TBS

(50 mM Tris, 50 mM NaCl, pH 7.4) for protein crystallography. Protein concentration was determined by Nanodrop ND-1000 (ThermoFisher) and protein was stored at 4°C until use (biophysical characterization) or used immediately (protein crystallography).

Characterisation of Thermal Stability

Thermal stability of purified FN3con and FnFN8 was measured by monitoring the circular dichroism (CD) signal at 222 nm to assess secondary structure content. Protein samples were used at a concentration of 0.2 mg/ml. T_m values were measured using a Jasco J-815 CD spectrophotometer with a peltier thermal control unit (CDF-426S). A quartz cuvette with a path length of 1 mm was used throughout. Samples were heated from 20°C to 110°C at a rate of 1°C per minute and monitored at 222 nm. Far-UV scans from 195 nm and 260 nm were collected in triplicate at 20°C before and after each melt. A buffer-only scan was collected in order to calculate a baseline. Following baseline removal, data was collected in triplicate, averaged and fit to a two-state unfolding model using a non-linear least squares fitting algorithm [159,160]. The melting temperature (T_m) was calculated as the G/Mg ratio.

Equilibrium Measurements

A 6 M solution of guanidine isothiocyantate (GITC) in TBS was combined in varying ratios with TBS buffer using a liquid handling robot to create a range of denaturant solutions from 0 – 6 M GITC. These solutions were subsequently mixed in an 8:1 ratio with 9 μM protein in TBS to give a final concentration of 1 μM protein. All solutions were left to equilibrate at 25°C for at least three hours, after which the fluorescence of each solution was measured on a Perkin Elmer LS55 fluorimeter using an excitation wavelength of 280 nm and an emission range of 300 – 400 nm. Readings were obtained from a 1 cm pathlength cuvette maintained at 25 ± 0.1°C. The experiment was repeated, but using 9 μM protein pre-unfolded in 5 M GITC to generate a refolding curve. These solutions were left to equilibrate for at least six hours before their fluorescence was ascertained.

Kinetic Measurements

Folding was monitored by changes in fluorescence using a 350 nm cut-off filter and an excitation wavelength of 280 nm. All experiments were performed using an Applied Photophysics (Leatherhead, UK) stopped-flow apparatus maintained at 25 ± 0.1°C. For unfolding experiments, one volume of 11 μM protein solution was mixed rapidly with ten volumes of a concentrated GITC solution. For refolding, one volume of denatured protein in 4 M GITC solution was mixed with ten volumes of low-concentration GITC. In all cases, both solutions contained TBS buffer and were equilibrated at 25°C for at least 30 minutes before use. Data collected from at least six experiments were averaged and traces were fit to a single or double exponential function as appropriate. Due to mixing effects, data collected in the first 2.5 ms were always removed before fitting.

Data analysis of equilibrium and kinetic measurements

An Excel spreadsheet was used to derive the fluorescence average emission wavelength (AEW) for each of the equilibrated denaturant solutions [160,161]. Excel was also used to convert each denaturant concentration into a denaturant activity since the two values are not directly proportional for GITC [161,162]. A plot of AEW against denaturant activity (Kaleidagraph, Synergy Software) yielded the expected sigmoidal plot, which was fitted to the standard two-state equation [163] to obtain the *m*-value ($m_{D\text{-}N}$), the denaturant activity 50% ($[D']_{50}$) and hence the stability of the protein in TBS buffer ($D\mathit{G}_{D\text{-}N}$). Both the unfolding and refolding AEW curves can be converted to Fraction Folded by first removing the baselines and then normalizing the resulting data.

All kinetic traces fitted well to a single exponential decay plus a linear drift term. Longer experiments indicated the presence of a much slower second refolding rate that was incompatible with the timescale of the stopped-flow apparatus. Since FN3con has 11 proline residues, we attribute this rate to proline isomerization although further experiments are required to confirm this hypothesis. An amplitude analysis suggests that this slower rate accounts for between 50% and 80% of all proteins at low concentrations of denaturant. The resulting chevron plot showed rollover in the refolding arm (indicating the presence of a refolding intermediate) and a kink in the unfolding

arm (indicating the presence of a high energy intermediate). It was fitted using Prism (Synergy Software) to the following equation to estimate all parameters:

$$\ln(k_{\text{obs}}) = \ln\left(\frac{1}{2}\left(-A_1 - \sqrt{{A_1}^2 - 4A_2}\right)\right)$$

where:

$$A_1 = -\left(k_f + k_{-1}e^{m_{-1}[D']} + k_2 e^{-m_2[D']} + k_{-2}e^{m_{-2}[D']}\right)$$

$$A_2 = \left(k_f\left(k_2 e^{-m_2[D']} + k_{-2}e^{m_{-2}[D']}\right) + k_{-1}e^{m_{-1}[D']} + k_{-2}e^{m_{-2}[D']}\right)$$

$$k_f = k_i e^{-m_i[D']}\left(\frac{1}{1 + \frac{k_i e^{-m_i[D']}}{k_d e^{-m_d[D']}}}\right)$$

k_i and m_i are the folding rate constant from the refolding intermediate (I) to the first transition state (TS1) and its associated *m*-value, k_d and m_d are from the denatured state (D) to the first transition state (TS1), k_{-1} and m_{-1} are unfolding from the high energy intermediate (I*) over TS1, k_2 and m_2 are folding from the high energy intermediate (I*) over TS2, k_{-2} and m_{-2} are unfolding from the native state (N) over TS2. By convention, k_{-1} is set as 100,000 s^{-1} and m_{-1} is set as 0 M^{-1}: *m*2 is thus the *m*-value between TS1 and TS2 while the ratio k_{-1}/k_2 informs on the difference in free energy between the two transition states.

Crystallization of FN3con

FN3con was purified in 50 mM Tris, 50 mM NaCl, pH 7.4 and was concentrated to 25 mg/ml (Milipore 3 kDa cutoff concentrator). Concentrated FN3con was filtered through a 0.22 µm centrifugal filter and crystals were obtained from 0.1M phosphate-citrate pH 4.2 and 40% PEG300

(JCSG+ Suite, Qiagen). Drops were prepared 1:1 in 1 µl. Small hexagonal or cubic crystals were formed within three days.

X-ray diffraction, structure determination and refinement

FN3con crystals were flash frozen in liquid nitrogen without further cryoprotection. Diffraction data was collected at the Australian Synchrotron on the MX1 beamline and initially processed with Blu-Ice. Diffraction data to 1.98 Å resolution was collected and processed with iMOSFLM [164]. Complete data collection statistics can be found in Table 2. The FN3con structure was determined by molecular replacement (MR) with Phaser [165] using PDB entry 2CK2 [166] as a search probe (following removal of solvent atoms and trimming of sidechains to create a poly-Ala model). The asymmetric unit contains one protein molecule. Model building and structure refinement was carried out with PHENIX v. 1.8.4-1496 [167] and Coot [168]. Coordinates of FN3con were deposited in the RCSB Protein Data Bank with PDB ID 4U3H.

Structure analysis

In analysis of FN3con, residue numbering was kept as per amino acid positions of the construct, due to non-ideal sequence to structure alignment with the other FN3 domains (sequences in Data S2). In analysis of FNfn8, residues 1238-1325 in 1FNF were renumbered 1-88 with residue P1238 as residue 1. In analysis of FNfn10, residues 1416-1509 of 1FNF were renumbered 1-94 with residue V1416 as residue 1. In analysis of TNfn3, residues 802 to 891 from the original PDB file of 1TEN were renumbered 1-90 with residue R802 as residue 1. In analysis of Fibcon and Tencon, residue numbering was unchanged and is per respective PDB files 3TEU and 3TES. C-terminal His tags from Fibcon and Tencon were removed for structural analysis and molecular dynamics simulations.

Structural alignments were performed using the Mustang-MR webserver [169]. H-bonds and salt-bridges (<7 Å) were calculated using the WHATIF server [170]. Accessible surface area (ASA)

was calculated using the ASA tool from CCP4 [171]. The grand average hydropathy (GRAVY) score was calculated using the ProtParam tool provided by ExPASy and uses the Kyte and Doolittle hydropathy value for each amino acid [159]. Total cavity volume was calculated using the CASTp web server [160] using a 1.4 Å probe radius. Mean occluded surface packing (OSP) was calculated using the OS software [161].

System setup for molecular dynamics simulations

Simulations of FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3 were based on the following crystal structures with PDB codes 4U3H, 3TEU, 1FNF, 3TES, 1FNF, 1TEN respectively. Coordinates were prepared by removal of crystal waters, N- or C-terminal His tags and extracted from their respective PDB files as per listings in the *Structure analysis* methods section. Residues with missing atoms were modelled using MODELLER [162], followed by capping of the N- and C-termini with the neutral *N*-methyl amide and acetyl groups. All residues were simulated at their dominant protonation state at pH 7. Completed structures were solvated in a cubical simulation box with a minimum distance of 1.4 nm from any protein atoms to the box wall, followed by the addition of sodium and chloride ions to neutralize the system. Extra NaCl was added to reach a final concentration of approximately 150 mM NaCl. System dimensions and compositions are listed in Table S1.

Table S1. Simulation system dimensions and composition

System	Dimensions	Sodium ions	Chloride ions	Water molecules	Total atoms (approximate)
FN3con	6.8 nm^3	32	29	10,262	31,700
Fibcon	7.3 nm^3	40	35	12,313	37,800
FNfn10	7.0 nm^3	31	31	10,888	33,600
Tencon	7.0 nm^3	36	31	11,019	31,700

FNfn8	7.1 nm^3	15	11	11,600	35,700
TNfn3	7.2 nm^3	44	34	11,905	36,700

Simulation protocol

All simulation systems were subjected to energy minimization, followed by equilibration in the NPT ensemble (26.85 °C (300 K), 1 bar (~1 atm)) or (94.85 °C (368 K), 1 bar (~1atm)), with 1,000 kJ mol^{-1} nm^{-2} positional restraints applied to all non-hydrogen atoms; restraints were stepped down 10 fold every 100 ps over 300 ps. Equilibrated systems were run at 300 K and 368 K for 1 µs and 2 µs, in triplicate, with each replicate starting from a different distribution of initial velocities. All simulations were performed using GROMACS ver 4.0.7 [162,172] in conjunction with the GROMOS 53A6 united-atom force field [172,173]. Water was represented explicitly using the simple-point-charge (SPC) model [173,174]. All simulation systems were performed in an NPT ensemble under periodic conditions. Temperature was maintained close to its reference value of 300 K or 368 K by V-rescale temperature coupling [165,175]. Pressure was maintained close to a reference value of 1 atm by isotropic coupling with a Berensden pressure bath [166,175]. Non-bonded interactions were evaluated using a twin-range cut-off scheme: interactions falling within the 0.8 nm short-range cutoff were calculated every 2 fs whereas interactions within the 1.4 nm long cutoff were updated every 10 fs, together with the pair list. A generalized reaction-field correction was applied to the electrostatic interactions beyond the long-range cutoff [167,176], using a relative dielectric permittivity constant of $\varepsilon RF = 62$ as appropriate for SPC water [168,177]. All bond lengths to hydrogen atoms were constrained using the P-LINCS algorithm [169,178] and water geometry was constrained using the SETTLE algorithm [170,179]. A leap-frog integrator [172] was used throughout, with a time step of 2 fs.

Simulation Analysis

Analyses of the simulations were performed using the tools provided in the GROMACS package 4.0.7 [172] and custom scripts in conjunction with ProDy [180]. Graphs and plots were produced

using Matplotlib [181]. Molecular graphics were prepared with PyMol ver. 1.3.2 [182] and Visual Molecular Dynamics (VMD) 1.9.2 [183].

SI Data, Movies, and Figures

Data S1. Fasta file containing the sequence alignment used in consensus design of FN3con.

Data S2. Fasta file containing a sequence alignment of FN3con, Fibcon, FNfn10, Tencon and FNfn8.

Movie S1. Molecular dynamics simulations of FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3 at 368 K (94.85°C) for 2 μs, highlighting the dynamic alignment of the hydrophobic core. Structures are represented in cartoon form, with Cα atoms of hydrophobic residues represented as red spheres.

Movie S2. Molecular dynamics simulations of FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3 at 368 K (94.85°C) for 2 μs, highlighting the positions of electrically charged residues.

Movie S3. Molecular dynamics simulations of FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3 at 368 K (94.85°C) for 2 μs, highlighting the dynamical motions of tyrosine residues.

Data S1, S2 and Movies S1, S2, S3 can be found at the following link:
https://peds.oxfordjournals.org/content/suppl/2015/02/17/gzv002.DC1/gzv002supp_data.zip

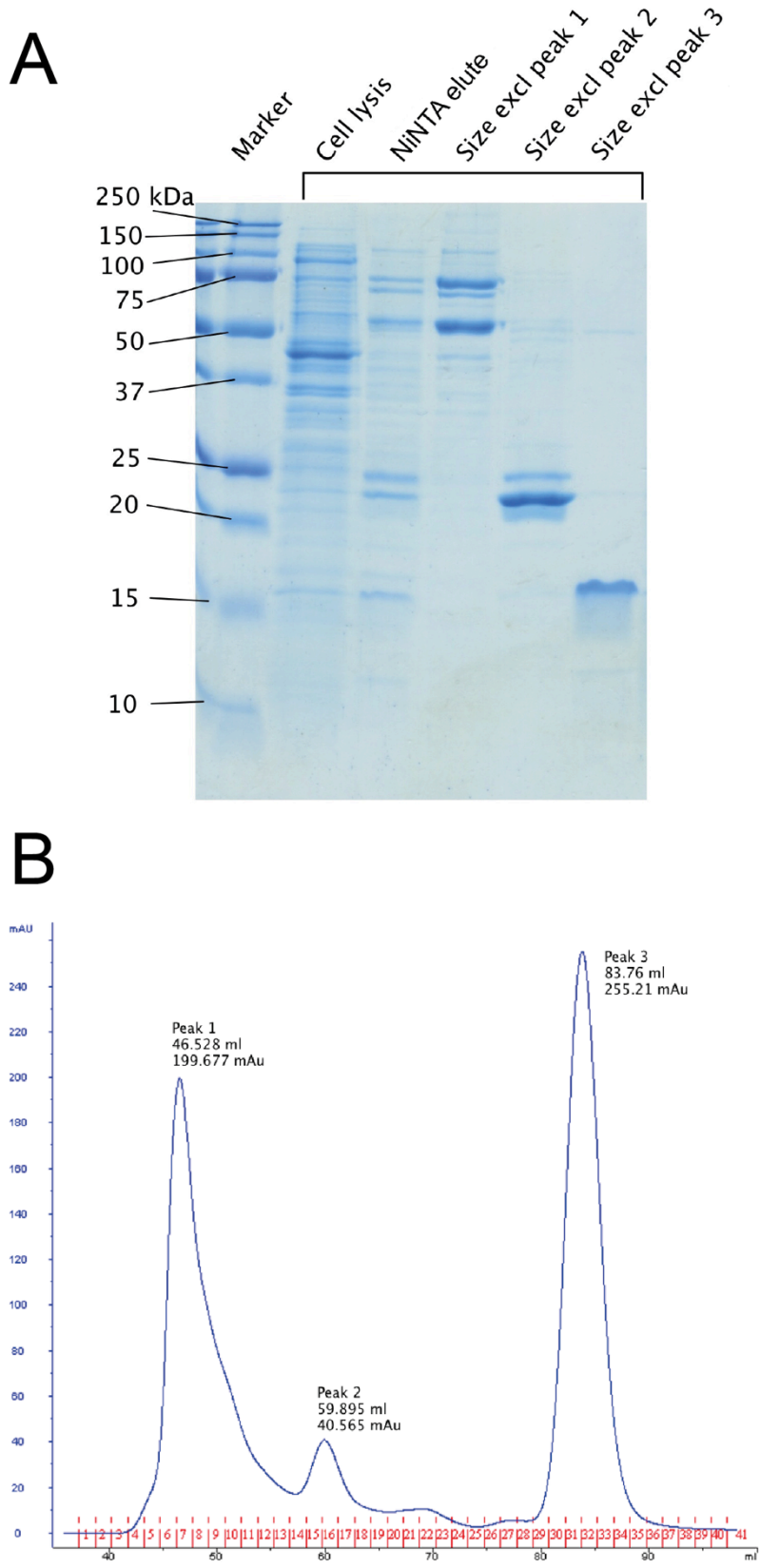


Figure S1. Purification of FN3con. (A) SDS page gel of the FN3con purification process, showing cell lysis, NiNTA elution fraction and the size exclusion peaks from B. (B) Size exclusion chromatography plot of FN3con from NiNTA elution, with peak 3 being FN3con.

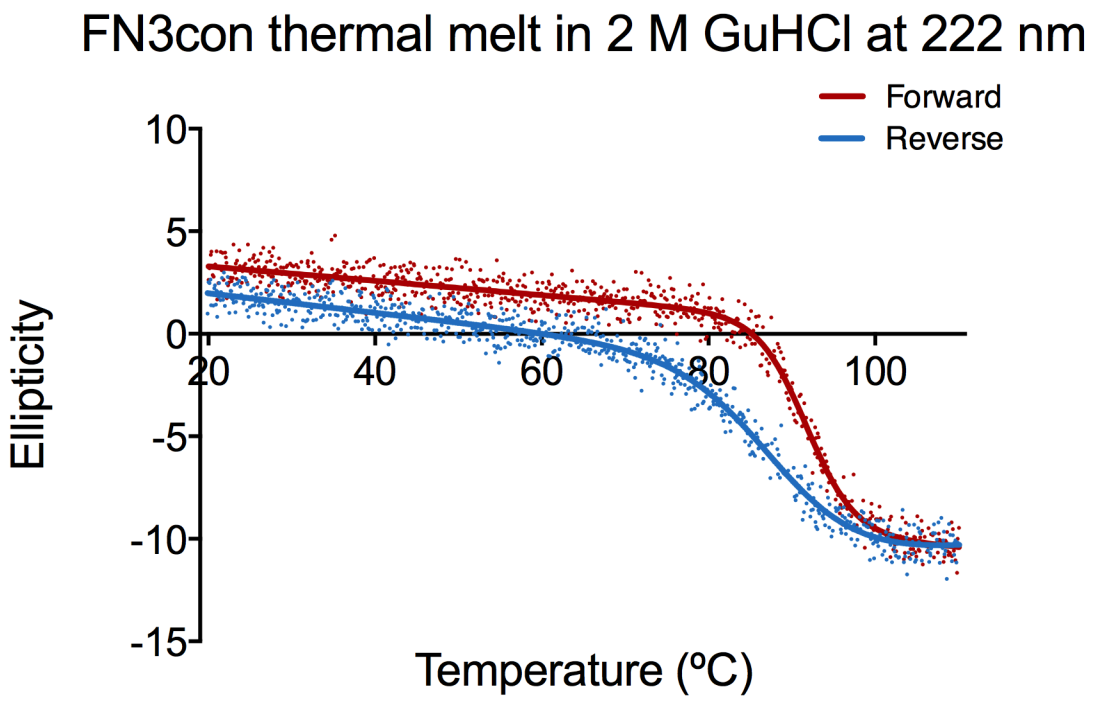


Figure S2. Reversible thermal folding of FN3con in 2 M GuHCl, monitored by CD at 222 nm. FN3con was heated from 20°C to 110°C (red) and cooled from 110°C to 20°C (blue). Respective non-linear fits were applied to the individual data points (R^2=*0.98* forwards and R^2=*0.98* reverse).

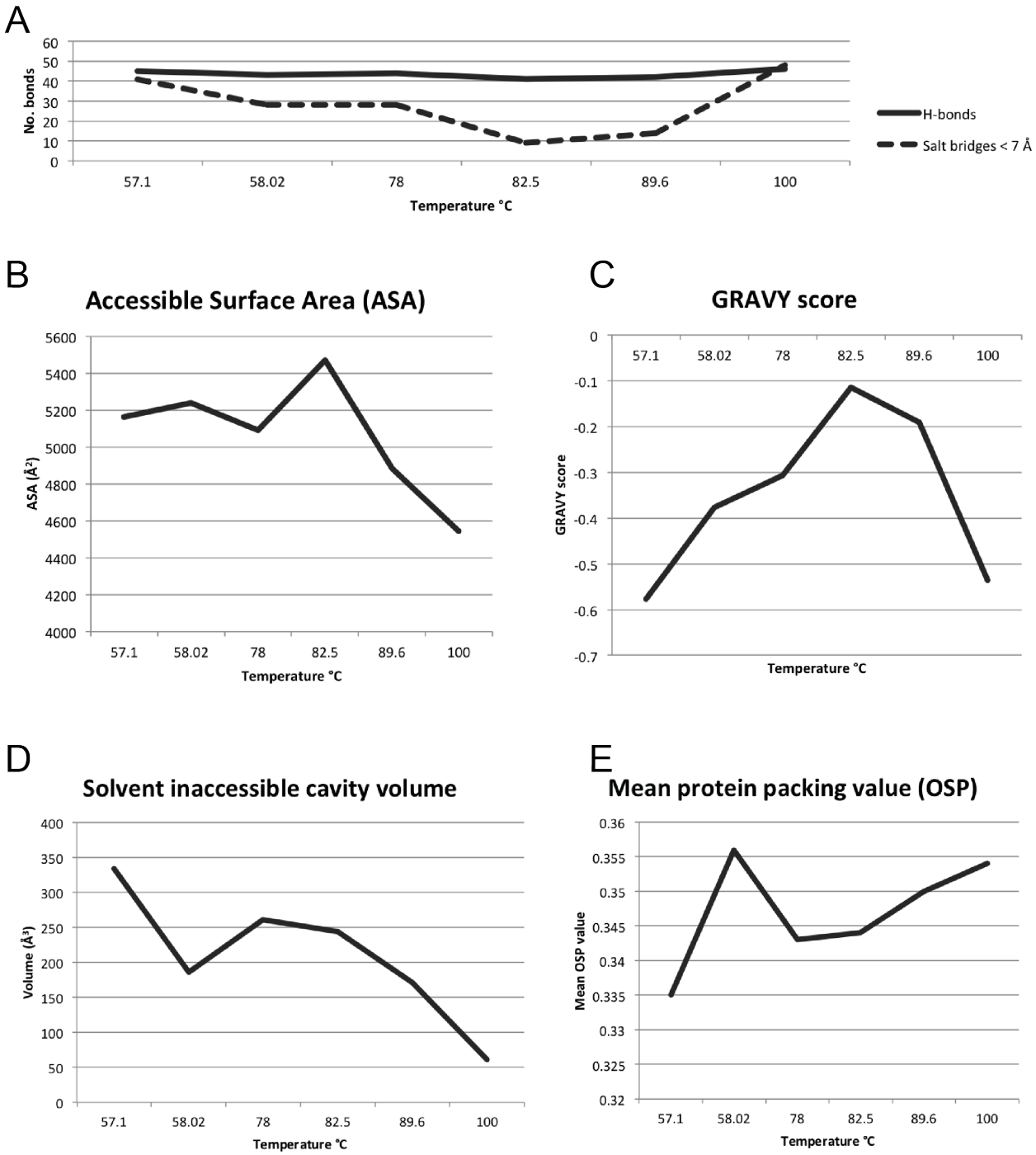


Figure S3. Plots of physiochemical properties from Tables 3 and 4 against determined melting temperatures of the FN3 domain. (A) Number of hydrogen bonds (solid line) and salt bridges (dashed line). (B) Solvent accessible surface area of respective FN3 domains plotted against temperature. (C) The grand average hydropathy (GRAVY) score of respective FN3 domains. (D) Solvent inaccessible cavity volume of FN3 domains in respect to their melting temperatures. Lower value indicated less cavity volume. (E) Mean protein packing value (OSP), larger value indicating better surface packing.

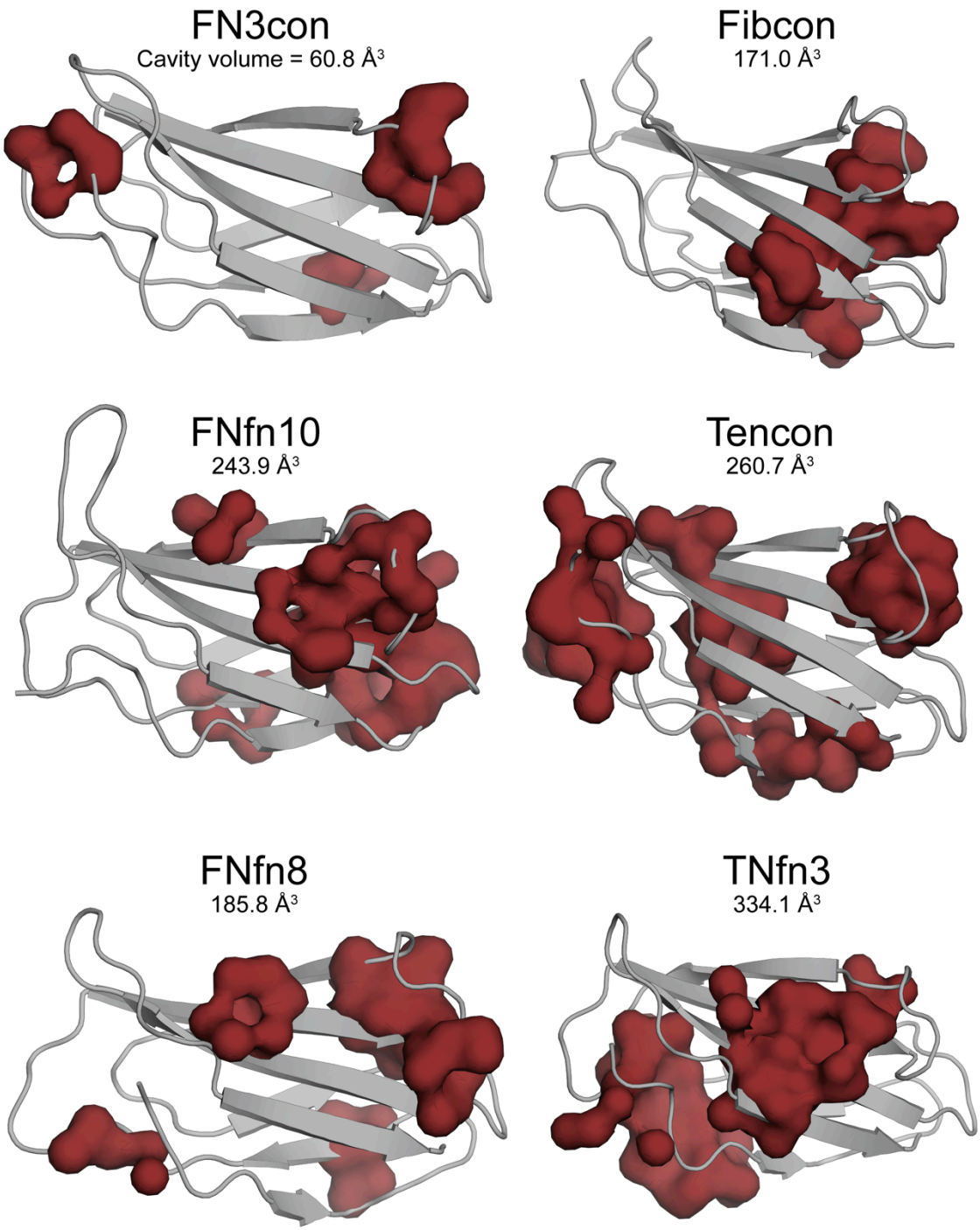


Figure S4. Approximate cavity volumes (red surface), including residues associated with the cavity volumes, as generated by the CASTp webserver using a 1.4 Å probe, for FN3con, Fibcon, FNfn10, Tencon, FNfn8 and TNfn3 (grey cartoon).

Chapter 3

Circumventing the stability-function trade-off in an engineered FN3 domain

Summary

In this chapter, I highlight the fact that as the biophysical properties of non-antibody scaffolds make them attractive alternatives to monoclonal antibodies, they too suffer from a stability function trade-off that results in marginal improvements. In order to assess whether FN3con (Chapter 2) provides any advantage as a non-antibody scaffold, I performed rational loop grafting from a previously engineered lysozyme binding FN3 domain. Rational design reveals the capacity to circumvent the stability-function trade-off, with FN3con exhibiting both high-affinity binding to lysozyme and a high level of thermodynamic stability.

This chapter has since been expanded and published in Protein Engineering Design and Selection – doi: 10.1093/protein/gzw046.

Circumventing the stability-function trade-off in an engineered FN3 domain

Authors

Benjamin T. Porebski[1], Paul J. Conroy[1], Nyssa Drinkwater[2], Morag R. Hunter[3], Rodrigo Vazquez-Lombardi[4,5], David E. Hoke[1], Daniel Christ[4,5], Sheena McGowan[2], Ashley M. Buckle[1,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia.

[2]Biomedicine Discovery Institute and Department of Microbiology, Monash University, Clayton, Victoria 3800, Australia.

[3]Department of Pathology, University of Cambridge, Cambridge, CB2 1QP, United Kingdom.

[4]Department of Immunology, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, NSW 2010, Australia

[5]Faculty of Medicine, St Vincent's Clinical School, The University of New South Wales, Darlinghurst, Sydney, NSW 2010, Australia

*To whom correspondence should be addressed:

Ashley M. Buckle

Biomedicine Discovery Institute and The Department of Biochemistry and Molecular Biology, Faculty of Medicine,

Monash University, Clayton, Victoria 3800 Australia

[redacted]

[redacted]

[redacted]

Key words: Protein engineering, Consensus design, stability-function trade-off, loop grafting, X-ray crystallography

Abstract

The favorable biophysical attributes of non-antibody scaffolds make them attractive alternatives to monoclonal antibodies. However, due to the well-known stability-function trade-off, these gains tend to be marginal after functional selection. A notable example is the fibronectin type III (FN3) domain, FNfn10, which has been previously evolved to bind lysozyme with 1 pM affinity (FNfn10-α-lys), but suffers from poor thermodynamic and kinetic stability. To explore this stability-function compromise further, we grafted the lysozyme-binding loops from FNfn10-α-lys onto our previously engineered, ultra-stable FN3 scaffold, *FN3con*. The resulting variant (FN3con-α-lys) bound lysozyme with a markedly reduced (1 million fold) affinity, but retained high levels of thermal stability. The crystal structure of FNfn10-α-lys in complex with lysozyme revealed unanticipated interactions at the protein-protein interface involving framework residues of FNfn10-α-lys, thus explaining the failure to transfer high affinity binding via loop grafting alone. Utilizing this structural information, we redesigned FN3con-α-lys and restored high-affinity binding to lysozyme, whilst maintaining thermodynamic stability (with a thermal melting temperature two-fold higher that that of FNfn10-α-lys). FN3con therefore provides an exceptional window of stability to tolerate deleterious mutations, providing a substantial advantage for functional design. This study emphasises the utility of consensus design for the generation of highly stable scaffolds for downstream protein engineering studies.

Introduction

A major goal of protein engineering is to produce novel proteins that bind to a specified target. Immunoglobulins are a natural scaffold for binding, and antibodies can be generated for virtually any given target [184-187]. Despite their success as a rapidly growing class of therapeutics [8], unmodified immunoglobulins nevertheless are subject to a range of limitations, such as their large size, challenges with over expression, solubility and stability that can limit their applicability [188,189]. To overcome the size and stability limitations of monoclonal antibodies, a large body of work has focused on the engineering of antibody single domains and fragments [186] and on increasing stability through mutation [190,191].

An alternative strategy relates to the generation of non-antibody scaffolds, which show great potential in terms of affinity, ease of production, target neutralization and stability for diagnostics, biotechnology and therapeutics [192-194]. Although the stability of parental non-antibody scaffolds is often considerable, this is not always observed for generated binders; indeed a stability-function tradeoff can often be observed, resulting in suboptimal candidates after functional selection [144,154,155,195,196]. This can result in non-antibody binders that are less stable or only marginally better than their monoclonal antibody counterparts [193].

One particular example of this tradeoff is observed in the fibronectin type III (FN3) domain. The benchmark scaffold from this family is the 10^{th} FN3 repeat from human fibronectin (FNfn10) [197]. This particular domain is chosen because it is the most stable human FN3 repeat with a midpoint of thermal denaturation (T_m) of 84°C [197] and has the capacity to tolerate a number of mutations in three surface-exposed loops (B/C, D/E and F/G which are analogous to the complementary determinant regions (CDRs) of antibodies) [197-200]. Combinatorial libraries have been built into these loops and specific binders selected for several different targets [199]. Whilst these engineered FN3 domains have been shown to display very high affinities to their targets, they often

exhibit a large reduction in thermodynamic stability, solubility and are prone to aggregation [198,199]. For example, Wittrup and colleagues evolved FNfn10 to bind lysozyme with high affinity [201]. The resulting clone (DE0.4.1, which we refer to as FNfn10-α-lys herein), bound lysozyme with an affinity of 1 pM, but had a T_m of 51±3°C that is 33°C lower than wild type FNfn10 (T_m of 84°C) [201].

Although many proteins have been shown to display a stability-function trade-off [13,154,155,196], it is possible for high stability and functionality to co-exist [50,55,202-205]. In order to further explore the stability-function trade-off in FN3 domains, we questioned whether the potent binding activity of FNfn10-α-lys could be achieved without a loss in stability. We previously reported the consensus design of a FN3 domain, *FN3con*, which exhibits an extremely high degree of thermodynamic and kinetic stability (T_m >100°C, reversible folding and aggregation resistant) [137]. In this study, we describe loop grafting from FNfn10-α-lys onto the stable FN3con scaffold (creating FN3con-α-lys). This variant, although destabilized by a relatively small amount, bound lysozyme with a significantly reduced affinity. To investigate the structural reason for reduced affinity, we determined the crystal structure of both FNfn10-α-lys in complex with lysozyme and of FN3con-α-lys alone, thereby enabling the rational redesign of FN3con-α-lys (FN3con-α-lys.v2). Redesign successfully restored binding affinity with a substantially smaller loss in thermodynamic stability, demonstrating that function and stability are not mutually exclusive. We discuss the implications of our findings for the use of consensus design in the generation of highly stable binding scaffolds that circumvent the stability-function trade-off.

Results and Discussion

Construction of FN3con-α-lys by loop grafting

Loop grafting is a common approach in the generation of humanized antibodies and affinity transfer across similar protein scaffolds [206]. Using this approach, we constructed FN3con-α-lys, in which the B/C, D/E and F/G loops from FNfn10-α-lys were grafted onto FN3con. FNfn10-α-lys was purified from insoluble inclusion bodies after recombinant expression in *E. coli*, with yields of 50 mg/L. Purified FNfn10-α-lys remained soluble for short periods of time (max. of 24 h at 4 °C) before visibly precipitating. While similar protein yields were obtained for FN3con-a-lys, this variant expressed as a soluble monomer that, after purification, remained in solution at 10 mg/ml for longer than 90 days at 4 °C. These preliminary observations served as an indication of the superior stability and aggregation resistance of the FN3con-a-lys variant.

Biophysical characterization of FNfn10-α-lys and FN3con-α-lys

FNfn10-α-lys undergoes irreversible thermal denaturation with a T_m of 43±2 °C, as measured by circular dichroism (CD), with complete loss of CD signal and visible precipitate upon cooling (Fig. 1A). This measurement is considerably lower than the previously reported T_m of 51±3 °C [201], likely due to the thermal denaturation assay used in the cited study; which assessed binding of yeast-displaying FNfn10-α-lys to lysozyme after heating. In striking contrast to FNfn10-α-lys, FN3con-α-lys unfolds reversibly with a T_m of 101±3 °C (Fig. 1B). Characterization of lysozyme-binding by size exclusion chromatography (SEC) and surface plasmon resonance (SPR) shows that FNfn10-α-lys forms a tight complex with lysozyme (Fig.1C and 1E), consistent with previous reports [201]. However, the loop-grafted construct FN3con-α-lys binds lysozyme weakly (Fig. 1D), with a fast dissociation-rate and low signal amplitude, as determined by SPR (Fig. 1F). We observed significant non-specific binding to lysozyme that could not be remedied in SPR by the use of a modified running buffer containing 12 mg/ml CM-Dextran. This prevented quantitative kinetic evaluation, since the measurement included both specific and non-specific binding.

However, qualitatively, these results demonstrate that direct grafting of the lysozyme-binding loops onto FN3con does not transfer high-affinity binding, suggesting that either the loops are unable to engage lysozyme in similar fashion or that interactions outside the grafted loops may play important roles.

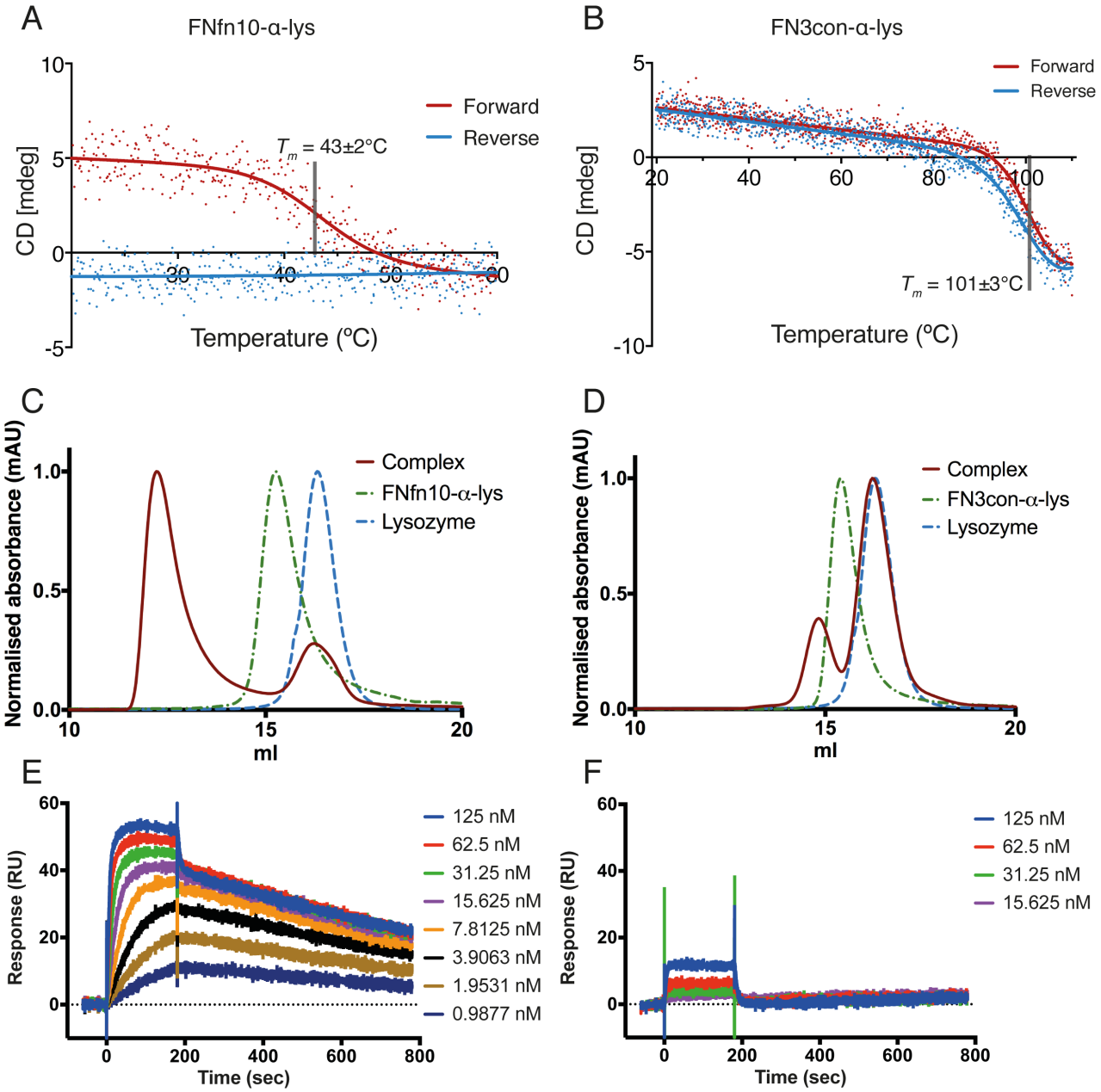


Figure 1. Biophysical characterization of FNfn10-α-lys and FN3con-α-lys. Circular dichroism (CD) thermal melts at 222 nm of **(A)** FNfn10-α-lys (T_m of 43±2 °C) and **(B)** FN3con-α-lys (T_m of 101±3 °C). Size exclusion chromatography (SEC), revealing complex formation for **(C)** FNfn10-α-lys and **(D)** FN3con-α-lys with lysozyme. Representative surface plasmon resonance (SPR) sensograms of **(E)** FNfn10-α-lys and **(F)** FN3con-α-lys with concentrations of lysozyme from 125 nM to 0.99 nM introduced during the mobile phase.

Direct loop grafting: Structural rationalization of suboptimal binding performance

In order to glean structural insight into our biophysical results, we determined the crystal structure of FNfn10-α-lys in complex with lysozyme at 2.54 Å resolution (Table 1 and Fig. 2). The asymmetric unit contains two copies of the FN3–lysozyme complex, arranged with a 1:1 binding stoichiometry. Both copies of the complex are highly conserved (root mean square deviation of 0.15 Å between the two complexes over 194 Cα atoms), and clear electron density for both subunits is observed at the protein-protein interface. The interface buries ~930 $Å^2$ and ~873 $Å^2$ surface area for FNfn10-α-lys and lysozyme respectively. The interface area is at the higher end of the scale of antibody-antigen interfaces [207] and other FN3-based variants [208].

The interaction utilises all three binding loops of FNfn10-α-lys, which packs well into the lysozyme active site cleft (Fig. 2A). The *Sc* statistic [209], which is a measure of the binding surface complementarity, is 0.79 (scale from 0.0 to 1.0, with 1.0 being perfect complementarity). This value is greater than the range observed for protease-protease inhibitors (0.71–0.76), oligomeric interfaces (0.70–0.74), antibody-antigen complexes (0.66–0.68) [209], and other FN3 domain-protein complexes (0.64-0.76) [208]. Although all three loops play a role in the interface, the structure reveals interactions between the framework residues (those that contribute to β-sheet secondary structure) of FNfn10-α-lys and lysozyme. These residues were not expected to form part of the interaction interface, and were therefore not grafted onto FN3con-α-lys, thus offering a simple explanation of poor affinity transfer upon sequence-based loop grafting (Fig. 2A, 2B and 2C).

The binding interface involves 14 hydrogen bonds mediated by seven residues (P27, A29, Y31, T77, R78, V79 and R81) and 6 salt bridges contributed by two residues (R78 and R81) in FNfn10-a-lys (Table S1). The number of interactions is on the upper end of the scale in comparison to seven other FN3 domain-protein complexes [208], which is consistent with the high affinity of FNfn10-α-lys for lysozyme and size of the interface.

A sequence alignment between FNfn10-a-lys and FN3con-a-lys (Fig. 2C) revealed nine framework positions that were not grafted (Table S1), that may be involved in recognition of the lysozyme epitope. In order to accurately discriminate the framework differences between FNfn10-α-lys and FN3con-α-lys, we determined the crystal structure of unbound FN3con-α-lys (Table I). There are eight molecules within the asymmetric unit, with prevalent interactions of the C-terminal His tag with strands from adjacent molecules (Fig. S1). Disordered electron density for the N-terminus and parts of the C-terminal His tag is observed in all 8 copies of the molecule, and density for the B/C and D/E loops is variable between the molecules. Further, density for the F/G loop is not observed in any monomer. By using all eight molecules and homology modeling (for the FG loop only), it was possible to generate a complete model of FN3con-α-lys (Fig. S1).

