IMPERIAL COLLEGE LONDON

FACULTY OF NATURAL SCIENCES

DEPARTMENT OF LIFE SCIENCES

THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

EXPLORING AND EXPANDING THE SOLENOID PROTEIN UNIVERSE

BY: GEORGI IVANOV NIKOV

SUPERVISOR: DR JAMES W. MURRAY

OCTOBER 2022

ABSTRACT

Solenoid proteins are a class of tandem repeat proteins with a unique elongated and modular architecture, which lends itself to design and engineering. Design and engineering efforts benefit from well-annotated solenoid repeat data. Although there are databases of solenoid proteins and methods for their detection, there is room for improvement in the coverage of solenoid protein space and the quality of solenoid repeat annotations. In this thesis, a new convolutional neural network method for solenoid protein detection and annotation was developed: SOLeNNoID, and competitive performance against current state-of-the-art-methods was demonstrated. SOLeNNoID was then used to discover hundreds to thousands of solenoid proteins in the Protein Data Bank and the AlphaFold Database and the trends in these findings were discussed. Finally, to extend beta-solenoid design work from the Murray group, multiple beta-solenoid designs were evaluated in the laboratory: consensus hexapeptide repeat proteins, and SynRFR variants with two adjacent loops, SpyTag/SpyCatcher moieties, and truncations designed for spontaneous self-assembly. Taken together the solenoid protein tools, databases and experimental data in this thesis help to explore and expand the solenoid protein universe and could have applications in solenoid protein biotechnology and bioinformatics. All of the code produced in this thesis can be found at https://github.com/gnik2018.

STATEMENT OF ORIGINALITY

I hereby declare that this thesis is my own work produced over the course of my PhD and that all other content provided herein is appropriately referenced and acknowledged.

COPYRIGHT

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ACKNOWLEDGEMENTS

Although this thesis bears my name, it would not have been possible to complete without the help and support of the many kind and brilliant people around me. Therefore, I would like to thank:

My supervisor, Dr. James W. Murray for his support and guidance on this journey and the knowledge and advice he has shared with me along the way. I have learned a lot and grown a lot over the course of my Ph.D.

The members of the Murray group and the wider 7th floor of SEC for the support, helpful discussions and happy memories over the years. It has been a pleasure to work alongside such talented and motivated people.

My progress review panel Dr. Morgan Beeby and Prof. Michael Sternberg for their constructive advice and direction along the way.

My friends for the encouragement and support over these four years.

My parents, Gergana and Ivan, for their unwavering support and belief in me through the thick and thin, and for the sacrifices they have made so I can be here.

And finally, my girlfriend, my shining star, Daniella Pretorius, for her endless kindness, patience, encouragement and love on this long journey.

TABLE OF CONTENTS

1. INTRODUCTION ----------14
1.1. TANDEM REPEATS, SOLENOIDS AND BETA-SOLENOIDS ----------15
1.1.1. TANDEM REPEAT PROTEINS ----------15
1.1.2. SOLENOID PROTEINS ----------17
1.1.3. BETA-SOLENOID PROTEINS ----------17
1.1.4. PENTAPEPTIDE REPEAT PROTEINS ----------19
1.1.5. SYNTHETIC RFR (SYNRFR) PROTEINS ----------20
1.2. METHODS FOR SOLENOID PROTEIN ANNOTATION/DETECTION ----------22
1.2.1. SEQUENCE METHODS ----------22
1.2.2. STRUCTURE METHODS ----------22
1.2.3. NEURAL NETWORKS ----------23
1.3. DATABASES OF PROTEIN STRUCTURES, REPEATS AND MODELS ----------24
1.3.1. PROTEIN DATA BANK (PDB) ----------24
1.3.2. REPEATSDB ----------24
1.3.3. DATABASE OF STRUCTURAL REPEATS IN PROTEINS (DBSTRIPS) ----------25
1.3.4. ALPHAFOLD 2 ----------25
1.3.5. ALPHAFOLD DATABASE ----------27
1.4. SOLENOID PROTEIN DESIGN ----------28
1.4.1. DESIGN OF IDEALISED SOLENOID PROTEINS ----------28
1.4.1.1. SEQUENCE-BASED (CONSENSUS) DESIGN ----------28
1.4.1.2. STRUCTURE-BASED DESIGN ----------28
1.4.2. DESIGNS BEYOND IDEALISED STRUCTURES ----------29
1.4.2.1. DESIGNS INCORPORATING SHAPE CONTROL ----------30
1.4.2.2. DESIGNS INCORPORATING MODULARITY ----------30
1.4.2.3. DESIGNS INTRODUCING FUNCTIONALITY ----------31
1.5. PROJECT AIMS AND MOTIVATION ----------33

2. NEURAL NETWORK-BASED SOLENOID PROTEIN DETECTION AND ANNOTATION ----------34
2.1. AIMS ----------35
2.2. MATERIALS AND METHODS ----------35
2.2.1. PROTEIN STRUCTURE REPRESENTATION ----------35
2.2.2. NEURAL NETWORK BACKGROUND ----------37
2.2.2.1. NEURAL NETWORKS ----------37
2.2.2.2. CONVOLUTIONAL NEURAL NETWORKS ----------37
2.2.2.3. U-NET ----------38
2.2.2.4. U-NET NEURAL NETWORK COMPONENTS ----------39
2.2.3. DATASETS ----------40
2.2.3.1. TRAINING/VALIDATION DATASET ----------40
2.2.3.2. TESTING DATASET ----------40
2.2.3.3. ADDITIONAL BETA-PROPELLER AND BETA-SANDWICH REPEAT DATASETS ----------42

2.2.4. MODEL AND EXPERIMENTAL SETUP ----43
2.2.4.1. DATA PROCESSING----43
2.2.4.2. MODEL ARCHITECTURE AND TRAINING ----44
2.2.4.3. MODEL OUTPUT ----45
2.2.4.4. BENCHMARK DATA COLLECTION AND PROCESSING----45
2.2.4.5. PREDICTION TIME TEST ----46
2.2.4.6. LARGE STRUCTURE TEST ----46
2.2.4.7. REPEAT SPLITTING FUNCTIONALITY----46
2.3. RESULTS----48
2.3.1. BENCHMARKING ON THE SOLENOID TESTING SET----48
2.3.2. BENCHMARKING ON THE FULL TESTING DATASET ----53
2.3.3. EXAMPLE VISUALISATION ----54
2.3.4. PROCESSING LARGE STRUCTURES----54
2.3.5. ADDITION OF EXTRA CLASSES ----57
2.3.6. DETECTING MULTIPLE TANDEM REPEAT CLASSES IN A SINGLE CHAIN ----58
2.3.7. REPEAT SPLITTING FUNCTIONALITY ----59
2.4. DISCUSSION ----63
3. SOLENOID PROTEINS IN PROTEIN STRUCTURE DATABASES ----65
3.1. AIMS----66
3.2. MATERIALS AND METHODS----66
3.2.1. ALPHAFOLD DB SEARCH----66
3.2.2. PDB SEARCH----67
3.2.3. DATA PROCESSING AND VISUALISATION ----67
3.3. RESULTS----68
3.3.1. ALPHAFOLD DATABASE SEARCH ----68
3.3.1.1. SOLENOID PROTEIN STATISTICS IN THE ALPHAFOLD DATABASE ----68
3.3.1.2. SOLENOID FRACTIONS OF THE ALPHAFOLD DATABASE PROTEOMES----68
3.3.1.3. SOLENOID PROTEIN LENGTHS ACROSS PROTEOMES ----72
3.3.1.4. UNCHARACTERISED PROTEINS ----75
3.3.1.5. SOLENOID PROTEINS OF INTEREST WITH NO PDB STRUCTURES----78
3.3.2. PDB SEARCH----88
3.3.2.1. SOLENOID PROTEIN STATISTICS IN THE PDB ----88
3.3.2.2. COMPARISON WITH OTHER REPEAT DATABASES ----89
3.3.2.3. DISTRIBUTION OF EXPERIMENTAL METHODS FOR SOLENOID STRUCTURE DETERMINATION ----91
3.3.2.4. DISTRIBUTION OF FUNCTIONS FOR SOLENOID PROTEINS ----93
3.3.2.5. DISTRIBUTION OF SOLENOID PROTEINS ACROSS SPECIES----98
3.3.2.6. EXAMPLES OF SOLENOID PROTEINS DETECTED BY MY METHOD ----103
3.4. DISCUSSION----109
4. BETA-SOLENOID PROTEIN DESIGN AND ENGINEERING----110
4.1. AIMS ----111
4.1.1. CONSENSUS HEXAPEPTIDE DESIGN----111

4.1.2. SYNRFR INSERTION/FUSION VARIANTS ----112
4.1.2.1. TWO-LOOP VARIANT DESIGN ----112
4.1.2.2. SPYTAG/CATCHER FUSIONS ----112
4.1.3. SPLIT SOLENOID VARIANTS ----114
4.2. MATERIALS AND METHODS ----115
4.2.1. GENERAL METHODS ----115
4.2.1.1. CLONING, EXPRESSION AND PURIFICATION ----115
4.2.1.2. X-RAY CRYSTALLOGRAPHY ----115
4.2.1.3. MOLECULAR WEIGHT ESTIMATES ----115
4.2.2. CONSTRUCT CLONING ----116
4.2.2.1. CONSENSUS HEXAPEPTIDES ----116
4.2.2.2. TWO-LOOP DESIGN ----116
4.2.2.3. SPYTAG/SPYCATCHER DESIGNS ----116
4.2.2.4. SPLIT SOLENOID DESIGNS ----116
4.2.3. CONSTRUCT SEQUENCES ----117
4.2.3.1. CONSENSUS HEXAPEPTIDE SEQUENCES (NUMBER CORRESPONDS TO DESIGN METHOD, CAPPING SEQUENCES IN BOLD) ----117
4.2.3.2. TWO-LOOP SEQUENCE (LOOPS IN BOLD) ----117
4.2.3.3. SPYTAG/SPYCATCHER CONSTRUCT SEQUENCES (SPYTAG/CATCHER SEQUENCES IN BOLD) ----118
4.2.3.4. SPLIT SOLENOID SEQUENCES ----119
4.2.4. SPYTAG/CATCHER REACTION ----119
4.3. RESULTS ----120
4.3.1. CONSENSUS HEXAPEPTIDE DESIGN ----120
4.3.1.1. REPEAT EXTRACTION ----120
4.3.1.2. SEQUENCE PROCESSING ----122
4.3.1.3. NEW HEXAPEPTIDE SEQUENCE GENERATION ----123
4.3.1.4. COLABFOLD MODELS OF DESIGNED SEQUENCES ----125
4.3.1.5. EXPERIMENTAL CHARACTERISATION ----127
4.3.2. TWO-LOOP (2_LOOP) VARIANT DESIGN ----134
4.3.2.1. BACKBONE DESIGN ----134
4.3.2.2. SEQUENCE DESIGN ----134
4.3.2.3. IN SILICO VALIDATION ----138
4.3.2.4. EXPERIMENTAL CHARACTERISATION ----142
4.3.3. SPYTAG/SPYCATCHER VARIANT DESIGN ----144
4.3.3.1. CONSTRUCT DESIGN ----144
4.3.3.2. SPYTAG/SPYCATCHER REACTION ----144
4.3.3.3. SIZE EXCLUSION CHROMATOGRAPHY ----146
4.3.3.4. ALPHAFOLD 2 MODELS ----151
4.3.3.5. X-RAY CRYSTALLOGRAPHY ----153
4.3.4. SPLIT SOLENOID DESIGN ----156
4.3.4.1. CONSTRUCT DESIGN ----156
4.3.4.2. PULL-DOWN ----157

4.3.4.3. SIZE EXCLUSION CHROMATOGRAPHY ----------162
4.3.4.4. ALPHAFOLD 2 MODELS ----------166
4.4. DISCUSSION----------169
5. CONCLUSIONS AND FUTURE WORK----------172
5.1. CONCLUSIONS----------173
5.2. FUTURE WORK----------174
BIBLIOGRAPHY ----------176

LIST OF FIGURES

Fig. 1. Examples of tandem repeat proteins according to the Kajava classification scheme. 16

Fig. 2. Beta-solenoid cross-sectional shapes. 18

Fig. 3. An example of a pentapeptide repeat protein: a Qnr (quinolone antibiotic resistance) protein from Aeromonas hydrophila. 19

Fig. 4. SynRFR design features. 21

Fig. 5. Intuition behind distance matrices. 36

Fig. 6. Schematic illustration of a convolution operation. 38

Fig. 7. Illustration of training data processing for input into the neural network. 43

Fig. 8. Multi-class confusion matrices for PRIGSA2, RepeatsDB-Lite, TAPO and SOLeNNoID on the solenoid testing set. 51

Fig. 9. Multi-class confusion matrix for the SOLeNNoID model on the full testing dataset. 53

Fig. 10. Example output for PDB ID 4OQT produced by SOLeNNoID. 54

Fig. 11. Relationship between protein length and prediction time for SOLeNNoID method on the YEAST v1 AlphaFoldDB dataset. 55

Fig. 12. SOLeNNoID alpha-solenoid predictions on chain A of PDB ID 7SUD. 57

Fig. 13. SOLeNNoID class predictions for additional repeat protein classes. 58

Fig. 15. Repeat splitting processing/output for PDB ID 3VBP chain E. 60

Fig. 20. Structural features of the AlphaFold model for human mesothelin. 79

Fig. 21. Structural and sequence feature of the AlphaFold model of P. falciparum HPR1. 81

Fig. 22. Structural features of AlphaFold models of dgf-1 family proteins and structurally related proteins from the PDB. 83

Fig. 23. Structural and sequence features of the AlphaFold models of CsgA and CsgB curlin subunits. 85

Fig. 24. Structural features of the AlphaFold model of G. max Ycf3 and the AlphaFold Multimer model of a complex between A. thaliana Ycf3 and Y3IP1. 87

Fig. 25. Venn diagrams showing overlap between the structures detected as solenoid by SOLeNNoID and solenoid structures in the RepeatsDB and DbStRiPs databases. 89

Fig. 26. Example of an incorrect LRR classification in DbStRiPs (PDB ID 5L13). 90

Fig. 27. Numbers of solenoid structures detected by SOLeNNoID determined by different experimental methods. 92

Fig. 31. Structural and sequence features of the D. melanogaster NOMPC channel (5VKQ). 104

Fig. 32. The structure of L. santarosai serovar Shermani rLRR20. 105

Fig. 33. Structural features of FwdC and FhcC. 107

Fig. 34. Prediction by SOLeNNoID for eIF2B subunit epsilon (PDB ID 6O85). 108

Fig. 35. Example of a natural RFR protein containing two loops - HetL (PDB ID 3DU1) and SpyCatcher/SpyTag (PDB ID 4MLI). 112

Fig. 36. Repeat extraction example output for three hexapeptide repeat proteins. 121

Fig. 37. Hexapeptide repeat sequence logo and structure. 122

Fig. 38. Structure prediction models for the consensus sequence and cap structures. 124

Fig. 39. AlphaFold structure prediction statistics for hexapeptide sequences generated by each of the four methods. 126

Fig. 40. Ni-NTA purification gels for hexapeptide 1-4 proteins. 128

Fig. 41. Size exclusion chromatography data for hexapeptide 2,3 and 4. 132

Fig. 42. Rosetta macromolecular models for two-loop designs. 136

Fig. 43. Next-generation KIC energy landscapes for designs 19_04242 and 19_05903. 139

Fig. 44. IUPred3 disorder predictions for SynRFR24.1 and the 19_04242 design. 140

Fig. 45. MolProbity validation of design 19_04242. 141

Fig. 46. Top-ranked AlphaFold model of design 19_04242. 141

Fig. 47. SDS-PAGE gel of 2_loop Ni-NTA IMAC fractions. 142

Fig. 48. Putative 2_loop spherulites found in the JCSG+ screen (Molecular Dimensions) in two different conditions. 143

Fig. 49. The beta-hairpin t1428 SynRFR variant. 144

Fig. 50. SpyTag/Catcher reaction gels showing individual proteins and reaction products. 145

Fig. 53. Examples of crystals for nSC-Beta1 and nST-Beta1. 153

Fig. 54. Crystal structures for nST-Beta1 and nSC-Beta1. 155

Fig. 55. The split solenoid split point shown on PDB ID 4YC5. 156

Fig. 56. Pull-down with simultaneous addition of parts to the resin. 158

Fig. 57. Pull-down with sequential addition of parts to the Ni-NTA resin. 161

Fig. 59. ColabFold models for part A- part B complexes with and without the N-terminal Met of part B, and SynRFR24.1. 167

Fig. 60. Alternative ColabFold models for the part A1-part B complex. 168

LIST OF TABLES

Table 1. Training, validation and testing set statistics. 41

Table 2. Class label values. 44

Table 3. Comparison of method performance on the solenoid testing set as a multi-class classification problem. 49

Table 4. Comparison of method performance on the solenoid testing set as a binary (solenoid/non-solenoid) classification problem. 50

Table 5. Fraction of solenoid testing set structures of each class with solenoid residues detected for each method. 50

Table 6. Performance of SOLeNNoID on the whole testing set as a multi-class classification problem. 53

Table 7. Comparison of the ability of the four methods to process structures with more than 1000 residues. 56

Table 8. Statistics for AFDB entries detected as solenoids by SOLeNNoID. 68

Table 9. Statistics for PDB entries detected as solenoids by SOLeNNoID. 88

Table 10. Number of structures detected by SOLeNNoID and not present in other databases and percentage increase over entries in RepeatsDB. 90

Table 11. Shared functional annotations across solenoid protein classes. 98

Table 12. Methods used for the design of consensus hexapeptide sequences. 123

Table 13. Metrics calculated over the 5 models produced for each sequence by ColabFold. 127

Table 14. Energy landscape features for designs 19_05903 and 19_04242. 139

ABBREVIATIONS

TR - tandem repeat; TRP - tandem repeat protein; HEAT - HEAT repeat named after **H**untingtin, elongation factor 3 (**E**F3), protein phosphatase 2 A (PP2**A**), and the yeast kinase TOR1 ; ARM - armadillo repeat ; LRR - leucine rich repeat; RFR - repeat five residue; SynRFR - synthetic RFR; PRP - pentapeptide repeat; sfGFP - super folder Green Fluorescent Protein; TRUST - Tracking Repeats Using Significance and Transitivity; T-REKS- identification of **T**andem **RE**peats in sequences with a **K**-mean**S** based algorithm; TRAL-Tandem Repeat Annotation Library; MSA - multiple sequence alignment; TAPO - **TA**ndem **P**r**O**tein detector; PRIGSA - Protein Repeat Identification by Graph. Spectral Analysis; ARD - Alpha-Rod Detector; PDB - Protein Data Bank; RepeatsDB - Repeats DataBase; DbStRiPs - Database of Structural Repeats in Proteins; cryo-EM - cryo-Electron Microscopy; CATH - **C**lass, **A**rchitecture, **T**opology, **H**omologous superfamily; SCOP - Structural Classification Of Proteins; ReUPred - Repeat Protein Unit Predictor; DSSP - Define Secondary Structure of Proteins; CASP - Critical Assessment of Protein Structure Prediction; GDT_TS - global distance test total score; pLDDT - predicted Local Distance Difference Test; pAE - predicted aligned error; ANK - Ankyrin Repeat; TPR - tetratricopeptide repeat; PPR - pentatricopeptide repeat; RTX - repeats in toxin; DARPin - designed ankyrin repeat protein; DHR - designed helical repeat; Fc - antibody constant region; MHC - major histocompatibility complex; CTPR - consensus tetratricopeptide repeat protein; hTNKS - human tankyrase proteins; CNN - convolutional neural network; ReLU - rectified linear unit; ELU - exponential linear unit; GPU - graphics processing unit; TIM-barrel - triose-phosphate-isomerase-barrel; MCC - Matthews correlations coefficient; EBI - European bioinformatics institute; HPC - high performance computing; RMSD - root mean square deviation; FNIP - folliculin-interacting protein; LPS - lipopolysaccharide; CA - colanic acid; IQR - interquartile range; MPF - megakaryocytic potentiating factor; HPR - heptatricopeptide repeat; OPR - octatricopeptide repeat; DGF-1 - dispersed gene family 1; ER - endoplasmic reticulum; NOMPC - no mechanoreceptor potential C; eIF2B - eukaryotic translation initiation factor 2B; VWM - Leukoencephalopathy with vanishing white matter; LB - lysogeny broth; IPTG - isopropyl β-D-1-thiogalactropyranoside; Ni-NTA - nickel-nitrilotriacetic acid; BLAST - basic local alignment search tool; LAM - loop analyser mover; KIC - kinematic closure; IMAC - immobilised metal ion affinity chromatography; PEG - polyethylene glycol; sfGFP - super folder green fluorescent protein; ST - SpyTag; SC - SpyCatcher; nST - N-terminal SpyTag; nSC - N-terminal SpyCatcher; iST - internal SpyTag; iSC - internal SpyCatcher; SDS-PAGE - sodium dodecyl sulphate polyacrylamide gel electrophoresis; ProteinMPNN - protein message passing neural network; GenKIC - generalised kinematic closure; pLDDT - predicted local-distance difference test; pAE - predicted aligned error; pTM - predicted TM-score to true structure

1.
INTRODUCTION

1.1. TANDEM REPEATS, SOLENOIDS AND BETA-SOLENOIDS

1.1.1. TANDEM REPEAT PROTEINS

Tandem repeat proteins (TRPs) consist of multiple identical or highly similar modules and are defined mostly by intra-repeat and inter-repeat interactions between residues adjacent in the primary sequence (Parmeggiani & Huang, 2017) as opposed to the longer-range interactions between residues distant in the primary sequence observed in globular proteins. According to Kajava's tandem repeat protein classification scheme there are five main classes of tandem repeat proteins (Fig. 1) (Kajava, 2012). Class I consists of short (1-to-2-residue) repeats which form crystallites, whereas class II comprises fibrous repeating structures such as collagen and alpha-helical coiled coils. Class III includes solenoid and non-solenoid 'open' structures, characterised by a modular elongated structure which is maintained by inter-repeat interactions. Class IV comprises modular 'closed' structures where the N- and C-terminal repeats of the protein interact. Class V is made up of beads-on-a-string proteins, where each repeat is a small independently folded domain connected to adjacent repeats via linker peptides. Examples of open tandem repeat structures (Class III) include solenoid proteins such as HEAT, ARM, LRR and pectate lyases, whereas beta-propellers and beta-barrels fall into the group of closed tandem repeat proteins (Class IV). Delucchi et al (Delucchi et al., 2020) estimate that 50.9% of proteins contain at least one tandem repeat, defined here as an adjacently repeated amino acid pattern. Tandem repeat proteins are thought to evolve through gain and loss of tandem repeat modules by replication slippage (DNA strand denaturation and mispairing, leading to the introduction or removal of di- and trinucleotides into the DNA sequence) and recombination, in addition to substitutions and indels, causing module diversification (Schaper & Anisimova, 2015). Tandem repeats are found both in prokaryotic and eukaryotic organisms and have important functions in health and disease (Andrade & Bork, 1995; Brody & Lee, 2016; Kajava, 2012; Kajava & Steven, 2006). Furthermore, due to their modular nature, TRPs can be more amenable to design and engineering compared to globular proteins. Designed TRPs have applications in therapy (Boersma & Plückthun, 2011) and nanotechnology (Gurunatha et al., 2016; Mejías et al., 2016; Sanchez-deAlcazar et al., 2018).

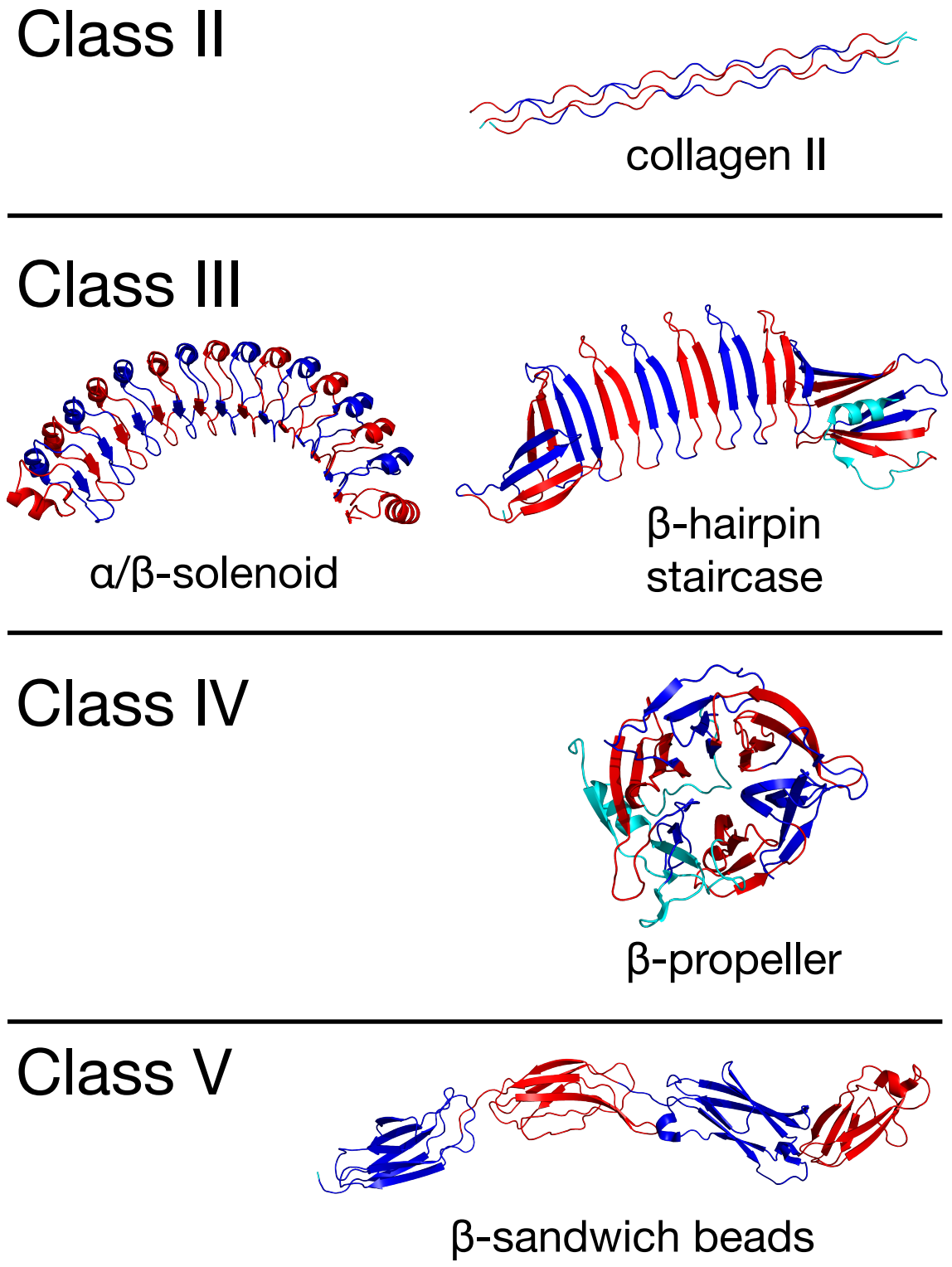


Fig. 1. Examples of tandem repeat proteins according to the Kajava classification scheme. Alternate repeats are coloured in blue and red. No crystal structures are available for class I so it is omitted. Class II (fibrous repeat) example: collagen II (PDB ID: 6HG7) (Gebauer et al., 2018): each Gly-X-Y motif is coloured in alternating blue and red. Class III (open) solenoid example: *Leptospira interrogans* LRR protein LIC11098 (PDB ID: 4U08) (Miras et al., 2015); class III non-solenoid example: *Borrelia* outer surface protein A (OspA) (PDB ID: 3EC5) (Biancalana et al., 2009). Class IV (closed) example: *Bacillus* phytase (PDB ID: 1H6L) (Shin et al., 2001). Class V (beads-on-a-string) example: human CD45 extracellular region, domains d1-d4 (PDB ID: 5FMV) (Chang et al., 2016).

1.1.2. SOLENOID PROTEINS

Solenoid repeat modules fold cooperatively into an extended structure with a continuous hydrophobic core and solvent accessible surface (Javadi & Itzhaki, 2013). The polypeptide chain in solenoid structures follows a helical trajectory, where individual repeats pack against each other and depend on each other for folding (Parmeggiani & Huang, 2017). Therefore, solenoid proteins deviate from the shape of globular proteins. Individual repeats (one turn around the solenoid axis) are made up of secondary structure elements ranging from alpha-helices and beta-strands to 310 and polyproline II helices connected by loops. Repeat lengths in solenoid proteins span between 12 and 45 amino acids. Longer solenoid repeats tend to have higher alpha-helical content, while shorter repeats have higher beta-strand content. The most conserved positions in solenoid sequences are the hydrophobic residues forming the extended core, as well as residues important for the topology of the polypeptide chain.

1.1.3. BETA-SOLENOID PROTEINS

The beta-solenoid architecture is comprised of beta strands and the turns, which connect them (Kajava & Steven, 2006). Beta-solenoid proteins form an elongated structure usually with limited curvature but with either a right-handed or a left-handed twist around the helical axis. The axial rise per turn is approximately 4.8 Å - equivalent to the alpha carbon distance between a pair of adjacent parallel beta strands. Beta-solenoid proteins have an extended hydrophobic core, which is sometimes protected by N- and C-terminal capping structures. Capping structures range from a regular repeat with an increased number of hydrophilic residues to a distorted repeat with a helix running across the cross-section of the end to cover the hydrophobic surface. Five main cross-sectional shapes are observed, based on the number of beta strands and the types of turns observed (Fig. 2): B-shaped, L-shaped, oval, rectangular (or square) and triangular. The number of residues per repeat range between 12 and 30 amino acids and can deviate within the same protein primarily due to insertions into the turns, which connect the beta strands, but other deviations in the repeating unit are also possible, such as a 3_{10}- or alpha-helical stretch. In addition, the solenoid structure can comprise three separate chains, or side-on trimeric solenoid assemblies can be formed.

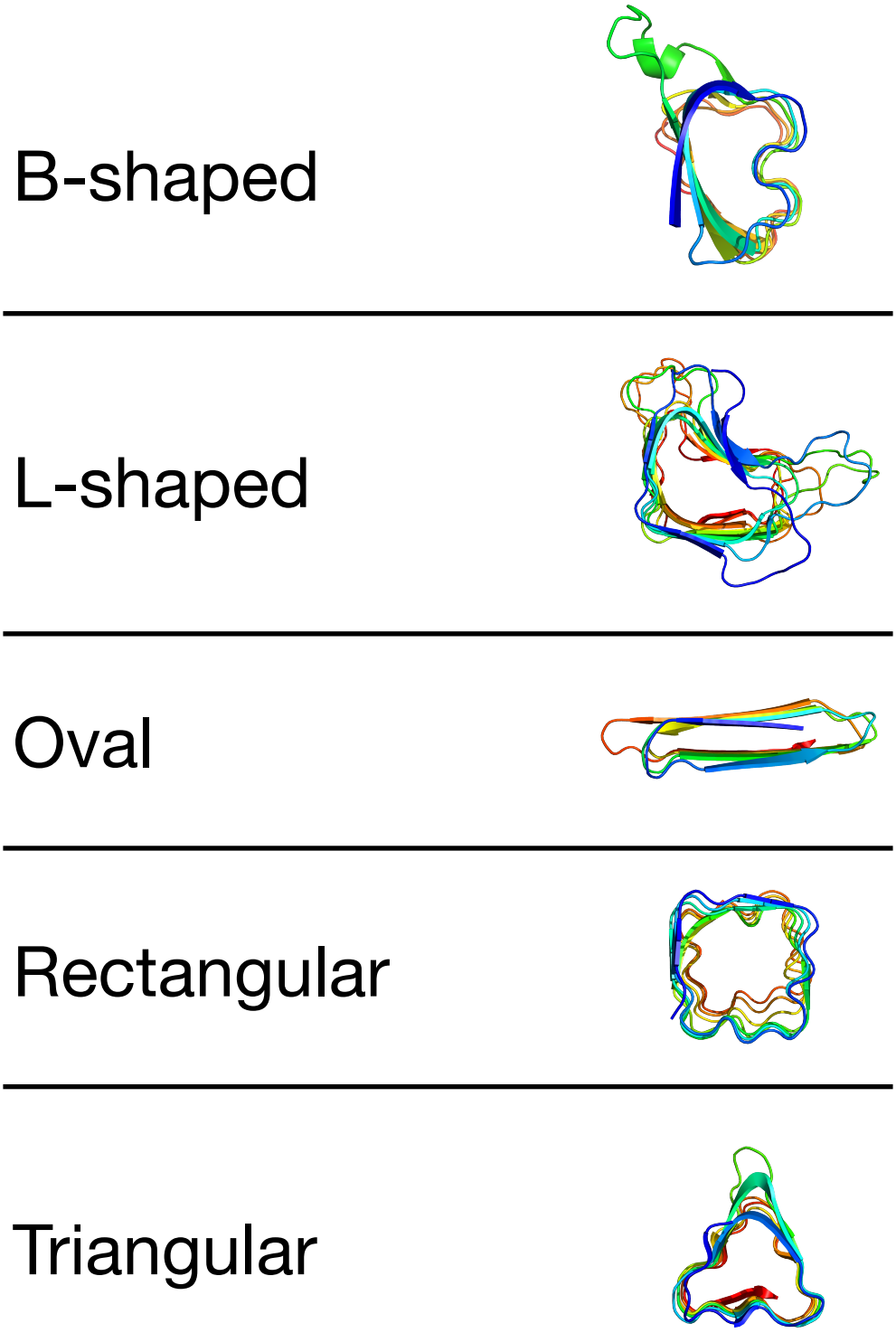


Fig. 2. Beta-solenoid cross-sectional shapes.
Fragments of the structure are shown, coloured N-to-C to highlight the cross-sectional shape. B-shaped example: alpha subunit of *Azospirillum brasilense* glutamate synthase (residues 1264-1396 from PDB ID 1EA0) (Binda et al., 2000). L-shaped example: monomer of tailspike protein of Sf6 bacteriophage (residues 235-503 from PDB ID 4URR) (Kang et al., 2016). Oval example: *Rhagium inquisitor* antifreeze protein (residues 21-101 from PDB ID 4DT5) (Hakim et al., 2013). Rectangular example: *Mycobacterium tuberculosis* MfpA (residues 1-148 from PDB ID 2BM5) (Hegde et al., 2005). Triangular example: monomer of gamma-class carbonic anhydrase (residues 11-134 from PDB ID 6IVE) (Wang et al., 2019).

1.1.4. PENTAPEPTIDE REPEAT PROTEINS

Pentapeptide repeat (also known as repeat-five-residue (RFR)) proteins are a class of beta-solenoid proteins originally identified in 1998 (Bateman et al., 1998). Pentapeptide repeat proteins have a right-handed beta-helical structure with a square cross-section and 20 residues per turn of the beta-helix. The turns connecting beta strands at a near 90 degree angle are either type I, II or IV beta-turns (R. Zhang & Kennedy, 2021).The parameters of the PRP beta-helix are determined by the types of beta turns between beta strands, as well as the presence of loops/insertions (R. Zhang et al., 2019).

Pentapeptide repeat proteins are found across all domains of life (Shah & Heddle, 2014) with nearly 39,000 members in the Pfam (Mistry et al., 2021; Punta et al., 2012) superfamily. Despite the abundance of sequence information about pentapeptide repeat proteins, there are only two examples of known functions for these proteins. The first is bacterial DNA gyrase inhibition by MfpA (Hegde et al., 2005) and Qnr (Vetting et al., 2011) proteins. These pentapeptide repeat proteins act as DNA mimics and block binding of fluoroquinolone to the complex between DNA and DNA gyrase. Loops extending from the beta-helix play a crucial role for the function of these proteins. The second known function is binding of the botulinum neurotoxin A to the eukaryotic synaptic vesicle protein 2C pentapeptide repeat luminal domain (Benoit et al., 2013) .Here, the interaction is between beta strands in the PRP domain and in the botulinum neurotoxin A protein.



Fig. 3. An example of a pentapeptide repeat protein: a Qnr (quinolone antibiotic resistance) protein from *Aeromonas hydrophila.*
This structure has PDB ID 3PSS (Xiong et al., 2011). The solenoid core is coloured pink, while loops are coloured teal. Left: rectangular cross-sectional shape of the RFR fold. Each side of the rectangle is made up of a single beta strand and the beta strands are connected by turns. Right: Right-handed beta-helical structure of the RFR fold. One full turn of the beta-helix comprises four beta-strands.

1.1.5. SYNTHETIC RFR (SYNRFR) PROTEINS

Initial design exploration of the pentapeptide repeat fold was carried out by MacDonald et al (MacDonald et al., 2016), producing synthetic 'SynRFR' pentapeptide repeat proteins (most notably PDB IDs 4YC5 and 5DNS). SynRFR scaffolds were designed by consensus sequence design using the pentapeptide repeats from PDB IDs 2J8K (Vetting et al., 2007) and 3DU1 (Ni et al., 2009). N- and C-terminal caps were derived from the *Nostoc sp.* HetL (Ni et al., 2009) and the the *Mycobacterium tuberculosis* MfpA protein (Hegde et al., 2005), respectively. The C-terminal cap contains a homodimerisation interface. SynRFR scaffold proteins are thermostable with a melting temperature (denaturation midpoint - Tm) above 65 °C and can support insertions into beta-turns ranging from a *de novo* designed beta-hairpin to a full sfGFP domain inserted inside the aforementioned beta-hairpin.

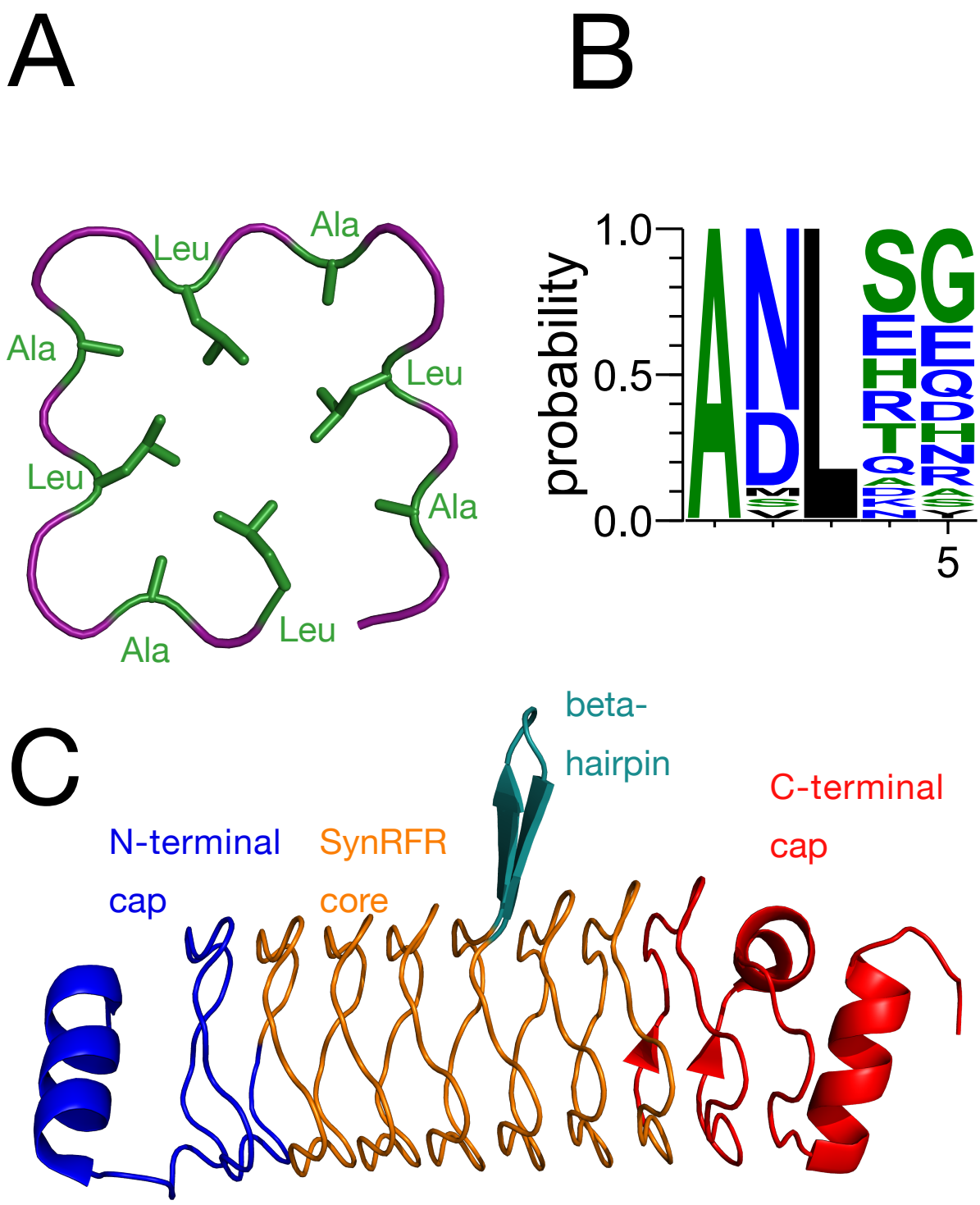


Fig. 4. SynRFR design features.
A) Square cross-section of the SynRFR24.1 (Beta1) protein (PDB ID 4YC5). Consensus hydrophobic core residues are shown in green in stick representation. B) Consensus pentapeptide repeat sequence generated using WebLogo (Crooks et al., 2004; Schneider & Stephens, 1990) for residues 50-194 of PDB ID 4YC5. The plot shows the probability of finding a specific amino acid at each of the 5 positions in the 5-residue repeat. C) The beta-hairpin t1428 SynRFR variant (PDB ID 5DNS). The N-terminal capping structure is shown in blue, the C-terminal capping structure is shown in red, the beta-hairpin is shown in teal and the SynRFR core is shown in orange.

1.2. METHODS FOR SOLENOID PROTEIN ANNOTATION/DETECTION

Both the study of naturally occurring tandem repeat proteins and design methodologies rely on pre-existing TRP data, particularly well-annotated and potentially parametrised (Woolfson et al., 2015) individual tandem repeats and whole TRP structures. Tandem repeat detection and analysis is primarily carried out on either the sequence or structure level.

1.2.1. SEQUENCE METHODS

Protein sequence information is much more abundant than structural information. For example, the 03 Aug 2022 release of the UniProtKB/TrEMBL (Bateman et al., 2021) protein sequence database contains 226,771,949 entries, while, according to the latest statistics, the PDB (Berman et al., 2000; Burley et al., 2021) comprises only 190,644 protein-containing structure entries, representing a 1190-fold difference.

A recent summary of sequence-based tandem repeat detection methods (Pellegrini, 2015) shows methods ranging from using self-alignment matrices (e.g. TRUST (Szklarczyk & Heringa, 2004)), k-means clustering (T-REKS (Jorda & Kajava, 2009)), Hidden Markov Model (HHRepID (Biegert & Söding, 2008)), discrete Fourier transform (REPETITA (Marsella et al., 2009)), to neural-network-based approaches (Fournier et al., 2013; Pagès et al., 2019; Palidwor et al., 2009). The TRAL library (Delucchi et al., 2021; Schaper et al., 2015) integrates many of the above tandem repeat detectors as well as post-processing, refinement, and annotation modules to evaluate and improve the detected repeats. In addition, the machine-learning-based Tally-2.0 method (Perovic et al., 2020) can be used for repeat validation by inputting a multiple sequence alignment (MSA) of previously predicted tandem repeats sequences.

However, due to the evolutionary divergence between tandem repeat units on the sequence level, sequence-level tandem repeat detection is more challenging than structural analysis due to deviations from ideality in tandem repeat proteins such as insertions within and between repeats, and sequence divergence between individual repeats (Chakrabarty & Parekh, 2020).

1.2.2. STRUCTURE METHODS

With the recent development of AlphaFold 2 (Jumper et al., 2021, Mirdita et al., 2022), RoseTTAFold (Baek et al., 2021) structural models can be derived from sequences lacking experimental structures (Tunyasuvunakool et al., 2021), thereby increasing the available pool of structural information in the form of high-confidence predictions. There are several structure-based methods which can either detect tandem-repeat-containing proteins or individual tandem repeats within proteins. They can be split into two broad classes: those based on algorithmic pipelines and those based on neural networks.

Two of the most recent tandem repeat detection methods are TAPO (do Viet et al., 2015) and RepeatsDB-lite (Hirsh et al., 2018). TAPO comprises an ensemble of 7 sub-methods, which use features such as periodicities of atomic coordinates, conformational alphabet strings, residue contact maps and secondary structure vectors. The outputs of these sub-methods are combined using a support vector machine (SVM) model, a non-neural-network machine learning algorithm, to detect tandem repeats. RepeatsDB-lite relies on iteratively matching tandem repeats from a pre-defined library to the input structure to expand the predicted repeating region. PRIGSA2 (Chakrabarty & Parekh, 2020) is another recent method for tandem repeat detection using analysis of the pattern of the eigenvector centrality profile of the contact network of a protein and secondary structure annotation. A further method, CE-Symm 2.0 (Bliven et al., 2019) is based on structure self-alignment and can report details about the type and axes of internal symmetry found in a protein, as well as repeat boundaries. ConSole is an older method for detecting solenoid protein regions using contact map template matching and SVM classification (Hrabe & Godzik, 2014), while Parra et al (Parra et al., 2013) perform an exhaustive alignment of all possible sub-structures in a protein to find 'tiles', which correspond to repeating units in the protein using the TopMatch algorithm (Sippl & Wiederstein, 2008).

1.2.3. NEURAL NETWORKS

While many tandem repeat prediction algorithms have been developed, there have been relatively few attempts to use neural networks. Deep neural network approaches have been very successful in other domains (Alom et al., 2018) and are increasingly finding applications in biology such as biomedical research (Baldi, 2018), protein structure prediction (AlQuraishi, 2019; Jumper et al., 2021; Senior et al., 2020), generative modelling for protein design (Strokach et al., 2020), prediction of the effects of mutations (Meier et al., 2021; Strokach et al., 2021), and protein 'hallucination' (Anishchenko et al., 2021; Jendrusch et al., n.d.) to name a few. Protein hallucination is the process of iteratively optimising a random amino acid sequence (e.g. by a Monte-Carlo approach) to arrive at a final sequence predicted to fold into a 3D protein structure, using feedback from a structure prediction neural network such as AlphaFold. In deep learning features and robustness to variation are automatically learned from the data to the point where entire human-engineered prediction pipelines can be replaced with an end-to-end differentiable model (AlQuraishi, 2019).

The ARD method (Fournier et al., 2013; Palidwor et al., 2009) is neural-network-based but specific to alpha solenoid proteins. It takes an amino acid sequence as an input and detects the periodical arrangement of hinge regions of alpha solenoid repeating units to globally classify a protein as containing alpha solenoid regions or not. The DeepSymmetry method (Pagès et al., 2019) is more generally applicable and processes 3D protein structures using a 3D convolutional neural network. This method detects internal symmetries and proteins containing tandem repeats.

1.3. DATABASES OF PROTEIN STRUCTURES, REPEATS AND MODELS

Databases of protein structures are crucial for the development of tandem repeat protein detection and annotation methods and for the design of new tandem repeat proteins. The main repository for experimentally determined protein structures is the PDB, the main database for predicted structural models is the AlphaFold databases, while RepeatsDB and DbStRiPs contain annotated tandem repeat annotations.

1.3.1. PROTEIN DATA BANK (PDB)

The Protein Data Bank (PDB) (Berman et al., 2000; "Crystallography: Protein Data Bank," 1971) is the first digital open access biological data resource (Burley et al., 2017) and currently a database of global significance. The number of structural entries has increased from about 10 in 1971 to over 188,000 in 2022 (rcsb.org statistics). The majority of yearly released structures are determined by X-ray crystallography (73.6% in 2021), with electron microscopy in second place (23.4% in 2021). Increasing numbers of structures are determined by cryo-EM because of the 'resolution revolution' (Kühlbrandt, 2014). The evolution of the PDB over the years is exemplified by the migration from the old fixed-width format, which traces back to punched cards and has a 99,999 limit to the number of atoms it can hold to the .mmcif format, which has no such limitations, has explicit relationships between data items and can be easily extended by adding more data items to the mmCIF dictionary. Each data deposition contains annotations and metadata in addition to atomic coordinates. The importance of the PDB as a global resource is underscored by the 1.5 million downloads/day performed by users during 2016 and the over 200 data resources, which depend on RCSB PDB data and services (Burley et al., 2017). The utility of the data provided by the PDB includes phasing of new structures by molecular replacement, analysis to determine common or differing features between structures and classification and analysis of protein domains in databases such as CATH (Sillitoe et al., 2021) and SCOP (Andreeva et al., 2014, 2020).An additional and increasingly important use case for PDB structural data is as input to machine learning prediction algorithms such as AlphaFold (Evans et al., 2021; Jumper et al., 2021; Senior et al., 2020).

1.3.2. REPEATSDB

The PDB provides fundamental structural information, but not tandem repeat annotations. This role is fulfilled by the RepeatsDB database (di Domenico et al., 2014; Paladin et al., 2017, 2021). While domain information databases exist for both the sequence (e.g. Pfam) and structure (e.g. CATH and SCOP) levels, Pfam can fail to detect a number of repeats, while CATH and SCOP do not provide explicit repeat annotations and rely on a secondary-structure-centric protein classification scheme, which does not appropriately reflect overall structural similarities between non-globular tandem repeat protein classes. The RepeatsDB database has gone through three

iterations so far and uses the RAPHAEL (Walsh et al., 2012), ReUPred (Hirsh et al., 2016) and RepeatsDB-Lite (Hirsh et al., 2018) repeat detection and annotation methods in addition to manual curation. The database in its current form integrates information from CATH, SCOP, Pfam, DSSP (Joosten et al., 2011; Kabsch & Sander, 1983) and UniProt (Bateman et al., 2021). The RepeatsDB database divides tandem repeat proteins into classes based on repeat length and the number of repeats in the protein, along with the extent of interaction between the repeating units, according to the Kajava schema (Kajava, 2012). The additional 'Clan' and 'Family' levels introduced in the last iteration (Paladin et al., 2021) require sequence similarity information in addition to structural information. The database contains tandem repeat predictions and statistics, including a repeating unit similarity matrix and multiple sequence alignment for each database entry. In its most recent iteration, the RepeatsDB database is populated using RepeatsDB-lite, a template-based method which matches individual repeats from a repeat unit library to novel structures. Therefore, the method is strongly dependent on the quality of its repeat unit library. This means that there is scope for the use of additional methods, which could find tandem repeat proteins which might be inaccessible to RepeatsDB-lite.

1.3.3. DATABASE OF STRUCTURAL REPEATS IN PROTEINS (DBSTRIPS)

The Database of Structural Repeats in Proteins (DbStRiPs) database (Chakrabarty & Parekh, 2021) is populated using the PRIGSA2 method (Chakrabarty & Parekh, 2020) and is a complementary resource to RepeatsDB. The DbStRiPs databases focuses on class III (elongated structures) and class IV (toroid structures) of the Kajava tandem repeat classification scheme. DbStRiPs provides information about the prediction confidence, secondary structure annotation of tandem repeats in the protein, in addition to linking to several protein databases such as Pfam and UniProt. Thirty percent of the structures found in DbStRiPs are also found in the RepeatsDB database, indicating that there is significant non-overlap between the structures found by the PRIGSA2 and RepeatsDB-Lite methods and therefore suggesting that additional methods could explore additional parts of tandem repeat protein space.

1.3.4. ALPHAFOLD 2

The performance of the AlphaFold 2 method in the CASP14 competition was a landmark achievement, surpassing the performance of the first AlphaFold (Senior et al, 2020). For the first time, protein structures could be predicted with accuracy comparable to experimental structures even without homologues (Jumper et al., 2021).The main metric of the CASP competition is the GDT_TS (global distance test total score), which approximately equates to the percentage of protein residues that have been predicted correctly (Bouatta et al., 2021). Even in the free modelling category, which requires *ab initio* prediction, a median GDT_TS of 87 was achieved.

The main advancements presented by the AlphaFold 2 model are discussed by Bouatta et al (Bouatta et al., 2021). The first is the use of transformers (Vaswani et al., 2017), which automatically learn what length scales are relevant from the input training data and can therefore find long-range dependencies. Convolutional neural networks used by preceding methods (Senior et al., 2020; J. Yang et al., 2020) instead focus on local neighbourhoods and can omit long-range dependencies. The second is the use of a novel rotationally and translationally equivariant transformer in the structure module of the AlphaFold 2 network, allowing AlphaFold 2 to 'reason' directly in 3D compared to previous methods operating in 2D and using a final physics-based module to reconstruct the 3D structure from the 2D restraints such as trRosetta (J. Yang et al., 2020). Finally, AlphaFold2 is end-to-end differentiable and so jointly optimises all components of the network, ensuring self-consistent results.

Despite its strong performance, AlphaFold 2 has some limitations. One of these limitations derives from the data used to train the model - the PDB. As the majority of structures in the PDB are determined by X-ray crystallography, this means that the model predicted by AlphaFold 2 effectively represents the conformation of a protein under crystallisation conditions, but it could also mean that AlphaFold2 is biased towards the sequence-to-structure relationship of crystallisable structures. Moreover, the monomer version of AlphaFold 2 incorrectly predicts the structure of proteins, which undergo a disorder-to-order transition on complex formation. As AlphaFold 2 uses multiple sequence alignment information, proteins with shallow MSAs or completely novel sequences might be predicted incorrectly (Perrakis & Sixma, 2021). In addition, in the case of sequences with multiple possible conformations or a conformational change on, for example, small molecule or ion binding, AlphaFold 2 will produce a single conformation as the output.

AlphaFold 2 produces two confidence measures: pLDDT and pAE. pLDDT (predicted local-distance difference test) is a prediction of the lDDT-Calpha (Mariani et al., 2013) measure - a superposition-free score, which compares local interatomic distances in the reference structure and the predicted model. A high pLDDT value indicates a high confidence in the prediction of the local environment around a residue. pAE (predicted aligned error) is, in contrast, a measure of inter-domain accuracy. It takes the form of a matrix where the value at position (x,y) indicates the position error predicted by AlphaFold 2 at residue x, when the ground truth and predicted structures are aligned on residue y. A low inter-domain pAE suggests packing between the domains, while a high inter-domain pAE indicates that the relative orientation of the domains cannot be confidently predicted. Therefore, for several of the above edge-cases, AlphaFold will produce low confidence values, indicating that the predictions are likely incorrect.

The predictions produced by AlphaFold 2 have several uses. For example, AlphaFold 2 models can be used for structure determination by molecular replacement (McCoy et al., 2022), for fitting models into cryo-EM maps, for testing hypotheses about structure/function relationships in monomers and large macromolecular machines (Humphreys et al., 2021; Qin et al., 2022).

Predictions on large databases of sequences can lead to the discovery of previously unobserved folds, new structure-based evolutionary insight, and more structural data for machine learning applications (*Great Expectations – the Potential Impacts of AlphaFold DB | EMBL*, 2021).

1.3.5. ALPHAFOLD DATABASE

A logical next step after the excellent performance of AlphaFold at CASP14 was to use the model to produce a large database of predicted structures (Varadi et al., 2022) to close the gap between sequences (approx. 220 million unique sequences in UniProt) and structures (approx 180,000 in the PDB). Tunyasuvunakool et al (Tunyasuvunakool et al., 2021) produced predictions for the human proteome. Even though humans are the best represented species in the PDB, the database contains only 35% of human proteins. 35.7% of all residues in the human proteome have AlphaFold 2 predictions with pLDDT above 90 (at which point side chain rotamers are 80% correct for a PDB test dataset), which is double the number of experimentally resolved residues. Porta-Pardo et al (Porta-Pardo et al., 2022) estimate that AlphaFold predictions increase structural coverage for the human proteome by 28% over experimental structures and homology models. However, the authors note that the impact of AlphaFold 2 predictions is likely to be greatest in non-human organisms. Indeed, the first version of the AlphaFold database (Varadi et al., 2022) contains predictions for the human proteome and 20 other species from all domains of life, totalling over 360,000 predicted structures.

1.4. SOLENOID PROTEIN DESIGN

1.4.1. DESIGN OF IDEALISED SOLENOID PROTEINS

Many solenoid proteins have been designed so far using different strategies. All design efforts depend on a wealth of structural and sequence information from sources such as the databases or tandem repeat methods described above. Solenoid protein designs fall into one of two categories: sequence-based and structure-based.

1.4.1.1. SEQUENCE-BASED (CONSENSUS) DESIGN

Consensus sequence design involves the selection of a domain of interest, sequence acquisition and pre-processing from databases such as Pfam and subsequent iterative MSA refinement to arrive at a consensus sequence (Porebski & Buckle, 2016), which reflects the most frequent amino acid(s) at each position in the domain sequence. A position in the consensus sequence is defined as conserved, if a single amino acid occurs with a probability greater than 50%, semi-conserved, if two to four amino acids occur with a higher probability and non-conserved otherwise (Chang et al., 2021; Mosavi et al., 2004). The consensus sequences produced by this approach display increased stability (as defined by melting temperature) compared to their natural counterparts. This is possibly due to a (partial) return to a more stable common ancestor (Porebski & Buckle, 2016). Tandem repeat protein families can further be summarised by a consensus repeat sequence, which can highlight residues important for the architecture of the fold, as well as residues, which are allowed to vary and are thus available for design. Tandem copies of the consensus repeat sequence then constitute an 'idealised' example of the fold. Examples of solenoid folds designed by consensus methods include ANK (Binz et al., 2003; Mosavi et al., 2002), ARM (Parmeggiani et al., 2008; Reichen et al., 2014), TPR (Main et al., 2003), PPR (Coquille et al., 2014; Gully et al., 2015), LRR (Lee et al., 2012; Parker et al., 2014), RFR (MacDonald et al., 2016) and RTX (Shur et al., 2013). Purely sequence-based consensus design becomes challenging when sequence conservation between individual repeats within a protein and between proteins is lower. Furthermore, consensus sequences are strongly dependent on the number of sequences they derive from and the quality of the MSA. However, with the advent of AlphaFold2 sequence-to-structure prediction, it is now possible to routinely generate structure-guided MSAs for repeat proteins.

1.4.1.2. STRUCTURE-BASED DESIGN

1.4.1.2.1. TEMPLATE-BASED

Template-based structure design involves the use of existing protein backbones at either the single repeat, full protein, or intermediate level. Rämisch et al selected, symmetrically docked and sequence optimised a two-repeat template module from porcine ribonuclease inhibitor to produce a novel, toroidal 'donut' design (Rämisch et al., 2014). Zhu et al instead threaded a putative TPR sequence generated from multiple copies of the RPS20 protein sequence onto a hybrid template

composed of a consensus designed TPR protein and RPS20 and selected the best scoring designs for experimental validation (Zhu et al., 2016). Voet et al successfully designed a symmetrical beta-propeller protein by selecting a single propeller blade from PDB ID 1RWL (Good et al., 2004), symmetrically docking into a circularly symmetrical assembly using RosettaDock (Gray et al., 2003; Marze et al., 2018) and mapping putative ancestral sequences onto the complete structural template (Voet et al., 2014). Although beta-propeller proteins do not follow a solenoid architecture, it is possible to extend this approach to solenoid proteins as well.

1.4.1.2.2. DE NOVO DESIGN

De novo solenoid design rests on the assumption that a tandem repeat protein can be summarised by a single repeat. Designing the backbone and sequence of the entire solenoid protein is then broken down to symmetrically propagating the design of its repeat. This method can be used to design new members of existing solenoid families (Parmeggiani et al., 2015) by incorporating protein family structure and sequence information, or to design novel topologies (Brunette et al., 2015; Doyle et al., 2015). Parmeggiani et al design new members of the ANK, ARM, TPR, HEAT, LRR and WD40 repeat families using RosettaRemodel (Huang et al., 2011) guided by family-specific structural constraints, followed by sequence design using a sequence profile term informed by the protein family MSA. Design steps are mirrored across all repeats of the proteins (Parmeggiani et al., 2015). Doyle et al successfully designed closed, left-handed alpha-solenoids, which were not previously observed in the PDB, by another symmetrical design process in Rosetta, which included a pseudo-energy term to favour desired inter-repeat geometry and clustering and resampling stages to focus on the conformational space around promising designs (Doyle et al., 2015). Brunette et al explored the helix-loop-helix-loop structural motif as a repeating unit. By varying the number of residues in each of the two helices and the two turns and performing symmetrical Monte Carlo fragment assembly in Rosetta, the authors generate 83 alpha-solenoid designs with sequences unrelated to known repeat proteins, suggesting that evolution might not have thoroughly explored the alpha-solenoid fold space (Brunette et al., 2015).

1.4.2. DESIGNS BEYOND IDEALISED STRUCTURES

Most of the above examples have involved the design of idealised, perfect solenoid structures with no associated function. However, natural solenoid structures display several deviations from an ideal, monomeric form. The shapes of solenoid proteins are controlled by structural idiosyncrasies in some of the repeats, they can form multimeric non-covalent assemblies by side-to-side and end-to-end contacts, and they have an array of functions such as protein/nucleic acid binding and enzymatic reactions.

1.4.2.1. DESIGNS INCORPORATING SHAPE CONTROL

Designs with control over the shape of solenoid or solenoid-like proteins have been achieved by a number of groups (Brunette et al., 2020; Elgamacy et al., 2018; Park et al., 2015; Youn et al., 2017).

Park et al designed LRR proteins with controlled curvature in a three-step approach (Park et al., 2015). First, they designed self-compatible building blocks, following the Parmeggiani et al protocol (Parmeggiani et al., 2015). Next, they superposed a 2-unit module from the previous step with a 1-unit junction module and used RosettaCM (Song et al., 2013) to generate a hybrid junction structure, which was then sequence optimised. Designs containing self-compatible building block modules and designs containing junction modules were crystallised and their structures used as new models for the final design stage. Here, the authors represented the design problem as a network where building block modules are the nodes and junction modules are the edges. Traversing the edges of the network then led to LRR proteins with unique curvature.

Brunette et al took curvature control further by de novo design of junctions between previously generated helical repeat domains (Brunette et al., 2020). The authors experimented with a helical superposition approach as well as a more general fragment assembly approach. In the fragment assembly case, all secondary structure definitions consistent with design criteria were generated and corresponding backbones were produced by Monte Carlo fragment assembly with the changes propagated to the repeat protein domain following the junction. Finally, sequence design was performed for the junction residues and the helical domain residues in contact with them and models satisfying design filters and *in silico* de novo structure prediction in Rosetta were selected for experimental characterisation.

1.4.2.2. DESIGNS INCORPORATING MODULARITY

As a solenoid protein chain consists of several self-complementary repeats, one way to design building blocks is to split the solenoid chain into multiple smaller fragments. Watson et al successfully split a consensus-designed armadillo repeat protein into two fragments, which could interact non-covalently to recapitulate the structure of the full-length protein (Watson et al., 2014). They found that the C-terminal fragment was well folded, while the N-terminal fragment exhibited a molten globule state but underwent coupled folding and binding in the presence of the C-terminal fragment. As armadillo repeat proteins recognise peptides in an extended conformation, Michel et al elaborated on the above approach by fusing a target peptide to a C-terminal ARM fragment and demonstrating selective full-length protein reconstitution with the N-terminal ARM fragment which has the highest affinity to the target peptide (Michel et al., 2018). In a further publication (Michel et al., 2019), the target peptide was added orthogonally, instead of being fused to one of the fragments and leads to predominant interaction between the fragments with the highest affinity for it from a mixture of fragments.

Beyond splitting an existing solenoid domain, novel interfaces can be introduced, enabling the controlled build-up of higher order structures. Mohan et al designed a novel homodimeric interface between two copies of an erythropoietin receptor-binding designed ankyrin repeat protein (Mohan et al., 2019). Insertion of non-binding consensus repeats into the dimeric DARPin then allowed for control of the angle and distance between erythropoietin receptors, which resulted in variation of the strength of receptor agonism.

Going further, large macromolecular assemblies can be built using solenoid components. Pyles et al designed tunable hexagonal lattices guided by a mica surface using pre-existing designed helical repeat proteins as the starting point (Pyles et al., 2019). The authors selected a C3 symmetry and *in silico* docked assemblies matching the mica lattice geometry were subjected to backbone remodelling and sequence design in Rosetta. To control the spacing of the resulting honeycomb lattice, DHR designs with varying numbers of repeats were designed.

Shen et al designed self-assembling helical filaments using DHR proteins as a starting point (Shen et al., 2018). The construction of a helical filament is achieved by repeatedly applying a rigid-body transform to the DHR building block to generate a helix, which is then repeated into a defined n-fold cyclic symmetry. Each building block must then form at least two interfaces and is sequence-optimised with propagation to all other monomers in the filament. Finally filters such as shape complementarity, low interaction energy and limited buried unsatisfied H-bonds are applied to reach the final designs.

Divine et al designed nano cages displaying antibodies (Divine et al., 2021). The nano cages consist of three components: proteins that bind the Fc region of an antibody, a designed helical repeat monomer and a helical homo-oligomer. Fusions of these three components that satisfy geometric requirements, contact requirements and have undergone sequence repacking and optimisation were then selected for experimental characterisation.

1.4.2.3. DESIGNS INTRODUCING FUNCTIONALITY

Idealised designed solenoid proteins commonly do not perform any function. There are two main avenues for the introduction of functionality - a) by using the extended cavities and surfaces formed by solenoid proteins to introduce ligand or cofactor binding sites and b) by using the regions connecting adjacent repeats to introduce loops or entire protein domains.

The work of Mejias et al and di Meo et al exemplifies the former approach (Mejias et al., 2019; di Meo et al., 2019). Mejias et al introduced [4Fe-4S] iron-sulfur cluster binding sites into a consensus designed tetratricopeptide repeat protein by making cysteine mutations into a tandem repeat pair. The resulting protein-[4Fe-4S] constructs were folded and redox-active, raising the possibility of using TPR proteins as scaffolds for electron transfer circuits. Di Meo et al used a

designed homodimeric alpha-solenoid protein as a starting point to construct Diels-Alderase enzymes. The two monomers were genetically fused and a single Cys mutation was introduced into the C-terminal domain of the resulting fusion protein. A copper (II)/phenantroline or copper(II)/terpyridine complex was then covalently attached to the Cys residue to yield a catalytic centre for the Diels-Alder reaction. Hence, closed solenoid architectures could be used to design non-globular enzymes.

The latter approach is taken by Correnti et al (Correnti et al., 2020). The authors designed self-assembling modules such as a module with 6 33-residue repeats, which tetramerises to form the full structure. Cargo domains were then incorporated into the N-termini and/or 4 internal surface loops of the modules. Examples of the cargo the authors fuse to the modules include SH2 peptide-binding domains, SpyCatcher domains (Zakeri et al., 2012), single chain class I MHC complexes and a single-chain antibody Fv domain targeting CD28. Thus, the system described by Correnti et al allows for modular and symmetric display of cargo domains from both prokaryotic and eukaryotic sources.

