Title

Protein Nanomaterials as Tools for Cryo-EM Structural Analysis

Author

Agdanowski, Matthew Paul

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Protein Nanomaterials as Tools for Cryo-EM Structural Analysis

A dissertation submitted in partial satisfaction of the requirements for Doctor of Philosophy in Biochemistry, Molecular, and Structural Biology

by

Matthew Paul Agdanowski

2024

ABSTRACT OF THE DISSERTATION

Protein Nanomaterials as Tools for Cryo-EM Structural Analysis

by

Matthew Paul Agdanowski

Doctor of Philosophy in Biochemistry, Molecular, and Structural Biology

University of California, Los Angeles, 2024

Professor Jose Alfonso Rodriguez, Chair

In the last few decades there has been tremendous technological and computational advances in the field of cryo-electron microscopy which has led to a phenomena referred to as “The Resolution Revolution,” in which the number of high resolution structures solved via this technique has exploded. Despite these advances, there still remains a size limitation for your target of interest, below which high resolution microscopy remains challenging. Adding to this, a vast majority of the biologically relevant proteins and nucleic acids inside of cells lie below this size limit. Parallel advances in protein design may provide an avenue for progress on this challenging problem. By designing large protein assemblies, we are able to artificially increase the size of a given target, making it amenable to cryo-EM studies. This thesis describes recent advances in the field as well as efforts to generate imaging scaffolds for both small cancer-related protein targets, and RNA molecules. The knowledge gained through these endeavors will better guide future design efforts.

The dissertation of Matthew Paul Agdanowski is approved.

David S. Eisenberg

Feng Guo

Steven G. Clarke

Jose Alfonso Rodriguez, Chair

University of California, Los Angeles

2024

DEDICATION

To my friends and family, thank you for supporting me through this entire process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2 of this dissertation is a version of a published review article: Yeates, TO., Agdanowski, MP., Liu, Y. Development of imaging scaffolds for cryo-electron microscopy. Current Opinion in Structural Biology. 60: 142-149 (2020). The manuscript was prepared by T.O.Y., M.P.A., and Y.L.

Chapter 3 of this dissertation is a version of a published research article: Agdanowski, MP., Castells-Graells, R., Sawaya, MR., Cascio, D., Yeates, TO., Arbing, MA. Acta Crystallographica Section F. 80(50): 107-115 (2024). The project was conceived by M.P.A., R.C.G., and M.A.A. The experiments were performed by M.P.A., M.R.S., D.C., and M.A.A. The data was analyzed by M.P.A., R.C.G., M.R.S., T.O.Y., and M.A.A. All authors were involved in the preparation and editing of the manuscript.

Appendix 1 of this dissertation is a version of a published research article: Miller, JE., Agdanowski, MP., Dolinsky, JL., Sawaya, MR., Cascio, D., Rodriguez, JA., and Yeates, TO. AlphaFold-assisted structure determination of a bacterial protein of unknown function using X-ray and electron crystallography. Acta Crystallographica Section D. 80(4): 270-278 (2024). The project was conceived by J.E.M., J.A.R., and T.O.Y. All experiments were performed by J.E.M., M.P.A., J.L.D., M.R.S., and D.C. The data was processed and analyzed by J.E.M., M.P.A., M.R.S., J.A.R., and T.O.Y. All authors were involved in the preparation and editing of the manuscript.

I would like to acknowledge the funding sources that made the work presented here possible: the US Department of Energy (DE-FC02-02ER63421), the National Institute of Health (RO1GM1298554), the UCLA Cellular and Molecular Biology Predoctoral Training Grant

(GM007185) and all the outside facilities, microscopes and beamlines (CNSI, NE-CAT, APS, UCLA-DOE, NSLS II, CMBS, and NIGMS).

Finally, I would like to acknowledge all the staff researchers, colleagues, and undergraduate mentees who have contributed their time, reagents and support. None of this would have been possible without the community we have built here.