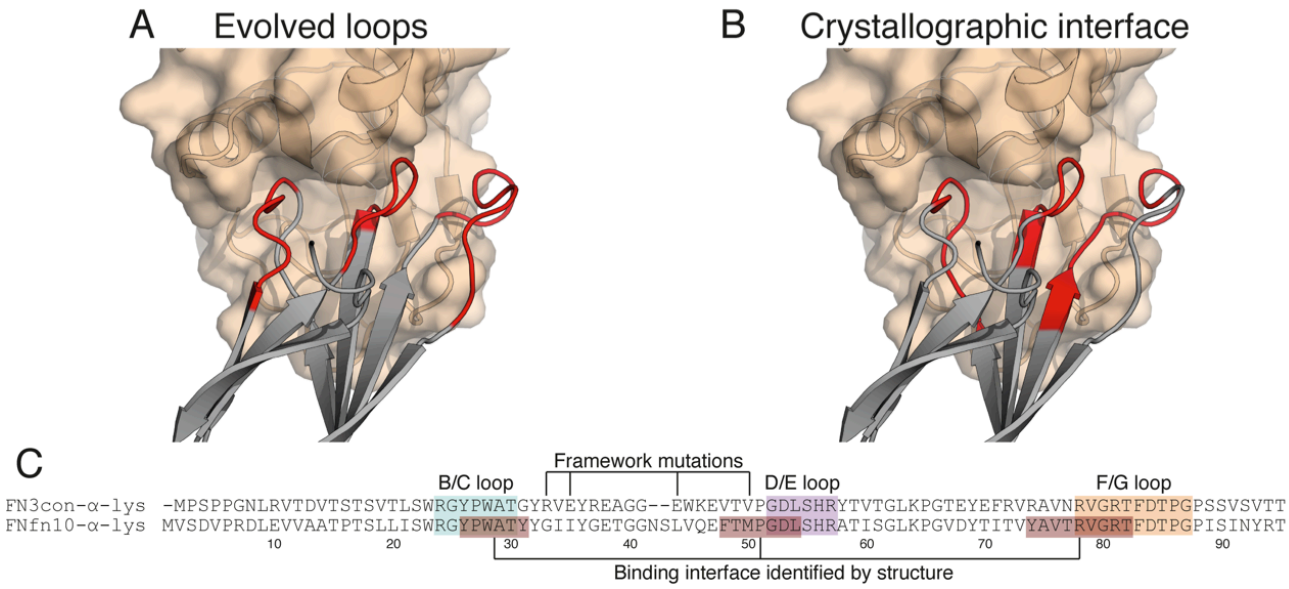


Figure 2. The FNfn10-α-lys-lysozyme complex reveals a tight binding interface that makes use of framework residues. **(A)** The loop residues (red) of FNfn10-α-lys (grey) as previously evolved for lysozyme (tan) binding [201]. **(B)** The actual binding interface residues (red) of FNfn10-α-lys (grey) with lysozyme (tan) as determined by crystal structure and the PDBePISA webserver [210]. **(C)** A sequence alignment of FN3con-α-lys with FNfn10-α-lys, highlighting the B/C, D/E and F/G loops (blue, purple and orange) that were previously evolved for lysozyme binding [201], the actual residues involved in the binding interface (red) and positions of the FNfn10-α-lys framework mutations previously introduced [201].

A structural alignment between FNfn10-α-lys and FN3con-α-lys reveals two features that are consistent with the poor affinity transfer. First, a conformational change is present in strand D of FNfn10-α-lys that results in a 180° flip and register shift (Fig. 3A) that is not present in FNfn10 (Fig. 3B) or FN3con-α-lys (Fig. 3C). Although this conformational change may be the result of induced fit on binding with lysozyme, we could not predict this event from sequence alone. Subsequently, the flip and shift alters the physiochemical properties of the binding surface (paratope) on strand D (Fig. 3D), moving from polar and charged residues in FNfn10 and FN3con-α-lys to hydrophobic and polar residues in FNfn10-α-lys (Fig. 3A, 3B and 3C). The flip and shift is particularly important as M50 from FNfn10-α-lys packs between two tryptophan residues (W62 and W63) from lysozyme (Fig. S1). As the FN3con-α-lys structure does not display this conformational change, the position of P48 in FN3con-α-lys may impose a steric clash with W62 from lysozyme when modeled into the same binding site, thereby preventing tight complex formation (Fig. S2).

Furthermore, when the entire framework-binding region of FNfn10-α-lys is compared to FN3con-α-lys, it becomes clear that the physiochemical incompatibilities with lysozyme binding extend beyond strand D. In particular, Y31 in FNfn10-α-lys plays a pivotal role in packing and forming hydrogen bonds with D48, S50 and N59 from lysozyme (Fig. 3D and S2). The analogous residue in FN3con-α-lys is G30, which not only results in a loss of three hydrogen bonds, but may also leave a large unfilled cavity in the binding interface (Fig. 3E and S2). In close proximity to G30 is R32 and R71, which belong to a stability-enhancing electrostatic mesh on the surface of FN3con [137] that were retained in the grafting process. As these are long and charged residues, we therefore hypothesize that residues R32 and R71 in FN3con-α-lys (G33 and Y74 in FNfn10-α-lys) may impose steric clashes with lysozyme, as suggested by modeling the complex, thereby further restricting tight complex formation (Fig. 3D, 3E and S2).

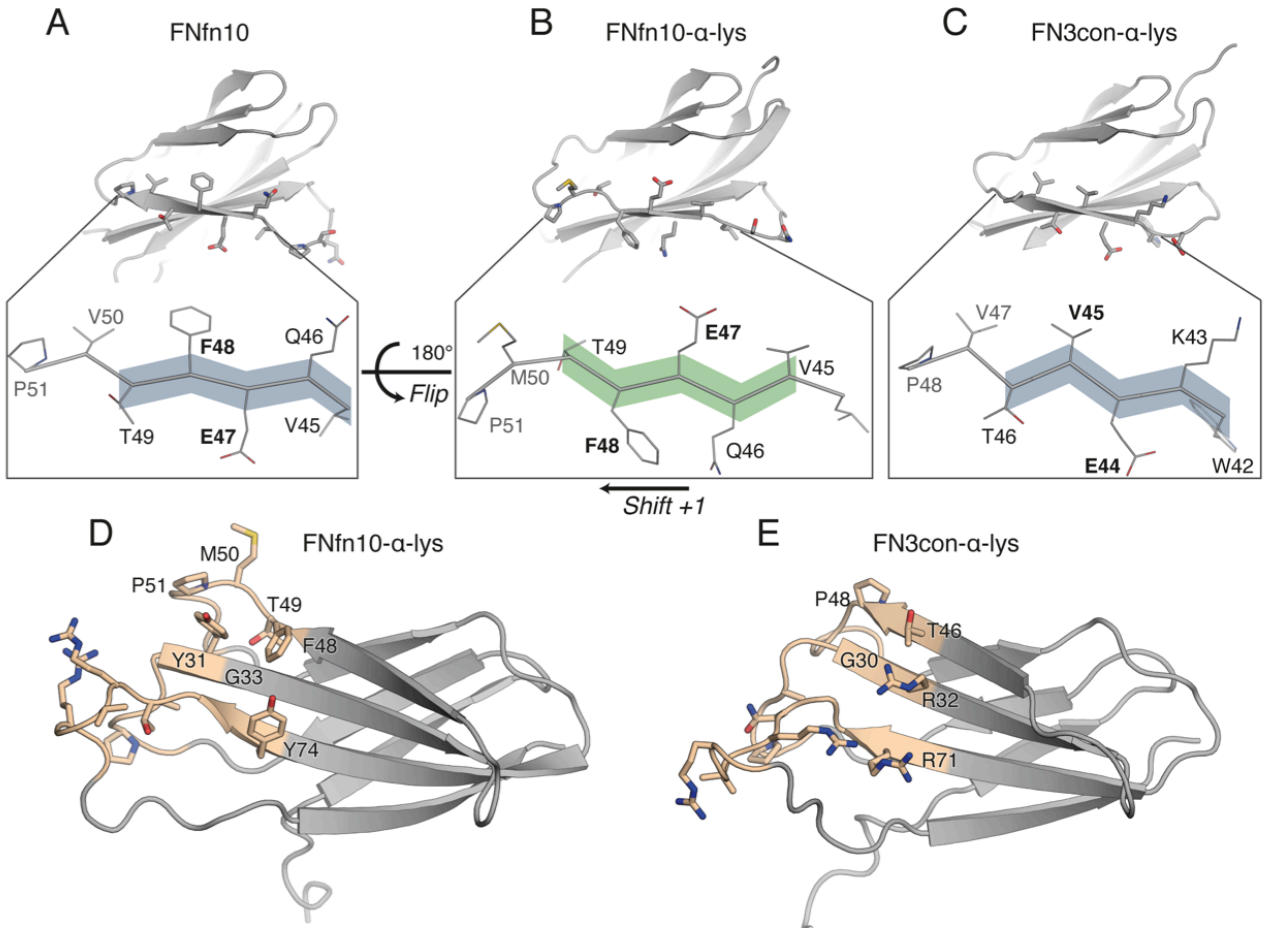


Figure 3. Structural comparison between FNfn10-α-lys and FN3con-α-lys reveals framework incompatibilities that likely prevent tight complex formation. A conformational change is observed between FNfn10 **(A)** and FNfn10-α-lys **(B)** resulting in a 180° flip and +1 register shift of strand D that is also lacking in the unbound FN3con-α-lys crystal structure **(C)**. Differences in framework residues of the lysozyme-binding interface (tan region) between FNfn10-α-lys **(D)** and FN3con-α-lys **(E)** highlight the potential for cavity formation due to the lack of Y31 (G30 in FN3con-α-lys) and steric clashes as a result of R32 and R71 in FN3con-α-lys. These characteristics may impact the formation of a tight binding interface.

Rational redesign of FN3con-α-lys restores binding

Given our high-resolution structural insights for the FNfn10-α-lys-lysozyme interface and direct comparison with a structure of the low-affinity FN3con-α-lys graft, we sought to rationally redesign FN3con-α-lys. Through redesign, we aimed to restore binding by addressing the conformational change seen in strand D, and the physiochemical incompatibilities of framework residues in the lysozyme-binding interface (Fig. 4A and 4B). We have previously shown FN3con to be extremely rigid due to an optimized hydrophobic core [137]. For this reason, we did not expect to be able to

easily engineer strand D to mimic the conformation seen in the FNfn10-α-lys structure (Fig. 3B and 3C). Instead, we decided to retain the hydrophobic core residues of FN3con and simply mutate the surface residues of strand D to match that of FNfn10-α-lys. This involved the mutation of E44Q, T46F, V47T and the insertion of a methionine residue between V47T and P48 (Fig. 4C). To mimic the remaining FNfn10-α-lys paratope, we made three more mutations (G30Y, R32G and R71Y), which predominantly removed charged residues from FN3con's stability enhancing electrostatic mesh [137]. We did so with the hypothesis that this will restore packing and remove the predicted steric clashes, thereby allowing for tighter complex formation. In total, seven mutations (Table S2) were made to FN3con-α-lys, producing the variant FN3con-α-lys.v2 (Fig. 4C and Data S1). FN3con-α-lys.v2 expressed as a soluble monomer to approximately 50 mg/L of culture and was soluble when concentrated to 10 mg/ml for over 30 days.

Assessment of binding to lysozyme was performed by SEC (Fig. 4D) and SPR, revealing excellent signal separation at low concentrations of lysozyme and a substantially slowed dissociation-rate that is more similar to FNfn10-α-lys (Fig. 4E and 1E). As lysozyme binding was successfully restored, these results support our hypothesis that steric hindrance and the presence of a cavity were limiting tight complex formation in the FN3con-α-lys graft. Subsequent characterization of the thermodynamic stability of FN3con-α-lys.v2 revealed a T_m of 87±2 °C (Fig. 4F). Although thermal denaturation is not completely reversible, as demonstrated by the small loss of CD signal on cooling (Fig. 4F), the remaining and exceptionally high T_m of 87±2 °C positions FN3con-α-lys.v2 well above the T_m of both FNfn10-α-lys (43±2 °C) and FNfn10 (84 °C). Together, these results highlight the remarkable capacity of consensus design to generate highly stable and mutationally tolerant scaffolds. As restoration of lysozyme binding required the degeneration of known stability enhancing features in FN3con, future directed evolution studies could be designed with this in mind, thereby retaining a greater degree of thermodynamic stability and favorable biophysical properties for the same level of function.

Table I. Data collection and refinement statistics[a].

Data collection	FNfn10-α-lys complex	FN3con-α-lys
Wavelength (Å)	0.9537	0.9537
Space group	P 2_1 2_1 2_1	P 1 2_1 1
Unit cell dimensions (Å)	54.857, 87.725, 100.895, 90.00, 90.00, 90.00	50.220, 71.229, 126.141, 90.00, 90.43, 90.00
Resolution (Å)	2.54	2.46
Number of measured reflections	111926 (12664)	117587 (13558)
Number of unique reflections	16719 (1940)	32118 (3627)
Completeness (%)	99.0 (96.0)	98.8 (98.8)
Redundancy	6.7 (6.5)	3.7 (3.7)
R_{pim}	0.139 (0.728)	0.062 (0.949)
<I/σI>	8.90 (2.13)	11.60 (1.70)
Structure refinement		
Number of reflections	16665 (1579)	32099 (3181)
Number of protein atoms	3430	4365
Number of water molecules	52	70
Number of ligands	0	11
R_{work} (%)	0.2176	0.2488
R_{free} (5% of data) (%)	0.2499	0.2811
CC1/2	0.993 (0.685)	0.997 (0.408)
RMSD bond lengths (Å)	0.003	0.012
RMSD bond angles (°)	0.56	1.53
Average B-factor ($Å^2$)	37.52	59.13
Protein	37.75	59.23
Solvent	22.68	44.64
Ramachandran		
Favoured (%)	99	95
Outliers (%)	0	0.71
MolProbity score	0.96, 100^{th} percentile[b] (N=6642, 2.535Å ± 0.25Å)	2.21, 89^{th} percentile[b] (N=6959, 2.46Å ± 0.25Å)
PDB ID	5J7C	5J7K

[a]Statistics for the highest-resolution shell are shown in parentheses.

[b]100^{th} percentile is the best among structures of comparable resolution; 0^{th} percentile is the worst.

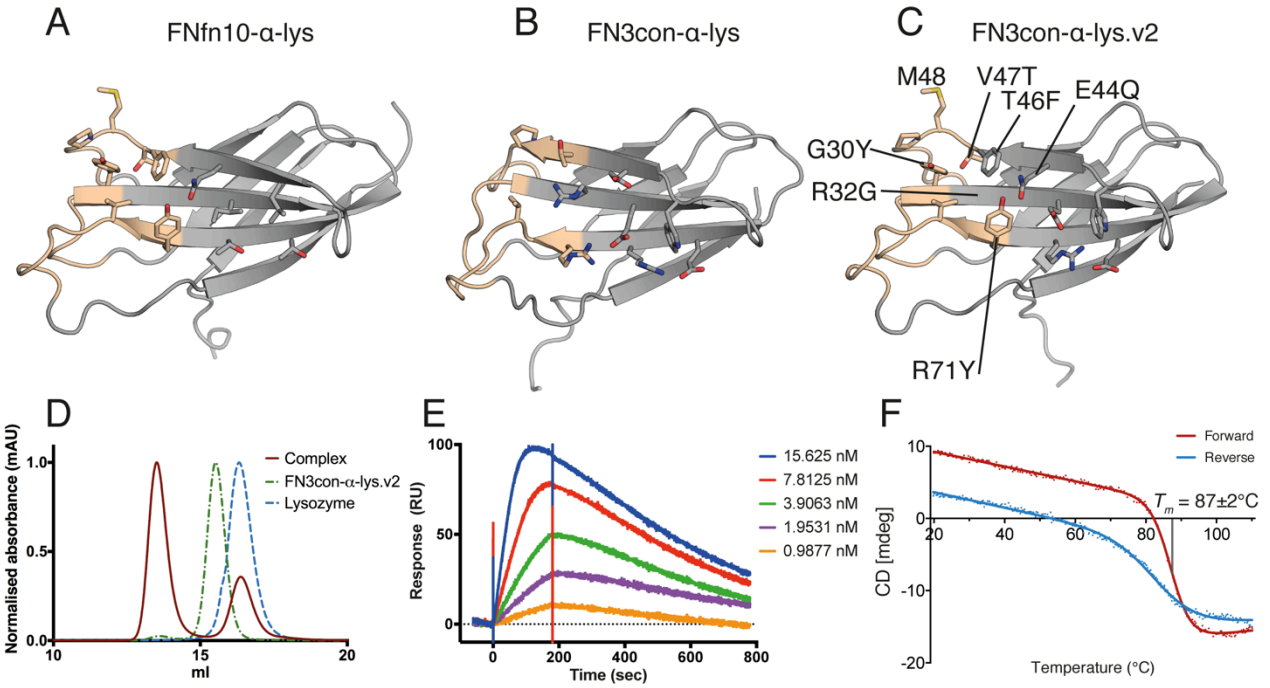


Figure 4. Framework residues in the lysozyme-binding interface of FN3con-α-lys were redesigned by alignment of FNfn10-α-lys and the FN3con-α-lys crystal structures. Redesign restored binding at the cost of thermodynamic stability. **(A)** The crystal structure of FNfn10-α-lys showing the paratope surface residues (tan) and surrounds; **(B)** The composite crystal structure of FN3con-α-lys showing the paratope surface residues (tan) and surrounds; **(C)** A homology model of the redesigned FN3con-α-lys.v2 based on FNfn10-α-lys showing the redesigned binding interface residues (tan); **(D)** SEC complex formation shift of FN3con-α-lys.v2; **(E)** Representative SPR sensograms of FN3con-α-lys.v2 with lysozyme titrations; **(F)** Variable temperature CD melt at 222 nm of FN3con-α-lys.v2 showing a T_m of 87±2 °C and incomplete reversible folding.

Conclusions

Non-antibody scaffolds are attractive alternatives to monoclonal antibodies, but experience stability-function trade-offs after selection, and are thus only marginally more stable than their antibody counterparts. This study sought to circumvent the stability-function trade-off in the FN3 domain. Crystallographic structure determination provided key structural insight into the binding between FNfn10-α-lys and lysozyme, which allowed for successful transfer of binding affinity onto the FN3con scaffold by a combination of loop grafting and rational design. Biophysical characterization subsequently showed FN3con to exhibit a smaller loss in thermodynamic stability after the engineering of function. These results therefore highlight that the effect of loop mutagenesis and stability-function trade-off is not equivalent across homologous proteins, with the FN3con scaffold imparting a greater resistance to destabilizing loop sequences (ΔT_m of 41 °C between FNfn10 and FNfn10-α-lys and ΔT_m of ~20 °C between FN3con and FN3con-α-lys.v2). This study underlines the utility of consensus design for the generation of highly stable and mutationally-tolerant scaffolds that may be suited to further protein engineering and directed evolution studies.

Methods

Loop grafting and homology modelling

Loop grafting was performed in PyMol V. 1.5.0.4 using structural alignments of FN3con (PDB: 4U3H) and an FNfn10-α-lys homology model based on FNfn10 (PDB: 1FNF). Loop boundaries were identified and grafting was conducted on the FN3con sequence. A homology model of FN3con-α-lys was generated based on FN3con (PDB: 4U3H) using Modeller V. 9.12 [162]. In each instance, 50 models were built and the lowest DOPE (Discrete Optimized Protein Energy) scoring model was selected for further analysis. Modeller was also used to complete the FN3con-α-lys crystal structure, and generate a model of FN3con-α-lys.v2.

Protein expression and purification

Genes encoding FNfn10-α-lys, FN3con-α-lys and FN3con-α-lys.v2 were chemically synthesized and provided in a pD444-CH (C-terminal 6x His tag, ampicillin resistance) vector by DNA2.0. Competent C41 *E. coli* cells were transformed with the resulting plasmids for expression. A single colony from each transformation was picked and grown overnight at 37°C in 100 ml of 2xYT (16.0 g/L tryptone, 10.0 g/L yeast extract, 5.0 g/L NaCl) media containing 100 µg/ml of ampicillin. These cultures were then used to seed 1 L of 2xYT media. Cultures were induced at an OD_{600} of 0.9 with IPTG (0.5 mM final concentration), and grown for a further 4 hours at 37°C. The cells were harvested by centrifugation. FN3con-α-lys and FN3con-α-lys.v2 had their cell pellets resuspended in 5 ml/g of native lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0) and were lysed by sonication. Cell debris was removed by centrifugation and recombinant protein was isolated from the supernatant by nickel affinity chromatography using loose NiNTA resin (Sigma). Protein eluted from NiNTA resin was filtered and then loaded onto a size exclusion column (Superdex 75 16/60, GE Healthcare) equilibrated in either PBS (140 mM NaCl, 2.7 mM KCl, 10 mM PO_4^{3-}, pH 7.4) for biophysical characterization or TBS (50 mM Tris, 200 mM NaCl, pH 7.4) for protein crystallography. Protein concentration was determined by Nanodrop ND-1000 (ThermoFisher) and protein was stored at 4°C until use (biophysical characterization) or used immediately (protein crystallography).

Refolding and purification of FNfn10-α-lys

FNfn10-α-lys expressed insolubly under the same conditions as FN3con-α-lys and FN3con-α-lys.v2 (above) to approximately 50 mg/ml. The culture was harvested and resuspended in 5 ml/g of native lysis buffer and lysed by sonication. The supernatant was discarded and the insoluble fraction resolublised in denaturing buffer (8 M urea, 50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0). Remaining insoluble material was cleared by centrifugation and filtration with a 0.8 µm syringe filter (Merk-Milipore), then loaded onto loose NiNTA resin (Sigma) that was equilibrated in denaturating lysis buffer. The resin was washed with 50 ml of denaturating buffer

with 20 mM imidazole and eluted in denaturing buffer with 300 mM imidazole. Eluted and denatured protein was diluted to ~1 mg/ml and refolded by overnight dialysis in 4 L of TBS. Visible precipitate was present if the protein concentration of the dialysis bag is >1mg/ml. Protein aggregate is readily separated by centrifugation and filtration with a 0.22 µm syringe filter (Merk-Milipore). Filtered refolded material is then loaded onto a size exclusion column (Superdex 75 16/60, GE Healthcare) equilibrated in either PBS for biophysical characterization or TBS for protein crystallography.

Circular dichroism thermal melts

Thermal stability of purified FNfn10-α-lys, FN3con-α-lys and FN3con-α-lys.v2 was measured by circular dichroism (CD). CD measurements were performed using a Jasco 815 spectropolarimeter with 0.2 mg/ml protein in PBS used in a 0.1 cm path length quartz cuvette. Thermal denaturation was measured by observing signal changes at 222 nm during heating at a rate of 1 °C/min. The melting temperature (T_m) was obtained by fitting to a sigmoidal dose-response (variable slope) equation.

Binding measurements

The binding affinities of FNfn10-α-lys, FN3con-α-lys and FN3con-α-lys.v2 were measured using surface plasmon resonance (BIAcore T-100, GE Healthcare). FN3 domains were immobilized (90 µl at 5 µl/min) on a NiNTA sensor chip at a concentration of 2.5 µg/ml. A 1:2 dilution series from 125 nM to 0.936 nM of lysozyme (Sigma) was injected at a flow rate of 50 µl/min. The NiNTA sensor chip was regenerated and the FN3 domain immobilized after each dilution of lysozyme. HBS-EP$^+$ (10 mM HEPES, 200 mM NaCl, 3mM EDTA, 0.05% (v/v) Tween 20, pH 7.4) was used as the running buffer and both FN3 domains and lysozyme were prepared in HBP-EP$^+$ with the addition of 12 mg/ml CM-Dextran to remove the non-specific binding of lysozyme.

Crystallisation, X-ray data collection, structure determination and refinement

All crystals were grown using the hanging drop vapor diffusion method, with 1:1 (v/v) ratio of protein to mother liquor (500 µl well volume). For the FNfn10-α-lys-lysozyme complex, purified FNfn10-α-lys at a concentration of 0.5 mg/ml was mixed in a roughly 1:1 molar ratio with hen egg-white lysozyme (HEL) (Sigma Aldrich) for a total volume of 5 ml, incubated for 30 minutes at room temperature, then purified by size exclusion chromatography (Superdex 75 16/60) in 20 mM Tris, 200 mM NaCl, pH 7.4. Peak fractions corresponding to the size of a FNfn10-α-lys-lysozyme complex were concentrated to 8.56 mg/ml. Long thin needle like crystals formed in 10% PEG 6000, 0.1 M Bicine, pH 8.8 within 4 hours. A single crystal was extracted and cryoprotected in 20% ethylene glycol, 12% PEG 6000, 0.1M Bicine, pH 8.8 prior to collection.

For the unbound FN3con-α-lys loop graft, purified protein was concentrated to 29 mg/ml. Large plates were formed in 5% glycerol, 10% 2-propanol, 0.2 M zinc acetate, 0.1 M sodium cacodylate, pH 6.0 within 2 days. Crystals were dehydrated by stepwise equilibration of the crystallisation drops over wells with progressively increasing concentration of glycerol and decreasing concentration of 2-propanol, with 24 hours between transfers. The final reservoir solution contained 15% glycerol, 0.2 M zinc acetate, 0.1 M sodium cacodylate pH 6.0. Crystals were subjected to a final soak in 20% glycerol, 0.2 M zinc acetate, 0.1 M sodium cacodylate pH 6.0 for 20 minutes prior to data collection.

Data for both crystals were collected at 100 K at the Australian Synchrotron micro crystallography MX2 beamline. FNfn10-α-lys-lysozyme complex crystals diffracted to 2.54 Å resolution, and given the small size of these crystals, radiation damage became a significant problem. This was mitigated by collecting 4x 45° wedges along the crystal and later merging the wedges together. Crystals for the unbound FN3con-α-lys diffracted to 2.46 Å. Diffraction images were processed using iMosflm [164] and Aimless from the CCP4 suite [171]. Each dataset was initially processed in *P*1 and Laue group determination was achieved using Pointless within

Aimless [171]. Datasets were reintegrated, scaled and merged in their respective space-group and 5% of each dataset was flagged for calculation of R_{Free}, with neither a sigma nor a low-resolution cut-off applied to any dataset. A summary of statistics is provided in Table I.

Structure determination proceeded using molecular replacement and the program PHASER [165]. A search model for the FNfn10-α-lys-lysozyme complex was constructed from the crystal structure of FNfn10 (PDB: 1FNF) by removing solvent molecules and loops that lack homology, and from the crystal structure of hen egg white lysozyme (PDB: 4Z98) that had solvent molecules, acetate ion and hydrogen atoms removed. PHASER identified two complexes in the asymmetric unit, for a total of two HEL and two FNfn10-α-lys molecules. A single clear peak for both the rotation and translation functions was evident and the molecules packed well within the asymmetric unit. Model building was conducted using COOT [168] and refined using Buster [211] and Phenix [167].

The FN3con-α-lys crystal was initially identified to be of the space group $P22_12_1$, with pseudo-translational non-crystallographic symmetry (NCS). A molecular replacement search model for FN3con-α-lys was constructed from the crystal structure of FN3con (PDB: 4U3H) by removing solvent molecules and loops that lack homology. PHASER identified four molecules in the asymmetric unit, however, electron density failed to align well with two of the four molecules, and refinement stalled with an R_{Free} above 0.4. We found that by lowering the symmetry of the space group to $P2_1$ (P 1 2_1 1) the crystal also exhibited twinning (twinned fraction of 0.48), which was not detected in P1 or $P22_12_1$ (twin law of *h, -k, -l*). With these corrections, molecular replacement was repeated with PHASER, identifying eight molecules in the asymmetric unit (two tetramers). Initial electron density maps were significantly improved, with all eight chains fitting well. Model building was conducted in COOT [168] and refinement using Buster [211] and Phenix [167].

Structural analysis

Interface analysis was performed using the PDBePISA webserver [210]. Shape complementarity calculations were performed using the Sc program [209] from the CCP4 software suite [171].

Author contributions

BTP and AMB designed the study. BTP performed loop grafting and design. BTP and MRH performed protein expression, refolding and purification. BTP performed the CD thermal melts and biophysical characterization. BTP, ND and SM performed the crystallography, data collection and structural determination. BTP and PJC performed the biacore experiments. BTP generated figures. BTP, DEH, MRH, RVL, DC and AMB wrote the paper.

Acknowledgements

We thank Harry Powell, Herman Schreuder, Mirko Velic, Blake Riley, Remy Robert, Andrew Ellisdon and the CCP4BB for helpful discussions and advice during the course of this research. We thank the Australian Synchrotron for beam-time and technical assistance. This work was supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (www.massive.org.au). We acknowledge the Monash Macromolecular Crystallization Facility. AMB and DC hold fellowships from the National Health and Medical Research Council (1022688, 1050146).

Supplementary information

Table S1. Interfacing residues of FNfn10-α-lys as identified by the PDBePISA webserver [210], highlighting (grey) those that exhibit hydrogen bonding (H) or salt bridge (S) interactions with lysozyme.

Residue	Location	Interaction	Buried surface area (%)
TYR26	B/C loop		10
PRO27	B/C loop	H	40
TRP28	B/C loop		50
ALA29	B/C loop	H	90
THR30	B/C loop		90
TYR31	Framework	H	100
PHE48	Framework		70
THR49	Framework		30
MET50	Framework		70
PRO51	Framework		100
GLY52	D/E loop		100
ASP53	D/E loop		30
LEU54	D/E loop		50
TYR74	Framework		30
ALA75	Framework		20
VAL76	Framework		100
THR77	Framework	H	100
ARG78	F/G loop	HS	100
VAL79	F/G loop	H	40
ARG81	F/G loop	HS	80
THR82	F/G loop		10

Table S2. Mutations introduced into FN3con-α-lys to restore binding.

Mutations
G30Y
R32G
E44Q
T46F
V47T
Insertion: M47-48
R71Y

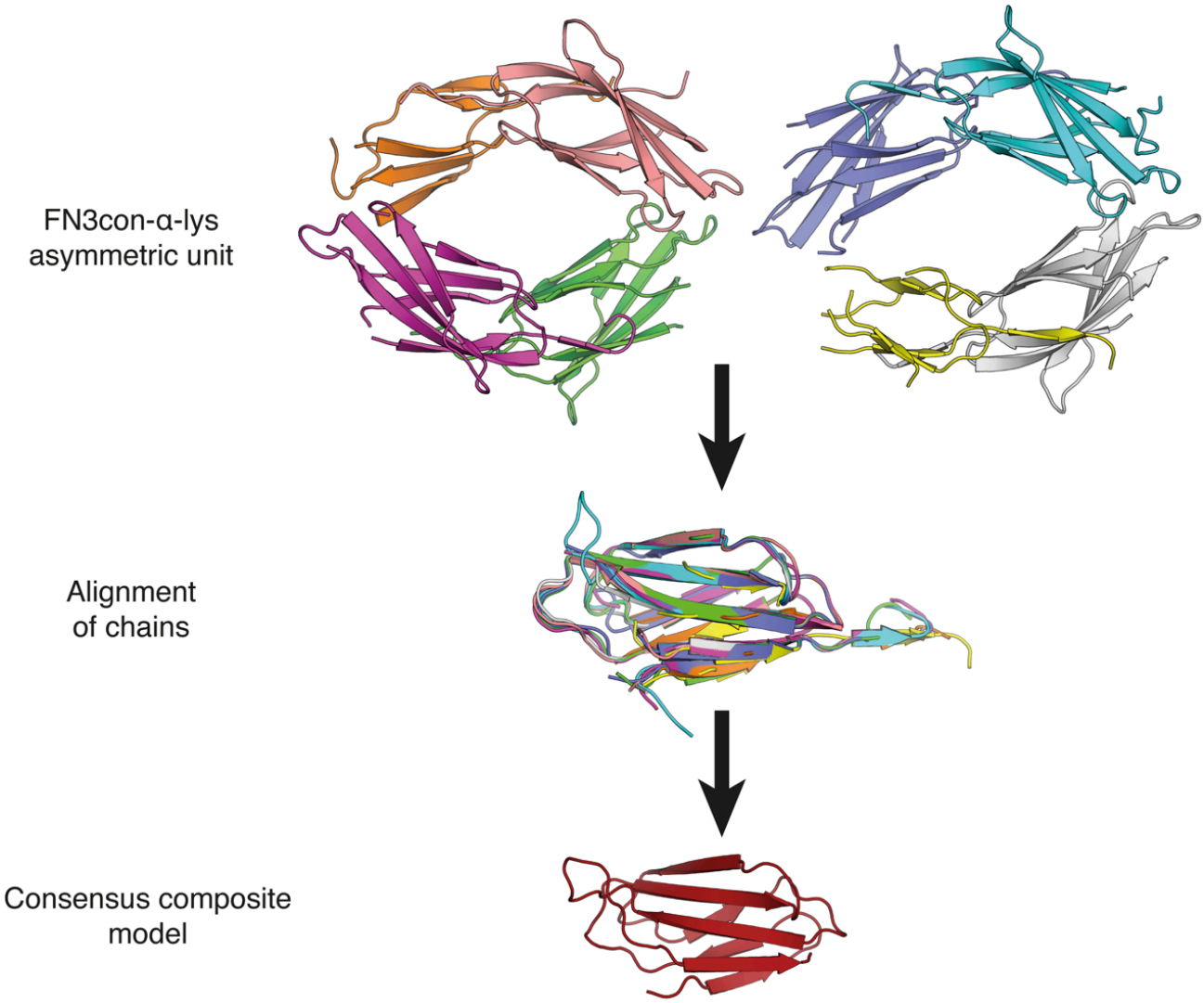


Figure S1. The asymmetric unit of FN3con-α-lys shows 8 molecules in two tetrameric groups of variable completeness with domain swapping of the C-terminal 6x His tag. An alignment of all 8 molecules results in a completed structure with the exception of the F/G loop, which was not observed in any monomer. Homology modeling was then implemented to construct a composite model using all 8 monomers that is used in all subsequent structural analysis.

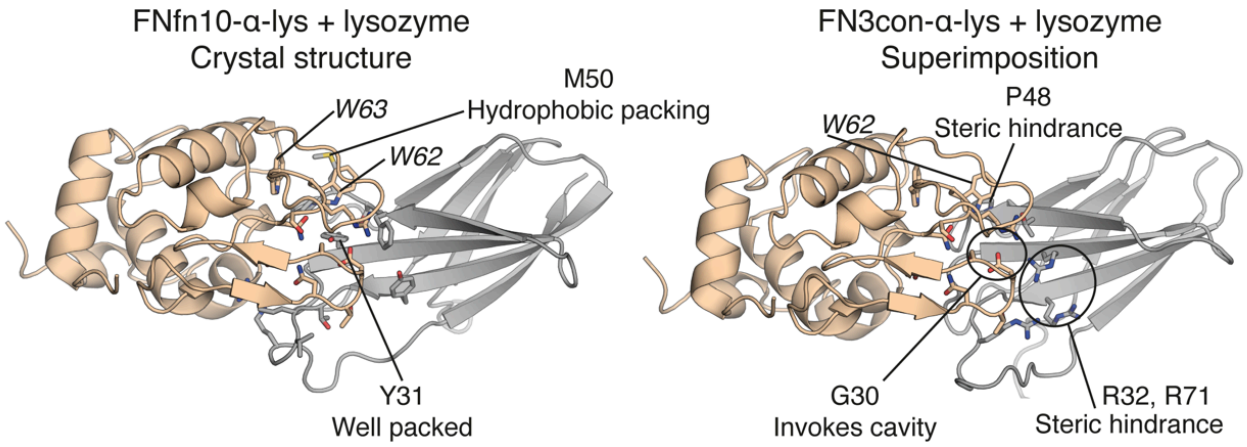


Figure S2. The complex crystal structure of FNfn10-α-lys-lysozyme (left) shows two major sites of packing (Y31 and M50). When the completed model of FN3con-α-lys is superimposed into the lysozyme-binding site (right), the analogous residue to Y31, G30, results in presence of a cavity. Further, the positions of P48, R32 and R71 in FN3con-α-lys may result in steric clashes with W62 and the region surrounding D48 in lysozyme. We hypothesise that the combination of these differences restricts tight and full complex formation of FN3con-α-lys with lysozyme.

Chapter 4

Exploring the evolvability of FN3con with yeast surface display

Summary

In this chapter, I hypothesized that the superior structural and biophysical properties of FN3con discussed in the previous two chapters improve its evolvability as a binding scaffold. To test this hypothesis, I randomised the solvent exposed B/C, D/E and F/G loops of FN3con through the construction of a yeast surface display (YSD) library and assessed binding affinity for lysozyme with fluorescence activated cell sorting (FACS). Data revealed that the naïve FN3con library contains clones that bind lysozyme with approximately low nanomolar affinity, without the need to conduct affinity maturation studies. These results support the evolvability hypothesis, therefore suggesting that FN3con may be able to tolerate and display a larger sequence space than other FN3 domains. Although the exact mechanism for improved evolvability has yet to be elucidated, it is highly likely that the myriad of stability enhancing features in FN3con provides a compensatory capacity to absorb unfavourable interactions without a loss of structure. This work serves to highlight consensus design as a means to generating robust protein molecules that are amenable to directed evolution and as a promising foundation for future biophysical characterisation and isolation of binders against practical targets.

Introduction

Chapter 3 provided insight into the mutational tolerance of FN3con as rational loop grafting was successful in transferring lysozyme binding without a significant loss in thermodynamic stability. With this in mind, I hypothesise that a randomized loop library based on the FN3con scaffold will provide the capacity to display a greater combination of amino acid sequences (sequence space) on its loops, compared to the wild type 10^{th} type III domain from human fibronectin (FNfn10). This chapter therefore explores the evolvability and applicability of FN3con as a binding scaffold by directed evolution with yeast surface display (YSD).

YSD is an effective system for displaying a target protein on the surface of *Saccharomyces cerevisiae* yeast cells, where the yeast cells function as a compartmentalized linker between the phenotype and genotype which can self-replicate [212]. To display a target protein on the surface of yeast, Boder and Wittrup commandeered the cell surface receptor *a*-agglutinin, which is a two-subunit, disulfide-linked glycoprotein consisting of Aga1p and Aga2p [212]. In this construct, the 725 residue Aga1p subunit anchors to the yeast cell wall via β-glucan covalent linkage, whilst the 69 residue Aga2p subunit is linked to the Aga1p subunit by two disulfide bonds [212,213]. By linking a target protein to the C-terminus of Aga2p, it is possible to display approximately 10^5 molecules on the surface of each yeast cell (Fig. 4.1) [214]. To enable display of the target protein, Wittrup and colleagues have developed several fusion construct plasmids [212,214], with the one of the most readily available being pCTcon2 [215]. The pCTcon2 plasmid is designed for use in the yeast strain EBY100, which is deficient in the machinery to synthesize the amino acid tryptophan and contains the Aga1 gene in the yeast genome [214]. The pCTcon2 plasmid then encodes for the gene TRP1, which is important for tryptophan synthesis, allowing for selection of transformed clones in minimal growth media. The plasmid also encodes for the Aga2p protein fused to a hemagglutinin (HA) tag, then a cloning site for the protein of interest (flanked by NheI and BamHI restriction sites), followed by a C-terminal c-Myc tag. Both Aga1p and Aga2p are under the control of a galactose-inducible promoter. Switching the yeast from glucose-rich to galactose-rich media

will induce display of the target protein. Therefore, the resulting surface-displayed target protein, which has two epitope tags, can be used for immunocytochemical detection.

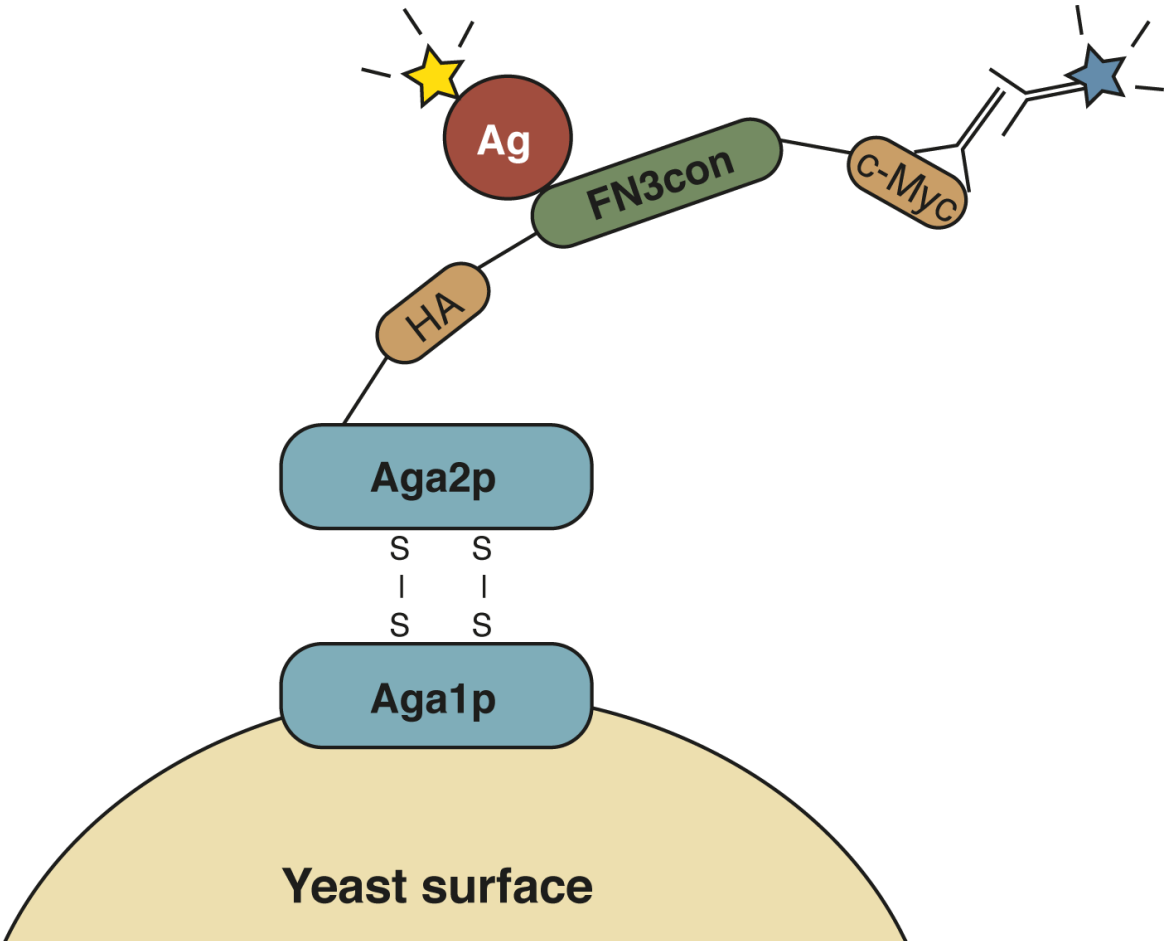


Figure 4.1. A schematic of FN3con displayed on the surface of yeast. FN3con resides as a C-terminal fusion of the Aga2p protein and two detection tags: hemagglutinin (HA) epitope tag at the N-terminus and a c-Myc epitope tag at the C-terminus. Aga2p forms disulfide bonds with Aga1p, which is anchored to the cell wall via β-glucan linkage. Binding of a biotinylated antigen (Ag) is detected by fluorophore-conjugated neutravidin. To ensure that FN3con does not exhibit any stop codons or frameshift mutations, full length FN3con is detected with a primary mouse anti-c-Myc antibody and a goat anti-mouse H+L fluorophore-conjugated secondary antibody.

Yeast surface display offers several advantages for directed evolution such as eukaryotic post-translational modification and in cell library assembly through homologous recombination, a mechanism that allows for the assembly of linear DNA fragments that share overlapping homology [216-218]. YSD further enables quantitative screening through the use of fluorescence-activated

cell sorting (FACS), allowing for equilibrium binding of the sample to be observed directly during the screening process [216].

The directed evolution process begins with the generation of a variant library for a desired protein. In this case, a library was generated for FN3con; yeast cells were transformed and surface display induced. Cells are then labelled with an anti-cMyc antibody (to detect full length FN3) and biotinylated antigen (in this case, lysozyme). Both anti-cMyc and biotinylated lysozyme are labelled with respective secondary fluorophore conjugates of different and non-overlapping colours, enabling detection of FN3 display and antigen binding by flow cytometry. These cells can then be sorted based on their antigen binding. The pCTcon2 plasmid can then be extracted from the sorted yeast that had successfully bound antigen, thereby allowing for characterisation of individual clones or further affinity maturation. Affinity maturation repeats the process of generating diversity from selected clones by utilising a mutagenesis regime such as error prone PCR (as discussed in Chapter 1).

This chapter describes the design and application of an FN3con yeast surface display library for the generation of binders against lysozyme.

Results and Discussion

Design of a NNS FN3con variant library for yeast surface display

The design of the FN3con variant library was inspired by several examples in the literature [201,212,214,216], but modified to take advantage of the recent reduction in costs for DNA synthesis. Accordingly, genes were commercially synthesized according to the following parameters. The variant library incorporated NNS degenerate codons (where N = A/C/G/T and S = C/G) into the B/C, D/E and F/G loops of FN3con, for 7, 5 and 8 amino acids, respectively (listed as X in the following protein sequence). The NNS codon was chosen as it still encodes all 20 amino acids with a 1/32 chance of encoding a stop codon, as opposed to a 3/64 chance for the NNN codon.