Perez-Riba et al performed a large-scale study of loop insertions into consensus designed tetratricopeptide repeat proteins (Perez-Riba et al., 2019). The results show decreased cooperativity between repeats and reduced enhancement of thermodynamic stability on addition of more repeats when long loops are inserted at each inter-repeat junction. However, even constructs with loop insertions at every inter-repeat junction were highly thermostable, raising the possibility of functionalising the CTPR proteins through designed loop insertions. In addition, Ripka et al show that a 68-residue two-repeat CTPR protein is thermostable to >60 °C when a range of intrinsically disordered peptides up to 58 residues in length are inserted at the inter-repeat junction (Ripka et al., 2021). As an example, the same group also produced CTPR proteins with the Keap1-binding peptide from Nrf2 inserted into the inter-repeat junction (Madden et al., 2019). The grafted peptide exhibited higher affinity for Keap1 than the native peptide. Going further, Diamante et al designed mono- and multivalent CTPR proteins, including CTPR trimers with human tankyrase binding peptides inserted into inter-repeat junctions (Diamante et al., 2021). The proteins were stable and downregulated hTNKS activity, showing the strength of this peptide loop insertion approach.

1.5. PROJECT AIMS AND MOTIVATION

As discussed above, there are many tandem repeat detection and annotation methods using different representations of protein structures and different methods to process this structural information. Each method has strengths and weaknesses and therefore the databases populated using these methods will also reflect this. Database entries in RepeatsDB and DbStRiPs are of variable quality, even when manually reviewed, highlighting the need for additional approaches and validation. Traditionally, structural tandem repeat databases derive from the PDB. However, with the advent of AlphaFold 2, there is now an additional database (with high-confidence structural predictions, instead of experimentally determined structures), which can be scanned for solenoid and other tandem repeat structures. The value of high-quality solenoid protein annotation such as is found in the above databases is that it allows a) detailed analysis of the sequence-structure relationship of TRPs, b) leveraging a large amount of data to be able to design not just 'perfect' or regular TRPs but proteins of arbitrary shape and insertions at positions known to be well tolerated. Alternatively, or concomitantly, it could enable the generation of a collection of LEGO-like TR building blocks which can then be assembled into designed nanostructures using methods such as Elfin (Yeh et al., 2018) and WORMS (Hsia et al., 2021). A well-annotated dataset could also be used by generative methods (Strokach & Kim, 2022), which could aid the exploration of TR structure/sequence space. The design/engineering of beta-solenoid proteins is especially important as they are relatively under-explored but have interesting and useful properties.

Therefore, the aims of this project are to:

1. develop a convolutional neural network-based method for solenoid protein detection and annotation and compare it to current state-of-the-art approaches - addressed in **Chapter 2**;
2. search protein structural databases (PDB and AlphaFold DB) to find solenoid proteins not covered by existing databases and survey the results to find trends and statistics - addressed in **Chapter 3**;
3. advance beta-solenoid protein design and engineering by designing insertions and fragments of the SynRFR protein scaffold, as well as designing consensus hexapeptide repeat proteins: - addressed in **Chapter 4**.

2. NEURAL NETWORK-BASED SOLENOID PROTEIN DETECTION AND ANNOTATION

2.1. AIMS

My aim is to develop a neural-network-based method for solenoid region prediction using structural information as input. The methods most closely related to this goal (and used as a benchmark) are RepeatsDB-lite, TAPO and PRIGSA2, as my goal is to provide high-quality residue-level solenoid region annotation. Other methods either produce only a global prediction of whether a protein contains tandem repeats or arrive at repeat definition from a more global property of the protein such as its symmetry type.

I focus on the solenoid sub-classes (alpha, alpha/beta, beta) (Kobe & Kajava, 2000) of tandem repeat proteins and propose a structure-based supervised deep learning method, SOLeNNoID, for high-quality annotation of repeating regions. I use a 2D distance matrix representation of 3D protein structure and a semantic segmentation approach, employing a U-Net convolutional neural network architecture (Ronneberger et al., 2015) to produce a probability that a given amino acid in the protein structure belongs to one of the solenoid classes, or is non-solenoid. To my knowledge, this is the most fine-grained approach to repeating region prediction to date.

In this chapter, I demonstrate that this approach is competitive with the current state-of-the-art in terms of annotation on a curated solenoid protein test dataset. In addition, I show that SOLeNNoID can find new solenoid proteins in a database, and more importantly ones that are inaccessible to other methods. My results demonstrate that this method can complement existing approaches in refining solenoid protein discovery in databases and solenoid protein annotation.

2.2. MATERIALS AND METHODS

2.2.1. PROTEIN STRUCTURE REPRESENTATION

A protein is a linear polymer of amino acids, which can fold into a specific three-dimensional shape. The position of each atom in a protein structure can be described by a set of xyz coordinates relative to a defined origin. The 3D shape of the protein can then be represented by a 2D all-to-all matrix of distances between Cα atoms (Holm, 2020). This distance matrix is orientation-independent: rotation or translation of the entire 3D structure will not produce a change in the distance matrix values. Cα atoms are chosen as they trace the path of the protein backbone. Each row **n** of the distance matrix contains the distances from amino acid $\mathbf{x_n}$ to itself and all other amino acids in the protein. Information about the structural environment of that amino acid is contained within the pattern of distances. In addition, protein structure can be recovered from the distance matrix, if the chirality of the structure is known.

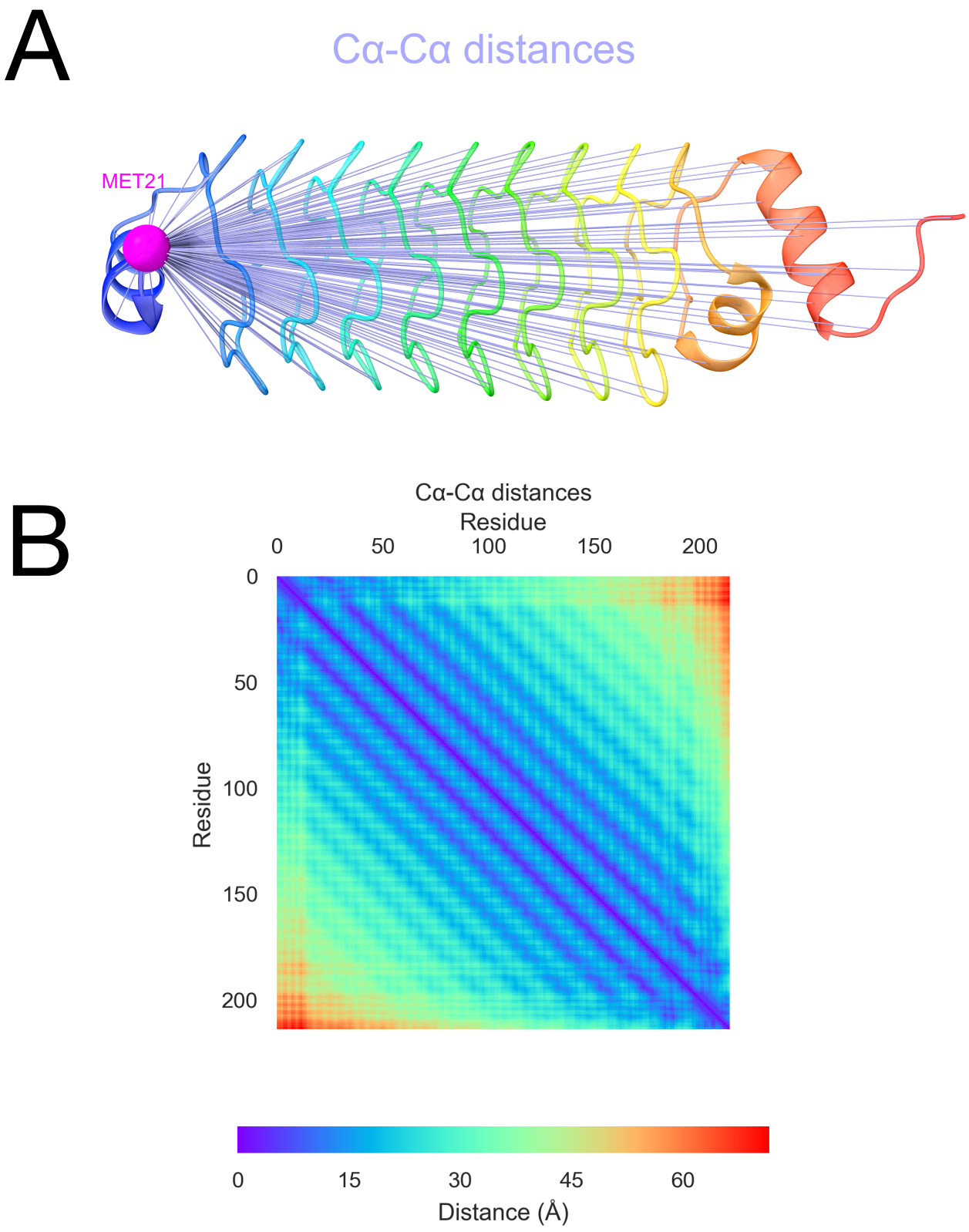


Fig. 5. Intuition behind distance matrices.
A) The distances from residue 21 to all other residues in the protein (PDB ID 4YC5, made using ChimeraX (Pettersen et al., 2021)). B) The distance matrix for PDB ID 4YC5 showing all interatomic distances (made using seaborn (Waskom, 2021)).

2.2.2. NEURAL NETWORK BACKGROUND

2.2.2.1. NEURAL NETWORKS

The process of machine learning consists of estimating and optimising the parameters of a statistical model so that the learned model can be applied to a specific task. In supervised deep learning, inputs and their corresponding labels are fed into the network. The neural network then aims to approximate the label based on the input by learning a set of weights and biases during the training process. The output of each layer goes through a non-linear function as the relationship between input features and labels that the neural network attempts to approximate is non-linear. An iterative process of parameter optimisation is then carried out to improve the approximation of the desired outputs using stochastic gradient descent and backpropagation (Rumelhart et al., 1986). Deep learning is a universal approach, which automatically learns features from the data and robustness to variation. Deep networks require large amounts of data to train but once trained, they can then very rapidly make predictions on new data. Although they comprise multiple layers and parameters, neural networks are not necessarily inscrutable 'black boxes' and it is possible to interpret their data representation and decision process (Olah et al., 2020).

2.2.2.2. CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks (CNNs) are a subset of neural networks, inspired by visual information processing systems in animals (Hubel & Wiesel, 1968). CNNs are particularly well-suited to processing image information, whether in 2 or 3 dimensions. The original application of CNNs was for character recognition tasks (Cun et al., 1989) but CNNs really came into prominence with AlexNet (Krizhevsky et al., n.d.) achieving state-of-the-art performance on the ImageNet classification challenge. Convolutional neural networks apply filters to an image or an input feature map to produce an output feature map. This process is repeated in multiple layers, enabling hierarchical feature extraction. In contrast to fully connected networks, a hidden layer neuron is connected to a subset of neurons in the previous layer and shares weights with neurons corresponding to the same feature map, decreasing the number of parameters in the network (Aloysius & Geetha, 2018).

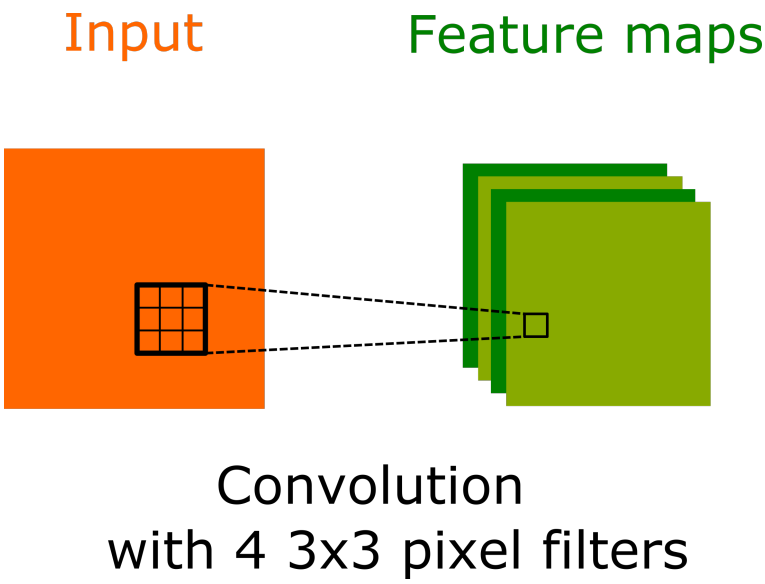


Fig. 6. Schematic illustration of a convolution operation.
A filter with dimensions 3 pixels by 3 pixels consisting of numerical weights (learnable parameters in a neural network) is overlaid with a filter-sized patch of the input (e.g. an image). The dot product between the the filter and the filter-sized patch is taken, resulting in a single value. Once the filter is applied to all overlapping patches in the input, the result is a map of all of the dot products, which is called a feature map. The feature map has reduced dimensions relative to the input. Applying four convolutional filters to the input results in a stack of four feature maps.

2.2.2.3. U-NET

U-Net (Ronneberger et al., 2015) presents a different use case for convolutional neural networks. Instead of classifying entire images, U-Net can provide pixel-wise class 'segmentation' labels. The architecture of U-net consists of a contracting path to capture context from the input image and a symmetric expanding, or upsampling path that increases the resolution of the output. Features from the contracting path are combined with output from the upsampling path, enabling more precise localisation. The authors demonstrate that this network can learn from limited training data and can be used to segment images of arbitrary size.

2.2.2.4. U-NET NEURAL NETWORK COMPONENTS

2.2.2.4.1. CONVOLUTIONAL LAYER

A convolutional layer consists of a number of filters of defined dimensions, comprising the parameters of the layer. In 2D convolutional operations, the filter is a 2D matrix of learnable weights. Each filter is scanned across each filter-sized patch of input data in an overlapping fashion. The result of this operation is a feature map, made up of the dot product of the filter with each input patch. The filter can find the feature it corresponds to anywhere in the input data, thereby showing translational invariance. Feature maps are then passed through a non-linear activation function such as ReLU. Multiple filter weights are learned by the network in parallel, whereas stacks of convolutional layers extract increasingly higher-order features from the input. For example, in image classification early convolutional layers might detect lines, while later convolutional layers might detect faces.

2.2.2.4.2. TRANSPOSED CONVOLUTION

In contrast to a pooling operation, a transposed convolution increases the size of the output feature map compared to the input feature map. The input of a transposed convolution can be thought of as the output of a direct convolution on some initial feature map. (Dumoulin et al., 2018). The transposed convolution allows the reconstruction of the shape of the initial feature map. An example of a use case for transposed convolutions is presented by auto encoders. Using auto encoders, dimensionality reduction of the input can be achieved by learning to reconstruct the input. The encoder arm learns a compressed representation of the input, while the decoder attempts to reconstruct the input from this compressed representation.

2.2.2.4.3. ACTIVATION FUNCTIONS

Activation functions are required to approximate the non-linear mapping of input features to the ground truth label.

The rectified linear unit (ReLU) activation function (Glorot et al., 2011) is commonly used in convolutional neural networks. It is expressed as:

$$f(x) = max(0,x)$$

The exponential linear unit (ELU) activation function is an improved version of the ReLU activation function (Clevert et al., 2015) and is expressed as:

$$f(x) = \begin{cases} x & \text{if} \quad x > 0 \\ \alpha(exp(x)-1) & \text{if} \quad x \leq 0 \end{cases}$$

The softmax activation function is frequently used as a final activation function for multi-class classification problems. It returns the probability that a given input belongs to each possible class in a multi-class classification problem and is expressed as:

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)}$$

2.2.2.4.4. DROPOUT

Dropout (Hinton et al., 2012; Srivastava et al., 2014) is a regularizing mechanism, whereby a randomly selected subset of activations within a layer is set to 0 during training of the neural network. This prevents co-adaptation of neurons in the neural network, whereby multiple neurons can extract the same or similar features and lead to overfitting on the training dataset. The proportion of all activations to be zeroed can be defined as the dropout rate hyperparameter.

2.2.2.4.5. POOLING

Pooling operations (Scherer et al., 2010) downscale their inputs. In Max Pooling, a filter of a defined size is overlapped with the inputs. The maximum value in the overlapping region is returned. In contrast to filters in convolutional layers, the filter is applied to non-overlapping filter-sized patches of the input, unless the stride value is set to 1. Pooling therefore downsamples the input feature maps, thereby reducing the sensitivity of the neural network to the location of features in either the input or intermediate feature maps.

2.2.3. DATASETS

2.2.3.1. TRAINING/VALIDATION DATASET

The training/validation dataset is a modified version of the dataset produced by Marsella et al (Marsella et al., 2009). The dataset consists of a set of protein PDB files and label files, which contain a label for each residue in the PDB file. The label for each residue is an integer value, corresponding to either of the four possible classes – non-solenoid, beta-solenoid, alpha/beta-solenoid, or alpha-solenoid. In total, the dataset consists of 246 non-solenoid, 52 beta-solenoid, 17 alpha/beta-solenoid and 33 alpha-solenoid proteins. The training and validation datasets were derived by a 80:20 (test:validation) random split.

2.2.3.2. TESTING DATASET

The testing dataset consists of 50 non-solenoid, 15 beta-solenoid, 7 alpha/beta-solenoid and 11 alpha-solenoid proteins. The non-solenoid proteins were derived from the PDBe (Velankar et al., 2016) while the solenoid proteins were derived from manually reviewed solenoid entries from RepeatsDB. All structures are non-redundant with respect to the training and validation structures to a TM-align (Y. Zhang & Skolnick, 2005) TM-score threshold of < 0.6. If an alignment between two proteins produces a TM-score of > 0.5, it can be assumed that the proteins generally have

the same fold in SCOP/CATH. Thus, as the neural network will be used to identify and annotate proteins broadly in the same folds as the structures in the training dataset, the less stringent threshold of TM-score < 0.6 was used. The RepeatsDB solenoid labels were amended by excluding loops from repeat definitions and adding parts of partial or imperfect repeats. Table 2 contains a number of statistics relating to these datasets, including the number of SCOP (Andreeva et al., 2014, 2020) families, for entries for which this information was available.

Training Set				
Class	**Alpha-solenoid**	**Alpha/beta-solenoid**	**Beta-solenoid**	**Non-solenoid**
Number of structures	26	13	41	196
Avg. number of residues	292	263	310	255
Avg. interresidue distance (Å)	28	26	24	20
Avg. % of solenoid residues	0.78	0.78	0.6	0
Number of SCOP families	19	7	20	181
Validation Set				
Class	**Alpha-solenoid**	**Alpha/beta-solenoid**	**Beta-solenoid**	**Non-solenoid**
Number of structures	7	4	11	50
Avg. number of residues	346	246	327	269
Avg. interresidue distance (Å)	28	23	24	19
Avg. % of solenoid residues	0.75	0.75	0.53	0
Number of SCOP families	5	5	10	50
Testing Set				
Class	**Alpha-solenoid**	**Alpha/beta-solenoid**	**Beta-solenoid**	**Non-solenoid**
Number of structures	11	7	15	49
Avg. number of residues	582	513	450	263
Avg. interresidue distance (Å)	35	34	32	23
Avg. % of solenoid residues	0.74	0.72	0.55	0
Number of SCOP families	11	3	6	24

Table 1. Training, validation and testing set statistics.

2.2.3.3. ADDITIONAL BETA-PROPELLER AND BETA-SANDWICH REPEAT DATASETS

Beta-propeller (example of closed repeat class (Type IV in the Kajava classification)) and beta-sandwich repeat (example of the beads on a string class (Type V in the Kajava classification)) structures were selected from 'Reviewed' RepeatsDB entries such that the TM-align TM-score of each structure to every other structure was less than or equal to 0.6. Then, structures with more than one tandem repeat type were removed. Label data was converted from the repeat-based RepeatsDB format to the residue-based format used for solenoid classes. Beta-propeller residues were labelled as class 4, beta-sandwich repeats as class 5 and missing data was moved to class 6. The model was retrained using this additional data.

2.2.4. MODEL AND EXPERIMENTAL SETUP

2.2.4.1. DATA PROCESSING

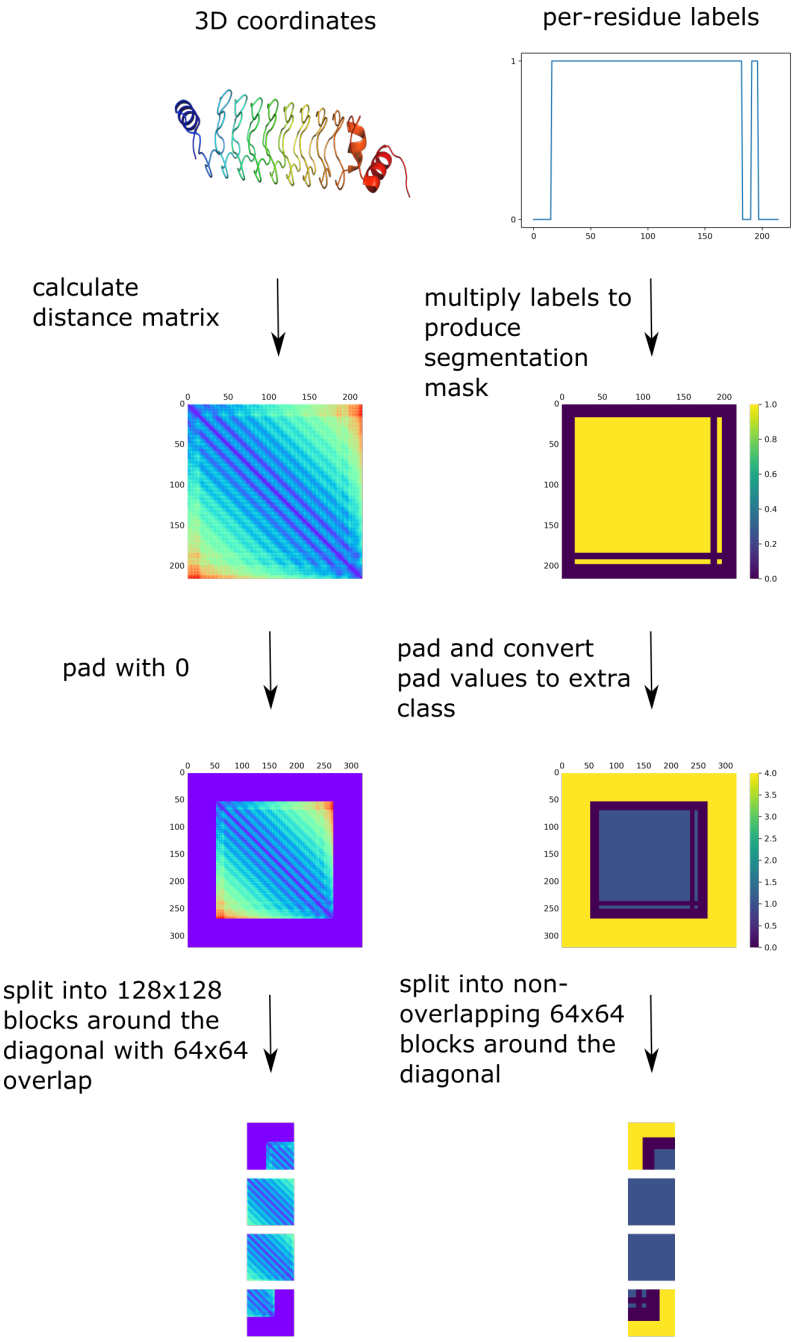


Fig. 7. Illustration of training data processing for input into the neural network. Left: structural data processing. An inter-residue distance matrix is calculated using Cα xyz coordinates, padded with 0 values and then split into overlapping 128x128 blocks centred on the diagonal. Right: label data processing. The per-residue label array for each structure is multiplied by itself elementwise to produce a label matrix (or 'segmentation mask'). The label matrix is padded, and the pad values are converted to a new class. The padded label matrix is then split into non-overlapping 64x64 blocks centred on the diagonal and corresponding to a 128x128 block of the padded distance matrix.

All scripts were written in Python3 (van Rossum & Drake, 2009). Each PDB file was processed into a distance matrix using BioPython (Cock et al., 2009) and NumPy (van der Walt et al., 2011). At points where protein chains were discontinuous, both the labels and the distance matrix entries were replaced with zero values to maintain the register of the amino acid indices. Each distance matrix was pre- and post-padded with zeroes in both x and y axes to a total length divisible by 128. The distance matrix was then normalised by the largest distance observed. Normalisation was performed for each individual structure, rather than the maximum distance observed in the dataset.

Class	Label
Non-solenoid	0
Beta-solenoid	1
Alpha/beta-solenoid	2
Alpha-solenoid	3
Missing data	4

Table 2. Class label values.

Labels were provided according to the scheme in Table 2. Label arrays were multiplied elementwise by themselves, then the square root was taken to produce a 2D label matrix (or 'segmentation mask'). The segmentation mask was then padded identically to the distance matrix. A separate class was used for missing data and padding to separate it from the non-solenoid class.

Training data consisted of randomly cropped 128x128 slices of distance matrix and segmentation masks but only the central 64x64 slice of the segmentation mask slice was used for training. The reasoning here is that the model could learn to better classify the pixels in the 64x64 slice by 'seeing' a larger 128x128 context. Random cropping was done to increase the amount of training data to improve model generalisation. Validation set distance matrices and segmentation masks were split into overlapping 128x128 slices with an overlap of 64, with final validation set segmentation masks comprising the central 64x64 slice of the 128x128 segmentation mask as in the training set. The segmentation mask labels were then converted to one-hot encoded values.

2.2.4.2. MODEL ARCHITECTURE AND TRAINING

The model code was written in Python using the Tensorflow (Abadi et al., 2016), Keras (Chollet, 2015), Scikit-learn (Pedregosa et al., 2011) and NumPy libraries. The model follows the U-Net architecture with a reduced number of filters per layer and a Gaussian noise layer before the first convolutional layer. Gaussian noise was added as a regularisation measure to reduce model overfitting. The final upsampling block of the U-Net was removed to yield a 64x64 segmentation mask prediction for each 128x128 input distance matrix slice. The network is fully convolutional, and the final convolutional layer has a softmax activation function. Early stopping and model

checkpointing callbacks were used with both callbacks tracking loss over the validation dataset. Training was performed for 100 epochs (approx. 1 day) with a batch size of 64 using the Adam (Kingma & Ba, 2015) optimiser on an NVIDIA P1000 GPU. The categorical crossentropy loss function was used. The model weights leading to the lowest loss value on the validation set were saved.

2.2.4.3. MODEL OUTPUT

An individual PDB ID and chain ID are specified. The structure (in mmCIF format) is loaded and Cα coordinates, and indices are stored. These are then used to produce the Cα distance matrix. The distance matrix is then processed as outlined in Section 2.2.4.1. to prepare it for prediction.

Predictions are made using the U-Net model, yielding segmentation masks. The masks are converted to sparse class labels and reshaped back into a size identical to the distance matrix. Finally, segmentation mask pixels are only considered if they fall within the length of the protein (are not padding) and if the x and y indices are identical (diagonal values). If these criteria are not met, then that residue is classed as non-solenoid. The non-solenoid and missing data classes are collapsed together. The resulting predictions are then split into the respective classes for visualisation. Predictions are mapped onto the structure and visualised using the nglview library (H. Nguyen et al., 2018), while the sequence is visualised using the Bokeh library (Bokeh Development Team, 2022).

2.2.4.4. BENCHMARK DATA COLLECTION AND PROCESSING

For all comparison methods, predictions were collected using the respective servers and adapted to the labelling scheme for this method. For TAPO, the top prediction was used. In contrast to SOLeNNoID, all comparison methods detect individual repeats instead of entire regions. RepeatsDB-Lite and PRIGSA2 provide class labels, whereas TAPO does not. In the cases of RepeatsDB-Lite and PRIGSA2 repeats were not included, if they were detected but predicted to belong to an incorrect class outside the scope of SOLeNNoID (for example, if the ground truth class is alpha-solenoid but instead TIM-barrel is predicted). Ground truth and comparison method labels were concatenated for all structures in the testing dataset. Model predictions were also collected and concatenated.

Benchmark metrics Used:

TP - true positives - cases where the model has predicted the positive class label

FP - false positives - cases where the model has incorrectly predicted the positive class label

FN - false negatives - cases where the model has incorrectly predicted the negative class

Precision: $P = \frac{TP}{TP + FP}$

Precision is probability that a predicted positive value is a true positive value.

Recall: $R = \frac{TP}{TP + FN}$

Recall is the probability that a true positive value is predicted as positive.

F1-score: $F_1 = 2 \times \frac{P \times R}{P + R}$

F1 score is the harmonic mean of precision and recall.

The definitions for Precision, Recall and F1-score above are for a binary classification problem but can also be applied to a multi class classification problem to obtain class-wise metrics.

Multi-class Matthews correlation coefficient (MCC) (Gorodkin, 2004; Jurman et al., 2012). In a binary classification problem, the MCC is a correlation coefficient between observed and predicted binary classifications, ranging from -1 (complete disagreement between observed and predicted classes) to +1 (complete agreement). MCC is a useful metric when assessing a model on an imbalanced dataset with classes with different numbers of samples.

The scikit-learn classification_report was used to produce Precision, Recall and F1-score metrics and matthews_corrcoef to produce the Matthews correlation coefficient measure. Confusion matrices were produced using scikit-learn confusion_matrix and ConfusionMatrixDisplay.

2.2.4.5. PREDICTION TIME TEST

The prediction time test was conducted using the 6,040 AlphaFold2-predicted *S. cerevisiae* structures (v1 database) in the EBI AlphaFold database. The script was executed on the Imperial College HPC with 16 CPU cores, 96 GB RAM and 4 RTX6000 GPUs. The prediction time was defined from before loading a structure to after obtaining solenoid predictions for the structure.

2.2.4.6. LARGE STRUCTURE TEST

10 solenoid structures with more than 1000 residues were selected from the PDB using the Advanced search option with "**Polymer Entity Sequence Length"** set to above **1000** and **"Polymer Entity Type"** set to **Protein**. Predictions were attempted for all structures using all four methods under comparison. If the resulting prediction was 'no repeats' or the structure was not processed, the method is considered unsuccessful on that structure. Thus, the test simply checks for any repeat predictions.

2.2.4.7. REPEAT SPLITTING FUNCTIONALITY

User-defined repeat splitting was achieved by recording clicked indices of equivalent positions in the nglview viewer. The recorded indices were then used to fill in repeats from the predicted solenoid residues. These repeats were then output in PDB format. The output repeats were

processed using mTM-align (Dong et al., 2018), producing a PDB file with superposed repeats, and self-similarity matrices of the repeats with TM-score or RMSD as the criterion. In addition, the fasta file produced by mTM-align was used to produce a logo plot of the repeat sequences with the logomaker library (Tareen & Kinney, 2020).

2.3. RESULTS

2.3.1. BENCHMARKING ON THE SOLENOID TESTING SET

SOLeNNoID was benchmarked against the other three methods on the solenoid testing set. The four metrics used to describe performance on this dataset were precision, recall, F1-score (harmonic mean of precision and recall) and multi-class Matthews correlation coefficient.

SOLeNNoID detects solenoid residues, rather than repeats. Therefore, repeat predictions of other methods were converted into a per-residue annotation. Unlike RDB-Lite and PRIGSA2, TAPO does not provide class labels. Therefore, there are two ways of comparing performance: one where it is assumed that TAPO would provide correct class labels and one where the predictions produced by SOLeNNoID and by TAPO are converted into a single solenoid class and benchmarked against each other thus.

If we assume that TAPO provides correct class labels, SOLeNNoID had the best precision on the non-solenoid class, the best recall on the alpha-solenoid class and the best F1-score on the non-solenoid and beta-solenoid classes (see Table 3). The PRIGSA2 method showed the highest precision on all solenoid classes but was outperformed by TAPO and SOLeNNoID on recall. Overall, SOLeNNoID outperformed PRIGSA2 and RDB-lite on F1-score and MCC and performed competitively with TAPO.

If all methods are instead converted to binary classifiers, SOLeNNoID outperformed TAPO in all metrics (see Table 4). PRIGSA2 had the highest precision on the solenoid class but it had very low recall, ultimately resulting in a low F1-score.

In addition, SOLeNNoID was able to predict solenoid residues for all structures in the dataset as shown in Table 5. However, it must be noted that the main class prediction for one of the structures, 1IVO, was beta-solenoid instead of the ground truth alpha/beta-solenoid. The poor performance of PRIGSA2 and RDB-lite on the alpha-solenoid and beta-solenoid classes, respectively, shows that these methods were unable to detect any solenoid residues for many structures in the testing set and helps to explain the poorer performance of these methods compared to TAPO and SOLeNNoID in previous comparisons (Table 3 and 4).

The full confusion matrices for all methods are shown in Fig. 8. Only SOLeNNoID and TAPO had large values across the entire diagonal (corresponding to residues with agreement between predicted and true label). For both PRIGSA2 and RDB-Lite, a large number of alpha- and beta-solenoid residues were incorrectly predicted as non-solenoid. PRIGSA2 exhibited little to no incorrect predictions between solenoid classes. This matches the high precision but poor recall for PRIGSA2 in Table 3. For RepeatsDB-Lite an especially large number of alpha-solenoid residues were predicted as non-solenoid. This is likely due to some alpha-solenoid residues not

being predicted as tandem repeat at all and some being predicted as a non-solenoid repeat type such as TIM-barrel or alpha-barrel. In addition, many alpha/beta-solenoid residues were incorrectly predicted as beta-solenoid. This is likely due to the similar repeat architecture and topology of the two classes. For TAPO and SOLeNNoID, there were still some incorrect predictions of solenoid residues as non-solenoid. The major type of incorrect predictions for TAPO were non-solenoid residues predicted as beta-solenoid, while for SOLeNNoID non-solenoid residues were incorrectly predicted as alpha-solenoid.

	Precision				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	**0.79**	0.78	0.52	0.51
	Beta-solenoid	0.82	0.77	**0.86**	0.50
	Alpha/beta-solenoid	0.95	0.86	**0.96**	0.78
	Alpha-solenoid	0.81	0.91	**0.97**	0.89
	Recall				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	0.77	0.74	**0.96**	0.89
	Beta-solenoid	0.81	**0.84**	0.16	0.55
	Alpha/beta-solenoid	0.8	**0.93**	0.73	0.15
	Alpha-solenoid	**0.91**	0.88	0.56	0.35
	F1-score				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	**0.78**	0.76	0.67	0.64
	Beta-solenoid	**0.81**	0.80	0.28	0.53
	Alpha/beta-solenoid	0.86	**0.89**	0.83	0.25
	Alpha-solenoid	0.86	**0.90**	0.71	0.50
	MCC				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
		0.75	**0.76**	0.56	0.39

Table 3. Comparison of method performance on the solenoid testing set as a multi-class classification problem.
The metrics used are precision, recall and F1-score for single classes and multi-class MCC as a global metric. Values in bold represent the best performance across methods.

	Precision				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	**0.79**	0.78	0.52	0.51
	Solenoid	0.86	0.85	**0.95**	0.89
		Recall			
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	0.77	0.74	**0.96**	0.89
	Solenoid	**0.88**	**0.88**	0.48	0.5
	F1-score				
	Method	**SOLeNNoID**	**TAPO**	**PRIGSA2**	**RDB-lite**
Class	**Non-solenoid**	**0.78**	0.76	0.67	0.64
	Solenoid	**0.88**	0.86	0.64	0.64
	MCC				
		SOLeNNoID	**TAPO**	**PRIGSA2**	**RDB-lite**
		0.65	0.62	0.45	0.39

Table 4. Comparison of method performance on the solenoid testing set as a binary (solenoid/non-solenoid) classification problem.
The metrics used are precision, recall and F1-score for single classes an multi-class MCC as a global metric. Values in bold represent the best performance across methods.

Method	Alpha-solenoid	Alpha/beta-solenoid	Beta-solenoid
SOLeNNoID	**1.00**	**1.00**	**1.00**
TAPO	0.91	**1.00**	**1.00**
RepeatsDB-Lite	0.45	0.86	0.67
PRIGSA2	0.73	0.86	0.27

Table 5. Fraction of solenoid testing set structures of each class with solenoid residues detected for each method.
Values in bold represent the best performance across methods.

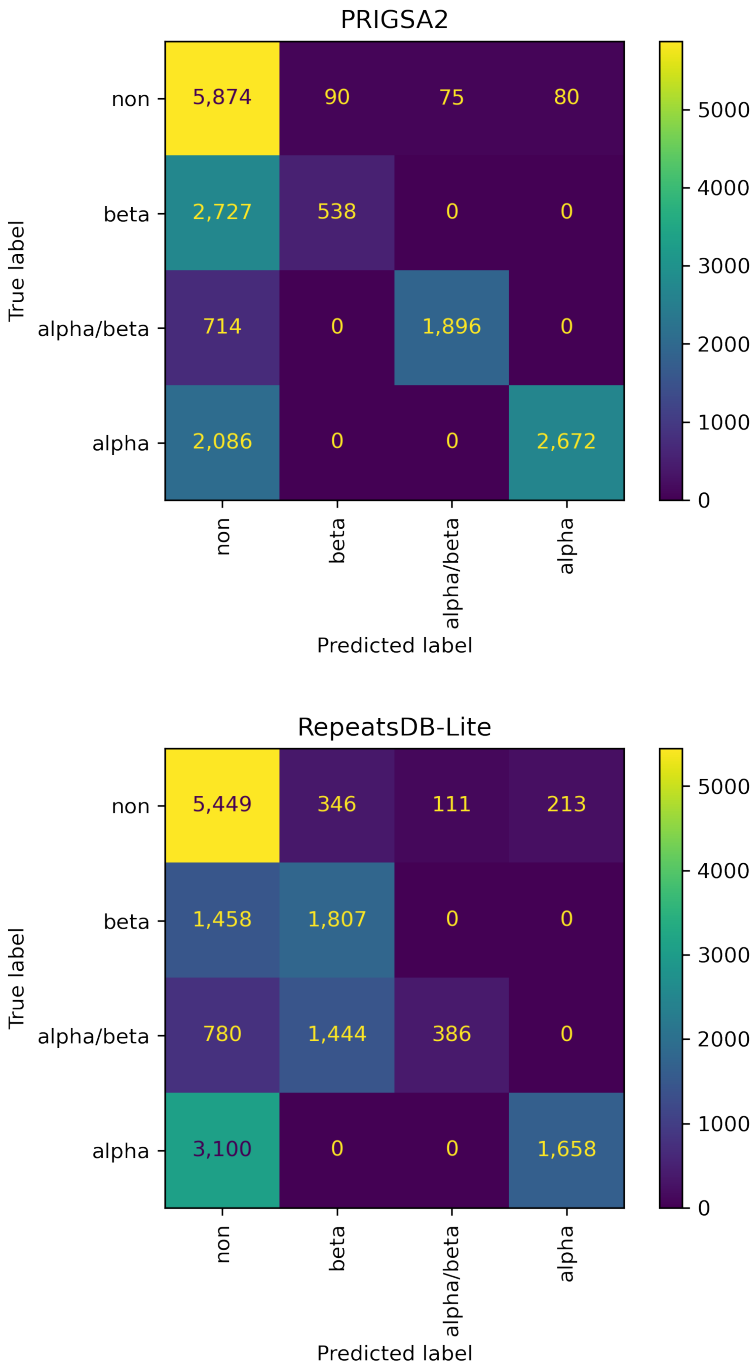


Fig. 8. Multi-class confusion matrices for PRIGSA2, RepeatsDB-Lite, TAPO and SOLeNNoID on the solenoid testing set.
True labels are shown on the y-axis and predicted labels are shown on the x-axis. The numbers represent the number of protein residues placed within each category of the confusion matrix. The colour palette goes from purple (0 residues) to yellow (large number of residues).

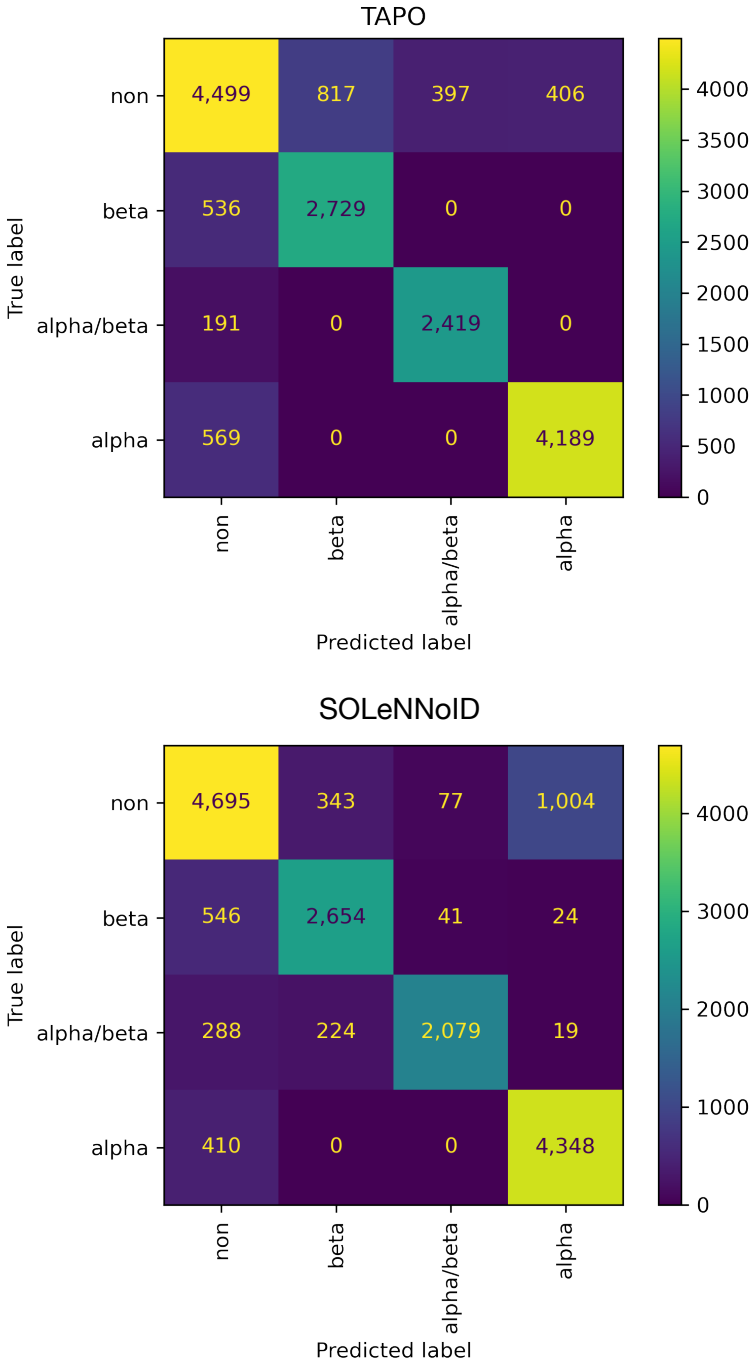


Fig. 8. Multi-class confusion matrices for PRIGSA2, RepeatsDB-Lite, TAPO and SOLeNNoID on the solenoid testing set, continued.

2.3.2. BENCHMARKING ON THE FULL TESTING DATASET

As shown in Table 6, SOLeNNoID also performs well on the full testing set, which includes non-solenoid structures. Precision, recall and F1-score values for solenoid classes were nearly identical to the values on the solenoid testing set only apart from precision on alpha-solenoid residues, which is now lower on introduction of non-solenoid structures. This can also be seen in the top right corner of the confusion matrix in Fig. 9. Comparison with the confusion matrix on the solenoid testing set alone indicates that the model incorrectly classifies an additional 256 alpha-solenoid residues as non-solenoid.

Class	Precision	Recall	F1-score
Non-solenoid	0.93	0.91	0.92
Beta-solenoid	0.82	0.81	0.82
Alpha/beta-solenoid	0.95	0.8	0.86
Alpha-solenoid	0.77	0.91	0.84
MCC			
0.8			

Table 6. Performance of SOLeNNoID on the whole testing set as a multi-class classification problem.

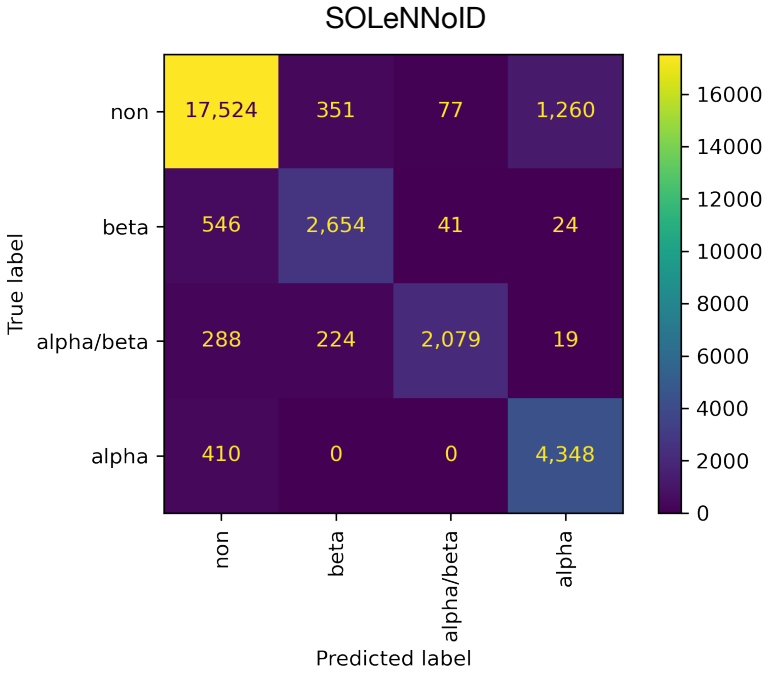


Fig. 9. Multi-class confusion matrix for the SOLeNNoID model on the full testing dataset. The color bar corresponds to the number of observations.

2.3.3. EXAMPLE VISUALISATION

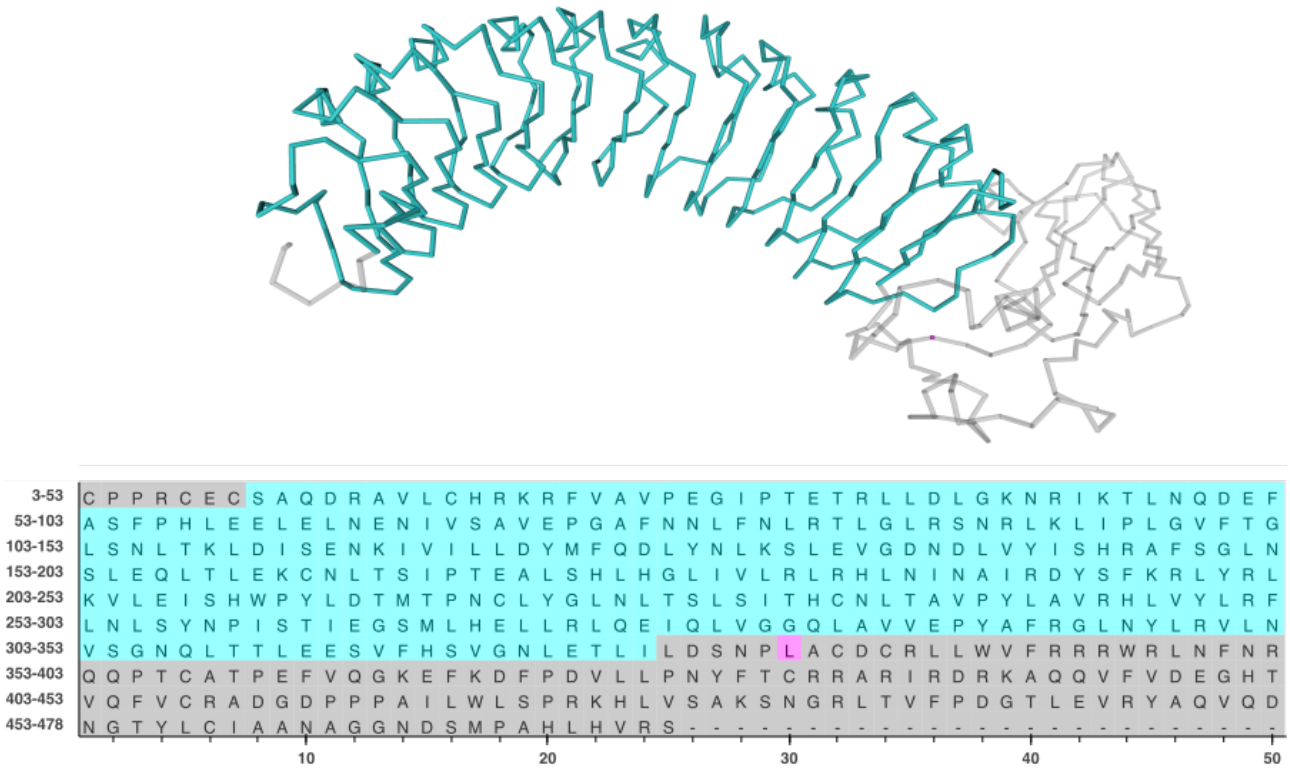


Fig. 10. Example output for PDB ID 4OQT produced by SOLeNNoID.
The structure is shown above in ribbon representation using the nglview viewer. The sequence is shown below using the Bokeh library. Non-solenoid residues are coloured grey, alpha/beta-solenoid residues are coloured cyan and beta-solenoid residues are coloured magenta.

The visualisation produced by SOLeNNoID includes a 3D structure view and a sequence view coloured according to solenoid class. This enables an intuitive understanding of the detected solenoid residues on both the sequence and structure level. In addition, the information can be downloaded in the form of a csv file, containing PDB residue numbering, amino acid identity and solenoid class label. An example is shown for PDB ID 4OQT (Pepinsky et al., 2014) in Fig. 10. A continuous alpha/beta-solenoid region is detected between residues 10 and 326 and residue 332 is incorrectly predicted as beta-solenoid.

2.3.4. PROCESSING LARGE STRUCTURES

To evaluate the time taken by SOLeNNoID to process a structure and make predictions, a test was conducted on the 6,040 v1 AlphaFoldDB predicted protein structures for *S. cerevisiae*. The result in Fig. 11 below shows that prediction time had a second order polynomial relationship with respect to the length of a protein. However, predictions for structures around 2500 amino acids

took less than a minute under the conditions used, whereas predictions for structures up to around 1200 amino acids took less than 10 seconds. This demonstrates that SOLeNNoID can rapidly produce predictions even for large structures. In comparison, the TAPO, PRIGSA2 and RepeatsDB-Lite servers can take on the order of minutes to produce a prediction for a structure over 1000 amino acids. A single outlier was also observed in this test. The outlier protein was alcohol dehydrogenase 5 (UniProt ID: P38113), which comprises 351 residues and SOLeNNoID took 16.7s to complete the prediction on this protein. However, a subsequent attempt to predict solenoid residues for this protein structure took less than a second, suggesting that the result in Fig.11 is an artefact of that particular HPC job.

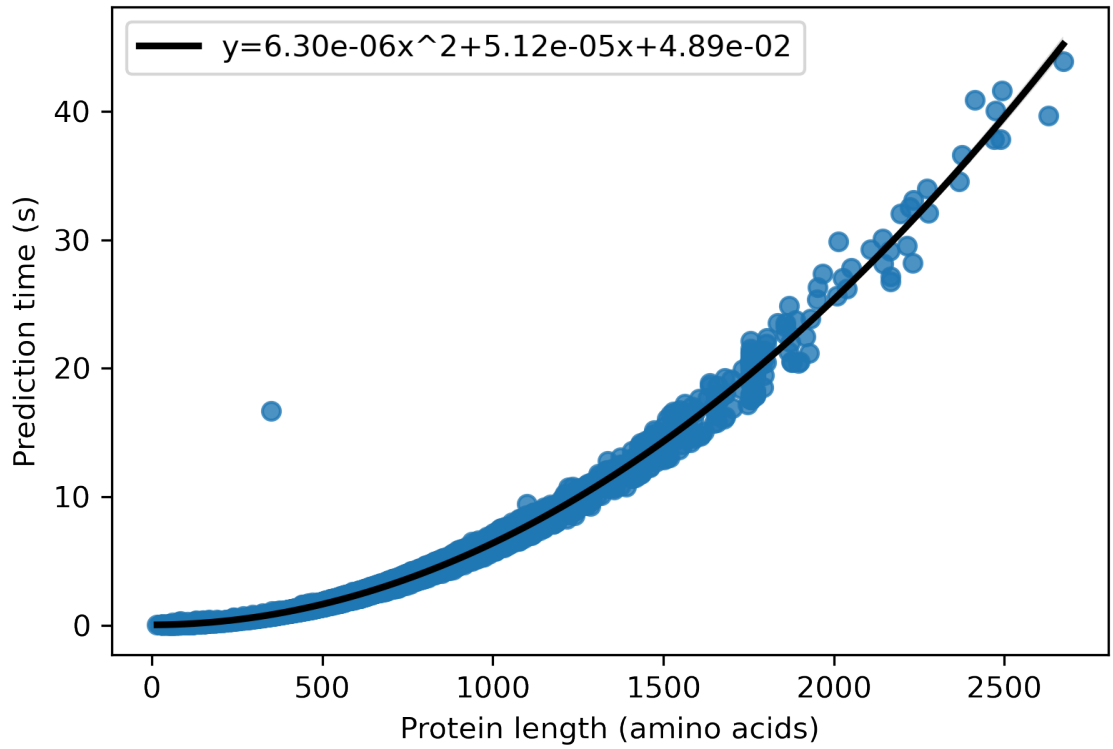


Fig. 11. Relationship between protein length and prediction time for SOLeNNoID method on the YEAST v1 AlphaFoldDB dataset.
Top left - second order polynomial equation fit to the data using numpy.polyfit.

As a further test of SOLeNNoID on large structures, 10 solenoid structures with more than 1000 amino acids were selected from the PDB and the four methods were compared for their ability to process and find solenoid residues in large structures. Only SOLeNNoID produced solenoid residue predictions for all 10 structures as shown in Table 7, demonstrating that SOLeNNoID can find large solenoid proteins, where other methods may fail or only have partial coverage.

PDB ID and Chain	Reference	Class	Number of residues	Method			
				TAPO	PRIGSA2	RDB-Lite	SOLeNNoID
3a6pA	(Okada et al., 2009)	Alpha	1,204	✗	✗	✗	✓
3j9pA	(Paulsen et al., 2015)	Alpha	1,528	✓	✓	✓	✓
3jbzA	(Lau et al., 2016)	Alpha	1,165	✓	✓	✓	✓
3syjA	(Meng et al., 2011)	Beta	1,011	✓	✓	✓	✓
1wxrA	(Otto et al., 2005)	Beta	1,048	✗	✓	✗	✓
7sudA	(Liu et al., 2022)	Alpha	4,128	✗	✗	✗	✓
7pe8A	(Wälchli et al., 2021)	Alpha	2,571	✗	✗	✗	✓
7ptxA	(Chen et al., 2021)	Alpha	1,253	✗	✗	✓	✓
7jluA	(Martin et al., 2020)	Alpha/ beta	1,328	✓	✗	✓	✓
7dxjA	(Huang et al., 2021)	Alpha	3,167	✗	✗	✗	✓

Table 7. Comparison of the ability of the four methods to process structures with more than 1000 residues.
Number of residues is equal to the sequence length provided in the PDB entry for each chain. A tick mark signifies that the structure has been processed, whereas an X signifies a prediction of no repeats or problem with processing the file

In particular, the structure of chain A of the DNA-PK complex VIII (PDB ID 7SUD (Liu et al., 2022)) offers an illustrative case study. This chain has more than 4000 amino acids and resulted in a processing error by PRIGSA2 and predictions that the chain does not contain repeats by TAPO and RepeatsDB-Lite. Notably, the RepeatsDB-Lite server took longer than two hours to return its prediction on this chain. In comparison, SOLeNNoID correctly predicted that the majority of residues in the protein were alpha-solenoid in about a minute (Fig. 12).

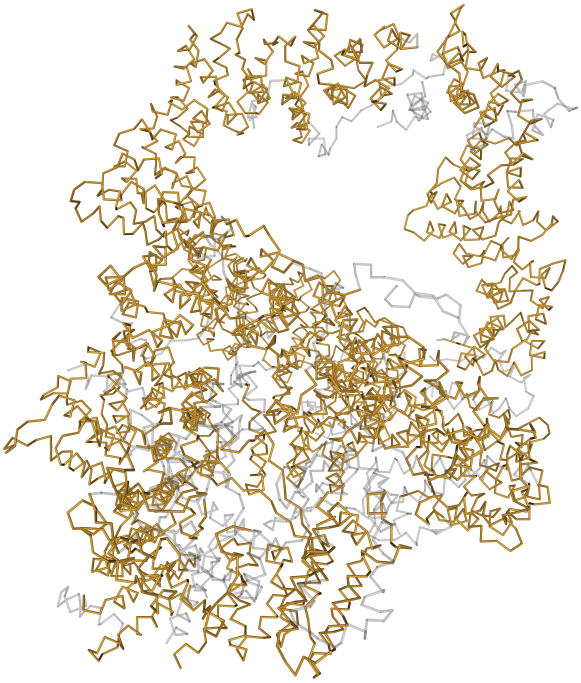


Fig. 12. SOLeNNoID alpha-solenoid predictions on chain A of PDB ID 7SUD.
The protein chain is shown in ribbon representation. Non-solenoid residues are coloured grey and alpha-solenoid residues (2997 predicted) are coloured gold.

2.3.5. ADDITION OF EXTRA CLASSES

Additional labelled data for beta propellers and beta sandwich repeat proteins were obtained from RepeatsDB. These tandem repeat proteins are representatives of the closed (Type IV according to Kajava's classification scheme) and beads-on-a-string (Type V) classes, respectively. Although the focus of SOLeNNoID is on detecting solenoid proteins, I trained the model on the original dataset and these additional data to assess whether SOLeNNoID can be extended to further, non-solenoid classes of repeat proteins. As shown in Fig. 13, once trained on the additional data, SOLeNNoID can successfully detect beta-propeller and beta-sandwich residues. For the beta-propeller structure (PDB ID 4CI8 (Richards et al., 2014)) some residues were incorrectly predicted as beta-sandwich, suggesting that further refinement of the labels might be required to achieve optimal prediction performance. In addition, there is some preliminary evidence that the model can generalise beyond these structures as residues in repeats recorded as alpha/beta-sandwich in RepeatsDB were also classified as beta-sandwich by SOLeNNoID. However, this might be due to the relatively small structural differences between these two classes of repeats.

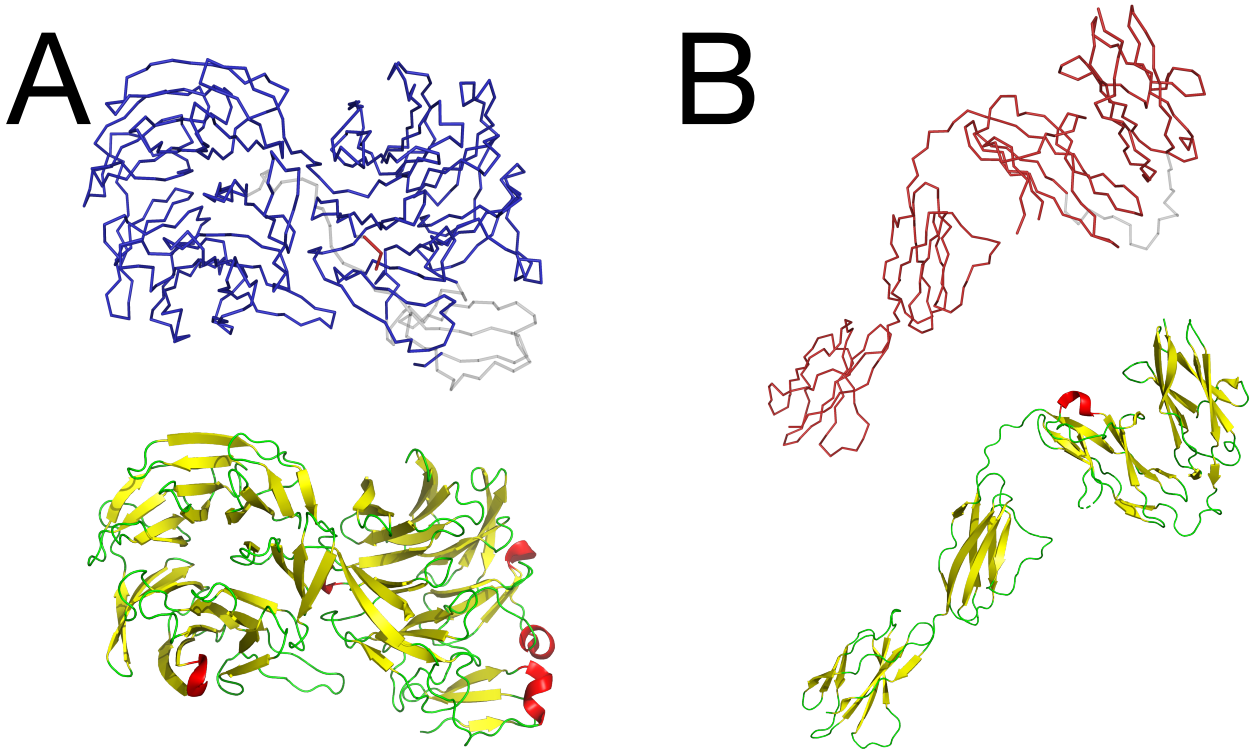


Fig. 13. SOLeNNoID class predictions for additional repeat protein classes.
Top: SOLeNNoID class predictions. Blue - residues predicted as beta-propeller; Red - residues predicted as beta-sandwich repeat. Bottom: secondary structure annotation. Red - alpha-helix, green - loop/turn, yellow - beta-strand. A) SOLeNNoID predictions and secondary structure for PDB ID 4CI8 (Richards et al., 2014), B) SOLeNNoID predictions and secondary structure for PDB ID 5XWT (Goto-Ito et al., 2018).

2.3.6. DETECTING MULTIPLE TANDEM REPEAT CLASSES IN A SINGLE CHAIN

As shown in Fig. 14 below, the SOLeNNoID model trained on solenoid as well as beta-propeller and beta-sandwich data could detect multiple tandem repeat classes in the same chain for two different proteins. In comparison, RepeatsDB-Lite only predicted beta-solenoid repeats for chain A of 3NB2 (Lin et al., 2011) but correctly predicted both beta-propeller and alpha-solenoid repeats for chain A of 6U62 (Rogala et al., 2019). PRIGSA2 predicted no repeats for 3NB2 and could not process 6U62. TAPO only predictsed alpha-solenoid repeats for 3NB2 and no repeats for 6U62. However, there were a few incorrect SOLeNNoID predictions, most evidently for PDB ID 6U62. Some residues in the beta-propeller domain were detected as beta-sandwich and a helix in a non-repeat-containing domain was detected as beta-sandwich.

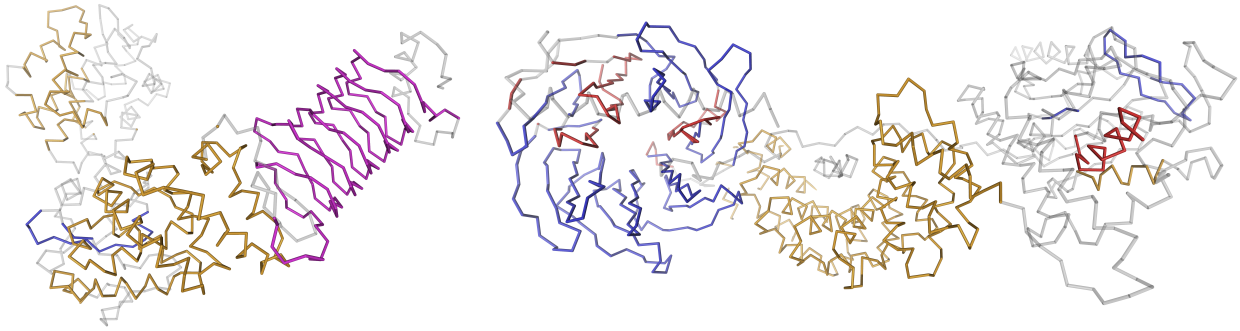


Fig. 14. Multi-class predictions using the SOLeNNoID trained on extra classes.

Left - SOLeNNoID predictions for PDB ID 3NB2 chain A. Two main distinct tandem repeat classes are detected - alpha-solenoid (orange) and beta-solenoid (magenta). Some residues are also detected as beta-propeller (blue). Right - SOLeNNoID predictions for PDB ID 6U62 chain A. Two main distinct tandem repeat classes are detected - alpha-solenoid (orange) and beta-propeller (blue). Some residues are also detected as beta-sandwich repeat (red).

2.3.7. REPEAT SPLITTING FUNCTIONALITY

Although SOLeNNoID does not produce repeat definitions, it is possible to quickly obtain this information through user input. By selecting equivalent residues in adjacent repeats in the interactive nglview viewer, a user can set reference indices (Fig. 15 A, green sticks), which are then used to automatically generate repeats, drawing from SOLeNNoID solenoid predictions (Fig. 15 A, magenta ribbon). The resulting repeats (Fig. 15 B) are then output as individual PDB files and processed using mTM-align and logomaker to produce informative visualisations of repeat similarity on the structure (structural similarity matrices with TM-score (Fig. 15 D) or RMSD (Fig. 15 E) as the criterion) and sequence (logo plot) (Fig. 15 F) levels. A similar output is provided for structures in the RepeatsDB database.