FN3con.NNS
MPSPPGNLRVTDVTSTSVTLSWE**XXXXXXX**GYRVEYREAGGEWKEVTVP**XXXXX**SYTVTGLKPGTE
YEFRVRA**XXXXXXXX**PSSVSVTT

Loop boundaries were selected based on designs reported in the literature and also to avoid disrupting key stability enhancing features of FN3con, such as the hydrophobic core, tyrosine corners and electrostatic mesh [137]. Although the entire library could have been synthesized in an assembled form, this particular design would make subsequent affinity maturation steps more difficult, as there would not be a simple way to target the loop regions. Instead, a modular design of three NNS loop cassettes, one for each loop, is assembled by yeast homologous recombination into a modified FN3 scaffold. This modified scaffold lacks the B/C, D/E and F/G loops (*FN3con.delta.loops*), and the missing loops have been replaced with a SmaI restriction site. The *FN3con.delta.loops* construct is cloned into the pCTcon2 vector, creating the pCTcon2-FN3con.delta.loops vector that is amenable to simple amplification and library generation (Fig. 4.2). During assembly, the pCTcon2-FN3con.delta.loops vector is linearized by blunt end digestion with SmaI, thus removing overhanging homology between linear fragments and restricting the

probability of reassembly (Fig. 4.2). As the NNS loop cassettes were chemically synthesized and share 50bp of 5' and 3' homology with the FN3con.delta.loop scaffold, their addition provides a means for recircularisation of the pCTcon2-FN3con.delta.loops vector by yeast homologous recombination and subsequent survival in selective media.

Assembly of the FN3con NNS library

Although a powerful assembly technique, yeast homologous recombination is an inherently inefficient process that requires large amounts of DNA (60 µg of insert and 40 µg of vector) for the generation of a library with a diversity of 5 x 10^7 variants [216-218]. I therefore thought it would be more efficient to sub-clone the FN3con.delta.loops construct from the manufacturer-supplied pUC57 vector into pCTcon2 to create pCTcon2-FN3con.delta.loops, which can be readily amplified in *E. coli* to meet the large DNA requirements. Assembly therefore proceeds by linearization of pCTcon2-FN3con.delta.loops vector with SmaI, heat inactivation of the SmaI enzyme, the addition of the NNS loops and transformation into competent EBY100 cells by electroporation. As the NNS loops were ordered as three ~120 bp fragments at 10 µg each, it was evident that I would not have a sufficient quantity for direct transformation. With an expected diversity of 10^{11} variants per loop and a resulting in a copy number of ~8.1x10^{13} molecules, I calculated that 130 ng of each loop (10x diversity) could be amplified to 100 µg by PCR.

As I was able to generate sufficiently large quantities of loop and vector DNA, I decided to assemble a larger library and subsequently scaled the transformations to match. I based the transformation protocol from Gera et al., [216], which indicated 4 µg of vector to 6 µg insert, per 50 ml of EBY100 overnight culture. I therefore scaled the transformation protocol 20 fold, to 1 L of EBY100 and 20 individual transformations. Transformed cells were propagated by overnight growth and aliquoted into cryotubes with 100-fold excess of library diversity for long-term storage.

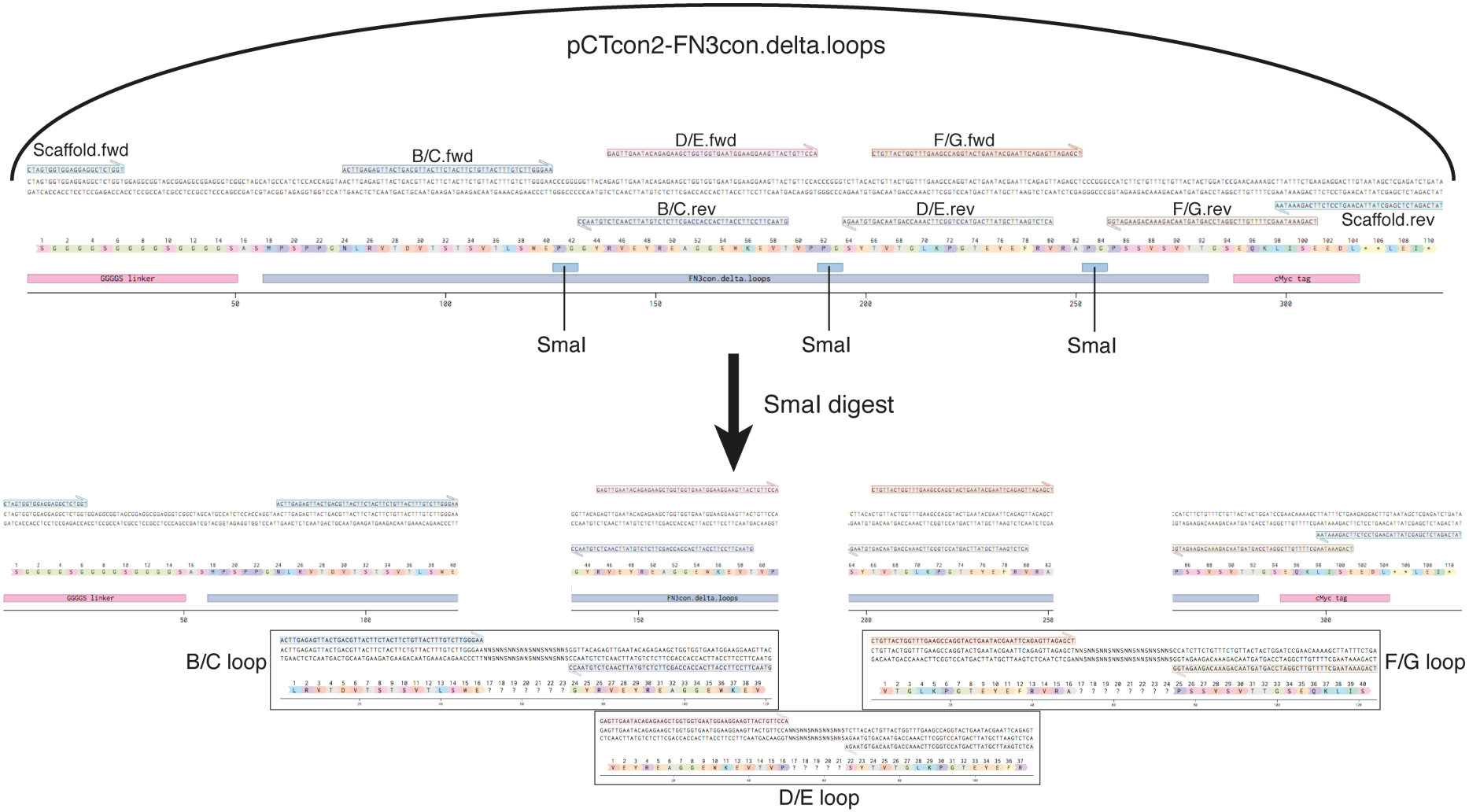


Figure 4.2. A schematic of the FN3con NNS loop library. The top half of the figure shows the pCTcon2-FN3con.delta.loops vector with SmaI restriction sites replacing the B/C, D/E and F/G loops. The bottom half of the figure shows the individual loop cassettes, which exhibit 50 bp homology with

Library transformation yielded approximately 2.3 × 10^8 yeast transformants, serial dilution. Fourteen (50%) of 28 clones sequenced matched the library de (39%) contained frameshift mutations, and three (10%) contained incc recombination. NNS diversification of the loops yields stop codons in approxima $(1-(31/32)^{20})$. Thus, 26.5% [14/28×(1–0.47)] of all transformants should displ domains. Therefore, I expect to see approximately 6.09 x 10^7 full-length clone the full library size is approximately 1.3 x 10^{36}, the functional library size is a gr of the possible sequence space. However, this may not necessarily be a pr sized libraries using the FNfn10 scaffold, it is likely that a much smaller subset will give rise to a fully folded protein due to the inherent stability-function trac Chapter 3. Conversely, I hypothesise that the increased stability of FN3con a library produced will tolerate a much larger sequence space; therefore enablin of novel epitopes without the need for extensive affinity maturation.

Library screening against lysozyme

As an initial proof of concept for this hypothesis, I screened and sorted FN3 hen egg white lysozyme. This was done using a two-step procedure involvi screening followed by FACS for analysis and cell sorting. In the first step, a fro thawed out, proliferated and induced. The induced cells (10^8 cells) were incuba lysozyme and then coated with micron-sized anti-biotin magnetic beads complexes were subsequently isolated by use of a powerful magnet. This proce with even low affinity for the target due to the avid yeast-bead interaction, throughput and highly effective initial screening strategy [216]. Flow cytom magnetic bead enrichment was conducted, titrating of lysozyme from 205 nM Magnetic enrichment reveals a significant population of lysozyme binding cells, background (no lysozyme) observed across the entire range titration range 2 (Fig. 4.3).

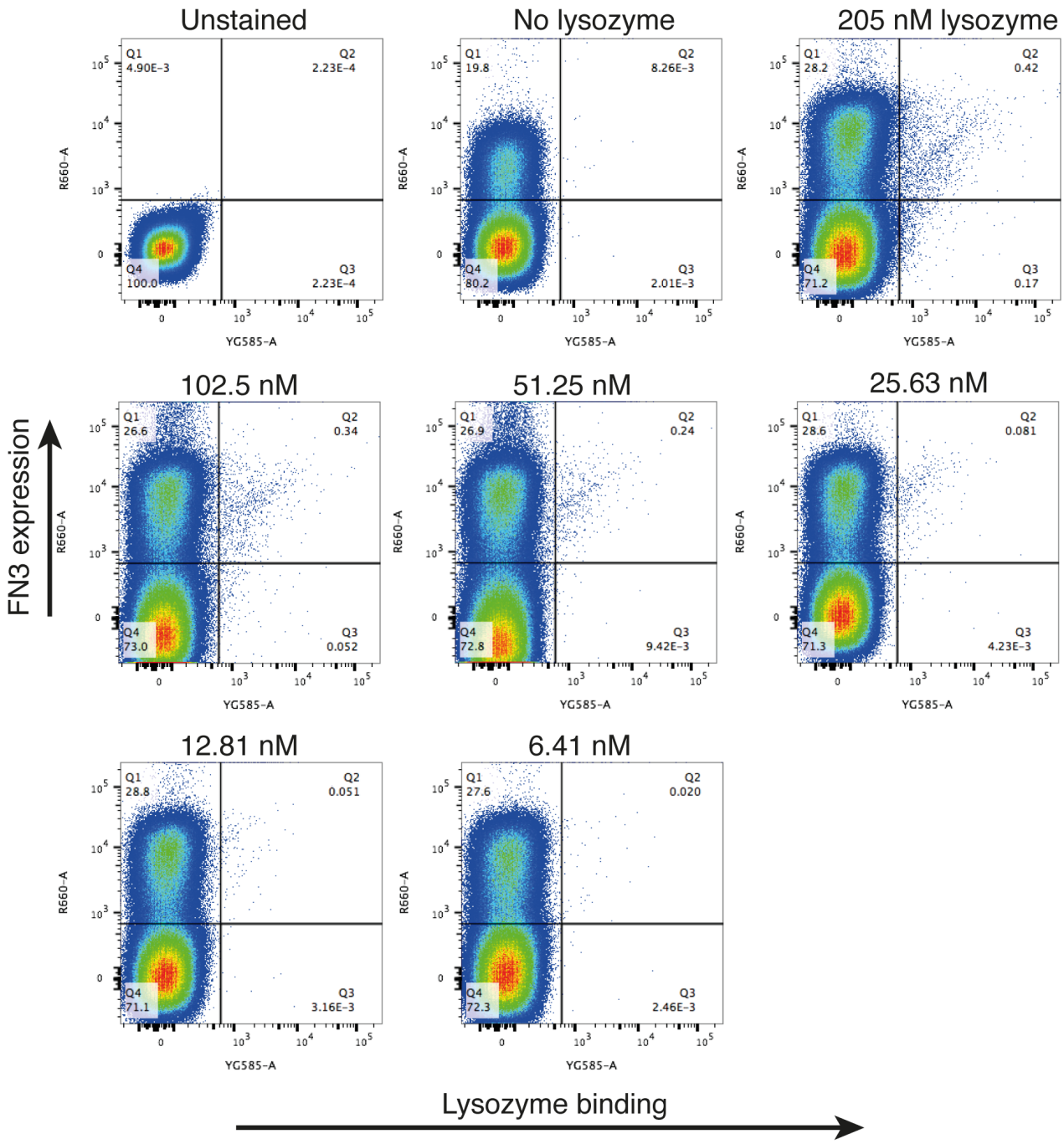


Figure 4.3. Flow cytometry analysis after one round of anti-biotin magnetic bead enrichment. Yeast cells expressing the FN3con library are labeled with biotinylated lysozyme in a concentration range of 205 nM to 6.41 nM and an anti-cMyc antibody (1/200). Detection of binding is achieved by secondary labeling with a goat anti-mouse fluorophore conjugate (1/400) and neutravidin-PE (1/400). Unstained and no lysozyme controls were used during analysis to assess background and affinity against the neutravadin-PE secondary.

The results seen after a single round of magnetic bead enrichment are encouraging, with 0.081% of double positive events (Q2) at a lysozyme concentration of 25.63 nM. To further enrich the double positive population in the Q2 gate, yeast cells were subjected to FACS with 25.63 nM of lysozyme (Fig. 4.4). As analysis and sorts were typically 1 day apart due to the heavy utilization of the flow cytometry facility, I would typically subculture the cells used during analysis. This however, appeared to result in a drastic reduction in the population of double positives, hence why the 25.63 nM plot from analysis (Fig. 4.3) is significantly different from the plot generated during FACS (Fig. 4.4). I suspect that the cause of problem is due to the starting small proportion of double positives (0.081%) that are essentially outgrown by the non-double positive population (99.919%). Regardless of this problem, the first round of FACS selected a double positive population of 0.009%, collecting 5,150 cells out of a total of 52.4 million (Fig. 4.4).

The 5,150 sorted cells were allowed to proliferate to sufficient density and subjected to a second round of analysis (Fig. 4.5). Analysis revealed further enrichment of the double positive clones in Q2, from 0.009% to 0.25% of the population at 25.63 nM of lysozyme. However, some (0.13%) background binding to neutravadin-PE is observed in the no lysozyme control. Undesired binding to the secondary fluorophores is a common and persistent problem in yeast surface display [105,214,216], although it may be resolved by alternating the use of different secondary fluorophores such as neutravidin-PE, streptavidin-PE or an anti-biotin antibody PE conjugate. Although time limitations restricted the ability to thoroughly resolve undesired binding to neutravadin-PE, and explore the biophysical characterization of isolated clones; analysis after magnetic bead enrichment and FACS fundamentally reveals a significant population of cells that display low nanomolar affinity (12.81-25.63 nM in Fig. 4.5) to lysozyme, without the need for any affinity maturation.

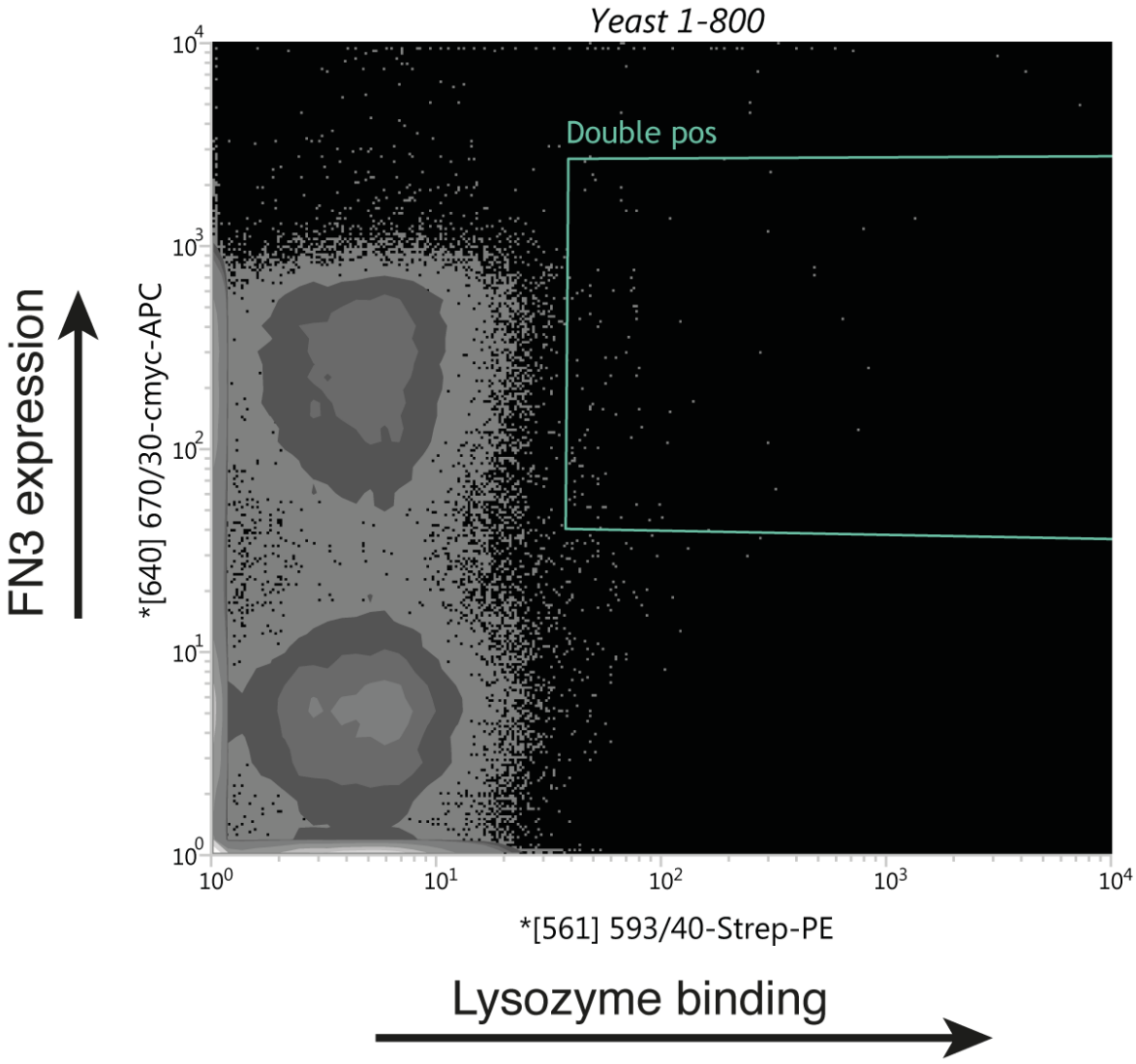


Figure 4.4. First round of screening with FACS. Yeast cells displaying an FN3con library have previously been enriched for lysozyme binders by magnetic bead screening (Fig. 4.3). Yeast cells expressing the FN3con library are labeled with the biotinylated lysozyme at a concentration of 25.63 nM and an anti-cMyc antibody (1/200 dilution). Detection of binding is achieved by secondary labeling with a goat anti-mouse fluorophore conjugate (1/400) and neutravidin-PE (1/400).

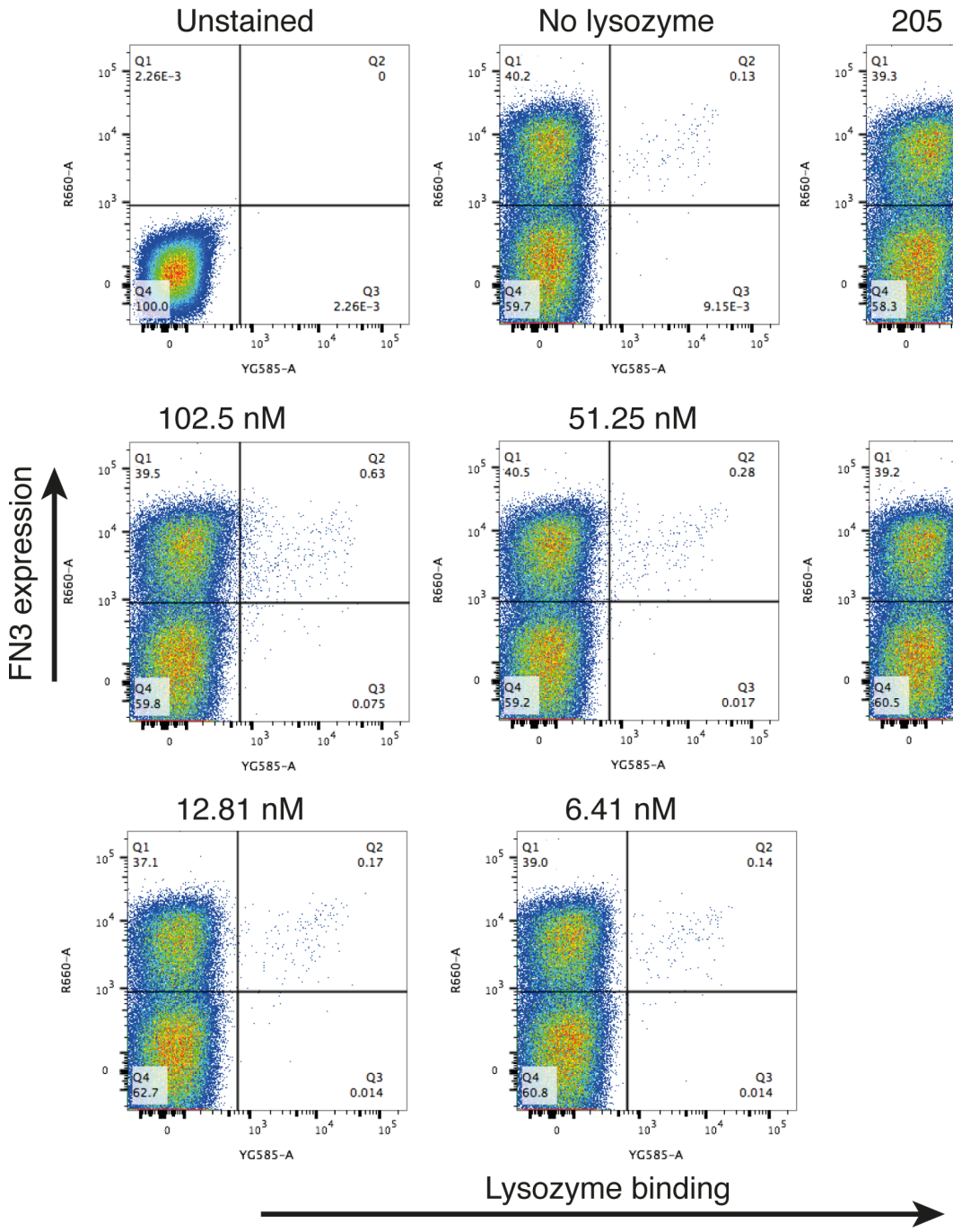


Figure 4.5. Flow cytometry analysis after one round of antibiotin magnetic bead enrichment and one round of FACS with neutravidin-PE. Yeast cells expressing a sorted FN3con library is labeled with biotinylated lysozyme in a concentration range of 205 nM to 6.41 nM and an anti-cMyc antibody (1/200). Detection of binding is achieved by secondary labeling with a goat anti-mouse fluorophore conjugate (1/400) and neutravidin-PE (1/400). Unstained and no lysozyme controls were used during analysis to assess background and affinity against the neutravadin-PE secondary.

Conclusions

As the thermodynamic and kinetic stability of naturally derived scaffolds such as FNfn10 imposes limitations on the sequence space available to the loops for display (Chapter 3), I hypothesised that the exceptional stability and mutational tolerance of FN3con provides a novel means for functional design. The preliminary results in this study suggest that FN3con is capable of displaying low nanomolar binders against lysozyme without the need for affinity maturation. This work serves as a promising foundation for which future biophysical characterization of isolated clones can be performed. Further research will also explore binding against more practical targets, the sequence space displayed on FN3con, whether isolated clones are amenable to other FN3 scaffolds, and how the stability-function trade off (discussed in Chapter 3) is affected. If future works prove this hypothesis to be true, it will further highlight the utility of consensus design for the generation of highly evolvable proteins.

Methods

DNA components of the library

DNA templates (10 µg)

Sample	Sequence
BC.loop.NNS Linear Diversity = ~10^{11}	ACTTGAGAGTTACTGACGTTACTTCTACTTCTGTTACTTTGTCTT GGGAANNSNNSNNSNNSNNSNNSNNSGGTTACAGAGTTGAATACA GAGAAGCTGGTGGTGAATGGAAGGAAGTTAC
DE.loop.NNS Linear Diversity = ~10^{11}	GAGTTGAATACAGAGAAGCTGGTGGTGAATGGAAGGAAGTTACTG TTCCANNSNNSNNSNNSNNSTCTTACACTGTTACTGGTTTGAAGC CAGGTACTGAATACGAATTCAGAGT
FG.loop.NNS Linear Diversity = ~10^{11}	CTGTTACTGGTTTGAAGCCAGGTACTGAATACGAATTCAGAGTTA GAGCTNNSNNSNNSNNSNNSNNSNNSCCATCTTCTGTTTCTG TTACTACTGGATCCGAACAAAAGCTTATTTCTGA
FN3con.delta.loops in pUC57	CTAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAG GGTCGGCTAGCATGCCATCTCCACCAGGTAACTTGAGAGTTACTG ACGTTACTTCTACTTCTGTTACTTTGTCTTGGGAACCCGGGGGTT ACAGAGTTGAATACAGAGAAGCTGGTGGTGAATGGAAGGAAGTTA CTGTTCCACCCGGGTCTTACACTGTTACTGGTTTGAAGCCAGGTA CTGAATACGAATTCAGAGTTAGAGCTCCCGGGCCATCTTCTGTTT CTGTTACTACTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACT TGTAATAGCTCGAGATCTGATA

Primers (100 nmol - ssDNA)

Sample	Sequence	
BC loop	5' primer BC.loop.fwd	ACTTGAGAGTTACTGACGTTACTTCTACTTCTGTTACTTTGTCTTGGGAA
	3' primer BC.loop.rev	GTAACTTCCTTCCATTCACCACCAGCTTCTCTGTATTCAACTCTGTAACC
DE loop	5' primer DE.loop.fwd	GAGTTGAATACAGAGAAGCTGGTGGTGAATGGAAGGAAGTTACTGTTCCA
	3' primer DE.loop.rev	ACTCTGAATTCGTATTCAGTACCTGGCTTCAAACCAGTAACAGTGTAAGA
FG loop	5' primer FG.loop.fwd	CTGTTACTGGTTTGAAGCCAGGTACTGAATACGAATTCAGAGTTAGAGCT
	3' primer FG.loop.rev	TCAGAAATAAGCTTTTGTTCGGATCCAGTAGTAACAGAAACAGAAGATGG
Scaffold	5' primer FN3con.delta.loop.fwd	CTAGTGGTGGAGGAGGCTCTGGT
	3' primer FN3con.delta.loop.rev	TATCAGATCTCGAGCTATTACAAGTCCTCTTCAGAAATAA

The entire construct is available on Benchling at: https://benchling.com/s/Fy6BrSWN/edit

pCTcon2-FN3con.delta.loops assembly

The pCTcon2-FN3con.delta.loops vector was assembled by subcloning FN3con.delta.loops from the pUC57 vector into pCTcon2. FN3con.delta.loops was chemically synthesized by GenScript (USA), whilst pCTcon2 was provided as a gift from Dane Wittrup (Addgene plasmid # 41843). As both pCTcon2 and FN3con.delta.loops contains NheI and BamHI restriction sites the construct was assembled by restriction digest (NEB products) and ligation (T4 ligase, NEB), then transformed into chemically competent DH5a *E. coli*. The new plasmid, pCTcon2-FN3con.delta.loops was then amplified in bacterial culture and 100 µg was purified by multiple minipreps. Prior to transformation of the amplified loop fragments, pCTcon2-FN3con.delta.loops needed to be linearized by overnight restriction digest with SmaI (NEB).

3x NNS loop preparation

The 3x NNS loops randomize the B/C, D/E and F/G loops of FN3con with 7, 5 and 8 amino acids, respectively. The NNS loops were synthesized as roughly 120bp cassettes that contain the randomized region and 50 bp of homology at the 5' and 3'. No changes were made to the loop length. The NNS region contains a diversity of approximately 10^{11} variants. Given the large amounts of DNA required for transformation and yeast homologous recombination (6 µg per transformation), the 10 µg of DNA per loop provided by Genscript was insufficient. I determined that 10x diversity equated to approximately 130 ng of DNA per loop. Therefore, amplification of 130 ng of DNA to 100 µg was sufficient for the transformation. Amplification from 130 ng of DNA to 100 µg is challenging in a single reaction. As replication accuracy is not necessarily important at this stage, I used the highly active MyTaq DNA polymerase (Bioline). Preliminary assessments found that a single 50 µl reaction was capable of amplifying 130 ng to 425 ng/µl after PCR clean up and elution in 20 µl of elution buffer. This roughly produced 8.5 µg of DNA per 50 µl reaction, therefore 100 µg was produced by 13 reactions per NNS loop. A master mix was created for all 13 reactions containing 5x MyTaq reaction buffer, 130 ng of template DNA, 6 mM of each forward and reverse primer, and 32 U of MyTaq DNA polymerase (Bioline). Prior to starting the reaction, the master mix was split into 13x 50 µl reactions in 200 µl PCR tubes. The mixture was denatured at 95°C for 30 seconds, followed by 35 cycles of 95°C for 15 s, 61°C for 15 s, and 72°C for 10 s, and a final extension at 72°C for 5 min. The PCR reaction was applied to a Wizard PCR cleanup kit (Promega) and ~1,000 ng/µl was eluted per loop in 100 µl, for a total of 100 µg of DNA per loop.

Library transformation

Twenty aliquots of 6 µg of each loop and 4 µg of linearized pCTcon2-FN3con.delta.loops were combined with 250 µl of electrocompetent EBY100 cells. Using the square wave protocol, a single pulse at 500 V with a 15-ms pulse duration was applied per aliquot. Pulsed cells were rescued with 1 mL of YPD (20g/L D-glucose, 20g/L peptone, 10g/L yeast extract) and transfered to a 50 ml tube. The cuvette was rinsed with an additional 1 mL of YPD and transfered to the same tube. Cells

were then incubated at 30°C without shaking for 1 hr. Serial dilutions of cells were plated on SD-CAA agar (20g/L D-glucose, 6.7g/L yeast nitrogen base, 5g/L casamino acids, 5.4g/L Na_2HPO_4, 8.6g/L NaH_2PO_4-H_2O, 16g/L agar, 182g/L sorbitol) to determine the number of transformants of the library. Plates were grown at 30°C for 2-3 days. Resulting in 2.3×10^8 transformants. The remaining cells were pelleted at 3000 x g for 3 min, resuspended in 1 L of SD-CAA media (20g/L D-glucose, 6.7g/L yeast nitrogen base, 5g/L casamino acids, 7.4g/L citric acid monohydrate, 10.4g/L sodium citrate, pH 4.5) and grown overnight at 30°C at 250 rpm.

Freezing yeast libraries

Yeast libraries were frozen for long-term storage at -80°C as a 15% glycerol stock. From an overnight culture, cells were harvested at 2,500 x g for 5 minutes at 4°C and the supernatant discarded. The pellet was resuspended in a freezing preparation of SD-CAA media to a cell concentration such that every vial has a 20- to 100-fold excess of library diversity. If the library produces $4x10^8$ clones, each vial should contain at least $4x10^9$ cells. At 100-fold excess, 40 storage vials can be created – the cell pellet was resispended in 24 mL of SD-CAA and 600 μl of cells were mixed in a cryotube containing 105 μl of sterile glycerol. Yeast vials were slowly frozen in an isopropanol bath and then transferred to a -80°C freezer. Revival of a frozen library was performed by thawing out a vial at 30°C or at room temperature, followed by growth in 500 ml SD-CAA overnight at 30°C shaking. Diversity induced cells were created by taking roughly 20 ml of the overnight culture and inoculating 500 mL of SG-CAA media (18g/L galactose, 2g/L D-glucose, 6.7g/L yeast nitrogen base, 5g/L casamino acids, 5.4g/L Na_2HPO_4, 8.6g/L NaH_2PO_4-H_2O, pH 6.0) followed by growth at 20°C, 250 rpm for 8-24 hours.

Library screening with magnetic beads

As an initial first step, magnetic bead enrichment provided a powerful and high throughput means to screen high and low affinity binders with the target antigen. $1x10^{10}$ diversity of induced yeast cells was washed with 25 ml of PBSA (0.01 M sodium phosphate, pH 7.4, 0.137 M sodium

chloride, 1 g/L bovine serum albumin). The cells were pelleted and resuspend in 10 ml PBSA. From a biotinylated lysozyme (hen egg white lysozyme, Sigma) stock at 300 µg/ml, lysozyme was added to a final concentration of 200 nM. Cells were incubated at room temperature for 60 minutes whilst on a roller. Cells were washed with 50 ml PBSA and resuspended in 5 ml PBSA, prior to the addition of 200 µl of magnetic anti-biotin beads (Macs Miltenyi Biotec). The cells were incubated at 4°C for 30 minutes. With an equilibrated LS column positioned in a MACS MultiStand, cells were added to the column and allowed to flow under gravity. Cells were eluted by removing the LS column from the MACS MultiStand and adding 7 ml of SD-CAA media to the column. Eluted cells were then used to inoculate 500 ml of SD-CAA with penstrep (1:100) (Sigma Aldrich) and grown at 30°C, 200 rpm for three days. Cells were used in subsequent FACS experiments, with a number of frozen vials created.

Library sorting with FACS

Prior to all sorting runs, analysis was performed to assess optimal lysozyme concentration for the given population of cells. Diversity induced yeast cells were pelleted and washed in 5 ml PBSA and resuspended in 1 ml PBSA with 1/200 of anti-cMyc antibody (mouse 9E11, ThermoFischer) and a required amount of biotinylated lysozyme (stock at 300 µg/ml). Analysis titrations ranged from 205 nM down to 6 nM. Cells were incubated at room temperature for 30 minutes on a roller. Cells were washed with 5 ml of PBSA, and resuspended in 1 ml PBSA with a 1/400 dilution of both Neutravadin PE (ThermoFischer) and goat anti-mouse IgG H+L Alexfluor 647 conjugate (ThermoFischer). Cells were then incubated at 4°C for 30 minutes, washed in 5 ml of PBSA and resuspended to a concentration of $1x10^8$ cells/ml. Analysis was used to determine the correct concentration for sorting. For sorting runs, double-positive cells were collected in 5 mL of SD-CAA media and added to 50 ml of SD-CAA. Sorted cells were grown at 30°C, 250 rpm until growth was visible.

Chapter 5

Smoothing a rugged protein folding landscape by sequence-based redesign

Summary

In this chapter, I explore the effect of consensus design on the serine protease inhibitor (serpin) family where stability and folding are directly linked to activity. Serpins exist in a metastable state that undergoes a major conformational change in order to inhibit target proteases. However, conformational lability of the native serpin fold renders them susceptible to misfolding and aggregation, and underlies misfolding diseases such as α_1-antitrypsin deficiency. As a result, a full characterization of the serpin folding pathway has proved impossible. Using a consensus approach we designed *conserpin*, a synthetic serpin that exhibits reversible two-state folding, is functional, thermostable, and resistant to polymerization. Characterization of its structure, folding and dynamics suggest that its remodeled folding landscape reduces the lifetime of the aggregation-prone intermediate ensemble. I propose that off-pathway folding and polymerization of the serpin is the result of independent evolutionary fine-tuning of the energy landscape for conformational control of function, and as such is eliminated by consensus design.

This chapter has since been modified and accepted by Nature Scientific Reports.

Smoothing a rugged protein folding landscape by sequence-based redesign

Benjamin T. Porebski[a,1], Shani Keleher[a,1], Jeffrey J. Hollins[b], Adrian A. Nickson[b], Emilia M. Marijanovic[a], Natalie A. Borg[a], Mauricio G.S. Costa[c], Mary A. Pearce[a], Weiwen Dai[a], Liguang Zhu[d], James Irving[e], David E. Hoke[a], Itamar Kass[a], James C. Whisstock[a,f], Stephen P. Bottomley[a], Geoffrey I. Webb[d], Sheena McGowan[a,g,2], Ashley M. Buckle[a,2]

Affiliations

[a]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia. [b]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom. [c]Programa de Computação Científica, Fundação Oswaldo Cruz, 21949900 Rio de Janeiro, Brazil. [d]Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia. [e]Wolfson Institute for Biomedical Research, University College London, Gower Street, London, WC1E 6BT, United Kingdom. [f]ARC Centre of Excellence in Advanced Molecular Imaging, Monash University, Clayton, Victoria 3800, Australia. [g]Department of Microbiology, Monash University, Clayton, Victoria 3800, Australia.

[1]B.T.P. and S.K contributed equally to this work

[2]To whom correspondence should be addressed:

Ashley M. Buckle. [redacted]

Sheena McGowan. [redacted]

Keywords: protein engineering, consensus design, protein stability, serpins, protein folding, aggregation

Abstract

Consensus design is one of the most successful sequence-based approaches for increasing protein stability. Here we apply this method to a relatively complex fold - the multi-domain serine proteinase inhibitor (serpin) fold, in which folding and activity are directly linked. Serpins exist in a metastable state that undergoes a major conformational change in order to inhibit target proteases. However, conformational lability of the native serpin fold renders them susceptible to misfolding and aggregation, and underlies misfolding diseases such as α_1-antitrypsin deficiency. As a result, a full characterization of the serpin folding pathway has proved impossible. Using a consensus approach we designed *conserpin*, a synthetic serpin that exhibits reversible two-state folding, is functional, thermostable, and resistant to polymerization. Characterization of its structure, folding and dynamics suggest that its remodeled folding landscape reduces the lifetime of the aggregation-prone intermediate ensemble. We propose that off-pathway folding and polymerization of the serpin is the result of independent evolutionary fine-tuning of the energy landscape for conformational control of function, and as such is eliminated by consensus design. This work indicates that consensus design is not only useful for increasing the stability of the native state of single domains, but is able to remodel the folding landscape. This may offer several benefits for protein engineering in general, including the removal of aggregation-prone intermediates and modifying protein scaffolds for use as protein therapeutics and diagnostic reagents.

Introduction

Consensus design, one of the most successful sequence-based approaches, is a probabilistic method based on the hypothesis that at a given position in a multiple sequence alignment of homologous proteins, the respective consensus amino acid contributes more than average to the stability of the protein than non-consensus amino acids [111,130,136,219]. The efficacy of consensus design has been demonstrated numerous times to increase the stabilities of a wide range of proteins [111,134,136-139,150,220].

To date, consensus design has typically been performed on single domain and relatively static proteins. In order to test its applicability to a highly complex fold, where folding and activity are directly linked, we applied this approach to the serine protease inhibitor (serpin) superfamily. The serpin superfamily contains over 1,500 members, which typically fold to a metastable native state that undergoes a major conformational change (termed the stressed [S] to relaxed [R] transition) central for the protease inhibitory mechanism [221]. The S to R transition is accompanied by a major increase in stability. The archetypal serpin fold is exemplified by α1-antitrypsin (α1-AT), a single domain protein consisting of 394 residues, which folds into 3 β-sheets (A → C) and 9 α-helices (A → I) that surround the central β-sheet scaffold [222]. The reactive center loop (RCL) protrudes from the main body of the molecule and contains the scissile bond (P1 and P1' residues), which mediates α1-AT's inhibitory specificity against the target protease, neutrophil elastase. The inhibitory mechanism of serpins is structurally well understood [221]. Briefly, a target protease initially interacts with and cleaves the RCL of the serpin. However, following RCL cleavage, but prior to the final hydrolysis of the acyl enzyme intermediate, the RCL inserts into the middle of β-sheet A to form an extra strand [221,223]. Since the protease is still covalently linked to the P1 residue, the process of RCL insertion results in the translocation of the protease to the opposite end of the molecule. In the final complex, the protease active site is distorted and trapped as the acyl enzyme intermediate [221,224].

In certain circumstances the serpin RCL can spontaneously insert, either partially (delta), or fully (latent) into the body of the serpin molecule without being cleaved [225]. Both latent and delta conformations are considerably more thermodynamically stable than the active, native state although they are inactive as protease inhibitors. Folding to the latent conformation is thought to occur via a late stage, irreversible folding step that is accessible from the native or highly native like state [226,227]. As such, change to the latent state can be triggered by perturbations to the native state via small changes in solution conditions such as temperature or pH [224,228,229], or by spontaneous formation over long time scales [230,231]. An additional consequence of serpin metastability is the polymerization events associated with the serpinopathies. Two mechanisms for serpin polymerization have been proposed. Early work suggested that the RCL of one molecule may insert into the A-sheet of another, thus forming a highly stable chain [232,233]. Later work, including structural studies on α1-AT suggested an alternative model for serpin polymerization via an extensive domain-swapping event [234,235]. Here it was suggested that mutations stabilize an ordinarily short-lived folding intermediate that has the propensity to form domain swapped polymers.

The puzzle of how the folding polypeptide chain of α1-AT achieves its metastable native state and avoids other thermodynamically more favorable conformations have proven challenging to solve. The unusual folding properties of serpins create significant challenges for isolating and studying the folding pathway to the native state. Equilibrium and kinetic unfolding experiments, which are generally more straightforward as they start with the well-defined native state, have been performed for a range of serpins including α1-AT [89,224,226-228,236-240]. A recent study of α1-AT folding using hydrogen-deuterium exchange mass spectrometry reveals relatively fast folding of the B/C barrel followed by much slower formation of the A β-sheet [227]. These observations are consistent with other domain-swapped models of α1-AT polymers and a mechanism of polymerization that involves a folding “race” between the B/C barrel and the A β-sheet [226,235]. After the initial fast folding of the B/C barrel, α1-AT populates an aggregation-prone intermediate

ensemble that is observed in all kinetics and some equilibrium unfolding/refolding studies [89,224,226-228,236-240].

Serpin redesign using a consensus approach is therefore interesting from two perspectives. First, serpins are multi-state proteins that have evolved a relatively complicated folding mechanism linked to its function. Coupled with sequence and structural diversity within the superfamily that reflects specialized functional and regulatory requirements, this presents challenges for consensus design, which has to date been restricted to relatively simple structures. Second, despite much effort, the aggregation-prone nature of wild-type serpins and poor refolding properties have hindered a rigorous characterization of the folding pathway. Construction of a synthetic serpin that reflects an optimal sequence conservation therefore offers a fresh avenue of exploration of folding behaviour, in addition to investigating how consensus design effects the folding landscape – an unexplored question. Using consensus design, therefore, we analysed a sequence alignment of the serpin superfamily and determined the prevalent amino acid residue at each position. In doing so, we generated a single sequence (396 residues in length) that was hypothesised to adopt a serpin fold and have a common biological behaviour – we termed this serpin *conserpin* (*con*sensus *serpin*). Crystallographic, equilibrium and kinetic folding studies, and molecular dynamics simulations reveal the characteristics of conserpin that likely dictate its remarkable stability and reversible folding behaviour, whilst retaining activity as a serine protease inhibitor. This work provides general insights into the effectiveness of consensus design for complex multi-state proteins, as well as the folding and misfolding mechanisms of serpins.

Results and Discussion

We constructed *conserpin* using a multiple alignment of 212 sequences from the serpin superfamily, starting from a previously reported alignment of 219 sequences [241]. After filtering to remove incomplete sequences and redundancy reduction, we aligned the resulting 212 sequences and generated a new protein sequence by selecting the most frequently observed residue at each column of the sequence alignment (the 'consensus method') (Dataset S1). N-terminal his-tagged conserpin was expressed in *Escherichia coli* and readily purified from the soluble fraction as a monomeric protein of expected molecular weight. Conserpin shares the highest sequence identity with α1-AT at 62 %. There is an overall loss of 10 residues in conserpin that are predominantly located at the N-terminus of the D-helix and C-terminus of the protein, and in total there are 137 sequence substitutions in comparison with α1-AT.

Conserpin is an inhibitory serpin

The consensus sequence of the RCL in conserpin contains 7 residue differences compared to α1-AT, notably an arginine at P1 compared to the methionine of α1-AT. Conserpin was found to inhibit trypsin with a stoichiometry of inhibition (SI) of 1.8 (Fig. S1A) and a k_{assapp} of 7.5×10^6 M^{-1} s^{-1} and hence a rate of association k_{ass} of 1.4×10^7 M^{-1} s^{-1} (Fig. S1B, C). Complex formation of conserpin with trypsin is observed on SDS PAGE (Fig. S1D). Partial degradation is seen, which is probably due to excess trypsin; however, there are also some unusual higher molecular weight species not seen in the inhibition of trypsin by α1-AT. This increased SI may be the consequence of conserpin having an RCL that is shortened on the 'prime' side of the recognition sequence for trypsin. Thus, further optimisation studies may provide improved kinetic parameters and target specificity [242].

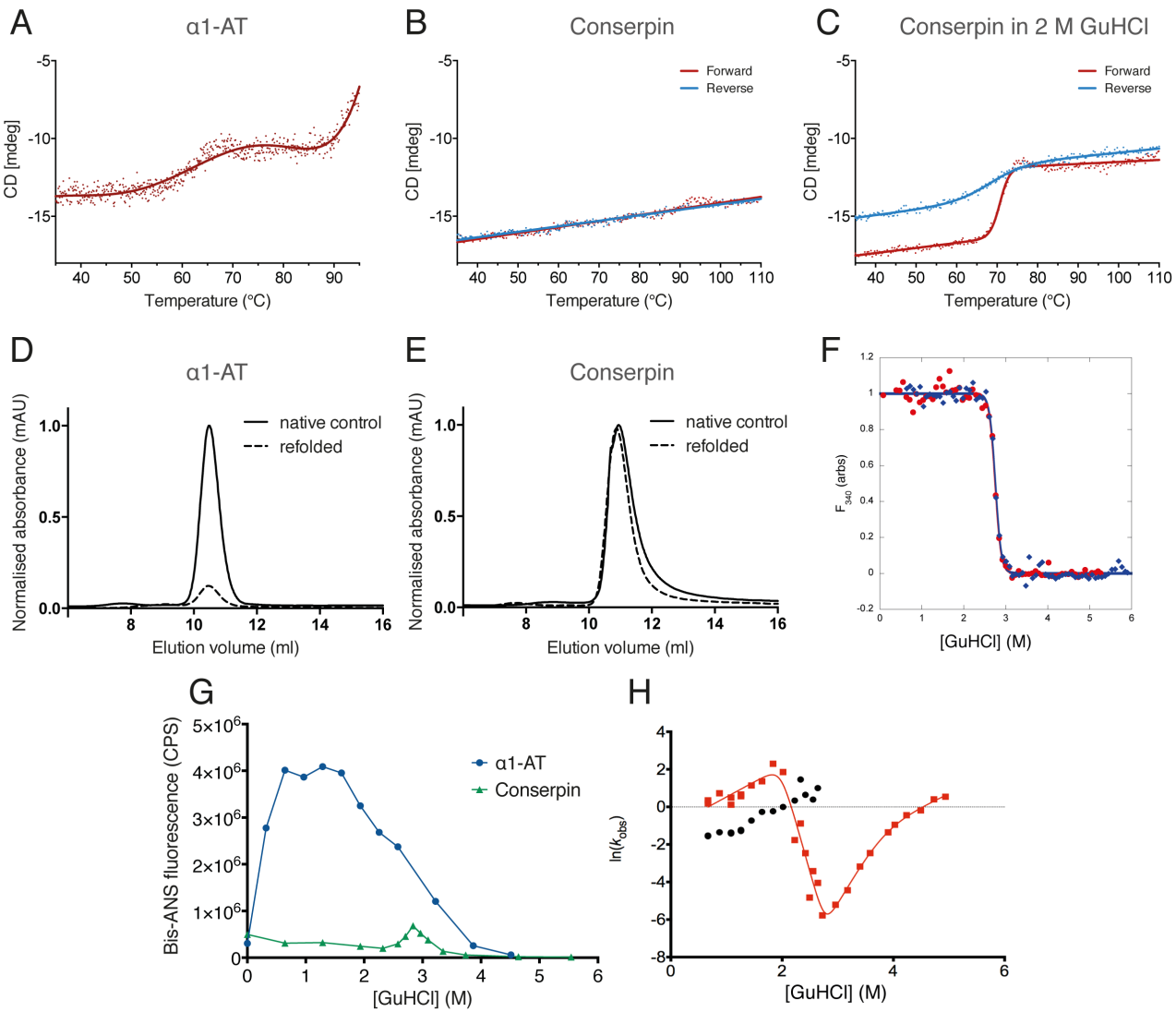


Figure 1. Conserpin has superior biophysical properties compared to α1-AT. Variable temperature thermal melts of **(A)** α1-AT, **(B)** conserpin and **(C)** conserpin in 2M GuHCl as measured by CD at 222 nm. Chemical refolding of **(D)** α1-AT and **(E)** conserpin shows that conserpin can refold to a monomer. Gel filtration chromatograms from a Superdex 75 10/300 size exclusion column are shown. Final protein concentrations loaded to column were 2 µM. Samples were unfolded in 5 M GuHCl and then diluted out to 0.5 M GuHCl (dotted line). Control samples of native protein are shown as the solid black line. **(F)** Intrinsic fluorescence equilibrium unfolding (red dots) and refolding (blue diamonds) curves of conserpin coincide, demonstrating the reversibility of folding. **(G)** Conserpin shows a significant reduction of intermediate formation during bis-ANS fluorescent equilibrium unfolding of α1-AT (blue circles) and conserpin (green triangles). **(H)** Chevron plot for conserpin showing two discernable refolding rates. The positive slope in each refolding arms suggests the presence of intermediate species that have to partially unfold to reach the native state. Red squares are the fast folding rate, black circles are the second folding rate.

Conserpin is thermostable and resists polymerization when heated

Variable temperature far-UV circular dichroism (CD) studies indicate that conserpin is highly thermostable. Upon heating, α1-AT undergoes a three-state transition with an initial midpoint temperature (T_m) of 61.8 °C (Fig. 1A) and an incomplete transition that starts at 90 °C. This process is not reversible and on cooling we observed the presence of white precipitate in the cuvette. The initial T_m of 61.8 °C fits well with what has been reported in the literature to rapidly induce aggregation of α1-AT [232,243]. We therefore postulate that the second transition reflects unfolding to an intermediate state, followed by the production of higher-order α1-AT aggregates. In contrast, conserpin showed a slight decrease in signal (~2.66 mdeg) at 222 nm by 110 °C, with no defined unfolding transition, and was fully reversible on cooling (Fig. 1B). Spectral scans of conserpin before and after the thermal melt indicate no change in secondary structure content, which was previously observed in comparisons of native and latent plasminogen activator inhibitor 1 (PAI-1) [236,244]. This therefore suggests that conserpin is capable of reversible folding to the native state under thermal denaturation conditions. Complete unfolding of conserpin was only achieved in the presence of 2 M guanidine hydrochloride (GuHCl) with a T_m of 72.5°C (Fig. 1C), however, the CD signal was not completely reversible. As we did not observe any visible aggregate on cooling, we suspect that a combination of thermal and chemical treatment induce a transition to the latent state or other higher order species. Refolding transverse urea gradient (TUG) gels further demonstrate that conserpin is more resistant to polymerization (Fig. S2A, B) than is α1-AT, which mostly formed polymers on refolding (Fig. S2C, D). Although some polymer formation is observed in the conserpin chemical refolding TUG gel (Fig. S2B), this was not seen when the experiment was repeated using gel filtration (Fig. 1D,E).

Conserpin folds reversibly to the native state

The majority of serpins unfold through an aggregation-prone intermediate ensemble and do not completely refold after thermal and/or chemical denaturation [89,227,236,245-248]. This is exemplified by α1-AT, which shows a very small amount of refolded monomer via chemical denaturation, rapid dilution and gel filtration (Fig. 1D). In contrast, conserpin can be completely refolded back to a monomeric state (Fig. 1E). Equilibrium unfolding and refolding of conserpin in the presence of GuHCl, measured by intrinsic fluorescence, fits well to a two-state equation and confirms complete and reversible refolding to a native-like state (Fig. 1F). The unfolding and refolding curves overlay well, revealing a midpoint of denaturation, $[D]_{50}$, of 2.75 ± 0.10 M, an equilibrium *m*-value, $m_{D\text{-}N}$, of 8.45 ± 0.65 kcal mol^{-1} M^{-1}, and hence a stability, $\Delta G_{D\text{-}N}$, of -23.2 ± 2.0 kcal mol^{-1}. The correlation of unfolding and refolding curves, the single unfolding transition, and the steep *m*-value all suggest that formation of an intermediate ensemble is minimal. In comparison, the serpin PAI-1 has a $\Delta G_{D\text{-}Native}$ of ~-12 kcal mol^{-1}, and a $\Delta G_{D\text{-}Latent}$ of ~-21 kcal mol^{-1} [236], which would be regarded as a high level of stability of the native state and an extreme level of stability of the latent state.

As structural analysis later reveals, the positions of tryptophan and tyrosine residues between the native and latent states of conserpin are identical, thus the intrinsic fluorescent methods used are unable to distinguish these states from one another. To overcome this ambiguity and further validate the initial CD analysis, we conducted catalytic assays on chemically refolded and thermally treated conserpin with trypsin. Conserpin when unfolded in 6 M GuHCl and rapidly diluted in TBS was found to inhibit with an SI of 2.29 (Fig. S1E), a slight decrease from native conserpin's SI of 1.8 (Fig. S1A), but within the range of error for this experiment, suggesting that conserpin is fully refoldable to the native state after chemical denaturation. However, after thermal treatment at 80°C for 20 minutes, conserpin suffered a complete loss in activity (Fig. S1F), which in contrast with our previous CD data, suggesting that conserpin transitions to the more stable latent state rather than entering polymerization and aggregation pathways.

Conserpin avoids polymerization by transient sampling of an intermediate ensemble during folding and unfolding

As intermediates are key species in the aggregation pathway, and minimal formation was observed using equilibrium unfolding/refolding via intrinsic fluorescence, we repeated the experiment using bis-ANS (4,4'-Dianilino-1,1'-Binaphthyl-5,5'-Disulfonic Acid, Dipotassium Salt) fluorescence – a dye that fluoresces on binding to hydrophobic regions of a protein. In native conditions (0 M GuHCl), both conserpin and α1-AT show similar levels of fluorescence (Fig. 1G). The unfolding transition of α1-AT reveals a high intensity fluorescent peak between 1 – 2 M GuHCl, indicating the presence of a folding intermediate. In contrast, bis-ANS binding in the unfolding of conserpin is reduced approximately 4-fold, with a relatively narrow peak at approximately twice the concentration of GuHCl compared to α1-AT. This is consistent with the $[D]_{50}$ determined by intrinsic fluorescence and reduced intermediate ensemble formation.

We next used rapid mixing techniques to see whether intermediates could be observed kinetically (Fig. 1H). Although the unfolding traces fitted well to a single exponential and were easy to obtain, the refolding traces were far more complicated. When the protein was refolded from an equilibrated denatured solution (single-jump), the resulting traces could not be fitted to fewer than three exponentials and showed inconsistencies between repeats. In contrast, when native conserpin was first unfolded and then refolded after a short delay (double-jump), the refolding traces were far more consistent and fitted to a double exponential. Both refolding rates were independent of the delay time. There are three alternative explanations for the presence of two refolding rates: (1) the presence of two denatured states folding on different timescales (for example, folding is limited by proline isomerization); (2) a fast rate of refolding to a structured intermediate, followed by a slow rate of refolding from that intermediate; or (3) two fluorophores reporting on independent folding events (e.g., two independently nucleating subdomains) [249].

For conserpin, it is most likely that we are detecting folding from two similarly structured ground states. If we were observing a fast rate, followed by a slow rate, we should expect the fast rate to become kinetically invisible when the two rates cross (~2 M GuHCl), which it does not. Similarly, if there are two independent folding events, then the relative amplitudes of each rate should be consistent, which they are not. Most interestingly, the refolding *m*-values are positive at low concentrations of denaturant (< 2 M), suggesting that the two populated ground states are more structured than the subsequent folding transition state(s). Therefore, the starting states cannot be denatured states, and must be structured intermediates. The fast folding rate (Fig. 1F, red squares) is consistent with the unfolding rate at the expected $[D]_{50,}$ (2.75 M), verifying that this rate shows folding over the major transition state. The "rollover" in this rate demonstrates that the first intermediate (I_1) is in rapid pre-equilibrium with the denatured state (D) and there is a switch in ground state from I_1 to D when the two species are of equal stability (2 M GuHCl). The second intermediate (I_2; Fig. 1F, black circles) shows an almost identical folding *m*-value and, assuming this also folds over the major transition state, is likely to be very similar in structure to the first intermediate. However, I_2 is more stable than I_1 and persists until the denaturant midpoint (2.75 M).

As such, we propose that I_1 is likely to be the previously observed polymerogenic folding branch point [226,227,237,240,250-252]. From the known relationship between *m*-value and the change in accessible surface area upon unfolding (SASA; [163]), we can estimate [253] that I_1 is very native-like in structure. As native-like species on the folding pathway are highly aggregation prone in other serpins, it is possible that I_2 is a multimer (possibly a dimer) of the first intermediate. Nevertheless, in contrast to other serpins, where the aggregation-prone intermediate ensemble is substantially populated, the conserpin intermediate ensemble is transient and this may contribute to the observed lack of polymerization.

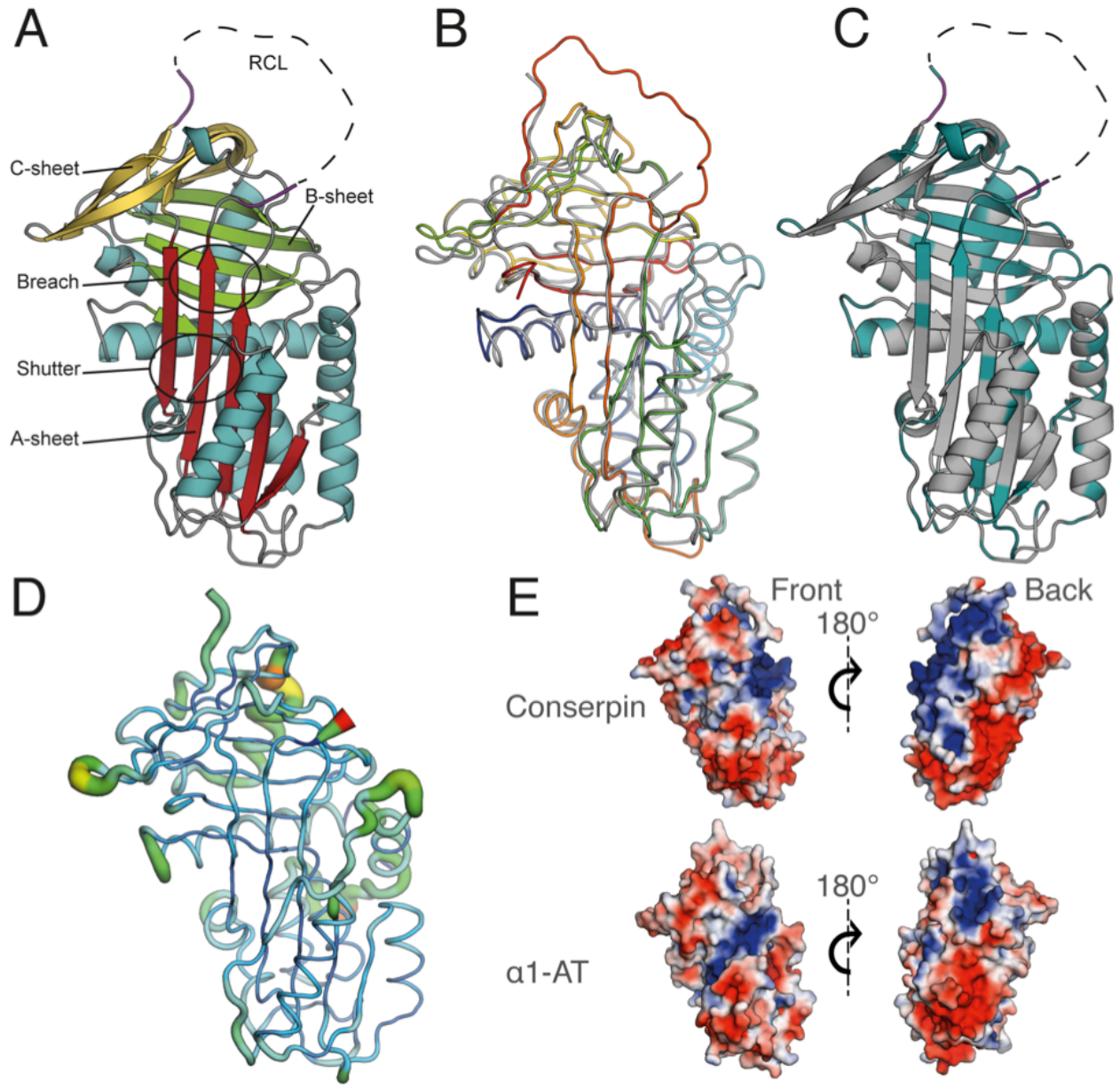


Figure 2. Structural analysis of conserpin. **(A)** Crystal structure of conserpin, showing the breach and shutter regions, the A, B and C sheets colored in red, green and yellow respectively, and the RCL stumps in magenta; **(B)** Structural alignment of conserpin (grey) with α1-AT (PDB: 3NE4; spectrum, blue to red); **(C)** Conserpin showing residue differences (cyan) with α1-AT; **(D)** Conserpin ribbon with color gradient according to RMSD with α1-AT (blue=low, green=medium, red=high); **(E)** Electrostatic potential surface of conserpin and α1-AT models (blue=+ve, red=-ve), in same orientation as other panels (front) and a 180° rotation.

The structure of native conserpin is not typical for a thermostable protein

Our kinetics data imply that the folding reversibility and low polymerization propensity of conserpin is due to minimal formation of a folding intermediate. However, to investigate whether the structure of conserpin in the native state also plays a role in its increased thermodynamic stability, we determined an X-ray crystal structure in the native state at 2.4 Å resolution (Table S1). Conserpin adopts the archetypal native serpin fold (Fig. 2A); structural alignment with α1-AT reveals a root mean square deviation (RMSD) of 0.91 Å across 296 backbone Cα atoms (Fig. 2B, D), reflecting their similar sequences (sequence ID=62%; Fig. 2C). Comparison with the native-state structures of α1-AT and thermostable serpins, thermopin [254] and tengpin [255], reveals that despite having the highest thermostability, conserpin has the fewest H-bonds and salt bridges (Table S2). Conserpin has the largest accessible surface area and largest solvent inaccessible cavity volume. These characteristics are unusual for thermostable proteins, which typically feature more interactions and optimized packing compared to their mesophilic counterparts [16,22,42,137,155,256-260]. A growing body of literature implicates the electrostatic surface potential with aggregation resistance, and subsequent engineering studies have emphasized this link by dramatically increasing the net surface charge [80,81]. Comparison of the electrostatic surface potential of conserpin with that of α1-AT reveals minor differences on the surface-exposed face of the A-sheet, whilst the back surface of the molecule reveals that conserpin is substantially more positively charged (blue region; Fig. 2E). However, these results suggest global analyses to not be a particularly reliable means for predicting stability.

Favorable interactions and reduced dynamics surrounding the D-helix

Given the conformational plasticity required for serpin function, we next investigated the differences in the dynamics between conserpin and α1-AT. We performed molecular dynamics (MD) simulations for 0.5 μs at 300 K in triplicate for both conserpin and α1-AT. RMSD of Cα atoms across the simulation trajectories show both systems reach equilibrium by 150 ns, with conserpin having an average RMSD of 2.60 Å and α1-AT of 2.34 Å (Fig. S3A). Although the increased

mobility of the RCL and the C-terminus of hA of conserpin leads to a higher overall RMSD, inspection of root mean square fluctuations (RMSFs) shows conserpin to exhibit an overall reduction in dynamics in the majority of regions, specifically the extended N-terminus of hA, hC/hD loop, hD, hE, hF, hG, hH (Fig. S3B, C).

The most notable reduction in the dynamics of conserpin, compared to α1-AT, is seen in the D-helix (hD; RMSD of 0.58 vs. 1.65 Å; Fig. 3A and S3B, C). The D-helix of α1-AT has been implicated in stability; notably two mutations ($T114F_{\alpha1\text{-}AT}$ and $G117F_{\alpha1\text{-}AT}$) stabilize the D-helix and rescue the polymerogenic Z-variant [261,262]. Structurally, the conserpin hD is shortened by the deletion of five residues (~1.5 turns); four residues at the N-terminal end (L84, E86, I87 and P88 in α1-AT) and one at the C-terminal end (Q109 in α1-AT; Fig. 3B and S4A). The deletion of $L84_{\alpha1\text{-}AT}$ and $I87_{\alpha1\text{-}AT}$ reduces overall hydrophobicity without affecting the packing of hD against the core of conserpin (Fig. S4A). Residue numbering will adhere to the following convention unless explicitly stated: $Q105_{\alpha1\text{-}AT}$ or $R79_{conserpin}$ or $Q105R^{79}$, where Q105 from α1-AT has been mutated to an R, which is residue number 79 in conserpin.

The stability of the hD in conserpin compared to α1-AT appears to arise from two main events: (i) formation of a salt bridge between $Q105R^{79}$ of hD and $E376^{346}$ and (ii) interactions of the N-terminus with hD. The salt bridge between the B-sheet and hD of conserpin is present throughout the entire MD simulation and may function to stabilize the top of the D-helix (Fig. 3B). In contrast, there are no similar salt bridges in the α1-AT crystal structure or during MD (Fig. 3B). Rather, hD in α1-AT undergoes significant conformational rearrangement and loss of secondary structure in one of the replicates (Fig. S4B). This is consistent with other reports, which indicates that minor changes to hD may accelerate or reduce polymer formation [261,262]. The second stabilizing factor is the presence of an extended N-terminus in conserpin that results from the addition of the N-terminal purification tag. Although an artifact to conserpin, N-terminal extensions have been observed in thermophilic serpins [254,255]. Four residues of the extension (residues -1 to -4) were

resolved in the X-ray crystal structure and a single H-bond is observed between the backbone of residue A-$1_{conserpin}$ and the N-terminus of hD ($D65_{conserpin}$; Fig 3C). Throughout MD, this H-bond is persistent and extends to form a small β-sheet (Fig. 3D). Given the addition of H-bonds, it is possible that our extended N-terminus of conserpin imparts stabilizing influences to hD and may reflect similar interactions seen in the naturally extended N-termini of thermophilic serpins [254,255]. Taken together, our observations suggest that optimized interactions in and around hD increase the stability of the native state.

The electrostatic network of the serpin breach region is extended in conserpin

The breach region, consisting of a highly conserved electrostatic network between residues $E342_{α1\text{-}AT}$, $K290_{α1\text{-}AT}$ and $D341_{α1\text{-}AT}$ at the top of the A-sheet of serpins is important for controlling the conformational change that drives protease inhibition [229,232,241]. This network is significantly extended in conserpin, compared to α1-AT (Fig. 4A). Specifically, the mutations of $T339E^{310}$ and $S292K^{264}$ contribute to a salt bridge network spanning s3A, s5A and s6A with $K191^{163}$. $T294E^{266}$ also forms a new salt bridge with $K335^{306}$ between s6A and s5A, whilst $D341N^{312}$ mediates an unfavorably charged cluster of $E310_{conserpin}$, $E313_{conserpin}$ and $E314_{conserpin}$ that is not present in α1-AT (Fig. 4A). These observations are interesting in the context of serpin polymerization, which involves insertion of the RCL and/or s5A from one molecule into the flexible A-sheet of another [226,227,234,235,251,263]. In particular, the disease-causing Z-variant, $E342K_{α1\text{-}AT}$ induces repulsion with $K290_{α1\text{-}AT}$, which either retards the formation of the A-sheet during folding, increasing the lifetime of the polymerogenic intermediate ensemble, or destabilizes the structure and increases the dynamics of the native state, allowing for s5A and s6A to separate, and reduce the energy barrier for polymerization [226,227,264-266]. The extended salt-bridge network in conserpin is consistent with these mechanisms, via either stabilization of the native state and reduction in dynamics, or reduction of the population of the polymerogenic intermediate ensemble by alteration of the folding energy landscape.

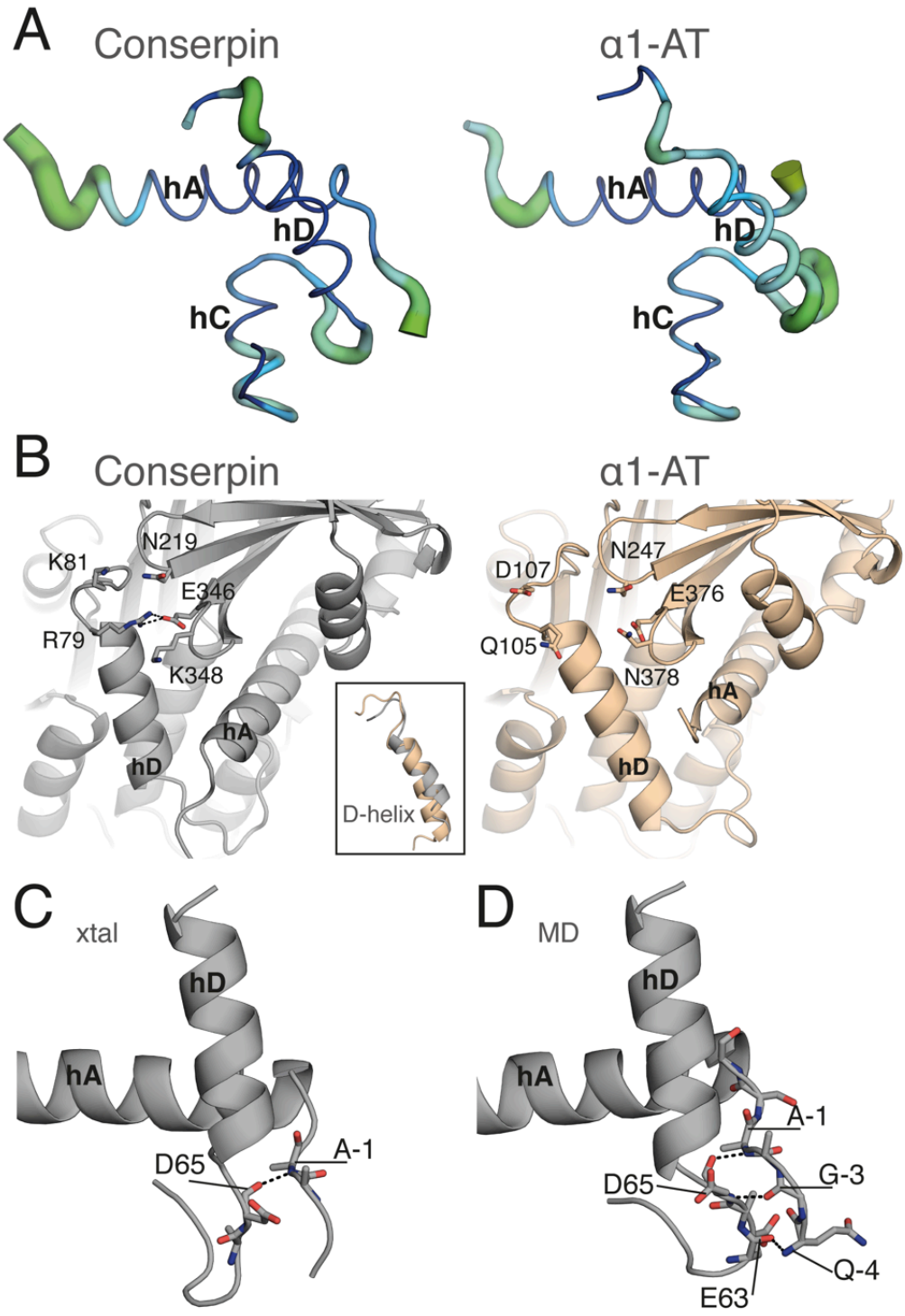


Figure 3. Structure and dynamics of helix-D in conserpin and α1-AT. **(A)** MD RMSF of each Cα atom mapped onto conserpin and α1-AT crystal structures as B-factor putty, highlighting differences surrounding hD. The width of the tubes and their colors (blue to red) are proportional to the magnitude of the RMSF (maximum: 6 Å); **(B)** The introduced salt bridge in hD of conserpin with residues Q105R[79] and E376[346]. There is no comparable interaction present in α1-AT. Inset shows the shortened D-helix in conserpin; **(C)** H-bonding between A-1 of the extended N-terminus and D65 of hD, as seen in the conserpin crystal structure; **(D)** Persistent hydrogen bonding between Q-4, G-3 and A-1 of the extended N-terminus and E63 and D65 of hD, in conserpin as seen in MD simulation.

It is difficult to truly ascertain the effect of conserpin's extended salt bridge network on A-sheet dynamics during folding as our MD simulations only described dynamics of the native state. Nevertheless, our simulations of α1-AT reveal its A-sheet salt bridge network to dissipate over time (Fig. 4B), allowing for the slight separation of strands s5A and s3A (Fig. S5). In contrast, the extensive network in conserpin is present throughout the majority of the simulation, with additional interactions being formed, notably an alternate conformation of $K264_{conserpin}$ mediating interactions between $E310_{conserpin}$, and $E266_{conserpin}$ (Fig. 4B). Furthermore, $E317_{conserpin}$ in the RCL of conserpin is able to adopt a stable conformation, mediating the salt-bridge between $K165_{conserpin}$ and $K217_{conserpin}$, with $K165_{conserpin}$ forming transient interactions to $E314_{conserpin}$. Equilibrium and kinetic folding studies of α1-AT provide compelling evidence for the late folding of s5A during transition through the polymerogenic intermediate state [226,227]. As such, our observations suggest an improved, energetically stable native state with possible improvements to the folding cooperativity in this region, which may also be augmented by the hydrophobic core behind the A-sheet.

Importance of A-sheet hydrophobic core packing

The hydrophobic core buried by the A-sheet is known to be important for serpin stability [243,267,268]. Amongst 19 mutations designed to probe the stability of α1-AT, 7 mutations in the hydrophobic core were found to be stabilizing [267]. Four of these mutations are found in conserpin (T59S[37], T68A[46], A70G[48] and M374I[244]). For the remaining 3 that are not in conserpin, the local environment adapts to improve packing or introduce favorable interactions. This is seen with F51[29], where adjacent mutations I340V[311] and L291F[263], M374I[244], and an alternate conformation of I188[160] optimize packing of the hydrophobic core (Fig. S6A). Conserpin also contains T59S[37], which is surrounded by additional mutations L30N[8], A58S[36] and S140A[113], together allowing for favorable non-polar and polar interactions that are not possible in α1-AT (Fig. S6B). Similarly, L291F[263] improves van der Waals packing against s6A (Fig. S6A). Both α1-AT and conserpin contain K387[357], buried in the core. However, in conserpin the neighboring mutation N46D[25] allows for the formation of a transient salt bridge between E264[236] of hH and K387[357] of s5B during MD

(Fig. S6C). In the context of folding, where it has been shown that s5A is late in folding to the native state, conserpin has no changes to any of the hydrophobic residues of s5A. As such, if there are any effects on the folding rate surrounding s5A, this may rather be a cooperative folding event contributed by other strands in sheet-A or hF via hydrophobic or electrostatic forces.

Improved packing of the F-helix may increase native state stability

Packing between hF and the A-sheet is known to stabilize the native state serpin fold, with hF acting as a physical barrier for RCL insertion into the A-sheet and subsequently during protease inhibition and polymerization [226,228,269-271]. Conserpin contains 3 key mutations in this region (Fig. 5A); Y187A[159] and G115A[88], which allows s2A to more tightly pack against hF, and Y160W[132], which further improves the packing density (collectively reducing the cavity volumes from 233.8 to 120.9 Å^3 (Fig. 5B)). This is consistent with mutagenesis studies of α1-AT, where Y160A resulted in a 5°C decrease in T_m and was attributed to the loss of a hydrogen bond and creation of a cavity [269]. In contrast Y160W raised the T_m of α1-AT to 65°C, and slowed the rate of polymerization (from 10 to 60 minutes at 60°C) [269]. MD reveals hF of conserpin to be slightly less flexible than that of α1-AT, with W160[132] remaining conformationally locked compared with Y160 of α1-AT, which frequently flips in and out of the hydrophobic core (Fig. 5C). We note that interactions within the “clasp” motif at the F-helix are structurally conserved in conserpin and maintained throughout simulation, consistent with its proposed role in regulating conformational change [272]. Taken together, these changes likely provide a significant improvement to the stability of the native state.

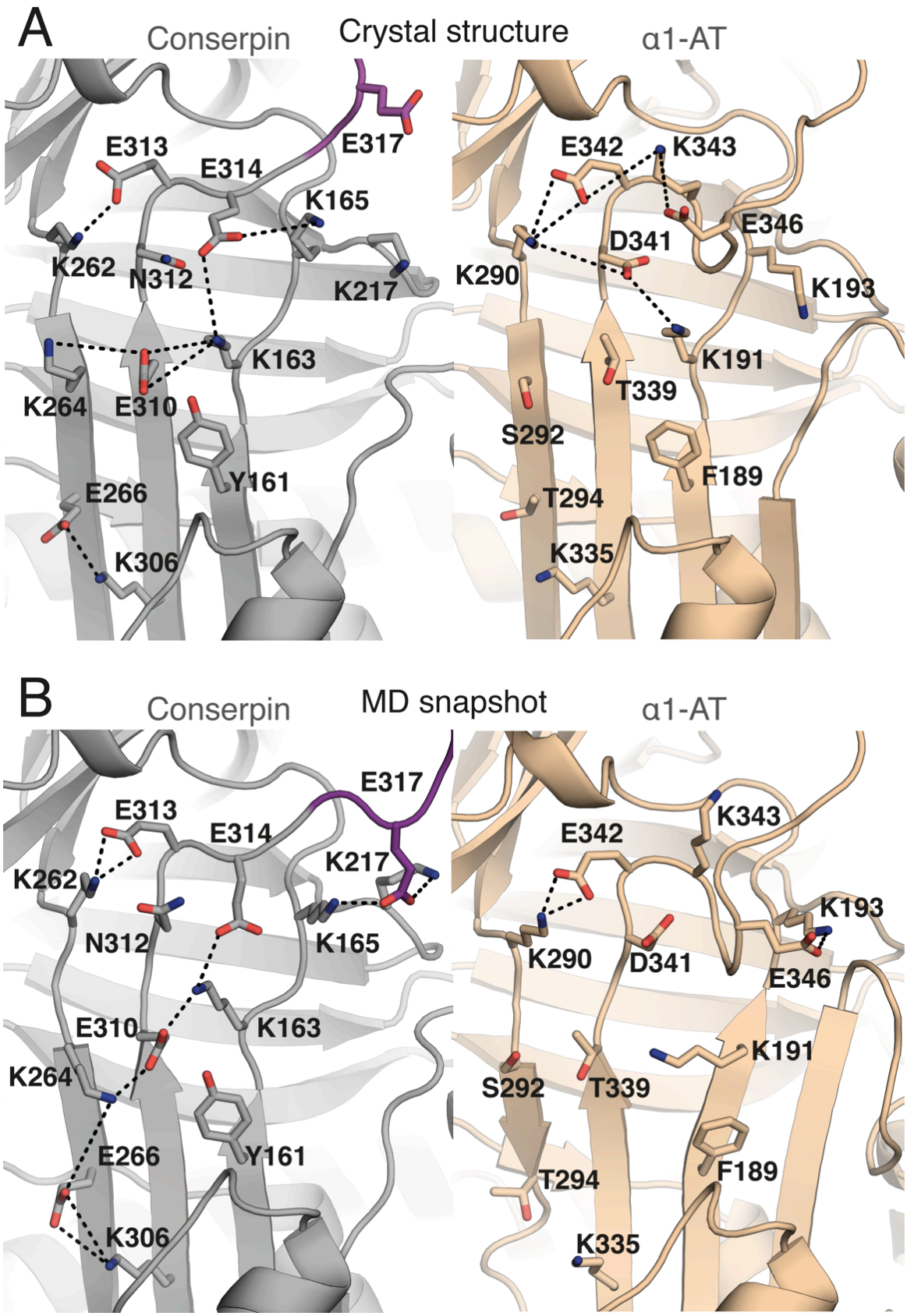


Figure 4. (A) A-sheet salt bridge interactions in the crystal structures of conserpin (grey) and α1-AT (wheat; PDB: 3NE4); **(B)** Simulation snapshot taken at 500 ns, showing A-sheet salt bridge interactions. The modeled RCL of conserpin is colored magenta.

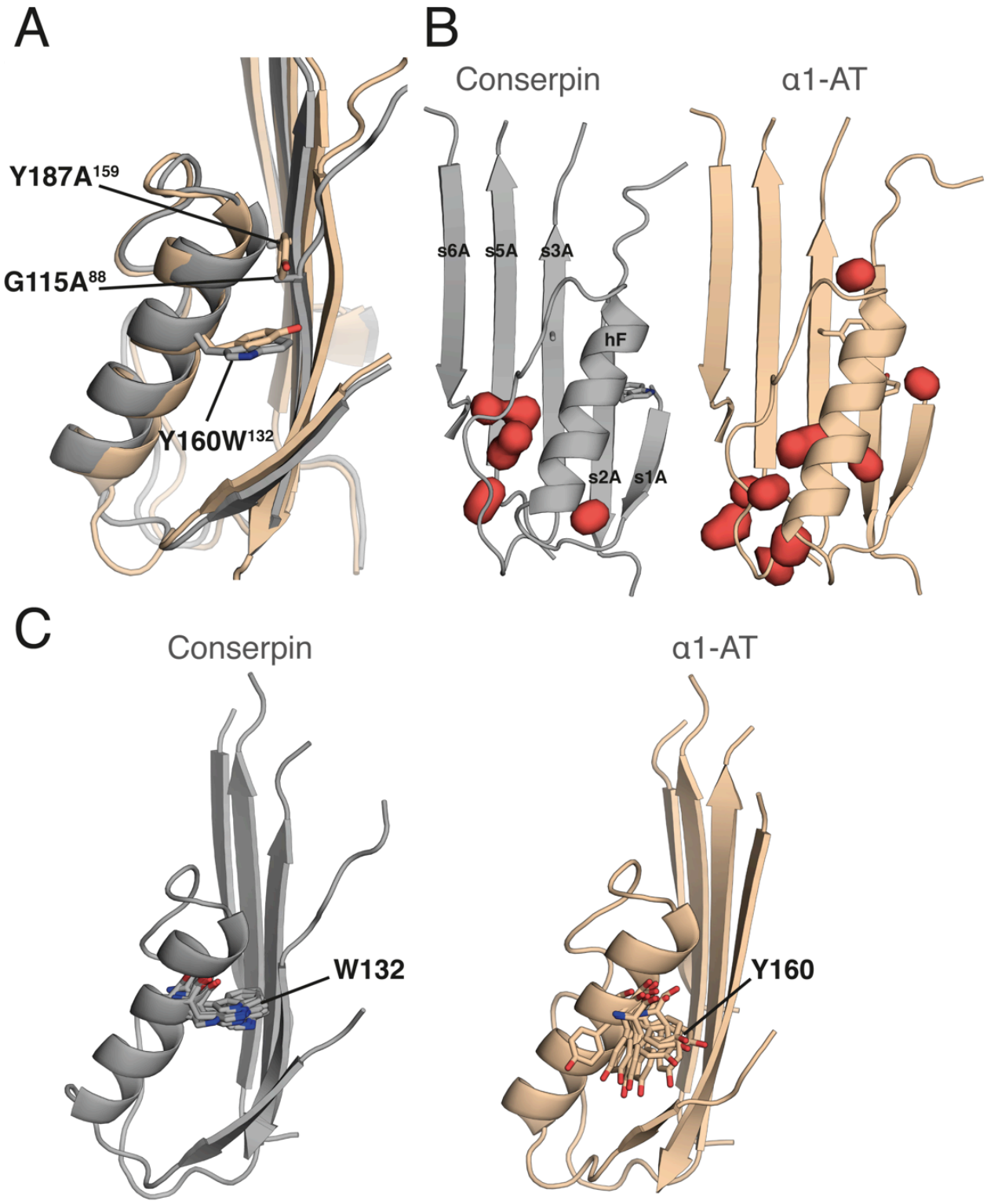


Figure 5. W160 stabilizes hF in conserpin. **(A)** A structural overlay of hF in conserpin (grey) and α1-AT (wheat), highlighting the positions of Y160W[132], Y187A[159] and G115A[88]**; (B)** Solvent inaccessible cavities (red blobs) surrounding hF of conserpin and α1-AT; **(C)** MD simulation frames (every 50 ns), highlighting the dynamic differences of W132 in conserpin and Y160 in α1-AT.

Remodeling of the B/C barrel, a putative folding nucleus

There is substantial evidence supporting the critical role of the B/C barrel in the folding pathway of α1-AT and stability of the metastable native state [224,226,227,238,273]. The B/C barrel of α1-AT is responsible for the fast folding rate seen by via the rapid formation of a folding nucleus that precedes formation of the A-sheet [227]. As part of this rapid folding event, the B/C barrel is proposed to function as a “kinetic trap” that captures the RCL and prevents folding to other more stable states, such as the latent state [224,226,227,238,273]. Mutations that destabilize the B/C barrel subsequently slow down folding and increase the propensity for folding to the latent state and formation of polymers [227,238]. Conversely, mutations and small molecules that stabilize the native state of the B/C barrel promote resistance to the formation of more stable conformations [267,273].

Conserpin contains several mutations in the B/C barrel that improve hydrophobic packing and form favorable interactions within the native state (Fig. 6A). Most notably is the tighter packing of hH as a result of F275W[247] and E279L[251], which is compensated by local repacking through F253I[225] and Q230Y[202]. In addition, the introduction of a small salt-bridge network between K274[246], C232D[204] and K234E[206] in hH may further strengthen the hydrophobic core of the B/C barrel and provide extra stability to the native state (Fig. 6A). Prior studies involving alanine scanning and mutagenesis of the B/C barrel have revealed single mutations that eliminate the kinetic trap in native α1-AT, favoring the transition to the latent state or aggregation (Table S3) [227,238]. Conserpin harbors the known destabilizing mutation F366A[336], which in isolation would result in the formation of a large, potentially destabilizing cavity; however, we observe the compensatory mutation V364F[334] and introduction of a coordinated salt-bridge network between D256[228], E257[229], K368R[238] and N367D[337] (Fig. 6B).

Intriguingly, conserpin also harbors the potentially destabilizing mutation W238K[210] (Fig. 6B), which would likely weaken hydrophobic packing and introduce a large cavity. However, W238K[210] now forms backbone polar contacts with E363[333], and in turn may function as a solvent barrier that

shields the hydrophobic core. Furthermore, the mutations I229Y[201] and A284V[256] may provide extra solvent protection and a small increase in hydrophobic packing within the B/C barrel core (Fig. 6B). MD simulation also shows the presence of a transient salt bridge between W238K[210] and D256[228]. Extending the analysis to the C-sheet, conserpin also contains L224K[196] and S285E[257], which staples s2C and s3C together, further stabilizating the native state (Fig. 6B). Finally, L241E[213] and N228Y[200] are within close proximity of the B-sheet hydrophobic core and the same region in which citrate was found to bind and stabilize α1-AT, thus potentially providing extra stability [273]. Taken together, these changes may infer a substantial increase in core nucleation rates during early protein folding and subsequent stabilization of the native state that resists unfolding, consistent with our kinetic unfolding and refolding data (Fig. 1).

Conserpin is less frustrated than α1-AT

We next investigated the distribution of energy within the structures of conserpin and α1-AT using a mutational frustration analysis, which explores the influence of localized sequence and conformational perturbations on the energetic frustrations [274]. The overall networks obtained for both proteins were similar, presenting a vast majority of favorable contacts around the protein core with small patches of frustrated interactions (Fig. S7A). These patches are connected by a few highly frustrated contacts that may be implicated in dynamic information transfer within the serpin fold. Despite the similarities, a more detailed inspection of the networks reveals important differences. Of the total contacts made by hF, 38 % are minimally frustrated in α1-AT compared to 47 % in conserpin. Further, the number of highly frustrated hF contacts is almost negligible in conserpin; 1.6 %, compared to 7 % in α1-AT. In particular the Y160W[132] mutation in conserpin that leads to more favorable interactions at hF also reduces the frustration of this residue in comparison to Y160 in α1-AT (Fig. S7B, C). A similar trend was observed in hC but no noticeable differences were detected in other regions (Table S3). These observations are consistent with our structural and dynamics data regarding hF and further implicate both hC and hF in the folding and stability of conserpin.