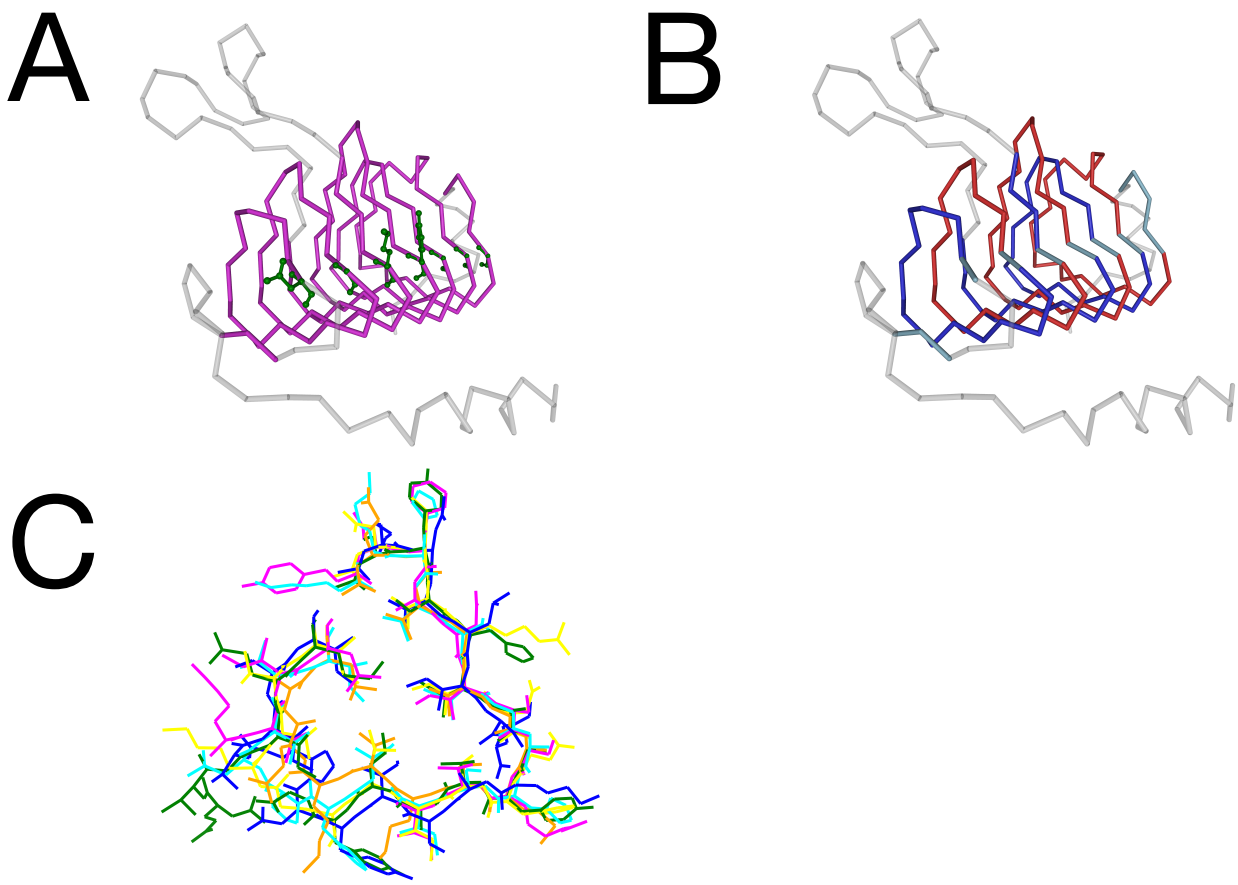
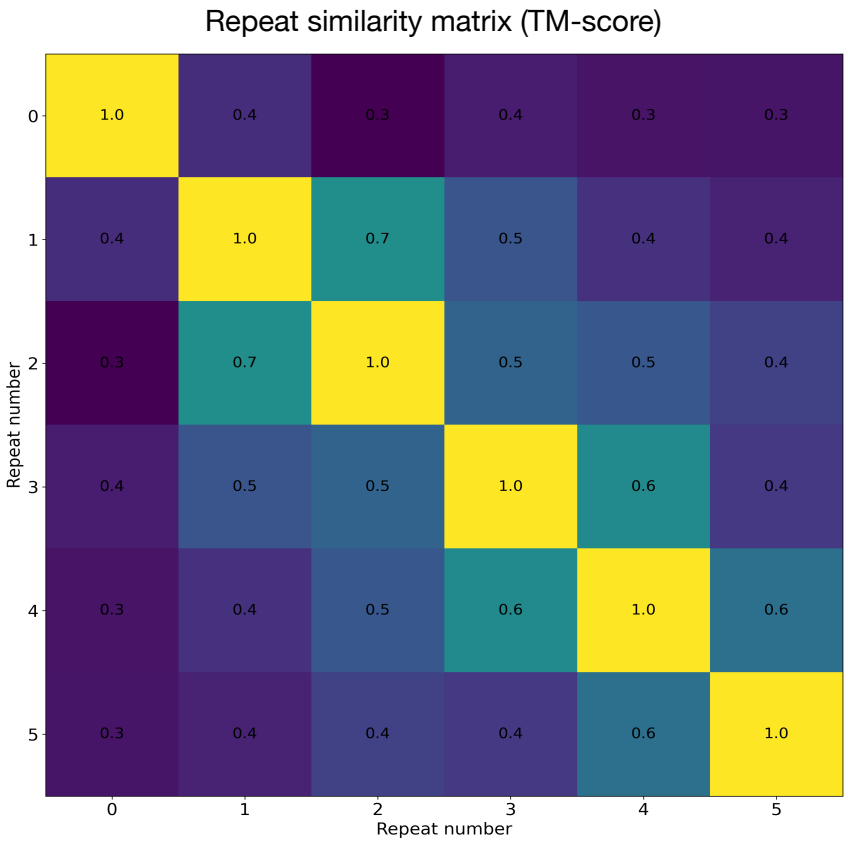


Fig. 15. Repeat splitting processing/output for PDB ID 3VBP chain E.
A) SOLeNNoID beta-solenoid predictions in magenta backbone representation and user-selected equivalent residues in each turn in green ball and stick representation. B) Automatically defined repeats based on the user input of equivalent residues in the SOLeNNoID predictions. Alternating repeats are shown in blue and red for ease of view. C) Superposition of all repeats defined from SOLeNNoID predictions and user input in line representation. Each repeat is given a different colour. D) mTM-align output repeat similarity matrix using TM-score. E) mTM-align output repeat similarity matrix using RMSD. F) Sequence logo for the repeats using the fasta file produced by mTM-align (generated using the logomaker library). Counts in the Y-axis correspond to the number of repeats of 3VBP chain E in which a residue is observed at a given position. For example, at position 0, Gly is observed in 3 repeats, while Tyr and Lys are observed in 1 repeat each.

D



E

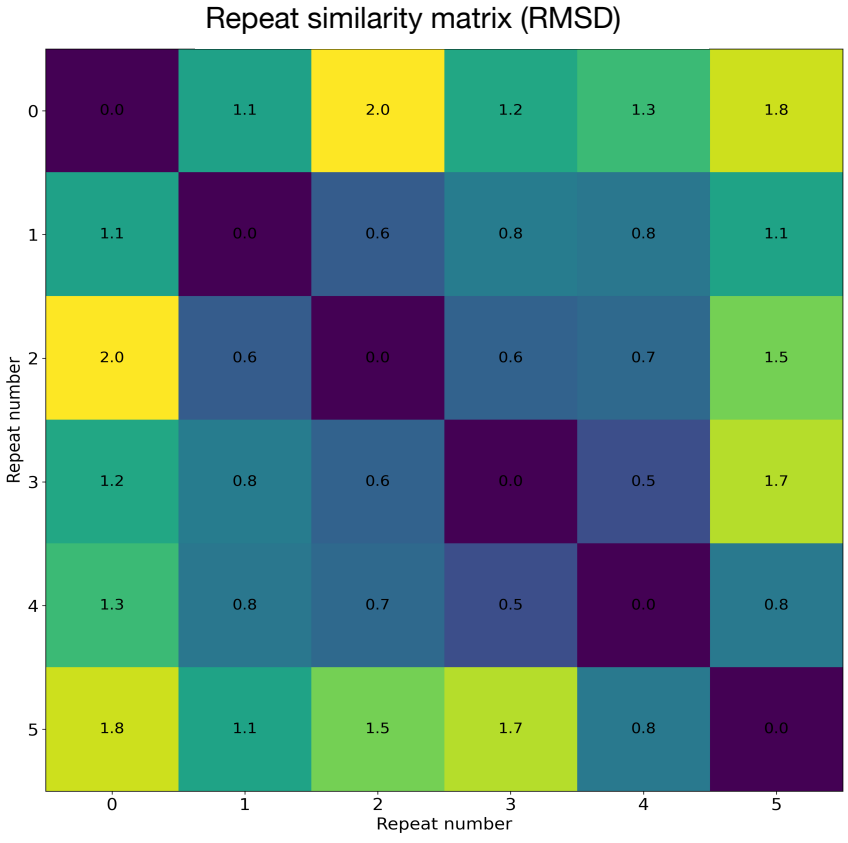


Fig.15, continued.

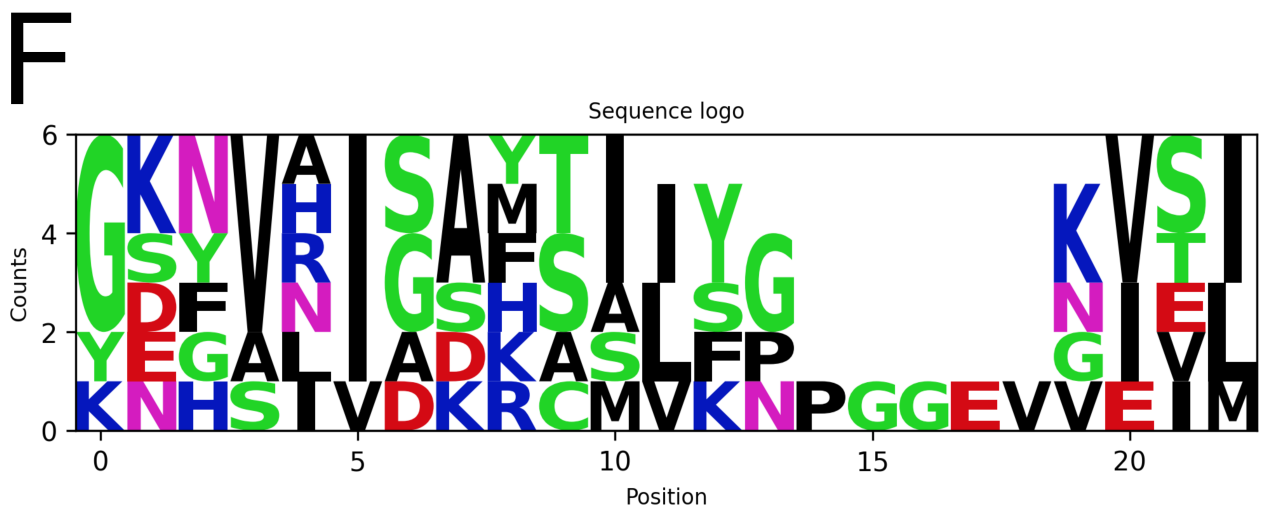


Fig.15, continued.

For example, for protein 3VBP (Kubiak and Holden, 2012) the repeat similarity matrices indicate that repeats 1-4 form a group with high similarity between members. The sequence alignment indicates that the PGGEV sequence is an insertion found only in one of the repeats. Consensus residues are apparent at some of the remaining positions such as a Gly at position 0 and Ile at position 5.

2.4. DISCUSSION

SOLeNNoID differs from other methods in that it highlights individual solenoid residues rather than repeats. Testing set benchmark data shows that SOLeNNoID is competitive with TAPO and outperforms RepeatsDB-Lite and PRIGSA2. Furthermore, SOLeNNoID has balanced precision and recall. Therefore, it can be used to both annotate solenoid regions and find new solenoid proteins, which may be omitted if either RepeatsDB-Lite or PRIGSA2 are used. PRIGSA2 could miss true solenoid proteins due to absent or inconsistent secondary structure assignment, which is used in its repeat identification and validation modules. In addition, PRIGSA2 uses a contact map/network representation of protein structure, as well as comparison with a consensus profile for known repeat types, which would bias the structures found by the method towards more 'ideal' tandem repeats. RepeatsDB-Lite depends heavily on the quality of the repeating unit library it uses to structurally align against a query protein, as well as the 'Master' unit, which defines to which class the repeating region belongs. Some incorrect repeat assignments by RepeatsDB-Lite are likely due to incorrect 'Master' unit assignments - such as alpha-solenoid proteins detected as alpha-barrels. TAPO, in contrast has a balanced performance as it uses several structural features in a SVM model. A key limitation of SOLeNNoID is that it is dependent on the quality of the training dataset and solenoid proteins far outside the structural space covered by this dataset may be incorrectly predicted or omitted. All methods show certain biases on the testing dataset presented here, whether predicting solenoid residues as non-solenoid, or vice versa, or confusing solenoid classes. This highlights the need for a consensus approach to solenoid region prediction by using the outputs of several different repeat detection methods.

SOLeNNoID rapidly outputs predictions in less than 10 seconds for a structure of 1000 residues. Structures of arbitrary size can be processed, and SOLeNNoID could find solenoid proteins inaccessible to other methods due to size constraints. This is illustrated using PDB ID 7SUD as an example. The protein is very large and clearly contains multiple alpha-solenoid residues, which were not detected by the other approaches. SOLeNNoID could therefore be useful for processing large structures such as those produced by cryo-EM or prediction methods such as AlphaFold2 and RoseTTAFold as discussed in the next chapter.

Initial data also shows that SOLeNNoID can be expanded to other structural classes such as closed repeats and beads-on-a-string. Therefore, it might be possible to extend the method to all tandem repeat types of the elongated, closed and beads-on-a-string superclasses or fold classes within these superclasses with more data collection and cleaning. SOLeNNoID can also accommodate multi-class repeat proteins, which are less frequent but can provide useful information about how residues from different repeat classes interact and connect to one another within a single chain.

The output from SOLeNNoID can be used in downstream applications - e.g. parametric analysis of the solenoid region(s), or to find positions which support loop insertions into the structure. In addition, it is possible to conduct user-defined repeat annotation with the output from SOLeNNoID, producing useful information about repeat structural and sequence similarity to complement that found in the RepeatsDB database. The output from the command-line version of mTM-align could also be used to determine the transformation matrices between adjacent repeats, which could be useful for the design of new solenoid proteins. In addition, SOLeNNoID could be used as a filter for ‘solenoid-ness’ when designing novel proteins.

3.
SOLENOID PROTEINS IN PROTEIN STRUCTURE DATABASES

3.1. AIMS

The wealth of protein structural information provides fertile ground for bioinformatic analysis, such as structural and functional annotation. My aim is to use my solenoid detection/annotation method (SOLeNNoID) described in **Chapter 2** to search the PDB and the the proteomes of all 20 species included in version 1 of the AlphaFold database for alpha-, alpha/beta- and beta-solenoid proteins. While the PDB has previously been searched for tandem repeat proteins by the authors of RepeatsDB and DbStRiPs, to my knowledge, this is the first attempt at mining for solenoid proteins in the AlphaFold database.

In this chapter, I present the results of these database searches, including a breakdown of solenoid protein frequency and size across AlphaFold database proteomes as well as a discussion of models of solenoid proteins with no experimentally determined structures. In addition, I demonstrate that SOLeNNoID can find a set of solenoid structures in the PDB that only partially overlaps with those in the RepeatsDB and DbStRiPs databases, as well as presenting the most frequent functions across solenoid classes and a species-wide distribution of solenoid protein entries detected by SOLeNNoID.

3.2. MATERIALS AND METHODS

3.2.1. ALPHAFOLD DB SEARCH

Version 1 of AlphaFoldDB was used in .pdb format (Varadi et al., 2022). Processing to a distance matrix and prediction were carried out as described in the previous chapter. All structures for all 21 species were processed. The data collected for each structure included the ID, name metadata, the number of residues for the structure and the number, percentage, and positions of solenoid predictions for all classes. Subsequently a mean pLDDT value was calculated for the lists of residues predicted to belong to each solenoid class (e.g. mean pLDDT for beta solenoids, mean pLDDT for alpha solenoids etc). Finally, structures for analysis were selected if they had over 50% solenoid residues of a given class and if the mean pLDDT for the respective class was above 70. The solenoid percentage cutoff was chosen to decrease spurious solenoid predictions, as many structures incorrectly predicted as solenoid had solenoid percentages below 50%. The 50% figure was also chosen to bias the search towards proteins where the solenoid domain accounts for the majority of the residues in the overall structure. The pLDDT cutoff was selected as residues with pLDDT above 70 generally have a correct backbone and correspond to ‘confident’ predictions according to the AlphaFold scheme.

3.2.2. PDB SEARCH

The PDB as of 30.06.2021 was used with files in .mmcif format. This was done as .pdb files are not available for all structures (.pdb files are limited to a maximum of 99,999 atoms). Predictions were attempted on all chains using the method described in the previous chapter. Not all chains were successfully processed, and data was collected on structures, where processing failed, to inform future enhancements to the protocol. The data collected for each structure included the 4-letter PDB ID, chain ID, name, deposition date, structure method and resolution metadata, the number of residues for the structure and the number, percentage, and positions of solenoid predictions for all classes. PDB IDs from the traning/validation and testing datasets from the previous chapter and the RepeatsDB dataset were removed from the final list of processed structures to limit the analysis to new protein chains. Four groups of proteins were selected: those with over 50% of their residues detected as alpha, alpha/beta or beta solenoid and a final group with under 50% of residues falling within any single solenoid group but 50 or more percent of residues falling into the solenoid category: this is the ‘combined’ group.

All database searches using SOLeNNoID were carried out on the Imperial HPC cluster by splitting the search into a number of jobs.

Data for RepeatsDB solenoid proteins was downloaded from https://www.repeatsdb.org/classification/3. Subsequently, only entries with a ‘Classification’ tab beginning with ‘3.’ were selected. The latest version of RepeatsDB was populated from the PDB at 2020-08-14. Data for DbStRiPs was downloaded from https://bioinf.iiit.ac.in/dbstrips/data.php, selecting Type ‘Structural class’ and Sub-type ‘Alpha Solenoids’, ‘Beta Solenoids’ and ‘Alpha-Beta Solenoids’. The latest version of DbStRiPs was populated from the PDB at 2019-06-03.

3.2.3. DATA PROCESSING AND VISUALISATION

Data analysis was carried out using the Pandas library (Mckinney, 2010) in Python3 (van Rossum & Drake, 2009). Data visualisation was carried out using the Seaborn (Waskom, 2021), matplotlib (Hunter, 2007), matplotlib-venn (https://github.com/konstantint/matplotlib-venn) and logomaker libraries. Protein structure visualisation was carried out using the nglview library in Python3 and the PyMOL molecular graphics software (Schrödinger LLC, 2015).

3.3. RESULTS

3.3.1. ALPHAFOLD DATABASE SEARCH

3.3.1.1. SOLENOID PROTEIN STATISTICS IN THE ALPHAFOLD DATABASE

	Alpha-solenoid	Alpha/beta-solenoid	Beta-solenoid
Total Number	21278	1527	653
Unique 'names'	9341	650	194
Uncharacterised	3814	380	152
Min number of solenoid residues	43	29	35
Average number of solenoid residues	509	445	341
Max number of solenoid residues	2615	1891	1053

Table 8. Statistics for AFDB entries detected as solenoids by SOLeNNoID.

All 365,198 structures in the AlphaFold database were processed. The number of structures per species range from a minimum of 2,888 for *S. aureus* to a maximum of 55,799 for *G. max* (soya). The total number of structures identified as solenoid-containing by SOLeNNoID according to the criteria outlined in the Section 3.2.1 were 23,458, representing approximately 6.4% of all entries. Using the 'name' section of the header of the .pdb files, the unique annotations for proteins detected as solenoid were 10,185, while solenoid entries with 'uncharacterised' in their 'name' annotation were 4,346. As shown in Table 8, alpha-solenoids were the most abundant in the database, followed by alpha/beta-solenoids and finally beta-solenoids. Alpha-solenoid protein regions were also on average larger than those of other solenoid types. On manual inspection of structures classified as solenoid by SOLeNNoID, a number of non-solenoid helical structures were found, such as long alpha-helices and coiled coil structures. Therefore, the large number of alpha-solenoids found by SOLeNNoID was likely an overestimate.

3.3.1.2. SOLENOID FRACTIONS OF THE ALPHAFOLD DATABASE PROTEOMES

3.3.1.2.1. ALPHA-SOLENOID PROTEINS

An important insight we can derive from the AlphaFold database based on my selection criteria is the percentage of solenoid structures for each proteome. As seen in Fig. 16 A, putative alpha-solenoids made up between 1 and 9 percent of the proteome, depending on the species. The bacteria and archaea in the database had a roughly 2 to 4 times lower percentage of alpha-

solenoids in their proteomes, compared to the eukaryotic organisms. This observation was supported by analysis by Fournier et al (Fournier et al., 2013). As discussed by Fournier et al, the increased complexity of eukaryotic cells compared to prokaryotic cells both necessitates and enables the synthesis and folding of larger, more extended proteins, capable for forming interactions with other proteins and nucleic acids. My finding that frequency of alpha-solenoids across eukaryotes was relatively similar also coincided with the findings of Fournier et al. In eukaryotic groups, the highest percentages of putative alpha-solenoids were observed in *H. sapiens* and *S. pombe*, while plant alpha-solenoid percentages were generally lower than those in animals. However, it is unclear whether these differences are biologically significant.

3.3.1.2.2. ALPHA/BETA-SOLENOID PROTEINS

Alpha/beta-solenoids were almost absent from the proteomes of the bacteria/archaea in the database but present in all eukaryotic species. Based on my analysis, alpha/beta-solenoid proteins made up a roughly 10 times smaller percentage of the proteome compared to alpha-solenoid proteins. Single-cell eukaryotes on average had fewer alpha/beta-solenoids compared to animals and plants. Soanes and Talbot (Soanes & Talbot, 2010) found that species of the kingdom Fungi do not possess LRR pattern recognition receptors, which are involved in the innate immune response in animals and plants. This could partly account for the lower frequency of alpha/beta-solenoid proteins in the *S. cerevisiae* and *S. pombe* proteomes compared to animals and plants. *C. albicans* had a higher percentage of alpha/beta-solenoid proteins but the majority were either uncharacterised or of unknown function. It is possible that the larger number of leucine-rich repeats in *C. albicans* compared to *S. cerevisiae* is linked to its pathogenicity. Plants had a higher percentage of alpha/beta-solenoid proteins than animals. Many of these proteins belong to the disease resistance protein (Belkhadir et al., 2004) (and the receptor-like protein subgroup thereof (Kruijt et al., 2005)), leucine-rich repeat receptor kinase (LRR-RK) (Chakraborty et al., 2019) and receptor-like kinase (Shiu & Bleecker, 2001) groups. Taken together, these LRR-containing proteins fulfil a variety of roles in plant growth and development, environmental interactions, wound healing, immune defense and others. *Dictyostelium discoideum* was the outlier in terms of alpha/beta-solenoid percentage. 109 out of the 118 detected alpha/beta-solenoid proteins in *D. discoideum* were uncharacterised. Altae-Tran et al (Altae-Tran et al., 2020) identified 90 LRR proteins in *Dictyostelium purpureum,* which belong to the FNIP/IP22 repeat family (O'Day et al., 2006) and suggest that these proteins could have a role in nucleic acid binding. A number of the uncharacterised proteins detected by SOLeNNoID in *D.discoideum* belong to the FNIP family, according to the Pfam database. Additionally, examining experimentally characterised LRR proteins in *Dictyostelium* might shed light on possible functions of these uncharacterised AlphaFold DB entries. LrrkA (Bodinier et al., 2021) is a kinase with leucine-rich repeats, involved in regulating phagocytosis and cell motility in *D. discoideum*. MPL1 is a phosphatase with LRRs and is involved in cell motility during development (Rodriguez et al., 2008). Another LRR protein, Leep1 (Y. Yang et al., 2021), is a polarity regulator, involved in pseudopod and macropinocytic cup

dynamics. GbpC is a multidomain protein, involved in cGMP-mediated chemotaxis (van Egmond et al., 2008). Therefore, a number of the uncharacterised *Dictyostelium* LRR proteins might also be involved in regulating cell motility/polarity. However, further research is required to experimentally determine the functions of these proteins.

3.3.1.2.3. BETA-SOLENOID PROTEINS

Beta-solenoids comprised a smaller percentage of the proteome in single-cell eukaryotes, animals and plants compared to alpha/beta-solenoids and alpha-solenoids, with a maximum percentage of the proteome of 1% in *T. cruzi*. In contrast, in bacteria and archaea, beta-solenoids were more frequent than the nearly absent alpha/beta-solenoids. Animals and single-cell eukaryotes generally had a lower percentage of beta-solenoids compared to plants. This is due to the abundance of pectate lyase (Uluisik & Seymour, 2020), pectin methylesterase (Jolie et al., 2010) and polygalacturonase (F. Wang et al., 2016) enzymes, which have a beta-solenoid structure and are involved in cell wall remodelling during plant developmental and physiological processes as well as fruit ripening. The outlier species in terms of beta-solenoid percentage were *E. coli*, *T. cruzi* and *D. discoideum*. *E. coli* beta-solenoid proteins clustered around a few key functions - there were putative enzymes involved in lipopolysaccharide (LPS) and colanic acid (CA) synthesis, large outer membrane proteins (possibly auto transporters or adhesins) and curlin subunits, which will be discussed in detail in Section 3.3.1.5.4. LPS is part of the outer membrane of *E. coli*, whereas CA is secreted, and both are involved in biofilm formation (C. Wang et al., 2020). Autotransporter proteins are associated with functions related to virulence such as adhesion between cells and biofilm formation in Gram-negative bacteria (Wells et al., 2007) and have a complicated folding/translocation mechanism (Yuan et al., 2018). Curlin subunits are the components of curli fibres - part of the *E. coli* extracellular matrix. Detected beta-solenoid proteins for *T. cruzi* were primarily made up of entries from the dgf-1 gene family (180 out of 193 detected proteins), which is discussed in more detail in Section 3.3.1.5.3. 44 out of 54 proteins detected as beta-solenoid in *D. discoideum* were uncharacterised. Searches with Foldseek (Kempen et al., 2022) for entries Q54QS6, Q54RK1 and Q551H1, yielded the HxuA protein from *H. influenzae* involved in binding to haemopexin to release haem (Zambolin et al., 2016), DFA-IIIase from *A. chlorophenolicus* (Yu et al., 2018) and the cyanophage Pam1 tailspike receptor-binding domain (Zhang et al, unpublished), respectively as the closest structural matches from the PDB based on TM-align TM-score. The functions of these *Dictyostelium* proteins need to be experimentally characterised but their structural similarity to bacterial/phage PDB entries suggests that they might be the result of horizontal gene transfer from bacteria (Eichinger et al., 2005).

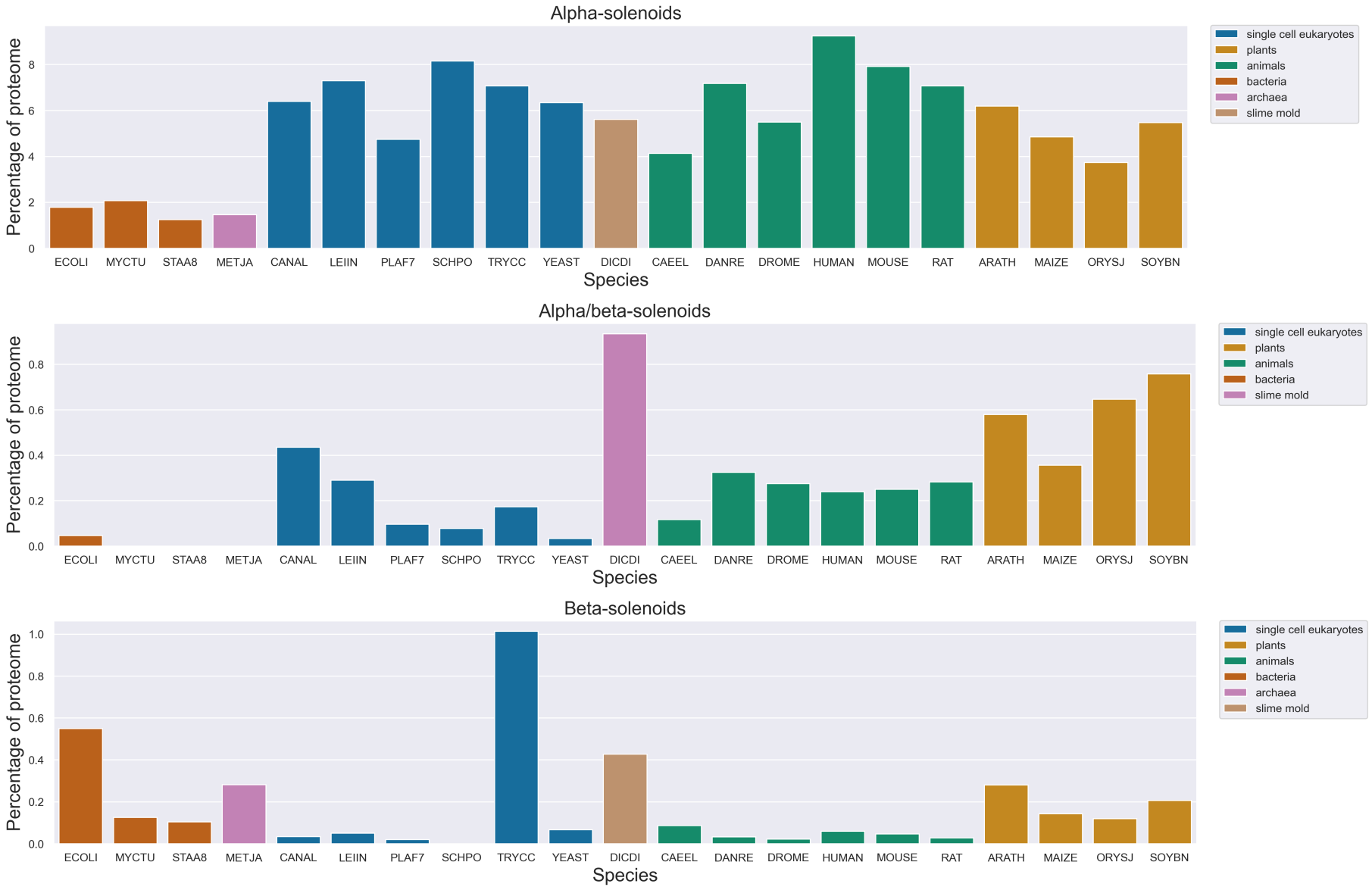


Fig. 16. Percentages of solenoid structures across AlphaFold DB proteomes.
Top - alpha-solenoid percentages, middle - alpha/beta-solenoid percentages, bottom - beta-solenoid percentages. For each plot, the legend in the top right corner indicates the colours of groups of organisms and the basis of the grouping.

3.3.1.3. SOLENOID PROTEIN LENGTHS ACROSS PROTEOMES

3.3.1.3.1. ALPHA-SOLENOID PROTEINS

The median and interquartile range were smaller in the bacteria and archaea in this dataset. The median number of alpha-solenoid residues per protein in prokaryotic species ranged from 174 in *M. jannaschii* to 313 in *M. tuberculosis*. For eukaryotic species the median number of alpha-solenoid residues per protein fell between 363 in *P. falciparum* to 591 in *H. sapiens* - double that of prokaryotes. Interquartile ranges were relatively consistent across eukaryotes at around 400 residues with *D. discoideum* and *H. sapiens* having the broadest and plants having the narrowest.

3.3.1.3.2. ALPHA/BETA-SOLENOID PROTEINS

Alpha/beta-solenoids were present mainly in eukaryotic species. However, the distributions for *P. falciparum*, *S. pombe* and *S. cerevisiae* could not be interpreted due to a low number of observations.The median number of alpha/beta-solenoid residues per protein for all other eukaryotic species was within the 200-500 range. Animal species had a narrower interquartile range of approximately 100-150 residues and lower median values compared to plants, which had interquartile ranges greater than 200 residues. For example, animal median values ranged from 230 in *H. sapiens* to 296 residues in *D. melanogaster*, while plant medians ranged from 277 in *Z. mays* to 425 in *G. max*. Coupled with the finding from the Section 3.3.1.2 of a greater percentage of alpha/beta-solenoids in plants compared to animals, this suggests that alpha/beta-solenoids are more broadly employed by plants compared to animals and display a greater range of sizes, possibly reflecting a broader range of functions. Single cell eukaryotes had a similar IQR compared to animals but higher median values. For example, C.albicans and T.cruzi had median alpha/beta-solenoid residues per protein values of 364 and 301 residues, respectively.

3.3.1.3.3. BETA-SOLENOID PROTEINS

The beta-solenoid residue number IQRs in *E. coli*, *D. discoideum*, *H. sapiens* and *M. musculus* were comparable at approximately 250-300 residues. However, in eukaryotes median values were larger, ranging from 214 in *O. sativa* to 819 in *T. cruzi*, compared to 177 for *E. coli*. Plants had median values between *E. coli* and animals, from 214 in *O. sativa* to 247 in *A. thaliana*. It is notable that plant IQRs were very narrow at less than 100 residues, likely reflecting a strong functional clustering of beta-solenoids in plants, possibly into the functions discussed in the Section 3.3.1.2.2. *T. cruzi* was the outlier, displaying the largest median value (819) and the largest IQR at over 750 residues. Almost all *T. cruzi* entries were from the dgf-1 family, which will be discussed in more detail in Section 3.3.1.5.3. The distribution for *T. cruzi* suggests that dgf-1 proteins are large but also varied in size. There were few to no beta-solenoid proteins for a number of species such as *P. falciparum*, *S. pombe*, *S. cerevisiae*, *C. albicans, S. aureus* and *D. melanogaster*. This suggests that the beta-solenoid fold is less common across domains compared to alpha- and alpha/beta-solenoids.

3.3.1.3.4. SUMMARY

Beta-solenoid residues distributions were more variable across species in comparison to alpha- and alpha/beta-solenoid distributions. The general observation of increased length in eukaryotes compared to prokaryotes is supported by the work of Brocchieri and Karlin (Brocchieri & Karlin, 2005), who found that for most protein families, eukaryotic proteins are larger than their prokaryotic counterparts. Furthermore, Delucchi et al (Delucchi et al., 2020) found that eukaryotes have more tandem repeats and larger proteins compared to prokaryotes.

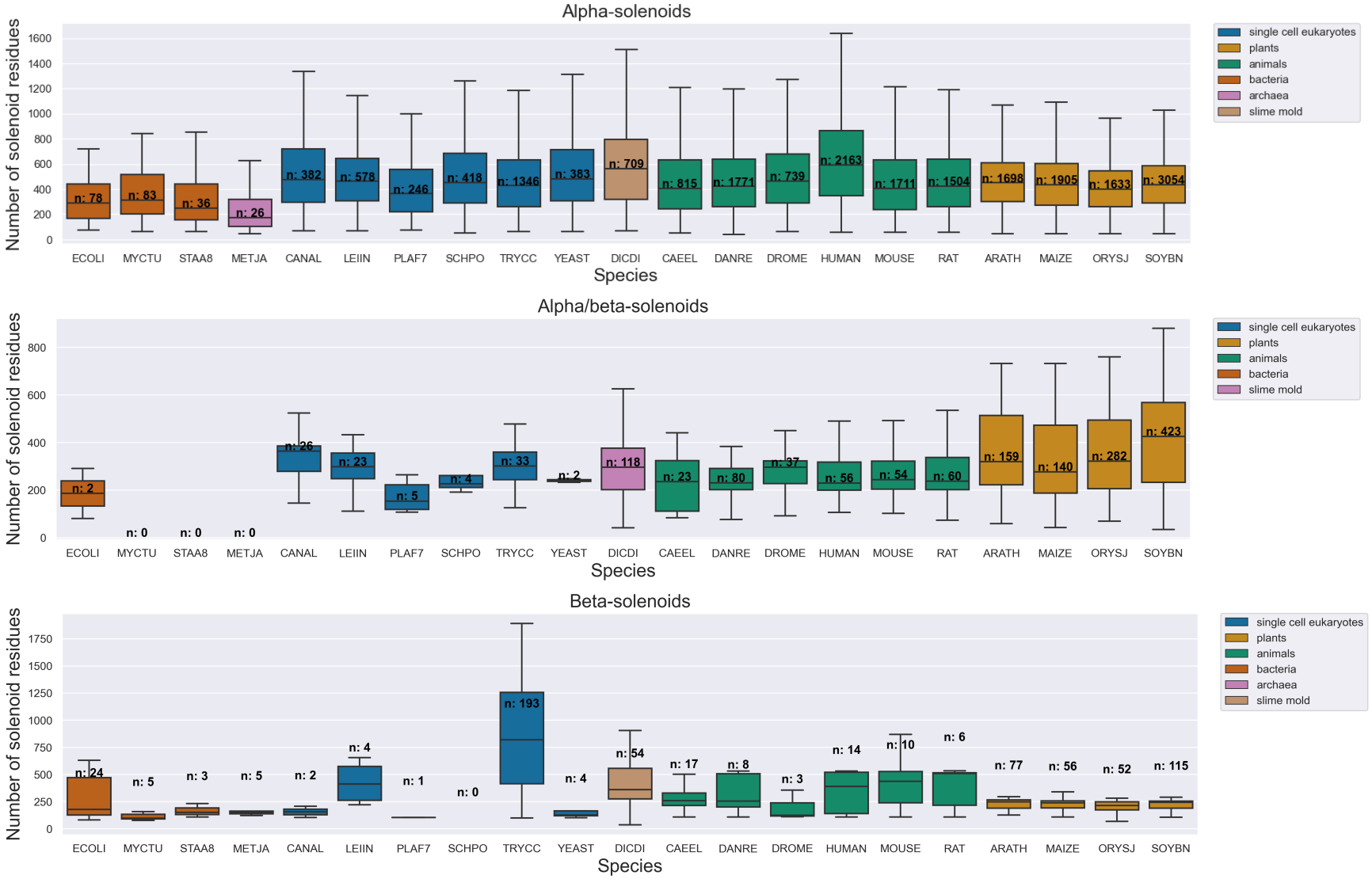


Fig. 17. Distribution of the number of solenoid residues across solenoid classes and AFDB proteomes.

Top - distribution of alpha-solenoid residue numbers, middle - distribution of alpha/beta-solenoid residue numbers, bottom - distribution of beta-solenoid residue numbers. For each plot, the legend in the top right corner indicates the colours of groups of organisms. In each box plot, the central line denotes the median value, the bottom and top line of the box denote the 25th and 75th percentiles, respectively, and the lower and upper whisker denote the minimum and maximum values, respectively. The minimum is defined as: 25th percentile minus 1.5 times IQR, the maximum value is defined as the 75th percentile plus 1.5 times the IQR. The number associated with each box denotes the number of observations.

3.3.1.4. UNCHARACTERISED PROTEINS

Searching the AlphaFold database with SOLeNNoID returned 4,346 structures with 'uncharacterised' in their annotation. The numbers for each solenoid class are shown in Table 8. Most structures (3,814 out of 4,346) are identified as putative alpha-solenoids. There were no proteins with an 'uncharacterised' annotation for *O. sativa*. The lowest number of uncharacterised proteins (3) was found in *M. musculus*, while the highest (1,948) was found in *G. max*. This difference is likely due to the large size of the *G. max* proteome (55,799 proteins). Other species with large numbers of uncharacterised proteins were *T. cruzi* (862) and *D. discoideum* (497). In the case of *G. max*, 87% of uncharacterised proteins were putative alpha-solenoids, in *T. cruzi* 97% of the uncharacterised proteins belonged to the alpha-solenoid class, whereas in *D. discoideum*, putative alpha-solenoids made up 69% of all uncharacterised entries. The species with the greatest number of uncharacterised alpha/beta-solenoid entries were *D. discoideum* and *G. max*. The only species with more than 10 uncharacterised beta-solenoid proteins were *D. discoideum* and *G. max*.

The distributions of the numbers of alpha-solenoid residues across species for uncharacterised species were generally consistent with the observations from section 3.2.1.3. Prokaryotic medians were lower than eukaryotic medians and the highest median values (502 and 629, respectively) and largest interquartile ranges (both over 400 residues) were observed for *L. infantum* and *D. discoideum.*

For most species with uncharacterised alpha/beta-solenoid entries, the median value for the number of alpha/beta-solenoid residues fell within the 200 to 400 residue range, consistent with the observations from section 3.2.1.3. *G. max* had the largest interquartile range at over 200 residues, while *C. albicans* had the highest median value at 374 residues.

Although both *D. discoideum* and *G. max* had many putative uncharacterised beta-solenoid proteins, *G. max* had a smaller median value for the number of beta-solenoid residues (248 residues compared to 360) as well as a much narrower interquartile range than *D. discoideum* (less than 100 compared to 300), consistent with the beta-solenoid distribution of characterised *G. max* entries, discussed in section 3.2.1.3.

The wealth of uncharacterised putative solenoid proteins found by SOLeNNoID could be further investigated using methods such as FoldSeek (Kempen et al., 2022) to find structural homologs of these uncharacterised proteins and therefore also an initial indication of possible function.

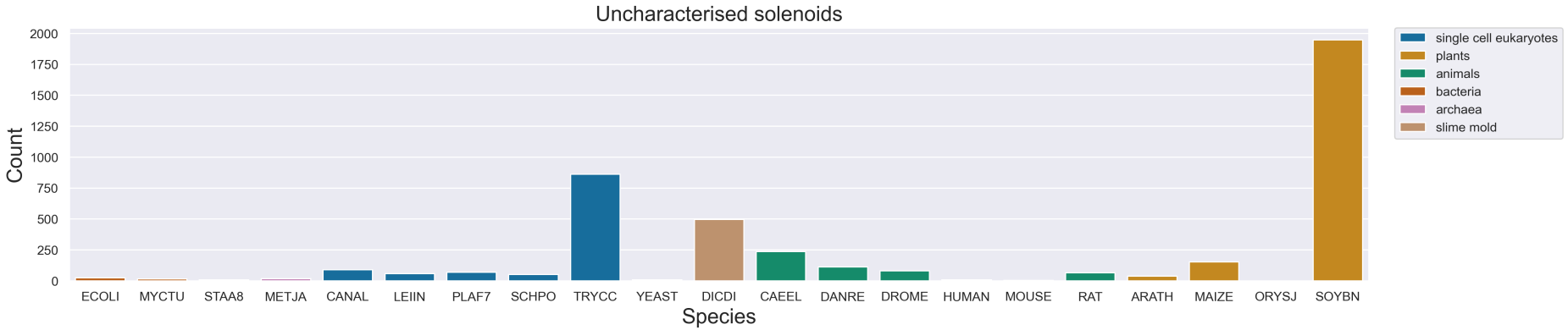


Fig. 18. Number of solenoid proteins containing 'uncharacterised' in their annotation across AFDB proteomes. For each plot, the legend in the top right corner indicates the colours of groups of organisms.

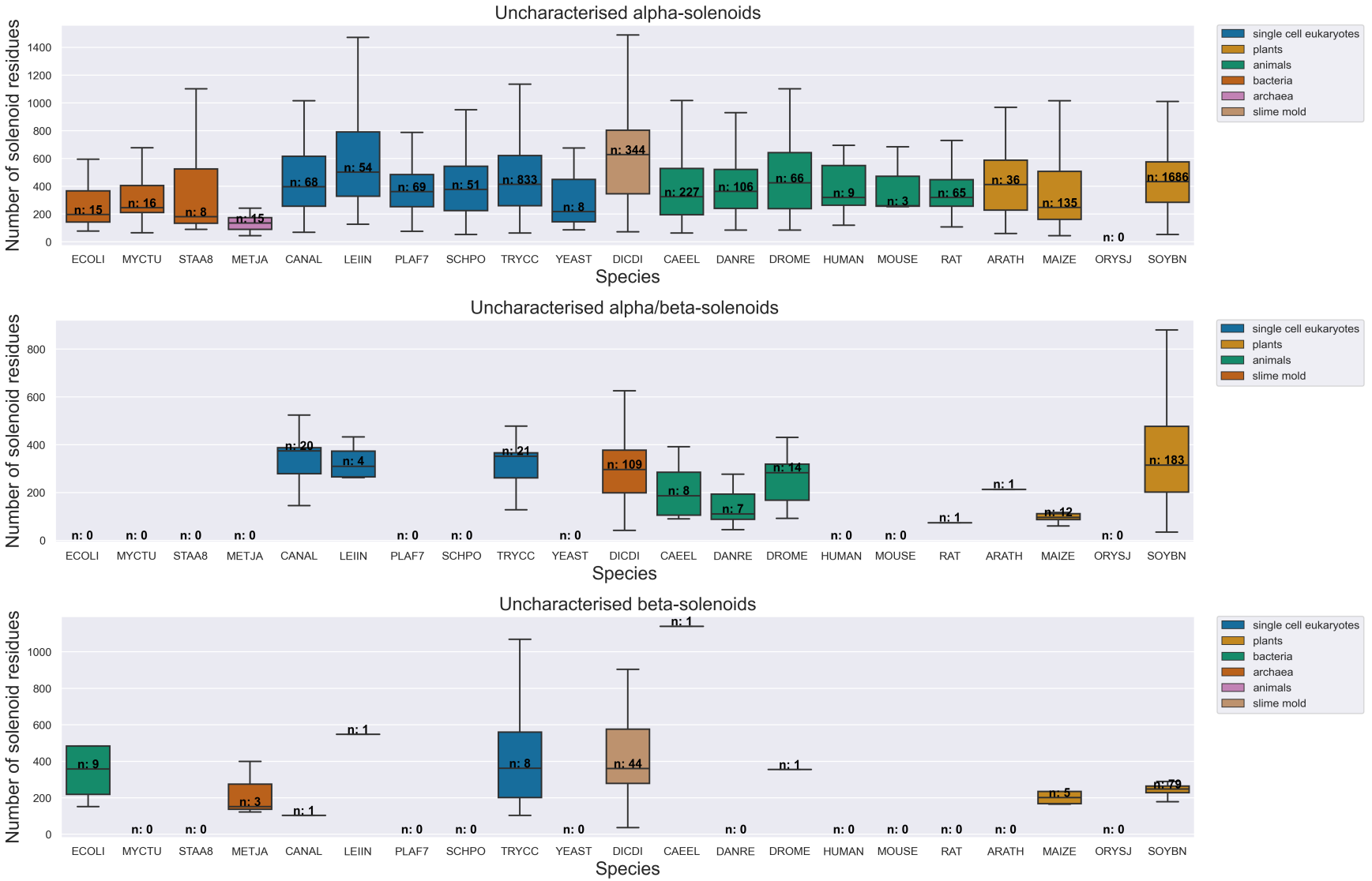


Fig. 19. Distribution of the number of solenoid residues for 'uncharacterised' proteins across solenoid classes and AFDB proteomes.
Top - distribution of alpha-solenoid residue numbers. Middle - distribution of alpha/beta-solenoid residue numbers. Bottom - distribution of beta-solenoid residue numbers. For each plot, the legend in the top right corner indicates the colours of groups of organisms. In each box plot, the central line denotes the median value, the bottom and top line of the box denote the 25th and 75th percentile, respectively, and the lower and upper whisker denote the minimum and maximum values, respectively. The minimum is defined as (25th percentile minus 1.5 times IQR, the maximum value is defined as the 75th percentile plus 1.5 times the IQR. The number associated with each box denotes the number of observations.

3.3.1.5. SOLENOID PROTEINS OF INTEREST WITH NO PDB STRUCTURES

3.3.1.5.1. MESOTHELIN

Mesothelin is a protein overexpressed in several human cancers such as mesotheliomas, pancreatic adenocarcinomas, ovarian cancers, and lung adenocarcinomas (Hassan & Ho, 2008). The 71-kDa precursor protein is processed to the 31 kDa megakaryocytic potentiating factor protein (MPF) and the 40-kDa GPI-anchored mesothelin protein. The only experimentally determined structure in the PDB is that of MORAb-009 (humanised monoclonal antibody against mesothelin) bound to a fragment of mesothelin from residue 7 to residue 64 (Ma et al., 2012). The MORAb-009 antibody blocks mesothelin binding to the MUC16/CA125 protein, which is thought to contribute to cancer metastasis (Rump et al., 2004).

The AlphaFold model of human mesothelin was detected by SOLeNNoID as an alpha-solenoid protein. The model in the AlphaFold database is of the full-length precursor protein and the pAE plot is consistent with two independent domains, corresponding to MPF and mesothelin. Furthermore, the solenoid parts of mesothelin follow an ARM/HEAT repeat fold (Sathyanarayana et al., 2009). A superposition of residues 7-64 of the mesothelin fragment in PDB ID 4F3F (Ma et al., 2012) and residues 302-359 of the AFDB model resulted in an RMSD of <0.5 Å, suggesting that the model successfully recapitulates the structure of that part of mesothelin. MORAb-009 is targeted to Region I of the mesothelin protein, while Y.-F. Zhang et al (Y.-F. Zhang et al., 2015) report antibodies binding to Regions II and III. The AlphaFold model suggestsed further structural features of note: disulfide bonds between residues 302 and 326, and 450 and 476; and a disordered loop spanning residues 403-415 (pLDDT<70). Mesothelin is an attractive cancer immunotherapy target (Hassan et al., 2016). The structural insight provided by the AlphaFold mesothelin model could inform structure-guided designs of anti-mesothelin antibodies or binders.

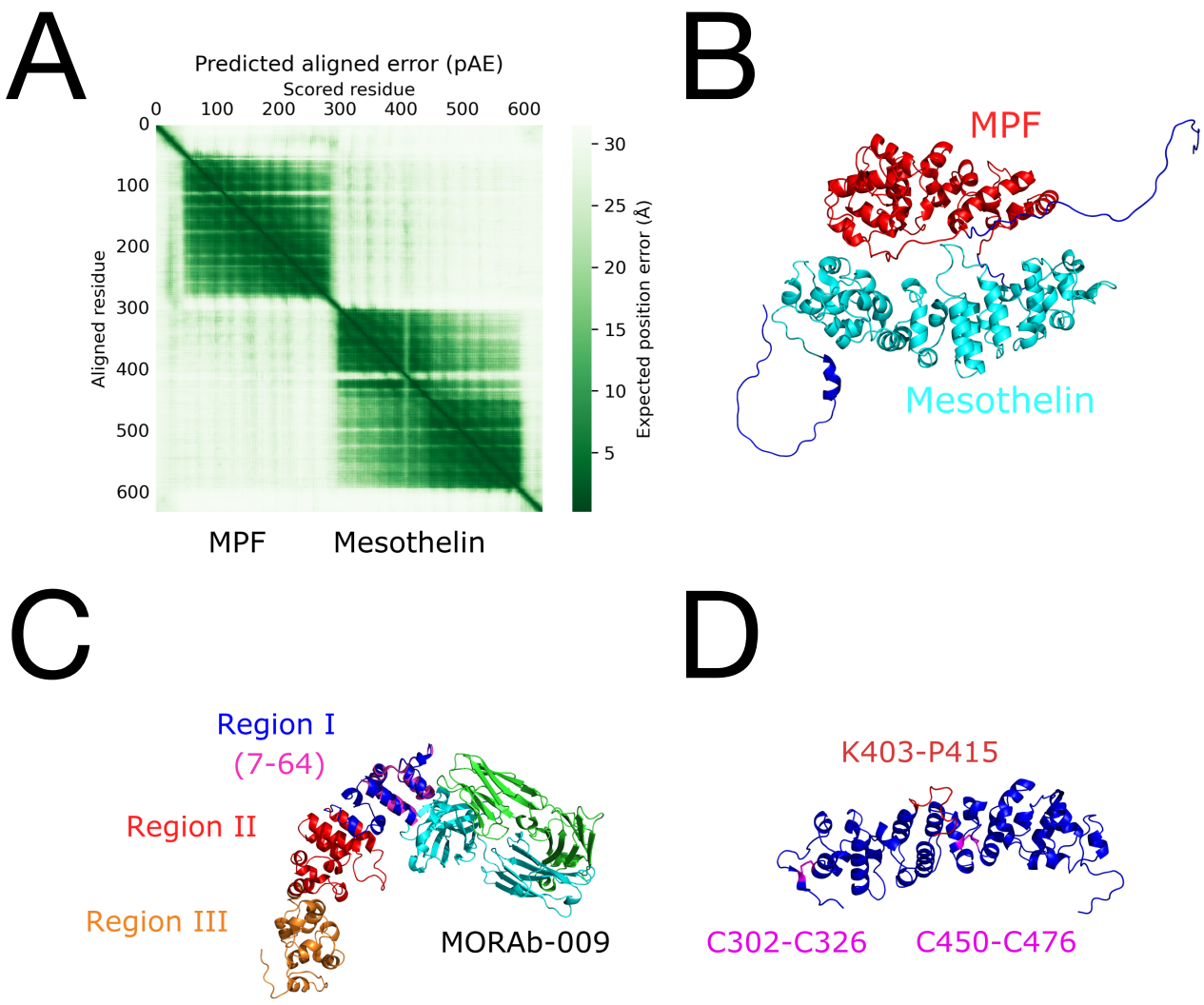


Fig. 20. Structural features of the AlphaFold model for human mesothelin.
A) Predicted aligned error plot for the model of the full-length precursor of mesothelin, showing two separate domains. B) The structure of the full-length precursor of mesothelin with the MPF domain highlighted in red and the mesothelin domain highlighted in cyan. The rest of the protein is shown in blue. C) Superposition of the AlphaFold mesothelin model with PDB ID 4F3F. Regions I, II and III of the mesothelin molecules are highlighted. The MORAb-009 Fab complex is shown in green and cyan. The mesothelin fragment from residue 7 to 64 from PDB ID 4F3F is shown in magenta. D) The structure of the AlphaFold mesothelin model. Disulfide bonds are highlighted in magenta and the disordered loop from residue 403 to 415 is highlighted in red.

3.3.1.5.2. HEPTATRICOPEPTIDE REPEAT PROTEIN 1 (HPR1)

Hillebrand et al (Hillebrand et al., 2018) identified a novel alpha-solenoid repeat family in *Plasmodium* - the heptatricopeptide repeat family. In addition, the authors showed mitochondrial targeting and in vitro RNA binding of a member of this family (PbHPR1) in *P. berghei*, suggesting a role for these repeat proteins in mitochondrial RNA processing. Heptatricopeptide repeat proteins have a 37-aa repeating unit and share similarities with pentatricopeptide repeat and octatricopeptide repeat proteins.

The AlphaFold model of the HPR1 protein in *P. falciparum* was detected by SOLeNNoID as an alpha-solenoid protein. The full model is shown in Fig. 21. A. The region from residue 35 to residue 295 was used for subsequent analysis. Consistent with the findings of Hillebrand et al, the protein contained both positively and negatively charged surface patches (Fig. 21 C), which are likely important for RNA binding. The repeat length in the model was 37 residues as predicted previously. Four repeats (73-109,110-146,147-183,184-220) were selected and analysed using my repeat splitting approach outlined in the previous chapter (Fig. 21 B, D). The consensus sequence found using my repeat annotation method was consistent with the findings of Hilebrand et al (2018). However, in my analysis three residue positions not annotated as helical in the original paper were predicted to be 310-helical (cyan in the sequence logo) by STRIDE (Heinig & Frishman, 2004). HPR proteins are essential for the survival of apicomplexan organisms like *Plasmodium* and *Toxoplasma*. The AlphaFold database contains a number of additional putative heptapeptide repeat protein entries for *P. falciparum*, which could extend the analysis by Hillebrand et al and provide information about the sequence-structure relationship of HPR proteins. RNA-binding proteins such as PPR and OPR proteins have been explored as components in inducible expression switches (Carrera-Pacheco et al., 2020; Rojas et al., 2019). HPR proteins could be used to further expand the repertoire of custom RNA binders for such applications.

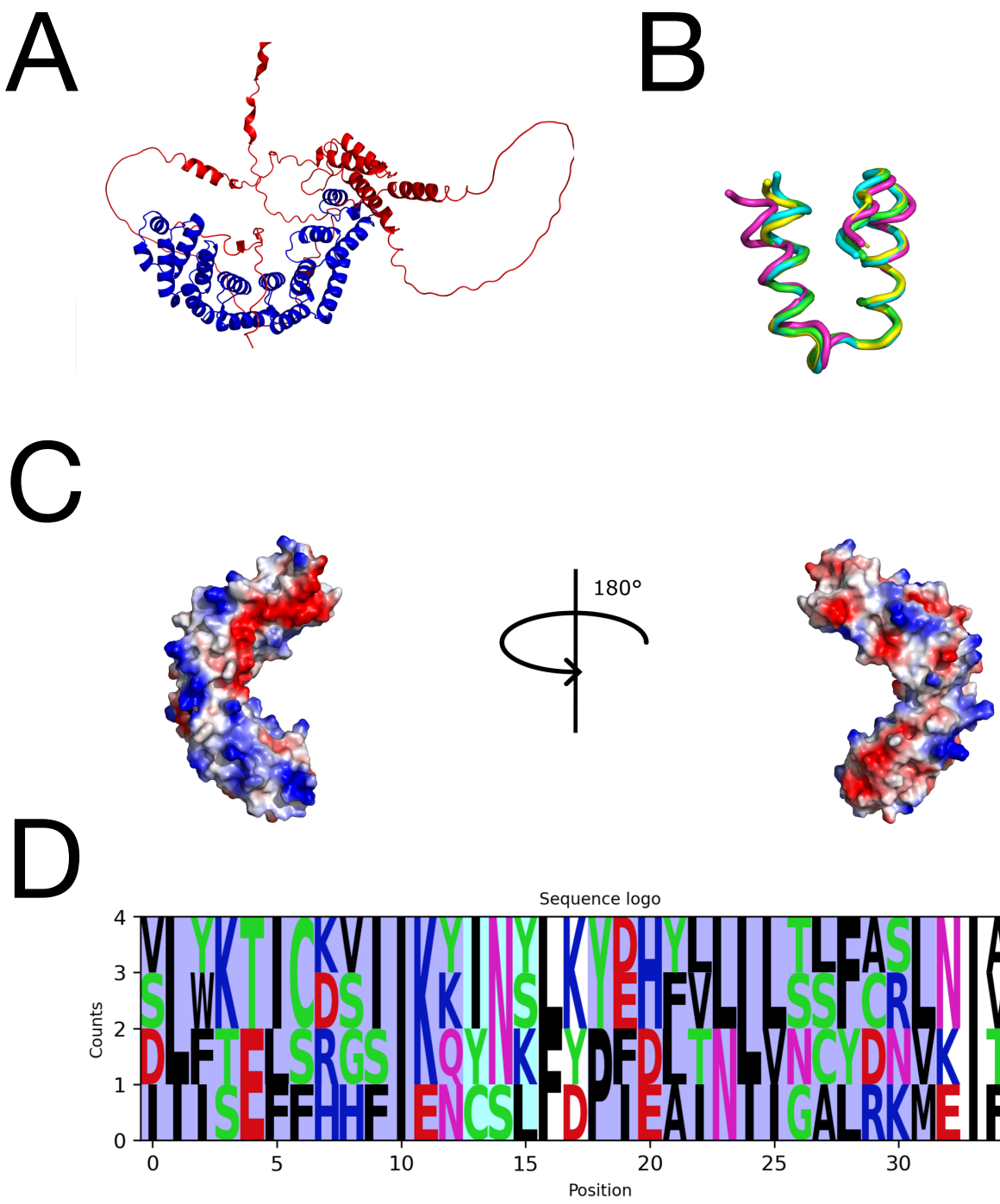


Fig. 21. Structural and sequence feature of the AlphaFold model of *P. falciparum* HPR1. **A) Full AlphaFold model for HPR1. Blue - region from 35 to 295 used for subsequent analysis. B) mTM-align superposition of four HPR1 repeats (73-109,110-146,147-183,184-220) C) Electrostatic surface representation generated using PyMOL showing positively and negatively charged patches on the surface of HPR1. D) Sequence logo produced using the four repeats from B). Helical residues predicted by STRIDE are highlighted in blue, 3-10-helical residues predicted by STRIDE are highlighted in cyan. Counts in the Y-axis correspond to the number of repeats in which a residue is observed at a given position. For example, at position 1, Leu is observed in 3 repeats, while Ile is observed in 1 repeat.**

3.3.1.5.3. DISPERSED GENE FAMILY 1 (DGF-1)

Trypanosoma cruzi is the organism, which causes Chagas disease. The Dispersed Gene Family-1 is the third largest *T. cruzi*-specific gene family. Kawashita et al (Kawashita et al., 2009) showed that the DGF-1 genes have conserved adhesin motifs and that DGF-1 members are expressed on the trypomastigote stage surface. Members of DGF-1 contain RGD integrin-binding motifs and carbohydrate binding sequences, suggesting a role in binding to extracellular matrix components (Rodríguez-Bejarano et al., 2021).

To my knowledge, no experimental structures or models of dgf-1 protein structure have been discussed in the literature. SOLeNNoID detected 180 entries for the *T. cruzi* proteome with the label 'dispersed gene family protein 1 (dgf-1), putative' in the AlphaFold database as beta-solenoid proteins. The DGF-1 entries had one or more beta-solenoid domains with an L-shaped cross section. In entry Q4CX45, there were 5 domains (Fig. 22 A), whereas in entry Q4CXB9, there was a single domain (Fig. 22 B). AlphaFold predicts the individual solenoid domains of Q4CX45 with high confidence but predicts disordered linkers between them, so the exact domain orientation is unclear. Q4CXB9 was used as a template for a structure similarity search using Foldseek (Kempen et al., 2022). Two of the most similar entries in the PDB (TM-align TM-score of >0.6) were PfbA (PDB ID 4MR0) (Fig. 22 C), a surface adhesin from *S. pneumoniae* which binds to human fibronectin and plasminogen (PDB ID 6K0U) (Fig. 22 D) (Beulin et al., 2014) and α-1,3-glucanase from *Paenibacillus glycanilyticus* (Itoh et al., 2020). These results were consistent with the earlier findings outlined above and suggest that DGF-1 member proteins use beta-helical structures to mediate cell adhesion. Further investigation into the structural diversity of these models is required as well as docking efforts to identify putative sites and mechanisms of interaction of DGF-1 proteins and extracellular matrix components.

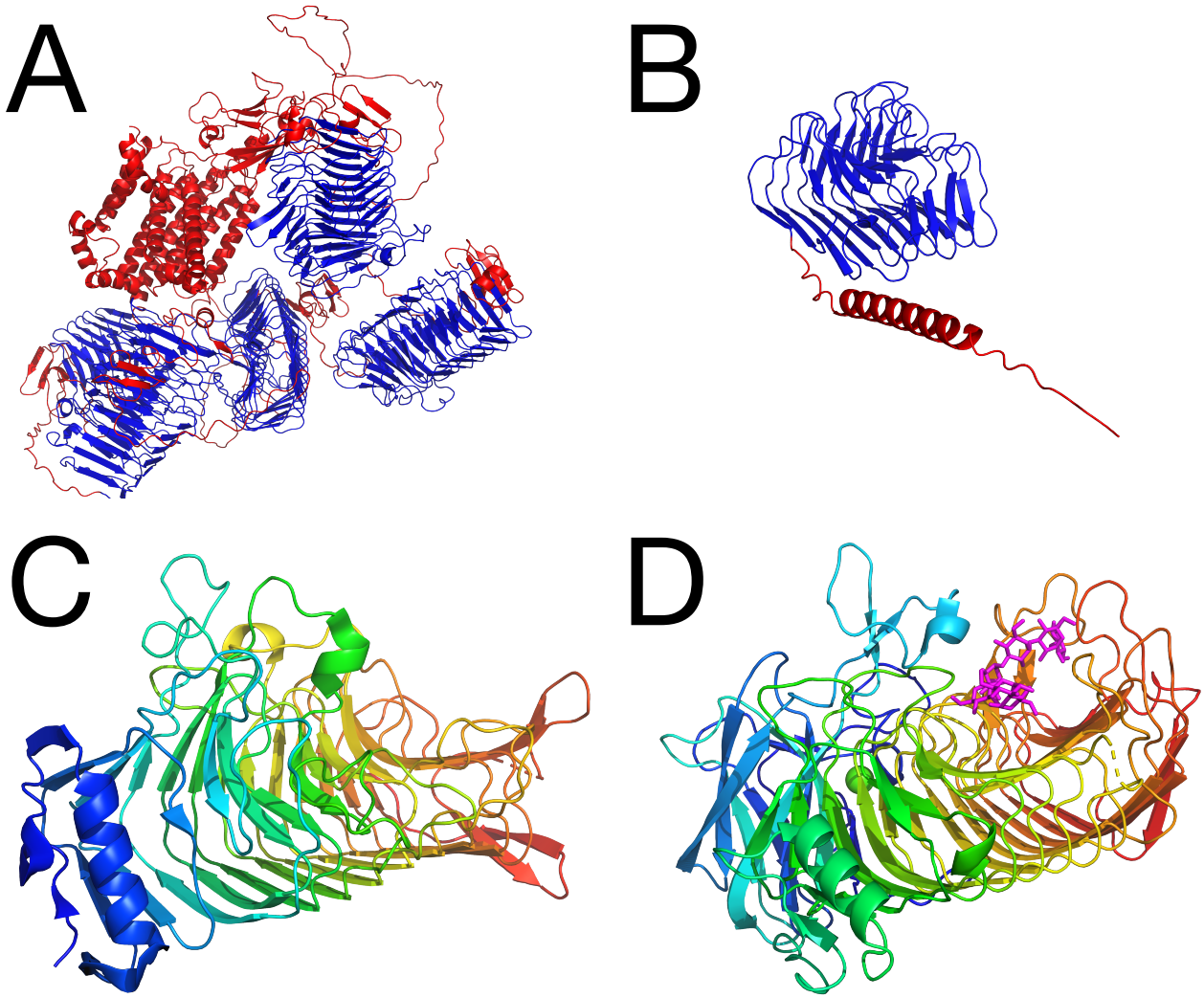


Fig. 22. Structural features of AlphaFold models of dgf-1 family proteins and structurally related proteins from the PDB.
A) AlphaFold model of Q4CX45. Beta-solenoid domains are coloured blue, the rest of the protein is coloured red. B) AlphaFold model of Q4CXB9. Beta-solenoid domains are coloured blue, the rest of the protein is coloured red. C) Structure of PfbA from *S. pneumoniae* (PDB ID 4MR0) coloured N-to-C. D) Structure of α-1,3-glucanase from *P. glycanilyticus* (PDB ID 6K0U) coloured N-to-C.

3.3.1.5.4. CURLIN SUBUNITS

E. coli produces an extracellular matrix with two main components: the curli protein polymer and cellulose. The extracellular matrix has a number of functions: it allows *E. coli* cells to adhere to a number of surfaces and provides resistance to immune and antimicrobial attack (Hufnagel et al., 2015). The components of the the curli polymer are CsgA and CsgB. Both are secreted across the bacterial outer membrane. CsgB binds to the cell surface and nucleates CsgA amyloid formation. Structural modelling predicts that CsgA and CsgB have beta helical structures with conserved glutamine and asparagine residues forming a stabilising hydrogen-bonding network (Barnhart & Chapman, 2006). Curli fibres are attractive from a biotechnological perspective as CsgA proteins can be functionalised, providing a biofilm that goes beyond normal biological functions. Indeed, this has been investigated by P.Q. Nguyen et al and Cui et al (P. Q. Nguyen et al., 2014; Cui et al., 2019), whereas Schmieden et al (Schmieden et al., 2018) have developed a method for 3D printing with a suspension of bacteria in a solution of alginate. The bacteria can then be induced to form a biofilm-like structure and stick to the underlying substrate.

SOLeNNoID detected both CsgA (major curlin subunit) and CsgB (minor curlin subunit) models in the AlphaFold database as beta-solenoids. The regions from residue 43-151 (CsgA) and 45-151 (CsgB) constitute the main beta-helix regions and were selected for further analysis. The alignment of CsgA and CsgB done using CEalign in PyMOL had an RMSD of 1.02 Å over 104 residues, suggesting that the two proteins are almost identical (Fig. 23 A). An analysis of polar contacts in PyMOL supports the previously predicted hydrogen bonding networks formed by the conserved glutamine and asparagine residues. Additionally, the conserved serines in position one of the repeat formed hydrogen bonds with residues from the following repeat. The observed hydrogen bonds involved both main chain and side chain atoms. Therefore, the stability of curli fibres is likely achieved by a combination of hydrophobic core interactions (Fig. 23 B, positions coloured cyan) and H-bonding networks (Fig. 23 B, positions coloured blue). Sequence logos obtained using the method outlined in the previous chapter confirmed that two of the repeats in CsgA have one additional residue (coloured orange in Fig. 23 C) and showed that CsgB has a uniform unit length and maintains 3 out of 4 consensus hydrogen-bonding residues (coloured blue in Fig. 23 D). It might be possible to introduce insertions into the two repeats with an extra residue to make functionalised CsgA mutants as most efforts so far have focused on terminal insertions.