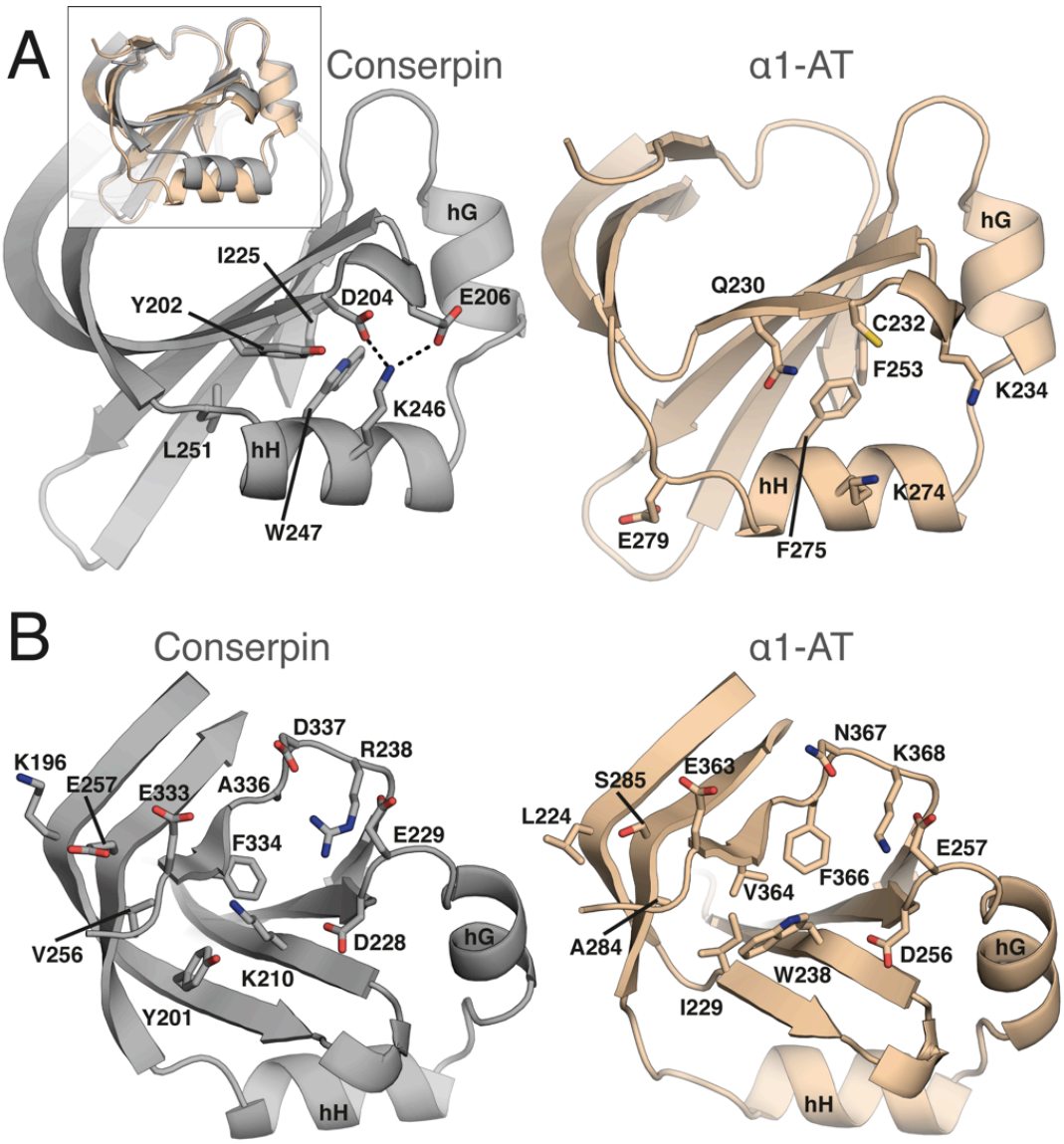


Figure 6. Structural analysis of the B/C barrel in conserpin (grey) and α1-AT (wheat). **(A)** Stabilizing hydrophobic mutations surrounding F275W[247]; **(B)** Remodeling of the inner barrel, surrounding W238K[210].

Biophysical and structural analysis of Z-conserpin

Intrigued by the stabilizing electrostatic interactions in the breach region of conserpin, we assessed the effect of introducing the disease-causing Z-mutation, E342K[313], into conserpin. $Glu342_{\alpha1\text{-}AT}$ is located in the breach, with mutation to lysine resulting in an increased propensity of α1-AT to polymerize in the endoplasmic reticulum of hepatocytes, leading to a lack of secretion into the circulation [275]. Studying the effects and mechanism of the Z-variant in α1-AT is particularly

difficult due to expression as insoluble aggregate and short half-life of soluble protein [261]. The most likely mechanism of Z-variant polymerization involves perturbation of the folding energy landscape, thus increasing the lifetime of the polymerogenic intermediate ensemble [227,235,265]. Although there is evidence to suggest that the Z-mutation also results in structural and dynamic changes to the native state [264,266,276], the crystal structure of Z α1-AT has not been determined. In order to investigate the effects of a destabilizing mutation on conserpin, we introduced E342K^{313} into conserpin to produce Z-conserpin. Z-conserpin expressed well as a soluble monomer in *E. coli*, which has not been possible with Z α1-AT [261].

Z-conserpin showed the same inhibitory profile as conserpin (Fig. S8F, G). We also determined an X-ray crystal structure of native Z-conserpin to 2.3 Å resolution (Table S1). The structure of Z-conserpin is essentially identical to native conserpin (backbone RMSD = 0.23 Å; Fig. S8H); the sole differences surrounding E342K^{313} are small side-chain shifts of K342^{313} and K290^{262}, most likely as a result of electrostatic repulsion (Fig. S8G). Interestingly, the single point mutation in Z-conserpin resulted in a significant loss of thermostability, supporting the important role of E342 in the serpin scaffold. Variable temperature far-UV CD melting curves in 2 M GuHCl, showed Z-conserpin to have a T_m of 60.7°C, whilst conserpin has a T_m of 72.5°C (Fig. S8A). Native PAGE shows conserpin to remain monomeric except when heated to 90°C for 10 minutes, whilst Z-conserpin has a complete loss of monomer at 80°C and appears to form a slightly higher molecular weight non-native species at 70°C when heated for 10 minutes (Fig. S8B). Gel filtration shows that Z-conserpin retains reversible folding after chemical denaturation and rapid dilution (Fig. S8C). We verified the reversible folding via intrinsic fluorescence equilibrium unfolding and refolding in the presence of GuHCl, which fits well to a two-state equation and indicates complete reversible refolding (Fig. S8D). The equilibrium unfolding and refolding curves overlay almost perfectly, revealing a midpoint of denaturation, $[D']_{50}$ to be 2.51 ± 0.01 M, an equilibrium *m*-value, $m_{D\text{-}N}$, of 5.18 kcal mol^{-1} M^{-1}, and a stability $\Delta G_{D\text{-}N}$, of -12.8 kcal mol^{-1}. In contrast to conserpin, all values have been lowered, with a significant reduction in ΔG and a ΔΔG of -10.04 kcal mol^{-1}. As with

conserpin, our equilibrium data did not reveal the presence of an intermediate species. We therefore performed equilibrium unfolding using bis-ANS fluorescence, observing the presence of a fluorescent peak at ~ 2.5 M GuHCl, that is slightly broader and more intense than observed in conserpin, but also consistent with the $[D']_{50}$ as determined by intrinsic fluorescence (Fig. S8E).

One important caveat to note is the lack of identical sequence conservation between α1-AT and conserpin surrounding E342K[313]. Specifically, the presence of K343E[314] in conserpin may partially negate the effects of E342K[313], where by K343E[314] forms a salt bridge with $K165_{conserpin}$. As such, future studies of the double mutant E342K[313] and $E314K_{conserpin}$ would be insightful.

The structure of Z-conserpin represents the first crystal structure of a serpin containing the Z-mutation, and reveals essentially no structural changes to the native state. This disagrees with reports of structural perturbations within the native state [264,266,276] and favors the mechanism of Z-variant polymerization via a folding intermediate. However, the intermediate versus native state polymerization mechanisms may be reconciled if the intermediate ensemble is native-like in structure. Considering the evidence in support of this for a wide range of proteins [277], our data is therefore consistent with the Z mutation altering the folding energy landscape, possibly by lowering the kinetic barrier of the unfolding transition to the intermediate ensemble [265], thus extending the lifetime of the polymerogenic intermediate ensemble (Fig. S8E).

Conserpin undergoes the transition to latent form

A consequence of serpin metastability is the presence of a highly stable latent state, in which the intact RCL is embedded within the A-sheet, formed when the serpin undergoes the S->R transition in the absence of protease [89]. We wondered whether the increased stability of the native state of conserpin would affect its ability to undergo the transition to the latent state. We were able to induce the formation of the latent state of conserpin by incubation at 76°C for 5 h and 68°C for 1 h in 10 μM in 50 mM Tris, 90 mM NaCl, pH 8.0 (Fig. S9A). These conditions are similar to those required for α1-AT, with the exception that citrate is necessary for α1-AT to resist

polymerization by binding a pocket between the A and B β-sheets [273,278]. The observed polymerization resistance of conserpin in the absence of citrate is consistent with its favorable folding properties and provides a unique precedent for future mutagenesis and folding studies. In order to confirm that conserpin forms a properly folded latent state, we determined its X-ray crystal structure to 1.45 Å resolution (Table S1). Structural analysis reveals successful adoption of the latent conformation, with the RCL inserted between s3A and s5A (Fig. S9B). In comparison to the latent structure of α1-AT, the newly inserted s4A of conserpin extends its β-sheet contacts, potentially providing a larger contribution to stability. This is particularly supportive given that the RCL of conserpin is shortened by one residue; warranting further study into the energy landscape and stability of conserpin's latent state, and how this relates to the insertion kinetics during activity, especially in the context of engineered RCL variants.

Conclusions

Our study reports for the first time, the successful engineering of a reversibly folding serpin that is highly resistant to polymerization and aggregation, even after the introduction of the polymerogenic, disease-causing Z-mutation. Structural analysis reveals the presence of many context dependant and stabilising interactions in regions that are known to be important for proper serpin folding. These include the stabilizing interactions surrounding the D-helix, introduction of a cooperative salt bridge network in the A-sheet that may resist aberrant RCL insertion, optimization of A-sheet hydrophobic core residues that have positive effects on folding and the energy landscape, stabilising mutations in the F-helix that should raise the energy barrier for RCL insertion, and potentially improved packing around the B/C barrel. Although some single mutations in these same regions have been reported to stabilise α1-AT, we found that many mutations within conserpin act together in a cooperative fashion.

Our analysis provides insights into important regions that may dictate the ruggedness of the serpin folding landscape. More generally, manipulation of the folding landscape in preference to stabilization of the native state may be a more effective strategy for optimizing the folding behavior of proteins, a process that may be informed by evolutionary sequence information. The successful design and synthesis of a highly stable, reversible folding, aggregation resistant and functional serpin begs the question of why such a molecule has not (yet, at least) been observed in nature. One possibility is that the conformational modulation of serpin activity by cofactors [279-281], for example in antithrombin [282,283] and PAI-1 [284], requires compromises in the folding energy landscape, reminiscent of the well-known stability-function trade-off observed for many enzymes [144,155,196]. Our findings support this hypothesis, where the divergent evolution of serpin clades and activities each developed their own regulatory capacity and compromises of their energy landscapes; thus, conformational modulation of activity is not highly conserved across the entire serpin superfamily. The subsequent application of consensus design effectively removes the regulatory mechanism of each member in our sequence alignment, resulting in a smoother energy and folding landscape that is likely associated with the observed two-state reversible folding and aggregation resistance. Finally, the fragile nature of serpin folding has thus far hindered residue-level kinetic characterization of all species on the folding pathway using phi-value analysis [285]; the robustness of conserpin may finally provide the basis, as a model system, for such characterization.

Methods

A curated representative structure guided alignment of 212 serpins was generated as previously described [241] and used to generate a consensus sequence as described previously [137]. Protein expression and purification, SI measurement and spectroscopic analysis was performed as described previously [228]. Protein Crystallography was performed at the MX1 and MX2 beamlines at the Australian Synchrotron [286]. All experimental and computational methods are described in detail in SI Methods.

Author contributions

BTP, JAI, SM, SPB and AMB designed the study. JAI, JCW, LZ, SPB and GIW performed the protein design. SK, EMM, MP, WD, and BTP performed the protein expression, purification and CD thermal melt experiments. BTP, NAB, SM and SK performed the crystallography. BTP performed molecular dynamics simulations and analysis with contributions from IK. DEH assisted in biophysical data analysis. MGSC performed the frustration analysis, structural analysis and contributed to discussion of data. SK, JH and AAN performed the folding kinetics and equilibrium measurement experiments. BTP and SK generated figures. BTP, SM, AAN and AMB wrote the manuscript.

Acknowledgements

We thank Jane Clarke, Andrew Ellisdon, and Gordon Lloyd for helpful discussions and advice. AAN and JJH are supported by the Wellcome Trust (grant number WT 095195). SM acknowledges fellowship support from the Australian Research Council (FT100100960). GIW is an Australian Research Council Discovery Outstanding Researcher Award Fellow (DP140100087). AMB is a National Health and Medical Research Senior Research Fellow (1022688). JCW is an NHMRC Senior Principal Research fellow and also acknowledges the support of an ARC Federation Fellowship. We thank the Australian Synchrotron for beam-time and technical assistance. This work was supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (www.massive.org.au). We acknowledge the Monash Protein Production Unit and Monash Macromolecular Crystallization Facility.

Supporting Information

SI Methods

Design of conserpin

Conserpin was designed from an alignment of 212 sequences within the serpin superfamily; this alignment was based on the alignment of 219 serpin sequences produced by Irving *et al.* [241]. In order to remove bias from the alignment, we removed redundant sequences above 95% similarity using CD-HIT [157], which resulted in 212 sequences. Application of a 100% consensus algorithm was applied over all 212 sequences (Dataset S1), resulting in a single sequence, which we call *conserpin*.

Protein expression and purification

The conserpin gene was synthesized by DNA 2.0 (CA, USA) and cloned into a pLIC-His vector using Ligation Independent Cloning. Proteins were expressed in *Escherichia coli* BL21 (DE3) using autoinduction media *(Overnight Express Instant TB* Medium, Novagen) at 28 °C until an OD_{600} of ≥10 was obtained. Cells were harvested by centrifugation and resuspended in lysis buffer (25 mM NaH_2PO_4, 300 mM NaCl, pH 8.0; 2 mM β-mercaptoethanol, PMSF and lysozyme). Cells were disrupted using sonication and whole cell lysate collected via centrifugation. 10 mM imidazole was added to the collected whole cell supernatant and applied to a 5 ml His Trap HP nickel-affinity column (GE Healthcare, UK). The column was washed with 20 mM imidazole in lysis buffer before bulk elution with 250 mM imidazole. Conserpin was further purified using size-exclusion chromatography (Superdex 200 16/60, GE Healthcare, UK) in 50 mM Tris, 90 mM NaCl, pH 8.0. Purified conserpin was concentrated to ~ 200 μM, as measured using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, DE, USA), and stored at -80 °C. Expression of conserpin was verified using mass spectrometry (MS) with a Matrix-Assisted Laser Desorption Ionization Time-of-Flight Time-of-Flight (MALDI-TOF-TOF). Approximately 10 mg of pure monomeric protein

was obtained from a 250 ml autoinduction culture. Z-conserpin was created by QuikChange method (Stratagene) against the conserpin plasmid using KOD DNA polymerase (Novagen). Oligonucleotides were synthesized by Geneworks (Australia). Z-conserpin was expressed and purified using the procedure outlined above. Wild-type α1-AT was purified from *E. coli* as per [287]. Z-α1-AT was purified from *P. pastoris* as previously described [288].

Gel filtration analysis of protein refolding

Samples of native and refolded protein were analyzed using a Superdex 75 10/300 column (GE Healthcare, UK) in 50 mM Tris, 90 mM NaCl, pH 8.0 and a 280 nm lamp. For refolding, the protein was unfolded for 1 h in 5 M GuHCl and then refolded by performing a 10 times dilution of the sample. Final samples were at a concentration of 2 µM and were centrifuged for 5 min to remove large protein aggregates before loading onto the column using a 500 µl loop.

Native PAGE

5 µl samples of protein at a concentration of 10 µM in 50 mM Tris, 90 mM NaCl, pH 8.0 were heated at various temperatures in a heating block and immediately placed on ice and mixed with ice-cold non-denaturing loading buffer to be run on 10 % (w/v) acrylamide native gels with a discontinuous buffer system. 9 M urea native gels were prepared as 10 % (w/v) native gels but with urea dissolved into the running gel to a final concentration of 9 M.

Transverse Urea Gradient (TUG) gel analysis

TUG gel analysis was performed as previously described [289], using a 0–9 M urea gradient perpendicular to the direction of electrophoresis. The running buffer used was 43 mM imidazole, 35 mM HEPES (pH 7.2-7.8). The unfolding of the protein was examined by applying native protein to the gel, and refolding by applying protein pre-denatured in urea.

Determination of protease inhibition kinetic parameters

The stoichiometry of inhibition (SI) and rate of association (k_{ass}) were measured at 37 °C in 50 mM Tris, 150 mM NaCl, 0.2% (w/v) PEG 8000, pH 7.4 as previously described [290]. Conserpin was incubated with trypsin for 30 min before measuring residual trypsin activity to calculate the SI. Trypsin (T1426, Sigma-Aldrich, MO, USA) activity was measured using the substrate S-2222 (Chromogenix, Italy). When measuring progress curves to calculate k_{ass} the final concentration of trypsin was 0.04 nM. For complex gels conserpin was incubated with trypsin in a 2:1 molar ratio of serpin:protease for 30 min at 37 °C in 50 mM Tris, 150 mM NaCl, 0.2% (w/v) PEG 8000, pH 7.4.

Characterization of Thermal Stability

Thermal stability of purified serpins was measured by circular dichroism. CD measurements were performed using a Jasco 815 spectropolarimeter; 0.2 mg/ml protein in PBS (140 mM NaCl, 2.7 mM KCl, 10 mM PO_4^{3-}, pH 7.4) was used in a 0.1 cm path length cuvette. Thermal denaturation was measured by observing signal changes at 222 nm during heating at a rate of 1 °C/min. The T_m was obtained by fitting to a sigmoidal dose-response (variable slope) equation.

Equilibrium Measurements

A 6 M solution of guanidine hydrochloride (GuHCl) in TBS was combined in varying ratios with TBS buffer using a liquid handling robot to create a range of denaturant solutions from 0 – 6 M GuHCl. These solutions were subsequently mixed in an 8:1 ratio with 9 μM protein in TBS to give a final concentration of 1 μM protein. All solutions were left to equilibrate at 25°C for at least three hours, after which the fluorescence of each solution was measured on a Perkin Elmer LS55 fluorimeter using an excitation wavelength of 280 nm and an emission range of 300 – 400 nm. Readings were obtained from a 1 cm pathlength cuvette maintained at 25 ± 0.1°C. The experiment was repeated, but using 9 μM protein pre-unfolded in 6 M GuHCl to generate a refolding curve. These solutions were left to equilibrate for at least six hours before their fluorescence was

ascertained. Bis-ANS unfolding experiments were conducted in a similar manner, except with the addition of bis-ANS to a final concentration of 5 μM.

Kinetic Measurements

Folding was monitored by changes in fluorescence using a 350 nm cut-off filter and an excitation wavelength of 280 nm. All experiments were performed using an Applied Photophysics (Leatherhead, UK) stopped-flow apparatus maintained at 25 ± 0.1°C. For unfolding experiments, one volume of 11 μM protein solution was mixed rapidly with ten volumes of a concentrated GuHCl solution. For single jump refolding, the protein (11 μM) was unfolded in 6 M GuHCl and left to equilibrate for at least 30 mins before use. This was then mixed 10:1 with buffer to give a final concentration of 1 μM, resulting in some very complex traces. To correct this, we conducted double jump refolding, where the protein (12 μM in 2M GuHCl) was diluted 1:1 with 6 M GuHCl buffer (to give 6 μM in 4 M GuHCl). This solution was left for a variable delay time (1 —> 50 s) and was then mixed 1:5 with buffer (to give 1 μM in 0.67 M GuHCl). Double jump resulted in much cleaner traces and allowed for data fitting to a double exponential. Data collected from at least six experiments were averaged and traces were fit to a single (unfolding) double (double jump refolding) or triple (single jump refolding) exponential as appropriate. Due to mixing effects, data collected in the first 2.5 ms were always removed before fitting.

Data analysis of equilibrium and kinetic measurements

An Excel spreadsheet was used to derive the fluorescence average emission wavelength (AEW) for each of the equilibrated denaturant solutions. A plot of AEW against denaturant concentration (Kaleidagraph, Synergy Software) yielded the expected sigmoidal plot, which was fitted to a standard two-state equation to obtain the *m*-value ($m_{D\text{-}N}$), the denaturant activity 50% ($[D']_{50}$) and hence the stability of the protein in TBS buffer ($D\mathit{G}_{D\text{-}N}$). Both the unfolding and refolding AEW curves can be converted to Fraction Folded by first removing the baselines and then normalizing the resulting data.

All kinetic traces fitted well to a single exponential decay plus a linear drift term. Single jump refolding yielded at least three different phases, which made data analysis problematic. We suspect that at least one of these (the slowest) phases may be a proline phase since it disappeared in the double jump experiments. The resulting chevron plot from our double jump experiments showed rollover in the refolding arm (indicating the presence of a refolding intermediate) and a kink in the unfolding arm (indicating the presence of a high energy intermediate). Double jump experiments also identified the presence of a second refolding rate that was distinguished by amplitude analysis. The main rate was then fitted using Prism (Synergy Software) to the following equation to estimate all parameters:

$$\ln(k_{\text{obs}}) = \ln\left(\frac{1}{2}\left(-A_1 - \sqrt{{A_1}^2 - 4A_2}\right)\right)$$

where:

$$A_1 = -\left(k_f + k_{-1}e^{m_{-1}[D]} + k_2 e^{-m_2[D]} + k_{-2}e^{m_{-2}[D]}\right)$$

$$A_2 = \left(k_f\left(k_2 e^{-m_2[D]} + k_{-2}e^{m_{-2}[D]}\right) + k_{-1}e^{m_{-1}[D]} + k_{-2}e^{m_{-2}[D]}\right)$$

$$k_f = k_i e^{-m_i[D]}\left(\frac{1}{1 + \frac{k_i e^{-m_i[D]}}{k_d e^{-m_d[D]}}}\right)$$

k_i and m_i are the folding rate constant from the refolding intermediate (I) to the first transition state (TS1) and its associated *m*-value, k_d and m_d are from the denatured state (D) to the first transition state (TS1), k_{-1} and m_{-1} are unfolding from the high energy intermediate (I*) over TS1, k_2 and m_2 are folding from the high energy intermediate (I*) over TS2, k_{-2} and m_{-2} are unfolding from the native state (N) over TS2. By convention, k_{-1} is set as 100,000 s^{-1} and m_{-1} is set as 0 M^{-1}: $m2$ is thus the *m*-value between TS1 and TS2 while the ratio k_{-1}/k_2 informs on the difference in free energy between the two transition states.

Crystallization, X-ray data collection, structure determination and refinement

All crystals were grown using the hanging drop vapor diffusion method, with 1:1 (v/v) ratio of protein to mother liquor (0.5 ml well volume). For native conserpin, the protein was concentrated to 10 mg/ml. Crystals appeared within 5 days in 0.2 M magnesium chloride hexahydrate, 16% PEG-3350, 10 mM bis Tris, pH 7.5. For native Z-conserpin, the protein was concentrated to 9.6 mg/ml. Large, rectangular crystals appeared within 5 days in 20% (v/v) polyethylene glycol (PEG) 3350, 10 mM bis Tris (pH 7.5) and 0.2 M Magnesium chloride (hexahydrate) and did not grow further. For latent conserpin, the protein was concentrated to 10 mg/ml. Crystals appeared with 5 days in 0.1 M Hepes pH 8.0, 0.2 M Ammonium Acetate, 35% v/v MPD. All crystals were cryo-protected by the addition of 10 % glycerol prior to data collection.

Data for all three crystals was collected at 100 K using at the Australian synchrotron macro crystallography MX1 beamline. A 1.7 Å dataset was collected for conserpin and resolution cut to 2.4 Å, 2.3 Å for Z-conserpin and 1.45 Å for latent conserpin. Diffraction images were processed using iMosflm [164] and Aimless from the CCP4 suite [171]. Each dataset was processed in *P*1 and Laue group determination was achieved using Pointless within Aimless. Datasets were scaled and merged in their respective space-group and 5% of each dataset was flagged for calculation of R_{Free}, with neither a sigma nor a low-resolution cut-off applied to any dataset. A summary of statistics is provided in Table S1.

Structure determination proceeded using the Molecular Replacement method and the program PHASER [165]. A search model for conserpin was constructed from the crystal structure of native α1-AT (PDB: 3NE4) by removing solvent molecules. All other structures used native conserpin as the search model. A single clear peak for both the rotation and translation functions was evident and the molecules packed well within the asymmetric unit. Together with the unbiased features in the initial electron density maps, the correctness of the molecular replacement solutions was confirmed. Automated model building was performed using AutoBuild in the Phenix package [167].

All subsequent model building and structural validation was done using Phenix [167] and COOT [168]. Solvent molecules were added only if they had acceptable hydrogen-bonding geometry contacts of 2.5 to 3.5 Å with protein atoms or with existing solvent and were in good $2F_o$-F_c and F_o-F_c electron density. The coordinates and structure factors are available from the Protein Data Bank (5CDX, 5CE0, 5CDZ).

Structure analysis

For all analysis and MD simulations, missing atoms, side chains and residues were rebuilt using Modeller V. 9.12 [162]. In each instance, 50 models were built and the lowest DOPE (Discrete Optimized Protein Energy) scoring model was selected for further analysis. Hydrogen bonding and salt bridge values were calculated using the WHAT-IF web-server (Hekkelman *et al.,* 2010). Solvent accessible surface area was calculated using AREAIMOL as part of the ccp4 package with a default probe radius of 1.4 Å [171]. Total cavity volumes and related structures were calculated using the Depth web server [291]. Molecular graphics were prepared with PyMol Molecular Graphics, Ver. 1.5.0.4

MD system setup and simulation protocol

Molecular dynamics simulations were carried out on native conserpin and native α1-AT. Missing atoms, side chains and residues were modeled as described above. Chain termini were capped with neutral groups (acetyl and methylamide). Residues were protonated according to their states at pH 7. Completed structures were solvated in a rectangular simulation box leaving at least 10 Å of water shell thickness on all sides of the protein. System charges were neutralized with respective sodium or chloride counter ions. Protein and ions were modeled using the AMBER ff99SB force field [292] and waters were represented using the 3-particle TIP3P model [293]. All bonds involving hydrogen atoms were constrained to their equilibrium lengths with the SHAKE algorithm [294]. The resulting systems were subjected to at least 10,000 energy minimization steps to remove any clashes, followed by an equilibration protocol. During equilibration, we applied

harmonic positional restraints of 10 $kcal^{-1}$ mol^{-1} $Å^{2\ -1}$ to the protein backbone atoms, pressure was kept at 1 atm using Berendsen algorithm [175] and the temperature was increased from 10 K to 300 K as a linear function of time over the course of 1.2 ns, with Langevin temperature coupling. Relaxation times for temperature and pressure were 0.5 ps. Subsequently, we removed the restraints and performed a 5-ns simulation at constant isotropic pressure of 1 atm and temperature of 300 K. Electrostatic interactions were computed using an 8-Å cutoff radius and the Particle Mesh Ewald method for long-range interactions [295]. All MD simulations (equilibration and production) were carried out under periodic boundary conditions.

Production simulations of native α1-AT and conserpin were carried out in the NPT ensemble. Temperature was kept at 300 K using the Langevin thermostat with a collision frequency of 2 ps, whilst Berensden pressure coupling was used to maintain the pressure at 1 atm with a 2ps relaxation time. The simulation time step was 2 fs and snapshots were taken every 100 ps. Simulations were run in duplicate with Amber 14 [296], using PMEMD on a Nvidia K20m GPU for 500 ns.

MD analysis

Simulation trajectories were processed and analysed using a combination of Amber Tools 14, custom scripts and ProDy [180,296]. Graphs and plots were produced with Matplotlib [181]. Molecular graphics were prepared with PyMol ver. 1.5.3 [182].

Local Frustration Analysis

Local frustration analysis was conducted with the Frustratometer web server [297], using a completed model based on the crystal structures of conserpin and α1-AT (PDB: 3NE4) in the native state. Essentially, the energetic frustration is obtained by the comparison of the native state interactions to a set of generated "decoy" states where the identities of each residue are mutated. A contact is defined as "minimally frustrated" or "highly frustrated" upon comparison of its frustration energy with values obtained from the decoy states, as described [274].

Table S1*: Data collection and refinement statistics.*

Data collection	Native conserpin	Native Z-conserpin	Latent conserpin
Wavelength (Å)	0.9537	0.9537	0.9537
Space group	C 2 2 2_1	P 1 2_1 1	P 62
Unit cell dimensions (Å)	68.14, 76.12, 150.24, 90, 90, 90	49.22, 150.27, 51.015, 90, 94.74, 90	101.679, 101.679, 62.72, 90, 90, 120
Resolution (Å)	2.4	2.3	1.449
Number of measured reflections	30073	252442	2493126
Number of unique reflections	15325	32616	65513
Completeness (%)	97.77	99.64	99.94
Redundancy	2.0	7.7	38.0
R_{pim}	0.04451		
<I/σI>	9.20	30.83	151.41
Structure refinement			
Number of reflections	15304	32565	65500
Number of protein atoms	2647	5345	2853
Number of water molecules	58	243	351
R_{work} (%)	0.1942	0.2012	0.1566
R_{free} (5% of data) (%)	0.2582	0.2294	0.1851
CC1/2	0.997	0.99	0.826
CC*	0.999	0.997	0.951
RMSD bond lengths (Å)	0.007	0.003	0.012
RMSD bond angles (°)	1.03	0.69	1.57
Average B-factor ($Å^2$)	55.10	51.00	25.30
Protein	55.20	51.20	23.60
Solvent	52.30	46.40	38.80
Ramachandran			
Favoured (%)	97	98	97
Outliers (%)	0	0	1.4
MolProbity score	1.65, 99th percentile (N=8058, 2.40Å ± 0.25Å)	1.18, 100th percentile (N=8909, 2.30Å ± 0.25Å)	1.56, 78th percentile (N=3441, 1.449Å ± 0.25Å)
PDB ID	5CDX	5CE0	5CDZ

Table S2: Global physiochemical properties of serpins

Serpin	PDB code	*Tm*	H-bonds	Salt bridges < 7 Å	Accessible surface area (Å^2)	Total cavity volume (Å^3)
Conserpin	5CDX	>90°C*	178	112	17,280.7	2,257.6
α1-AT	3NE4	61.8°C	206	126	16,599.3	1,928.6
Thermopin	1SNG	67°C	178	207	16,852.1	2,227.9
Tengpin	2PEE	N/A	207	91	16,667.1	1,844.1

* A true melting tempterature (*Tm*) was not determined in our experiments. All other values were determined as per SI methods.

Table S3: Localized frustration in the C and F-helices. The degree of frustration by contact is defined in Ferreiro et al., 2007 [274].

		Frustrated interactions (% of total helix contacts		
Protein		*Minimally*	*Neutral*	*Highly*
C-helix	α1-AT	33.5	53.5	13.0
	Conserpin	37.9	59.6	2.5
F-helix	α1-AT	38.7	54.2	7.1
	Conserpin	46.9	51.5	1.6

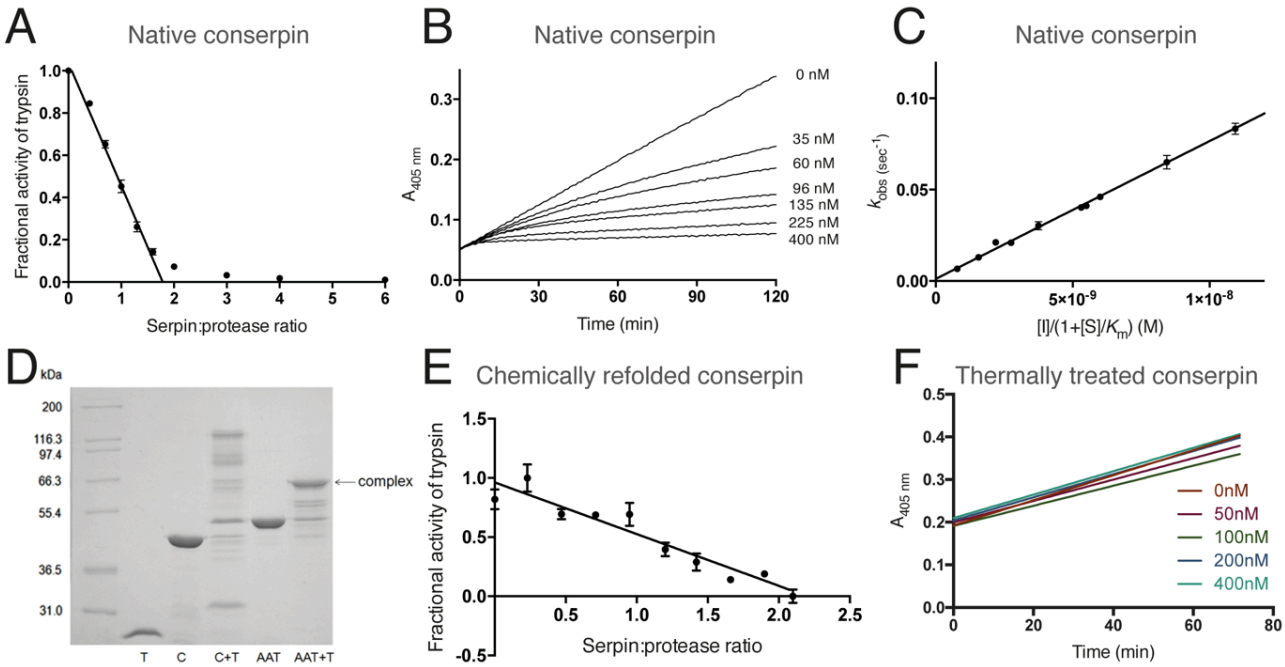


Figure S1. Activity of conserpin against trypsin. **(A)** SI (mean and standard error shown, n=6); **(B)** Representative traces of progress curves of trypsin activity (100 nM) in the presence of conserpin. The concentration of conserpin (inhibitor) is indicated next to each curve. $A_{405\ nm}$ is the absorbance measured at 405 nm (arbitrary units); **(C)** Rates of the progress curves were determined (observed rate constant, k_{obs}). k_{obs} for each inhibitor concentration [I] is plotted against [I]/(1+[S]/K_m) to find k_{ass} (mean and standard error shown, n=3). K_m is the Michaelis constant for trypsin cleavage of the substrate (S). The slope of the linear function was taken as the k_{assapp}, which was then multiplied by the SI to give the k_{ass}; **(D)** Formation of a serpin-protease complex and other species seen in SDS-PAGE formed at a ratio of 2:1 (serpin:protease) under reducing conditions. Lane 1, trypsin alone (T); lane 2, conserpin alone (C); lane 3, conserpin incubated with trypsin (C+T); lane 4, α1-AT alone (α1-AT); lane 5, α1-AT incubated with trypsin (α1-AT+T). Partial degradation of the complex is seen, which is probably due to excess trypsin. The expected MW of the conserpin-trypsin complex is 66.3 kDa (the MW of trypsin is 23.8 kDa). There are also some unusual higher molecular weight species not seen in the inhibition of trypsin by α1-AT; **(E)** SI (mean and standard error shown, n=3) of conserpin after chemical denaturation in 6 M GuHCl and rapid dilution into TBS; **(F)** Representative traces of progress curves of trypsin activity (100 nM) in the presence of conserpin that has been heated at 80°C for 20 minutes. The concentration of conserpin (inhibitor) is indicated next to each curve. $A_{405\ nm}$ is the absorbance measured at 405 nm (arbitrary units).

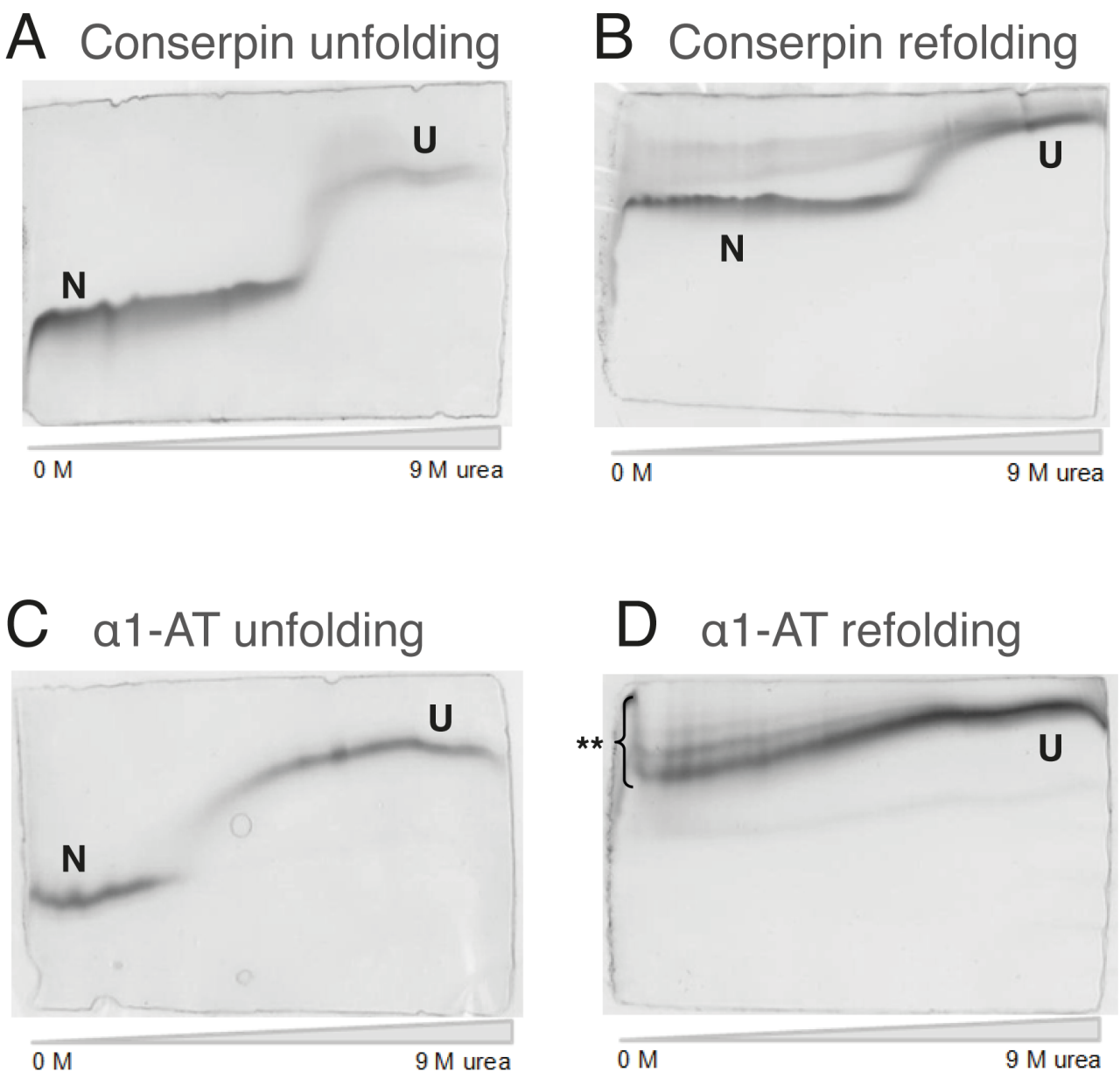


Figure S2. TUG gels show that conserpin is more stable than α1-AT in urea and more resistant to polymerization upon refolding. The **(A)** unfolding and **(B)** refolding of conserpin, and the **(C)** unfolding and **(D)** refolding of α1-AT. The TUG gels contain a gradient of 0 M to 9 M urea from left to right. N indicates the native species, U indicates the unfolded species and ** indicates polymers. α1-AT polymers can be seen as laddering of higher molecular weight species near the top of the gel. Differences between the position of N in **(A)** and **(B)** are the result of non-equal run time of the gel.

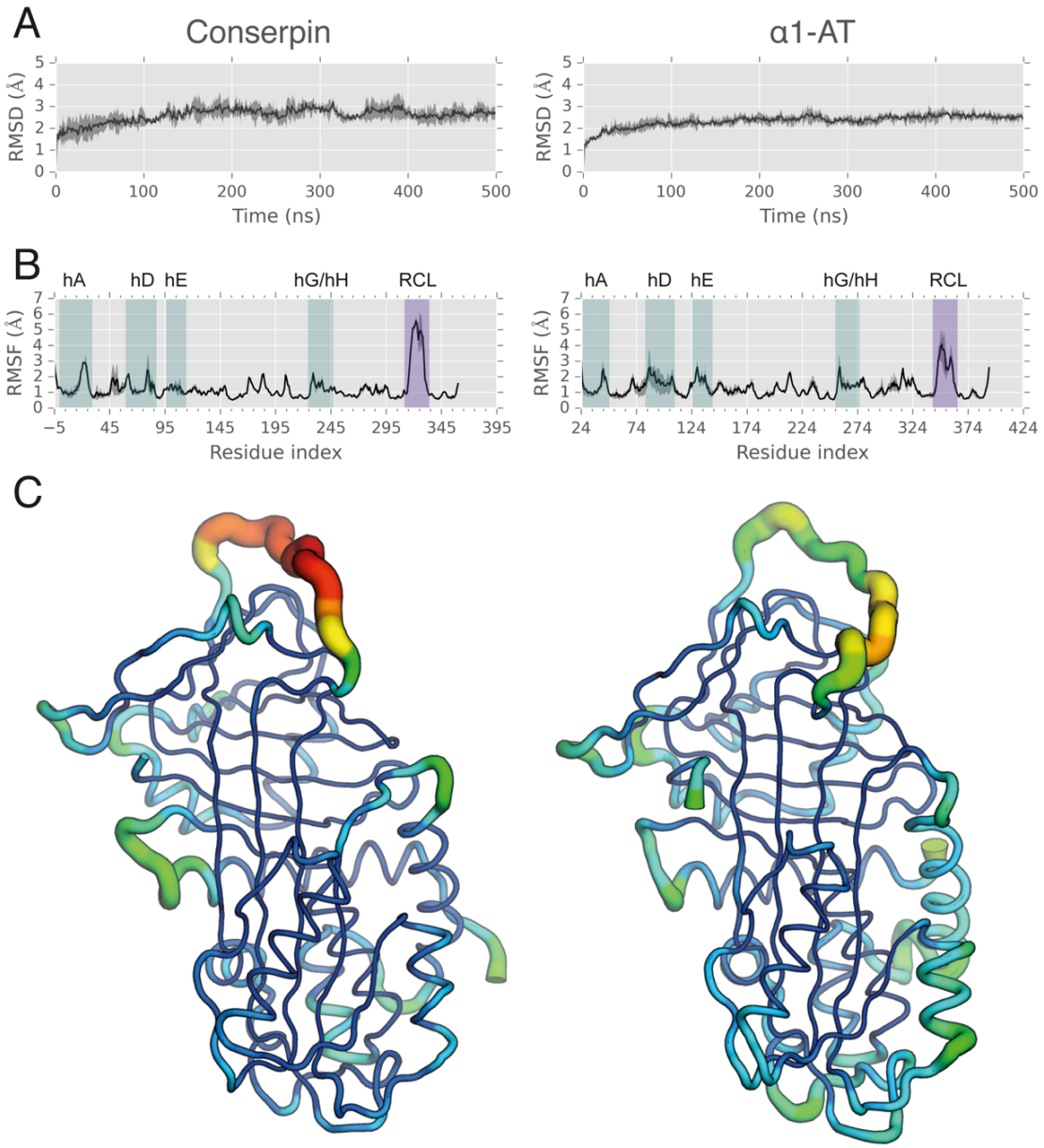


Figure S3. Molecular dynamics simulation of conserpin and α1-AT. **(A)** Root mean square deviation (RMSD) plots of Cα atoms in conserpin and α1-AT over 500 ns at 300 K. Plots show the mean RMSD (solid line) with the min/max variation (n=2); **(B)** Root mean square fluctuation (RMSF) plots of Cα atoms in conserpin and α1-AT over 500 ns at 300 K (n=2); **(C)** Conserpin (left) and α1-AT RMSF values embedded onto their corresponding crystal structures. Colors and representations as given in Figure 3A.