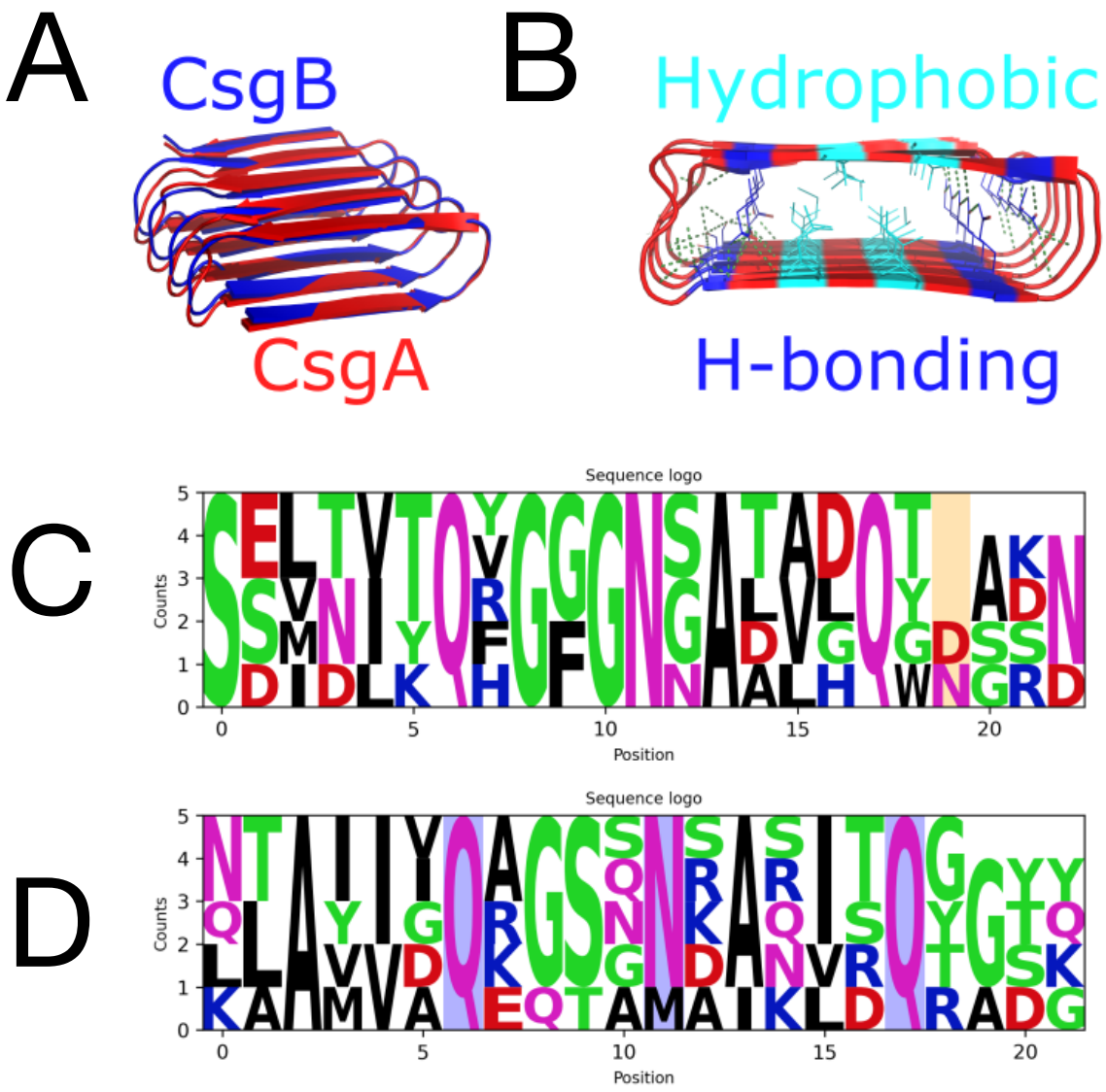


Fig. 23. Structural and sequence features of the AlphaFold models of CsgA and CsgB curlin subunits.
A) Superposition of CsgA and CsgB. B) View of the cross-section of CsgA highlighting structurally important hydrogen-bonding and hydrophobic residues. Hydrogen-bonding residues are coloured blue, hydrophobic core residues are coloured cyan. Hydrogen bonds are shown in green. C) Sequence logo produced using CsgA repeats. The position of the extra residue observed in two of the repeats is highlighted in orange. D) Sequence logo produced using CsgB repeats. The positions of highly conserved residues present in both CsgA and CsgB are highlighted in blue. In both C) and D) counts in the Y-axis correspond to the number of repeats in which a residue is observed at a given position. For example, at position 0 in CsgA (C) Ser is observed in all 5 repeats.

3.3.1.5.5. PHOTOSYSTEM I ASSEMBLY PROTEIN YCF3

Ycf3 is essential for stable accumulation of PSI and acts at the post-translational level. The amino acid sequence contains 3 tetratricopeptide (TPR) repeats and is conserved from cyanobacteria to higher plants (Naver et al., 2001). Naver et al found that Ycf3 is required for assembly but not stabilisation of PSI through a mutagenesis screen while specific interactions with PsaA and PsaD were detected by immunoprecipitation. Mutations in Tyr95-Tyr96 and Tyr142-Trp143 led to impaired growth on minimal medium in *C. reinhardtii* (Naver et al., 2001). Albus et al (Albus et al., 2010) identified Y3IP1 - a nucleus-encoded thylakoid protein which specifically and stably interacts with Ycf3. These findings have received further support by Nellaepalli et al (Nellaepalli et al., 2018), who also report that the Ycf3-Y3IP1 module binds to PsaA and PsaB PSI subunits, suggesting a role in the assembly of reaction centre subunits.

SOLeNNoID detected the Ycf3 AlphaFold model from *G. max* and other plant species as an alpha-solenoid protein. Consistent with previous findings, the model adopted a structure with 3 TPR repeats with additional helices flanking the TPR domain N- and C-terminally (Fig. 24 A). The residues mutated by Naver et al are shown in red. Residues Y94 and Y142 were exposed in the AlphaFold model, suggesting a possible involvement in protein-protein interactions. Residues Y95 and W143 could instead be involved in the packing between adjacent TPR repeats or helices. Using the *A. thaliana* Ycf3 and Y3IP1 models, the multimer version of AlphaFold 2 was run to identify possible interaction positions. The highest scoring model was selected and low confidence predictions for Y3IP1 were removed to yield the region 103-191 (Fig. 24 B, blue). This model suggests a possible role for Y94 in interacting with Y3IP1 and some shape complementarity could be observed between the first and second TPR repeats and Y3IP1. Further investigations could include more extensive modelling of the Ycf3-Y3IP1 complex and possibly modelling Ycf3 or Ycf3-Y3IP1 interactions with PsaA.

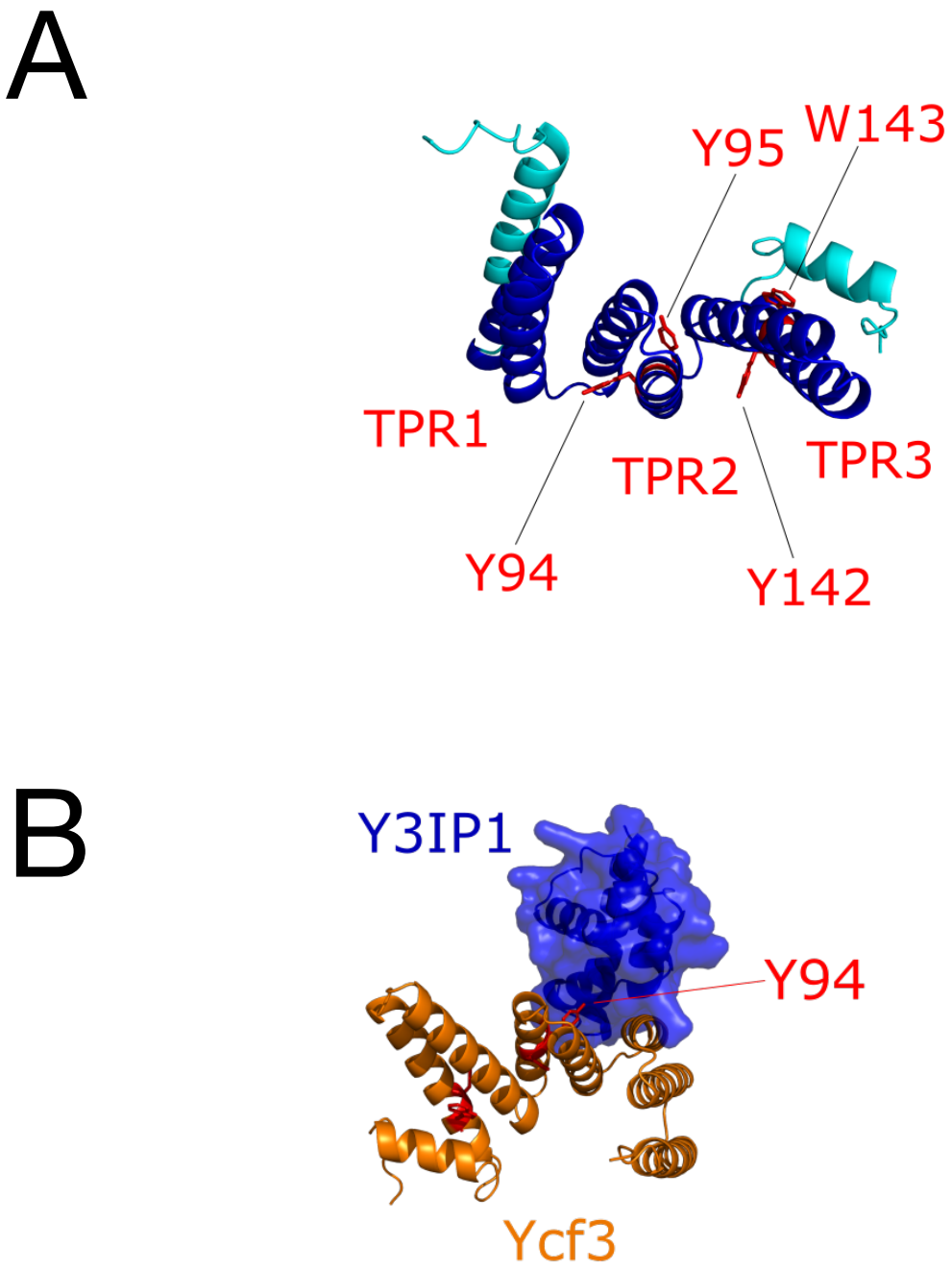


Fig. 24. Structural features of the AlphaFold model of *G. max* Ycf3 and the AlphaFold Multimer model of a complex between *A. thaliana* Ycf3 and Y3IP1.
A) *G. max* Ycf3. The TPR region is coloured blue. Mutations leading to impaired growth on minimal medium in *C. reinhardtii* are coloured red. B) Ycf3-Y3IP1 complex prediction. The mutations from A) are shown in red.

3.3.2. PDB SEARCH

3.3.2.1. SOLENOID PROTEIN STATISTICS IN THE PDB

568,479 chains from the PDB were processed using SOLeNNoID. This corresponds to 168,572 unique PDB IDs. The total number of chains identified as solenoid-containing by SOLeNNoID according to the criteria outlined in Section 3.2.2 was 11,582, about 2% of all entries in the PDB. This figure agrees with the work of Chakrabarty and Parekh (Chakrabarty & Parekh, 2021), which also found that approximately 2% of the PDB entries contain solenoids. Considering only unique PDB IDs, 4,124 PDB IDs were identified as solenoid-containing, representing 2.44% of the 168,572 processed PDB IDs. The total number of non-redundant functional annotations (classification structure keyword) was 406. As shown in Table 9, alpha-solenoids had the largest number of solenoid residues, followed by alpha/beta and then beta-solenoids. SOLeNNoID predictions also suggest that alpha-solenoid entries were the most abundant in the PDB. However, the incorrect classification of non-solenoid structures as alpha-solenoid by SOLeNNoID discussed previously also applies here.

	Alpha-solenoid	Alpha/beta-solenoid	Beta-solenoid	Combined
Number of chains	9275	950	957	400
Number of PDB IDs	3063	468	360	233
Min. number of solenoid residues	36	49	28	49
Mean number of solenoid residues	284	275	197	293
Max. number of solenoid residues	3334	732	631	1940
Unique functional annotations	335	87	39	46

Table 9. Statistics for PDB entries detected as solenoids by SOLeNNoID.

3.3.2.2. COMPARISON WITH OTHER REPEAT DATABASES

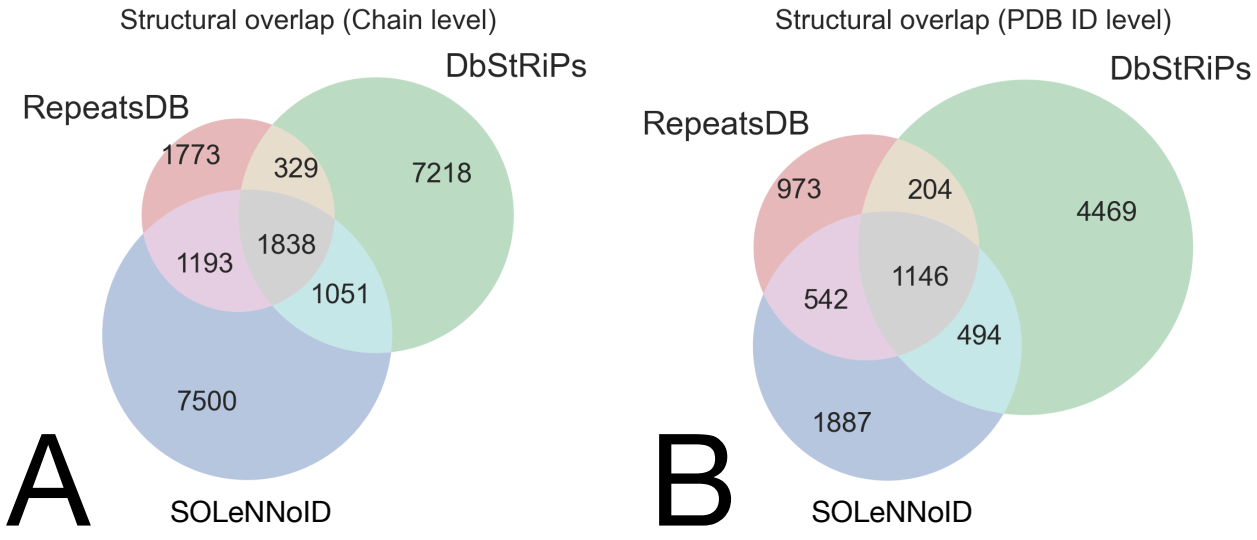


Fig. 25. Venn diagrams showing overlap between the structures detected as solenoid by SOLeNNoID and solenoid structures in the RepeatsDB and DbStRiPs databases.
Red - solenoid structures found only in the RepeatsDB database, green - solenoid structures found only in the DbStRiPs database, blue - solenoid structures found only using SOLeNNoID. Grey - solenoid structures found in all databases, pink - solenoid structures found both in RepeatsDB and by SOLeNNoID, cyan - solenoid structures found both in DbStRiPs and by SOLeNNoID, yellow - solenoid structures found both in RepeatsDB and DbStRiPs. A) Overlap between databases and SOLeNNoID discoveries when considering individual chains. B) Overlap between databases and SOLeNNoID discoveries when considering unique PDB IDs.

As shown in Fig. 25, there was significant overlap between the PDB structures found by SOLeNNoID and those found in the RepeatsDB and DbStRiPs databases. However, SOLeNNoID also found many structures not covered by the databases. SOLeNNoID identified more new chains than PDB IDs, likely due to detecting multiple chains from large protein complexes. The large number of PDB IDs in DbStRiPs is largely due to entries from the LRR repeat type (4217). However, on closer inspection of entries from this class, a number of chains with a prediction score <0.8 were found not to belong to the LRR repeat type. An example shown below is chain B from PDB ID 5L13 (Buchman et al., 2017). The prediction score for this entry was 0.79 and consecutive repeats are shown in alternating red and blue in Fig. 26. The repeats detected for 5L13 do not follow the topology of the LRR fold. I estimate that up to 1000 (or more) DbStRiPs entries for this repeat type might be misclassified so the real number of solenoid chains and PDB IDs in DbStRiPs is likely lower than in Fig. 25 B.

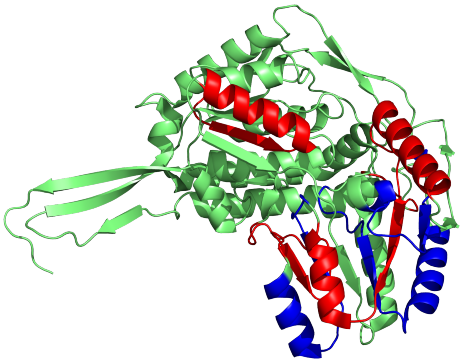


Fig. 26. Example of an incorrect LRR classification in DbStRiPs (PDB ID 5L13). Consecutive repeats are shown in alternating red and blue.

Table 10 shows the number of putative solenoid chains and PDB IDs found by SOLeNNoID that were not present in RepeatsDB and DbStRiPs. A total of 6927 new chains and 1887 unique PDB IDs were found with most belonging to the alpha-solenoid class. The discovery of new solenoid structures by SOLeNNoID is due to two factors: 1) the ability of SOLeNNoID to detect examples that are not accessible to other methods and 2) processing a more recent version of the PDB with a larger number of chains and IDs. The scale of my contribution to solenoid protein annotation is demonstrated as a percentage increase over the "gold standard" RepeatsDB database. These results further highlight the need for multiple approaches for finding solenoid proteins as different methods show significant overlap but also complementarity and together, they enable a better coverage of solenoid protein structures.

	Alpha-solenoid	Alpha/beta-solenoid	Beta-solenoid	Combined	Total
Number of chains	6032	310	332	253	6927
Number of PDB IDs	1517	148	119	141	1887
% increase over RepeatsDB (PDB IDs)	81.3	45.7	27.7	N/A	72

Table 10. Number of structures detected by SOLeNNoID and not present in other databases and percentage increase over entries in RepeatsDB.

3.3.2.3. DISTRIBUTION OF EXPERIMENTAL METHODS FOR SOLENOID STRUCTURE DETERMINATION

The most common structure determination methods for alpha, alpha/beta, beta-solenoids, and combined predictions are shown in Fig. 27. The prevalent method for alpha/beta and beta-solenoids was X-ray diffraction, while for alpha-solenoids a larger number of structures were determined by EM compared to X-ray diffraction. This is possibly due to the participation of alpha-solenoid proteins in large protein complexes, which are more amenable to structure determination using cryo-EM.

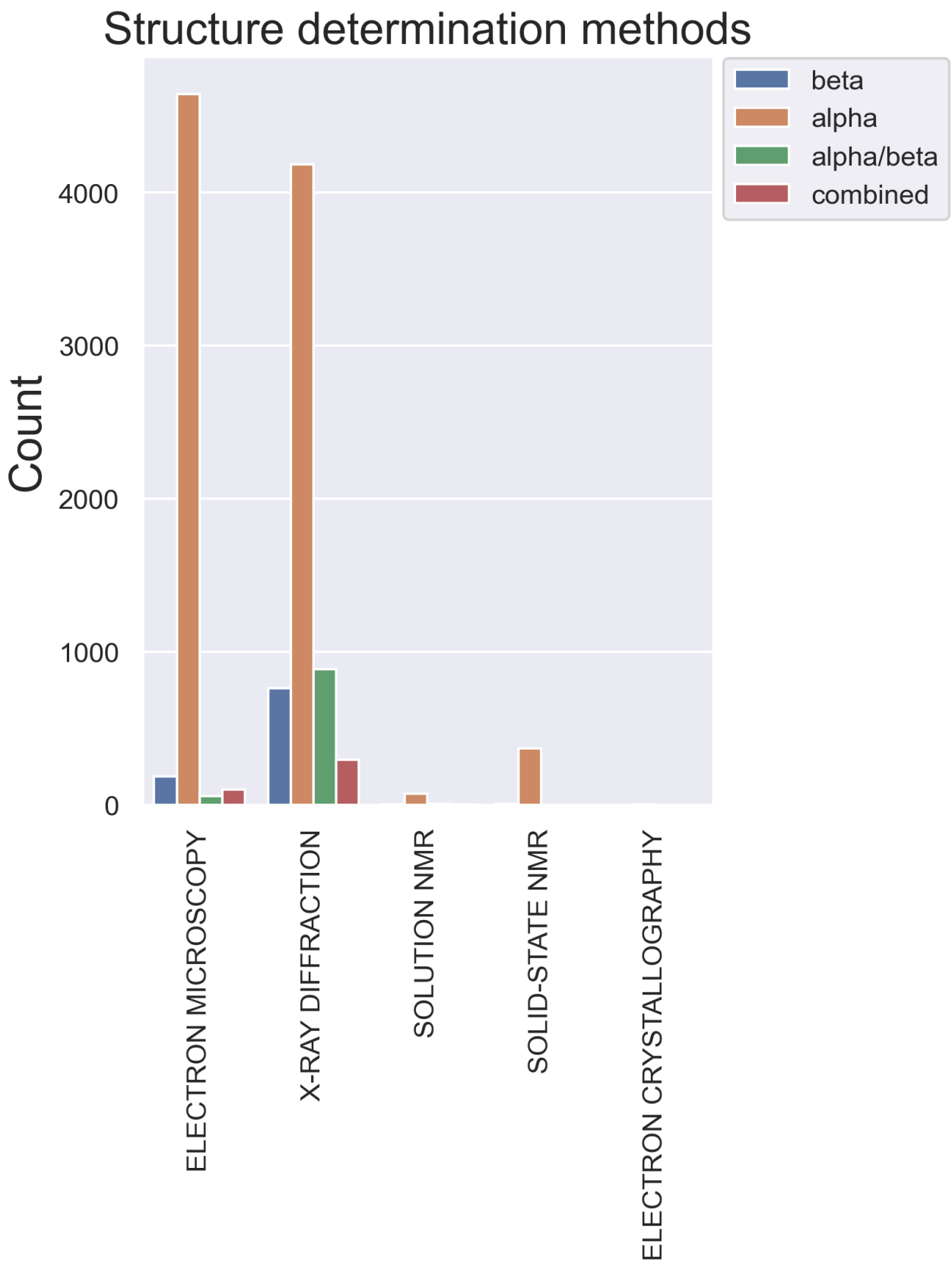


Fig. 27. Numbers of solenoid structures detected by SOLeNNoID determined by different experimental methods.
Top right - legend showing colours corresponding to solenoid classes.

3.3.2.4. DISTRIBUTION OF FUNCTIONS FOR SOLENOID PROTEINS

Alpha-solenoid proteins are frequently found as part of protein complexes, and this was reflected in the distribution of keywords for the alpha-solenoid PDB entries detected by SOLeNNoID. For example, alpha-solenoids were most commonly involved in transport. Examples of this include dsubunits of the COPI coat (Dodonova et al., 2015) and importins such as Importin13 (Grünwald et al., 2013). Protein transport functions included the UBAP1 subunit of the endosomal sorting complexes required for transport (Agromayor et al., 2012) and subunits of the ER membrane protein complex, involved in membrane protein biogenesis (O'donnell et al., 2020). Among protein binding functions were cytoskeleton-associated proteins such as *D. melanogaster* Msps (Howard et al., 2015) and KANK1 (Pan et al., 2018). Transcription and ribosome functions were exemplified by the Ctr9 subunit of Paf1, a general transcription elongation factor of RNA polymerase II (Deng et al., 2018) and the mS27 ribosomal subunit of the human mitochondrial ribosome (class 1) (Amunts et al., 2015). There were also a number of computationally designed alpha-solenoids including designed armadillo repeat proteins (Reichen et al., 2016) and designed ankyrin repeat proteins (Batyuk et al., 2016).

The functions of alpha/beta-solenoid proteins, in contrast, largely clustered around signalling. Alpha/beta-solenoids were most commonly associated with immune system roles. Examples included Toll-like receptors (Kang et al., 2009) and the variable lymphocyte receptors found in lampreys (Han et al., 2008). In terms of signalling functions, alpha/beta-solenoids comprised the ectodomains of leucine-rich repeat-containing G protein-coupled receptors (Peng et al., 2013) and fibronectin leucine-rich repeat transmembrane proteins, which are involved in cell adhesion and receptor signalling (Jackson et al., 2016), to name a few. Cell adhesion functions included chondroadherin, a protein which interacts with collagen and cell-surface proteoglycans and integrins (Rämisch et al., 2017), and leucine-rich repeat transmembrane neuronal proteins, which are involved in synaptic organisation (Yamagata et al., 2018). An alpha/beta-solenoid protein involved in splicing is U2 small ribonucleoprotein A', a component of the spliceosome (Zhan et al., 2018).

Beta-solenoid proteins primarily had enzymatic roles. The transferase function was exemplified by the *E. coli* GlmU acetyltransferase domain (Green et al., 2012) and *P. aeruginosa* LpxA (UDP-N-acetylglucosamine acyltransferase) (Smith et al., 2015). Beta-solenoid lyases included the PelI pectate lyase from *E. chrysanthemi* (Creze et al., 2008) and the KflA lyase from K5A coliphage (Thompson et al., 2010). Examples of hydrolases included pectin methylesterase from *A. niger* (Kent et al., 2016) and polygalacturonase from human colonic *Bacteroides* (Luis et al., 2017). A large number of detected beta-solenoid structures were viral proteins such as bacteriophage tailspike proteins (Broeker et al., 2013) and the snake adenovirus cementing protein LH3, which also has a phage tailspike-like fold (Menéndez-Conejero et al., 2017). Beta-solenoids were also the only solenoid class that could perform antifreeze functions. Examples of this were TisAFP8

from *T. ishikariensis* (Cheng et al., 2016), which has a triangular cross-sectional shape, and RiAFP from *R. inquisitor* (Hakim et al., 2013), which has an oval cross-sectional shape.

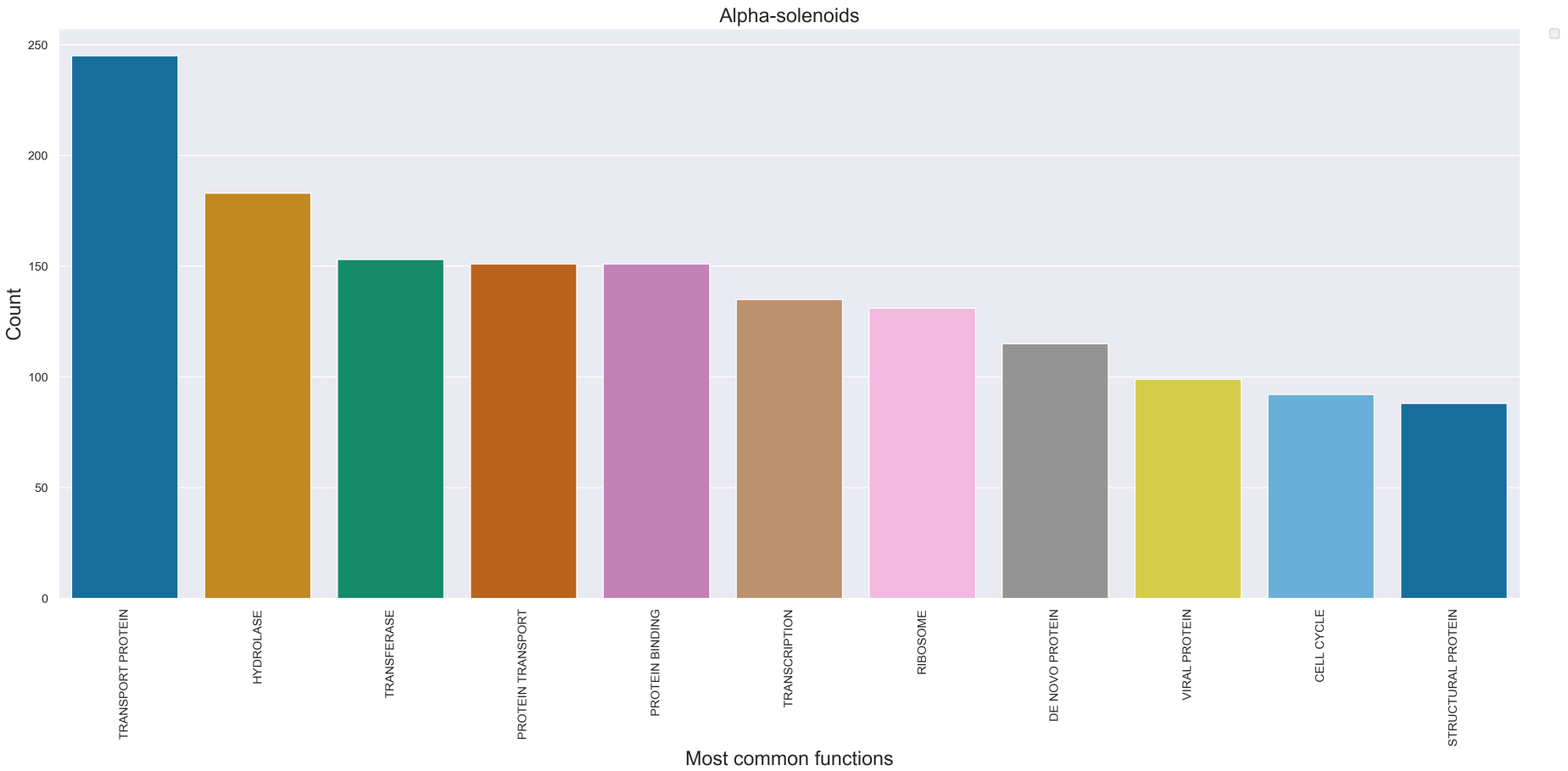


Fig. 28. The ten most frequent structure keywords in the PDB across solenoid classes.
The most frequent annotations for alpha-solenoid proteins.

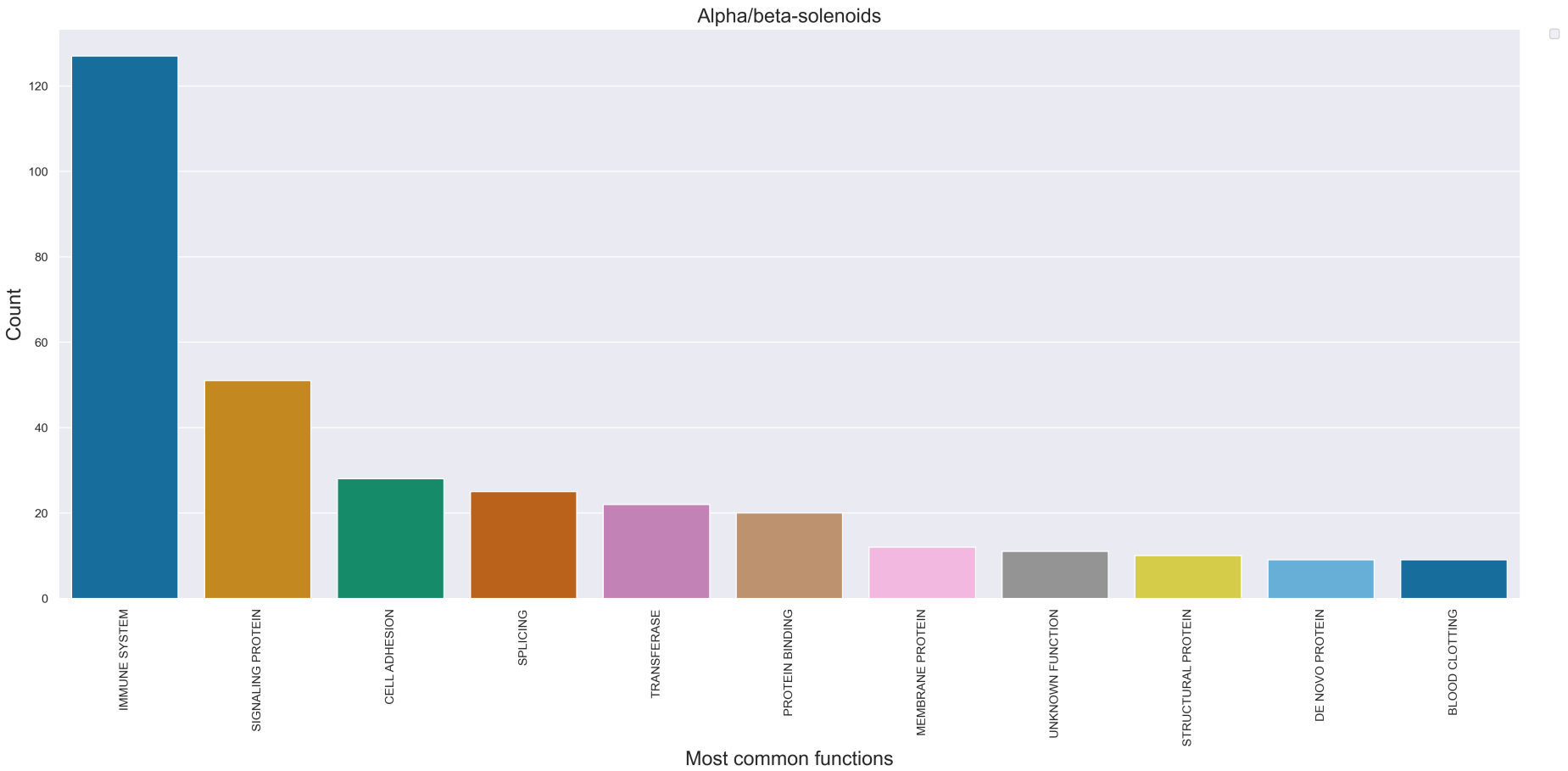


Fig. 28. The ten most frequent structure keywords in the PDB across solenoid classes.
The most frequent annotations for alpha/beta-solenoid proteins.

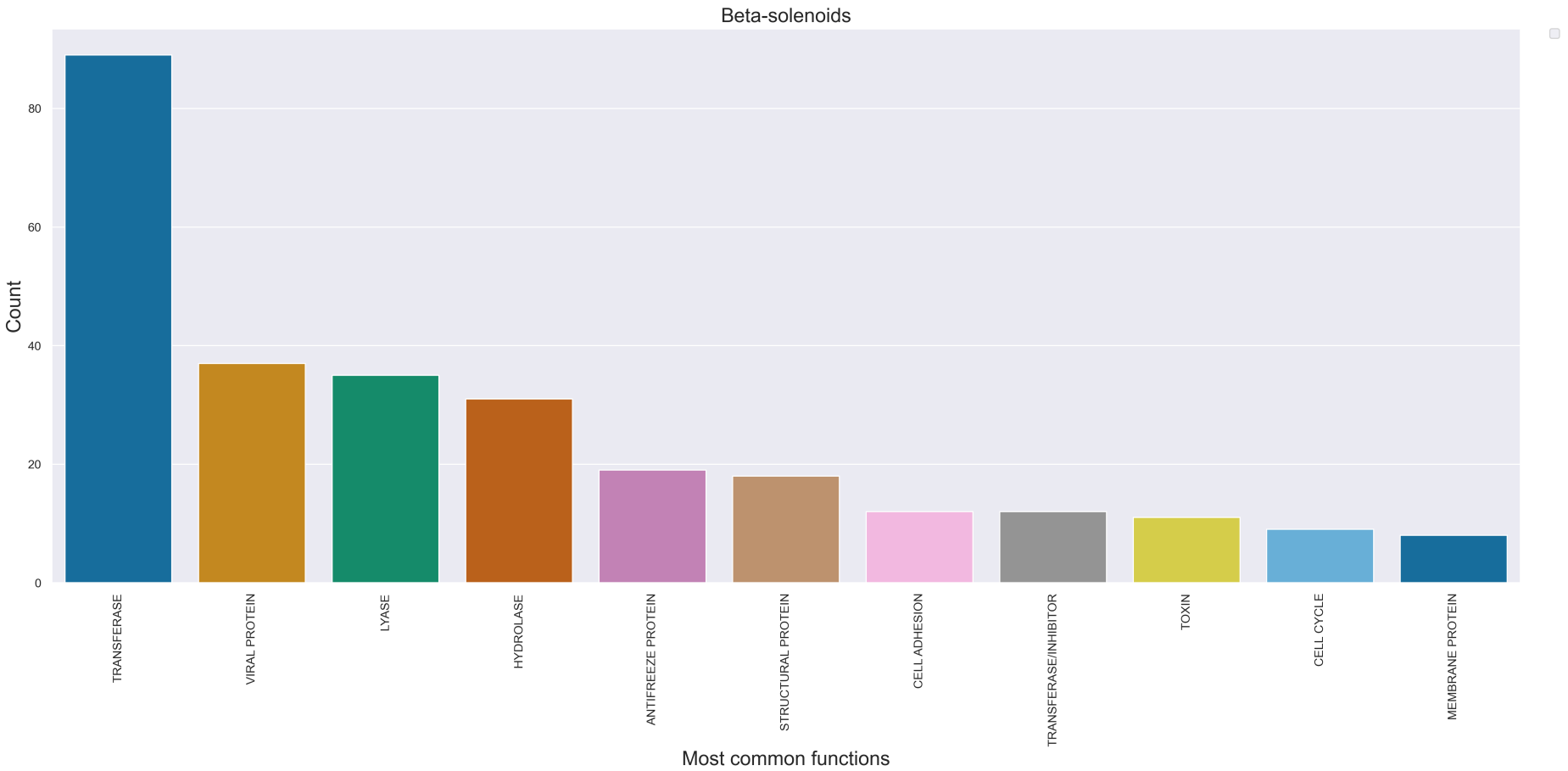


Fig. 28. The ten most frequent structure keywords in the PDB across solenoid classes.
The most frequent annotations for beta-solenoid proteins.

Using the ten most common structural keywords for each solenoid class, a table of shared functional annotations was produced (Table 11). According to this analysis, the transferase and structural protein functions were shared between all solenoid classes, while the remaining functions in Table 11 were shared between two out of three solenoid classes. Therefore, different solenoid classes are involved in overlapping but non-identical sets of functions.

Function	Alpha-solenoid	Alpha/beta-solenoid	Beta-solenoid
Hydrolase	✓	✗	✓
Transferase	✓	✓	✓
Structrural Protein	✓	✓	✓
Cell Adhesion	✗	✓	✓
Cell Cycle	✓	✗	✓
Protein Binding	✓	✓	✗
Viral Protein	✓	✗	✓
Membrane Protein	✗	✓	✓

Table 11. Shared functional annotations across solenoid protein classes.

3.3.2.5. DISTRIBUTION OF SOLENOID PROTEINS ACROSS SPECIES

The twenty most common species for all classes of solenoid proteins are shown in Fig. 29, whereas Fig. 30 shows a breakdown of the percentages of eukaryotic, prokaryotic, and viral species in the twenty most common species for the three solenoid classes. Alpha- and alpha/beta-solenoid structures in the PDB derived primarily from eukaryotic organisms, whereas beta-solenoid structures originated from prokaryotic and viral (most frequently bacteriophage) sources. Most alpha-solenoid and alpha/beta-solenoid structures were of human proteins, followed by proteins from *S. cerevisiae* and *A. thaliana*, respectively. Most beta-solenoid structures were of *E. coli* proteins, followed by *M. thermophila*. This shows that the distribution of solenoid PDB entries is unequal across species, which could reflect biases in research interest but could also be a result of differences in the distribution of solenoid proteins across species as discussed in Section 3.3.1.

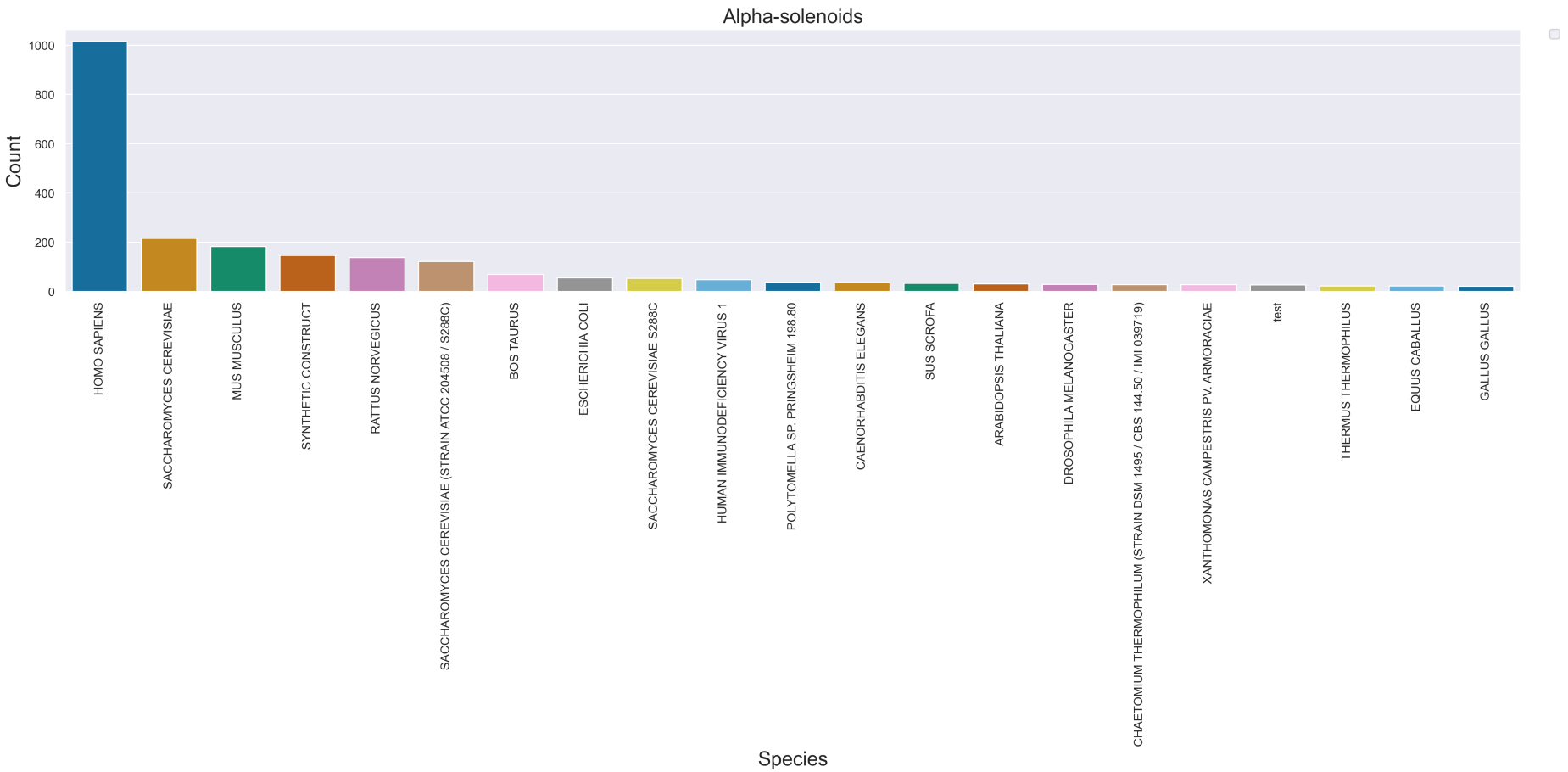


Fig. 29. The twenty species with the most solenoid protein entries in the PDB.
The species with the most alpha-solenoid entries.

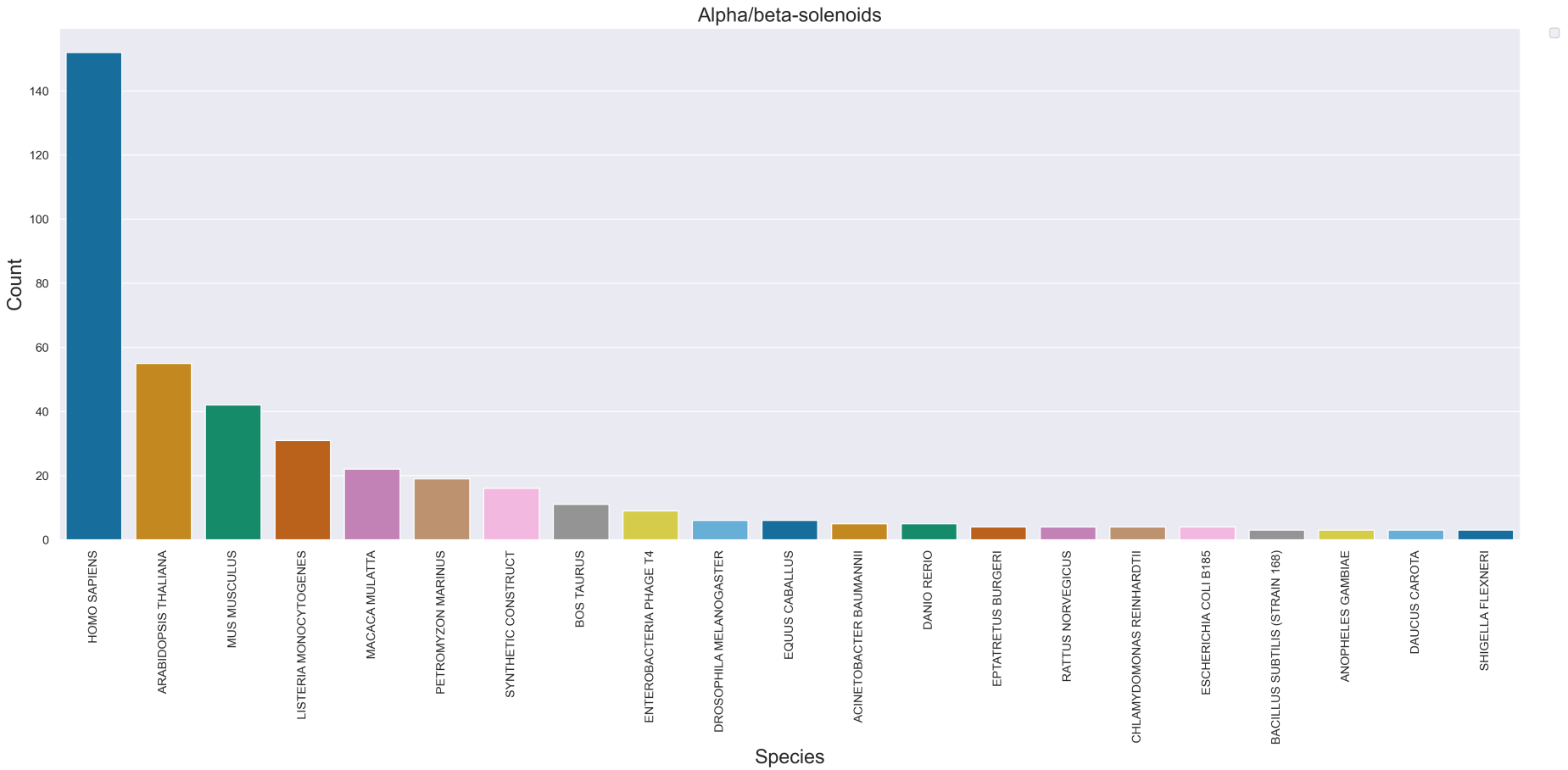


Fig.29. The twenty species with the most solenoid protein entries in the PDB.
The species with the most alpha/beta-solenoid entries.

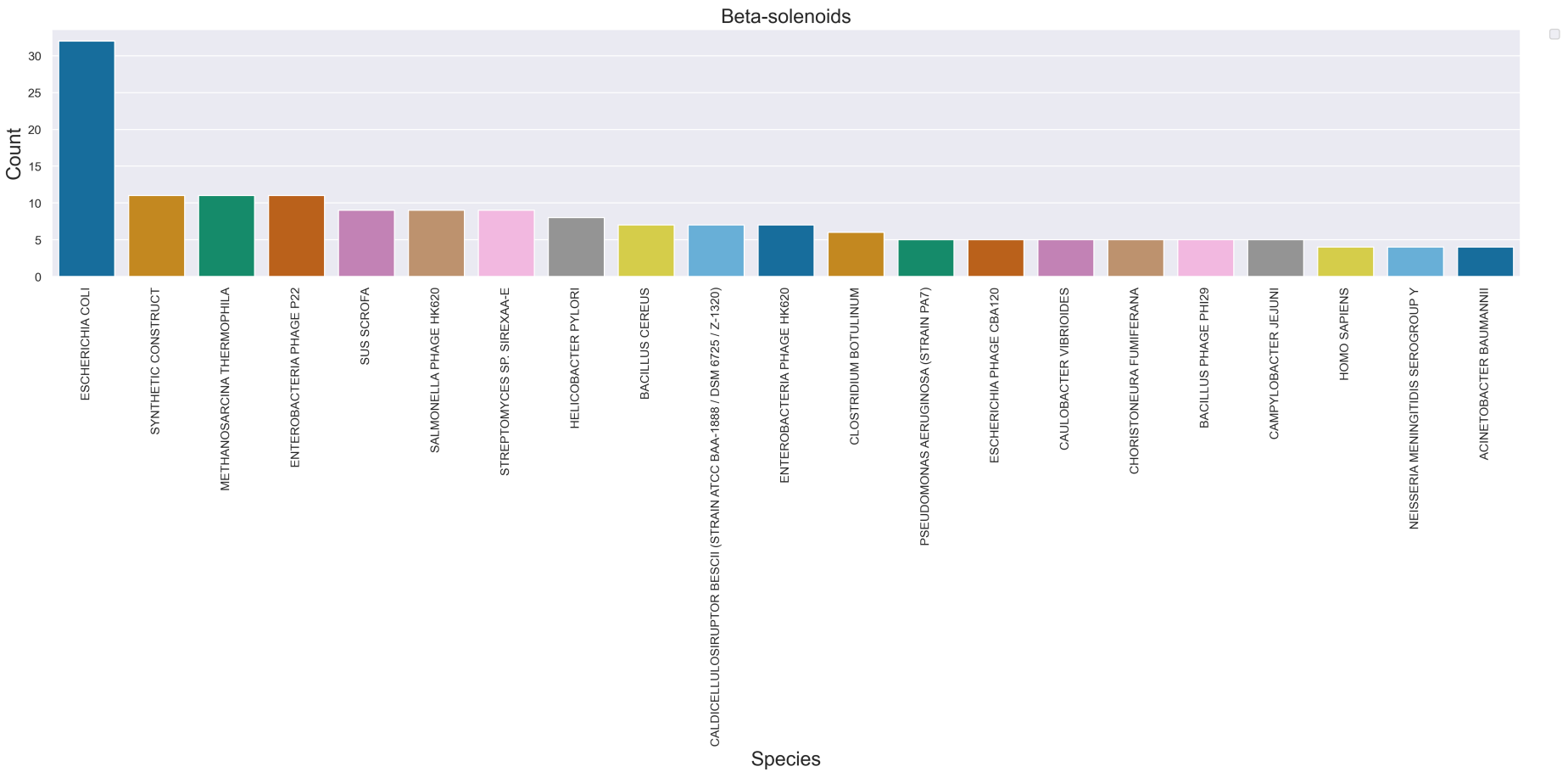


Fig. 29. The twenty species with the most solenoid protein entries in the PDB.
The species with the most beta-solenoid entries.

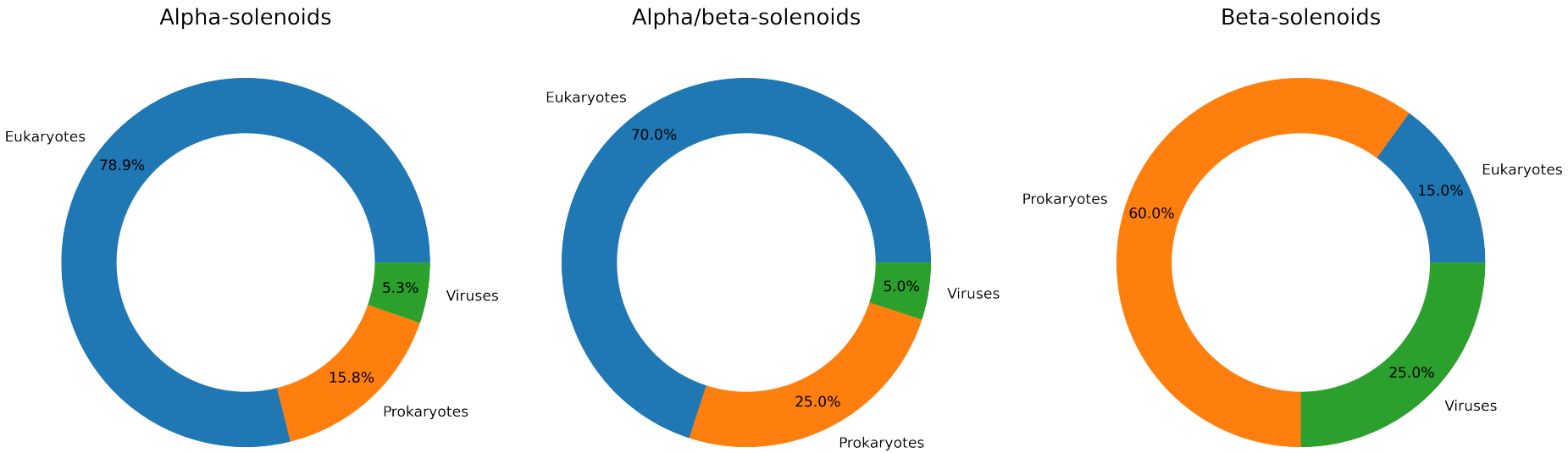


Fig. 30. The percentages of eukaryotic, prokaryotic and viral structures across solenoid classes.

3.3.2.6. EXAMPLES OF SOLENOID PROTEINS DETECTED BY MY METHOD

In this section I describe examples solenoid protein structures of note found in the PDB. A member of each solenoid class is discussed, as well as a protein containing residues from two solenoid classes in the same chain.

3.3.2.6.1. ALPHA-SOLENOID EXAMPLE (NOMPC)

The cryo-EM structure of the mechanotransduction channel NOMPC from *D. melanogaster* (PDB ID 5VKQ) (Jin et al., 2017) was detected as alpha-solenoid-containing by SOLeNNoID. This channel is important for locomotion, touch, and sound sensation in a number of invertebrates and vertebrates. NOMPC is tetrameric and each chain contains 29 ankyrin repeats. The authors speculate that the ankyrin domains could act as a spring connected to microtubules and thus enable channel gating in response to mechanical displacement. Analysis of 28 out of the 29 ankyrin repeats of chain A of the structure using my repeat analysis method yielded a sequence logo of the most common positions. A number of the ankyrin repeats deviated from the 33-residue idealised motif. Most notable were ankyrin repeats 13,14 and 24, which have a length of 43,43 and 46 residues, respectively. Repeat 24 is part of one of the inter-chain interaction clusters but the loop section is not involved in this interaction. The functional significance (if any) of these deviations in repeat length is unclear.

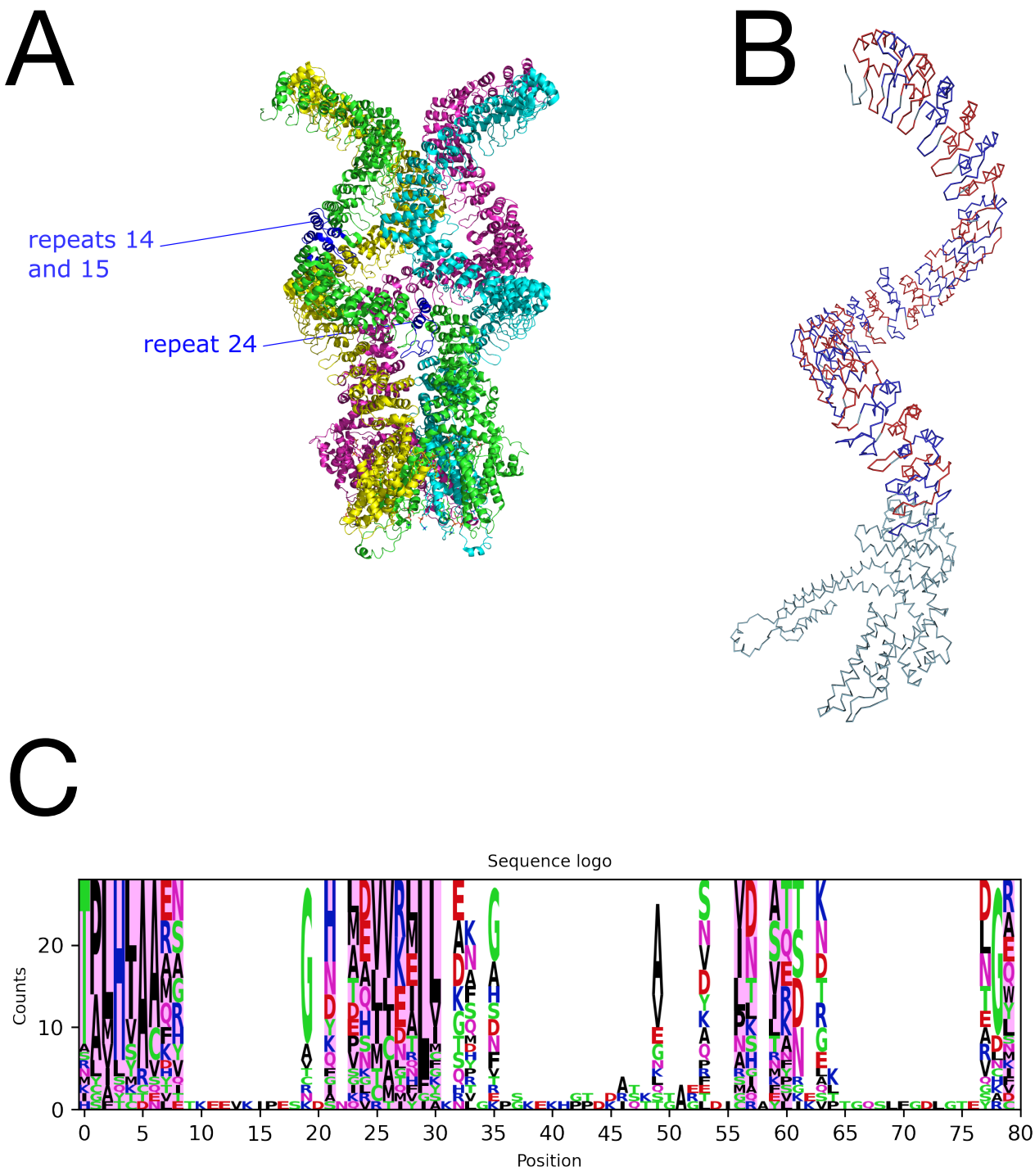


Fig. 31. Structural and sequence features of the *D. melanogaster* NOMPC channel (5VKQ). A) Structure of the channel tetramer. Chain A is coloured green and the ankyrin repeats with the largest deviation from the idealised 33-residue motif are coloured blue. B) 28 of the 29 ankyrin repeats highlighted using my repeat splitting method in alternating red and blue. C) Sequence logo produced using the 28 repeats from B). Common core regions according to mTM-align are highlighted in magenta. Counts in the Y-axis correspond to the number of repeats in which a residue is observed at a given position. For example, at position 0, Thr is observed in more than 10 repeats.

3.3.2.6.2. ALPHA/BETA-SOLENOID EXAMPLE (LRR20)

The X-ray diffraction structure of *Leptospira santarosai serovar Shermani* leucine-rich repeat 20 (PDB ID 6LX0) (Hsu et al., 2020) was detected as alpha/beta-solenoid by SOLeNNoID. *Leptospira* species cause the zoonotic disease leptospirosis in humans and pathogenic species contain LRRs, which are rarely expressed in non-pathogenic strains. The authors show that the rLRR20 protein interacts with the EC1 domain of E-cadherin via negatively charged residues on its concave surface, thus facilitating attachment of *Leptospira* to host cells.

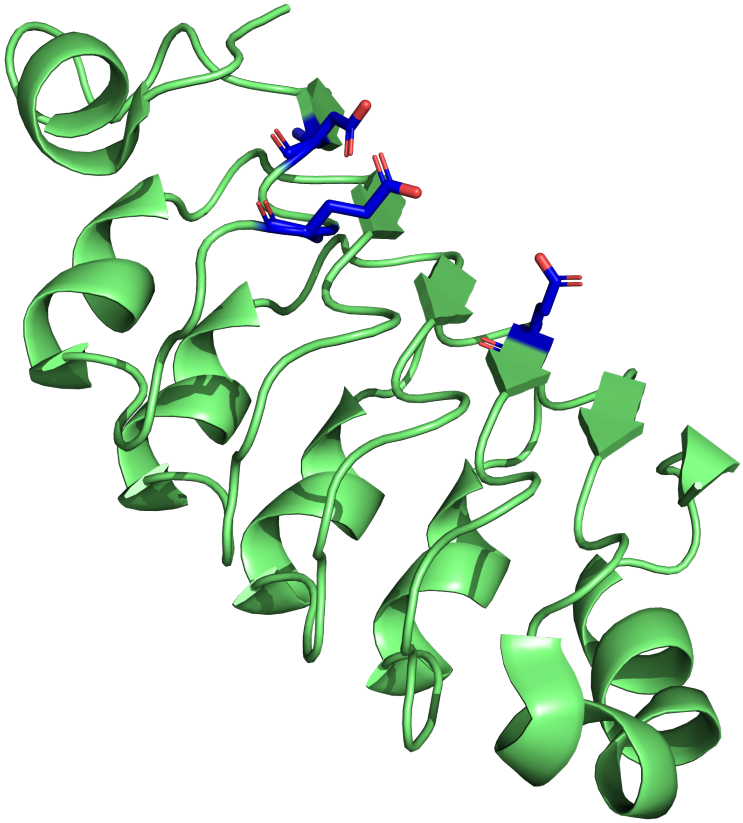


Fig. 32. The structure of *L. santarosai serovar Shermani* rLRR20.
Residues important for its interaction with the EC1 domain of E-cadherin (D56, E59 and E123) are shown in blue in stick representation.

3.3.2.6.3. BETA-SOLENOID EXAMPLE (FWDC, FHCC)

The X-ray diffraction structures of the FwdC subunit of formyl-methanofuran dehydrogenase from *Methanothermobacter wolfeii* (PDB ID 5T5I) (Wagner et al., 2016) and the FhcC subunit of the formylmethyltransferase/hydrolase complex from *Methylobacterium extorquens* (PDB ID 6S6Y) (Hemmann et al., 2019) were detected as beta-solenoid by SOLeNNoID (Fig. 33). FwdC and FhcC have structural scaffolding functions in their respective complexes. The alignment of FwdC and FhcC in PyMOL using CEalign resulted in an RMSD of 2.09 Å over 248 residues. The main differences between the two proteins are the C-terminal region, where FwdC has a beta-hairpin, while FhcC has a helix-loop-helix structure, and the surface charge distribution. FwdC flanks the tunnel between the FwdA and FwdBD active sites but its function is unknown. FhcC forms an interaction with FhcD and provides the bulk of the positively charged residues for methylofuran binding. The authors suggest that the structural differences between FhcC and FwdC reveal how the bacterial formate-producing formylmethyltransferase/hydrolase complex might have evolved from the CO_2-fixing formyl-methanofuran dehydrogenase. A further reason why these structures are interesting is that they are the only other known examples of B-shaped cross-section beta-solenoids, apart from the C-terminal domain of bacterial glutamate synthase, with which they have a distant sequence similarity.

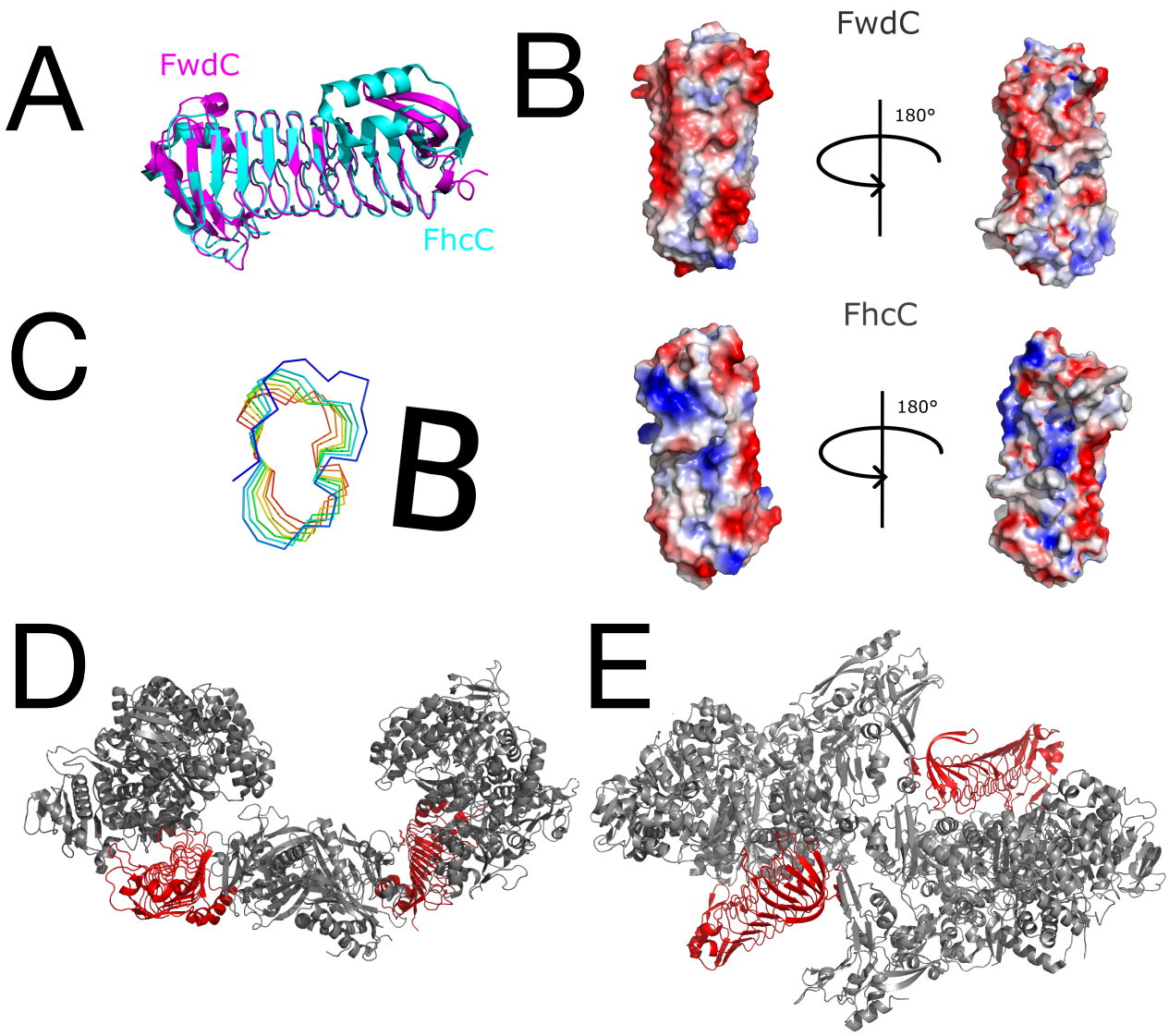


Fig. 33. Structural features of FwdC and FhcC.
A) Superposition of FwdC and FhcC. The C-terminal region of FwdC forms a beta-hairpin, while that of FhcC forms a helical hairpin. B) Electrostatic surface representation generated using PyMOL showing positively and negatively charged patches on the surfaces of FwdC and FhcC. C) The B-shaped cross-section of FwdC coloured N-to-C. D) Formylmethyltransferase/hydrolase complex from *Methylobacterium extorquens*, FhcC highlighted in red (PDB ID 6S6Y). E) formyl-methanofuran dehydrogenase from *Methanothermobacter wolfeii*, FwdC highlighted in red (PDB ID 5T5I).

3.3.2.6.4. MULTI-SOLENOID EXAMPLE (EIF2B)

SOLeNNoID identified an example of a single chain containing two solenoid regions: eukaryotic translation initiation factor 2B (eIF2B) subunit epsilon. eIF2B is crucial for protein synthesis and consists of a catalytic sub-complex formed by the gamma and epsilon subunits and a regulatory complex formed by the alpha, beta, and delta subunits. SOLeNNoID detected an alpha-solenoid and beta-solenoid region in the epsilon subunit (PDB ID 6O85) (Kenner et al., 2019). The alpha-solenoid region consists of HEAT repeats, while the beta-solenoid region has a left-handed beta-helical hexapeptide repeat structure. The HEAT domain is involved in the guanine nucleotide exchange factor function of eIF2B. Based on mutagenesis (X. Wang et al., 2012) and mass spectrometry and cross-linking (Gordiyenko et al., 2014) studies, the left-handed beta-helix domain of eIF2B epsilon is involved in eIF2B complex assembly and mutations in it can also affect the activity of eIF2B, independently of complex assembly. The left-handed beta-helix domain is also implicated in the autosomal recessive neurodegenerative disease leukoencephalopathy with vanishing white matter (VWM) as there are several surface mutations in this region of the protein (Kashiwagi et al., 2016).

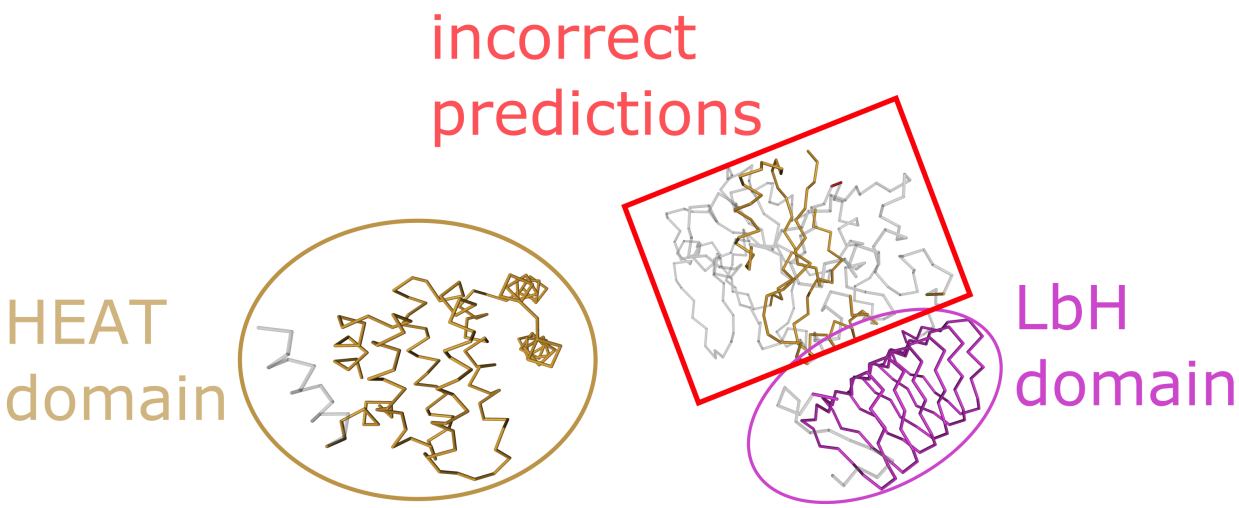


Fig. 34. Prediction by SOLeNNoID for eIF2B subunit epsilon (PDB ID 6O85).
Alpha-solenoid predictions are shown in orange, beta-solenoid predictions are shown in magenta. The HEAT and LbH domains are highlighted, as well as residues incorrectly predicted as alpha-solenoid.

3.4. DISCUSSION

Using SOLeNNoID, I have detected hundreds to thousands of proteins for the different solenoid classes from the largest databases of structural models (AFDB) and experimentally determined structures (PDB), respectively. 50% of residues in a protein must be detected as solenoid - this is a stringent threshold used to remove false positive solenoid protein hits. However, this will also have the effect of under-estimating true solenoid protein numbers as solenoids with partial coverage are omitted. To my knowledge, this is the first attempt to study solenoid percentages and sizes across species and domains and for all three solenoid protein subclasses. In my analysis, alpha-solenoids were the most abundant, followed by alpha/beta-solenoids and finally beta-solenoids. Alpha-solenoid regions were also larger than those of other solenoid classes. Alpha-solenoids and alpha/beta-solenoids were also larger and more frequent in eukaryotes compared to prokaryotes. The distribution across species and sizes was more varied for beta-solenoids with large numbers found in *E. coli* and plants. At least some of the differences across species in terms of solenoid protein distribution and size could be linked to function, as was the case for alpha/beta- and beta-solenoid proteins in plants. I have also provided structural annotation for a number of uncharacterised protein entries from the AlphaFold database, which narrows down the range of possible functions for these proteins. I have discussed the AlphaFold models for important proteins with no or limited experimental structural data and, where possible, compared the models to previously obtained experimental findings. In cases where partial previous experimental information is available, the AlphaFold models enable hypothesis testing (such as for mesothelin and Ycf3), whereas when this information is absent, the AlphaFold models enable hypothesis generation, such as for dgf-1 proteins and their structural similarity to bacterial adhesion proteins.

The PDB search I conducted revealed substantial overlap with existing databases of solenoid proteins but also complementarity of the different methods. Therefore, the best way to address the challenge of finding new solenoid proteins is to combine multiple approaches. I present information about the solenoid entries discovered by SOLeNNoID including the most common structure determination approaches, functions, and species across solenoid classes. Finally, my discussion of examples from the PDB illustrates the wealth of solenoid structures, including a rare case of two different classes of solenoid within the same chain. The solenoid protein databases produced by SOLeNNoID from the AFDB and PDB could be useful for downstream analysis such as structure-based clustering and determination of evolutionary relationships. These findings could also aid the design of novel solenoid proteins by expanding the pool of known solenoid structures. However, alpha-solenoid numbers are likely overrepresented as SOLeNNoID erroneously detects some additional alpha-helical and repetitive structures as alpha-solenoids. Therefore, further training of the model with examples of these structures (e.g. long helices and coiled coils) in the non-solenoid class must be performed.

4.
BETA-SOLENOID PROTEIN DESIGN AND ENGINEERING

4.1. AIMS

My goal was to expand on the work in the original SynRFR design paper (MacDonald et al., 2016) by 1) developing a more automated and generalisable consensus sequence design method, 2) further testing the insertion/fusion tolerance of SynRFR proteins with two-loop variant and N-terminal and internal SpyTag/Catcher fusions, 3) splitting the SynRFR protein into smaller modules and testing them for self-assembly. In this chapter I discuss the progress I have made towards these goals.

4.1.1. CONSENSUS HEXAPEPTIDE DESIGN

The original consensus sequence design approach only utilised the pentapeptide repeats from two proteins. My aim was to automate and streamline this approach. First, I developed a method for structure-based repeat extraction from proteins, inspired by the work of Parra et al (Parra et al., 2013). An MSA can be made using the resulting sequences and refined, leading to a consensus sequence and frequency table. New repeat sequences can then be designed with a simple script, drawing from the frequency table and any other user-defined constraints, and validated using AlphaFold 2. I tested this approach using hexapeptide repeat proteins (a sub-class of beta-solenoid proteins with a triangular cross-section and 18 residues per turn of the beta-helix) as an example fold.

4.1.2. SYNRFR INSERTION/FUSION VARIANTS

4.1.2.1. TWO-LOOP VARIANT DESIGN

The two-loop design was inspired by the beta-hairpin insertion into SynRFR24.1 (Beta1), the natural pentapeptide protein HetL (Ni et al., 2009) (Fig. 35), mini protein design work (Baker et al., 2017) and loop insertion work done in TPR proteins by the Itzhaki group as described above. The aim was to use SynRFR as a stabilising scaffold, onto which two interacting loops can be designed by a fragment-free approach as a modular means of building up complexity.

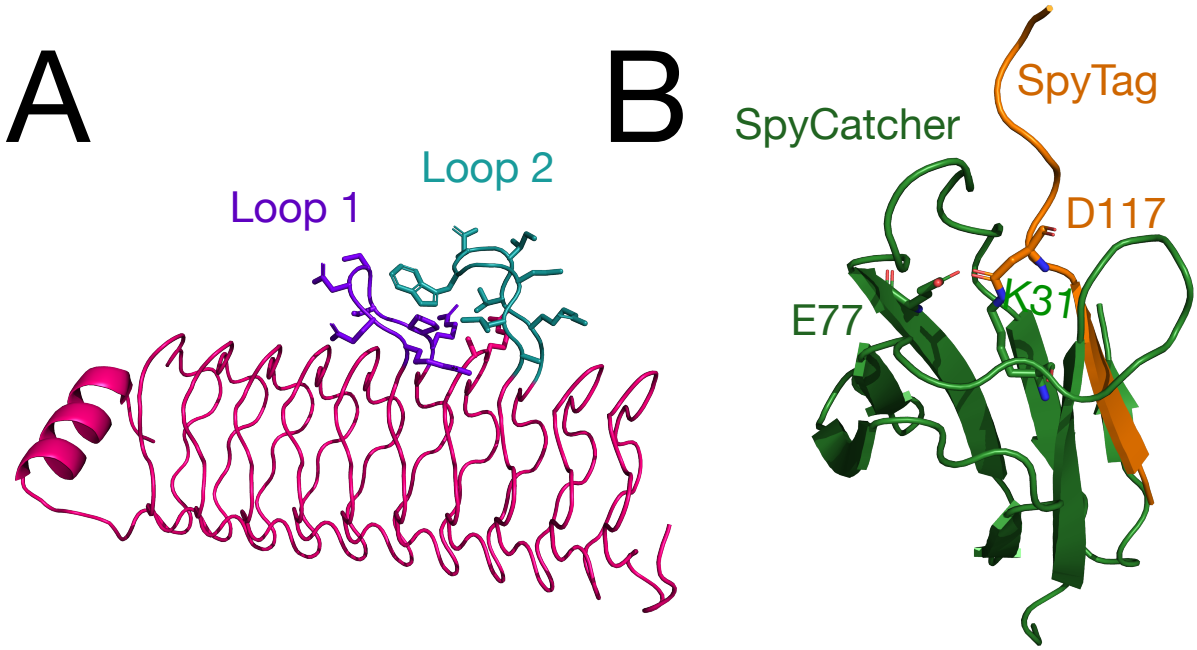


Fig. 35. Example of a natural RFR protein containing two loops - HetL (PDB ID 3DU1) and SpyCatcher/SpyTag (PDB ID 4MLI).
A) HetL. The SynRFR core and capping structure are shown in pink, loop 1 is shown in purple and loop 2 is shown in teal. B) SpyCatcher/SpyTag covalent complex. SpyCatcher is shown in green, SpyTag is shown in orange. The isopeptide bond is formed between K31 on SpyCatcher and D117 on SpyTag. E77 on SpyCatcher catalyses isopeptide bond formation.

4.1.2.2. SPYTAG/CATCHER FUSIONS

The SpyTag/Catcher system (Keeble et al., 2017; Zakeri et al., 2012), derived from the CnaB2 domain of *S.pyogenes* fibronectin-binding protein FbaB, offers a means for covalent attachment of two target proteins. An isopeptide bond forms between an Asp residue on the SpyTag and a Lys residue on the SpyCatcher protein. SpyTags can be placed terminally or internally (Giessen & Silver, 2016; Khairil Anuar et al., 2019; Zakeri et al., 2012). Similar to the work of Correnti et al (Correnti et al., 2020) described above as well as that of P. Q. Nguyen et al (P. Q. Nguyen et al., 2014), my aim was to introduce SpyTag and SpyCatcher domains into the beta-hairpin SynRFR variant to allow the SynRFR protein to be covalently functionalised with an arbitrary domain of

interest. N-terminal SpyTag/Catcher fusions were also designed with a view to building synthetic nano fibres, inspired by the work of Veggiani et al (Veggiani et al., 2016).

4.1.3. SPLIT SOLENOID VARIANTS

The design of self-assembling solenoid modules was inspired primarily by the work of Watson et al (Watson et al., 2014) on reconstituting an ARM protein from split variants. Vetting et al (Vetting et al., 2007) perform a putative ancestral reconstruction of a PRP-coding gene by removing the stop codon between the Np275 and Np276 genes from *N.punctiforme* and compare it to the smaller Np275 gene product, positing that the original gene was split into two fragments during evolution. Further inspiration for these designs came from *E.coli* curli fibres, made up of multiple copies of the CsgA protein (Barnhart & Chapman, 2006; Sleutel et al., 2022) and the rational engineering of beta-solenoid proteins to form amyloid fibrils (Peralta et al., 2015). My aim, however, was to keep the capping structures to prevent the formation of fibrils and make initial steps towards the design of compatible SynRFR building blocks.

4.2. MATERIALS AND METHODS

4.2.1. GENERAL METHODS

4.2.1.1. CLONING, EXPRESSION AND PURIFICATION

100 µg/ml Ampicillin was used for selection in all media. Plasmids were transformed into NEB10β *E. coli* cells and grown overnight on LB agar plates. Colonies were picked and grown overnight in 10 ml LB medium. The colonies were plasmid-miniprepped (NEB Monarch Kit; NEB) and subsequently sequenced (Eurofins Genomics). Sequenced plasmids were transformed into chemically competent KRX cells (Promega). For each variant, between 1 and 4 L of LB medium were inoculated with 10 ml/L of overnight cultures. Cultures were grown at 37 °C until OD_{600} = 0.6. Expression was then induced with 1mM isopropyl β-D-1-thiogalactropyranoside (IPTG) and 0.1% w/v rhamnose. After induction, the cells were grown overnight at 18 °C. The cells were then harvested and resuspended in lysis buffer (100 mM bicine, 150 mM NaCl pH 9.0) and 1 EDTA SIGMAFAST protease inhibitor tablet was added per L of culture (Sigma). The cells were lysed using a cell disruptor and clarified by centrifugation at 35,000 x g for 35 min. Proteins were purified by a batch purification method using nickel-nitrilotriacetic acid (Ni-NTA) resin. The bound proteins were washed with wash buffer (100 mM bicine, 150 mM NaCl, 25 mM imidazole, pH 9.0) and eluted with elution buffer (100 mM bicine, 150 mM NaCl, 250 mM imidazole, pH 9.0). The proteins were further purified by size exclusion chromatography using Superdex HiLoad 16/600 200 pg (Cytiva) at a flow rate of 1 ml/min and amount of protein loaded between 2 and 20 mg. Approximate molecular weights were estimated by running gel filtration standards (Bio-Rad).

4.2.1.2. X-RAY CRYSTALLOGRAPHY

Proteins for which crystallisation was attempted were concentrated to at least 10 mg/ml and used to set up vapour diffusion sparse-matrix crystallisation trials with a Mosquito robot (TTP Labtech). The crystallisation screens used were JCSG+ (Molecular Dimensions) and Wizard 3 and 4 (Rigaku Reagents/Molecular Dimensions). Crystals were cryoprotected by the addition of 30% glycerol or PEG400 and flash-cooled in liquid nitrogen. Diffraction data were collected at Diamond Light Source.

4.2.1.3. MOLECULAR WEIGHT ESTIMATES

Molecular weight calculations for all sequences were performed using ExPASy ProtParam (Walker et al., 2005) with the N-terminal Met residue of the sequence removed.

4.2.2. CONSTRUCT CLONING

4.2.2.1. CONSENSUS HEXAPEPTIDES

Sequences were ordered from Twist Bioscience in a pET29b(+) expression vector.

4.2.2.2. TWO-LOOP DESIGN

The DNA sequence of a part of the SynRFR core with the two loops incorporated was ordered as a linear gBlock from IDT. The pET11a SynRFR24.1 plasmid was linearised by PCR and the final constructs were generated by Gibson assembly (NEB) of the gBlock and the linearised plasmid.

4.2.2.3. SPYTAG/SPYCATCHER DESIGNS

4.2.2.3.1. iST-t1428

The DNA sequence of a part of the SynRFR core with the SpyTag incorporated was ordered as a linear gBlock from IDT. The pET11a SynRFR24.1 plasmid was linearised by PCR and the final constructs were generated by Gibson assembly (NEB) of the gBlocks and the linearised plasmid.

4.2.2.3.2. iSC-t1428 and nSC-beta1

The MBPx-SpyCatcher pET21b plasmid (Veggiani et al., 2016) was ordered from AddGene. The SpyCatcher region was amplified by PCR. The t1428 SynRFR variant and the SynRFR24.1 plasmids were linearised by PCR and the final constructs were generated by Gibson assembly of the SpyCatcher coding sequence and the linearised plasmids.

4.2.2.3.3. nST-beta1

The sequence for nST-beta1 was ordered from Twist Bioscience in a pET29b(+) expression vector.

4.2.2.4. SPLIT SOLENOID DESIGNS

The pET11a SynRFR plasmid was linearised and part of coding sequence was removed by PCR to produce part A1,A2 or B. The circular plasmid was then reformed by reaction with KLD enzyme mix (NEB). A C-terminal thrombin cleavage site and 6xHis-Tag sequence was added to the pET11a plasmid encoding part B by PCR and KLD.

4.2.3. CONSTRUCT SEQUENCES

4.2.3.1. CONSENSUS HEXAPEPTIDE SEQUENCES (NUMBER CORRESPONDS TO DESIGN METHOD, CAPPING SEQUENCES IN BOLD)

>Hexapeptide_1 MW - 18860.64 Da

MGSSHHHHHHSSGLVPRGSH**MASRLERVYQSEQAEKLLLAGVMLRDPARFDLRGTLTHGRDVE**I
GDGVDIGNHVTIGTHVVIGNGVIIGRQVQIGADVLIGAQVLIGIGVIIGHGVCIGEDVTIGQYVMIGAQVEI
GMGVIIGAGVT**LEPRSAAIGVTIPDGRYIPAGMVVTSQAEADKLPE**

>Hexapeptide_2 MW - 19155.71 Da

MGSSHHHHHHSSGLVPRGSH**MASRLERVYQSEQAEKLLLAGVMLRDPARFDLRGTLTHGRDVE**I
GSHVLIGANVHIGDGVRIGPGVKIGSNVEIGDNVCIGHQVSIGSDVYIGDSVDIGMFVHIGNKVIIGANVH
IGSNVTIGYNVF**LEPRSAAIGVTIPDGRYIPAGMVVTSQAEADKLPE**

>Hexapeptide_3 MW - 19440.21 Da

MGSSHHHHHHSSGLVPRGSH**MASRLERVYQSEQAEKLLLAGVMLRDPARFDLRGTLTHGRDVE**F
GHNVYILINAYIGDRVLVGEGSTIGDRVRISANTHVATMCVIIADVALGSTLKLVKRTMLGDGAEVGINNE
FGANVIIGKFSI**LEPRSAAIGVTIPDGRYIPAGMVVTSQAEADKLPE**

>Hexapeptide_4 MW - 19304.98 Da

MGSSHHHHHHSSGLVPRGSH**MASRLERVYQSEQAEKLLLAGVMLRDPARFDLRGTLTHGRDVE**L
GSDTYIGPGVKIGPDVFILAFAVLASGSAIHMGCYIGDNAVIADRIHLKPNGVLGSLSQIGDGVEIYEDVH
LGHGVIIHLYAI**LEPRSAAIGVTIPDGRYIPAGMVVTSQAEADKLPE**

The above sequences were ordered from Twist Bioscience in a pET29b(+) expression vector.

4.2.3.2. TWO-LOOP SEQUENCE (LOOPS IN BOLD)

>2-loop **MW - 26810.27 Da**

MGSSHHHHHHSSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHADL
TAADLRSADLRNANLSHANLG**TEDLEPNALP**GANLEEANLEAANLRGADLKGANLSGADLQEANLT
QANLG**ADETPEDVQF**AANLSDANLEQADLAGADLAFANLDGANLHGANLNNANLQNAMLTRANLE
QADLSGARTTGARLDDADLRGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCLAGGSGC

4.2.3.3. SPYTAG/SPYCATCHER CONSTRUCT SEQUENCES (SPYTAG/CATCHER SEQUENCES IN BOLD)

>nSC-beta1 **MW - 37593.08 Da**

MGSSHHHHHHSSGLVPRGSHGSG**MVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDEDGKELA GATMELRDSSGKTISTWISDGQVKDFYLYPGKYTFVETAAPDGYEVATAITFTVNEQGQVTVNGKA TKGDAHI**SGGGGMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHANLSRANLRSA DLRNANLSHANLSGANLEEANLEAANLRGADLHEANLSGADLQEANLTQANLKDANLSDANLEQAD LAGADLQGAVLDGANLHGANLNNANLSEAMLTRANLEQADLSGARTTGARLDDADLRGATVDPVLW RTASLVGARVDVDQAVAFAAAHGLCLAGGSGC

>nST-beta1 **MW - 26949.53 Da**

MGSSHHHHHHSSGLVPRGSHGSG**AHIVMVDAYKPTK**GSGESGMNVGEILRHYAAGKRNFQHINLQ EIELTNASLTGADLSYANLHHANLSRANLRSADLRNANLSHANLSGANLEEANLEAANLRGADLHEAN LSGADLQEANLTQANLKDANLSDANLEQADLAGADLQGAVLDGANLHGANLNNANLSEAMLTRANL EQADLSGARTTGARLDDADLRGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCLAGGSGC

>iSC-t1428 **MW - 38415.11 Da**

MGSSHHHHHHSSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHANL SRANLRSADLRNANLSHANLSGANLEEANLEAANLRGANLVGANLSGADLQEANLTQAMLGQTYPG GGGS**MVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDEDGKELAGATMELRDSSGKTISTWISD GQVKDFYLYPGKYTFVETAAPDGYEVATAITFTVNEQGQVTVNGKATKGDAHI**SGGGGWGTVGAD LSDANLEQADLAGAQLVGAVLDGANLHGANLNNANLDNAMLTRANLEQADLSGARTTGARLDDADL RGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCLAGGSGC

>iST-t1428 **MW - 27612.43 Da**

MGSSHHHHHHSSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHANL SRANLRSADLRNANLSHANLSGANLEEANLEAANLRGANLVGANLSGADLQEANLTQAMLGQTYPG GGGS**AHIVMVDAYKPTK**SGGGGWGTVGADLSDANLEQADLAGAQLVGAVLDGANLHGANLNNAN LDNAMLTRANLEQADLSGARTTGARLDDADLRGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCL AGGSGC

4.2.3.4. SPLIT SOLENOID SEQUENCES

>part A1 **MW - 12742.87 Da**

MGSSHHHHHHSSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHANL
SRANLRSADLRNANLSHANLSGANLEEANLEAANLRGADLHEANLSGADLQE

>part A2 **MW - 12613.75 Da**

MGSSHHHHHHSSGLVPRGSHMNVGEILRHYAAGKRNFQHINLQEIELTNASLTGADLSYANLHHANL
SRANLRSADLRNANLSHANLSGANLEEANLEAANLRGADLHEANLSGADLQ

>part B **MW - 12094.32 Da**

MANLTQANLKDANLSDANLEQADLAGADLQGAVLDGANLHGANLNNANLSEAMLTRANLEQADLSG
ARTTGARLDDADLRGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCLAGGSGC

>part BT **MW - 13758.11 Da**

MANLTQANLKDANLSDANLEQADLAGADLQGAVLDGANLHGANLNNANLSEAMLTRANLEQADLSG
ARTTGARLDDADLRGATVDPVLWRTASLVGARVDVDQAVAFAAAHGLCLAGGSGCLVPRGSGSSHH
HHHH

4.2.4. SPYTAG/CATCHER REACTION

The SpyCatcher/Tag reaction was carried out for 2h at 10 μM concentration in 100 mM bicine 150 mM NaCl pH 9.0. At 2h, a sample of the reaction mixture for each reaction was taken and heated in SDS loading buffer for 7 min in a 95 °C heating block to terminate the reaction. The results were visualised on an 4-20% gradient Tris-Gly polyacrylamide gel (SERVA) by SDS-PAGE. For native PAGE, samples of the reaction mixtures were taken in loading buffer without SDS and visualised on a 4-20% gradient Tris-Gly polyacrylamide gel (SERVA) using 250 mM Tris 190 mM glycine.

4.3. RESULTS

4.3.1. CONSENSUS HEXAPEPTIDE DESIGN

4.3.1.1. REPEAT EXTRACTION

As a first step towards consensus design, I wrote a Python3 script(van Rossum & Drake, 2009) , which automatically finds a set of repeats in a repeat protein with the lowest inter-repeat RMSD. The intuition behind this is to try to capture the most structurally 'ideal' set of repeats from each protein. Parsing through many structures in this way then allows the production of a library of repeats, which can be used for consensus design downstream.

Using the repeat extraction script, a protein PDB file was parsed, and beta carbon 3D coordinates extracted using BioPython. Euclidean distances between beta carbon atoms were calculated using NumPy and contacting beta carbon pairs were selected. Two beta carbon atoms were defined as 'in contact' if the Euclidean distance between them was < 10 Å and the distance between them along the primary sequence was > 10 amino acids. For all contacting pairs, their distance along the primary sequence was recorded as a contact length and the 15 most frequent contact lengths are saved. For each of these contact lengths, all continuous fragments of that length were generated using a sliding window approach. Subsequently, all-vs-all superposition of all fragments of a given contact length was performed using SVDSuperimposer in BioPython. The result was a Pandas DataFrame with the spans of all pairs of repeats as well as the RMSD value between the two repeats in each pair. All entries with an RMSD > 4 Å were discarded. Then, repeat 'reading frames' of non-overlapping repeats were built up using this information. For example, if residues 103-123 match 144-164, and 144-164 match 184-204, the full reading frame would contain repeats 103-123, 144-164 and 184-204. Each reading frame has an associated repeat length, the residue spans of individual repeats, the total number of amino acids in the reading frame and the average RMSD for all repeats in the reading frame (Fig.36). Finally, a contact length of interest was selected (e.g. 18 in the case of hexapeptide repeat proteins) and repeat amino acid sequences for the reading frame with the lowest average RMSD for that tile length were output as a .fasta file.

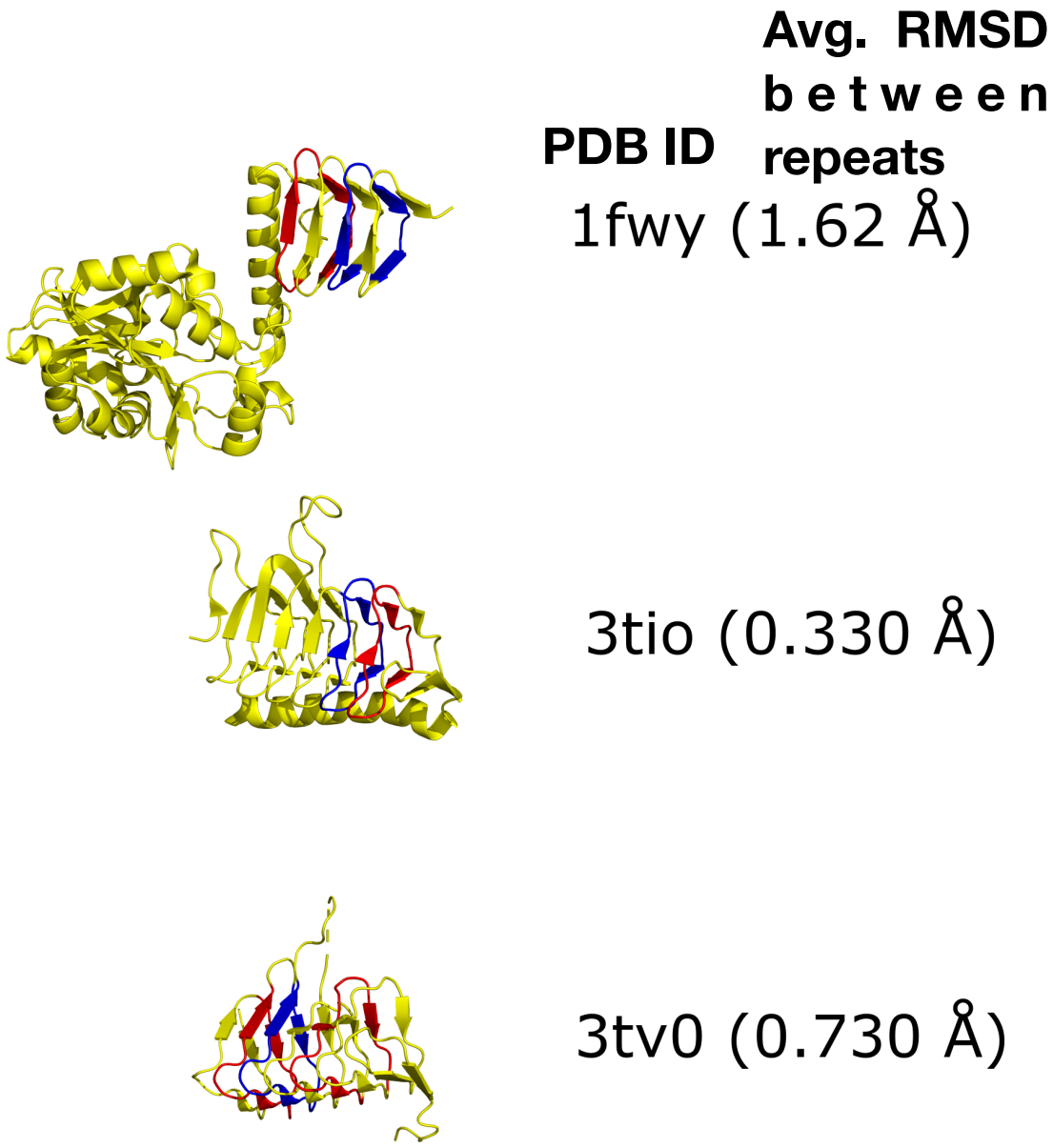


Fig. 36. Repeat extraction example output for three hexapeptide repeat proteins.
The entire protein is shown in yellow and 18-residue repeats selected by the repeat extraction script are shown in alternating red and blue for ease of visualisation. For each protein the highlighted repeats denote a specific repeat 'reading frame', in this case the repeat 'reading frame' with the lowest average inter-repeat RMSD. The value in the brackets denotes the average inter-repeat RMSD for the repeats selected by the repeat extraction script. The proteins comprise PDB ID 1FWY (Brown et al., 1999), 3TIO (Park et al., 2012) and 3TV0 (Derewenda et al., to be published).

4.3.1.2. SEQUENCE PROCESSING

A single chain was selected for each hexapeptide repeat protein PDB ID from 127 hexapeptide repeat protein entries in the RepeatsDB database. The structures were processed using the repeat extractor described above and 258 unique 18-residue repeat sequences were saved into a single .fasta file. An MSA was then produced using MUSCLE (Edgar, 2004). The sequence logo generated using WebLogo (Crooks et al., 2004; Schneider & Stephens, 1990) was inspected and hexapeptide stretches in the MSA were selected to produce a new MSA of hexapeptide repeats, comprising 344 unique 6-residue repeats. The MSA of the 6-residue repeats was used to construct a residue frequency table. The sequence logo generated from the 6-residue repeat sequences is shown below in Fig. 37. The amino acids isoleucine, glycine and valine appeared with a frequency of 0.52, 0.49 and 0.29 at positions 1,2 and 5, respectively. The consensus hexapeptide sequence was therefore IGxxVx. Fig. 37 B illustrates the structural context of each position in the consensus sequence. Conserved positions 1 and 5 form the hydrophobic core of the protein, while position 2 is part of the Type I' A (de Brevern, 2016) beta-turn, which changes the direction of the polypeptide chain.

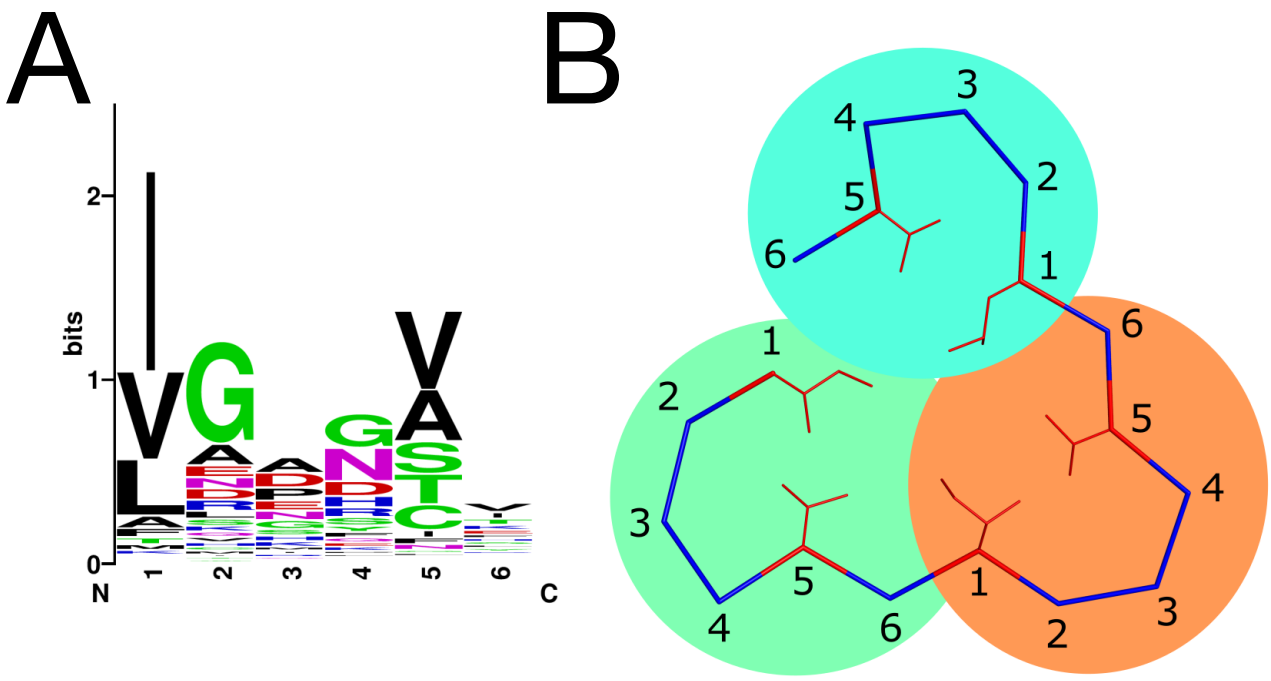


Fig. 37. Hexapeptide repeat sequence logo and structure.
A) Hexapeptide repeat sequence logo generated using WebLogo. The Y-axis shows information content in bits. Information content is related to probability. The more probable an event is, the less information it contains. Hence, conserved residues have a higher information content in bits. B) Hexapeptide repeat structure. Each hexapeptide region is highlighted and numbered according to the numbering scheme from A). Hydrophobic core residues are coloured red and shown in line representation using PyMOL.

4.3.1.3. NEW HEXAPEPTIDE SEQUENCE GENERATION

Analogous to the SynRFR design approach, capping sequences were selected from natural proteins. The N-terminal cap was derived from PDB ID 3TWD (Buurman et al., 2011) (residues 232-273 in PDB numbering), while the C-terminal cap was derived from PDB ID 1THJ (Kisker et al., 1996) (residues 150-184 in PDB numbering). A central 84-residue consensus hexapeptide sequence was then generated by applying one of the four hexapeptide generation methods outlined in Table 12 14 times. In method 1, positions 1, 2 and 5 are fixed as I, G and V, respectively, following the consensus sequence for the hexapeptide motif. The remaining residues are chosen randomly, biased by the frequency table obtained in the previous section. Method 2 is similar to method 1 but also specifies that no two adjacent positions can be identical within a given hexapeptide sequence. Method 3 allows all positions to be chosen randomly biased by the frequency table. Method 4 is similar to method 3 but also specifies that no two adjacent positions can be identical within a given hexapeptide sequence.

	Hexapeptide Position					
Method	**1**	**2**	**3**	**4**	**5**	**6**
1	**I**	**G**	Draw from frequency table	Draw from frequency table	**V**	Draw from frequency table
2	**I**	**G**	• Draw from frequency table • Amino acid identity cannot be the same as positions 2 and 4	• Draw from frequency table • Amino acid identity cannot be the same as positions 3 and 5	**V**	• Draw from frequency table • Amino acid identity cannot be the same as position 5
3	Draw from frequency table	Draw from frequency table	Draw from frequency table	Draw from frequency table	Draw from frequency table	Draw from frequency table
4	• Draw from frequency table • Amino acid identity cannot be the same as position 2	• Draw from frequency table • Amino acid identity cannot be the same as positions 1 and 3	• Draw from frequency table • Amino acid identity cannot be the same as positions 2 and 4	• Draw from frequency table • Amino acid identity cannot be the same as positions 3 and 5	• Draw from frequency table • Amino acid identity cannot be the same as positions 4 and 6	• Draw from frequency table • Amino acid identity cannot be the same as position 6

Table 12. Methods used for the design of consensus hexapeptide sequences.

The full sequence was generated by combining the N- and C-terminal caps with the generated consensus hexapeptide sequence (Fig. 38 B) 50 sequences were generated by each of the 4 methods outlined above.

An example 6-turn (or 18-hexapeptide) sequence was generated according to method 1, outlined in Section 4.1.2. Structure prediction was carried out for this sequence using both ColabFold (Mirdita et al., 2022) and trRosetta (J. Yang et al., 2020). In both cases, the top model had a hexapeptide structure. However, trRosetta and ColabFold differ in the handedness of the predicted structure (Fig. 38 A). A BLAST search (Altschul et al., 1990) against the PDB with the

generated sequence returned exclusively left-handed beta-helical structures. Further structure prediction attempts with AlphaFold 2 and RoseTTAFold (Baek et al., 2021) also resulted in a left-handed hexapeptide structure. Hence, the ColabFold handedness prediction is likely incorrect. An AlphaFold 2 model incorporating the caps and another generated hexapeptide sequence is shown in Fig. 38 B. This model showcases the intended design and position of the capping structures.

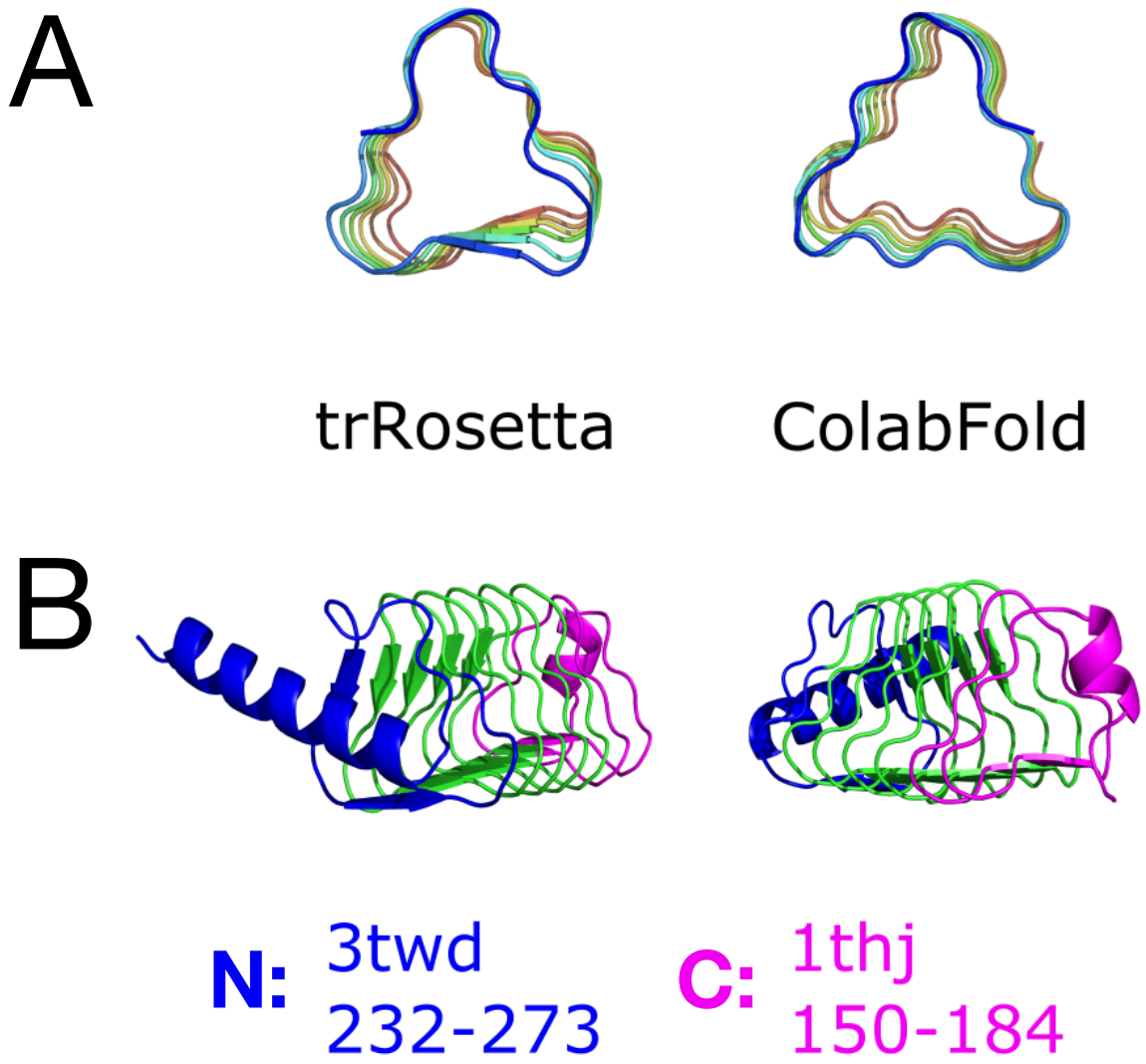


Fig. 38. Structure prediction models for the consensus sequence and cap structures.
A) trRosetta and ColabFold predictions for an example sequence generated by method 1. Structures are coloured N-to-C. trRosetta predicts a left-handed beta-helix, while ColabFold predicts a right-handed beta-helix. B) AlphaFold 2 prediction for a full sequence comprising an N-terminal cap (residues 232-273 from PDB ID 3TWD) coloured blue, a generated consensus hexapeptide sequence coloured green, and a C-terminal cap (residues 150-184 from PDB ID 1THJ) coloured magenta.

4.3.1.4. COLABFOLD MODELS OF DESIGNED SEQUENCES

50 sequences were generated by each method outlined above and structure predictions were made using the ColabFold implementation (Mirdita et al., 2022) of AlphaFold 2 (Jumper et al., 2021) with the use of MSAs allowed. ColabFold uses the AlphaFold 2 model but replaces the multiple sequence alignment generation step from the original method with MMseqs (Hauser et al., 2016). The top models for each sequence were ranked by average pLDDT, and average pAE over all residues, and pTM. pLDDT is a metric recording model confidence in a prediction. Therefore, a model with high average pLDDT will have high-confidence structure predictions across the entire structure. pAE is a metric recording the expected position error at a given residue when predicted and true structures are aligned on a different residue. A model with low average pAE will therefore consist of a cohesive, single domain. pTM is the predicted TM-score on aligning the predicted and true structures. Models with high pTM again signal high confidence in the structure prediction.

A similar distribution was observed for all metrics (Fig. 39) (inverted in the case of average pAE as lower values indicate greater confidence). Sequences designed by methods 1 and 2 had the fewest high-scoring models (above 0.9 as measured by pLDDT and above 0.8 as measured by pTM), while methods 3 and 4 produced the highest-scoring models. Method 4 produced the largest number of high-scoring models. For example, 125 models in total had an average pLDDT score above 0.9. Of these sequences, 18.4% were designed by method 1, 16.8% by method 2, 29.6 % by method 3 and 35.2% by method 4.

ColabFold produces 5 structural models for each sequence. The average pLDDT over all residues in the top-ranked model out of 5 was chosen as the criterion to pick the top-scoring sequence designed by each method. Hence, in total four sequences were chosen for experimental characterisation - one designed by each method. Table 12 details metrics calculated over the 5 structural models produced for each sequence by ColabFold. Hexapeptide 3 and 4 had the best metrics and displayed the most consistency across all 5 models, whereas hexapeptide 1 and 2 had models with variable pAE and pLDDT. Some hexapeptide 1 and 2 models matched the design structure, while others had significant deviations from it.

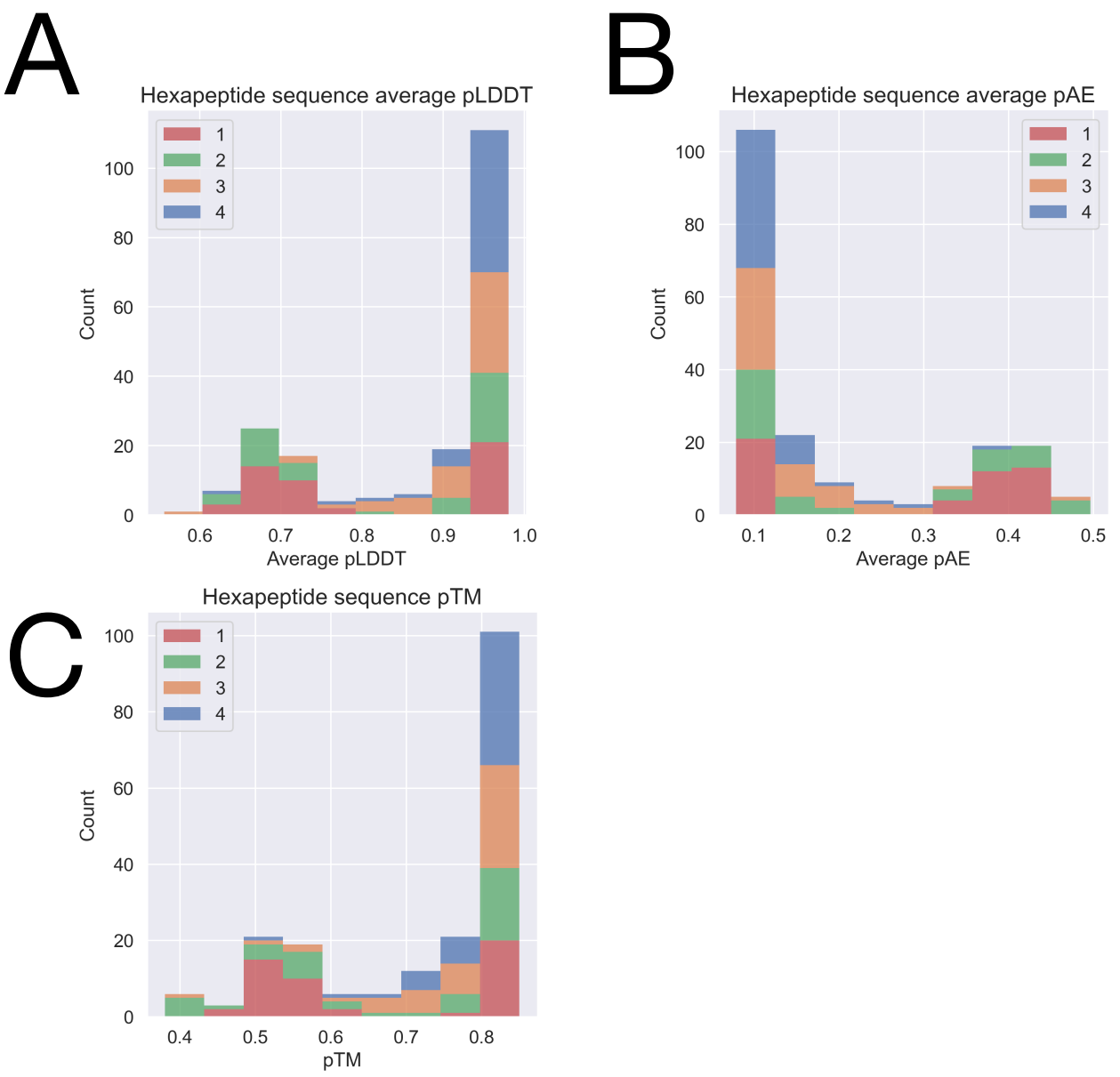


Fig. 39. AlphaFold structure prediction statistics for hexapeptide sequences generated by each of the four methods.
Each subfigure is a stacked count plot. Data on sequences generated using method 1 is coloured red, method 2 - green, method 3 - orange and method 4 - blue. High average pLDDT and high pTM, and low average pAE are desirable. In all metrics, the largest proportion of well-scoring sequences (average pLDDT > 90, or average pTM > 0.8, or average pAE < 0.15) was due to designs from methods 3 and 4. A) Average pLDDT over all residues across sequence design methods. B) Average pAE over all residues across sequence design methods. C) pTM across sequence design methods.

	Mean PAE (Å)	Mean pLDDT	Mean RMSD to best model (Å)
Hexapeptide 1	8.90	77.6	5.32
Hexapeptide 2	11.2	69.2	9.59
Hexapeptide 3	3.37	96.1	0.192
Hexapeptide 4	**2.79**	**97.5**	**0.174**

Table 13. Metrics calculated over the 5 models produced for each sequence by ColabFold. PAE and pLDDT are averaged over the 5 models. Mean RMSD is calculated using the RMSD of the remaining four models to the best model.

4.3.1.5. EXPERIMENTAL CHARACTERISATION

4.3.1.5.1. NI-NTA PURIFICATION

No soluble expression was observed for hexapeptide 1, while hexapeptide 2, 3 and 4 were expressed and could be purified. For hexapeptide 2, 3 and 4 the observed bands were consistent with the predicted molecular weights of around 19 kDa. The yield of hexapeptide 2 was greater than that of hexapeptide 3 and 4. However, a double band was observed for hexapeptide 2, possibly due to proteolytic cleavage.

Hexapeptide 1

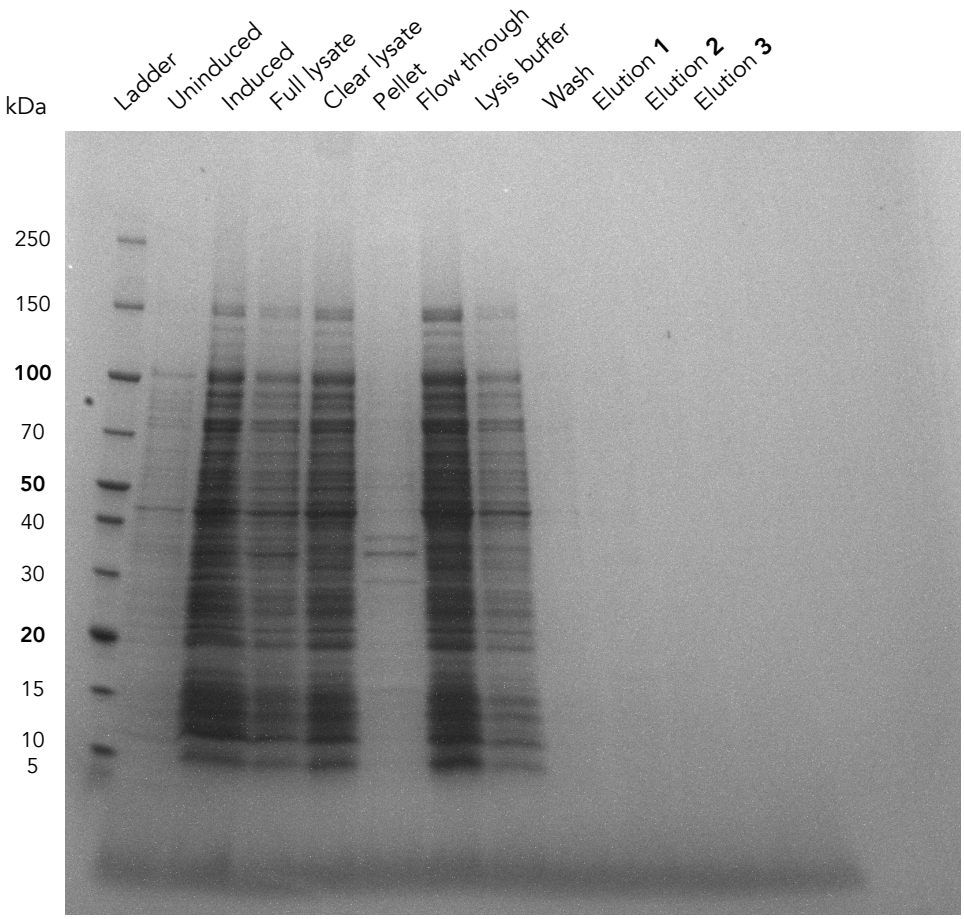


Fig. 40. Ni-NTA purification gels for hexapeptide 1-4 proteins.
Uninduced - sample from the bacterial culture prior to induction with rhamnose and IPTG. Induced - sample after induction. Full lysate - sample after lysis by cell disruption. Clear lysate - sample of supernatant after ultracentrifugation of full lysate. Pellet - sample of pellet after ultracentrifugation of full lysate. Flow through - sample of flow-through after incubation with Ni-NTA resin. Lysis buffer - sample of resin wash with lysis buffer (no imidazole). Wash - sample of resin wash with wash buffer (25 mM imidazole). Elution - elution samples. Arrows pointing to band(s) of expected molecular weight for each construct.

Hexapeptide 2

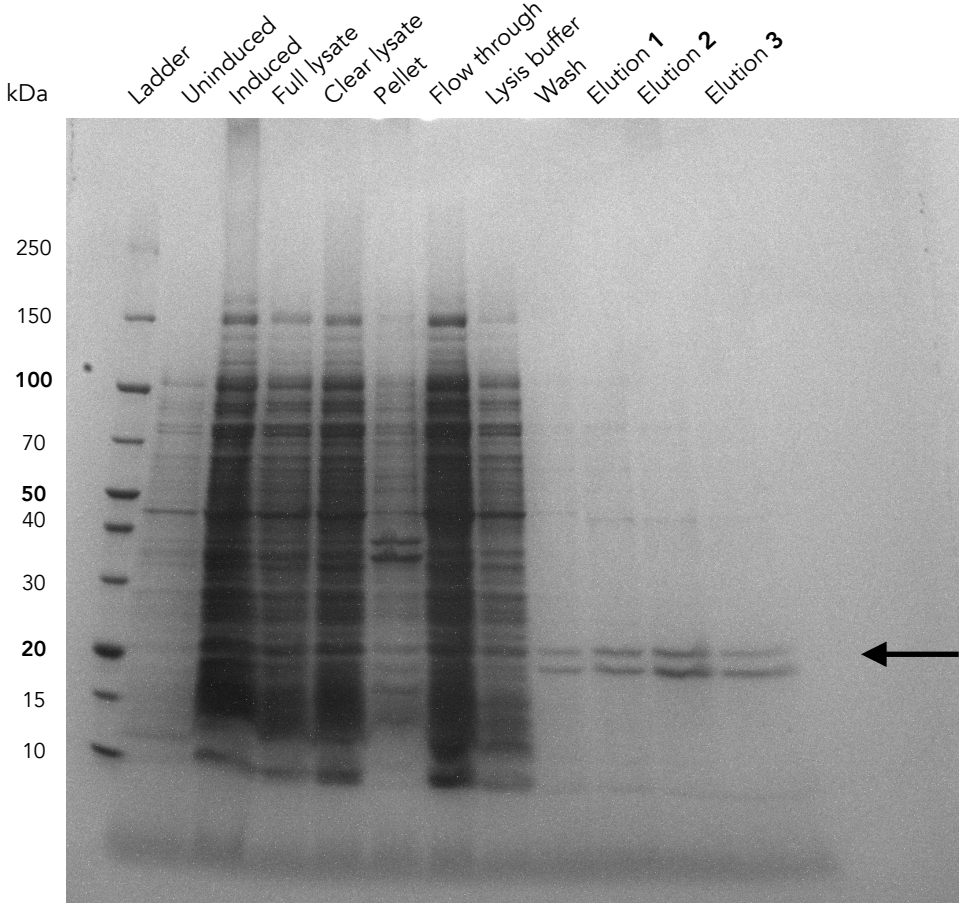


Fig. 40. Ni-NTA purification gels for hexapeptide 1-4 proteins, continued.

Hexapeptide 3

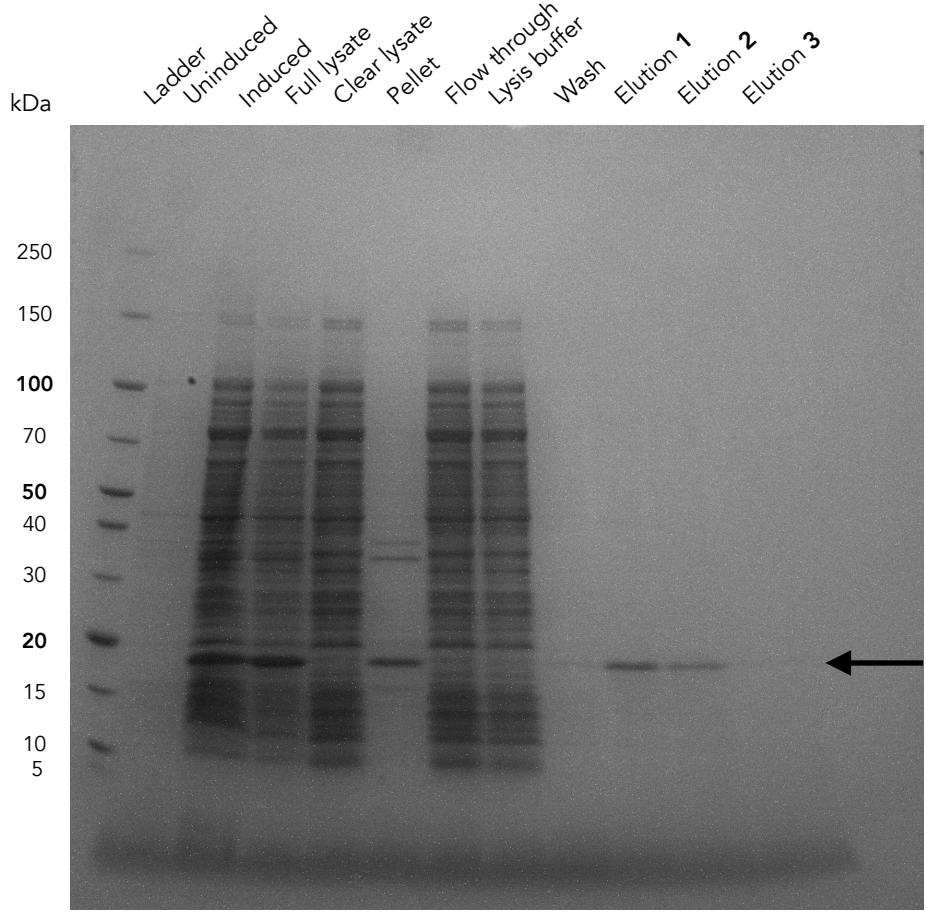


Fig. 40. Ni-NTA purification gels for hexapeptide 1-4 proteins, continued.

Hexapeptide 4

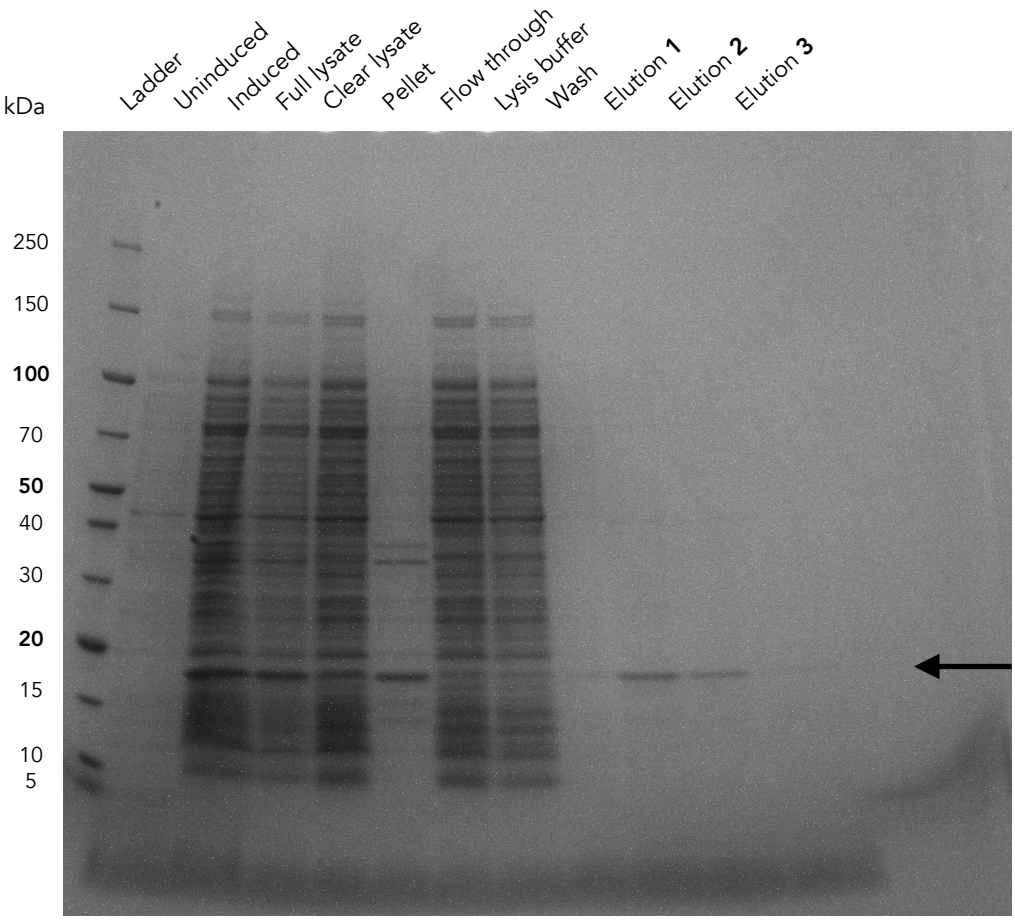


Fig. 40. Ni-NTA purification gels for hexapeptide 1-4 proteins, continued.

4.3.1.5.2. SIZE EXCLUSION CHROMATOGRAPHY

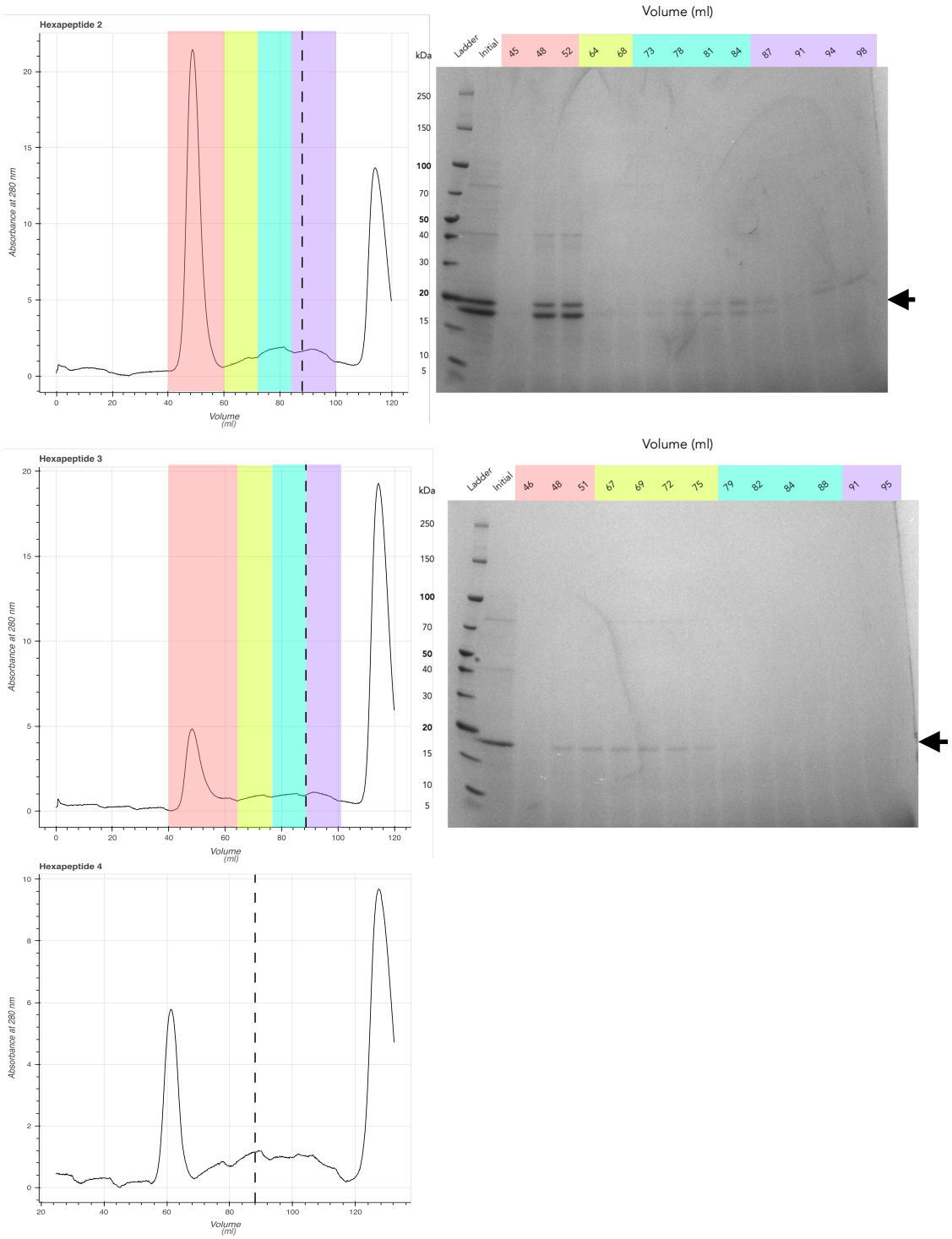


Fig. 41. Size exclusion chromatography data for hexapeptide 2,3 and 4.
A), B) Chromatogram and accompanying SDS-PAGE gel of elution fractions with the approximate elution volume recorded on top of each gel lane. Fractions are colour-coded according to the peaks in the chromatogram. The dashed line denotes the approximate expected elution volume based on the predicted molecular weight of the protein and a standard curve using protein standards of known molecular weight. C) Chromatogram for Hexapeptide 4. Arrows pointing to band(s) of expected molecular weight for each construct.

Size exclusion chromatograms were obtained for hexapeptides 2,3 and 4. SDS-PAGE gels were run for the size exclusion elution fractions for hexapeptides 2 and 3 but not hexapeptide 4 due to time constraints. In all cases a similar elution profile was observed. There was a significant aggregate peak (the protein appears in the void volume), and no clear bands were observed at the expected elution volume. This could be due to smaller soluble aggregates or due to some disulfide bond formation as the hexapeptide 2,3 and 4 sequences all contain a single cysteine residue. For hexapeptide 2, most of the protein sample was in the aggregated fraction but some was found at larger elution volumes, including at the approximate expected elution volume for the protein based on its molecular weight. The expected elution volume is calculated using a calibration curve from running protein standards of known molecular weight on the S200 column. For hexapeptide 3, bands were observed only for low elution volume (and correspondingly high molecular weight) fractions.

4.3.2. TWO-LOOP (2_LOOP) VARIANT DESIGN

The goal in the two-loop design was to design two structured loops, which can pack against each other using a fragment-free backbone design approach. If the design is successful, the two loops can be used to scaffold a functional site, e.g. a metal binding site.