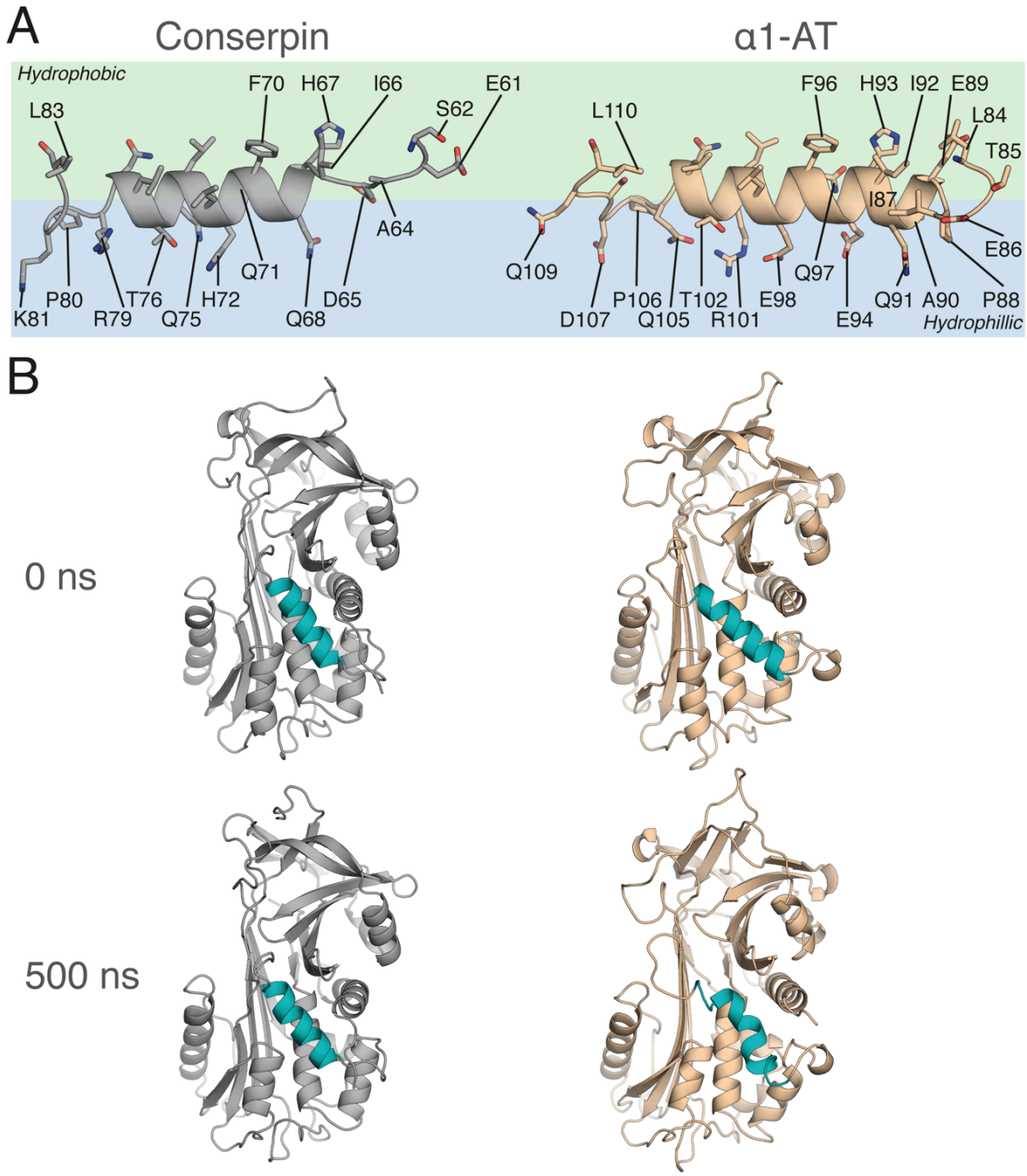


Figure S4. Inspection of the D-helix. **(A)** Differences in the D-helix between conserpin (grey) and α1-AT (wheat). The hydrophobic (green) section faces and packs against the hydrophobic core, whilst the hydrophilic (blue) section faces into solvent; **(B)** Partial loss of structure in the D-helix (cyan) of α1-AT after 500 ns of MD.

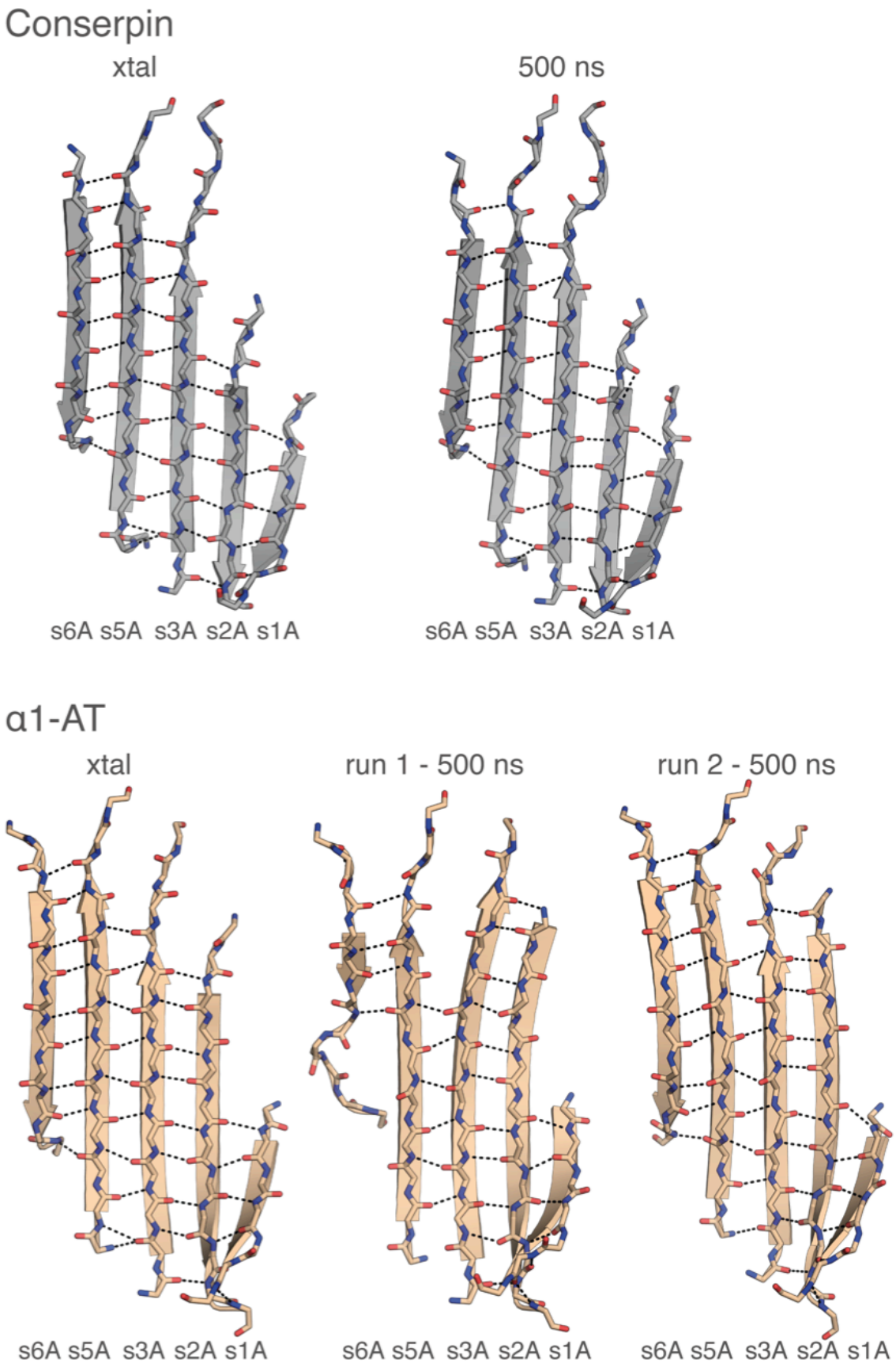


Figure S5. Hydrogen bonding differences between the backbone of sheet-A in conserpin (grey) and α1-AT (wheat) from crystal structure and 500 ns of MD, highlighting the increased propensity for separation between s3A and s5A in α1-AT.

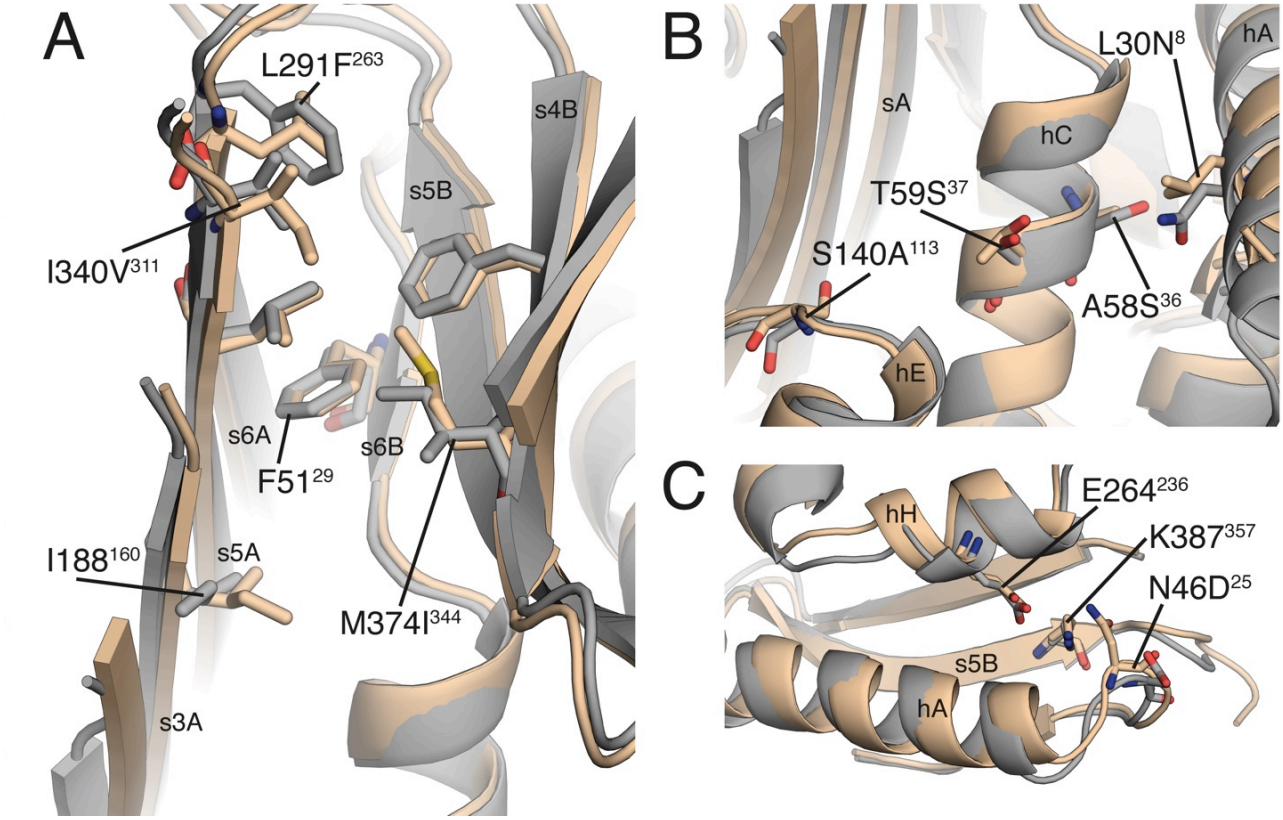


Figure S6. A-sheet hydrophobic core interactions in conserpin (grey) and α1-AT (wheat). **(A)** Comparison of the hydrophobic core surrounding F51[29] in conserpin and α1-AT; **(B)** Comparison surrounding the T59S[37] mutation; **(C)** Comparison surrounding K387[357], showing the salt bridge introduced by the N46D[25] mutation.

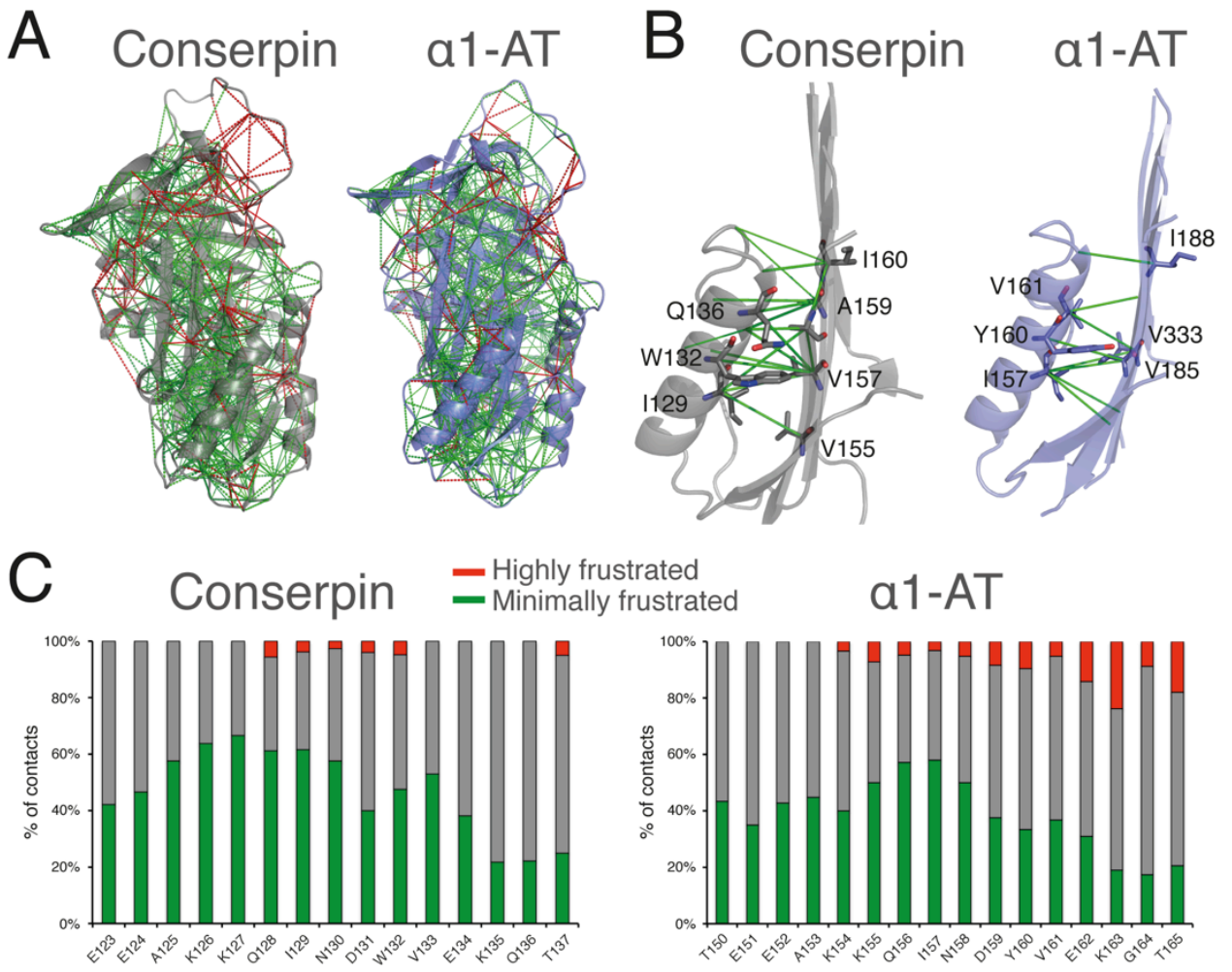


Figure S7. Analysis of energetic frustrations on conserpin and α1-AT. **(A)** Frustration networks embedded onto α1-AT and conserpin structures; **(B)** Close view of the F-helix contacts highlighting the positions of interacting residues; **(C)** Fraction of frustrated contacts per F-helix residue. Minimally, neutral and highly frustrated contacts are represented in green, grey and red respectively.

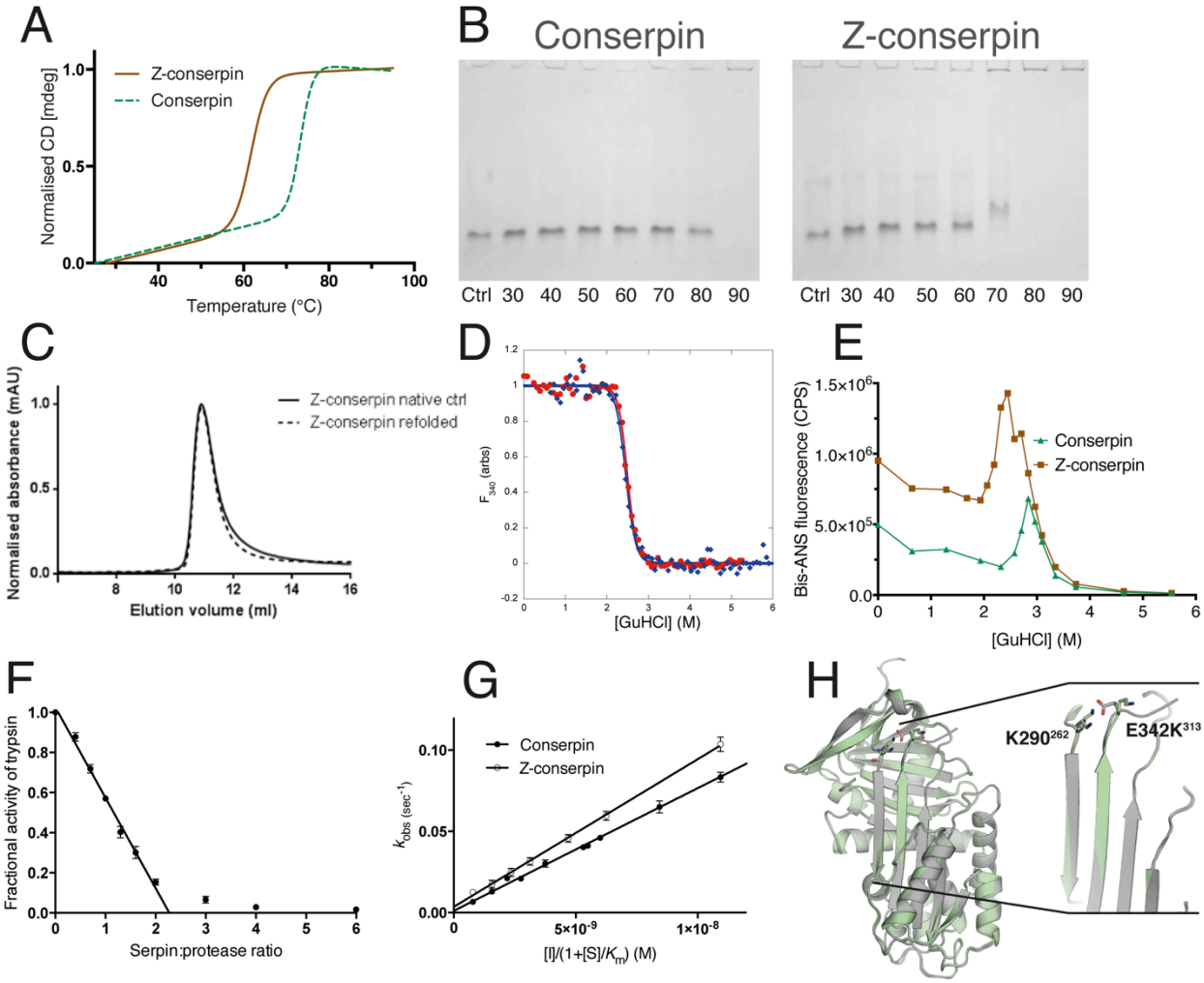
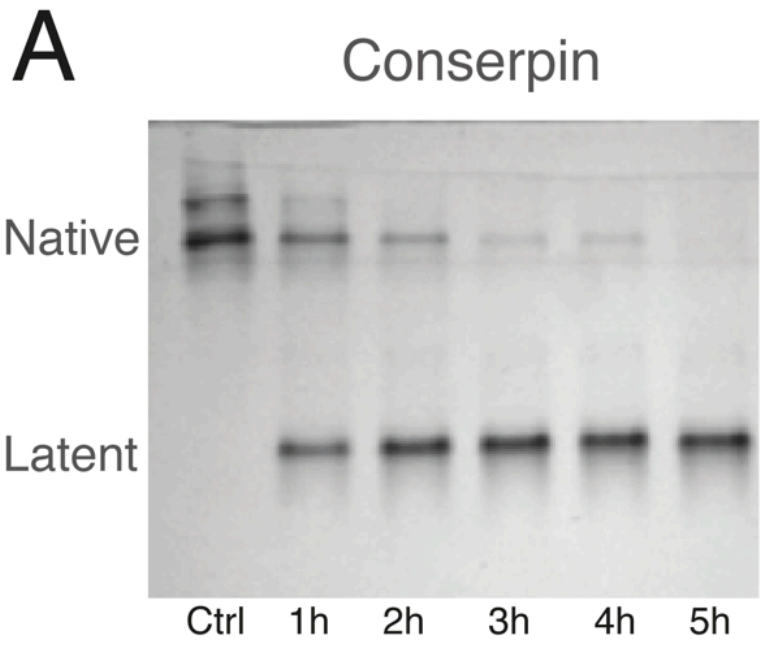


Figure S8. Folding, polymerization and structure of Z-conserpin. **(A)** Thermal unfolding of conserpin and Z-conserpin in 2M GuHCl; **(B)** Native PAGE of conserpin (left) and Z-conserpin (right), where samples were heated for 10 mins from 30C to 90C; **(C)** Gel filtration of native and refolded Z-conserpin as described in Fig. 1; **(D)** Equilibrium unfolding and refolding of Z-conserpin using intrinsic fluorescence at 280 nm; **(E)** Equilibrium unfolding of Z-conserpin under the presence of bis-ANS; **(F)** SI (mean and standard error shown, n=6) of Z-conserpin against trypsin; **(G)** k_{obs} for each inhibitor concentration [I] is plotted against [I]/(1+[S]/K_m) to find k_{ass} (mean and standard error shown, n=3). K_m is the Michaelis constant for trypsin cleavage of the substrate. The slope of the linear function was taken as the k_{assapp}, which was then multiplied by the SI to give the k_{ass}; **(H)** Structural alignment of conserpin (grey) and Z-conserpin (pale green), with a close up alignment, highlighting rotamer differences at E342K^{313}, resulting in the loss of a salt bridge.



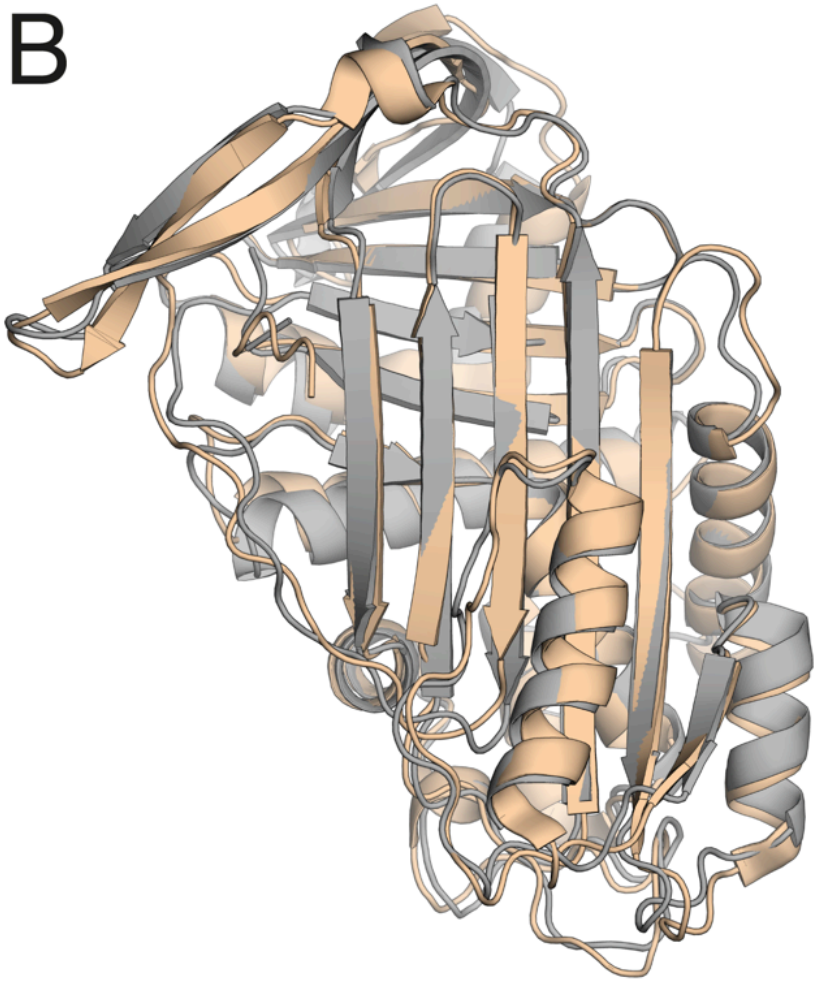


Figure S9. Formation and structure of latent conserpin. **(A)** Native gel containing 9 M urea after conserpin is heated at 76 °C for 1 to 5 h; **(B)** Structural alignment of latent conserpin (grey) and latent α1-AT (wheat; PDB: 1IZ2).

Chapter 6
Discussion and conclusion

Summary

In this chapter I draw together the results of this thesis along with the broader literature to present an in-depth analysis and comparison of consensus design as a method for engineering protein stability. Essential elements of design are thoroughly explored and their resulting effects are discussed. In context of the work in this thesis I am able to put forward a detailed strategy of application. Further, this analysis suggests the capacity for consensus design to affect biophysical properties other than thermodynamic stability, with smoothening of the protein energy landscape and reduction of immunogenicity. Finally, this chapter catalogues over thirty years of research in consensus design and highlights the value of this protein engineering methodology.

6.1 Overview of consensus design

Throughout this thesis, I have presented the consensus design and subsequent analysis of two proteins of unrelated structure and function. This experience has provided me with the unique position of not only being able to critique and appraise the approach, but to also put forward a detailed strategy of application. This chapter explores and discusses the essential elements of consensus design and their resulting effects on the biophysical properties.

Although the consensus design algorithm is a simple calculation to implement and compute, a firm understanding of how to best prepare and curate the input multiple sequence alignment (MSA) is lacking. In the initial stages of design, one must be very clear about whether they intend to mutate a target protein or generate a new protein sequence de novo, as this will greatly influence the composition of sequences used in the MSA and the optimal approach to implementation. From the work conducted in Chapters 2 and 5, along with examples in the literature (Table 6.1), two alternate implementations exist: (1) mutagenesis of a target with conserved residues, or (2) full-sequence design, each with subtle effects on the resulting biophysical and functional properties. In general, consensus design has four major components (Figure 6.1). First, it must be decided whether a specific target protein or a general protein family/fold will be subjected to consensus design, as this has ramifications for the second step of acquiring homologous sequences and how these sequences will be processed and curated. The third step is an iterative process requiring assessment of several multiple sequence alignment (MSA) regimes and removal of disruptive sequences from the MSA. Ideally, this process is repeated until the alignment best reflects the major structural features of the target protein or fold – a crucial and often difficult task depending on the level of homology and conservation of the composing sequences [298,299]. Finally, amino acid conservation is calculated by the standard consensus algorithm [111]. Conservation may then be used directly for mutagenesis, further filtered by computational/statistical methods, be used to weight directed evolution libraries or to compute a *de novo* consensus sequence - all depending on the initial design question.

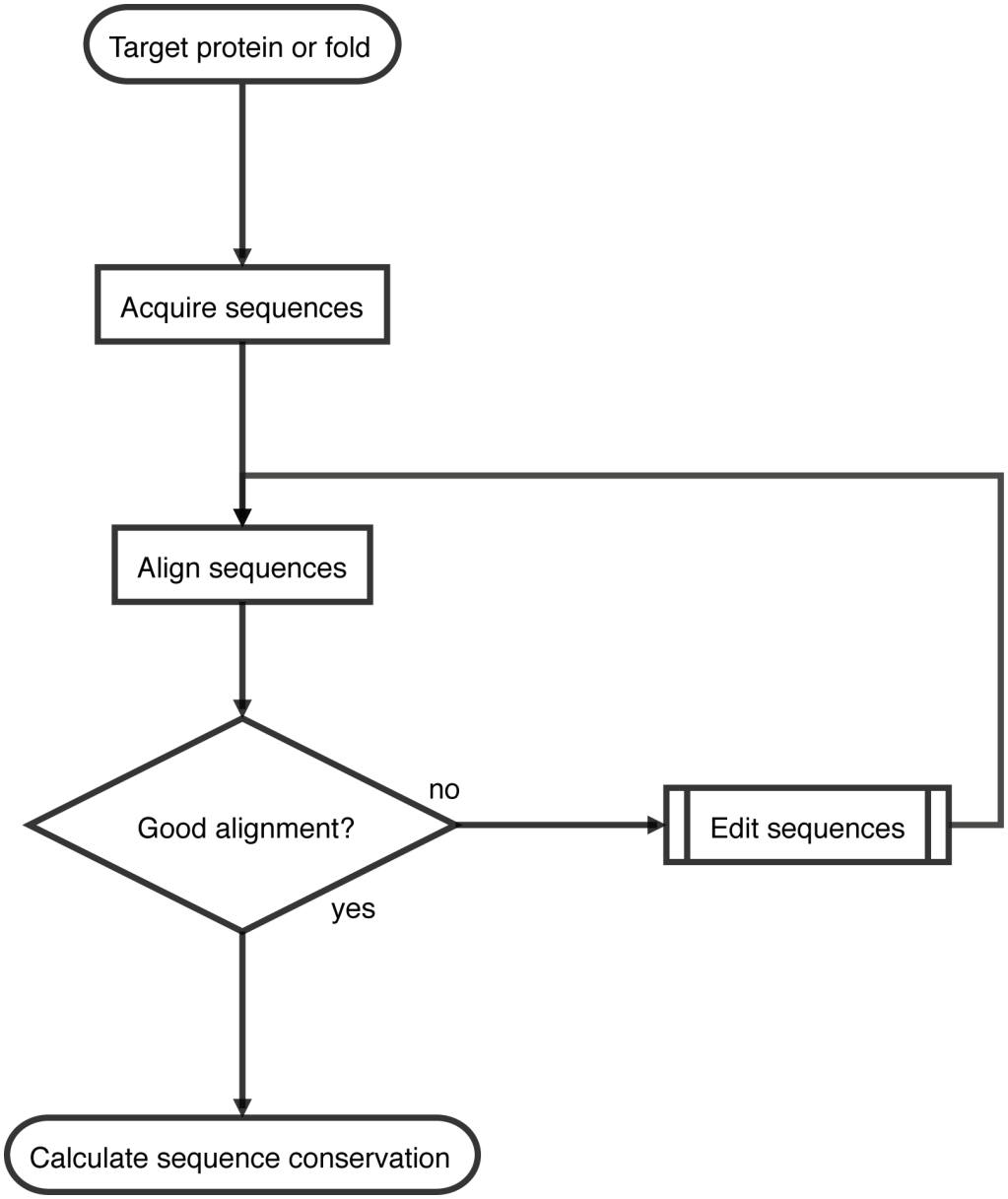


Figure 6.1. A basic flow chart for consensus design.

6.2 Factors to consider during consensus design

6.2.1 Acquisition of homologous sequences

The most efficient way to acquire sequences is via sequence alignment databases such as Pfam [117,300], Prosite [116], SMART [301] and Superfamily [118]. These databases contain small, manually curated seed alignments for the development of hidden markov model (HMM)

profiles [302] or motif specific rules and patterns [115], which are then be applied to larger collections such as the UniProtKB/Swiss-Prot [114], Protein Data Bank (PDB) and NCBI sequence [113] databases. Database coverage as of 2016 is typically high, with Pfam entries matching 76.1% of the UniProtKB [117], and the majority of common folds and large protein families being well represented [116-118,301]. This particular approach was used in the development of the two consensus designed proteins in this thesis, FN3con (Chapter 2) and conserpin (Chapter 5), producing exceptional improvements to stability. In particular, the alignment used for FN3con was derived from a hand curated seed alignment from the Prosite database [116], that was expanded to the UniProtKB database. Whereas, conserpin was derived from a smaller hand curated sequence alignment developed by Irving et al., 2000 [241].

If a protein target is not well represented in existing alignment databases, the best approach is to query the UniProtKB or NCBI NR sequence databases for a small number (< 20) of the most homologous sequences; which can be used to curate a representative alignment that can be subjected to HMM profiling with the HMMER suite [302], and subsequently align to more distant homologues. In the unfortunate instance that the target protein is minimally or not represented in the UniProtKB or NCBI NR sequence databases, the next option is to generate diversity through the use of neutral drift studies [119,120]. In neutral drift experiments, a target protein is subjected to rounds of random mutagenesis and selected purely on whether it is folded and/or functional, and then is sequenced. This approach was used as a means of generating unbiased sequence diversity in the consensus design of a chorismate mutase, showing the method to be successful with fewer than 30 selected sequences [119]. Therefore, this approach could also be utilized in difficult cases such as those lacking sequence database representation or poor sequence homology.

6.2.2 Homology

In general, random mutations are deleterious to thermodynamic stability, with surface residues contributing a mildly destabilizing mean ΔΔG of ~0.6 kcal mol^{-1}. Mutations to core buried residues contribute to a wider distribution of ΔΔG values with a stronger destabilizing mean ΔΔG of ~1.4 kcal mol^{-1} [13,153]. As covered in the introduction (Chapter 1.7.2), consensus design likely works because conserved residues are on average more likely to be stabilizing (50%) [141] than a random mutation (8-29%) [153]. Consensus design is therefore capable of combining stabilizing features across the MSA of protein homologues that no single protein needs or has been able to amass through evolution. Fundamentally, consensus design can only extract features that exist predominantly in each position of the alignment, meaning that the quality of the composing sequences and resulting alignment is important to successful design. Homology describes how sequentially similar members of an alignment are to one another and is related to their evolutionary distance. The effect of sequence homology on consensus design is poorly understood and highly likely to be a function of the target protein's biophysical properties, evolutionary history and the taxonomic representation in sequencing databases. Theoretically, inclusion of evolutionarily distant or diverse sequences should improve the probability of identifying more conserved features, as increased distance may imply increased sampling of sequence space. Although there are reports that that too little [150] and too much diversity of the input MSA is problematic [119,148,303], this area has not been thoroughly explored.

Determining the right amount of diversity is challenging. Sullivan and colleagues noted this in the consensus design of a triosephosphate isomerase (TIM) using comparisons of two Pfam alignments from database versions 18 and 22 [148]. The input sequences of version 18 were a roughly even mixture of bacterial and eukraryotic sequences, resulting in a weakly active and poorly folded consensus protein. However the version 22 alignment was composed of predominantly bacterial sequences and resulted in a well folded and fully active protein [148]. To further complicate matters, Sullivan filtered the version 22 sequences to be roughly the same

length, and removed duplicate entries, which may have had other effects and therefore reduce the general applicability of this approach. Highly successful designs such as FN3con (Chapter 2) [137] and cLRRTM2 [135] used sequences that were predominantly or exclusively from higher order eukaryotes without the need to filter based on sequence length, suggesting that spanning the MSA over taxonomic domains or kingdoms may negatively affect results. Parmeggiani also observed a similar problem as a result of too broad protein family selections of armadillo repeat proteins [303]. However, rather than filter or remove sequences, they sub-classified their MSA into closer taxonomic groups and combined conserved residues from each sub-classification.

Extending homology too far may result in poor conservation, which can prevent accurate alignment and lead to design failure. For example, sequence conservation within the β-defensin family is less than 33%, even though structural similarity is very high [304]. Here, alignment within a specific species is challenging and consensus design would likely be impossible using natural sequences [305] – leaving neutral drift studies as the only solution for generating homology and a chance of successful alignment [119]. Managing homology is therefore a balance between sequence similarity, which is good for computing a MSA, and sequence diversity, which provides a greater coverage of sequence space that can be sampled during design.

6.2.3 Bias

In contrast to diversity, the weighting or skew of the MSA may bias consensus design towards a predominant clade, such as a taxon, species or protein classification [119]. This is typically the result of preferences from genome sequencing projects, which tends to over-represent particular species or proteins in sequence databases. Bias is more likely to be an issue for domains, motifs or repeat proteins that are found within larger proteins. In some instances, bias may be intentional, as to preserve functional networks of a protein family from a single species or sub-classification. In the interest of purely identifying robustness and stability, it is reasonable to assume that bias and over representation should generally be avoided; these traits may mask conserved and possibly

stabilizing features from other less represented evolutionary lineages. Bias reduction of natural sequences can be performed with relative ease using the sequence clustering software CD-HIT [157] or by using likelihood-based methods to account for phylogeny [306]. The current body of consensus design literature is relatively sparse in the application of bias or redundancy reduction methods, with the few representative cases including Chapters 2 and 5 and refs [137,152,307-309]. In these cases, at least 90% redundancy reduction was used, where sequences with more than 90% sequence identity to one another were removed. However, the effect of these methods in combination with varying levels of sequence homology and sequence counts has yet to be thoroughly explored and understood. Indeed, consensus design appears to function regardless of whether the alignments are biased.

6.2.4 Sequence count

One of the key advantages of consensus design over other sequence-based methods is its ability to identify stability enhancing mutations from a MSA with as few as four members. Examples include Subtilisin BPN' (from 4 members) [146], and FN3 repeats (15 members) [150]. In the latter study, the top 10 most stable sequences were less successful in promoting thermodynamic stability than all 15 members [150], demonstrating that even the less stable sequences contribute to the overall stability of the resulting consensus design. In this case more sequences provides greater diversity, thus improving the signal to noise ratio, and therefore the detection of conserved residues in weakly conserved regions. This effect is exemplified by recent consensus designs using very large alignments, such as FN3con (2,123-sequences, ΔT_m of >27°C) (Chapter 2) [137], and cLRRTM2 (6,271-sequences, ΔT_m of 32°C) [135]. However, it is still poorly understood how the number of sequences utilized in consensus design affects the resulting protein product. This has led to researchers using trial-and-error design processes. Further work on this question may resolve what is currently a frustrating and inefficient iterative design process.

6.2.5 Quality of the sequence alignment

After the acquisition of homologous sequences, they must then be aligned before conservation can be calculated. Difficulties arise with multiple sequence alignments containing sequences of varying length, or when there are clusters of sequences that are locally, but not globally homologous [305,310]. Large insertions and deletions between members can affect the identification of weakly conserved positions, for example resulting in the design of a weakly active and poorly folded protein [148]. In this specific case filtering the homologous sequences to be roughly the same length and removing duplicate entries, the design was greatly improved, resulting in a well-folded and fully active protein. Interestingly, the only sequence differences between the "raw" versus the filtered design were in predominantly non-conserved stretches of the protein. By sequence assessment alone, there was no obvious reason for why these differences resulted in vastly different biophysical properties. It is possible that filtering sequences to those that are more homologous improved the alignment, which allows for better identification of weakly conserved residues [148].

Generating a "good" multiple sequence alignment can be difficult, and may actually be considered more art than science [298]. Unfortunately, MSA methods tend to vary significantly and there is currently no quantitative measure for the quality of alignment [310-312]. This is further compounded by homology, bias and sequence count and its convoluted interplay with the particular evolutionary history of a target protein and its family. Therefore it is highly recommended to carefully examine resulting alignments prior to consensus design, possibly with overlay of secondary structure to gauge conservation boundaries and gaps [313]. Iterative rounds of phylogenetic assessment and sequence pruning can improve alignment quality, which should be inspected for aligned columns that correspond with structural motifs or secondary structure elements that have few insertions, deletions, and gaps.

6.2.6 The fundamental limitations of protein folds

As variations in the success of consensus design may arise from quality and quantity of the MSA, it is worthwhile to also consider the physiochemical limitations of particular three-dimensional topologies and how this affects the attainable stability. Indeed, this property of protein structure has been thoroughly explored, where it has been shown that certain structures or folds can be encoded by more sequences than others [314-326]. This is particularly so for the immunoglobulin (Ig) fold, of which the FN3 domain is a member of. The Ig fold is considered to be one of the most evolvable due to its tight packing, numerous convergent evolution events and high degree of sequence variability [156,327]. The ability of a topology to be encoded by a large number of sequences is advantageous, as it reduces the chance of a random mutation invoking a deleterious effect [314-316,324], which for antibodies and Chapters 3 and 4 is highly advantageous.

This is commonly referred to the “designability” or mutational tolerance of a protein fold and is thought to be a function of the interaction space in a particular topology, thereby dictating the possible sequence space available [314,316]. In turn, designability also imparts limitations on the number of physically possible protein folds. Interestingly, Zhang et al. [317] attempted to calculate this number in 2006 using a number of computational methods that created an ensemble of all possible folds. They found that essentially all members of their calculated ensemble had determined structural analogues in the Protein Data Bank (PDB); suggesting that examples of almost every possible structural fold have been available since at least 2006. This is heavily supported by the lack of any new folds being reported since 2008, for a total of 1,393 known protein folds in 2016 (http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-scop accessed 4/2/2016).

The designability of a protein fold can be predicted to some extent by assessing the relationship between the contact matrix of a protein structure and the shape of its energy spectrum in sequence space [328]. England and Shakhnovich show that protein folds with a higher number of

intramolecular interactions may have greater potential for thermal stability and mutational tolerance. Further, they hypothesise that thermophilic organisms exhibit a genomic bias towards protein folds of higher intramolecular interactions, which was later supported by further studies [329,330]. As a result of the physiochemical limitations amongst protein folds, designability is an important aspect to consider during target selection, as it may greatly impact the upper limit in which consensus design can improve stability. Interestingly, when consensus design was applied to the metastable serine protease inhibitor family, we observed an absolutely remarkable improvement over naturally evolved variants. From these results, I speculate that evolution may often under utilize the full potential of a fold.

6.2.7 Approaches to the implementation of consensus design

With an adequate MSA, it is next necessary to consider how consensus design will be applied. In the simplest context, either point mutations are made to a target sequence or a consensus sequence is calculated *de novo*. Both approaches have yielded notable success in generating stability (Table 6.1), however the original design question will dictate which is most applicable. If stabilisation of a specific target protein is sought, or if the preservation of function is important, there is a higher chance that protein function will be retained if fewer mutations that are made. Therefore, it is recommended that mutations that are ranked in descending order of conservation, and implemented in a stepwise fashion [138,139,143].

However, one problem with this approach is that a large number of conserved residues tend to be destabilizing (~40%) [111,130,136,138-142], leading to moderately sized mutagenesis studies to identify which mutations are genuinely stabilizing. The search space can be reduced by computational methods, which have extensively been tested for consensus design, however, they frequently fail to deliver accurate results [93,303]. An alternative approach is the use of a directed evolution library that is spiked with consensus residues, however, this comes with its own costs and challenges as discussed in Chapter 1 [142,152].

If function is not particularly important, such as the case for downstream directed evolution studies or for non-functional proteins, a full *de novo* sequence calculation may avoid the need to identify true stabilizing residues, as was done for FN3con and conserpin. That being said, full sequence designs often retain function, as seen in conserpin (Chapter 5), albeit at reduced levels [134,140,146-149]. The reason for why function is reduced after design is presently unknown, however, it may be the result of a stability-function trade-off [144,155,196,331], the mixing of different covarying mutations, or the lack of conservation in the mechanisms that govern high levels of specificity and activity; thereby being omitted during consensus design.

6.2.8 Statistical enhancements to consensus design

It is intriguing that consensus design is successful despite its assumption of amino acid independence, ignoring the known importance of cooperativity and coupling of amino acids [68,332]. Furthermore, successes rival and often exceed those of rational design and directed evolution, which is impressive given the relative ease in which consensus design can be performed. Coupling manifests as simple pairwise interactions, through to dense and complex inter-atomic networks [333-335]. For consensus design to work, coupling must be encoded into the evolutionary history and represented by amino acid conservation to some extent, which might explain why ~40% of reported consensus mutations are destabilizing [111,130,136,138-142].

Attempts to improve consensus design have typically utilized additional statistical analysis that identifies coupling or covariation [145,336-339], and have generally been very successful in the engineering of stability [93,148,313,340,341]. Ranganathan and colleagues highlighted the necessity of statistical coupling analysis (SCA) to include both conserved and coupled mutations in the design of a WW domain, as consensus design alone was insufficient in creating a protein that folded correctly [145]. However, two previous studies had no difficulty in generating folded and stable WW domains [342,343], suggesting that failure of consensus design may have been a result of the MSA composition rather than a limitation of the WW domain itself. Another approach

employed by Sullivan and colleagues used the mutual information method to calculate the pairwise statistical interactions between positions in the MSA and chose to avoid making mutations to those positions, thereby improving the accuracy of identifying stabilizing mutations from ~50% to 90% [93]. However, this approach may not always be necessary; the pairwise covariation within and between ankyrin repeat motifs was found to be well represented by consensus design alone [344].

The role of covarying residues is even less understood than those of consensus mutations, although it appears that in some instances conserved residues encode most, if not all cooperativity. Therefore, consensus design and its enhancement by filtering correlated residues is dependent on how well the cooperativity is encoded into conserved residues, and whether other such correlations are statistically discernable from the alignment. Consensus design also appears to suffer when there are incompatible conserved residues and couplings as a result of divergent evolution, although this can be corrected by covariation methods [145,339,340]. However, covariation methods may not work in all scenarios; they typically require large MSAs to discern mutual amino acid dependencies [145,339], and are not applicable in situations where neutral drift studies are required, due to the rare event of coevolution. Interestingly, covaried residues in many cases actually have no physiospatial interactions with one another, recently sparking debate over what these methods are actually measuring [339]. Covarying substitutions are often found on different branches of the phylogenetic tree, and are perhaps independent events that may or may not be attributable to molecular coevolution [339]. In the case of consensus design, highly conserved residues tend to be found within the protein core, evolve slowly, and are therefore unlikely to be detected by covariation analysis even in very large alignments [339,345,346]. Regardless, covariation methods overall do seem to have utility and appear to generally identify favourable pairs of residues that can be used on their own and in conjunction with consensus design.

6.3 Biophysical effects of consensus design

6.3.1 Thermodynamic stability

Much of the discussion about consensus design focuses primarily on the general trend of improving thermostability [78,130-132]. Indeed, consensus design reports a wide range of improvements to melting temperature from the modest increase of the marginally stable antibody V_H domain (T_m of 36.4°C) by 6.1°C [133], the modest increase of the highly stable Azami green fluorescent protein (T_m of ~90°C) by 5.5°C [134], through to the large increase of the moderately stable Mouse Leucine Rich Repeat Transmembrane Neuronal 2 (LRRTM2) (T_m of ~50°C) by 32°C [135]. In this thesis, Chapter 2 describes the large increase in melting temperature of the highly stable Fibronectin Type III (FN3) domain, FNfn10 (T_m of 82°C), by at least 27°C [137]. With respect to ΔG, there is in general a strong correlation between improvements to T_m resulting in improvements to ΔG [220], with some expected outliers [150]. However, improvements to thermodynamic stability are not necessarily the only observed effects of consensus design.

6.3.2 Protein evolvability

A consequence of improved stability is often improved protein evolvability. Similar to designability [318,328], evolvability is different in that it explores the mutational tolerance or robustness of a specific protein as a function of its thermodynamic stability rather than the sequence space available to its fold or topology as a function of interaction space [13,14,156,347,348]. Proteins are often mutationally robust, with more than half of random single point mutants retaining native function [13,153,349]. However, extra thermodynamic stability is known to increase the robustness of the native structure to random mutations by increasing the fraction of variants that continue to possess the minimal stability required to fold [14,349,350]. The mechanism by which this occurs is not fully understood, although is thought to involve a combination of raw stability and ‘global suppressor’ residues that buffer the effect of deleterious mutations [155,219,350,351].

In the context of consensus design, raw stability is definitely observed (Table 6.1), however, without extensive mutagenesis studies it is unclear whether conserved residues infer global suppressor like properties. Given that global suppressor residues appear to be transferrable across protein homologues, such as the case in TEM β-lactamases [352], it is reasonable to suggest that conserved residues which happen to be global suppressors will induce similar effects when made in consensus design. Tawfik and colleagues have shown the utility of consensus design to enhance the evolvability of a computationally designed Kemp eliminase (KE59) [142]. In this study, optimisation of activity by directed evolution was sought, however the stability of KE59 was insufficient to tolerate mutations, rapidly producing unfolded proteins, thereby trapping the evolutionary trajectory in a local minimum. To boost KE59's evolvability, conserved residues were spiked into the directed evolution library, thereby improving protein stability and allowing for fresh downhill evolution of function. This scenario can be described with the computationally designed KE59 sitting very close to the bottom of the neutral zone of stability (Fig. 6.2A), where evolution of existing or new function (Fig. 6.2B) results in a deleterious effect. Like all optimisation processes, directed evolution of improved function can suffer from being trapped in local minima. When this happens, there significantly few mutations that can be made without shifting the stability outside of the neutral zone or to improve function [96,331]. However, consensus mutations or consensus-weighted libraries allow for a novel approach in the evolution of function, because they can be selected to ignore functional residues, whilst still increasing protein stability (Fig. 6.2C).

In Chapters 3 and 4, I explored the utility, mutational tolerance and evolvability of FN3con as a binding scaffold. Chapter 3 used a rational design approach, grafting the binding loops from the 10th FN3 repeat of human fibronectin (FNfn10) that had been previously evolved to bind lysozyme at 1 pM (DE0.4.1/FNfn10-α-lys) [201] onto FN3con (FN3con-α-lys.v2). These results showed that whilst FNfn10-α-lys was able to bind lysozyme in the sub-nanomolar range with a slow off rate, the protein expressed insolubly, had a relatively low melting temperature (43.2°C) and visibly precipitated at a concentration greater than 1 mg/ml and on thermal challenge. In contrast,

FN3con-α-lys.v2 was able to bind lysozyme with a similar SPR binding profile, whilst retaining soluble expression, a high melting temperature (87.2°C) and reversible folding. Although FN3con provides an exceptional starting point for protein engineering, the loss in stability on generation of function between FNfn10-α-lys (loss of 39.3°C) and FN3con-α-lys.v2 (loss of 17-25°C) was not the same. These results indicate that the FN3con scaffold is significantly more robust than FNfn10. At present, I am unsure about the source of the increased mutational robustness, however it may be due to FN3con's improved kinetic stability, which reduces the propensity for aggregation – a feature that is not selected for in naturally evolved proteins.

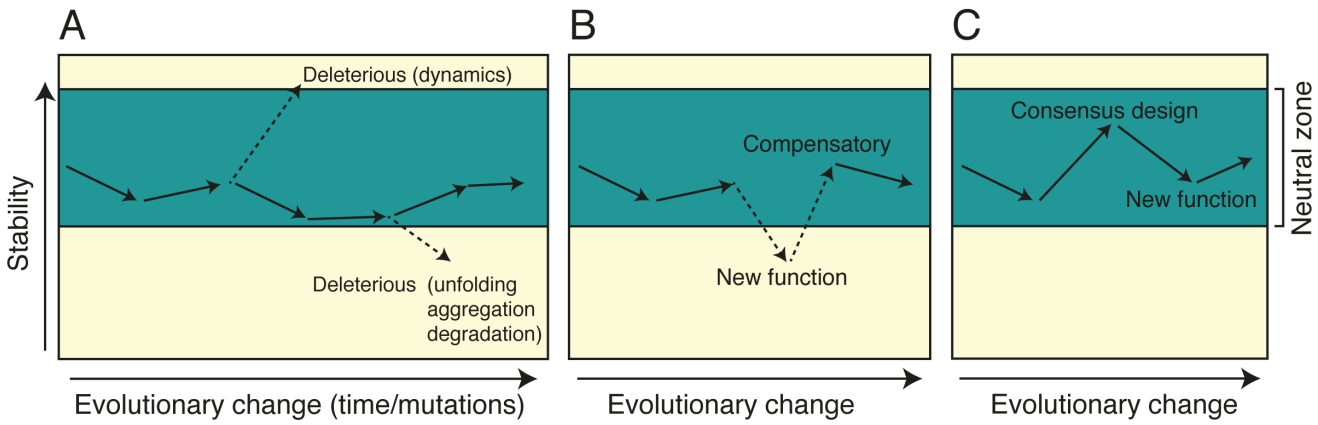


Figure 6.2. Evolutionary changes and subsequent changes in stability and fitness. **(A)** Proteins evolve and acquire mutations within a neutral zone of stability that maintains their stability (turquoise area). Mutations that do not alter stability outside of the neutral zone (cream area) are tolerated, however about a third of all mutations are destabilizing and may reduce stability to the extent of unfolding, misfolding and aggregation. Rarely, a mutation may increase stability above the neutral zone, reducing function by negatively affecting dynamics or regulatory networks. **(B)** The evolution of existing or new functionality tends to be destabilizing, requiring compensatory mutations to restore stability, or the use of chaperone proteins to extend the size of the neutral zone. **(C)** A powerful approach for the evolution of new functionality is to create a highly stable starting point for evolution via consensus design and allow for downhill divergence towards the desired function.

Chapter 4 follows on from Chapter 3 with a directed evolution approach, using yeast surface display to evolve an FN3con derivative that binds to lysozyme. Data collected during this study indicates that the naïve FN3con library contains low nanomolar clones against lysozyme, without the need to conduct affinity maturation studies. I therefore hypothesise that the improved biophysical properties of FN3con enables the display of a greater fraction of sequence space, than alternative FN3 scaffolds. Future work will validate this hypothesis for FN3con and expand the theory to other functional designs, such as conserpin. These results are promising and suggest that consensus design has great potential for the generation of highly evolvable proteins.

6.3.3 Protein folding and kinetic stability

Protein folding and the kinetic stability is an often overlooked property in protein design projects due to many proteins exhibiting irreversible folding on denaturation and the associated complexities of studying multistate folding pathways [87]. However, thermodynamic stability alone does not guarantee that the protein will fold or remain folded in the native state for extended periods of time under biological or arduous industrial conditions. *In vivo*, the biological function of many proteins requires a rugged energy landscape, which puts them at risk of misfolding and aggregation [23,24,87,274,353,354]. The delicate balance between function and misfolding is exemplified by members of the serine protease inhibitor or serpin superfamily [225,226,281,355]. Inhibitory members fold to a metastable native state that undergoes a major conformational change in order to inhibit target proteases [221]. As such, serpins have evolved a relatively complicated folding mechanism required for their function, with sequence and structural diversity within the superfamily reflecting specialized functional and regulatory requirements.

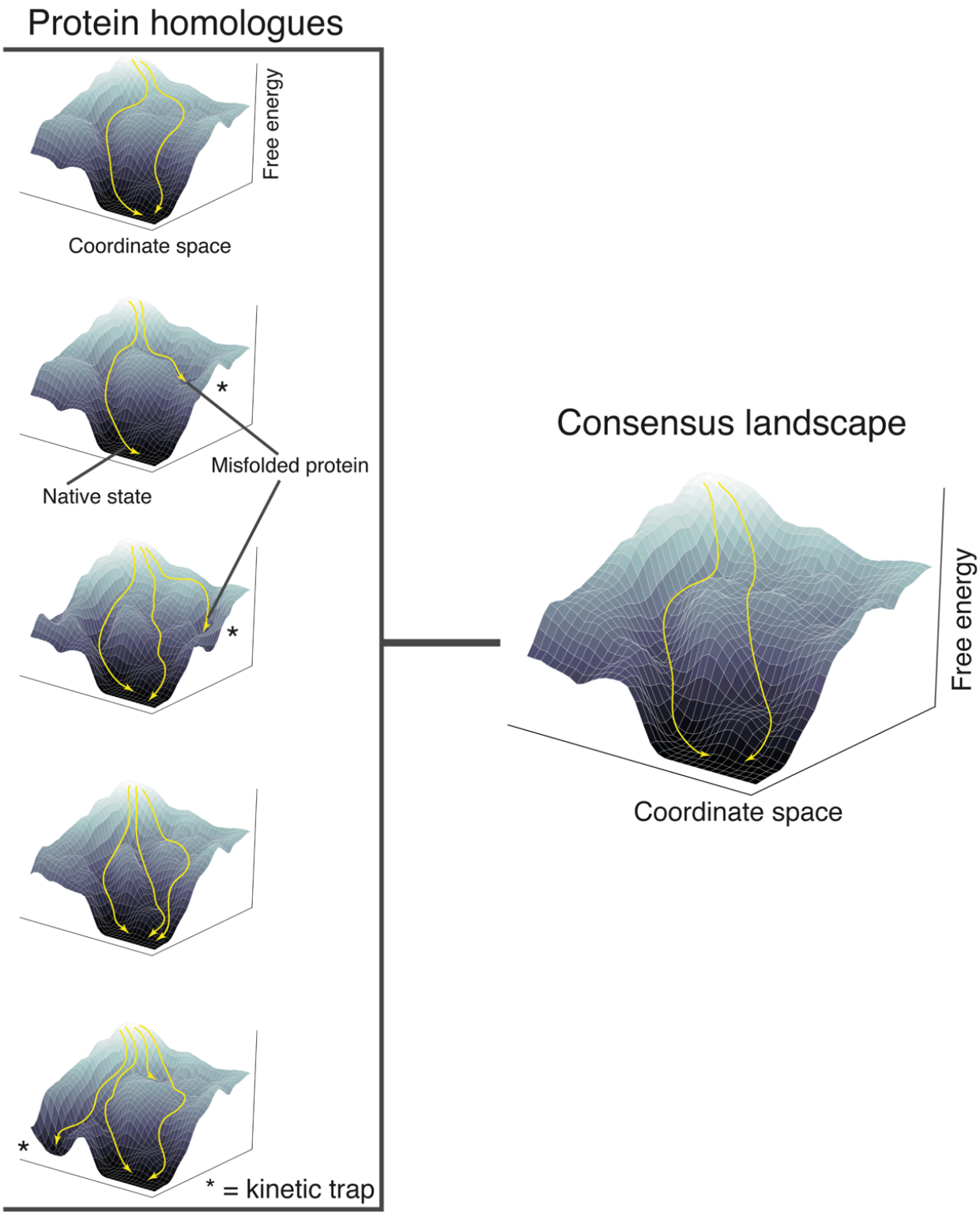


Figure 6.3. Smoothing of five hypothetical energy landscapes by consensus design. Five protein homologues exhibit differences in their energy landscapes, with three containing kinetic traps that present a propensity for misfolding. As the kinetic traps are not conserved across all five homologues, consensus design is capable of smoothing out the energy landscape to eliminate non-conserved features.

In chapter 5, I used consensus design to create a synthetic serpin, based on the hypothesis that a serpin molecule reflecting optimal sequence conservation may offer insight into the serpin folding-function trade-off. Remarkably, the consensus serpin uniquely exhibits reversible two-state folding, is functional, thermostable, and resistant to polymerization. Structural and biophysical analysis suggests that consensus design remodelled its folding landscape, thereby reducing the lifetime of aggregation-prone intermediates. I also observed similar, though less dramatic effects with FN3con (Chapter 2) [137], where consensus design led to a large increase in the folding rate and decrease in the unfolding rate (high kinetic stability), suggesting a more smooth and funnel like energy landscape.

Although consensus design nearly always modifies energy landscapes, improvements to folding and kinetic stability are unlikely to be universal (for example see [148,308,356]). Success likely depends on the specific functional requirements and evolutionary history of the MSA, as these will dictate the consensus energy landscape. In the case of serpins, off-pathway folding and polymerization is probably the result of independent evolutionary fine-tuning of the energy landscape as a means for conformational control of function. As these independent pathways are not highly conserved in a sequence alignment, they are removed by consensus design, thus remodelling the energy landscape to be smoother and funnel like, whilst still retaining conserved conformational properties necessary for function (Fig. 6.3). It is tempting to speculate that consensus design may prove to be a fruitful avenue for investigating and engineering the risky energy landscapes of functional proteins [354,357].

6.3.4 Function

The function of consensus designed proteins can be preserved, and is influenced by the implementation of design (Table 6.1). In general, consensus mutations, especially those that are distal from the catalytic site, give the highest odds of completely preserving function [129,141], whilst full sequence (*de novo*) designs are likely to reduce catalytic rates and specificity, as was

the case with conserpin (Chapter 5) and can be seen in Table 6.1. Although full sequence designs often reduce catalytic rates and specificity, they tend to retain function at elevated temperatures and wider ranges of pH [140,148,358]. Full sequence designs likely yield these results in a similar manner to the proposed energy landscape smoothing; with the finer features of catalytic activity being less conserved across all homologues, and are therefore removed during design. Fundamentally a goal needs to be set during the initial stages of the project, which must consider whether any additional directed evolution and optimization of function is to be undertaken. In the case of conserpin, I was more interested in the effect of consensus design on the protein fold itself; therefore the presence of any activity was an added bonus. We are still uncertain about the true inhibitory specificity of conserpin, as the reactive centre loop (RCL) contains a non-natural sequence. Ongoing work is presently exploring the rational engineering of the RCL, with results indicating that the restoration of near native activities and specificities is possible.

6.3.5 Immunogenicity

The application of consensus design to the reduction of immunogenicity, an important factor in the design of protein therapeutics [359-361], remains largely unexplored. Interestingly, what appears to be the first consensus designed protein, alfacon-1 (Infergen), is less immunogenic and significantly more active than recombinant interferon-α (IFN-α) [147,362]. Created using a sequence alignment of 25 IFN-α subtypes, alfacon-1 is the only known consensus designed protein that has been marketed as a therapeutic drug. Although alfacon-1 is the only reported experimental study of immunogenicity for a consensus designed protein, computational predictions using FN3 domains suggest that Tencon is also of low immunogenic potential [150]. However, it is of course possible that the absence of reports may reflect a general failure in immunogenicity reduction by consensus design. Regardless, the possibility that consensus design can reduce immunogenicity warrants further investigation.

6.4 Concluding remarks

Consensus design is a proven and highly effective sequence based method that is typically overlooked in protein engineering in favor of directed evolution and rational design methodologies. With a present inability to model entropy and non-native states, consensus design provides an additional tool for the protein engineer to not only stabilize the native state, but also modify the folding landscape. This thesis has highlighted a large number of successes in improving thermodynamic and kinetic stability through the combination of positive and negative design, with FN3con and conserpin exhibiting some of the most impressive stability improvements reported to date. The consensus approach can be implemented on its own, as single point mutations, or full sequence designs. It can also be used in combination with computational and or experimental methods. Consensus design can be extended further with statistical calculations to identify correlated or coupled pairs, often surpassing the improvements of consensus design alone. This thesis described the development and characterization of two full-sequence consensus designed proteins, FN3con and conserpin (Chapters 2 and 5).

Chapter 2 explored the effect of a large curated sequence alignment on the consensus design of a fibronectin type III (FN3) domain, which resulted in a molecule of immense thermodynamic and kinetic stability (FN3con). This study also explored the effect of protein folding on a consensus design protein, leading to the conclusion that the protein energy landscape was substantially smoother after design. As the FN3 domain can be applied as a binding scaffold, I sought to explore the utility of FN3con. This was done initially by rational design (Chapter 3), revealing a circumvention of the stability-function trade-off in the FN3 domain, with FN3con shown to provide a substantial mutational tolerance advantage over FNfn10 for the binding of lysozyme. Chapter 4 followed this line of research, asking the question as to whether FN3con provides evolvability advantages as a binding scaffold. A randomized loop library was developed and directed evolution by yeast surface display assessed binding affinity for lysozyme. Data collected during this study

indicates that the naïve FN3con library contains low nanomolar clones against lysozyme, without the need to conduct affinity maturation studies. These results therefore suggest that FN3con may be able to tolerate and display a larger sequence space than other alternative FN3 domains, warranting future investigation.

Chapter 5 explored the effect of consensus design on the serine protease inhibitor (serpin) family where stability and folding are directly linked to activity. I hypothesized that a serpin molecule reflecting optimal sequence conservation may offer insight into the folding-function trade-off that is notorious to serpins. By using consensus design, we designed *conserpin*, a synthetic serpin that exhibits reversible two-state folding, is functional, thermostable, and resistant to polymerization. Characterization of its structure, folding and dynamics suggest a remodeled folding landscape that reduces the lifetime of an aggregation-prone intermediate ensemble, subsequently smoothing the energy landscape.

Taken together, the integration of results in this thesis with the wider literature provides an extension to the general explanation for how consensus design improves stability. I therefore propose a general capacity for consensus design to smooth out energy landscape features that are not conserved across the entire input MSA; thereby resulting in improved thermodynamic and kinetic stability, with decreased activity and specificity. The discussed results and proposed hypothesis highlight the utility of consensus design for the generation of proteins with exceptional biophysical properties, which may serve as highly valuable starting positions for subsequent engineering studies or as novel model systems.

Table 6.1. An exhaustive catalogue of consensus design

Parent	*Function*	*Engineered protein*	*Type of design*[a]	*Experimental summary*	*Reference*
Subtilisin BPN'	Catalysis, Peptidase	N/A	Mutations (4)	One of the earliest examples of consensus design. Introduced 1 consensus mutation that resulted in a ΔΔG of -0.48 kcal mol^{-1} and ΔT_m of 1.8°C in comparison to wild type.	[146]
Antibody V_L domain	Binding	N/A	Mutations (~2,707)	Theoretical explanation of consensus design. Predicted ten individual stabilizing mutations, six were found to individually be stabilizing, three were found to be neutral, two were destabilising. ΔΔG of single stabilizing mutations ranged from -1 to -5.7 kcal mol^{-1}.	[111]
Antibody V_H domain	Binding	IcaH-501	Mutations (~6,319)	5 mutations were made to a V_H domain, which raised the T_m from 36.4°C to 42.5°C – ΔT_m of 6.1°C. Survival time at temperature is increased 20 fold. Improved protein expression and reversible folding observed.	[133]
Antibody Fv	Binding	Ica-Fv20	Mutations (N/A)	A consensus V_L domain (IcaL-14) was coupled with a native V_H domain (IcaH-01). This produced a fully reversible folding construct that expressed well.	[363]
β-sheet proteins	N/A	N/A	N/A (75)	The authors explored how natural β-sheet proteins avoid edge-to-edge aggregation by using a multitude of methods including homologous sequence alignment and assessment of conserved residues. No design was made.	[364]
FN3 domain	Structural, Binding	Fibcon	Full sequence (15)	Authors used 15 FN3 sequences from human fibronectin for full sequence design, which resulted in a ΔG of -11.4 kcal mol^{-1} and T_m of 89.6°C. In comparison to FNfn10, this is a ΔΔG of -2 kcal mol^{-1} and ΔT_m of 7.1°C.	[150]
		FibconB	Full sequence (10)	Authors used the top 10 most stable FN3 sequences from the Fibcon dataset, resulting in a ΔG of -6.7 kcal mol^{-1} and T_m of 85.3°C. In comparison to FNfn10, this is a ΔΔG of 2.68 kcal mol^{-1} and ΔT_m of 2.8°C.	[150]
		Tencon	Full sequence (15)	Authors collected 15 FN3 sequences from human tenascin	[150]

				[1], T_m is in excess of 100°C. In comparison to FNfn10, this is a ΔΔG of -6.1 kcal mol^{-1} and ΔT_m of >27°C. FN3con is the most stable, aggregation resistant, fastest folding and slowest unfolding FN3 domain reported.	
NOD receptor leucine rich repeat (LRR)	Binding	CLRR2	Full sequence (311)	Authors created CLRR2, which is a stable, monomeric, cysteine free scaffold protein that retains physiological binding. Does not follow two state folding and does not reversibly fold. Two unfolding transitions observed at 30°C and 68°C – no comparison to wild type.	[356]
Serine protease inhibitor (serpin)	Protease inhibition, conformational change	Conserpin	Full sequence (212)	Full sequence design of a serine protease inhibitor, which typically aggregate at 60°C. Conserpin is thermostable, aggregation resistant, retains function (less than wild type comparison) and reversibly folds after thermal and chemical denaturation. This is the first description of a reversibly folding serpin, which has a ΔG of -23.2 kcal mol^{-1} and a T_m in excess of 100°C. In comparison to PAI-1, this is a ΔΔG of -11.2 kcal mol^{-1} and ΔT_m of ~48°C.	Chapter 5
SH3 domain	Protein binding, signalling	N/A	Mutations (350)	Authors explored amino acid conservation at two sites, E24 and S41. A significant correlation was found between amino acid conservation and thermodynamic stability. Mutating E24 to the most conserved residue, E24D, resulted in a ΔG of -5.67 kcal mol^{-1} and a T_m of 82.2°C. In comparison to WT fyn SH3, this is a ΔΔG of -0.68 kcal mol^{-1} and ΔT_m of 2.1°C.	[220]
		N/A	Mutations (266)	The authors identified eight positions in the Abp1p SH3 domain that could be replaced by the most conserved residue at that position. Three out of eight mutations made were found to be stabilizing, with an additive triple mutant resulting in a ΔG of -6.43 kcal mol^{-1} and a T_m of 92.2°C. In comparison to WT Abp1p, this is a ΔΔG of -3.35 kcal mol^{-1} and ΔT_m of 31.9°C.	[307]
p53	Binding, DNA repair, cell cycle regulation	N/A	Mutations (23)	20 non-conserved residues were individually mutated to their respective consensus residue. Theoretical sum of all 20 produced a minor increase in stability. Four of the most stabilizing mutations were combined, which resulted in a ΔG of -11.29 kcal mol^{-1} and a T_m of 47.2°C. In comparison to WT p53, this is a ΔΔG of -2.65 kcal mol^{-1} and ΔT_m of 5.6°C.	[138]

GroEL	Chaperone, conformational change	M1 and M2	Mutations (130)	34 single conserved mutations were made and assessed. Six stabilizing mutations were combined to create two constructs M1 (ΔG of -11.7 kcal mol^{-1} and a T_m of 85.7°C) and M2 (ΔG of -10.6 kcal mol^{-1} and a T_m of 81.3°C). In comparison to wild type GroEL, this is a ΔΔG of -6.99 kcal mol^{-1}, ΔT_m of 18.6°C and a ΔΔG of -6.15 kcal mol^{-1}, ΔT_m of 14.2°C, respectively.	[139]
Green Fluorescent protein (GFP)	Fluorescence	Consensus green protein (CGP)	Full sequence (31)	Full sequence design with 31 GFP homologues. Comparisons with Azami green (mAG) FP was used to resolve ambiguous positions. The resulting design was extremely well expressed, monomeric, brighter and fluorescent with red shifted absorption and emission characteristics. CGP was slightly less stable than mAG, with a T_m of 79°C and a ΔT_m of -5.5°C.	[134]
Fungal phytase	Catalysis	Consensus phytase-1 (CFP-1)	Full sequence (13)	Full sequence consensus design was performed on 13 fungal phytase sequences. The resulting protein, CFP-1, has a T_m of 78°C, with a ΔT_m of 15-22°C in comparison to all parental wild types. Normal function was retained, at somewhat lower levels, but CFP-1 was functional at a wider range of pH and temperature.	[140]
		CFP-10, CFP-10-thermo[5]	Full sequence, mutations (19)	Full sequence consensus design was performed on 19 fungal phytase sequences. The resulting protein, CFP-10, has a T_m of 85.4°C, which is a ΔT_m of 7.4°C in comparison to CFP-1 and ΔT_m of 22.4°C in comparison to the most stable parental phytase. CFP-10 has 32 mutations in comparison to CFP-1, of which 10 were found to be stabilizing, 10 were destabilizing and 8 we neutral. Four residues were not tested. By back mutation the 8 destabilizing residues in CFP-10 and adding an additional stabilizing residues, CFP-10-thermo[5] was produced which has a T_m of 90.4°C, and ΔT_m of 27.4°C in comparison to the most stable parent.	[136]
Type I interferon (IFN)	Immune regulation	Alfacon-1	Full sequence (25)	Created by Amgen in 1981, consensus interferon (Alfacon-1) was shown to have activity that is 5-fold to 20-fold higher than that of IFN-α. Aflacon-1 is less immunogenic than recombinant IFN, and as far as we know, the only consensus protein to become a therapeutic drug, where it	[147]

				was commonly used to treat patients with chronic hepatitis C [365].	
Triosephosphate isomerase (TIM)	Catalysis (role in glycolysis)	cTIM, ccTIM	Full sequence (639, 781)	Authors compared the effect a sequence alignment from an unmodified Pfam (v18) alignment versus a length filtered Pfam (v22) alignment. The raw consensus TIM (cTIM) is weakly active, poorly folded, and monomeric, in contrast to nearly all known natural TIMs, which are dimers. The curated consensus TIM (ccTIM) is dimeric, well folded, and fully active. cTIM does not refold after heating to 95°C, whilst ccTIM does. Thermal melts were performed, but a T_m was not calculated. Authors found the differences between cTIM and ccTIM to occur exclusively in non-conserved regions. Reduction of sequences from the ccTIM alignment produced roughly the same protein. Authors suspect that the curated sequences produced a better alignment, especially on weakly conserved stretches.	[148]
		AlgoTIM	Mutations, statistical correlation (781)	Authors used the same, curated alignment as ccTIM and found that by avoiding residues with high statistical correlation with one another during consensus design, it was possible to improve the accuracy of stability enhancing mutations to 90%. By using this method, algoTIM was created by making 15 consensus mutations to wild type Sc. TIM, resulting in a T_m of 67.2°C, a ΔT_m of 8°C. Interestingly, making 14 stabilising mutations without considering correlation resulted in a loss of 2°C.	[93]
Zinc finger	DNA binding	N/A	Full sequence (131)	Stability was not assessed. Consensus protein was used as a scaffold for engineering specificity and affinity with DNA.	[366]
Ankyrin repeat	Binding	N/A	Full sequence (4,400)	A single consensus ankyrin domain was generated and used to explore the biophysical properties when repeated from 1-4 per polypeptide chain. 1 and 2 repeats were not folded in solution. 3 and 4 repeats were folded and resulted in a T_m of 69.4°C and 81.3°C respectively. No comparison to wild type was provided.	[344]
IκBα (Ankrytin repeat)	Binding	N/A	Mutations (N/A)	Authors made 3 mutations that improved T_m by ~11°C. It is not clear as to the number of sequences used in the alignment.	[143]
Tetratrico peptide repeat	Binding	CTPR1,	Full sequence	Authors created CTPR and explored the effects of linking 1	[308]

(TPR)		CTPR2, CTPR3	(1,837)	to 3 repeats together. This revealed a stepwise increase in thermostability, T_m of 47°C for the three-TPR domain of PP5 and 83°C for CTPR3. Reversible folding was not present.	
Chorismate mutase	Catalysis (conversion of chorismate to prephenate)	N/A	Combinatorial library, full sequence design (30)	Authors created several libraries of homologues by random mutagenesis, sequenced these libraries and utilized them as an alternate source of diversity for consensus design. In each library, the consensus protein was more stable than the original wild type or library members. Activity ranged significantly from 2-fold higher to 30-fold lower. This work emphasizes that diversity can be created *in vitro* and is not reliant on sequence databases.	[119]
Staphylococcal nuclease	Catalysis (cleavage of phosphodiester bonds)	N/A	Mutations (42)	Authors explored multiple combinations and permutations of hydrophobic core residues and found that there is a general trend between consensus side chains and favourable interaction energies. Further, they highlighted the importance of hydrophobic packing in protein stability, which is also seen in chapter 2.	[367]
β-lactamase (BLA)	Catalysis, antibiotic resistance	NA04.17	Combinatorial consensus library (38), mutational design	Authors aligned 38 homologues and identified 29 positions where BLA differed. A combinatorial library was generated for all 29 positions. The library was screened for thermostable members. Most stable variants improved T_m by 9.1°C. Variants showed irreversible folding.	[151]
		ALL-CON, GN-CON, GBP-CON	Full sequence (75, 26, 14)	Authors created three different full sequence consensus designs of BLAs using all 75 sequences and two extant clades. This study compared the results of consensus design with that of three ancestral reconstructions. The results conclusively found that ancestral reconstruction was better suited at invoking greater stability and retention of activity than consensus design.	[129]
TEM-1 BLA	Catalysis, antibiotic resistance	N/A	Consensus library	Authors subjected TEM-1 to 18 rounds of intense mutational drift and found that in response to high rates of mutation, sequences drifted towards the family consensus sequence. These consensus mutations were better suited to suppressing many different deleterious mutations.	[120]
Armadillo repeat protein	Binding	N/A	Full sequence (319)	Consensus design was used to construct a general peptide-binding scaffold. Results produced a well expressed and	[303]

				stable, but dimeric or molten globule protein. This was not desired, but was resolved by parallel computational optimization of the hydrophobic core.	
WW domain	Binding	WW-prototype	Mutations (NA)	Produced what appeared to be a properly folded WW domain with stability measured to be roughly the same as naturally occurring variants (T_m of 44.2°C).	[342]
		N/A	Mutations (>200)	Authors made a triple mutant to the wild type *h*Yap based on a sequence alignment of >200 WW domain homologues. The three mutations were rationally selected using structural comparison. The triple mutant increased the ΔG by 2.5 kcal mol^{-1} and the T_m by 28°C.	[343]
		N/A	Full sequence by Monte-Carlo (120)	Authors showed that consensus design failed to generate folded protein. This was restored to near wild type levels by using coupled conservation via statistical coupling analysis (SCA).	[145]
Albumin binding domain (ABD)	Binding	ABDcon	Full sequence (20)	Created ABDcon from 20 ABD sequences. ABDcon is highly expressed in the soluble fraction, highly stable (T_m of 81.5°C) and binds human, monkey and mouse serum albumins with affinity as high as 61 pM. Authors successfully explored tuning pharmacokinetic parameters. An N-terminal extension to ABDcon improved stability to 90.9°C.	[309]
Glucose 1-dehydrogenase (GDH)	Catalysis	N/A	Mutations (N/A)	Authors created a several GDH variants with combinatorial mutations that were stable and active in solutions with high concentrations of kosmotropic and chaotropic salts and water-miscible organic solvents.	[149]
Mouse Leucine Rich Repeat Transmembrane Neuronal 2 (LRRTM2)	Binding	cLRRTM2	Mutations (6,271)	The authors made 115 conserved mutations to mouse LRRTM2 resulting in cLRRTM2 with a T_m of 82°C, which is 32°C higher than wild type LRRTM2 (50°C). cLLRTM2 was also engineered and shown to be capable of forming synapse-like interactions in cell culture.	[135]
Affibody (3-helix bundle)	Binding	N/A	Mutations (942), Combinatorial library, synthetic population (38)	The authors used standard consensus design on 942 Affibody homologues to create several constructs between 6 and 10 mutations that resulted in constructs with stability that is equal to or less than that of the parental clone. Use of a combinatorial library resulted in a distribution of clones, but was successful in identifying one with good binding	[152]

				affinity and stability that was reduced by 2°C. The authors then utilized a neutral drift method and subsequent 38 sequences for consensus sequence design. The synthetic population consensus design yielded similar results to both other methods.	
Endoglucanase	Catalysis	G238P, QM	Mutations (18)	Authors identified a single mutation G328P that resulted in a T_m of 84.2°C and a ΔT_m of 3.5°C in comparison to wild type. The incorporation of this mutation into a previously identified triple mutant endoglucanase was additive and resulted in a T_m of 90.2°C and a ΔT_m of 9.5°C.	[368]
Kemp eliminase	Catalysis	KE59	Computational design stabilized by consensus mutations (7)	The authors used computational design methods for the *de novo* design of a Kemp eliminase, KE59, but found it to be too unstable for directed evolution. To resolve this problem, the design was spiked with consensus mutations to improve stability and evolvability, allowing for the directed evolution of catalytic activity.	[142]
Penicillin G acylase (PGA)	Catalysis, antibiotic resistance	N/A	Computationally selected mutations (8)	The authors aligned 8 PGA sequences and used a structure driven approach to selecting mutations that were not obviously disruptive to structure or function. 20 single point mutations were identified and found that only 10 were stabilizing (50%), 2 were neutral (10%) and 8 were destabilizing (40%). This paper highlights inaccuracies in computational stability predictions.	[141]
Split intein, (DnaE intein)	Catalysis, protein splicing	Cfa	Full sequence (73)	105 DnaE inteins were identified through a BLAST search of the JGI and NCBI databases. Alignment was sub-classified into 73 theoretically fast splicing inteins, based on the presence of 4 residues at specified positions. The 73 sequences were aligned and full consensus design was employed. Cfa is highly expressed, has an increased splicing rate as a function of temperature and is consistently faster than wild type Npu (2.5-fold at 30°C). Cfa maintains activity at 80 °C, albeit with reduced yield of splice products, while Npu is inactive at this temperature. Cfa can also splice in the presence of up to 4 M GuHCl, with little decrease in activity seen up to 3 M. For reference, most proteins unfold in < 3 M GuHCl.	[358]

[a]The number in brackets is the number of homologous sequences used in the design.

References

1. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. *Annu Rev Biophys* 2008; **37**:289–316.

2. Fersht AR, Matouschek A, Serrano L. The folding of an enzyme:: I. Theory of protein engineering analysis of stability and pathway of protein folding. *Journal of Molecular Biology* 1992; **224**:771–782.

3. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. *FASEB J* 1996; **10**:75–83.

4. Dill KA. Dominant forces in protein folding. *Biochemistry* 1990; **29**:7133–7155.

5. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design and Selection* 2009; **22**:553–560.

6. Fersht AR, Shi J-P, Knill-Jones J, *et al.* Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* 1985; **314**:235–238.

7. Brekke OH, Sandlie I. Therapeutic antibodies for human diseases at the dawn of the twenty-first century. *Nat Rev Drug Discov* 2003; **2**:52–62.

8. Walsh G. Biopharmaceutical benchmarks. *Nat Biotechnol* 2014; **32**:992–1000.

9. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. *Nature* 2012; **485**:185–194.

10. Ferdjani S, Ionita M, Roy B, *et al.* Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnology Letters* 2011; **33**:1215–1219.

11. Gao D, Narasimhan DL, Macdonald J, *et al.* Thermostable Variants of Cocaine Esterase for Long-Time Protection against Cocaine Toxicity. *Molecular Pharmacology* 2009; **75**:318–323.

12. Daniel RM, Cowan DA, Morgan HW, Curran MP. A correlation between protein thermostability and resistance to proteolysis. *Biochem J* 1982; **207**:641–644.

13. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* 2009; **19**:596–604.

14. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences* 2006; **103**:5869–5874.

15. Besenmatter W, Kast P, Hilvert D. Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins* 2007; **66**:500–506.

16. Fersht AR, Serrano L. Principles of protein stability derived from protein engineering experiments. *Current Opinion in Structural Biology* 1993; **3**:75–83.

17. Pace CN. Contribution of the hydrophobic effect to globular protein stability. *Journal of Molecular Biology* 1992; **226**:29–35.

18. Kauzmann W. Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 1959; **14**:1–63.

19. Dill KA, Chan HS. From Levinthal to pathways to funnels. *Nat Struct Biol* 1997; **4**:10–19.

20. Pace CN, Fu H, Fryar KL, *et al.* Contribution of hydrophobic interactions to protein stability. *Journal of Molecular Biology* 2011; **408**:514–528.

21. Tanford C. How protein chemists learned about the hydrophobic factor. *Protein Sci* 1997; **6**:1358–1366.

22. Horovitz A, Serrano L, Avron B, Bycroft M, Fersht AR. Strength and co-operativity of contributions of surface salt bridges to protein stability. *Journal of Molecular Biology* 1990; **216**:1031–1044.

23. Dobson CM. Protein folding and misfolding. *Nature* 2003; **426**:884–890.

24. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem Sci* 2000; **25**:331–339.

25. Shirley BA, Stanssens P, Hahn U, Pace CN. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 1992; **31**:725–732.

26. Fersht AR. Conformational equilibria in -and -chymotrypsin. The energetics and importance of the salt bridge. *Journal of Molecular Biology* 1972; **64**:497–509.

27. Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry* 1990; **29**:2403–2408.

28. Hendsch ZS, Tidor B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci* 1994; **3**:211–226.

29. Gong H, Freed KF. Electrostatic solvation energy for two oppositely charged ions in a solvated protein system: salt bridges can stabilize proteins. *Biophys J* 2010; **98**:470–477.

30. Roth CM, Neal BL, Lenhoff AM. Van der Waals interactions involving proteins. *Biophysj* 1996; **70**:977–987.

31. Clarke J, Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. *Biochemistry* 1993; **32**:4322–4329.

32. Matsumura M, Signor G, Matthews BW. Substantial increase of protein stability by multiple disulphide bonds. *Nature* 1989; **342**:291–293.

33. Matthews BW, Nicholson H, Becktel WJ. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proceedings of the National Academy of Sciences* 1987; **84**:6663–6667.

34. Mansfeld J, Vriend G, Dijkstra BW, *et al.* Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J Biol Chem* 1997; **272**:11152–11156.

35. Nicholson H, Becktel WJ, Matthews BW. Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles. *Nature* 1988; **336**:651–656.

36. Nicholson H, Anderson DE, Dao-pin S, Matthews BW. Analysis of the interaction between charged side chains and the alpha-helix dipole using designed thermostable mutants of phage T4

lysozyme. *Biochemistry* 1991; **30**:9816–9828.

37. Sali D, Bycroft M, Fersht AR. Stabilization of protein structure by interaction of alpha-helix dipole with a charged side chain. *Nature* 1988; **335**:740–743.

38. Serrano L, Fersht AR. Capping and alpha-helix stability. *Nature* 1989; **342**:296–299.

39. Marshall SA, Morgan CS, Mayo SL. Electrostatics significantly affect the stability of designed homeodomain variants. *Journal of Molecular Biology* 2002; **316**:189–199.

40. Blaber M, Zhang X, Matthews B. Structural basis of amino acid alpha helix propensity. *Science* 1993; **260**:1637–1640.

41. Serrano L, Neira J-L, Sancho J, Fersht AR. Effect of alanine versus glycine in α-helices on protein stability. *Nature* 1992; **356**:453–455.

42. Serrano L, Horovitz A, Avron B, Bycroft M, Fersht AR. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry* 1990; **29**:9343–9352.

43. Sun DP, Sauer U, Nicholson H, Matthews BW. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry* 1991; **30**:7142–7153.

44. Waldburger CD, Schildbach JF, Sauer RT. Are buried salt bridges important for protein stability and conformational specificity? *Nat Struct Biol* 1995; **2**:122–128.

45. Strop P, Mayo SL. Contribution of Surface Salt Bridges to Protein Stability †,‡. *Biochemistry* 2000; **39**:1251–1255.

46. Nick Pace C, Alston RW, Shaw KL. Charge-charge interactions influence the denatured state ensemble and contribute to protein stability. *Protein Sci* 2000; **9**:1395–1398.

47. Burley S, Petsko G. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* 1985; **229**:23–28.

48. Anderson DE, Hurley JH, Nicholson H, Baase WA, Matthews BW. Hydrophobic core repacking and aromatic-aromatic interaction in the thermostable mutant of T4 lysozyme ser 117 → phe. *Protein Sci* 1993; **2**:1285–1290.

49. Puchkaev AV, Koo LS, Ortiz de Montellano PR. Aromatic stacking as a determinant of the thermal stability of CYP119 from Sulfolobus solfataricus. *Archives of Biochemistry and Biophysics* 2003; **409**:52–58.

50. Giver L, Gershenson A, Freskgard PO, Arnold FH. Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences* 1998; **95**:12809–12813.

51. Zhao H, Arnold FH. Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Engineering Design and Selection* 1999; **12**:47–53.

52. Miyazaki K, Wintrode PL, Grayling RA, Rubingh DN, Arnold FH. Directed evolution study of temperature adaptation in a psychrophilic enzyme. *Journal of Molecular Biology* 2000; **297**:1015–1026.

53. Wintrode PL, Arnold FH. Temperature adaptation of enzymes: lessons from laboratory

evolution. *Adv Protein Chem* 2001; **55**:161–225.

54. Martin A, Sieber V, Schmid FX. In-vitro selection of highly stabilized protein variants with optimized surface. *Journal of Molecular Biology* 2001; **309**:717–726.

55. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A. How enzymes adapt: lessons from directed evolution. *Trends Biochem Sci* 2001; **26**:100–106.

56. Eijsink VGH, Veltman OR, Aukema W, Vriend G, Venema G. Structural determinants of the stability of thermolysin-like proteinases. *Nat Struct Biol* 1995; **2**:374–379.

57. Hoseki J, Yano T, Koyama Y, Kuramitsu S, Kagamiyama H. Directed Evolution of Thermostable Kanamycin-Resistance Gene: A Convenient Selection Marker for Thermus thermophilus. *J Biochem* 1999; **126**:951–956.

58. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat Struct Biol* 2000; **7**:380–383.

59. Perl D, Schmid FX. Electrostatic stabilization of a thermophilic cold shock protein. *Journal of Molecular Biology* 2001; **313**:343–357.