4.3.2.1. BACKBONE DESIGN

Using the sequence with PDB ID 4YC5 as the base scaffold, two ten-residue loops separated by one turn of the SynRFR beta helix were introduced between residues 88 and 89, and 128 and 129, respectively. The loop backbone structures were designed using the fragment-free generalized kinematic closure protocol (Bhardwaj et al., 2016) in the Rosetta3 suite (Leaver-Fay et al., 2011). A modified energy function with a down-weighted Lennard-Jones potential term and upweighted hydrogen bonding term was used as backbone-backbone hydrogen bonds are more important than side chain packing at the backbone conformation sampling stage. All backbone omega angles in the loop were set to 180° to avoid cis proline conformers. The peptide bond geometry between the middle residues of each loop was set to ideal values and both loops were represented as polyalanine stretches to allow close packing of the loops against each other. Selection criteria for kinematic closure solutions enforced a maximum distance of 15 Å between the middle residues of the first loop and the middle residues of the second loop and a maximum width of each loop of 8 Å. These criteria were chosen to bias loops towards a more compact conformation. 164,000 loop closure attempts resulted in 6699 structures which satisfied the requirements.

Next, all backbone-only structures were relaxed with the ref2015 Rosetta all-atom energy function (Alford et al., 2017) and subjected to energy-based clustering. In this clustering approach the lowest energy structure is defined as the first cluster centre and removed from the unclustered list. Further structures are added to the cluster, if they fall within the specified RMSD cut-off relative to the cluster centre, otherwise a new cluster centre is defined, and the process repeated. No unclustered structures remained at the end. The clustering cut-off was set to 1.25 Å, yielding 328 cluster centres.

4.3.2.2. SEQUENCE DESIGN

10, 000 sequences were designed for each of the top 20 cluster centres using the Rosetta FastDesign protocol. The energy function was modified to penalise more than two identical residues in a row and to have 1.5 times increase in hydrogen bonding strength. Due to the observed biases towards aromatic amino acids, proline and glycine, composition constraints were defined with a maximum of 2 prolines and glycines per loop and a maximum of 1 Trp and 1 Tyr per loop. Amino acid identities in the loop region excluded CHKMRS. Cysteine was avoided due to potential for intra- intra or intermolecular disulfide bonds, whereas lysine and arginine were removed due to an apparent bias towards them in unrestrained sequence design trials. SynRFR residues in close proximity to the loops were also redesigned.

All 100 lowest energy designs (in Rosetta Energy Units) were derived from cluster centre 19. Therefore, two sequence designs from backbone cluster centre 19: 19_04242 and 19_05903 (Fig. 42) were selected for *in silico* validation The design with the lowest total score was 19_05903 and the design with the lowest loop analyser mover (LAM) score was 19_04242. The LAM score is a loop quality metric in Rosetta.

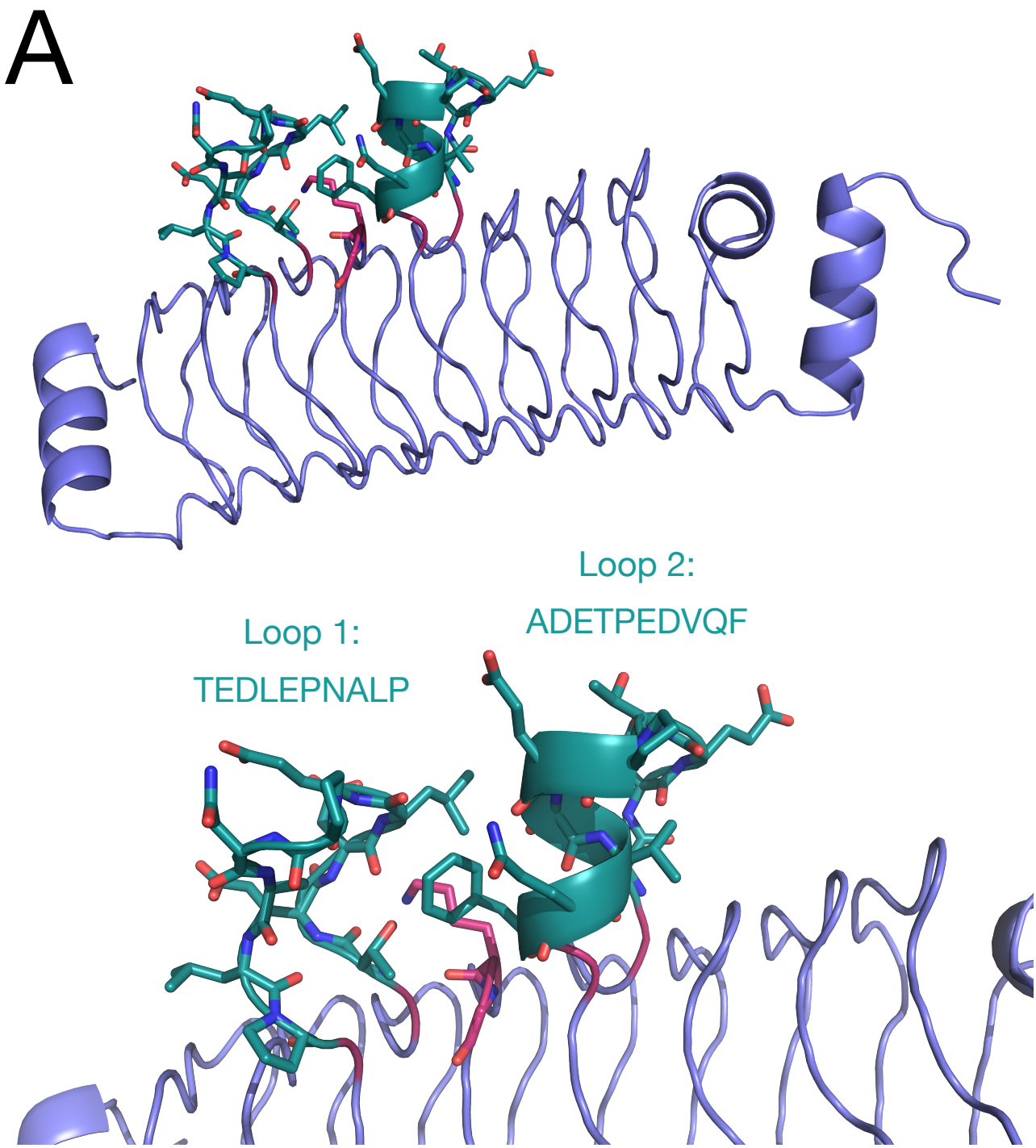


Fig. 42. Rosetta macromolecular models for two-loop designs.
A) Design 19_04242, B) Design 19_5903 (B). Purple - components of the original starting construct, Fuchsia - residues on either side of the loop insertions and residues at the vertex of the intervening turn of the beta-helix, Teal - loop insertions. The sequence of each loop is shown above it.

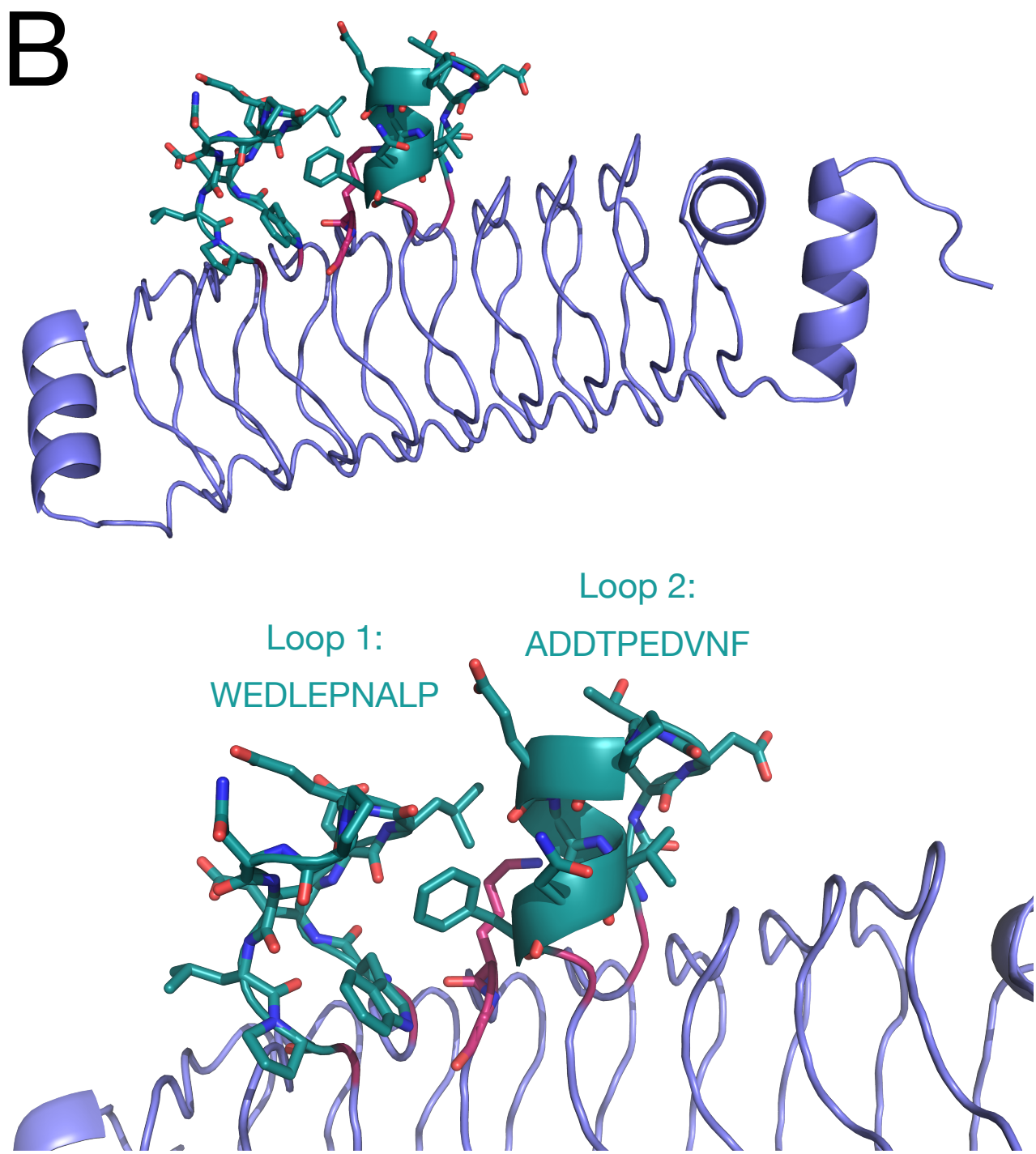


Fig. 42. Rosetta macromolecular models for design 19_04242 (A) and design 19_5903 (B), continued.

4.3.2.3. IN SILICO VALIDATION

1000 independent folding trajectories of the loop sections of the 19_04242 and 19_5903 designs were run using next-generation KIC (Stein & Kortemme, 2013). Cα RMSD was calculated for each loop relative to the designed structure and the overall RMSD score was obtained by summing the RMSD values of the two loops. The assumption was made that the SynRFR scaffold is stably folded so only the loops need to be refolded. This is because the Rosetta *ab initio* relax protocol uses fragment assembly, which itself relies on secondary structure prediction. The solenoid beta helix has a non-standard beta secondary structure, resulting in poor fragment assembly solutions, even for the original SynRFR24.1 protein. In all cases the ref2015 score function, or modified versions thereof were used (Alford et al., 2017). The Imperial College High Performance Computing cluster was used for protein design calculations.

As shown in Fig. 43, a folding funnel is observed for both sequences with conformations near the intended design having the lowest energy. Further analysis of the folding trajectories from next-generation kinematic closure (Table 13) revealed that design 19_04242 has a larger energy gap relative to the lowest energy trajectory with an RMSD greater than 1 Å and a larger percentage of trajectories sampling sub- Å near-native conformations compared to design 19_05903. Therefore, 19_04242 was chosen for experimental characterisation as these metrics suggest that the designed sequence preferentially folds into the designed conformation. The structural features of both designs can be seen in Fig. 42. The two loops interact by via the stacking of a Leu residue on loop 1 and a Phe residue on loop 2. Each loop contains backbone-backbone, backbone-side-chain and side-chain-side-chain hydrogen bonds, in addition to hydrogen bonds to the solenoid scaffold.

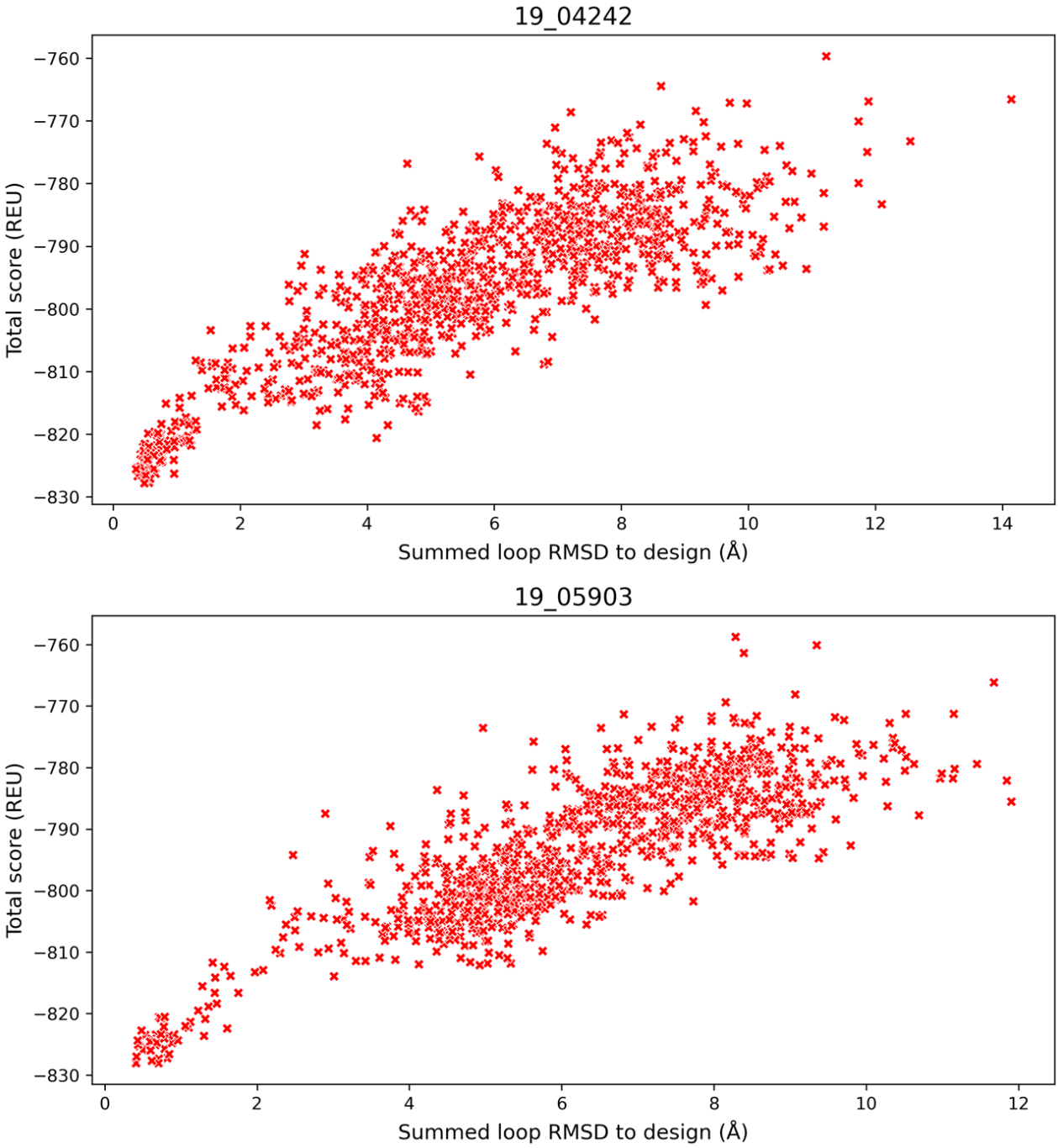


Fig. 43. Next-generation KIC energy landscapes for designs 19_04242 and 19_05903. A folding funnel is observed towards the designed loop conformations.

Structure	Percentage of sub-Å trajectories	Energy gap (REU)
19_04242	6.7	6.3
19_05903	2.8	3.6

Table 14. Energy landscape features for designs 19_05903 and 19_04242.

Disorder prediction with IUPred3 (Erdos et al., 2021) using the 19_04242 sequence suggested that the loop insertions had a greater disorder propensity than the original starting sequence as shown in Fig. 44. In addition, all loop residues had a score above 0.5, which classifies them as disordered (Dosztányi et al., 2005).

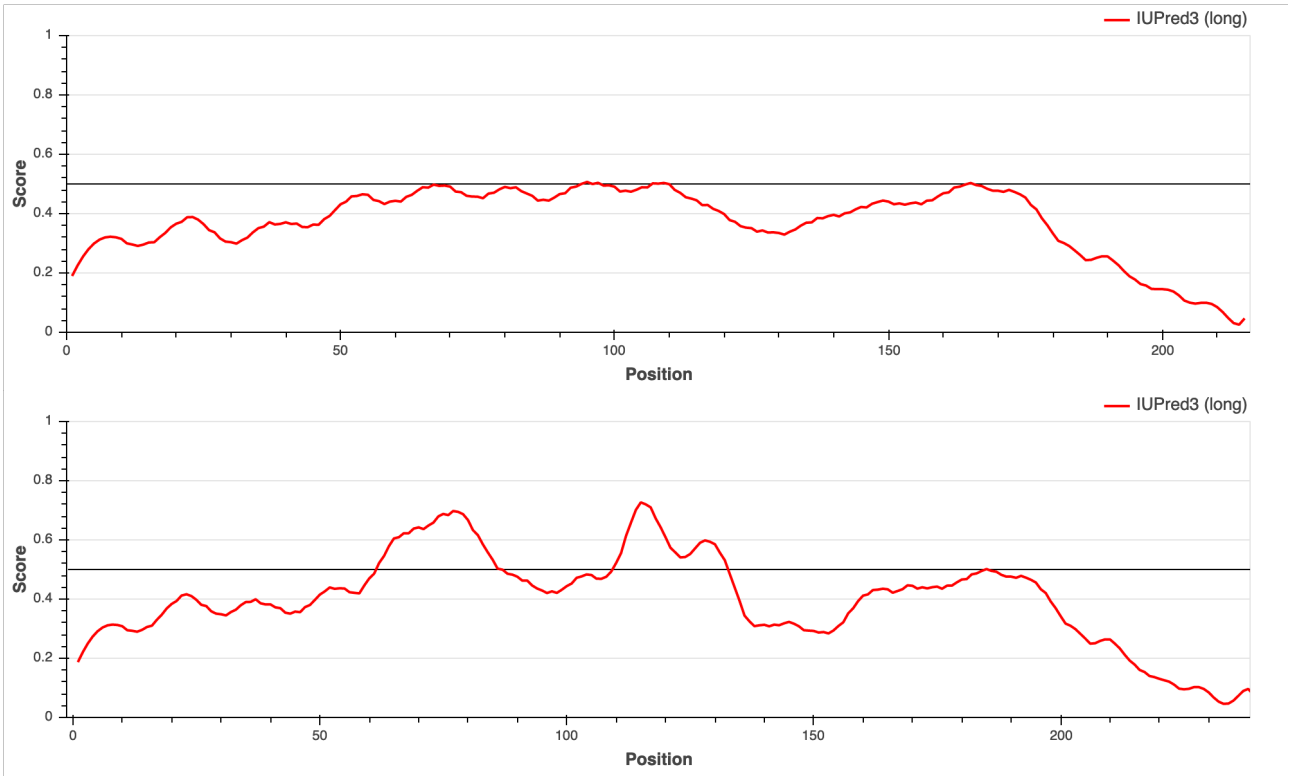


Fig. 44. IUPred3 disorder predictions for SynRFR24.1 and the 19_04242 design. A score above 0.5 denotes disorder.

Validation with MolProbity (Williams et al., 2018) showed no residues in forbidden regions of the Ramachandran plot, suggesting that the designed loop conformations are physically plausible.

Protein Geometry	Poor rotamers	0	0.00%	Goal: <0.3%
	Favored rotamers	169	99.41%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	221	94.85%	Goal: >98%
	Rama distribution Z-score	-0.71 ± 0.49		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	0 / 1746	0.00%	Goal: 0%
	Bad angles:	0 / 2379	0.00%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 4	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	0	0.0%	Goal: <1.0%
	CA Geometry outliers	0	0.00%	Goal: <0.5%
Additional validations	Chiral volume outliers	0/278		

Fig. 45. MolProbity validation of design 19_04242.

On availability of AlphaFold 2, further *in silico* validation was performed. In the two top-ranked models, loop 2 was predicted to be in a conformation nearly identical to the design (Cα RMSD of 0.716 Å), while loop 1 was predicted to be incorporated into the beta-solenoid structure. Indeed, the loop 1 sequence contains two Leu residues, which are in register relative to the hydrophobic core Leu residues in the rest of the solenoid structure.

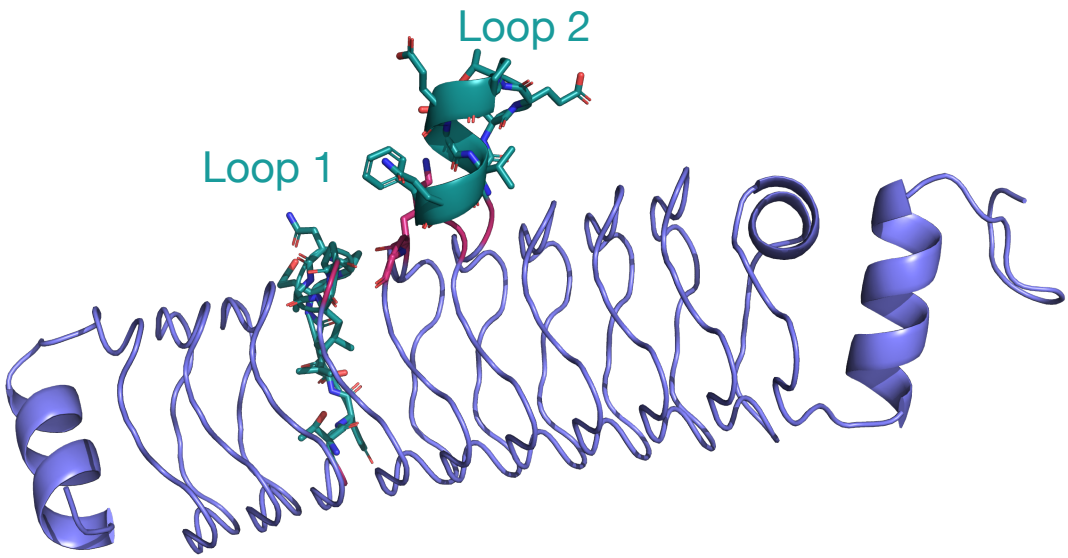


Fig. 46. Top-ranked AlphaFold model of design 19_04242.
Purple - components of the original starting construct, Fuchsia - residues on either side of the loop insertions and residues at the vertex of the intervening turn of the beta-helix, Teal - loop insertions.

4.3.2.4. EXPERIMENTAL CHARACTERISATION

The 2_loop variant was cloned, expressed and successfully purified by Ni-NTA IMAC. As seen in Fig. 47, the protein expressed solubly, suggesting that the introduced loops do not cause severe aggregation. Crystallisation trials were attempted but failed. However, the conditions of 0.1 M phosphate/citrate pH 4.2 with 40% v/v, and 0.1 M bicine pH 9.0 with 10% w/v PEG 20,000 and 2% v/v 1,4-Dioxane from the JCSG+ screen (Molecular Dimensions) produced putative spherulites, suggesting a route for future optimisation (Fig. 48).

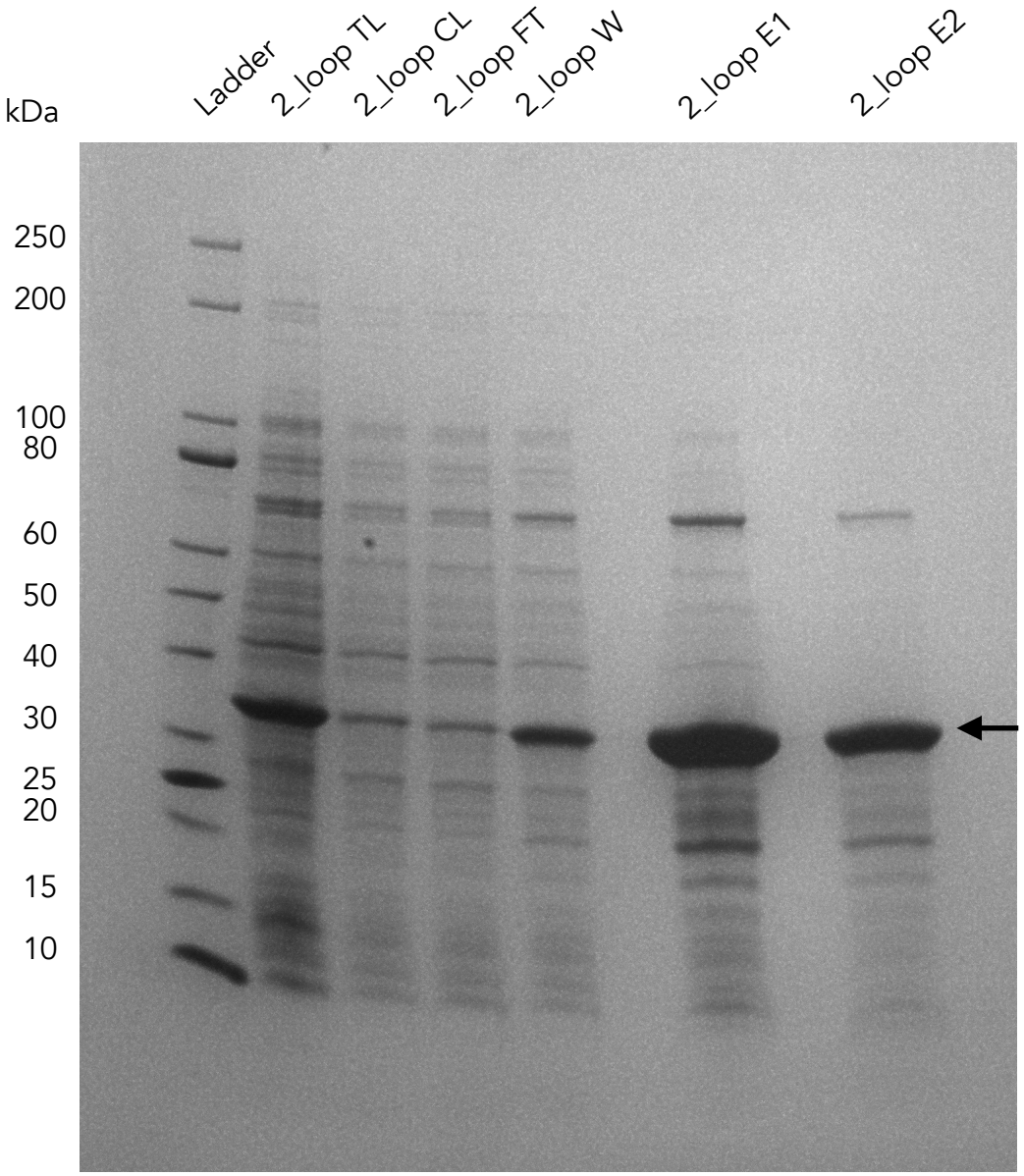


Fig. 47. SDS-PAGE gel of 2_loop Ni-NTA IMAC fractions.
TL – total lysate after sonication, CL – clear lysate after centrifugation of TL, FT – flow-through from resin incubation, Wash – wash with buffer with 25 mM imidazole, Elution – elution fraction eluted with buffer with 250 mM imidazole. The arrow points to the band of the expected molecular weight.

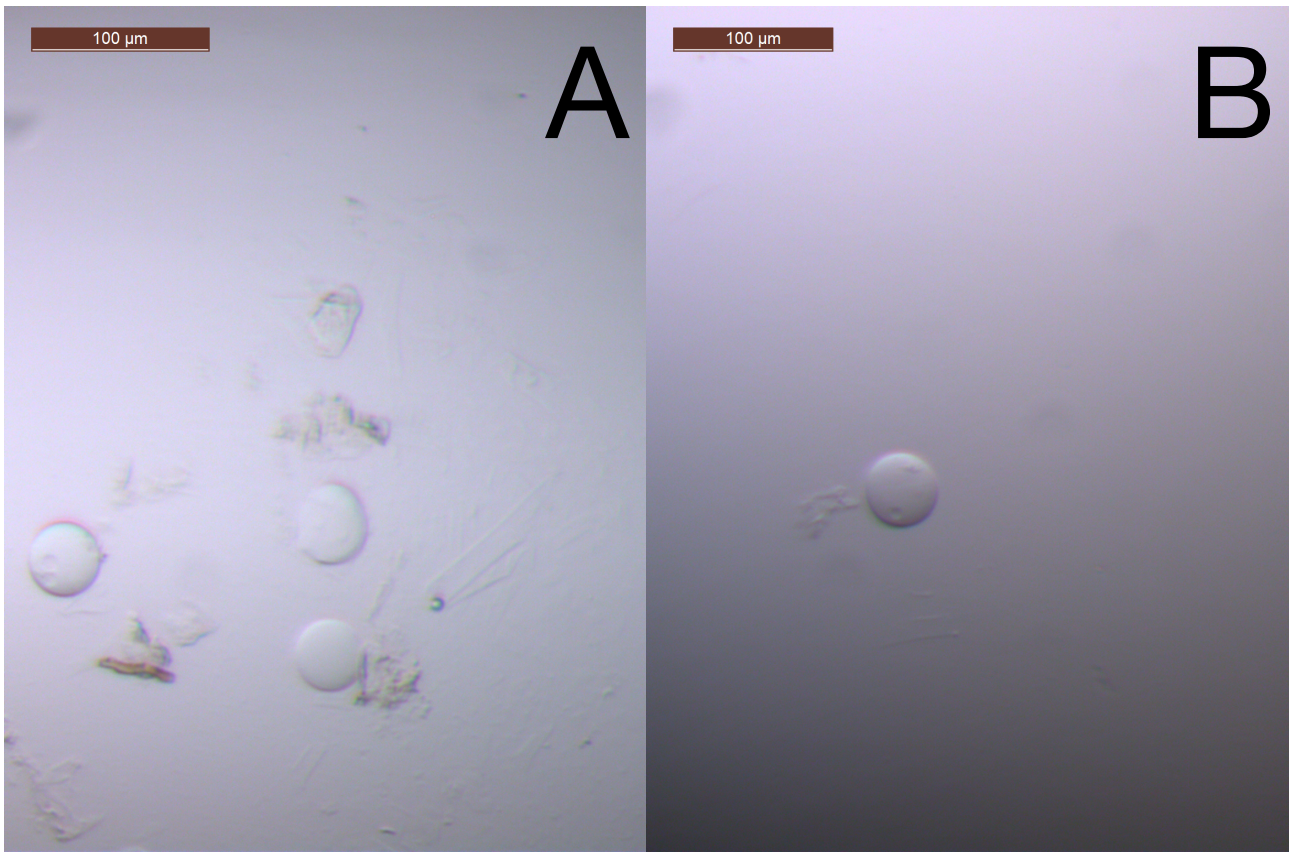


Fig. 48. Putative 2_loop spherulites found in the JCSG+ screen (Molecular Dimensions) in two different conditions.
A) 0.1 M phosphate/citrate pH 4.2 with 40% v/v B) 0.1 M bicine pH 9.0 with 10% w/v PEG 20,000 and 2% v/v 1,4-Dioxane.

4.3.3. SPYTAG/SPYCATCHER VARIANT DESIGN

4.3.3.1. CONSTRUCT DESIGN

For internal SpyTag and SpyCatcher insertions, the design methodology for the previously designed sfGFP fusion construct (MacDonald et al., 2016) was followed. The G_4S-SpyTag/SpyCatcher-SG_4 sequence was inserted between residues P132 and W133 in the beta-hairpin of the original t1428 variant (PDB ID 5DNS, Fig.49). For N-terminal SpyTag and SpyCatcher fusions, the Tag/Catcher sequence was inserted after the thrombin cleavage site of SynRFR24.1 (Beta1) with PDB ID 4YC5 and a short glycine-rich linker was added between the Tag/Catcher and the SynRFR sequence.

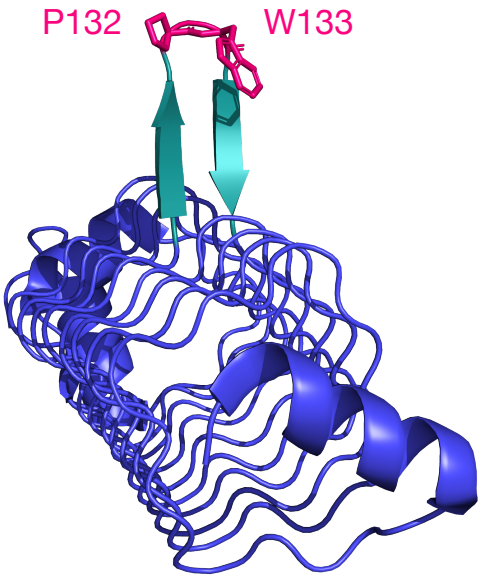


Fig. 49. The beta-hairpin t1428 SynRFR variant.
The SynRFR core is shown in purple, the beta-hairpin loop is shown in teal, and residues P132 and W133 are highlighted in fuchsia, showing the insertion site for the internal SpyTag and SpyCatcher fusions.

4.3.3.2. SPYTAG/SPYCATCHER REACTION

There was soluble expression for all SpyTag/Catcher constructs. In addition, the SpyTag and SpyCatcher moieties in all four constructs were reactive, which can be seen as the appearance of a band of greater molecular weight than both components (Fig. 50 A) on addition of the appropriate interaction partner. The iST-t1428 protein expressed solubly and could be purified by Ni-NTA but precipitated over time, while no such behaviour was observed for the other three constructs. Fig. 50 B shows a number of higher molecular weight bands on reaction between nST-beta1 and nSC-beta1, suggesting the formation of higher-order assemblies through covalent attachment between N-terminal SpyTags/Catchers and beta1 non-covalent C-terminal interfaces.

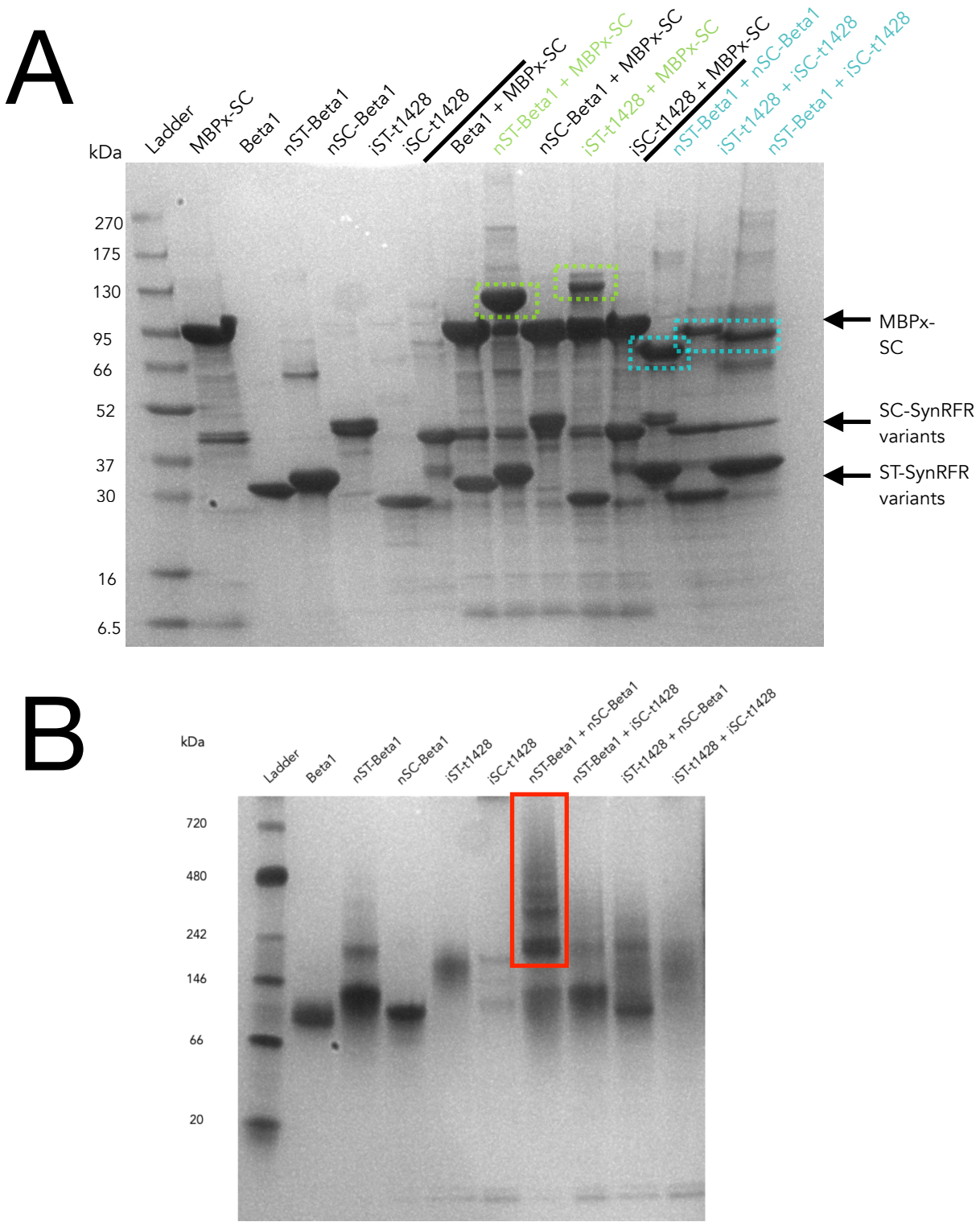


Fig. 50. SpyTag/Catcher reaction gels showing individual proteins and reaction products. A) SDS-PAGE gel of SpyTag/Catcher reactions. The arrows point to the three main types of individual proteins. Successful reactions are highlighted and the band shift on forming the SpyTag/Catcher product is surrounded by a dotted box. Successful reactions with MBPx-SC are highlighted in lime green, while successful reactions between SC- and ST-SynRFR variants are highlighted in cyan. B) Native gel of SpyTag/SpyCatcher reactions. The most notable observation is the putative formation of higher-order assemblies (seen as multiple bands of increasing molecular weight) in the nST-Beta1 + nSC-Beta1 reaction - highlighted with a red rectangle.

4.3.3.3. SIZE EXCLUSION CHROMATOGRAPHY

Size exclusion chromatography was carried out to assess the aggregation and oligomeric state of the ST- and SC-SynRFR variants. As iST-t1428 precipitated progressively, it was not purified by SEC. The Beta1 chromatogram (Fig. 51 A) had a single large peak at an elution volume consistent with a dimer (cyan) and a minimal soluble aggregate peak (red), indicating that most of the protein sample was soluble and of the expected oligomeric state. A similar profile was observed for nST-Beta1 (Fig. 51 B). However, there was now a shoulder to the main peak, suggesting some sample heterogeneity, possibly due to degradation. The nSC-Beta1 chromatogram (Fig. 51 C) showed a large soluble aggregate peak and a smaller peak at an elution volume consistent with a dimer. The accompanying SDS-PAGE gel suggestsed that the aggregate peak is comprised of impurities and the peak in cyan contains the nSC-Beta1 protein. The iSC-t1428 chromatogram had a broad and large aggregate peak, containing most of the protein sample as well as a smaller, broad peak at approximately the elution volume consistent with a dimer. The accompanying SDS-PAGE gel showed three main bands, which are not separated by SEC and could be the result of degradation (Fig. 51 D).

A

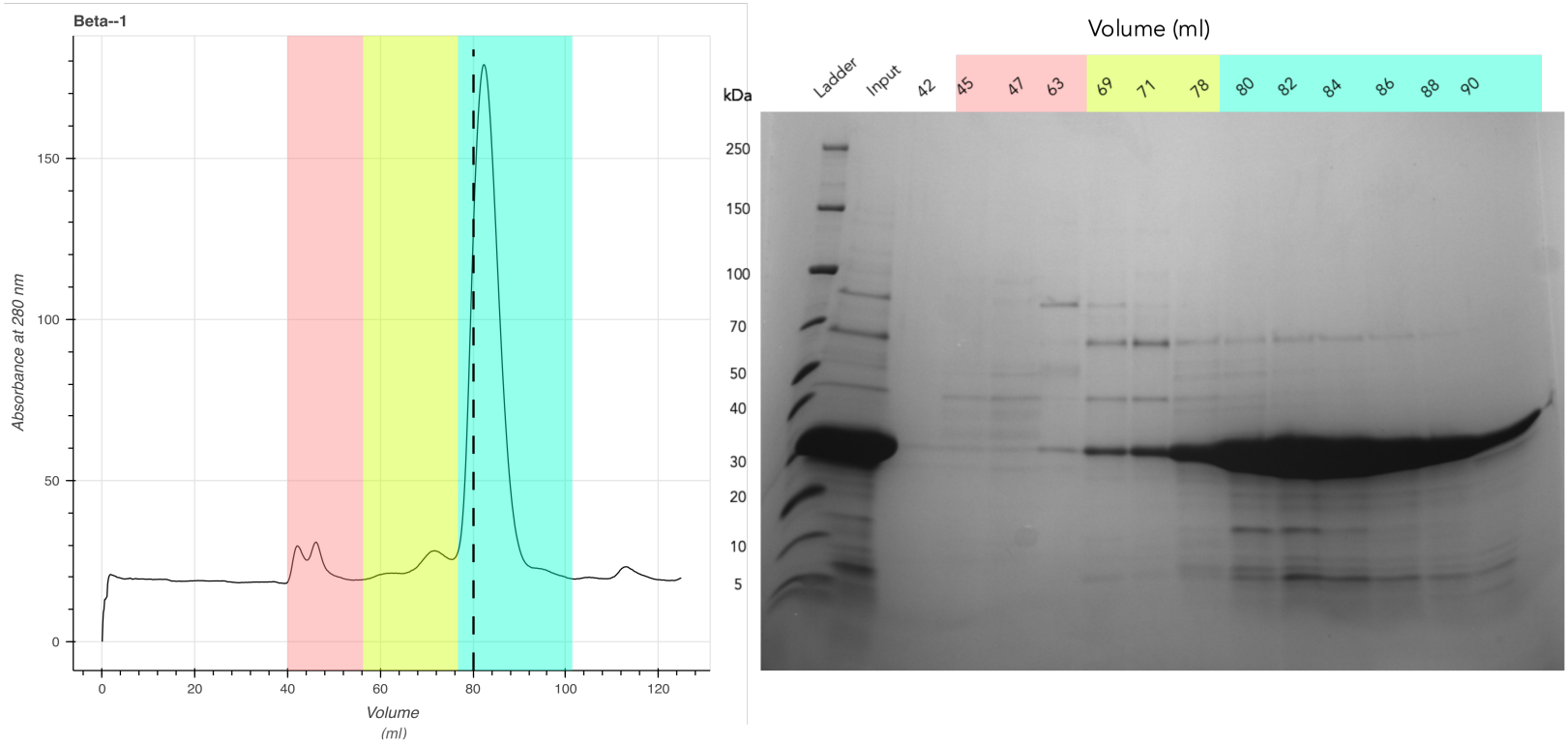


Fig. 51. Size exclusion chromatography data for beta1,nST-beta1,nSC-beta1,iSC-t1428 .

A), B) ,C),D) Chromatogram and accompanying SDS-PAGE gel of elution fractions with the approximate elution volume recorded on top of each gel lane. Fractions are colour-coded according to the peaks in the chromatogram. The dashed line denotes the approximate expected elution volume for a dimer of each protein based on the predicted molecular weight of the protein and a standard curve using protein standards of known molecular weight. Fractions are colour-coded according to the peaks in the chromatogram.

B

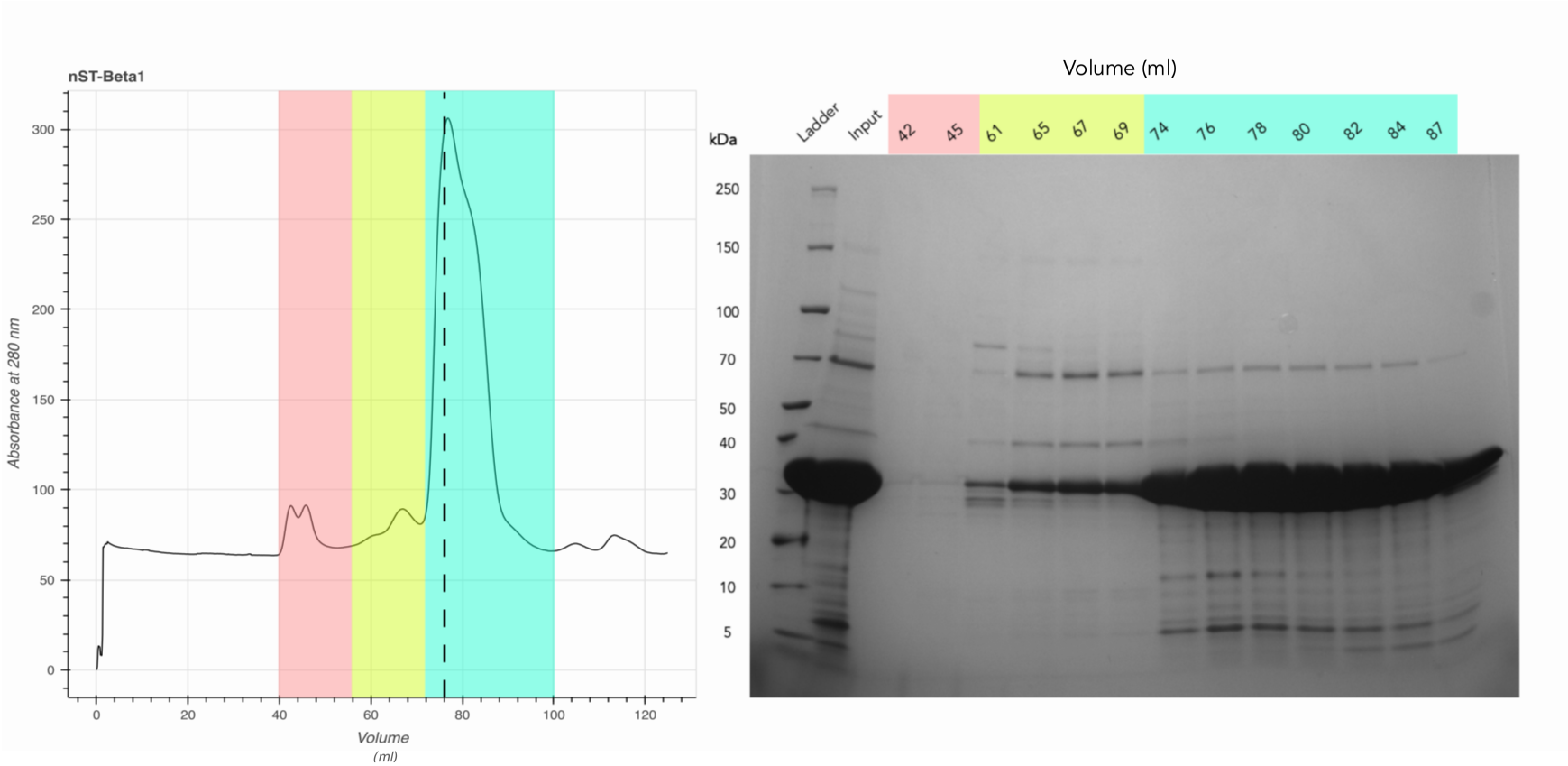


Fig. 51. Size exclusion chromatography data for beta1,nST-beta1,nSC-beta1,iSC-t1428, continued.

C

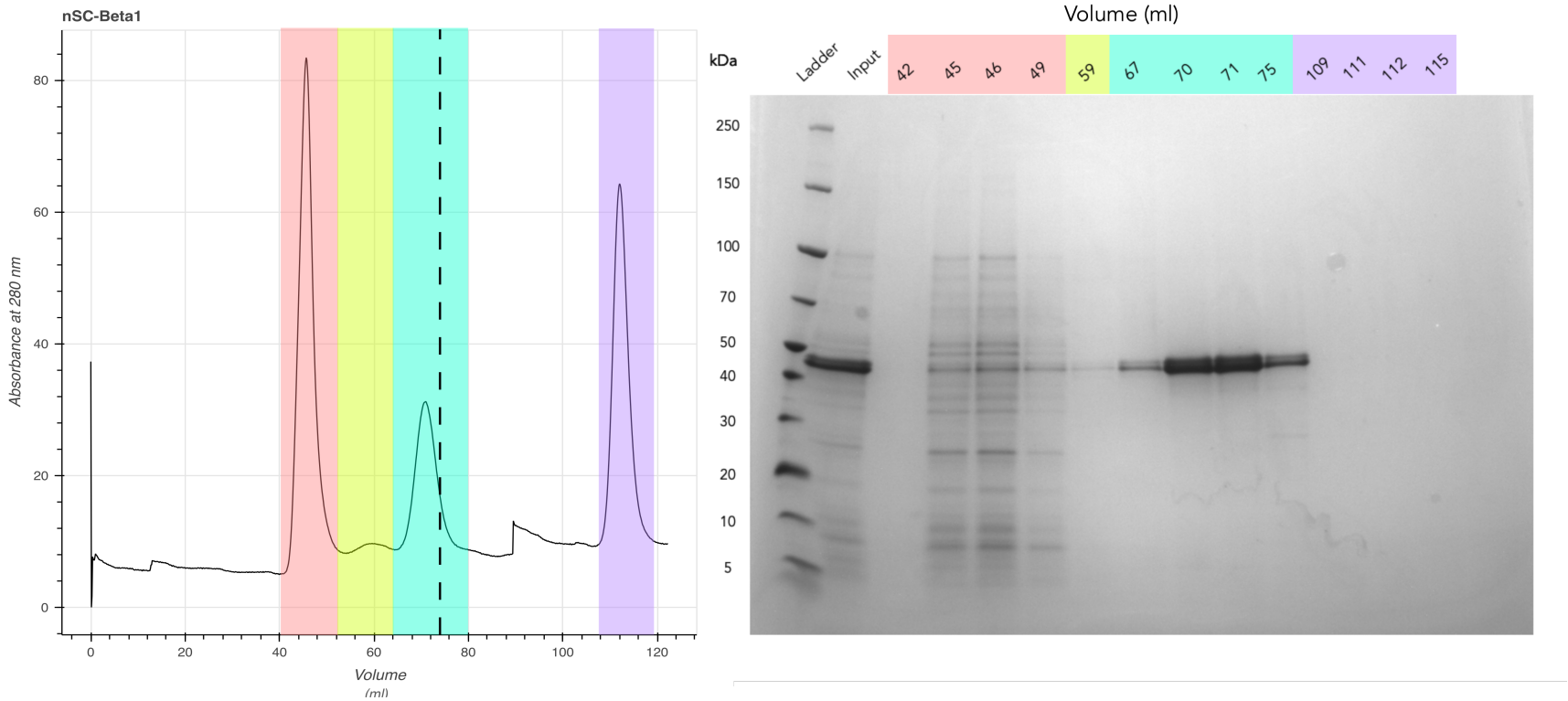


Fig. 51. Size exclusion chromatography data for beta1,nST-beta1,nSC-beta1,iSC-t1428, continued.

D

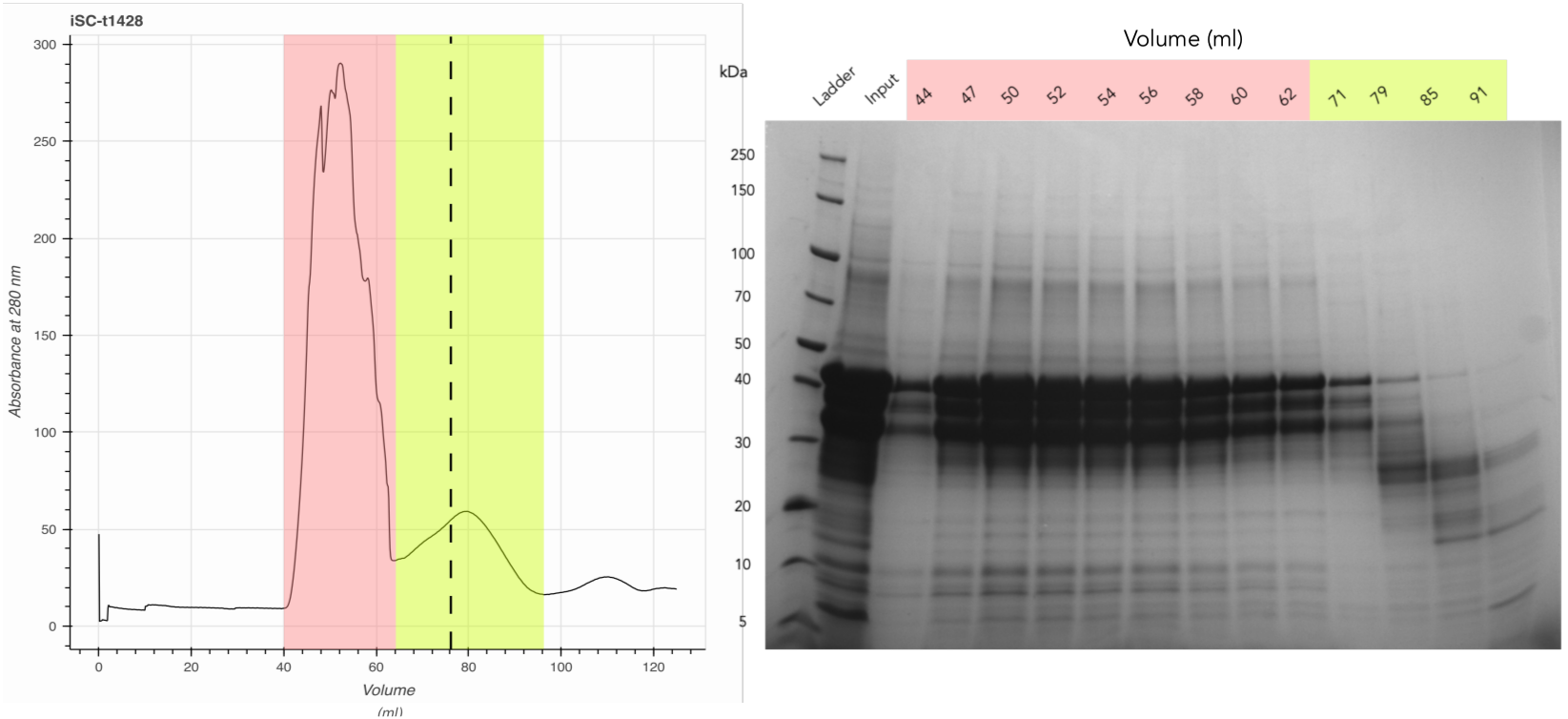


Fig. 51. Size exclusion chromatography data for beta1,nST-beta1,nSC-beta1,iSC-t1428, continued.

4.3.3.4. ALPHAFOLD 2 MODELS

ColabFold predictions without the use of structural templates were produced for all SpyTag/SpyCatcher constructs. In all cases but iST-t1428 the top-ranked model out of 5 is shown in Fig. 52. For all constructs but iST-t1428 the solenoid region of the top model was predicted to be intact and aligned to the crystal structure of the parent construct (PDB ID 4YC5 for N-terminal fusions and PDB ID 5DNS for internal fusions) with sub-Å RMSD using the PyMOL align function. In some of the models for iST-t1428 the SpyTag was predicted to interfere with the solenoid structure. Overall, the models agreed well with the intended design.

Fig. 52. ColabFold models of ST/SC constructs compared to the crystal structure of the starting construct, and a LEGO(R)-like representation. Blue spheres - N-terminus, Red spheres - C-terminus, Purple - components of the original starting construct, Fuchsia - G4S linkers, Teal - SpyTag/ Catcher, Coral - C-terminal dimerisation interface (LEGO representation only).

4.3.3.5. X-RAY CRYSTALLOGRAPHY

Crystals were obtained for the nST-Beta1 and nSC-Beta1 proteins with examples shown in Fig. 53. The crystals shared the same space group as PDB ID 4YCQ (P 32 2 1). Crystal structures were obtained for both proteins. However, the N-terminal regions containing the SpyTag and SpyCatcher, respectively, could not be resolved. The most N-terminal resolved residue for nST-Beta1 is His 20, whereas for nSC-Beta1 it is Leu 52 (Fig. 54), according to the numbering scheme for PDB ID 4YC5.

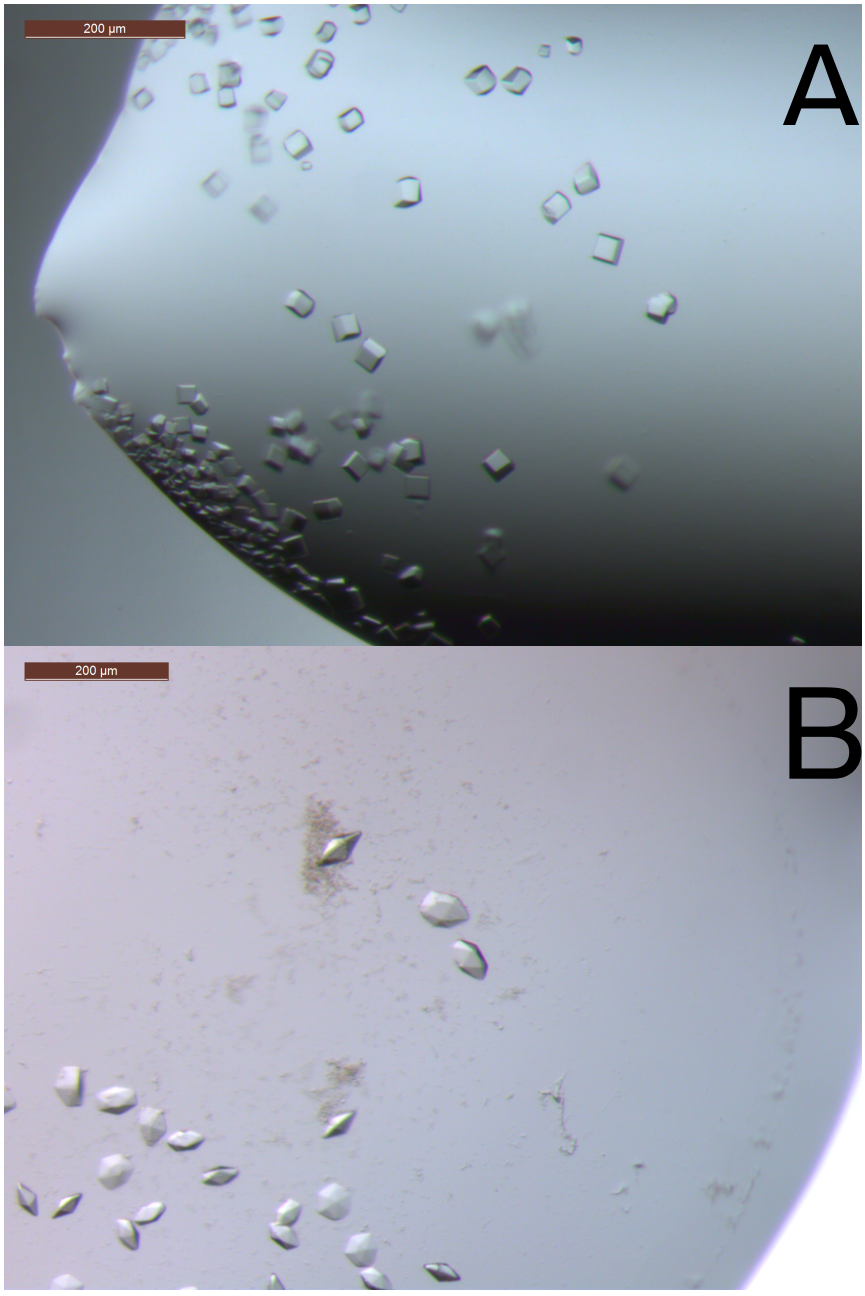


Fig. 53. Examples of crystals for nSC-Beta1 and nST-Beta1.
A) nSC-Beta1 crystals. 0.2 M Magnesium chloride hexahydrate, 0.1 M Sodium HEPES pH 7.5, 30% v/v PEG 400 B) nSC-Beta1 crystals. 0.1 M Sodium cacodylate pH 6.5, 40 % v/v MPD C) nST-Beta1 crystals. 0.2 M Magnesium chloride, 0.1 M Sodium HEPES pH 7.5, 30% v/v PEG 400 D) nST-Beta1 crystals. 0.2 M Magnesium chloride hexahydrate, 0.1 M Tris pH 7.0, 10% w/v PEG 8000.

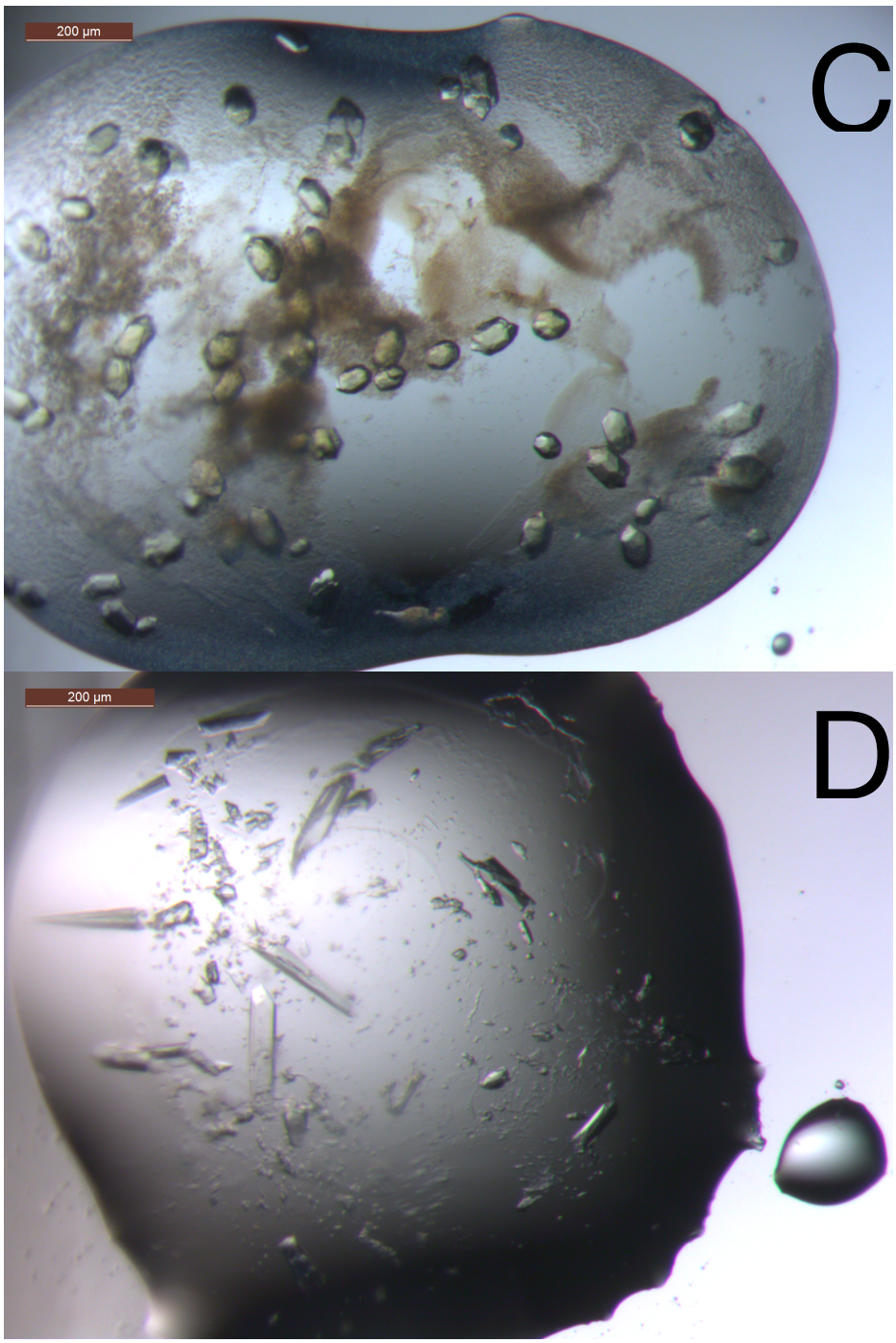


Fig. 53. Examples of crystals for nSC-Beta1 and nST-Beta1, continued.

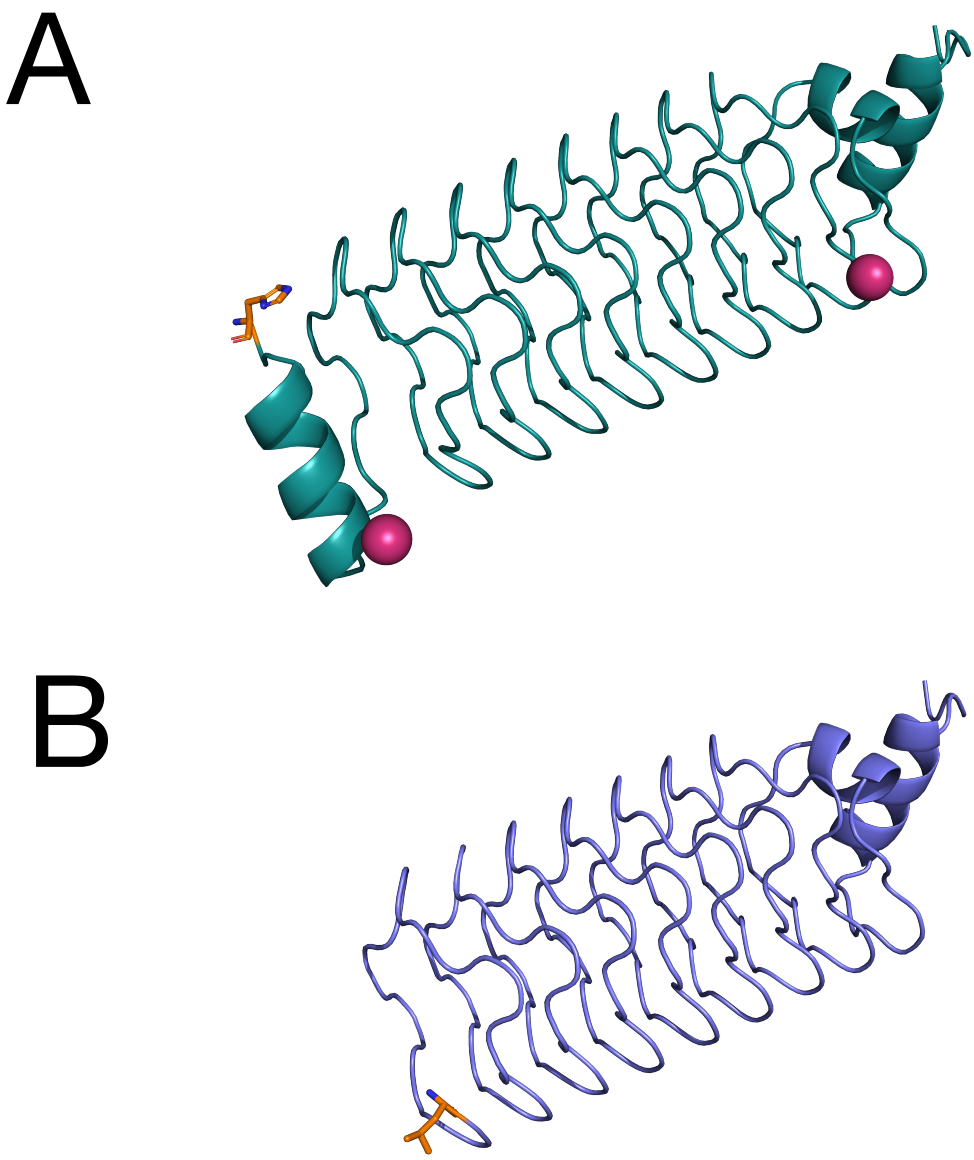


Fig. 54. Crystal structures for nST-Beta1 and nSC-Beta1.
A) nST-Beta1 structure. Pink - chloride ions, orange - residue His 20. B) nSC-Beta1 structure. Orange - residue Leu 52.

4.3.4. SPLIT SOLENOID DESIGN

4.3.4.1. CONSTRUCT DESIGN

The approach taken here contrasts with the fibrillar assembly approach of Peralta et al (Peralta et al., 2015). The capping structures at the N- and C-terminus of the rationally designed SynRFR24.1 protein were retained, and the protein was split into two parts (A and B) to test for spontaneous self-assembly of the two units. The split point is at a vertex of the rectangular cross-section repeating unit. This point was chosen as the vertex has previously supported loop insertions. To generate part A1, the native sequence is terminated after residue E119, whereas to generate part A2, the native sequence is terminated before E119, at residue Q118 (Fig. 55). In part B, residue A120 of the original full-length sequence is preceded by an N-terminal methionine. Both part A variants have an N-terminal thrombin-cleavable hexahistidine tag. The amino acid sequence of part B was modified to include an N-terminal Met residue, following the protocol of Peralta et al (Peralta et al., 2015). An additional part B construct with a C-terminal thrombin-cleavable hexahistidine tag was also generated.

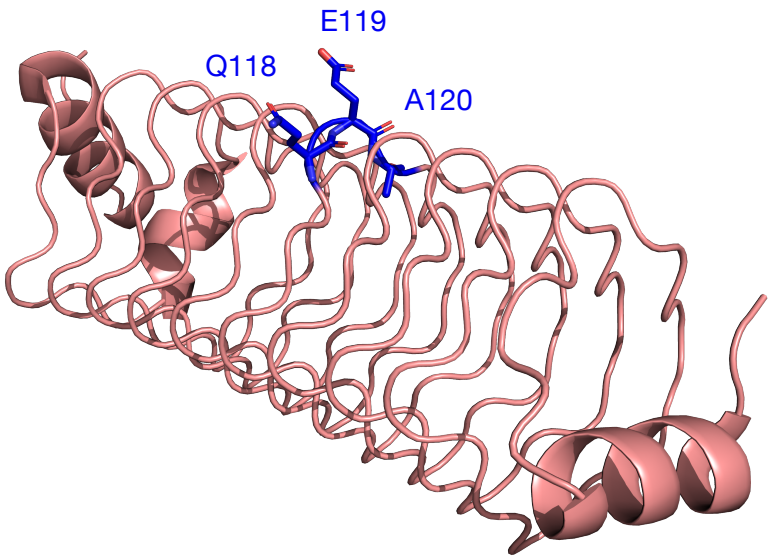


Fig. 55. The split solenoid split point shown on PDB ID 4YC5.
The three residues at the vertex split point are shown in blue, in stick representation. The solenoid sequence is split into two halves either between residue E119 and A120, or residue E119 is deleted so the first half of the solenoid sequence terminates at residue Q118.

4.3.4.2. PULL-DOWN

To test for association between parts A1 and A2, and part B, a pull-down strategy was selected. The pull-downs were carried out either with simultaneous or sequential addition of parts to Ni-NTA resin. This was followed by visualisation by SDS-PAGE.

4.3.4.2.1. SIMULTANEOUS ADDITION OF PARTS TO THE RESIN

Part A cleared lysate and part B cleared lysate were mixed and incubated with Ni-NTA resin for 1h. The resin was then washed with 40 column volumes of 100 mM bicine 150 mM NaCl pH 9.0, followed by 2 column volumes of 100 mM bicine 150 mM NaCl 25 mM imidazole pH 9.0 and 4 column volumes of 100 mM bicine 150 mM NaCl 250 mM imidazole pH 9.0. The same protocol was also performed in the absence of part B as a control. The results were visualised on 4-20% gradient Tris-Gly polyacrylamide gels (SERVA) by SDS-PAGE.

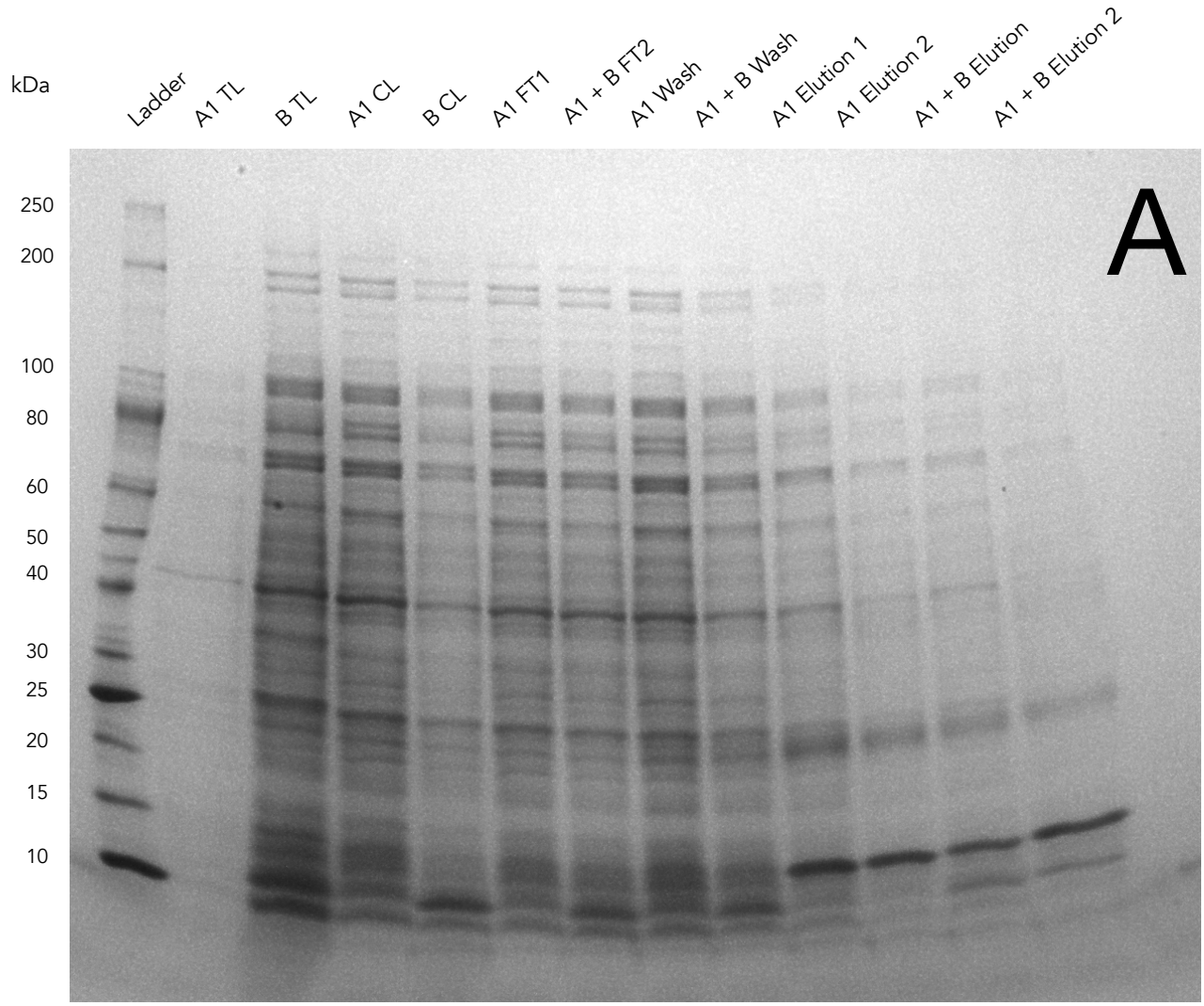


Fig. 56. Pull-down with simultaneous addition of parts to the resin.
A) SDS-PAGE gel of fractions collected from the pull-down experiment using part A1 and part B. TL – total lysate (after sonication), CL – cleared lysate (after ultracentrifugation of TL), FT1 – flow-through after resin incubation with part A, FT2 – flow-through after part A-bound resin incubation with part B, Wash – 25 mM imidazole buffer wash after incubation. B) SDS-PAGE gel of fractions collected from the pull-down experiment using part A2 and part B. An additional band due to part B appears on incubation with part B.

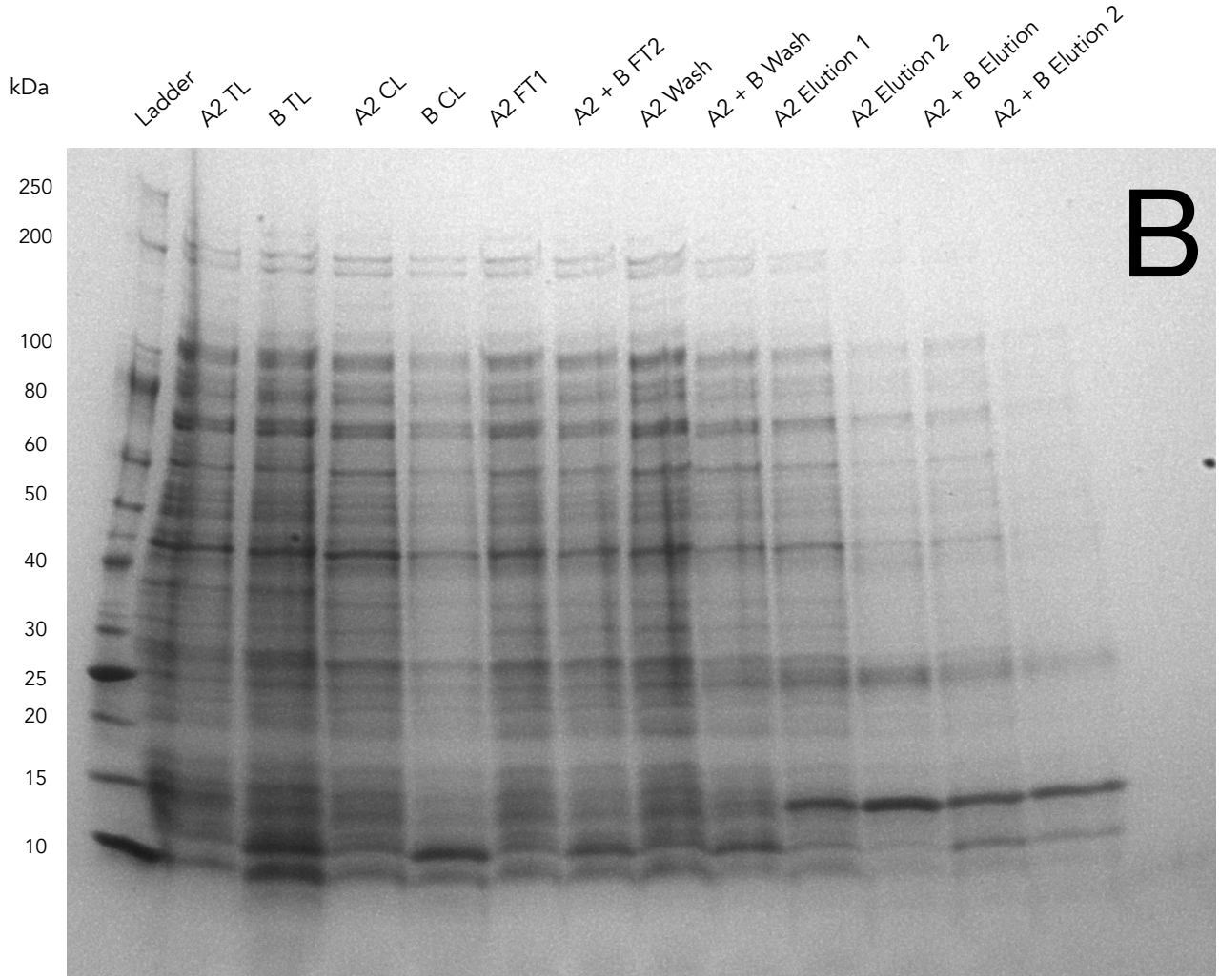


Fig. 56. Pull-down with simultaneous addition of parts to the resin, continued.

4.3.4.2.2. SEQUENTIAL ADDITION OF PARTS TO THE RESIN

0.5 ml of Ni-NTA resin were used in a batch-method gravity column. 5 ml of part A cleared lysate and 3 ml of part B cleared lysate were used. Part A1 and A2 cleared lysates were incubated with resin for 45 min, then the resin was washed with 40 column volumes of 100 mM bicine 150 mM NaCl pH 9.0, 40 column volumes of 100 mM bicine 150 mM NaCl 25 mM imidazole pH 9.0 and finally 40 column volumes of 100 mM bicine 150 mM NaCl pH 9.0 again. After washing, half of the samples of part A1- and part A2-bound resins were incubated with part B cleared lysate for 45 minutes. Then both part A1-bound and part A2-bound resins with and without incubation with part B cleared lysate were washed with 40 column volumes of 100 mM bicine 150 mM NaCl pH 9.0, 40 column volumes of 100 mM bicine 150 mM NaCl 25 mM imidazole pH 9.0 and finally 40 column volumes of 100 mM bicine 150 mM NaCl pH 9.0 again. An elution was then performed with 2 column volumes of 100 mM bicine 150 mM NaCl 250 mM imidazole pH 9.0 buffer. Additionally, part B cleared lysate was incubated with resin and washed and eluted as above to test for non-specific binding to the Ni-NTA resin. Samples of the cleared lysate, flow-through and elutions were collected for SDS-PAGE. A 1:10 dilution was used for the cleared lysate and flow-through fractions. The elution fractions were undiluted. The samples were run at 200V for 70 min on an 8-16% gradient Tris-Gly polyacrylamide gel (SERVA). Ladder: Elite Pre-stained Protein Ladder (Protein Ark).

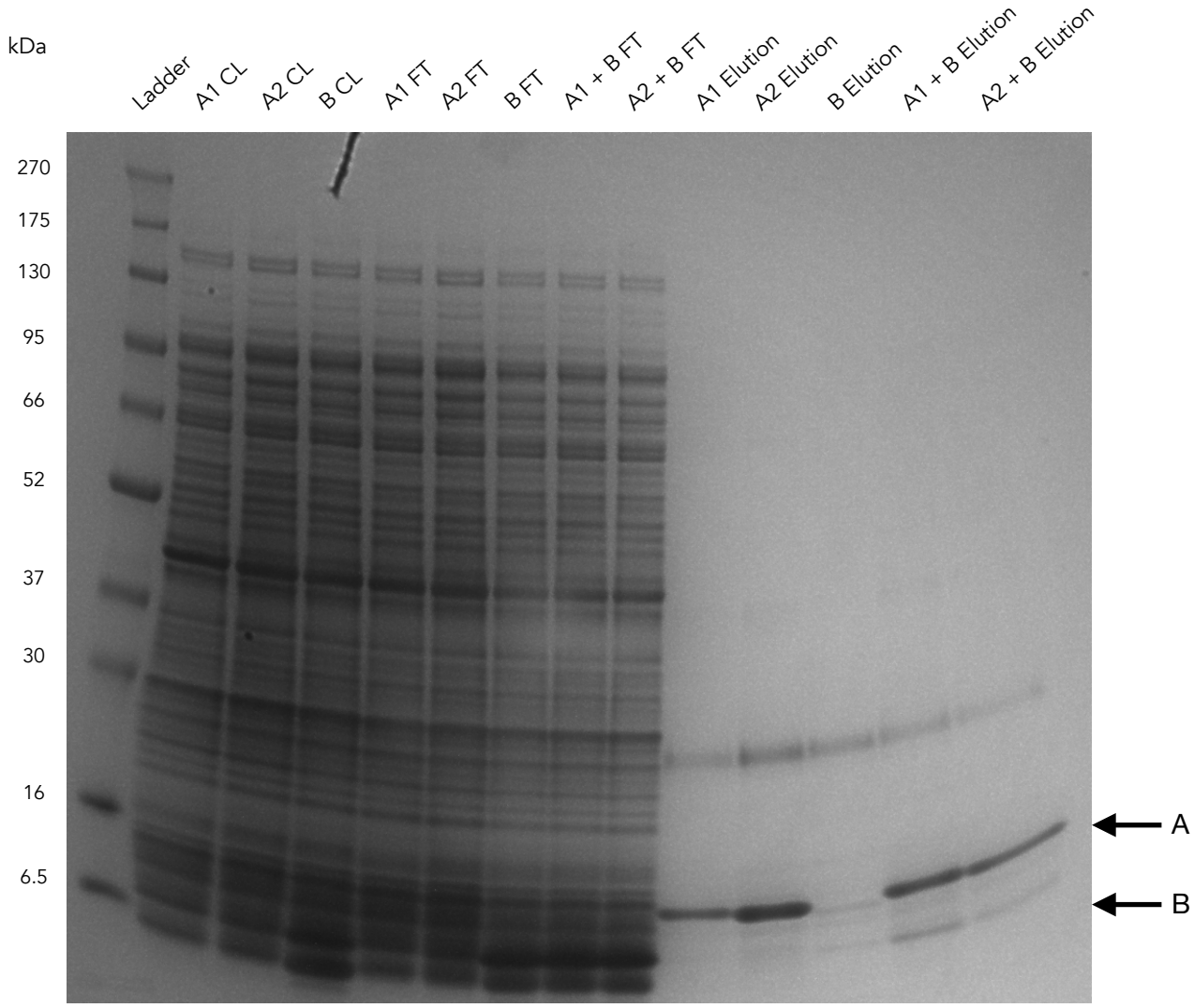


Fig. 57. Pull-down with sequential addition of parts to the Ni-NTA resin.
Fractions collected from the pull-down experiment. CL - cleared lysate (after ultracentrifugation of the total lysate). FT - flow-through from incubation of CL with resin. Arrows pointing to parts A and B. An additional band due to part B appears on incubation with part B.

Parts A1, A2 and B expressed solubly with part B showing a more intense expression band compared to parts A1 and A2. In both pull-down tests an additional band, consistent with part B (as judged from the part B CL lanes), appeared in the elution fractions of samples containing both part A1/A2 and part B. Non-specific binding of part B to the Ni-NTA resin was minimal, as can be seen from the faint bands in the B Elution lane of the SDS-PAGE gel in Fig. 56. Therefore, the SDS-PAGE gels suggest that some part B was being captured onto resin with part A1 or part A2 bound. Part A1 and part A2 appeared to bind to similar amounts of part B based on band intensities in Fig. 57. In addition, the gels suggested that a part A1-part B or part A2-part B complex could be built sequentially by part B binding to resin-captured part A1 or A2 or the complex could form in solution and then be captured via the part A1/part A2 His-tag onto the Ni-NTA resin.

4.3.4.3. SIZE EXCLUSION CHROMATOGRAPHY

Size exclusion chromatography was carried out on a different prep of parts A1, A2 and BT to test for aggregation and oligomeric state. As can be seen in Fig. 58, there were large aggregate peaks in the size exclusion chromatograms for part A1 and part A2, as well as a persistent contaminant at a molecular weight of approximately 30 kDa as seen in the accompanying SDS-PAGE gels of the fractions. In addition, part A1 and part A2 do not contain any tryptophan residues, making absorbance measurements at 280 nm less reliable. Therefore, it is difficult to ascertain the oligomeric state of part A1 and A2 from this prep. In contrast, the size exclusion chromatogram for part BT showed minimal aggregation and a large peak at an elution volume consistent with a dimer. Therefore, part BT expressed solubly, was associated with minimal impurities and was dimeric. Crystallisation trials were attempted for part BT but were unsuccessful.

A

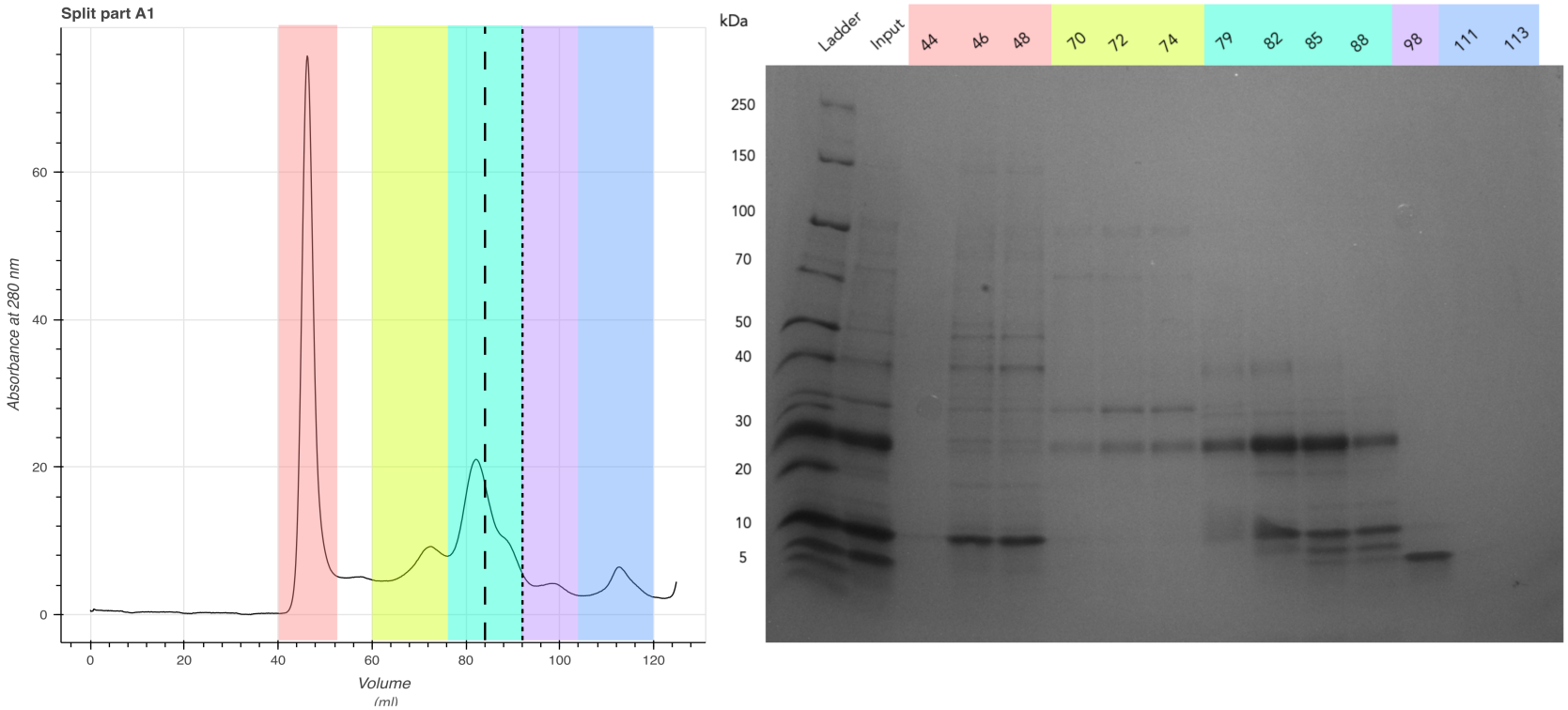


Fig. 58. Size exclusion chromatography data for part A1, part A2 and part BT .

A), B) ,C) Chromatogram and accompanying SDS-PAGE gel of elution fractions with the approximate elution volume recorded on top of each gel lane. Fractions are colour-coded according to the peaks in the chromatogram. The dashed line denotes the approximate expected elution volume for a dimer of each protein, whereas the dotted line represents the approximate expected elution volume for a monomer of each protein based on the predicted molecular weight of the protein and a standard curve using protein standards of known molecular weight. Fractions are colour-coded according to the peaks in the chromatogram.

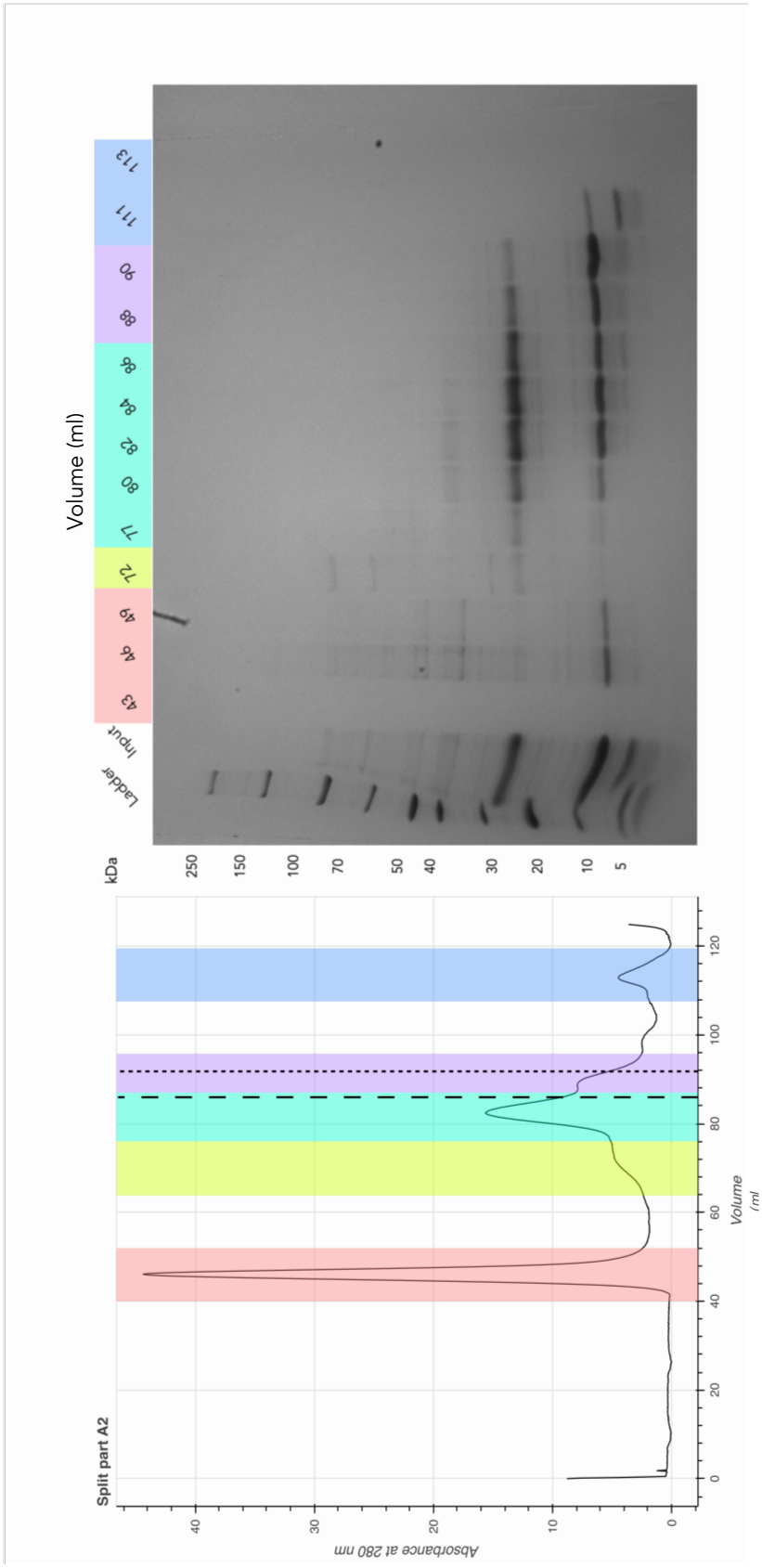


Fig. 58. Size exclusion chromatography data for part A1, part A2 and part BT, continued.

C

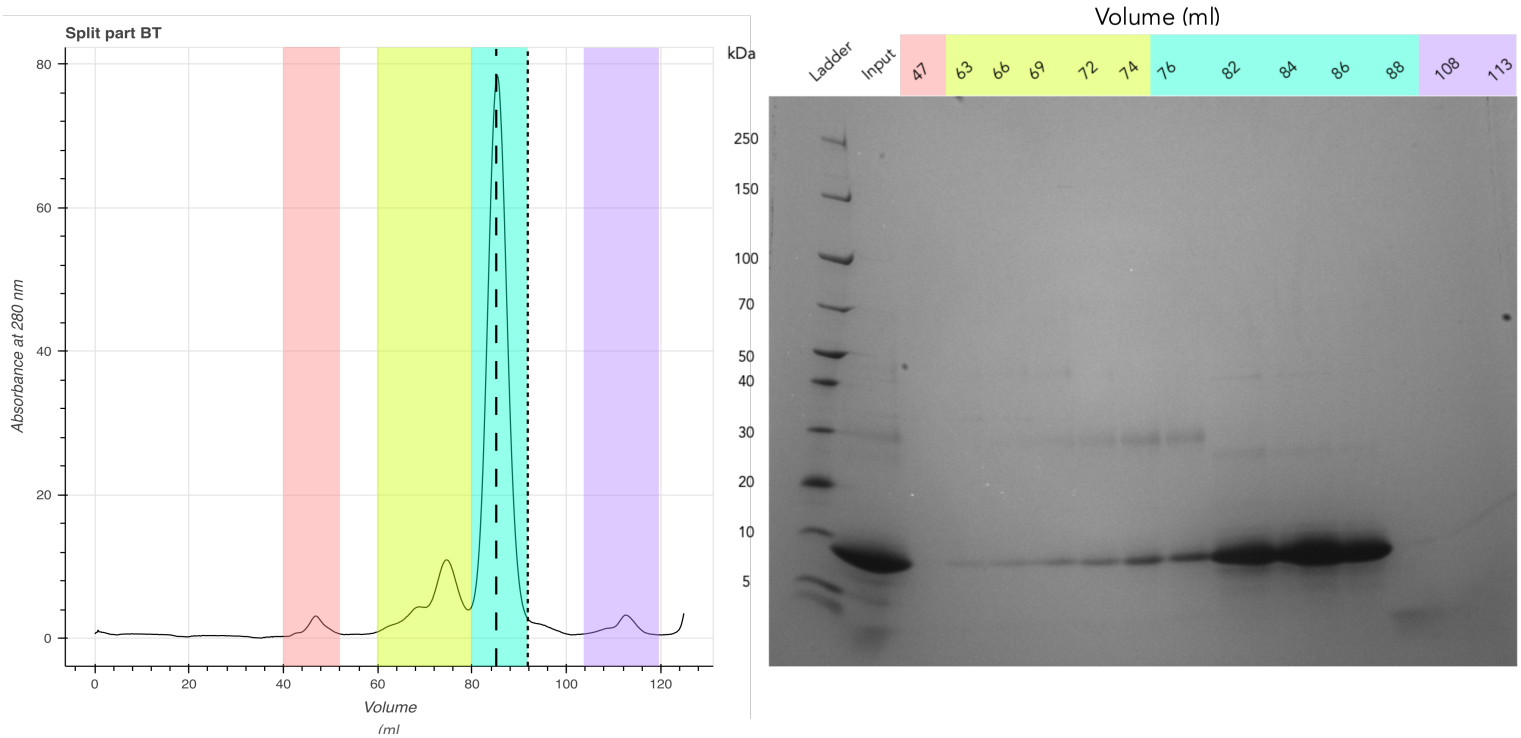


Fig. 58. Size exclusion chromatography data for part A1, part A2 and part BT, continued.

4.3.4.4. ALPHAFOLD 2 MODELS

As shown in Fig. 59, ColabFold models for the part A- part B complexes followed the design intention and had a solenoid region almost identical to that of PDB entry 4YC5. Models with or without the N-terminal Met residue of part B had similar backbone structures but models without the N-terminal Met formed an additional H-bond between the N-terminal amino group of the Ala residue and the C-terminal carboxyl group of the Q residue at the C-terminus of part A1/A2. Alternative ColabFold models for the part A1-part B complex (Fig. 60) suggested different possible binding modes between the two parts. In Fig. 60 A) the C-terminal five residues of part A1 were looped out, creating a change in the register of the beta-solenoid coils. In Fig. 60 B) and C) an alternative orientation of the two parts was observed.

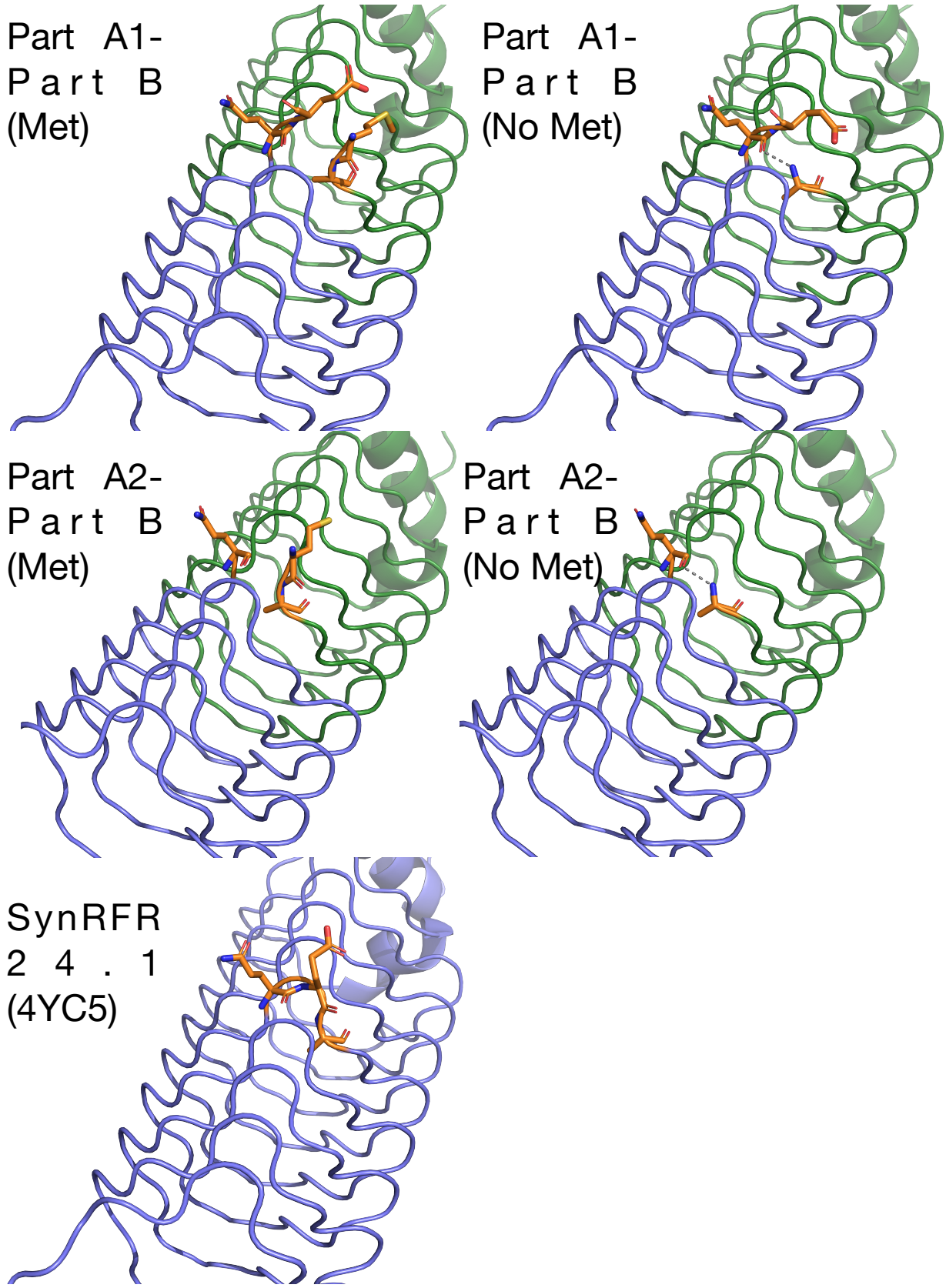


Fig. 59. ColabFold models for part A- part B complexes with and without the N-terminal Met of part B, and SynRFR24.1.
Purple - part A1/ 4yc5. Green - part B. Orange - residues at split site - QE (part A1), Q (part A2), MA or A (part B). Grey dotted line - H-bond.

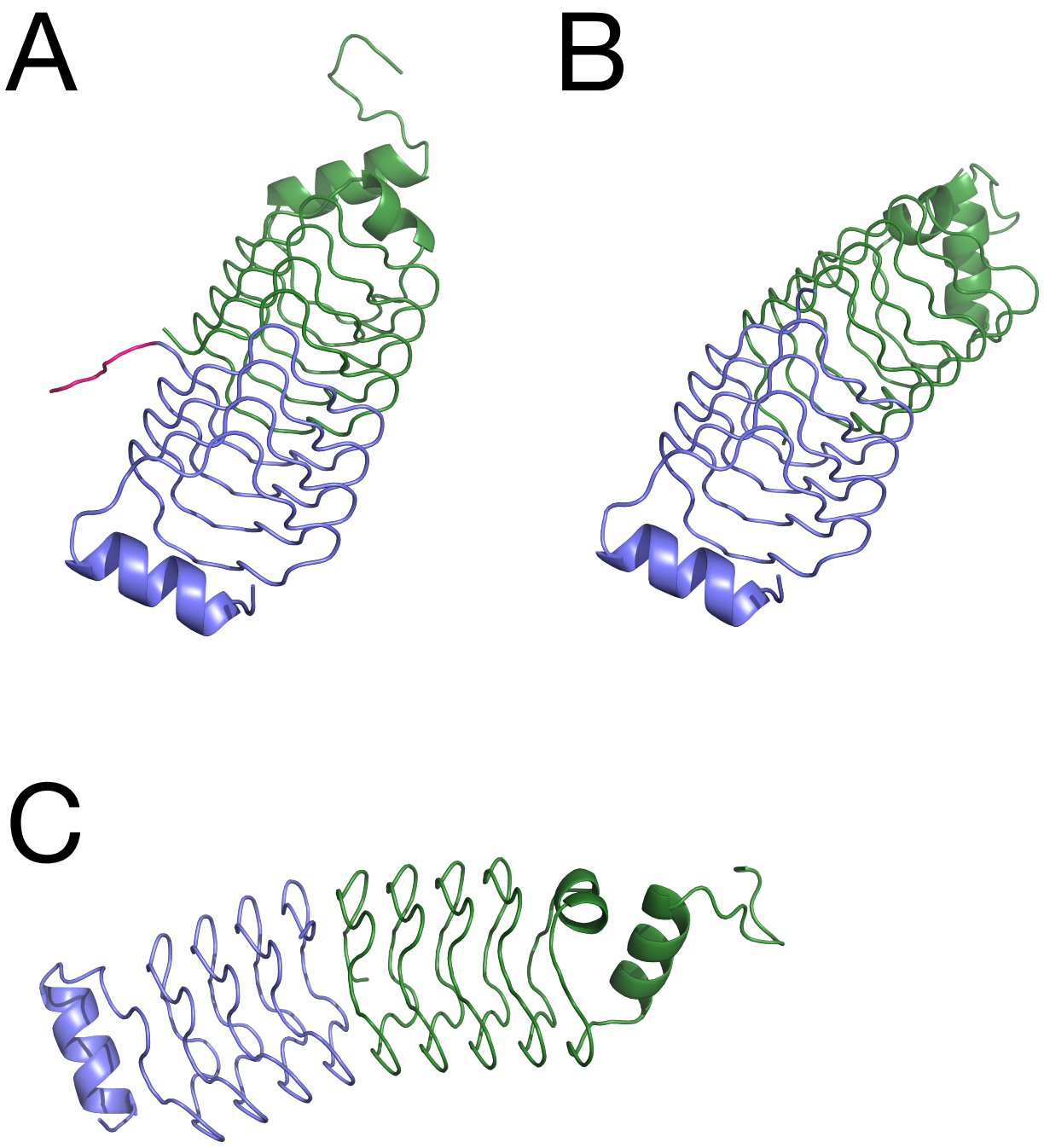


Fig. 60. Alternative ColabFold models for the part A1-part B complex. Purple - part A1. Green - part B., red - looped out five residues.

4.4. DISCUSSION

4.4.1. CONSENSUS HEXAPEPTIDE DESIGN

In contrast to the original SynRFR paper (MacDonald et al., 2016), which involved manual alignment of pentapeptide repeat sequences from 10 RFR proteins, the approach presented here allows for the automated processing of hundreds to thousands of protein structures to obtain a consensus sequence and a residue frequency table. The structure-based repeat alignment and extraction presented in this work can be performed on any repeat type with a defined repeat length and the repeat definitions obtained can be used for alternative downstream applications, such as defining the transformation matrices that superpose equivalent residues in two adjacent repeats. The consensus sequence for the hexapeptide repeat fold obtained using our method is supported by a review on left-handed beta-helix proteins (Prabha & Balaji, 2021). I generated sequences by four different approaches as outlined in Section 4.1.2 to test the effect of constraints on residue identities in conserved positions of the hexapeptide repeat (positions 1, 2 and 5). Instead of using Rosetta tools for sequence validation, I opted for the ColabFold implementation of AlphaFold, inspired by the work of Jendrusch et al (Jendrusch et al., 2021). The highest percentage of models passing the threshold of average pLDDT > 0.9 were produced by method 4, followed in second place by method 3. Hence, hexapeptide sequences without restrictions in residue identities at conserved hexapeptide positions produce a greater number of high-scoring AlphaFold2 models compared to sequences with constraints. However, sequences with no identity restrictions in positions 1,2 and 5 but with the added constraint that no two adjacent residues within each hexapeptide repeat can be identical produce the greatest number of high-scoring AlphaFold2 models. This could be due to some bias of the AlphaFold2 model towards natural or natural-like sequences as these sequences comprise its training dataset. Alternatively, it could suggest that sequences produced by method 4 have the highest likelihood of producing stable, folded structures. To test this, the top-scoring sequence produced by each method was expressed and purified. Due to time constraints, this project could not be pursued further. Only the sequences produced by methods 2,3 and 4 expressed and could be purified. However, the size exclusion profile of all proteins revealed a large aggregation peak and little to no soluble protein at the expected elution volume. Some soluble protein around the expected elution volume was observed for the sequence produced by method 2 but not for the sequence produced by method 3. These results suggest that good performance on AlphaFold 2 metrics does not necessarily correlate with soluble, mono disperse protein in the laboratory. Therefore, additional constraints and *in silico* validation are required, such as threading the sequences onto the desired backbone in Rosetta (Noguchi et al., 2019) and using filters based on surface hydrophobicity and aggregation propensity. Sequences can also be considered on the 18-residue turn level, rather than just the hexapeptide level and junction positions between individual hexapeptide repeats can be redesigned. Alternatively, or concomitantly with additional constraints and filters, ProteinMPNN (Dauparas et al., 2022) can be used to redesign these sequences. The authors observe that even when using a Monte Carlo loop with AlphaFold prediction to design

sequences, the sequences produced are largely insoluble. However, redesigning the sequences with the deep-learning-based ProteinMPNN method can ‘rescue’ the sequences obtained from the AlphaFold loop by increasing the number of lysine and glutamate residues, thus increasing solubility.

4.4.2. TWO-LOOP VARIANT DESIGN

The design of the 2-loop variant involved a conceptually similar approach to the design of the original SynRFR loop extensions - i.e. the loops were designed by a fragment-free approach. The idea behind fragment-free loop protocols is to more thoroughly explore conformational space rather than being constrained to conformations of peptide fragments from protein structures in the PDB. However, instead of using a CG potential energy function to sample loop conformations (MacDonald et al., 2013) as in the original paper, a loop-closure-based approach (GenKIC) was used. Additional differences include the conformational clustering step after backbone generation, as well as using next generation kinematic closure, instead of sequence threading onto backbones as a validation step. The disadvantage of using next generation kinematic closure for validation is that it does not consider the structure of the solenoid scaffold and performs loop closure simulations for each loop individually. However, if each loop displays a folding funnel in loop closure simulations, this suggests that the loops could fold correctly when combined as well. The final Rosetta model for design 19_04242 displays limited packing between the loops. However, there are a number of hydrogen bonds within each loop, which could maintain the conformation of the loops. Further *in silico* validation using IUPred3 and AlphaFold 2 suggests that at least one of the loops might not adopt its designed conformation. IUPred3 predicts both loops to be disordered, whereas AlphaFold2 predicts that only loop 2 folds to the designed conformation, while loop 1 is incorporated into the solenoid scaffold This suggests that further *in silico* filters are required such as disorder prediction and a measure of sequence similarity to the underlying SynRFR consensus sequence. Although the AlphaFold 2 model may not be representative of the true structure, it raises the possibility that loop designs can become part of the solenoid structures - a consideration to be incorporated in future design efforts. The protein expressed solubly and putative spherulites were obtained, suggesting that it might be possible to obtain crystals with further optimisation. Alternatively, circular dichroism can be used to assess the secondary structure content of the protein before and after the introduction of the two loops. However, due to time constraints, this project was not pursued further.

4.4.3. SPYTAG/SPYCATCHER VARIANT DESIGN

In the original paper an sfGFP domain was inserted into the beta-hairpin of the t1428 variant. This prompted the idea of inserting SpyTag/Catcher domains into the beta-hairpin of t1428 to allow functionalisation of the SynRFR scaffold with arbitrary proteins of interest, instead of generating custom fusion constructs for each protein of interest. In addition, constructs with N-terminal SpyTag/Catcher were also designed. While all constructs were able to form an isopeptide bond with their respective interaction partner, insoluble (iST-t1428) and soluble (iSC-t1428) aggregate

formation suggest that the internal fusion constructs need to be optimised e.g. by flexible linker redesign. The single G_4S linker on either side of the SpyTag/Catcher could be too short and leading to instability and aggregation. Alternatively, as discussed by Perez-Riba et al (Perez-Riba et al., 2019), these insertions could lead to cooperative unfolding due to the disruption of inter-repeat interfaces at the insertion site. In addition, multiple protein bands are observed for iSC-t1428 on an SDS-PAGE gel suggesting some level of proteolytic degradation. An alternative optimisation route is to use an internal DogTag, entirely replacing the beta-hairpin as DogTag already forms a hairpin structure (Keeble et al., 2022). The N-terminal SpyTag/Catcher fusion constructs, on the other hand, expressed solubly and could be successfully purified. Both constructs yielded crystals. However, the only structural insight that could be obtained from the resulting crystal structures is that the SynRFR scaffold is intact. Native PAGE data suggests that, when reacted together, nST-Beta1 and nSC-Beta1 could form higher order assemblies. This is a key direction for future work as there is scope for building nano fibres using SpyTag/Catcher at the N-terminus and the homodimerisation interface at the C-terminus. Additionally, SpyTag/Catcher / DogTag/Catcher combinations can be investigated to allow for further functionalisation of the SynRFR scaffold with more than one protein of interest.

4.4.4. SPLIT SOLENOID DESIGN

The idea to split the SynRFR structure into two halves was inspired by the work of Peralta et al (Peralta et al., 2015), who used different beta-solenoid proteins to form fibrils. However, my goal was to arrive at compatible solenoid building blocks, rather than long fibrils. A naive approach of splitting the SynRFR sequence into two parts was attempted. The C-terminal part with (BT) and without a His-Tag (B) has more soluble expression and sample homogeneity than the N-terminal parts (A1 and A2), suggesting that the N-terminal parts are less stable than the C-terminal SynRFR fragment. In addition, based on the SEC data, the C-terminal fragment retains a functional homodimerisation interface. This suggests that at least part of the structure might retain its solenoid fold and the absence of an N-terminal capping structure does not negatively impact solubility and oligomeric state. However, no crystallisation attempts of BT were successful. Instead, circular dichroism can be used to assess the secondary structure of the parts as they might be forming molten globules rather than folded beta-rich structures. Pull-down experiments were inconclusive but suggest that there could be some binding between the two parts. AlphaFold2 models of complexes between part A1/A2 and part B raise the possibility of alternative binding modes to the intended design, where the two parts can still bind to each other, albeit out of register. The addition of salt bridges to the interface as in (Peralta et al., 2015) could increase the enthalpic contribution to the free energy change of interaction and favour a preferred orientation between the two parts. In addition, deep-learning-driven redesign of the interface using AlphaFold and ProteinMPNN could lead to a functional heterodimeric interface. A further future direction would be the design of orthogonal solenoid parts.

5. CONCLUSIONS AND FUTURE WORK

5.1. CONCLUSIONS

The aim of this thesis was to develop bioinformatic tools and databases for solenoid proteins, as well as advancing design/engineering efforts on beta-solenoid proteins.

A new, structural, neural-network-based solenoid detector was developed and trained on previously published data. The method was tested on a new curated testing dataset and was competitive with state-of-the-art methods in solenoid detection so far. This method is fast, provides residue-level solenoid annotation, can process arbitrarily large structures, and is generalisable to other tandem repeat classes on re-training with data for these additional classes. This development coincides with the explosion of high-quality predicted structures by AlphaFold 2 and similar neural-network based methods. An interactive tool for repeat splitting and visualisation was also developed. Taken together, these tools will enable researchers to search for solenoid structures and obtain and refine solenoid region and repeat annotations to expand the solenoid protein knowledge base. These tools could also aid in downstream applications such as design or analysis of solenoid folds. For example, SOLeNNoID could be repurposed as a design filter to select for proteins with a solenoid fold.

The utility of SOLeNNoID was demonstrated by searching the PDB and AlphaFold databases, discovering hundreds to thousands of new solenoid proteins and discussing features of these newly discovered datasets. To my knowledge, this is the first search of the AlphaFold database for solenoid structures to date. The PDB search provided another way to validate the SOLeNNoID method, as the detector finds structures that overlap with entries from existing databases but also new structures outside their scope. The AlphaFold database search allowed the discovery of solenoid structures, which could support findings from previous research or pave the way for new hypotheses. The discovery and analysis of solenoid proteins from the AlphaFold database could enable new developments in protein design, such as the expansion of designed RNA-binding proteins using insights from HPR proteins, and investigation into key proteins in health and disease such as the dgf-1 family from *T. cruzi*.

Finally, multiple beta-solenoid designs were attempted: consensus hexapeptide repeat proteins, loop and domain insertions into the SynRFR scaffold and splitting the SynRFR into two interacting parts. These designs intersect into a beta-solenoid toolkit: the hexapeptide designs are an attempt to generate a new beta-solenoid scaffold similar to SynRFR, the split solenoid designs are an attempt to separate the SynRFR scaffold into smaller, interacting modules, and the loop and domain insertions are an attempt to introduce new structure and function to an otherwise inert domain. Computational methodology was developed for structural repeat extraction and the design of consensus hexapeptide and two-loop SynRFR proteins, albeit with partial experimental validation. In the years since the initial two-loop design, the protein design field has grown and developed at a rapid pace. Many Rosetta design methodologies have now been superseded by

machine learning approaches. With the advent of these new promising machine-learning design methods such as ProteinMPNN, it may be possible to modify and improve these design pipelines. SpyTag and SpyCatcher moieties were successfully inserted into the SynRFR core and at the N-terminus of the protein, whereas split solenoid data was inconclusive but suggested that the C-terminal SynRFR part is likely stable. Putative higher order assemblies of N-terminal SpyTag- and SpyCatcher-SynRFR fusion proteins were observed but warrant further investigation. Overall, these designs would benefit from further and higher-throughput optimisation and investigation. Additional techniques such as circular dichroism and the thermofluor assay can be used to assess secondary structure content as a proxy for correct folding, and melting temperature as a measure of stability, respectively.

5.2. FUTURE WORK

There is scope for improvement in many aspects of the work discussed in this thesis and a range of future directions.

The SOLeNNoID method can be improved by the use of an equivariant transformer architecture, such as the one employed by AlphaFold 2 to directly 'reason' in 3D, rather than using a 2D distance matrix representation of protein structure. This could enhance solenoid prediction performance by better capturing longer-range dependencies between residues in the protein structure. This may also enable direct prediction of repeats rather than individual solenoid residues. In addition, the method can be expanded to more tandem repeat classes with further training data. Additional training data or post-processing of predictions could also reduce the number of false positive alpha-solenoid predictions.

The search for solenoid proteins could also be expanded across the entirety of later versions of the AlphaFold database. Additionally, the current databases of solenoid predictions discussed in this thesis will be made available to researchers. Deeper analysis and a more thorough organisation of the solenoid prediction databases could identify new important or interesting structures and trends across proteomes and provide further infrastructure to support research on solenoid proteins.

Consensus hexapeptide designs could be improved by making a multiple sequence alignment of entire 18-residue repeats rather than individual hexapeptides. Furthermore, protein language models can be fine-tuned on hexapeptide multiple sequence alignments to enhance sequence design. ProteinMPNN can also be used to redesign sequences to increase solubility. Filtering of designs can also be improved by using aggregation propensity and surface charge predictions. In addition, a high-throughput experimental characterisation pipeline can be adopted to test multiple designs in a 96-well format. Circular dichroism could be used to filter designs experimentally and only select proteins with beta-strand structure. The thermofluor assay can then be used to select

the most thermostable designs for further experimental characterisation by X-ray crystallography. Similar improvements can be made to the design pipeline for two-loop SynRFR variants. However, instead of focusing on structured loops, an interesting future direction might be the scaffolding of unstructured but functional loops using the SynRFR protein to present, for example, multiple loop antigens. This could lead to applications of the SynRFR protein as a vaccine scaffold, depending on the innate immunogenicity of SynRFR. Analogous to this, multiple Tag/Catcher systems can be engineered into SynRFR to enable scaffolding together of proteins of arbitrary functionality. For example, using SpyTag, DogTag and SnoopTag in tandem on a SynRFR protein could allow the formation of a multi-enzyme assembly with each enzyme provided as a fusion protein with the respective Catcher. Alternatively, or concomitantly Catcher/Tag systems could allow the formation of SynRFR meshes or materials. Further research into assemblies of N-terminal SpyTag/Catcher fusions is warranted to assess their ability to form fibrous structures. Split solenoid variants could be achieved by modifying the design method. A future direction would be the design of heterodimeric SynRFR bricks which could then be functionalised to arrive at an RFR LEGO®-like system.

Taken together, the results of this thesis comprise a number of computational tools and databases, in addition to experimental results, which provide a launchpad for the future exploration and expansion of the solenoid protein universe.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. http://arxiv.org/abs/1603.04467

Agromayor, M., Soler, N., Caballe, A., Kueck, T., Freund, S. M., Allen, M. D., Bycroft, M., Perisic, O., Ye, Y., McDonald, B., Scheel, H., Hofmann, K., Neil, S. J. D., Martin-Serrano, J., & Williams, R. L. (2012). The UBAP1 subunit of ESCRT-I interacts with ubiquitin via a SOUBA domain. *Structure*, *20*(3), 414–428. https://doi.org/10.1016/J.STR.2011.12.013/ATTACHMENT/17A6B282-8EA7-429D-98E6-BB454A49AA3E/MMC1.PDF

Albus, C. A., Ruf, S., Schöttler, M. A., Lein, W., Kehr, J., & Bock, R. (2010). Y3IP1, a Nucleus-Encoded Thylakoid Protein, Cooperates with the Plastid-Encoded Ycf3 Protein in Photosystem I Assembly of Tobacco and Arabidopsis. *The Plant Cell*, *22*(8), 2838. https://doi.org/10.1105/TPC.110.073908

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. v., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, *13*(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*. http://arxiv.org/abs/1803.01164

Aloysius, N., & Geetha, M. (2018). A review on deep convolutional neural networks. *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, *2018-January*, 588–592. https://doi.org/10.1109/ICCSP.2017.8286426

AlQuraishi, M. (2019). End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, *8*(4), 292-301.e3. https://doi.org/10.1016/j.cels.2019.03.006

Altae-Tran, H., Gao, L., Strecker, J., Macrae, R. K., & Zhang, F. (2020). Computational Identification of Repeat-Containing Proteins and Systems. *QRB Discovery*, *1*. https://doi.org/10.1017/QRD.2020.14

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amunts, A., Brown, A., Toots, J., Scheres, S. H. W., & Ramakrishnan, V. (2015). The structure of the human mitochondrial ribosome. *Science*, *348*(6230), 95–98. https://doi.org/10.1126/SCIENCE.AAA1193/SUPPL_FILE/AMUNTS-SM.PDF

Andrade, M. A., & Bork, P. (1995). HEAT repeats in the Huntington's disease protein. *Nature Genetics 1995 11:2*, *11*(2), 115–116. https://doi.org/10.1038/ng1095-115

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, *42*(D1), D310–D314. https://doi.org/10.1093/NAR/GKT1242

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, *48*(D1), D376–D382. https://doi.org/10.1093/NAR/GKZ1064

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., & Baker, D. (2021). De novo protein design by deep network hallucination. *Nature 2021 600:7889*, *600*(7889), 547–552. https://doi.org/10.1038/s41586-021-04184-w

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. v., van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Baker, E. G., Bartlett, G. J., Porter Goff, K. L., & Woolfson, D. N. (2017). Miniprotein Design: Past, Present, and Prospects. *Accounts of Chemical Research*, *50*(9), 2085–2092. https://doi.org/10.1021/acs.accounts.7b00186

Baldi, P. (2018). Deep Learning in Biomedical Data Science. *Annual Review of Biomedical Data Science*, *1*(1), 181–205. https://doi.org/10.1146/annurev-biodatasci-080917-013343

Barnhart, M. M., & Chapman, M. R. (2006). Curli Biogenesis and Function. *Http://Dx.Doi.Org/10.1146/Annurev.Micro.60.080805.142106*, *60*, 131–147. https://doi.org/10.1146/ANNUREV.MICRO.60.080805.142106

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: the

universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/NAR/GKAA1100

Bateman, A., Murzin, A. G., & Teichmann, S. A. (1998). Structure and distribution of pentapeptide repeats in bacteria. *Protein Science : A Publication of the Protein Society*, *7*(6), 1477–1480. https://doi.org/10.1002/PRO.5560070625

Batyuk, A., Wu, Y., Honegger, A., Heberling, M. M., & Plückthun, A. (2016). DARPin-Based Crystallization Chaperones Exploit Molecular Geometry as a Screening Dimension in Protein Crystallography. *Journal of Molecular Biology*, *428*(8), 1574–1588. https://doi.org/10.1016/J.JMB.2016.03.002

Belkhadir, Y., Subramaniam, R., & Dangl, J. L. (2004). Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Current Opinion in Plant Biology*, *7*(4), 391–399. https://doi.org/10.1016/J.PBI.2004.05.009

Benoit, R. M., Frey, D., Hilbert, M., Kevenaar, J. T., Wieser, M. M., Stirnimann, C. U., McMillan, D., Ceska, T., Lebon, F., Jaussi, R., Steinmetz, M. O., Schertler, G. F. X., Hoogenraad, C. C., Capitani, G., & Kammerer, R. A. (2013). Structural basis for recognition of synaptic vesicle protein 2C by botulinum neurotoxin A. *Nature 2013 505:7481*, *505*(7481), 108–111. https://doi.org/10.1038/nature12732

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. In *Nucleic Acids Research* (Vol. 28, Issue 1, pp. 235–242). Oxford University Press. https://doi.org/10.1093/nar/28.1.235

Beulin, D. S. J., Yamaguchi, M., Kawabata, S., & Ponnuraj, K. (2014). Crystal structure of PfbA, a surface adhesin of Streptococcus pneumoniae, provides hints into its interaction with fibronectin. *International Journal of Biological Macromolecules*, *64*, 168–173. https://doi.org/10.1016/J.IJBIOMAC.2013.11.035

Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., Buchko, G. W., Pulavarti, S. V. S. R. K., Kaas, Q., Eletsky, A., Huang, P.-S., Johnsen, W. A., Greisen, P. J., Rocklin, G. J., Song, Y., Linsky, T. W., Watkins, A., Rettie, S. A., Xu, X., … Baker, D. (2016). Accurate de novo design of hyperstable constrained peptides. *Nature*, *538*(7625), 329–335. https://doi.org/10.1038/nature19791

Biancalana, M., Makabe, K., Koide, A., & Koide, S. (2009). Molecular Mechanism of Thioflavin-T Binding to the Surface of β-Rich Peptide Self-Assemblies. *Journal of Molecular Biology*, *385*(4), 1052–1063. https://doi.org/10.1016/J.JMB.2008.11.006

Biegert, A., & Söding, J. (2008). De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, *24*(6), 807–814. https://doi.org/10.1093/BIOINFORMATICS/BTN039

Binda, C., Bossi, R. T., Wakatsuki, S., Arzt, S., Coda, A., Curti, B., Vanoni, M. A., & Mattevi, A. (2000). Cross-talk and ammonia channeling between active centers in the unexpected domain arrangement of glutamate synthase. *Structure*, *8*(12), 1299–1308. https://doi.org/10.1016/S0969-2126(00)00540-2

Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P., & Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *Journal of Molecular Biology*, *332*(2), 489–503. https://doi.org/10.1016/S0022-2836(03)00896-9

Bliven, S. E., Lafita, A., Rose, P. W., Capitani, G., Prlić, A., & Bourne, P. E. (2019). Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLOS Computational Biology*, *15*(4), e1006842. https://doi.org/10.1371/journal.pcbi.1006842

Bodinier, R., Sabra, A., Leiba, J., Marchetti, A., Lamrabet, O., Ayadi, I., Filić, V., Kawata, T., Weber, I., & Cosson, P. (2021). Role of LrrkA in the Control of Phagocytosis and Cell Motility in Dictyostelium discoideum. *Frontiers in Cell and Developmental Biology*, *9*, 447. https://doi.org/10.3389/FCELL.2021.629200/BIBTEX

Boersma, Y. L., & Plückthun, A. (2011). DARPins and other repeat protein scaffolds: advances in engineering and applications. *Current Opinion in Biotechnology*, *22*(6), 849–857. https://doi.org/10.1016/j.copbio.2011.06.004

Bokeh Development Team (2022). Bokeh: Python library for interactive visualization

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by *AlphaFold 2*: are attention and symmetries all you need? *Acta Crystallographica Section D Structural Biology*, *77*(8), 982–991. https://doi.org/10.1107/S2059798321007531

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, *33*(10), 3390. https://doi.org/10.1093/NAR/GKI615

Brody, M. J., & Lee, Y. (2016). The role of leucine-rich repeat containing protein 10 (LRRC10) in dilated cardiomyopathy. *Frontiers in Physiology*, *7*(AUG), 337. https://doi.org/10.3389/FPHYS.2016.00337/BIBTEX

Broeker, N. K., Gohlke, U., Müller, J. J., Uetrecht, C., Heinemann, U., Seckler, R., & Barbirz, S. (2013). Single amino acid exchange in bacteriophage HK620 tailspike protein results in thousand-fold increase of its oligosaccharide affinity. *Glycobiology*, *23*(1), 59–68. https://doi.org/10.1093/GLYCOB/CWS126

Brown, K., Pompeo, F., Dixon, S., Mengin-Lecreulx, D., Cambillau, C., & Bourne, Y. (1999). Crystal structure of the bifunctional N-acetylglucosamine 1-phosphate uridyltransferase from Escherichia coli: a paradigm for the related pyrophosphorylase superfamily. The EMBO Journal, 18(15), 4096–4107. https://doi.org/10.1093/EMBOJ/18.15.4096

Brunette, T. J., Bick, M. J., Hansen, J. M., Chow, C. M., Kollman, J. M., & Baker, D. (2020). Modular repeat protein sculpting using rigid helical junctions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(16), 8870–8875. https://doi.org/10.1073/PNAS.1908768117/-/DCSUPPLEMENTAL

Brunette, T. J., Parmeggiani, F., Huang, P. S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., & Baker, D. (2015). Exploring the repeat protein universe through computational protein design. *Nature 2015 528:7583*, *528*(7583), 580–584. https://doi.org/10.1038/nature16162

Buchman, C. D., & Hurley, T. D. (2017). Inhibition of the Aldehyde Dehydrogenase 1/2 Family by Psoralen and Coumarin Derivatives. Journal of Medicinal Chemistry, 60(6), 2439–2455. https://doi.org/10.1021/ACS.JMEDCHEM.6B01825/SUPPL_FILE/JM6B01825_SI_002.CSV

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods in Molecular Biology*, *1607*, 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. v., Christie, C. H., Dalenberg, K., di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., … Zhuravleva, M. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, *49*(D1), D437–D451. https://doi.org/10.1093/NAR/GKAA1038

Buurman, E. T., Andrews, B., Gao, N., Hu, J., Keating, T. A., Lahiri, S., Otterbein, L. R., Patten, A. D., Stokes, S. S., & Shapiro, A. B. (2011). In vitro validation of acetyltransferase activity of GlmU as an antibacterial target in Haemophilus influenzae. Journal of Biological Chemistry, 286(47), 40734–40742. https://doi.org/10.1074/jbc.M111.274068

Carrera-Pacheco, S. E., Hankamer, B., & Oey, M. (2020). Light and heat-shock mediated TDA1 overexpression as a tool for controlled high-yield recombinant protein production in Chlamydomonas reinhardtii chloroplasts. *Algal Research*, *48*, 101921. https://doi.org/10.1016/J.ALGAL.2020.101921

Chakrabarty, B., & Parekh, N. (2020). PRIGSA2: Improved version of protein repeat identification by graph spectral analysis. *Journal of Biosciences*, *45*(1), 1–16. https://doi.org/10.1007/S12038-020-00058-X/FIGURES/8

Chakrabarty, B., & Parekh, N. (2021). DbStRiPs: Database of structural repeats in proteins. *Protein Science*. https://doi.org/10.1002/PRO.4052

Chakraborty, S., Nguyen, B., Wasti, S. D., & Xu, G. (2019). Plant Leucine-Rich Repeat Receptor Kinase (LRR-RK): Structure, Ligand Perception, and Activation Mechanism. *Molecules (Basel, Switzerland)*, *24*(17). https://doi.org/10.3390/MOLECULES24173081

Chang, M. P., Huang, W., & Mai, D. J. (2021). Monomer-scale design of functional protein polymers using consensus repeat sequences. *Journal of Polymer Science*, *59*(22), 2644–2664. https://doi.org/10.1002/POL.20210506

Chang, V. T., Fernandes, R. A., Ganzinger, K. A., Lee, S. F., Siebold, C., McColl, J., Jönsson, P., Palayret, M., Harlos, K., Coles, C. H., Jones, E. Y., Lui, Y., Huang, E., Gilbert, R. J. C., Klenerman, D., Aricescu, A. R., & Davis, S. J. (2016). Initiation of T cell signaling by CD45 segregation at "close contacts." *Nature Immunology 2016 17:5*, *17*(5), 574–582. https://doi.org/10.1038/ni.3392

Chen, M., Blum, D., Engelhard, L., Raunser, S., Wagner, R., & Gatsogiannis, C. (2021). Molecular architecture of black widow spider neurotoxins. Nature Communications 2021 12:1, 12(1), 1–14. https://doi.org/10.1038/s41467-021-26562-8

Cheng, J., Hanada, Y., Miura, A., Tsuda, S., & Kondo, H. (2016). Hydrophobic ice-binding sites confer hyperactivity of an antifreeze protein from a snow mold fungus. *Biochemical Journal*, *473*(21), 4011–4026. https://doi.org/10.1042/BCJ20160543

Chollet, F. (2015) keras, GitHub. https://github.com/fchollet/keras

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. https://arxiv.org/abs/1511.07289v5