60. Martin A, Kather I, Schmid FX. Origins of the high stability of an in vitro-selected cold-shock protein. *Journal of Molecular Biology* 2002; **318**:1341–1349.

61. Vetriani C, Maeder DL, Tolliday N, *et al.* Protein thermostability above 100 C: A key role for ionic interactions. *Proceedings of the National Academy of Sciences* 1998; **95**:12300–12305.

62. Xiao L, Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology* 1999; **289**:1435–1444.

63. Thoma R, Hennig M, Sterner R, Kirschner K. Structure and function of mutationally generated monomers of dimeric phosphoribosylanthranilate isomerase from Thermotoga maritima. *Structure* 2000; **8**:265–276.

64. Clantin B, Tricot C, Lonhienne T, Stalon V, Villeret V. Probing the role of oligomerization in the high thermal stability of Pyrococcus furiosusornithine carbamoyltransferase by site-specific mutants. *European Journal of Biochemistry* 2001; **268**:3937–3942.

65. Vieille C, Zeikus GJ. Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews* 2001; **65**:1–43.

66. Walden H, Bell GS, Russell RJM, Siebers B, Hensel R, Taylor GL. Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase. *Journal of Molecular Biology* 2001; **306**:745–757.

67. Maeda N. The Unique Pentagonal Structure of an Archaeal Rubisco Is Essential for Its High Thermostability. *J Biol Chem* 2002; **277**:31656–31662.

68. Matthews BW. Structural and genetic analysis of protein stability. *Annu Rev Biochem* 1993; **62**:139–160.

69. Eijsink VGH, Bjørk A, Gåseidnes S, *et al.* Rational engineering of enzyme stability. *Journal of Biotechnology* 2004; **113**:105–120.

70. Reetz MT, Carballeira JD, Vogel A. Iterative saturation mutagenesis on the basis of B factors

as a strategy for increasing protein thermostability. *Angew Chem Int Ed Engl* 2006; **45**:7745–7751.

71. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 2002; **9**:646–652.

72. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. *Science* 2011; **334**:517–520.

73. Joo JC, Pack SP, Kim YH, Yoo YJ. Thermostabilization of Bacillus circulans xylanase: computational optimization of unstable residues based on thermal fluctuation analysis. *Journal of Biotechnology* 2011; **151**:56–65.

74. Pikkemaat MG, Linssen ABM, Berendsen HJC, Janssen DB. Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng* 2002; **15**:185–192.

75. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology* 2002; **320**:369–387.

76. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res* 2005; **33**:W382–8.

77. Raman S, Vernon R, Thompson J, *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009; **77 Suppl 9**:89–99.

78. Magliery TJ. Protein stability: computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology* 2015; **33**:161–168.

79. Gribenko AV, Patel MM, Liu J, McCallum SA, Wang C, Makhatadze GI. Rational stabilization of enzymes by computational redesign of surface charge–charge interactions. *Proc Natl Acad Sci USA* 2009; **106**:2601.

80. Lawrence MS, Phillips KJ, Liu DR. Supercharging proteins can impart unusual resilience. *J Am Chem Soc* 2007; **129**:10110–10112.

81. Miklos AE, Kluwe C, Der BS, *et al.* Structure-based design of supercharged, highly thermoresistant antibodies. *Chemistry & Biology* 2012; **19**:449–455.

82. Kim T, Joo JC, Yoo YJ. Hydrophobic interaction network analysis for thermostabilization of a mesophilic xylanase. *Journal of Biotechnology* 2012; **161**:49–59.

83. Borgo B, Havranek JJ. Automated selection of stabilizing mutations in designed and natural proteins. *Proc Natl Acad Sci USA* 2012; **109**:1494–1499.

84. Baker D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci* 2010; **19**:1817–1819.

85. Murphy GS, Mills JL, Miley MJ, Machius M, Szyperski T, Kuhlman B. Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure* 2012; **20**:1086–1096.

86. Khan S, Vihinen M. Performance of protein stability predictors. *Hum Mutat* 2010; **31**:675–684.

87. Sanchez-Ruiz JM. Protein kinetic stability. *Biophysical Chemistry* 2010; **148**:1–15.

88. Dobson CM. Getting out of shape. *Nature* 2002; **418**:729–730.

89. Whisstock JC, Bottomley SP. Molecular gymnastics: serpin structure, folding and misfolding. *Current Opinion in Structural Biology* 2006; **16**:761–768.

90. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science* 2012; **338**:1042–1046.

91. Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 1994; **370**:389–391.

92. Chen K, Arnold FH. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences* 1993; **90**:5618–5622.

93. Sullivan BJ, Nguyen T, Durani V, *et al.* Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *Journal of Molecular Biology* 2012; **420**:384–399.

94. Wijma HJ, Floor RJ, Janssen DB. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current Opinion in Structural Biology* 2013; **23**:588–594.

95. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel* 2014; **27**:49–58.

96. Packer MS, Liu DR. Methods for the directed evolution of proteins. *Nat Rev Genet* 2015; **16**:379–394.

97. McCullum EO, Williams BAR, Zhang J, Chaput JC. Random mutagenesis by error-prone PCR. *Methods Mol Biol* 2010; **634**:103–109.

98. Cadwell RC, Joyce GF. Randomization of genes by PCR mutagenesis. *PCR Methods Appl* 1992; **2**:28–33.

99. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth* 2009; **6**:343–345.

100. Swers JS. Shuffled antibody libraries created by in vivo homologous recombination and yeast surface display. *Nucleic Acids Res* 2004; **32**:36e–36.

101. Forrer P, Jung S, Plückthun A. Beyond binding: using phage display to select for structure, folding and enzymatic activity in proteins. *Current Opinion in Structural Biology* 1999; **9**:514–520.

102. Wilson DS, Keefe AD, Szostak JW. The use of mRNA display to select high-affinity protein-binding peptides. *Proceedings of the National Academy of Sciences* 2001; **98**:3750–3755.

103. Hanes J, Plückthun A. In vitro selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Sciences* 1997; **94**:4937–4942.

104. Daugherty PS, Chen G, Olsen MJ, Iverson BL, Georgiou G. Antibody affinity maturation using bacterial surface display. *Protein Eng* 1998; **11**:825–832.

105. Boder ET, Wittrup KD. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Methods Enzymol* 2000; **328**:430–444.

106. Ghadessy FJ, Ong JL, Holliger P. Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences* 2001; **98**:4552–4557.

107. Ellefson JW, Meyer AJ, Hughes RA, Cannon JR, Brodbelt JS, Ellington AD. Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nat Biotechnol* 2014; **32**:97–101.

108. Tamakoshi M, Nakano Y, Kakizawa S, Yamagishi A, Oshima T. Selection of stabilized 3-isopropylmalate dehydrogenase of Saccharomyces cerevisiae using the host-vector system of an extreme thermophile, Thermus thermophilus. *Extremophiles* 2001; **5**:17–22.

109. Sieber V, Plückthun A, Schmid FX. Selecting proteins with improved stability by a phage-based method. *Nat Biotechnol* 1998; **16**:955–960.

110. Agresti JJ, Antipov E, Abate AR, *et al.* Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc Natl Acad Sci USA* 2010; **107**:4004–4009.

111. Steipe B, Schiller B, Plückthun A, Steinbacher S. Sequence statistics reliably predict stabilizing mutations in a protein domain. *Journal of Molecular Biology* 1994; **240**:188–192.

112. Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 1995; **141**:1641–1650.

113. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2014; **42**:D7–17.

114. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res* 2008; **36**:D190–5.

115. Sigrist CJA, Cerutti L, Hulo N, *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinformatics* 2002; **3**:265–274.

116. Sigrist CJA, de Castro E, Cerutti L, *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res* 2013; **41**:D344–7.

117. Finn RD, Coggill P, Eberhardt RY, *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016; **44**:D279–85.

118. Wilson D, Pethica R, Zhou Y, *et al.* SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* 2009; **37**:D380–6.

119. Jäckel C, Bloom JD, Kast P, Arnold FH, Hilvert D. Consensus protein design without phylogenetic bias. *Journal of Molecular Biology* 2010; **399**:541–546.

120. Bershtein S, Goldin K, Tawfik DS. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *Journal of Molecular Biology* 2008; **379**:1029–1044.

121. Thornton JW. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* 2004; **5**:366–375.

122. Pauling L, Zuckerkandl E, Henriksen T, Lövstad R. Chemical Paleogenetics. Molecular “Restoration Studies” of Extinct Forms of Life. *Acta Chemica Scandinavica* 1963; **17 supl.**:9–16.

123. Thornton JW, Need E, Crews D. Resurrecting the ancestral steroid receptor: ancient origin of

estrogen signaling. *Science* 2003; **301**:1714–1717.

124. Whitfield JH, Zhang WH, Herde MK, *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci* 2015; **24**:1412–1422.

125. Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 2008; **451**:704–707.

126. Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J Am Chem Soc* 2013; **135**:2899–2902.

127. Gaucher EA, Thomson JM, Burgan MF, Benner SA. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 2003; **425**:285–288.

128. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* 2006; **2**:e69.

129. Risso VA, Gavira JA, Gaucher EA, Sanchez-Ruiz JM. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins* 2014; **82**:887–896.

130. Lehmann M, Wyss M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Current Opinion in Biotechnology* 2001; **12**:371–375.

131. Lutz S. Beyond directed evolution--semi-rational protein engineering and design. *Current Opinion in Biotechnology* 2010; **21**:734–743.

132. Magliery TJ, Lavinder JJ, Sullivan BJ. Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Current Opinion in Chemical Biology* 2011; **15**:443–451.

133. Wirtz P, Steipe B. Intrabody construction and expression III: engineering hyperstable V(H) domains. *Protein Sci* 1999; **8**:2245–2250.

134. Dai M, Fisher HE, Temirov J, *et al.* The creation of a novel fluorescent protein by guided consensus engineering. *Protein Eng Des Sel* 2007; **20**:69–79.

135. Paatero A, Rosti K, Shkumatov AV, *et al.* Crystal Structure of an Engineered LRRTM2 Synaptic Adhesion Molecule and a Model for Neurexin Binding. *Biochemistry* 2016.

136. Lehmann M, Loch C, Middendorf A, *et al.* The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* 2002; **15**:403–411.

137. Porebski BT, Nickson AA, Hoke DE, *et al.* Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. *Protein Eng Des Sel* 2015; **28**:67–78.

138. Nikolova PV, Henckel J, Lane DP, Fersht AR. Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc Natl Acad Sci USA* 1998; **95**:14675–14680.

139. Wang Q, Buckle AM, Foster NW, Johnson CM, Fersht AR. Design of highly stable functional GroEL minichaperones. *Protein Sci* 1999; **8**:2186–2193.

140. Lehmann M, Kostrewa D, Wyss M, *et al.* From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng*

2000; **13**:49–57.

141. Polizzi KM, Chaparro-Riggers JF, Vazquez-Figueroa E, Bommarius AS. Structure-guided consensus approach to create a more thermostable penicillin G acylase. *Biotechnol J* 2006; **1**:531–536.

142. Khersonsky O, Kiss G, Röthlisberger D, *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci USA* 2012; **109**:10358–10363.

143. Ferreiro DU, Cervantes CF, Truhlar SME, Cho SS, Wolynes PG, Komives EA. Stabilizing IkappaBalpha by "consensus" design. *Journal of Molecular Biology* 2007; **365**:1201–1216.

144. Schreiber G, Buckle AM, Fersht AR. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure/Folding and Design* 1994; **2**:945–951.

145. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. *Nature* 2005; **437**:512–518.

146. Pantoliano MW, Whitlow M, Wood JF, *et al.* Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry* 1989; **28**:7205–7213.

147. Blatt LM, Davis JM, Klein SB, Taylor MW. The biologic activity and molecular characterization of a novel synthetic interferon-alpha species, consensus interferon. *J Interferon Cytokine Res* 1996; **16**:489–499.

148. Sullivan BJ, Durani V, Magliery TJ. Triosephosphate isomerase by consensus design: dramatic differences in physical properties and activity of related variants. *Journal of Molecular Biology* 2011; **413**:195–208.

149. Vazquez-Figueroa E, Yeh V, Broering JM, Chaparro-Riggers JF, Bommarius AS. Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media. *Protein Engineering Design and Selection* 2008; **21**:673–680.

150. Jacobs SA, Diem MD, Luo J, *et al.* Design of novel FN3 domains with high stability by a consensus sequence approach. *Protein Eng Des Sel* 2012; **25**:107–117.

151. Amin N, Liu AD, Ramer S, *et al.* Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Engineering Design and Selection* 2004; **17**:787–793.

152. Case BA, Hackel BJ. Synthetic and natural consensus design for engineering charge within an affibody targeting epidermal growth factor receptor. *Biotechnol Bioeng* 2016.

153. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology* 2007; **369**:1318–1332.

154. Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences* 1995; **92**:452–456.

155. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new functions trade off. *PLoS Comput Biol* 2008; **4**:e1000002.

156. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. *Science* 2009; **324**:203–207.

157. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; **26**:680–682.

158. Larkin MA, Blackshields G, Brown NP, *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; **23**:2947–2948.

159. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 1982; **157**:105–132.

160. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* 2006; **34**:W116–8.

161. Fleming PJ, Richards FM. Protein packing: dependence on protein size, secondary structure and amino acid composition. *Journal of Molecular Biology* 2000; **299**:487–498.

162. Eswar N, Webb B, Marti-Renom MA, *et al.* Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* 2007; **Chapter 2**:Unit 2.9.

163. Myers JK, Pace CN, Scholtz JM. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* 1995; **4**:2138–2148.

164. Battye TGG, Kontogiannis L, Johnson O, Powell HR, Leslie AGW. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* 2011; **67**:271–281.

165. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr* 2007; **40**:658–674.

166. Ng SP, Billings KS, Ohashi T, *et al.* Designing an extracellular matrix protein with enhanced mechanical stability. *Proc Natl Acad Sci USA* 2007; **104**:9633–9637.

167. Adams PD, Afonine PV, Bunkóczi G, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 2010; **66**:213–221.

168. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 2004; **60**:2126–2132.

169. Konagurthu AS, Reboul CF, Schmidberger JW, *et al.* MUSTANG-MR Structural Sieving Server: Applications in Protein Structural Analysis and Crystallography. Fernandez-Fuentes N, ed. *PLoS ONE* 2010; **5**:e10048.

170. Hekkelman ML, Beek te TAH, Pettifer SR, Thorne D, Attwood TK, Vriend G. WIWS: a protein structure bioinformatics Web service collection. *Nucleic Acids Res* 2010; **38**:W719–23.

171. Winn MD, Ballard CC, Cowtan KD, *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 2011; **67**:235–242.

172. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* 2008; **4**:435–447.

173. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* 2004; **25**:1656–1676.

174. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J. Interaction Models for Water in Relation to Protein Hydration. In: *Intermolecular Forces*.Vol 14. The Jerusalem Symposia on Quantum Chemistry and Biochemistry. Dordrecht: Springer Netherlands; 1981:331–342.

175. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984; **81**:3684.

176. Tironi IG, Sperb R, Smith PE, van Gunsteren WF. A generalized reaction field method for molecular dynamics simulations. *J Chem Phys* 1995; **102**:5451.

177. Heinz TN, van Gunsteren WF, Hünenberger PH. Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. *J Chem Phys* 2001; **115**:1125.

178. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 1997; **18**:1463–1472.

179. Miyamoto S, Kollman PA. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* 1992; **13**:952–962.

180. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* 2011; **27**:1575–1577.

181. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 2007; **9**:90–95.

182. DeLano WL. The PyMOL Molecular Graphics System. *(2002)* 2002.

183. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996; **14**:33–8– 27–8.

184. Zolot RS, Basu S, Million RP. Antibody–drug conjugates. *Nat Rev Drug Discov* 2013; **12**:259–260.

185. Demarest SJ, Glaser SM. Antibody therapeutics, antibody engineering, and the merits of protein stability. *Curr Opin Drug Discov Devel* 2008; **11**:675–687.

186. Holliger P, Hudson PJ. Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* 2005; **23**:1126–1136.

187. Beck A, Wurch T, Bailly C, Corvaia N. Strategies and challenges for the next generation of therapeutic antibodies. *Nat Rev Immunol* 2010; **10**:345–352.

188. Birch JR, Racher AJ. Antibody production. *Adv Drug Deliv Rev* 2006; **58**:671–685.

189. Rouet R, Lowe D, Christ D. Stability engineering of the human antibody repertoire. *FEBS Letters* 2014; **588**:269–277.

190. Dudgeon K, Rouet R, Kokmeijer I, *et al.* General strategy for the generation of human antibody variable domains with increased aggregation resistance. *Proc Natl Acad Sci USA* 2012; **109**:10879–10884.

191. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci USA* 2009; **106**:11937–11942.

192. Binz HK, Amstutz P, Plückthun A. Engineering novel binding proteins from

nonimmunoglobulin domains. *Nat Biotechnol* 2005; **23**:1257–1268.

193. Vazquez-Lombardi R, Phan TG, Zimmermann C, Lowe D, Jermutus L, Christ D. Challenges and opportunities for non-antibody scaffold drugs. *Drug Discov Today* 2015; **20**:1271–1283.

194. Stern LA, Case BA, Hackel BJ. Alternative Non-Antibody Protein Scaffolds for Molecular Imaging of Cancer. *Curr Opin Chem Eng* 2013; **2**:425–432.

195. Nagatani RA, Gonzalez A, Shoichet BK, Brinen LS, Babbitt PC. Stability for Function Trade-Offs in the Enolase Superfamily “Catalytic Module” †,‡. *Biochemistry* 2007; **46**:6688–6695.

196. Beadle BM, Shoichet BK. Structural bases of stability-function tradeoffs in enzymes. *Journal of Molecular Biology* 2002; **321**:285–296.

197. Koide A, Bailey CW, Huang X, Koide S. The fibronectin type III domain as a scaffold for novel binding proteins. *Journal of Molecular Biology* 1998; **284**:1141–1151.

198. Parker MH, Chen Y, Danehy F, *et al.* Antibody mimics based on human fibronectin type three domain engineered for thermostability and high-affinity binding to vascular endothelial growth factor receptor two. *Protein Eng Des Sel* 2005; **18**:435–444.

199. Lipovsek D. Adnectins: engineered target-binding protein therapeutics. *Protein Eng Des Sel* 2011; **24**:3–9.

200. Koide A, Gilbreth RN, Esaki K, Tereshko V, Koide S. High-affinity single-domain binding proteins with a binary-code interface. *Proceedings of the National Academy of Sciences* 2007; **104**:6632–6637.

201. Hackel BJ, Kapila A, Dane Wittrup K. Picomolar Affinity Fibronectin Domains Engineered Utilizing Loop Length Diversity, Recursive Mutagenesis, and Loop Shuffling. *Journal of Molecular Biology* 2008; **381**:1238–1252.

202. Taverna DM, Goldstein RA. Why are proteins marginally stable? *Proteins* 2002; **46**:105–109.

203. Serrano L, Day AG, Fersht AR. Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *Journal of Molecular Biology* 1993; **233**:305–312.

204. Arnold FH. Enzyme engineering reaches the boiling point. *Proceedings of the National Academy of Sciences* 1998; **95**:2035–2036.

205. Sánchez IE, Tejero J, Gómez-Moreno C, Medina M, Serrano L. Point mutations in protein globular domains: contributions from function, stability and misfolding. *Journal of Molecular Biology* 2006; **363**:422–432.

206. Nicaise M, Valerio-Lepiniec M, Minard P, Desmadril M. Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. *Protein Sci* 2004; **13**:1882–1891.

207. Sheriff S. Some Methods for Examining the Interactions between Two Molecules. *ImmunoMethods* 1993; **3**:191–196.

208. Ramamurthy V, Krystek SR, Bush A, *et al.* Structures of adnectin/protein complexes reveal an expanded binding footprint. *Structure* 2012; **20**:259–269.

209. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. *Journal of*

Molecular Biology 1993; **234**:946–950.

210. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* 2007; **372**:774–797.

211. Bricogne G, Blanc E, Brandl M, Flensburg C, Keller P. *BUSTER v. 2.10.2*. Global Phasing Ltd; 2011.

212. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* 1997; **15**:553–557.

213. Cappellaro C, Baldermann C, Rachel R, Tanner W. Mating type-specific cell-cell recognition of Saccharomyces cerevisiae: cell wall attachment and active sites of a- and alpha-agglutinin. *EMBO J* 1994; **13**:4737–4744.

214. Chen TF, de Picciotto S, Hackel BJ, Wittrup KD. Engineering fibronectin-based binding proteins by yeast surface display. *Methods Enzymol* 2013; **523**:303–326.

215. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* 2006; **1**:755–768.

216. Gera N, Hussain M, Rao BM. Protein selection using yeast surface display. *Methods* 2013; **60**:15–26.

217. Hua SB, Qiu M, Chan E, Zhu L, Luo Y. Minimum length of sequence homology required for in vivo cloning by homologous recombination in yeast. *Plasmid* 1997; **38**:91–96.

218. Gibson DG, Benders GA, Axelrod KC, *et al.* One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc Natl Acad Sci USA* 2008; **105**:20404–20409.

219. Steipe B. Evolutionary approaches to protein engineering. *Curr Top Microbiol Immunol* 1999; **243**:55–86.

220. Maxwell KL, Davidson AR. Mutagenesis of a Buried Polar Interaction in an SH3 Domain: Sequence Conservation Provides the Best Prediction of Stability Effects †. *Biochemistry* 1998; **37**:16172–16182.

221. Huntington JA, Read RJ, Carrell RW. Structure of a serpin-protease complex shows inhibition by deformation. *Nature* 2000; **407**:923–926.

222. Elliott PR, Lomas DA, Carrell RW, Abrahams JP. Inhibitory conformation of the reactive loop of alpha 1-antitrypsin. *Nat Struct Biol* 1996; **3**:676–681.

223. Stratikos E, Gettins PG. Major proteinase movement upon stable serpin-proteinase complex formation. *Proceedings of the National Academy of Sciences* 1997; **94**:453–458.

224. Tew DJ, Bottomley SP. Probing the equilibrium denaturation of the serpin α1-antitrypsin with single tryptophan mutants; evidence for structure in the urea unfolded state. *Journal of Molecular Biology* 2001; **313**:1161–1169.

225. Gettins PGW. Serpin Structure, Mechanism, and Function. *Chem Rev* 2002; **102**:4751–4804.

226. Krishnan B, Gierasch LM. Dynamic local unfolding in the serpin α-1 antitrypsin provides a mechanism for loop insertion and polymerization. *Nature Structural & Molecular Biology* 2011;

18:222–226.

227. Tsutsui Y, Cruz Dela R, Wintrode PL. Folding mechanism of the metastable serpin α1-antitrypsin. *Proc Natl Acad Sci USA* 2012; **109**:4467–4472.

228. Cabrita LD, Dai W, Bottomley SP. Different Conformational Changes within the F-Helix Occur during Serpin Folding, Polymerization, and Proteinase Inhibition †. *Biochemistry* 2004; **43**:9834–9839.

229. James EL, Bottomley SP. The mechanism of alpha 1-antitrypsin polymerization probed by fluorescence spectroscopy. *Archives of Biochemistry and Biophysics* 1998; **356**:296–300.

230. Dupont DM, Madsen JB, Kristensen T, *et al.* Biochemical properties of plasminogen activator inhibitor-1. *Front Biosci (Landmark Ed)* 2009; **14**:1337–1361.

231. Mushunje A, Evans G, Brennan SO, Carrell RW, Zhou A. Latent antithrombin and its detection, formation and turnover in the circulation. *J Thromb Haemost* 2004; **2**:2170–2177.

232. Dafforn TR, Mahadeva R, Elliott PR, Sivasothy P, Lomas DA. A kinetic mechanism for the polymerization of alpha1-antitrypsin. *J Biol Chem* 1999; **274**:9548–9555.

233. Ekeowa UI, Freeke J, Miranda E, *et al.* Defining the mechanism of polymerization in the serpinopathies. *Proc Natl Acad Sci USA* 2010; **107**:17146–17151.

234. Yamasaki M, Li W, Johnson DJD, Huntington JA. Crystal structure of a stable dimer reveals the molecular basis of serpin polymerization. *Nature* 2008; **455**:1255–1258.

235. Yamasaki M, Sendall TJ, Pearce MC, Whisstock JC, Huntington JA. Molecular basis of α1-antitrypsin deficiency revealed by the structure of a domain-swapped trimer. *EMBO Rep* 2011; **12**:1011–1017.

236. Wang Z, Mottonen J, Goldsmith EJ. Kinetically controlled folding of the serpin plasminogen activator inhibitor 1. *Biochemistry* 1996; **35**:16443–16448.

237. Pearce MC, Rubin H, Bottomley SP. Conformational change and intermediates in the unfolding of alpha 1-antichymotrypsin. *J Biol Chem* 2000; **275**:28513–28518.

238. Im H. Interactions Causing the Kinetic Trap in Serpin Protein Folding. *J Biol Chem* 2002; **277**:46347–46354.

239. Pearce MC, Cabrita LD, Rubin H, Gore MG, Bottomley SP. Identification of residual structure within denatured antichymotrypsin: implications for serpin folding and misfolding. *Biochem Biophys Res Commun* 2004; **324**:729–735.

240. Kim D, Yu MH. Folding pathway of human alpha 1-antitrypsin: characterization of an intermediate that is active but prone to aggregation. *Biochem Biophys Res Commun* 1996; **226**:378–384.

241. Irving JA, Pike RN, Lesk AM, Whisstock JC. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Research* 2000; **10**:1845–1864.

242. Zhou A, Carrell RW, Huntington JA. The Serpin Inhibitory Mechanism Is Critically Dependent on the Length of the Reactive Center Loop. *J Biol Chem* 2001; **276**:27541–27547.

243. Kwon KS, Kim J, Shin HS, Yu MH. Single amino acid substitutions of alpha 1-antitrypsin that confer enhancement in thermal stability. *J Biol Chem* 1994; **269**:9627–9631.

244. Sancho E, Declerck PJ, Price NC, Kelly SM. Conformational studies on plasminogen activator inhibitor (PAI-1) in active, latent, substrate, and cleaved forms. *Biochemistry* 1995; **34**:1064–1069.

245. Kwon KS, Lee S, Yu MH. Refolding of alpha 1-antitrypsin expressed as inclusion bodies in Escherichia coli: characterization of aggregation. *Biochim Biophys Acta* 1995; **1247**:179–184.

246. Shirai N, Tani F, Higasa T, Yasumoto K. Linear polymerization caused by the defective folding of a non-inhibitory serpin ovalbumin. *J Biochem* 1997; **121**:787–797.

247. Takehara S, Zhang J, Yang X, Takahashi N, Mikami B, Onda M. Refolding and polymerization pathways of neuroserpin. *Journal of Molecular Biology* 2010; **403**:751–762.

248. Onda M, Hirose M. Refolding mechanism of ovalbumin: investigation by using a starting urea-denatured disulfide isomer with mispaired CYS367-CYS382. *J Biol Chem* 2003; **278**:23600–23609.

249. Kim PS, Baldwin RL. Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu Rev Biochem* 1982; **51**:459–489.

250. James EL, Whisstock JC, Gore MG, Bottomley SP. Probing the unfolding pathway of alpha1-antitrypsin. *J Biol Chem* 1999; **274**:9482–9488.

251. Knaupp AS, Keleher S, Yang L, Dai W, Bottomley SP, Pearce MC. The Roles of Helix I and Strand 5A in the Folding, Function and Misfolding of α1-Antitrypsin. Crowther DC, ed. *PLoS ONE* 2013; **8**:e54766.

252. Tran ST, Shrake A. The folding of alpha-1-proteinase inhibitor: kinetic vs equilibrium control. *Archives of Biochemistry and Biophysics* 2001; **385**:322–331.

253. Geierhaas CD, Nickson AA, Lindorff-Larsen K, Clarke J, Vendruscolo M. BPPred: A Web-based computational tool for predicting biophysical parameters of proteins. *Protein Sci* 2006; **16**:125–134.

254. Fulton KF, Buckle AM, Cabrita LD, *et al.* The high resolution crystal structure of a native thermostable serpin reveals the complex mechanism underpinning the stressed to relaxed transition. *J Biol Chem* 2005; **280**:8435–8442.

255. Zhang Q, Buckle AM, Law RHP, *et al.* The N terminus of the serpin, tengpin, functions to trap the metastable native state. *EMBO Rep* 2007; **8**:658–663.

256. Karpusas M, Baase WA, Matsumura M, Matthews BW. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proc Natl Acad Sci USA* 1989; **86**:8237–8241.

257. Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. *Annu Rev Biochem* 1990; **59**:1007–1039.

258. DeDecker BS, O'Brien R, Fleming PJ, Geiger JH, Jackson SP, Sigler PB. The crystal structure of a hyperthermophilic archaeal TATA-box binding protein. *Journal of Molecular Biology* 1996; **264**:1072–1084.

259. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. *Annu Rev Biochem* 1997; **66**:549–579.

260. Kellis JT, Nyberg K, Sali D, Fersht AR. Contribution of hydrophobic interactions to protein stability. *Nature* 1988; **333**:784–786.

261. Parfrey H, Mahadeva R, Ravenhill NA, *et al.* Targeting a Surface Cavity of 1-Antitrypsin to Prevent Conformational Disease. *J Biol Chem* 2003; **278**:33060–33066.

262. Gooptu B, Miranda E, Nobeli I, *et al.* Crystallographic and Cellular Characterisation of Two Mechanisms Stabilising the Native Fold of α1-Antitrypsin: Implications for Disease and Drug Design. *Journal of Molecular Biology* 2009; **387**:857–868.

263. Sivasothy P, Dafforn TR, Gettins PGW, Lomas DA. Pathogenic 1-Antitrypsin Polymers Are Formed by Reactive Loop- -Sheet A Linkage. *J Biol Chem* 2000; **275**:33663–33668.

264. Kass I, Knaupp AS, Bottomley SP, Buckle AM. Conformational properties of the disease-causing Z variant of α1-antitrypsin revealed by theory and experiment. *Biophys J* 2012; **102**:2856–2865.

265. Knaupp AS, Levina V, Robertson AL, Pearce MC, Bottomley SP. Kinetic instability of the serpin Z alpha1-antitrypsin promotes aggregation. *Journal of Molecular Biology* 2010; **396**:375–383.

266. Hughes VA, Meklemburg R, Bottomley SP, Wintrode PL. The Z mutation alters the global structural dynamics of α1-antitrypsin. Rezaei H, ed. *PLoS ONE* 2014; **9**:e102617.

267. Lee KN, Park SD, Yu MH. Probing the native strain in alpha1-antitrypsin. *Nat Struct Biol* 1996; **3**:497–500.

268. Kim J, Lee KN, Yi GS, Yu MH. A thermostable mutation located at the hydrophobic core of alpha 1-antitrypsin suppresses the folding defect of the Z-type variant. *J Biol Chem* 1995; **270**:8597–8601.

269. Cabrita LD, Whisstock JC, Bottomley SP. Probing the Role of the F-Helix in Serpin Stability through a Single Tryptophan Substitution †. *Biochemistry* 2002; **41**:4575–4581.

270. Gettins PGW. The F-helix of serpins plays an essential, active role in the proteinase inhibition mechanism. *FEBS Letters* 2002; **523**:2–6.

271. Gooptu B, Hazes B, Chang WS, *et al.* Inactive conformation of the serpin alpha(1)-antichymotrypsin indicates two-stage insertion of the reactive loop: implications for inhibitory function and conformational disease. *Proc Natl Acad Sci USA* 2000; **97**:67–72.

272. Nyon MP, Segu L, Cabrita LD, *et al.* Structural dynamics associated with intermediate formation in an archetypal conformational disease. *Structure* 2012; **20**:504–512.

273. Pearce MC, Morton CJ, Feil SC, *et al.* Preventing serpin aggregation: the molecular mechanism of citrate action upon antitrypsin unfolding. *Protein Sci* 2008; **17**:2127–2133.

274. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci USA* 2007; **104**:19819–19824.

275. Lomas DA, Evans DL, Finch JT, Carrell RW. The mechanism of Z alpha 1-antitrypsin accumulation in the liver. *Nature* 1992; **357**:605–607.

276. Knaupp AS, Bottomley SP. Structural change in β-sheet A of Z α(1)-antitrypsin is responsible for accelerated polymerization and disease. *Journal of Molecular Biology* 2011; **413**:888–898.

277. Best RB, Hummer G, Eaton WA. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc Natl Acad Sci USA* 2013; **110**:17874–17879.

278. Lomas DA, Elliott PR, Chang WS, Wardell MR, Carrell RW. Preparation and characterization of latent alpha 1-antitrypsin. *J Biol Chem* 1995; **270**:5282–5288.

279. Carrell RW, Evans DL, Stein PE. Mobile reactive centre of serpins and the control of thrombosis. *Nature* 1991; **353**:576–578.

280. Gettins P, Patston PA, Schapira M. The role of conformational change in serpin structure and function. *Bioessays* 1993.

281. Law RHP, Zhang Q, McGowan S, *et al.* An overview of the serpin superfamily. *Genome Biol* 2006; **7**:216.

282. Li W, Johnson DJD, Esmon CT, Huntington JA. Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nature Structural & Molecular Biology* 2004; **11**:857–862.

283. Johnson DJD, Li W, Adams TE, Huntington JA. Antithrombin-S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation. *EMBO J* 2006; **25**:2029–2037.

284. Lindahl TL, Sigurdardottir O, Wiman B. Stability of plasminogen activator inhibitor 1 (PAI-1). *Thromb Haemost* 1989; **62**:748–751.

285. Fersht AR. Characterizing transition states in protein folding: an essential step in the puzzle. *Current Opinion in Structural Biology* 1995; **5**:79–84.

286. Cowieson NP, Aragao D, Clift M, *et al.* MX1: a bending-magnet crystallography beamline serving both chemical and macromolecular crystallography communities at the Australian Synchrotron. *Journal of Synchrotron Radiation* 2015; **22**:187–190.

287. Chang WS, Whisstock J, Hopkins PC, Lesk AM, Carrell RW, Wardell MR. Importance of the release of strand 1C to the polymerization mechanism of inhibitory serpins. *Protein Sci* 1997; **6**:89–98.

288. Levina V, Dai W, Knaupp AS, *et al.* Expression, purification and characterization of recombinant Z alpha(1)-antitrypsin--the most common cause of alpha(1)-antitrypsin deficiency. *Protein Expr Purif* 2009; **68**:226–232.

289. Dafforn TR, Pike RN, Bottomley SP. Physical characterization of serpin conformations. *Methods* 2004; **32**:150–158.

290. Le Bonniec BF, Guinto ER, Stone SR. Identification of Residues in Thrombin-Modulating Interactions with Antithrombin III and. alpha. 1-Antitrypsin. *Biochemistry* 1995; **34**:12241–12248.

291. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res* 2013; **41**:W314–21.

292. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 2006; **65**:712–725.

293. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983; **79**:926–11.

294. Lippert RA, Bowers KJ, Dror RO, *et al.* A common, avoidable source of error in molecular dynamics integrators. *J Chem Phys* 2007; **126**:046101.

295. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 1993; **98**:10089.

296. Case DA, Babin V, Berryman J, Betz RM, Cai Q. *Amber 14*. 2014.

297. Jenik M, Parra RG, Radusky LG, Turjanski A, Wolynes PG, Ferreiro DU. Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res* 2012; **40**:W348–51.

298. Morrison DA. Is Sequence Alignment an Art or a Science? *Systematic Botany* 2015; **40**:14–26.

299. Aniba MR, Poch O, Thompson JD. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res* 2010; **38**:7353–7363.

300. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 1998; **26**:320–322.

301. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* 2015; **43**:D257–60.

302. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011; **39**:W29–37.

303. Parmeggiani F, Pellarin R, Larsen AP, *et al.* Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *Journal of Molecular Biology* 2008; **376**:1282–1304.

304. Bauer F, Schweimer K, Klüver E. Structure determination of human and murine β-defensins reveals structural conservation in the absence of significant sequence similarity. *Protein …* 2001.

305. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999; **12**:85–94.

306. Bloom JD, Glassman MJ. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput Biol* 2009; **5**:e1000349.

307. Rath A, Davidson AR. The design of a hyperstable mutant of the Abp1p SH3 domain by sequence alignment analysis. *Protein Sci* 2000; **9**:2457–2469.

308. Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of Stable α-Helical Arrays from an Idealized TPR Motif. *Structure* 2003; **11**:497–508.

309. Jacobs SA, Gibbs AC, Conk M, *et al.* Fusion to a highly stable consensus albumin binding domain allows for tunable pharmacokinetics. *Protein Eng Des Sel* 2015.

310. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics* 2013; **Chapter 3**:Unit3.1.

311. Nuin PAS, Wang Z, Tillier ERM. The accuracy of several multiple sequence alignment

programs for proteins. *BMC Bioinformatics* 2006; **7**:471.

312. Kemena C, Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 2009; **25**:2455–2465.

313. Durani V, Magliery TJ. Protein engineering and stabilization from sequence statistics: variation and covariation analysis. *Methods Enzymol* 2013; **523**:237–256.

314. Govindarajan S, Goldstein RA. Why are some proteins structures so common? *Proceedings of the National Academy of Sciences* 1996; **93**:3341–3345.

315. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. *Science* 1996; **273**:666–669.

316. Goldstein RA. The structure of protein evolution and the evolution of protein structure. *Current Opinion in Structural Biology* 2008; **18**:170–177.

317. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences* 2006; **103**:2605–2610.

318. Wingreen NS, Li H, Tang C. Designability and thermal stability of protein structures. *Polymer* 2004; **45**:699–705.

319. Woolfson DN, Bartlett GJ, Burton AJ, *et al.* De novo protein design: how do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology* 2015; **33**:16–26.

320. Finkelstein AV, Gutin AM, Badretdinov AYa. Boltzmann-like statistics of protein architectures. Origins and consequences. *Subcell Biochem* 1995; **24**:1–26.

321. Finkelstein AV, Gutun AM, Badretdinov AYa. Why are the same protein folds used to perform different functions? *FEBS Letters* 1993; **325**:23–28.

322. Finkelstein AV, Badretdinov AYa, Gutin AM. Why do protein architectures have Boltzmann-like statistics? *Proteins* 1995; **23**:142–150.

323. Shakhnovich E. Protein design: a perspective from simple tractable models. *Fold Des* 1998; **3**:R45–58.

324. Shakhnovich EI, Gutin AM. Influence of point mutations on protein structure: probability of a neutral mutation. *J Theor Biol* 1991; **149**:537–546.

325. Broglia RA, Tiana G, Roman HE, Vigezzi E, Shakhnovich E. Stability of Designed Proteins against Mutations. *Phys Rev Lett* 1999; **82**:4727–4730.

326. Baker D. A surprising simplicity to protein folding. *Nature* 2000; **405**:39–42.

327. Halaby DM, Mornon JP. The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. *J Mol Evol* 1998; **46**:389–400.

328. England JL, Shakhnovich EI. Structural determinant of protein designability. *Phys Rev Lett* 2003; **90**:218101.

329. England JL, Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proceedings of the National Academy of Sciences* 2003;

100:8727–8731.

330. Berezovsky IN, Shakhnovich EI. Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences* 2005; **102**:12742–12747.

331. Tokuriki N, Jackson CJ, Afriat-Jurnou L, Wyganowski KT, Tang R, Tawfik DS. Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nat Commun* 2012; **3**:1257.

332. Horovitz A, Fersht AR. Co-operative interactions during protein folding. *Journal of Molecular Biology* 1992; **224**:733–740.

333. Chen J, Stites WE. Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. *Biochemistry* 2001; **40**:14012–14019.

334. LiCata VJ, Ackers GK. Long-range, small magnitude nonadditivity of mutational effects in proteins. *Biochemistry* 1995; **34**:3133–3139.

335. Luque I, Leavitt SA, Freire E. The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu Rev Biophys Biomol Struct* 2002; **31**:235–256.

336. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins* 1994; **18**:309–317.

337. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular Biology and Evolution* 2000; **17**:164–178.

338. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999; **286**:295–299.

339. Talavera D, Lovell SC, Whelan S. Covariation Is a Poor Measure of Molecular Coevolution. *Molecular Biology and Evolution* 2015; **32**:2456–2468.

340. Magliery TJ, Regan L. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *Journal of Molecular Biology* 2004; **343**:731–745.

341. Ozer HG, Ray WC. MAVL/StickWRLD: analyzing structural constraints using interpositional dependencies in biomolecular sequence alignments. *Nucleic Acids Res* 2006; **34**:W133–6.

342. Macias MJ, Gervais V, Civera C, Oschkinat H. Structural analysis of WW domains and design of a WW prototype. *Nat Struct Biol* 2000; **7**:375–379.

343. Jiang X, Kowalski J, Kelly JW. Increasing protein stability using a rational approach combining sequence homology and structural alignment: Stabilizing the WW domain. *Protein Sci* 2001; **10**:1454–1465.

344. Mosavi LK, Minor DL, Peng Z-Y. Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences* 2002; **99**:16029–16034.

345. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology* 2002; **324**:105–121.

346. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology* 1987;

195:957–961.

347. Wagner A. Robustness, evolvability, and neutrality. *FEBS Letters* 2005; **579**:1772–1778.

348. Bloom JD, Wilke CO, Arnold FH, Adami C. Stability and the Evolvability of Function in a Model Protein. *Biophys J* 2004; **86**:2758–2764.

349. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences* 2005; **102**:606–611.

350. Nikolova PV, Wong KB, DeDecker B, Henckel J, Fersht AR. Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. *EMBO J* 2000; **19**:370–378.

351. Poteete AR, Rennell D, Bouvier SE, Hardy LW. Alteration of T4 lysozyme structure by second-site reversion of deleterious mutations. *Protein Sci* 1997; **6**:2418–2425.

352. Brown NG, Pennington JM, Huang W, Ayvaz T, Palzkill T. Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM β-lactamases. *Journal of Molecular Biology* 2010; **404**:832–846.

353. Gianni S, Camilloni C, Giri R, *et al.* Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein. *Proc Natl Acad Sci USA* 2014; **111**:14141–14146.

354. Gershenson A, Gierasch LM, Pastore A, Radford SE. Energy landscapes of functional proteins are inherently risky. *Nat Chem Biol* 2014; **10**:884–891.

355. Lomas DA, Carrell RW. Serpinopathies and the conformational dementias. *Nat Rev Genet* 2002; **3**:759–768.

356. Parker R, Mercedes-Camacho A, Grove TZ. Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci* 2014; **23**:790–800.

357. Barrick D, Ferreiro DU, Komives EA. Folding landscapes of ankyrin repeat proteins: experiments meet theory. *Current Opinion in Structural Biology* 2008; **18**:27–34.

358. Stevens AJ, Brown ZZ, Shah NH, Sekar G, Cowburn D, Muir TW. Design of a Split Intein with Exceptional Protein Splicing Activity. *J Am Chem Soc* 2016:jacs.5b13528.

359. Chirino AJ, Ary ML, Marshall SA. Minimizing the immunogenicity of protein therapeutics. *Drug Discov Today* 2004; **9**:82–90.

360. De Groot AS, Scott DW. Immunogenicity of protein therapeutics. *Trends Immunol* 2007; **28**:482–490.

361. Jawa V, Cousens LP, Awwad M, Wakshull E, Kropshofer H, De Groot AS. T-cell dependent immunogenicity of protein therapeutics: Preclinical assessment and mitigation. *Clinical Immunology* 2013; **149**:534–555.

362. Alton K, Stabinsky Y, Richards R, Ferguson B. Production, characterization and biological effects of recombinant DNA derived human IFN-α and IFN-γ analogs. *The Biology of the Interferon System* 1983:119–128.

363. Ohage EC, Wirtz P, Barnikow J, Steipe B. Intrabody construction and expression. II. A synthetic catalytic Fv fragment. *Journal of Molecular Biology* 1999; **291**:1129–1134.

364. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* 2002; **99**:2754–2759.

365. Witthöft T. Review of consensus interferon in the treatment of chronic hepatitis C. *Biologics* 2008; **2**:635–643.

366. Desjarlais JR, Berg JM. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proceedings of the National Academy of Sciences* 1993; **90**:2256–2260.

367. Chen J, Stites WE. Packing Is a Key Selection Factor in the Evolution of Protein Hydrophobic Cores †. *Biochemistry* 2001; **40**:15280–15289.

368. Anbar M, Gul O, Lamed R, Sezerman UO, Bayer EA. Improved Thermostability of Clostridium thermocellum Endoglucanase Cel8A by Using Consensus-Guided Mutagenesis. *Applied and Environmental Microbiology* 2012; **78**:3458–3464.