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). BioPython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Coquille, S., Filipovska, A., Chia, T., Rajappa, L., Lingford, J. P., Razif, M. F. M., Thore, S., & Rackham, O. (2014). An artificial PPR scaffold for programmable RNA recognition. *Nature Communications 2014 5:1*, *5*(1), 1–9. https://doi.org/10.1038/ncomms6729

Correnti, C. E., Hallinan, J. P., Doyle, L. A., Ruff, R. O., Jaeger-Ruckstuhl, C. A., Xu, Y., Shen, B. W., Qu, A., Polkinghorn, C., Friend, D. J., Bandaranayake, A. D., Riddell, S. R., Kaiser, B. K., Stoddard, B. L., & Bradley, P. (2020). Engineering and functionalization of large circular tandem repeat protein nanoparticles. *Nature Structural & Molecular Biology 2020 27:4*, *27*(4), 342–350. https://doi.org/10.1038/s41594-020-0397-5

Creze, C., Castang, S., Derivery, E., Haser, R., Hugouvieux-Cotte-Pattat, N., Shevchik, V. E., & Gouet, P. (2008). The Crystal Structure of Pectate Lyase Pell from Soft Rot Pathogen Erwinia

chrysanthemi in Complex with Its Substrate. *Journal of Biological Chemistry*, *283*(26), 18260–18268. https://doi.org/10.1074/JBC.M709931200

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, *14*(6), 1188–1190. https://doi.org/10.1101/GR.849004

Crystallography: Protein data bank. (1971). *Nature New Biology*, *233*(42), 223–223. https://doi.org/10.1038/NEWBIO233223B0

Cui, M., Qi, Q., Gurry, T., Zhao, T., An, B., Pu, J., Gui, X., Cheng, A. A., Zhang, S., Xun, D., Becce, M., Briatico-Vangosa, F., Liu, C., Lu, T. K., & Zhong, C. (2019). Modular genetic design of multi-domain functional amyloids: insights into self-assembly and functional properties. *Chemical Science*, *10*(14), 4004–4014. https://doi.org/10.1039/C9SC00208A

Cun, L., Henderson, J., le Cun, Y., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, *2*.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., … Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, *378*(6615), 49–56. https://doi.org/10.1126/SCIENCE.ADD2187

de Brevern, A. G. (2016). Extension of the classical classification of β-turns. *Scientific Reports 2016 6:1*, *6*(1), 1–15. https://doi.org/10.1038/srep33191

Delucchi, M., Näf, P., Bliven, S., & Anisimova, M. (2021). TRAL 2.0: Tandem Repeat Detection With Circular Profile Hidden Markov Models and Evolutionary Aligner. *Frontiers in Bioinformatics*, *0*, 20. https://doi.org/10.3389/FBINF.2021.691865

Delucchi, M., Schaper, E., Sachenkova, O., Elofsson, A., & Anisimova, M. (2020). A new census of protein tandem repeats and their relationship with intrinsic disorder. *Genes*, *11*(4), 407. https://doi.org/10.3390/genes11040407

Deng, P., Zhou, Y., Jiang, J., Li, H., Tian, W., Cao, Y., Qin, Y., Kim, J., Roeder, R. G., Patel, D. J., & Wang, Z. (2018). Transcriptional elongation factor Paf1 core complex adopts a spirally wrapped solenoidal topology. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9998–10003. https://doi.org/10.1073/PNAS.1812256115/SUPPL_FILE/PNAS.1812256115.SAPP.PDF

di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A. v., & Tosatto, S. C. E. (2014). RepeatsDB: A database of tandem repeat protein structures. *Nucleic Acids Research*, *42*(D1), 1–6. https://doi.org/10.1093/nar/gkt1175

di Meo, T., Kariyawasam, K., Ghattas, W., Valerio-Lepiniec, M., Sciortino, G., Maréchal, J. D., Minard, P., Mahy, J. P., Urvoas, A., & Ricoux, R. (2019). Functionalized artificial bidomain proteins based on an α-solenoid protein repeat scaffold: A new class of artificial diels-alderases. *ACS Omega*, *4*(2), 4437–4447. https://doi.org/10.1021/ACSOMEGA.8B03448/ASSET/IMAGES/LARGE/AO-2018-03448Y_0005.JPEG

Diamante, A., Chaturbedy, P. K., Rowling, P. J. E., Kumita, J. R., Eapen, R. S., McLaughlin, S. H., Roche, M. de la, Perez-Riba, A., & Itzhaki, L. S. (2021). Engineering mono- and multi-valent inhibitors on a modular scaffold. *Chemical Science*, *12*(3), 880–895. https://doi.org/10.1039/D0SC03175E

Divine, R., Dang, H. v., Ueda, G., Fallas, J. A., Vulovic, I., Sheffler, W., Saini, S., Zhao, Y. T., Raj, I. X., Morawski, P. A., Jennewein, M. F., Homad, L. J., Wan, Y. H., Tooley, M. R., Seeger, F., Etemadi, A., Fahning, M. L., Lazarovits, J., Roederer, A., … Baker, D. (2021). Designed proteins assemble antibodies into modular nanocages. *Science (New York, N.Y.)*, *372*(6537). https://doi.org/10.1126/SCIENCE.ABD9994

do Viet, P., Roche, D. B., & Kajava, A. v. (2015). TAPO: A combined method for the identification of tandem repeats in protein structures. In *FEBS Letters* (Vol. 589, Issue 19, pp. 2611–2619). Elsevier B.V. https://doi.org/10.1016/j.febslet.2015.08.025

Dodonova, S. O., Diestelkoetter-Bachert, P., von Appen, A., Hagen, W. J. H., Beck, R., Beck, M., Wieland, F., & Briggs, J. A. G. (2015). A structure of the COPI coat and the role of coat proteins in membrane vesicle assembly. *Science*, *349*(6244), 195–198. https://doi.org/10.1126/SCIENCE.AAB1121/SUPPL_FILE/DODONOVA.SM.PDF

Dong, R., Peng, Z., Zhang, Y., & Yang, J. (2018). mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, *34*(10), 1719–1725. https://doi.org/10.1093/BIOINFORMATICS/BTX828

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, *21*(16), 3433–3434. https://doi.org/10.1093/BIOINFORMATICS/BTI541

Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L., & Bradley, P. (2015). Rational design of α-helical tandem repeat proteins with closed architectures. *Nature 2015 528:7583*, *528*(7583), 585–588. https://doi.org/10.1038/nature16191

Dumoulin, V., Visin, F., & Box, G. E. P. (2018). *A guide to convolution arithmetic for deep learning*. http://ethanschoonover.com/solarized

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792. https://doi.org/10.1093/NAR/GKH340

Eichinger, I., Pachebat, J. A., Glöckner, G., Rajandream, M. A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B. A., Rivero, F., Bankier, A. T., Lehmann, R., Hamlin, N., Davies, R., … Kuspa, A. (2005). The genome of the social amoeba Dictyostelium discoideum. Nature, 435(7038), 43. https://doi.org/10.1038/NATURE03481

Elgamacy, M., Coles, M., Ernst, P., Zhu, H., Hartmann, M. D., Plückthun, A., & Lupas, A. N. (2018). An Interface-Driven Design Strategy Yields a Novel, Corrugated Protein Architecture. *ACS Synthetic Biology*, *7*(9), 2226–2235. https://doi.org/10.1021/ACSSYNBIO.8B00224/SUPPL_FILE/SB8B00224_SI_001.PDF

Erdos, G., Pajkos, M., & Dosztányi, Z. (2021). IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, *49*(W1), W297–W303. https://doi.org/10.1093/NAR/GKAB408

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, 2021.10.04.463034. https://doi.org/10.1101/2021.10.04.463034

Fournier, D., Palidwor, G. A., Shcherbinin, S., Szengel, A., Schaefer, M. H., Perez-Iratxeta, C., & Andrade-Navarro, M. A. (2013). Functional and genomic analyses of alpha-solenoid proteins. *PLoS ONE*, *8*(11), 79894. https://doi.org/10.1371/journal.pone.0079894

Gebauer, J. M., Köhler, A., Dietmar, H., Gompert, M., Neundorf, I., Zaucke, F., Koch, M., & Baumann, U. (2018). COMP and TSP-4 interact specifically with the novel GXKGHR motif only found in fibrillar collagens. *Scientific Reports 2018 8:1*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-35447-8

Giessen, T. W., & Silver, P. A. (2016). A Catalytic Nanoreactor Based on in Vivo Encapsulation of Multiple Enzymes in an Engineered Protein Nanocompartment. *ChemBioChem*, *17*(20), 1931–1935. https://doi.org/10.1002/cbic.201600431

Glorot, X., Bordes, A., & Statistics, Y. B. B. T.-P. of the F. I. C. on A. I. and. (2011). *Deep Sparse Rectifier Neural Networks* (G. Gordon, D. Dunson, & M. Dudík, Eds.; pp. 315–323). PMLR . http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf

Good, M. C., Greenstein, A. E., Young, T. A., Ng, H. L., & Alber, T. (2004). Sensor Domain of the Mycobacterium tuberculosis Receptor Ser/Thr Protein Kinase, PknD, forms a Highly Symmetric β Propeller. Journal of Molecular Biology, 339(2), 459–469. https://doi.org/10.1016/J.JMB.2004.03.063

Gordiyenko, Y., Schmidt, C., Jennings, M. D., Matak-Vinkovic, D., Pavitt, G. D., & Robinson, C. v. (2014). eIF2B is a decameric guanine nucleotide exchange factor with a γ2ε2 tetrameric core. *Nature Communications 2014 5:1*, *5*(1), 1–12. https://doi.org/10.1038/ncomms4902

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, *28*(5–6), 367–374. https://doi.org/10.1016/J.COMPBIOLCHEM.2004.09.006

Goto-Ito, S., Yamagata, A., Sato, Y., Uemura, T., Shiroshima, T., Maeda, A., Imai, A., Mori, H., Yoshida, T., & Fukai, S. (2018). Structural basis of trans-synaptic interactions between PTPδ and SALMs for inducing synapse formation. Nature Communications 2017 9:1, 9(1), 1–9. https://doi.org/10.1038/s41467-017-02417-z

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, *331*(1), 281–299. https://doi.org/10.1016/S0022-2836(03)00670-3

Great expectations – the potential impacts of AlphaFold DB | EMBL. (n.d.). Retrieved March 16, 2022, from https://www.embl.org/news/science/alphafold-potential-impacts/

Green, O. M., McKenzie, A. R., Shapiro, A. B., Otterbein, L., Ni, H., Patten, A., Stokes, S., Albert, R., Kawatkar, S., & Breed, J. (2012). Inhibitors of acetyltransferase domain of N-acetylglucosamine-1-phosphate-uridyltransferase/glucosamine-1-phosphate-acetyltransferase (GlmU). Part 1: Hit to lead evaluation of a novel arylsulfonamide series. *Bioorganic & Medicinal Chemistry Letters*, *22*(4), 1510–1519. https://doi.org/10.1016/J.BMCL.2012.01.016

Grünwald, M., Lazzaretti, D., & Bono, F. (2013). Structural basis for the nuclear export activity of Importin13. *The EMBO Journal*, *32*(6), 899–913. https://doi.org/10.1038/EMBOJ.2013.29

Gully, B. S., Shah, K. R., Lee, M., Shearston, K., Smith, N. M., Sadowska, A., Blythe, A. J., Bernath-Levin, K., Stanley, W. A., Small, I. D., & Bond, C. S. (2015). The design and structural characterization of a synthetic pentatricopeptide repeat protein. *Acta Crystallographica Section D: Biological Crystallography*, *71*(2), 196–208. https://doi.org/10.1107/S1399004714024869/KW5103SUP1.PDF

Gurunatha, K. L., Fournier, A. C., Urvoas, A., Valerio-Lepiniec, M., Marchi, V., Minard, P., & Dujardin, E. (2016). Nanoparticles Self-Assembly Driven by High Affinity Repeat Protein Pairing. *ACS Nano*, *10*(3), 3176–3185. https://doi.org/10.1021/acsnano.5b04531

Hakim, A., Nguyen, J. B., Basu, K., Zhu, D. F., Thakral, D., Davies, P. L., Isaacs, F. J., Modis, Y., & Meng, W. (2013). Crystal Structure of an Insect Antifreeze Protein and Its Implications for Ice Binding. *Journal of Biological Chemistry*, *288*(17), 12295–12304. https://doi.org/10.1074/JBC.M113.450973

Han, B. W., Herrin, B. R., Cooper, M. D., & Wilson, I. A. (2008). Antigen recognition by variable lymphocyte receptors. *Science*, *321*(5897), 1834–1837. https://doi.org/10.1126/SCIENCE.1162484/SUPPL_FILE/HAN.SOM.PDF

Hassan, R., & Ho, M. (2008). Mesothelin targeted cancer immunotherapy. *European Journal of Cancer*, *44*(1), 46–53. https://doi.org/10.1016/J.EJCA.2007.08.028

Hassan, R., Thomas, A., Alewine, C., Le, D. T., Jaffee, E. M., & Pastan, I. (2016). Mesothelin Immunotherapy for Cancer: Ready for Prime Time? *Journal of Clinical Oncology*, *34*(34), 4171. https://doi.org/10.1200/JCO.2016.68.3672

Hauser, M., Steinegger, M., & Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics (Oxford, England), 32(9), 1323–1330. https://doi.org/10.1093/BIOINFORMATICS/BTW006

Hegde, S. S., Vetting, M. W., Roderick, S. L., Mitchenall, L. A., Maxwell, A., Takiff, H. E., & Blanchard, J. S. (2005). A Fluoroquinolone Resistance Protein from Mycobacterium tuberculosis That Mimics DNA. *Science*, *308*(5727), 1480–1483. https://doi.org/10.1126/science.1110699

Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, *32*(suppl_2), W500–W502. https://doi.org/10.1093/NAR/GKH429

Hemmann, J. L., Wagner, T., Shima, S., & Vorholt, J. A. (2019). Methylofuran is a prosthetic group of the formyltransferase/hydrolase complex and shuttles one-carbon units between two active sites. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(51), 25583–25590. https://doi.org/10.1073/PNAS.1911595116/SUPPL_FILE/PNAS.1911595116.SAPP.PDF

Hillebrand, A., Matz, J. M., Almendinger, M., Müller, K., Matuschewski, K., & Schmitz-Linneweber, C. (2018). Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. *Nucleic Acids Research*, *46*(19), 10417–10431. https://doi.org/10.1093/NAR/GKY710

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*. https://doi.org/10.48550/arxiv.1207.0580

Hirsh, L., Paladin, L., Piovesan, D., & Tosatto, S. C. E. (2018). RepeatsDB-lite: A web server for unit annotation of tandem repeat proteins. *Nucleic Acids Research*, *46*(W1), W402–W407. https://doi.org/10.1093/nar/gky360

Hirsh, L., Piovesan, D., Paladin, L., & Tosatto, S. C. E. (2016). Identification of repetitive units in protein structures with ReUPred. *Amino Acids*, *48*(6), 1391–1400. https://doi.org/10.1007/s00726-016-2187-2

Holm, L. (2020). DALI and the persistence of protein shape. *Protein Science*, *29*(1), 128–140. https://doi.org/10.1002/pro.3749

Howard, A. E., Fox, J. C., & Slep, K. C. (2015). Drosophila melanogaster Mini Spindles TOG3 Utilizes Unique Structural Elements to Promote Domain Stability and Maintain a TOG1- and TOG2-like Tubulin-binding Surface *. *Journal of Biological Chemistry*, *290*(16), 10149–10162. https://doi.org/10.1074/JBC.M114.633826

Hrabe, T., & Godzik, A. (2014). ConSole: Using modularity of Contact maps to locate Solenoid domains in protein structures. *BMC Bioinformatics*, *15*(1), 119. https://doi.org/10.1186/1471-2105-15-119

Hsia, Y., Mout, R., Sheffler, W., Edman, N. I., Vulovic, I., Park, Y. J., Redler, R. L., Bick, M. J., Bera, A. K., Courbet, A., Kang, A., Brunette, T. J., Nattermann, U., Tsai, E., Saleem, A., Chow, C. M., Ekiert, D., Bhabha, G., Veesler, D., & Baker, D. (2021). Design of multi-scale protein complexes by hierarchical building block fusion. *Nature Communications 2021 12:1*, *12*(1), 1–10. https://doi.org/10.1038/s41467-021-22276-z

Hsu, S.-H., Chu, C.-H., Tian, Y.-C., Chang, M.-Y., Chou, L.-F., Sun, Y.-J., & Yang, C.-W. (2020). *Crystal structure of Leptospira leucine-rich repeat 20 reveals a novel E-cadherin binding protein to induce NGAL expression in HK2 cells*. https://doi.org/10.1042/BCJ20200547

Huang, B., Guo, Q., Niedermeier, M. L., Cheng, J., Engler, T., Maurer, M., Pautsch, A., Baumeister, W., Stengel, F., Kochanek, S., & Fernández-Busnadiego, R. (2021). Pathological polyQ expansion does not alter the conformation of the Huntingtin-HAP40 complex. Structure, 29(8), 804-809.e5. https://doi.org/10.1016/j.str.2021.04.003

Huang, P. S., Ban, Y. E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R., & Baker, D. (2011). RosettaRemodel: a generalized framework for flexible backbone protein design. *PloS One*, *6*(8). https://doi.org/10.1371/JOURNAL.PONE.0024109

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/JPHYSIOL.1968.SP008455

Hufnagel, D. A., Depas, W. H., & Chapman, M. R. (2015). The Biology of the Escherichia coli Extracellular Matrix . *Microbiology Spectrum*, *3*(3). https://doi.org/10.1128/MICROBIOLSPEC.MB-0014-2014

Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X. H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernández, I. S., Szakal, B., … Baker, D. (2021). Computed structures of core eukaryotic protein complexes. *Science*, *374*(6573). https://doi.org/10.1126/SCIENCE.ABM4805/SUPPL_FILE/SCIENCE.ABM4805_PREDICTED_PROTEIN_PROTEIN_INTERACTIONS.ZIP

Itoh, T., Intuy, R., Suyotha, W., Hayashi, J., Yano, S., Makabe, K., Wakayama, M., & Hibi, T. (2020). Structural insights into substrate recognition and catalysis by glycoside hydrolase family 87 α-1,3-glucanase from Paenibacillus glycanilyticus FH11. *The FEBS Journal*, *287*(12), 2524–2543. https://doi.org/10.1111/FEBS.15161

Jackson, V. A., Mehmood, S., Chavent, M., Roversi, P., Carrasquero, M., del Toro, D., Seyit-Bremer, G., Ranaivoson, F. M., Comoletti, D., Sansom, M. S. P., Robinson, C. v., Klein, R., & Seiradake, E. (2016). Super-complexes of adhesion GPCRs and neural guidance receptors. *Nature Communications 2016 7:1*, *7*(1), 1–13. https://doi.org/10.1038/ncomms11184

Javadi, Y., & Itzhaki, L. S. (2013). Tandem-repeat proteins: Regularity plus modularity equals design-ability. In *Current Opinion in Structural Biology* (Vol. 23, Issue 4, pp. 622–631). Elsevier Current Trends. https://doi.org/10.1016/j.sbi.2013.06.011

Jendrusch, M., Korbel Id, J. O., Kashif, S., & Id, S. (n.d.). *AlphaDesign: A de novo protein design framework based on AlphaFold*. https://doi.org/10.1101/2021.10.11.463937

Jin, P., Bulkley, D., Guo, Y., Zhang, W., Guo, Z., Huynh, W., Wu, S., Meltzer, S., Cheng, T., Jan, L. Y., Jan, Y. N., & Cheng, Y. (2017). Electron cryo-microscopy structure of the mechanotransduction channel NOMPC. *Nature 2017 547:7661*, *547*(7661), 118–122. https://doi.org/10.1038/nature22981

Jolie, R. P., Duvetter, T., van Loey, A. M., & Hendrickx, M. E. (2010). Pectin methylesterase and its proteinaceous inhibitor: a review. *Carbohydrate Research*, *345*(18), 2583–2595. https://doi.org/10.1016/J.CARRES.2010.10.002

Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(suppl_1), D411–D419. https://doi.org/10.1093/NAR/GKQ1105

Jorda, J., & Kajava, A. v. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, *25*(20), 2632–2638. https://doi.org/10.1093/BIOINFORMATICS/BTP482

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLOS ONE*, *7*(8), e41882. https://doi.org/10.1371/JOURNAL.PONE.0041882

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/BIP.360221211

Kajava, A. v. (2012). Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, *179*(3), 279–288. https://doi.org/10.1016/j.jsb.2011.08.009

Kajava, A. v., & Steven, A. C. (2006). β-Rolls, β-Helices, and Other β-Solenoid Proteins. In *Advances in Protein Chemistry* (Vol. 73, pp. 55–96). https://doi.org/10.1016/S0065-3233(06)73003-0

Kang, J. Y., Nan, X., Jin, M. S., Youn, S. J., Ryu, Y. H., Mah, S., Han, S. H., Lee, H., Paik, S. G., & Lee, J. O. (2009). Recognition of Lipopeptide Patterns by Toll-like Receptor 2-Toll-like Receptor 6 Heterodimer. *Immunity*, *31*(6), 873–884. https://doi.org/10.1016/J.IMMUNI.2009.09.018

Kang, Y., Gohlke, U., Engström, O., Hamark, C., Scheidt, T., Kunstmann, S., Heinemann, U., Widmalm, G., Santer, M., & Barbirz, S. (2016). Bacteriophage Tailspikes and Bacterial O-Antigens as a Model System to Study Weak-Affinity Protein-Polysaccharide Interactions. Journal of the American Chemical Society, 138(29), 9109–9118. https://doi.org/10.1021/JACS.6B00240/SUPPL_FILE/JA6B00240_SI_001.PDF

Kashiwagi, K., Takahashi, M., Nishimoto, M., Hiyama, T. B., Higo, T., Umehara, T., Sakamoto, K., Ito, T., & Yokoyama, S. (2016). Crystal structure of eukaryotic translation initiation factor 2B. *Nature 2016 531:7592*, *531*(7592), 122–125. https://doi.org/10.1038/nature16991

Kawashita, S. Y., da Silva, C. v., Mortara, R. A., Burleigh, B. A., & Briones, M. R. S. (2009). Homology, paralogy and function of DGF-1, a highly dispersed Trypanosoma cruzi specific gene family and its implications for information entropy of its encoded proteins. *Molecular and Biochemical Parasitology*, *165*(1), 19–31. https://doi.org/10.1016/J.MOLBIOPARA.2008.12.010

Keeble, A. H., Banerjee, A., Ferla, M. P., Reddington, S. C., Anuar, I. N. A. K., & Howarth, M. (2017). Evolving Accelerated Amidation by SpyTag/SpyCatcher to Analyze Membrane Dynamics. *Angewandte Chemie International Edition*, *56*(52), 16521–16525. https://doi.org/10.1002/anie.201707623

Keeble, A. H., Yadav, V. K., Ferla, M. P., Bauer, C. C., Chuntharpursat-Bon, E., Huang, J., Bon, R. S., & Howarth, M. (2022). DogCatcher allows loop-friendly protein-protein ligation. *Cell Chemical Biology*, *29*(2), 339-350.e10. https://doi.org/10.1016/J.CHEMBIOL.2021.07.005

Kempen, M. van, Kim, S. S., Tumescheit, C., Mirdita, M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. *BioRxiv*, 2022.02.07.479398. https://doi.org/10.1101/2022.02.07.479398

Kenner, L. R., Anand, A. A., Nguyen, H. C., Myasnikov, A. G., Klose, C. J., McGeever, L. A., Tsai, J. C., Miller-Vedam, L. E., Walter, P., & Frost, A. (2019). EIF2B-catalyzed nucleotide exchange and

phosphoregulation by the integrated stress response. *Science*, *364*(6439), 491–495. https://doi.org/10.1126/SCIENCE.AAW2922/SUPPL_FILE/AAW2922S1.MOV

Kent, L. M., Loo, T. S., Melton, L. D., Mercadante, D., Williams, M. A. K., & Jameson, G. B. (2016). Structure and Properties of a Non-processive, Salt-requiring, and Acidophilic Pectin Methylesterase from Aspergillus niger Provide Insights into the Key Determinants of Processivity Control. *Journal of Biological Chemistry*, *291*(3), 1289–1306. https://doi.org/10.1074/JBC.M115.673152

Khairil Anuar, I. N. A., Banerjee, A., Keeble, A. H., Carella, A., Nikov, G. I., & Howarth, M. (2019). Spy&Go purification of SpyTag-proteins using pseudo-SpyCatcher to access an oligomerization toolbox. *Nature Communications*, *10*(1), 1734. https://doi.org/10.1038/s41467-019-09678-w

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kisker, C., Schindelin, H., Alber, B. E., Ferry, J. G., & Rees, D. C. (1996). A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon Methanosarcina thermophila. The EMBO Journal, 15(10), 2323–2330. https://doi.org/10.1002/J.1460-2075.1996.TB00588.X

Kobe, B., & Kajava, A. v. (2000). When protein folding is simplified to protein coiling: The continuum of solenoid protein structures. In *Trends in Biochemical Sciences* (Vol. 25, Issue 10, pp. 509–515). https://doi.org/10.1016/S0968-0004(00)01667-4

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (n.d.). *ImageNet Classification with Deep Convolutional Neural Networks*. Retrieved February 7, 2022, from http://code.google.com/p/cuda-convnet/

Kruijt, M., de Kock, M. J. D., & de Wit, P. J. G. M. (2005). Receptor-like proteins involved in plant disease resistance. *Molecular Plant Pathology*, *6*(1), 85–97. https://doi.org/10.1111/J.1364-3703.2004.00264.X

Kubiak, R. L., & Holden, H. M. (2012). Structural studies of AntD: An N-acyltransferase involved in the biosynthesis of d-anthrose. Biochemistry, 51(4), 867–878. https://doi.org/10.1021/BI201650C/ASSET/IMAGES/MEDIUM/BI-2011-01650C_0011.GIF

Kühlbrandt, W. (2014). The Resolution Revolution. *Science*, *343*(6178), 1443–1444. https://doi.org/10.1126/SCIENCE.1251652

Lau, W. C. Y., Li, Y., Liu, Z., Gao, Y., Zhang, Q., & Huen, M. S. Y. (2016). Structure of the human dimeric ATM kinase. Http://Dx.Doi.Org/10.1080/15384101.2016.1158362, 15(8), 1117–1124. https://doi.org/10.1080/15384101.2016.1158362

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., … Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, *487*(C), 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6

Lee, S. C., Park, K., Han, J., Lee, J. J., Kim, H. J., Hong, S., Heu, W., Kim, Y. J., Ha, J. S., Lee, S. G., Cheong, H. K., Jeon, Y. H., Kim, D., & Kim, H. S. (2012). Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3299–3304. https://doi.org/10.1073/PNAS.1113193109/SUPPL_FILE/PNAS.1113193109_SI.PDF

Li, C., Guo, X., Jia, Z., Xia, B., & Jin, C. (2005). Solution Structure of an Antifreeze Protein CfAFP-501 from Choristoneura fumiferana. Journal of Biomolecular NMR 2005 32:3, 32(3), 251–256. https://doi.org/10.1007/S10858-005-8206-3

Lin, D. Y. W., Diao, J., Zhou, D., & Chen, J. (2011). Biochemical and structural studies of a HECT-like ubiquitin ligase from Escherichia coli O157:H7. Journal of Biological Chemistry, 286(1), 441–449. https://doi.org/10.1074/jbc.M110.167643

Liu, L., Chen, X., Li, J., Wang, H., Buehl, C. J., Goff, N. J., Meek, K., Yang, W., & Gellert, M. (2022). Autophosphorylation transforms DNA-PK from protecting to processing DNA ends. Molecular Cell, 82(1), 177-189.e4. https://doi.org/10.1016/j.molcel.2021.11.025

Luis, A. S., Briggs, J., Zhang, X., Farnell, B., Ndeh, D., Labourel, A., Baslé, A., Cartmell, A., Terrapon, N., Stott, K., Lowe, E. C., McLean, R., Shearer, K., Schückel, J., Venditto, I., Ralet, M. C., Henrissat, B., Martens, E. C., Mosimann, S. C., … Gilbert, H. J. (2017). Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nature Microbiology 2017 3:2*, *3*(2), 210–219. https://doi.org/10.1038/s41564-017-0079-1

Ma, J., Tang, W. K., Esser, L., Pastan, I., & Xia, D. (2012). Recognition of Mesothelin by the Therapeutic Antibody MORAb-009: STRUCTURAL AND MECHANISTIC INSIGHTS *. *Journal of Biological Chemistry*, *287*(40), 33123–33131. https://doi.org/10.1074/JBC.M112.381756

MacDonald, J. T., & Freemont, P. S. (2016). Computational protein design with backbone plasticity. *Biochemical Society Transactions*, *44*(5), 1523. https://doi.org/10.1042/BST20160155

MacDonald, J. T., Kabasakal, B. v, Godding, D., Kraatz, S., Henderson, L., Barber, J., Freemont, P. S., & Murray, J. W. (2016). Synthetic beta-solenoid proteins with the fragment-free computational

design of a beta-hairpin extension. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(37), 10346–10351. https://doi.org/10.1073/pnas.1525308113

MacDonald, J. T., Kelley, L. A., & Freemont, P. S. (2013). Validating a Coarse-Grained Potential Energy Function through Protein Loop Modelling. *PLOS ONE*, *8*(6), e65770. https://doi.org/10.1371/JOURNAL.PONE.0065770

Madden, S. K., Perez-Riba, A., & Itzhaki, L. S. (2019). Exploring new strategies for grafting binding peptides onto protein loops using a consensus-designed tetratricopeptide repeat scaffold. *Protein Science*, *28*(4), 738–745. https://doi.org/10.1002/PRO.3586

Main, E. R. G., Jackson, S. E., & Regan, L. (2003). The folding and design of repeat proteins: reaching a consensus. *Current Opinion in Structural Biology*, *13*(4), 482–489. https://doi.org/10.1016/S0959-440X(03)00105-2

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722–2728. https://doi.org/10.1093/BIOINFORMATICS/BTT473

Marsella, L., Sirocco, F., Trovato, A., Seno, F., & Tosatto, S. C. E. (2009). REPETITA: Detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, *25*(12), 289–295. https://doi.org/10.1093/bioinformatics/btp232

Martin, R., Qi, T., Zhang, H., Liu, F., King, M., Toth, C., Nogales, E., & Staskawicz, B. J. (2020). Structure of the activated ROQ1 resistosome directly recognizing the pathogen effector XopQ. Science, 370(6521). https://doi.org/10.1126/SCIENCE.ABD9993/SUPPL_FILE/ABD9993-MARTIN-SM.PDF

Marze, N. A., Roy Burman, S. S., Sheffler, W., & Gray, J. J. (2018). Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics (Oxford, England)*, *34*(20), 3461–3469. https://doi.org/10.1093/BIOINFORMATICS/BTY355

McCoy, A. J., Sammito, M. D., & Read, R. J. (2022). Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallographica. Section D, Structural Biology*, *78*(1), 1–13. https://doi.org/10.1107/S2059798321012122/QG5003SUP1.PDF

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *BioRxiv*, 2021.07.09.450648. https://doi.org/10.1101/2021.07.09.450648

Mejias, S. H., Bahrami-Dizicheh, Z., Liutkus, M., Sommer, D. J., Astashkin, A., Kodis, G., Ghirlanda, G., & Cortajarena, A. L. (2019). Repeat proteins as versatile scaffolds for arrays of

redox-active FeS clusters. *Chemical Communications (Cambridge, England)*, *55*(23), 3319. https://doi.org/10.1039/C8CC06827E

Mejías, S. H., López-Andarias, J., Sakurai, T., Yoneda, S., Erazo, K. P., Seki, S., Atienza, C., Martín, N., & Cortajarena, A. L. (2016). Repeat protein scaffolds: ordering photo- and electroactive molecules in solution and solid state. *Chemical Science*, *7*(8), 4842–4847. https://doi.org/10.1039/C6SC01306F

Menéndez-Conejero, R., Nguyen, T. H., Singh, A. K., Condezo, G. N., Marschang, R. E., van Raaij, M. J., & San Martín, C. (2017). Structure of a Reptilian Adenovirus Reveals a Phage Tailspike Fold Stabilizing a Vertebrate Virus Capsid. *Structure*, *25*(10), 1562-1573.e5. https://doi.org/10.1016/J.STR.2017.08.007

Meng, G., Spahich, N., Kenjale, R., Waksman, G., & St Geme, J. W. (2011). Crystal structure of the Haemophilus influenzae Hap adhesin reveals an intercellular oligomerization mechanism for bacterial aggregation. The EMBO Journal, 30(18), 3864–3874. https://doi.org/10.1038/EMBOJ.2011.279

Michel, E., Plückthun, A., & Zerbe, O. (2018). Peptide-Guided Assembly of Repeat Protein Fragments. *Angewandte Chemie International Edition*, *57*(17), 4576–4579. https://doi.org/10.1002/ANIE.201713377

Michel, E., Plückthun, A., & Zerbe, O. (2019). Peptide binding affinity redistributes preassembled repeat protein fragments. *Biol. Chem*, *400*(3), 395–404. https://doi.org/10.1515/hsz-2018-0355

Miras, I., Saul, F., Nowakowski, M., Weber, P., Haouz, A., Shepard, W., & Picardeau, M. (2015). Structural characterization of a novel subfamily of leucine-rich repeat proteins from the human pathogen Leptospira interrogans. Acta Crystallographica Section D: Biological Crystallography, 71(6), 1351–1359. https://doi.org/10.1107/S139900471500704X/MH5169SUP2.PDF

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods 2022 19:6*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/NAR/GKAA913

Mohan, K., Ueda, G., Kim, A. R., Jude, K. M., Fallas, J. A., Guo, Y., Hafer, M., Miao, Y., Saxton, R. A., Piehler, J., Sankaran, V. G., Baker, D., & Garcia, K. C. (2019). Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. *Science (New York, N.Y.)*, *364*(6442). https://doi.org/10.1126/SCIENCE.AAV7532

Mosavi, L. K., Cammett, T. J., Desrosiers, D. C., & Peng, Z. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Protein Science : A Publication of the Protein Society*, *13*(6), 1435–1448. https://doi.org/10.1110/PS.03554604

Mosavi, L. K., Minor, D. L., & Peng, Z. yu. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 16029–16034. https://doi.org/10.1073/PNAS.252537899

Naver, H., Boudreau, E., & Rochaix, J.-D. (2001). Functional Studies of Ycf3 Its Role in Assembly of Photosystem I and Interactions with Some of Its Subunits. *The Plant Cell*, *13*(12), 2731–2745. https://doi.org/10.1105/TPC.010253

Nellaepalli, S., Ozawa, S. I., Kuroda, H., & Takahashi, Y. (2018). The photosystem I assembly apparatus consisting of Ycf3–Y3IP1 and Ycf4 modules. *Nature Communications 2018 9:1*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-04823-3

Nguyen, H., Case, D. A., & Rose, A. S. (2018). NGLview–interactive molecular graphics for Jupyter notebooks. *Bioinformatics*, *34*(7), 1241–1242. https://doi.org/10.1093/BIOINFORMATICS/BTX789

Nguyen, P. Q., Botyanszki, Z., Tay, P. K. R., & Joshi, N. S. (2014). Programmable biofilm-based materials from engineered curli nanofibres. *Nature Communications 2014 5:1*, *5*(1), 1–10. https://doi.org/10.1038/ncomms5945

Ni, S., Sheldrick, G. M., Benning, M. M., & Kennedy, M. A. (2009). The 2 Å resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp. strain PCC 7120. *Journal of Structural Biology*, *165*(1), 47–52. https://doi.org/10.1016/J.JSB.2008.09.010

Noguchi, H., Addy, C., Simoncini, D., Wouters, S., Mylemans, B., van Meervelt, L., Schiex, T., Zhang, K. Y. J., Tame, J. R. H., & Voet, A. R. D. (2019). Computational design of symmetrical eight-bladed β-propeller proteins. *IUCrJ*, *6*(1), 46–55. https://doi.org/10.1107/S205225251801480X/JT5028SUP1.PDF

O'Day, D. H., Suhre, K., Myre, M. A., Chatterjee-Chakraborty, M., & Chavez, S. E. (2006). Isolation, characterization, and bioinformatic analysis of calmodulin-binding protein cmbB reveals a novel tandem IP22 repeat common to many Dictyostelium and Mimivirus proteins. *Biochemical and Biophysical Research Communications*, *346*(3), 879–888. https://doi.org/10.1016/J.BBRC.2006.05.204

O'donnell, J. P., Phillips, B. P., Yagita, Y., Juszkiewicz, S., Wagner, A., Malinverni, D., Keenan, R. J., Miller, E. A., & Hegde, R. S. (2020). The architecture of EMC reveals a path for membrane protein insertion. *ELife*, *9*, 1–56. https://doi.org/10.7554/ELIFE.57887

Okada, C., Yamashita, E., Lee, S. J., Shibata, S., Katahira, J., Nakagawa, A., Yoneda, Y., & Tsukihara, T. (2009). A high-Resolution structure of the pre-microrna nuclear export machinery.

Science, 326(5957), 1275–1279. https://doi.org/10.1126/SCIENCE.1178705/SUPPL_FILE/OKADA.SOM.PDF

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), e00024.001. https://doi.org/10.23915/distill.00024.001

Otto, B. R., Sijbrandi, R., Luirink, J., Oudega, B., Heddle, J. G., Mizutani, K., Park, S. Y., & Tame, J. R. H. (2005). Crystal structure of hemoglobin protease, a heme binding autotransporter protein from pathogenic Escherichia coli. Journal of Biological Chemistry, 280(17), 17339–17345. https://doi.org/10.1074/jbc.M412885200

Pagès, G., Grudinin, S., & Valencia, A. (2019). DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*, *35*(24), 5113–5120. https://doi.org/10.1093/bioinformatics/btz454

Paladin, L., Bevilacqua, M., Errigo, S., Piovesan, D., Mičetić, I., Necci, M., Monzon, A. M., Fabre, M. L., Lopez, J. L., Nilsson, J. F., Rios, J., Menna, P. L., Cabrera, M., Buitron, M. G., Kulik, M. G., Fernandez-Alberti, S., Fornasari, M. S., Parisi, G., Lagares, A., … Tosatto, S. C. E. (2021). RepeatsDB in 2021: Improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*, *49*(D1), D452–D457. https://doi.org/10.1093/nar/gkaa1097

Paladin, L., Hirsh, L., Piovesan, D., Andrade-Navarro, M. A., Kajava, A. v., & Tosatto, S. C. E. (2017). RepeatsDB 2.0: Improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Research*, *45*(D1), D308–D312. https://doi.org/10.1093/nar/gkw1136

Palidwor, G. A., Shcherbinin, S., Huska, M. R., Rasko, T., Stelzl, U., Arumughan, A., Foulle, R., Porras, P., Sanchez-Pulido, L., Wanker, E. E., & Andrade-Navarro, M. A. (2009). Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Computational Biology*, *5*(3), 1000304. https://doi.org/10.1371/journal.pcbi.1000304

Pan, W., Sun, K., Tang, K., Xiao, Q., Ma, C., Yu, C., & Wei, Z. (2018). Structural insights into ankyrin repeat-mediated recognition of the kinesin motor protein KIF21A by KANK1, a scaffold protein in focal adhesion. *Journal of Biological Chemistry*, *293*(6), 1944–1956. https://doi.org/10.1074/JBC.M117.815779/ATTACHMENT/35F1E5A8-0D65-417E-B554-C69621E05FA8/MMC1.DOCX

Park, H. M., Park, J. H., Choi, J. W., Lee, J., Kim, B. Y., Jung, C. H., & Kim, J. S. (2012). Structures of the -class carbonic anhydrase homologue YrdA suggest a possible allosteric switch. Acta Crystallographica Section D: Biological Crystallography, 68(8), 920–926. https://doi.org/10.1107/S0907444912017210/RR5008SUP1.PDF

Park, K., Shen, B. W., Parmeggiani, F., Huang, P. S., Stoddard, B. L., & Baker, D. (2015). Control of repeat-protein curvature by computational protein design. *Nature Structural & Molecular Biology 2014 22:2*, *22*(2), 167–174. https://doi.org/10.1038/nsmb.2938

Parker, R., Mercedes-Camacho, A., & Grove, T. Z. (2014). Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Science : A Publication of the Protein Society*, *23*(6), 790. https://doi.org/10.1002/PRO.2461

Parmeggiani, F., & Huang, P. S. (2017). Designing repeat proteins: a modular approach to protein design. In *Current Opinion in Structural Biology* (Vol. 45, pp. 116–123). Elsevier Ltd. https://doi.org/10.1016/j.sbi.2017.02.001

Parmeggiani, F., Huang, P. S., Vorobiev, S., Xiao, R., Park, K., Caprari, S., Su, M., Seetharaman, J., Mao, L., Janjua, H., Montelione, G. T., Hunt, J., & Baker, D. (2015). A General Computational Approach for Repeat Protein Design. *Journal of Molecular Biology*, *427*(2), 563. https://doi.org/10.1016/J.JMB.2014.11.005

Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A., & Plückthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *Journal of Molecular Biology*, *376*(5), 1282–1304. https://doi.org/10.1016/J.JMB.2007.12.014

Parra, R. G., Espada, R., Sánchez, I. E., Sippl, M. J., & Ferreiro, D. U. (2013). Detecting repetitions and periodicities in proteins by tiling the structural space. *Journal of Physical Chemistry B*, *117*(42), 12887–12897. https://doi.org/10.1021/JP402105J/SUPPL_FILE/JP402105J_SI_002.PDF

Paulsen, C. E., Armache, J. P., Gao, Y., Cheng, Y., & Julius, D. (2015). Structure of the TRPA1 ion channel suggests regulatory mechanisms. Nature 2015 520:7548, 520(7548), 511–517. https://doi.org/10.1038/nature14367

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pellegrini, M. (2015). Tandem repeats in proteins: Prediction algorithms and biological role. In *Frontiers in Bioengineering and Biotechnology* (Vol. 3, Issue SEP). Frontiers Media S.A. https://doi.org/10.3389/fbioe.2015.00143

Peng, W. C., deLau, W., Forneris, F., Granneman, J. C. M., Huch, M., Clevers, H., & Gros, P. (2013). Structure of Stem Cell Growth Factor R-spondin 1 in Complex with the Ectodomain of Its Receptor LGR5. *Cell Reports*, *3*(6), 1885–1892. https://doi.org/10.1016/J.CELREP.2013.06.009

Pepinsky, R. B., Arndt, J. W., Quan, C., Gao, Y., Quintero-Monzon, O., Lee, X., & Mi, S. (2014). Structure of the LINGO-1–Anti-LINGO-1 Li81 Antibody Complex Provides Insights into the Biology of LINGO-1 and the Mechanism of Action of the Antibody Therapy. Journal of Pharmacology and Experimental Therapeutics, 350(1), 110–123. https://doi.org/10.1124/JPET.113.211771

Peralta, M. D. R., Karsai, A., Ngo, A., Sierra, C., Fong, K. T., Hayre, N. R., Mirzaee, N., Ravikumar, K. M., Kluber, A. J., Chen, X., Liu, G. Y., Toney, M. D., Singh, R. R., & Cox, D. L. (2015). Engineering amyloid fibrils from β-solenoid proteins for biomaterials applications. *ACS Nano*, *9*(1), 449–463. https://doi.org/10.1021/nn5056089

Perez-Riba, A., Komives, E., Main, E. R. G., & Itzhaki, L. S. (2019). Decoupling a tandem-repeat protein: Impact of multiple loop insertions on a modular scaffold. *Scientific Reports 2019 9:1*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-49905-4

Perovic, V., Leclercq, J. Y., Sumonja, N., Richard, F. D., Veljkovic, N., & Kajava, A. v. (2020). Tally-2.0: Upgraded validator of tandem repeat detection in protein sequences. *Bioinformatics*, *36*(10), 3260–3262. https://doi.org/10.1093/bioinformatics/btaa121

Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology. *EMBO Reports*, *22*(11). https://doi.org/10.15252/embr.202154046

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Science : A Publication of the Protein Society, 30(1), 70–82. https://doi.org/10.1002/PRO.3943

Porebski, B. T., & Buckle, A. M. (2016). Consensus protein design. *Protein Engineering Design and Selection*, *29*(7), 245–251. https://doi.org/10.1093/protein/gzw015

Porta-Pardo, E., Ruiz-Serra, V., Valentini, S., & Valencia, A. (2022). The structural coverage of the human proteome before and after AlphaFold. *PLoS Computational Biology*, *18*(1). https://doi.org/10.1371/JOURNAL.PCBI.1009818

Prabha, A., & Balaji, P. v. (2021). Characterization of left-handed beta helix-domains, and identification and functional annotation of proteins containing such domains. *Proteins: Structure, Function, and Bioinformatics*, *89*(1), 6–20. https://doi.org/10.1002/PROT.25990

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., & Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(Database issue). https://doi.org/10.1093/NAR/GKR1065

Pyles, H., Zhang, S., de Yoreo, J. J., & Baker, D. (2019). Controlling protein assembly on inorganic crystals through designed protein interfaces. *Nature 2019 571:7764*, *571*(7764), 251–256. https://doi.org/10.1038/s41586-019-1361-6

Qin, L., Tang, L., Xu, J., Zhang, X., Zhu, Y., Zhang, C., Wang, M., Liu, X., Li, F., Sun, F., Su, M., Zhai, Y., & Chen, Y. (2022). Cryo-EM structure and electrophysiological characterization of ALMT from Glycine max reveal a previously uncharacterized class of anion channels. *Science Advances*, *8*(9), 3238. https://doi.org/10.1126/SCIADV.ABM3238

Rämisch, S., Pramhed, A., Tillgren, V., Aspberg, A., & Logan, D. T. (2017). Crystal structure of human chondroadherin: Solving a difficult molecular-replacement problem using de novo models. *Acta Crystallographica Section D: Structural Biology*, *73*(1), 53–63. https://doi.org/10.1107/S205979831601980X/HTTPS://JOURNALS.IUCR.ORG/SERVICES/RSS.HTML

Rämisch, S., Weininger, U., Martinsson, J., Akke, M., & André, I. (2014). Computational design of a leucine-rich repeat protein with a predefined geometry. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(50), 17875–17880. https://doi.org/10.1073/PNAS.1413638111/SUPPL_FILE/PNAS.1413638111.SD01.DOCX

Reichen, C., Hansen, S., Forzani, C., Honegger, A., Fleishman, S. J., Zhou, T., Parmeggiani, F., Ernst, P., Madhurantakam, C., Ewald, C., Mittl, P. R. E., Zerbe, O., Baker, D., Caflisch, A., & Plückthun, A. (2016). Computationally Designed Armadillo Repeat Proteins for Modular Peptide Recognition. *Journal of Molecular Biology*, *428*(22), 4467–4489. https://doi.org/10.1016/J.JMB.2016.09.012

Reichen, C., Madhurantakam, C., Plückthun, A., & Mittl, P. R. E. (2014). Crystal structures of designed armadillo repeat proteins: Implications of construct design and crystallization conditions on overall structure. *Protein Science : A Publication of the Protein Society*, *23*(11), 1572. https://doi.org/10.1002/PRO.2535

Richards, M. W., Edward, E. W., Rennalls, L. P., Busacca, S., O'Regan, L., Fry, A. M., Fennell, D. A., & Bayliss, R. (2014). Crystal structure of EML1 reveals the basis for Hsp90 dependence of oncogenic EML4-ALK by disruption of an atypical β-propeller domain. Proceedings of the National Academy of Sciences of the United States of America, 111(14), 5195–5200. https://doi.org/10.1073/PNAS.1322892111/SUPPL_FILE/PNAS.201322892SI.PDF

Ripka, J. F., Perez-Riba, A., Chaturbedy, P. K., & Itzhaki, L. S. (2021). Testing the length limit of loop grafting in a helical repeat protein. *Current Research in Structural Biology*, *3*, 30–40. https://doi.org/10.1016/J.CRSTBI.2020.12.002

Rodriguez, M., Kim, B., Lee, N. S., Veeranki, S., & Kim, L. (2008). MPL1, a novel phosphatase with leucine-rich repeats, is essential for proper ERK2 phosphorylation and cell motility. *Eukaryotic Cell*, *7*(6), 958–966. https://doi.org/10.1128/EC.00442-07/ASSET/5BDD4B76-4680-458B-B8E4-A0E578D6A443/ASSETS/GRAPHIC/ZEK0060831450007.JPEG

Rodríguez-Bejarano, O. H., Avendaño, C., & Patarroyo, M. A. (2021). Mechanisms Associated with Trypanosoma cruzi Host Target Cell Adhesion, Recognition and Internalization. *Life 2021, Vol. 11, Page 534*, *11*(6), 534. https://doi.org/10.3390/LIFE11060534

Rogala, K. B., Gu, X., Kedir, J. F., Abu-Remaileh, M., Bianchi1, L. F., Bottino1, A. M. S., Dueholm1, R., Niehaus1, A., Overwijn1, D., Priso Fils1, A. C., Zhou1, S. X., Leary, D., Laqtom1, N. N., Brignole, E. J., & Sabatini, D. M. (2019). Structural basis for the docking of mTORC1 on the lysosomal surface. Science, 366(6464), 468–475. https://doi.org/10.1126/SCIENCE.AAY0166/SUPPL_FILE/PAPV2.PDF

Rojas, M., Yu, Q., Williams-Carrier, R., Maliga, P., & Barkan, A. (2019). Engineered PPR proteins as inducible switches to activate the expression of chloroplast transgenes. *Nature Plants*, *5*(5), 505–511. https://doi.org/10.1038/S41477-019-0412-1

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241. https://arxiv.org/abs/1505.04597v1

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature 1986 323:6088*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Rump, A., Morikawa, Y., Tanaka, M., Minami, S., Umesaki, N., Takeuchi, M., & Miyajima, A. (2004). Binding of Ovarian Cancer Antigen CA125/MUC16 to Mesothelin Mediates Cell Adhesion *. *Journal of Biological Chemistry*, *279*(10), 9190–9198. https://doi.org/10.1074/JBC.M312372200

Sanchez-deAlcazar, D., Mejias, S. H., Erazo, K., Sot, B., & Cortajarena, A. L. (2018). Self-assembly of repeat proteins: Concepts and design of new interfaces. *Journal of Structural Biology*, *201*(2), 118–129. https://doi.org/10.1016/j.jsb.2017.09.002

Sathyanarayana, B. K., Hahn, Y., Patankar, M. S., Pastan, I., & Lee, B. (2009). Mesothelin, stereocilin, and otoancorin are predicted to have superhelical structures with ARM-type repeats. *BMC Structural Biology*, *9*(1), 1–11. https://doi.org/10.1186/1472-6807-9-1/FIGURES/4

Schaper, E., & Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytologist*, *206*(1), 397–410. https://doi.org/10.1111/nph.13184

Schaper, E., Korsunsky, A., Pečerska, J., Messina, A., Murri, R., Stockinger, H., Zoller, S., Xenarios, I., & Anisimova, M. (2015). TRAL: tandem repeat annotation library. *Bioinformatics*, *31*(18), 3051–3053. https://doi.org/10.1093/BIOINFORMATICS/BTV306

Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *6354 LNCS*(PART 3), 92–101. https://doi.org/10.1007/978-3-642-15825-4_10

Schmieden, D. T., Basalo Vázquez, S. J., Sangüesa, H., van der Does, M., Idema, T., & Meyer, A. S. (2018). Printing of Patterned, Engineered E. coli Biofilms with a Low-Cost 3D Printer. *ACS Synthetic Biology*, *7*(5), 1328–1337. https://doi.org/10.1021/ACSSYNBIO.7B00424/SUPPL_FILE/SB7B00424_SI_001.PDF

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097. https://doi.org/10.1093/NAR/18.20.6097

Schrödinger LLC. (2015). *The PyMOL Molecular Graphics System, Version~1.8*.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shah, S., & Heddle, J. G. (2014). Squaring up to DNA: pentapeptide repeat proteins and DNA mimicry. *Applied Microbiology and Biotechnology*, *98*(23), 9545–9560. https://doi.org/10.1007/s00253-014-6151-3

Shen, H., Fallas, J. A., Lynch, E., Sheffler, W., Parry, B., Jannetty, N., Decarreau, J., Wagenbach, M., Vicente, J. J., Chen, J., Wang, L., Dowling, Q., Oberdorfer, G., Stewart, L., Wordeman, L., de Yoreo, J., Jacobs-Wagner, C., Kollman, J., & Baker, D. (2018). De novo design of self-assembling helical protein filaments. *Science (New York, N.Y.)*, *362*(6415), 705–709. https://doi.org/10.1126/SCIENCE.AAU3775

Shin, S., Ha, N. C., Oh, B. C., Oh, T. K., & Oh, B. H. (2001). Enzyme mechanism and catalytic property of β propeller phytase. *Structure*, *9*(9), 851–858. https://doi.org/10.1016/S0969-2126(01)00637-2

Shiu, S.-H., & Bleecker, A. B. (2001). Plant receptor-like kinase gene family: diversity, function, and signaling. *Science's STKE : Signal Transduction Knowledge Environment*, *2001*(113). https://doi.org/10.1126/STKE.2001.113.RE22

Shur, O., Dooley, K., Blenner, M., Baltimore, M., & Banta, S. (2013). A designed, phase changing RTX-based peptide for efficient bioseparations. *BioTechniques*, *54*(4), 197–206. https://doi.org/10.2144/000114010

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: increased structural coverage of

functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/NAR/GKAA1079

Sippl, M. J., & Wiederstein, M. (2008). A note on difficult structure alignment problems. *Bioinformatics*, *24*(3), 426–427. https://doi.org/10.1093/BIOINFORMATICS/BTM622

Sleutel, M., Pradhan, B., & Remaut, H. (2022). Structural analysis of the bacterial amyloid curli. *BioRxiv*, 2022.02.28.482343. https://doi.org/10.1101/2022.02.28.482343

Smith, E. W., Zhang, X. J., Behzadi, C., Andrews, L. D., Cohen, F., & Chen, Y. (2015). Structures of Pseudomonas aeruginosa LpxA Reveal the Basis for Its Substrate Selectivity. *Biochemistry*, *54*(38), 5937–5948. https://doi.org/10.1021/ACS.BIOCHEM.5B00720/ASSET/IMAGES/ACS.BIOCHEM.5B00720.SOCIAL.JPEG_V03

Soanes, D. M., & Talbot, N. J. (2010). Comparative Genome Analysis Reveals an Absence of Leucine-Rich Repeat Pattern-Recognition Receptor Proteins in the Kingdom Fungi. *PLOS ONE*, *5*(9), e12725. https://doi.org/10.1371/JOURNAL.PONE.0012725

Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T., Thompson, J., & Baker, D. (2013). High-Resolution Comparative Modeling with RosettaCM. *Structure*, *21*(10), 1735–1742. https://doi.org/10.1016/J.STR.2013.08.005

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stein, A., & Kortemme, T. (2013). Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS ONE*, *8*(5), e63090. https://doi.org/10.1371/journal.pone.0063090

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., & Kim, P. M. (2020). Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Systems*, *11*, 402-411.e4. https://doi.org/10.1016/j.cels.2020.08.016

Strokach, A., & Kim, P. M. (2022). Deep generative modeling for protein design. *Current Opinion in Structural Biology*, *72*, 226–236. https://doi.org/10.1016/J.SBI.2021.11.008

Strokach, A., Lu, T. Y., & Kim, P. M. (2021). ELASPIC2 (EL2): Combining Contextualized Language Models and Graph Neural Networks to Predict Effects of Mutations. *Journal of Molecular Biology*, *433*(11), 166810. https://doi.org/10.1016/J.JMB.2021.166810

Szklarczyk, R., & Heringa, J. (2004). Tracking repeats using significance and transitivity. *Bioinformatics*, *20*(suppl_1), i311–i317. https://doi.org/10.1093/BIOINFORMATICS/BTH911

Tareen, A., & Kinney, J. B. (2020). Logomaker: beautiful sequence logos in Python. *Bioinformatics*, *36*(7), 2272–2274. https://doi.org/10.1093/BIOINFORMATICS/BTZ921

Thompson, J. E., Pourhossein, M., Waterhouse, A., Hudson, T., Goldrick, M., Derrick, J. P., & Roberts, I. S. (2010). The K5 Lyase KflA Combines a Viral Tail Spike Structure with a Bacterial Polysaccharide Lyase Mechanism. *Journal of Biological Chemistry*, *285*(31), 23963–23969. https://doi.org/10.1074/JBC.M110.127571

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature 2021 596:7873*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Uluisik, S., & Seymour, G. B. (2020). Pectate lyases: Their role in plants and importance in fruit ripening. *Food Chemistry*, *309*, 125559. https://doi.org/10.1016/J.FOODCHEM.2019.125559

van Raaij, M. J., Schoehn, G., Burda, M. R., & Miller, S. (2001). Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre. Journal of Molecular Biology, 314(5), 1137–1146. https://doi.org/10.1006/JMBI.2000.5204

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

van Egmond, W. N., Kortholt, A., Plak, K., Bosgraaf, L., Bosgraaf, S., Keizer-Gunnink, I., & van Haastert, P. J. M. (2008). Intramolecular activation mechanism of the Dictyostelium LRRK2 homolog Roco protein GbpC. *The Journal of Biological Chemistry*, *283*(44), 30412–30420. https://doi.org/10.1074/JBC.M804265200

van Rossum, G., & Drake, F. L. (2009). Python3 Reference Manual; CreateSpace. *Scotts Valley, CA*, 242. https://www.python.org/

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://doi.org/10.48550/arxiv.1706.03762

Veggiani, G., Nakamura, T., Brenner, M. D., Gayet, R. v., Yan, J., Robinson, C. v., & Howarth, M. (2016). Programmable polyproteams built using twin peptide superglues. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(5), 1202–1207. https://doi.org/10.1073/pnas.1519214113

Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., Hendrickx, P. M. S., Lagerstedt, I., Mir, S., Montecelo, M. A. F., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Sanz-García, E., Sen, S., … Kleywegt, G. J. (2016). PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Research*, *44*(D1), D385–D395. https://doi.org/10.1093/NAR/GKV1047

Vetting, M. W., Hegde, S. S., Hazleton, K. Z., & Blanchard, J. S. (2007). Structural characterization of the fusion of two pentapeptide repeat proteins, Np275 and Np276, from Nostoc punctiforme: Resurrection of an ancestral protein. *Protein Science*, *16*(4), 755–760. https://doi.org/10.1110/ps.062637707

Vetting, M. W., Hegde, S. S., Wang, M., Jacoby, G. A., Hooper, D. C., & Blanchard, J. S. (2011). Structure of QnrB1, a Plasmid-mediated Fluoroquinolone Resistance Factor *. *Journal of Biological Chemistry*, *286*(28), 25265–25273. https://doi.org/10.1074/JBC.M111.226936

Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S. Y., Zhang, K. Y. J., & Tame, J. R. H. (2014). Computational design of a self-assembling symmetrical β-propeller protein. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(42), 15102–15107. https://doi.org/10.1073/PNAS.1412768111/SUPPL_FILE/PNAS.201412768SI.PDF

Wagner, T., Ermler, U., & Shima, S. (2016). The methanogenic CO2 reducing-and-fixing enzyme is bifunctional and contains 46 [4Fe-4S] clusters. *Science*, *354*(6308), 114–117. https://doi.org/10.1126/SCIENCE.AAF9284/SUPPL_FILE/WAGNER.SM.PDF

Wälchli, M., Berneiser, K., Mangia, F., Imseng, S., Craigie, L. M., Stuttfeld, E., Hall, M. N., & Maier, T. (2021). Regulation of human mTOR complexes by DEPTOR. ELife, 10. https://doi.org/10.7554/ELIFE.70871

Walker, J. M., Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. *The Proteomics Protocols Handbook*, 571–607. https://doi.org/10.1385/1-59259-890-0:571

Walsh, I., Sirocco, F. G., Minervini, G., di Domenico, T., Ferrari, C., & Tosatto, S. C. E. (2012). RAPHAEL: Recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, *28*(24), 3257–3264. https://doi.org/10.1093/bioinformatics/bts550

Wang, C., Zhang, H., Wang, J., Chen, S., Wang, Z., Zhao, L., & Wang, X. (2020). Colanic acid biosynthesis in Escherichia coli is dependent on lipopolysaccharide structure and glucose availability. *Microbiological Research*, *239*. https://doi.org/10.1016/J.MICRES.2020.126527

Wang, F., Sun, X., Shi, X., Zhai, H., Tian, C., Kong, F., Liu, B., & Yuan, X. (2016). A Global Analysis of the Polygalacturonase Gene Family in Soybean (Glycine max). *PLOS ONE*, *11*(9), e0163012. https://doi.org/10.1371/JOURNAL.PONE.0163012

Wang, W., Zhang, Y., Wang, L., Jing, Q., Wang, X., Xi, X., Zhao, X., & Wang, H. (2019). Molecular structure of thermostable and zinc-ion-binding γ-class carbonic anhydrases. *BioMetals*, *32*(2), 317–328. https://doi.org/10.1007/S10534-019-00190-8/FIGURES/7

Wang, X., Wortham, N. C., Lius, R., & Prouds, C. G. (2012). Identification of Residues That Underpin Interactions within the Eukaryotic Initiation Factor (eIF2) 2B Complex. *Journal of Biological Chemistry*, *287*(11), 8263–8274. https://doi.org/10.1074/JBC.M111.331553

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

Watson, R. P., Christen, M. T., Ewald, C., Bumbak, F., Reichen, C., Mihajlovic, M., Schmidt, E., Güntert, P., Caflisch, A., Plückthun, A., & Zerbe, O. (2014). Spontaneous self-assembly of engineered armadillo repeat protein fragments into a folded structure. *Structure*, *22*(7), 985–995. https://doi.org/10.1016/J.STR.2014.05.002/ATTACHMENT/AB9ACA96-E297-4B6E-AC8E-683C9D9474A1/MMC2.MP4

Wells, T. J., Tree, J. J., Ulett, G. C., & Schembri, M. A. (2007). Autotransporter proteins: novel targets at the bacterial cell surface. *FEMS Microbiology Letters*, *274*(2), 163–172. https://doi.org/10.1111/J.1574-6968.2007.00833.X

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, *27*(1), 293–315. https://doi.org/10.1002/PRO.3330

Woolfson, D. N., Bartlett, G. J., Burton, A. J., Heal, J. W., Niitsu, A., Thomson, A. R., & Wood, C. W. (2015). De novo protein design: how do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology*, *33*, 16–26. https://doi.org/10.1016/J.SBI.2015.05.009

Xiong, X., Bromley, E. H. C., Oelschlaeger, P., Woolfson, D. N., & Spencer, J. (2011). Structural insights into quinolone antibiotic resistance mediated by pentapeptide repeat proteins: conserved surface loops direct the activity of a Qnr protein from a Gram-negative bacterium. Nucleic Acids Research, 39(9), 3917–3927. https://doi.org/10.1093/NAR/GKQ1296

Yamagata, A., Goto-Ito, S., Sato, Y., Shiroshima, T., Maeda, A., Watanabe, M., Saitoh, T., Maenaka, K., Terada, T., Yoshida, T., Uemura, T., & Fukai, S. (2018). Structural insights into modulation and selectivity of transsynaptic neurexin–LRRTM interaction. *Nature Communications 2018 9:1*, *9*(1), 1–11. https://doi.org/10.1038/s41467-018-06333-8

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National*

Academy of Sciences of the United States of America, *117*(3), 1496–1503. https://doi.org/10.1073/PNAS.1914677117/SUPPL_FILE/PNAS.1914677117.SAPP.PDF

Yang, Y., Li, D., Chao, X., Singh, S. P., Thomason, P., Yan, Y., Dong, M., Li, L., Insall, R. H., & Cai, H. (2021). *Leep1 interacts with PIP 3 and the Scar/WAVE complex to regulate cell migration and macropinocytosis*. https://doi.org/10.1083/jcb.202010096

Yeh, C. T., Brunette, T. J., Baker, D., McIntosh-Smith, S., & Parmeggiani, F. (2018). Elfin: An algorithm for the computational design of custom three-dimensional structures from modular repeat protein building blocks. *Journal of Structural Biology*, *201*(2), 100–107. https://doi.org/10.1016/J.JSB.2017.09.001

Youn, S. J., Kwon, N. Y., Lee, J. H., Kim, J. H., Choi, J., Lee, H., & Lee, J. O. (2017). Construction of novel repeat proteins with rigid and predictable structures using a shared helix method. *Scientific Reports 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-02803-z

Yu, S., Shen, H., Cheng, Y., Zhu, Y., Li, X., & Mu, W. (2018). Structural and Functional Basis of Difructose Anhydride III Hydrolase, Which Sequentially Converts Inulin Using the Same Catalytic Residue. *ACS Catalysis*, *8*(11), 10683–10697. https://doi.org/10.1021/ACSCATAL.8B02424/SUPPL_FILE/CS8B02424_SI_001.PDF

Yuan, X., Johnson, M. D., Zhang, J., Lo, A. W., Schembri, M. A., Wijeyewickrema, L. C., Pike, R. N., Huysmans, G. H. M., Henderson, I. R., & Leyton, D. L. (2018). Molecular basis for the folding of β-helical autotransporter passenger domains. *Nature Communications 2018 9:1*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-03593-2

Zakeri, B., Fierer, J. O., Celik, E., Chittock, E. C., Schwarz-Linek, U., Moy, V. T., & Howarth, M. (2012). Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(12). https://doi.org/10.1073/PNAS.1115485109/SUPPL_FILE/APPENDIX.PDF

Zambolin, S., Clantin, B., Chami, M., Hoos, S., Haouz, A., Villeret, V., & Delepelaire, P. (2016). Structural basis for haem piracy from host haemopexin by Haemophilus influenzae. *Nature Communications 2016 7:1*, *7*(1), 1–12. https://doi.org/10.1038/ncomms11590

Zhan, X., Yan, C., Zhang, X., Lei, J., & Shi, Y. (2018). Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Research 2018 28:12*, *28*(12), 1129–1140. https://doi.org/10.1038/s41422-018-0094-7

Zhang, R., & Kennedy, M. A. (2021). Current Understanding of the Structure and Function of Pentapeptide Repeat Proteins. *Biomolecules 2021, Vol. 11, Page 638*, *11*(5), 638. https://doi.org/10.3390/BIOM11050638

Zhang, R., Ni, S., & Kennedy, M. A. (2019). Type I beta turns make a new twist in pentapeptide repeat proteins: Crystal structure of Alr5209 from Nostoc sp. PCC 7120 determined at 1.7

angström resolution. *Journal of Structural Biology: X*, *3*. https://doi.org/10.1016/J.YJSBX.2019.100010

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/NAR/GKI524

Zhang, Y.-F., Phung, Y., Gao, W., Kawa, S., Hassan, R., Pastan, I., & Ho, M. (2015). New High Affinity Monoclonal Antibodies Recognize Non-Overlapping Epitopes On Mesothelin For Monitoring And Treating Mesothelioma. *Nature Publishing Group*. https://doi.org/10.1038/srep09928

Zhu, H., Sepulveda, E., Hartmann, M. D., Kogenaru, M., Ursinus, A., Sulz, E., Albrecht, R., Coles, M., Martin, J., & Lupas, A. N. (2016). Origin of a folded repeat protein from an intrinsically disordered ancestor. *ELife*, *5*(September2016). https://doi.org/10.7554/ELIFE.16761