VITA

Matthew Paul Agdanowski

Education

University of California, San Diego - 2015

B.S., Biochemistry and Cell Biology

Provost Honors

University of California, Los Angeles - 2019

M.S., Biochemistry, Molecular, and Structural Biology

University of California, Los Angeles - 2024 (Expected)

Ph.D. Biochemistry, Molecular, and Structural Biology

Awards and Honors

NIH Cellular and Molecular Biology Training Grant 2018-2020

Publications

Agdanowski, MP., Castells-Graells, R., Sawaya, MR., Cascio, D., Yeates, TO, and Arbing, MA. *X-ray crystal structure of a designed rigidified imaging scaffold in the ligand-free conformation.* Acta Crystallographica Section F. (2024). PMID: 38767964

Miller, JE., **Agdanowski, MP**., Dolinsky, J., Sawaya, MR., Cascio, D., Rodriguez, JA., and Yeates, TO. *AlphaFold-Assisted Structure Determinations of Bacterial Protein of Unknown Function Using X-Ray and Electron Crystallography.* Acta Crystallogr D Struct Biol **80:** 270-278 (2024*).* PMID: 38451205

Yeates TO, **Agdanowski MP**, Liu Y. *Development of imaging scaffolds for cryo-electron microscopy*. Curr. Opin. Struct. Biol. **60:**142-149. (2020). PMID: 32066085

Posters, and Presentations

Miller, JE., **Agdanowski, MP**., Dolinsky, J., Cascio, D., Cannon, KA., Yeates, TO. *AlphaFold-Assisted Molecular Replacement of a Novel Protein Domain*. West Coast Structural Biology Workshop, 2023

Jose A. Rodriguez, David Eisenberg, Marcus Gallagher-Jones, Logan Richards, Ambarneil Saha, **Matthew Agdanowski**, Roger Castells-Graells, and Todd O. Yeates. *Development and Deployment of Enabling Technologies at the UCLA-DOE Institute,* DOE Genomic Sciences Program Principal Investigators Meeting, 2021

A Designed Scaffold for Near-Atomic Resolution Cryo-EM Imaging of Small Proteins, DOE Genomic Sciences Program Principal Investigators Meeting, 2019

Teaching and Mentoring Experience

Teaching Assistant, Biochemistry: Introduction of Structure	Winter 2021
Teaching Assistant, Biochemical Lab Techniques II	Spring 2018
Teaching Assistant, Biochemistry: Introduction of Structure	Winter 2018
Graduate Assistant, Talos F200C Electron Microscope	2022 2023
Rotation Student and Undergraduate Mentor	2018-2024

Chapter One: Introduction

1.1: A Brief History of Electron Microscopy

Electron microscopy as a technique has its roots over a hundred years ago when German physicist Ernst Ruka had the idea of using electrons instead of light as the energy source to create a microscope. Together with electrical engineer Max Knoll, they were credited with creating the world's first electron microscope[1]. The idea to use electrons was due to the fact that resolution, the ability to distinguish between two objects close together, is a function of the wavelength of energy used to probe those objects; if scientists wanted to probe the atomic structures inside of cells, then they would need energy with a much shorter wavelength than that of visible light.

Although almost as old as X-ray crystallography, electron microscopy never quite caught hold of the structural biology community and was primarily used to study well ordered and stable structures like metals and alloys[2,3]. It would be many decades before the technique would be successfully and routinely applied to biological macromolecules. The reason for x-ray crystallography's dominance in the structural biology field is primarily due to its ability to achieve a very high resolution of a broad range of biological materials with a high-throughput and scalability potentially requiring a small amount of material to generate many crystals[4,5]. Despite being a powerful and robust technique, x-ray crystallography's main obstacle has always been coaxing your molecule of interest into forming stable, well-ordered crystals capable of diffracting out to high resolutions[6]. To form these crystals, oftentimes various crowding agents or heavy metal additives may be used during the experiment which may introduce unwanted artifacts[7]. One of the powerful aspects of Cryo-EM is its ability to view your sample in its native environment in a vitrified solution rather than a solid-state crystal.

A lot of the advances that have enabled x-ray technology to become the workhorse that it is today are being applied to and being paralleled for electron microscopy, resulting in a phenomenon being heralded as the “Resolution Revolution”[8], a few of which are described below.

When an electron interacts with a specimen, two types of scattering events can occur - elastic scattering, where no appreciable energy is transferred upon interaction, and inelastic scattering, when energy is deposited into the sample after interaction with the electron beam[9]. For the purposes of biological electron microscopy, it is the inelastic scattering we are especially interested in, and its consequences manifest in two forms. Firstly, as the electrons travel down the column, they may interact with the various molecules in the atmosphere, creating many scattering events, and thus introducing a substantial amount of noise into your images. To combat this, the electron microscope column is held under vacuum, eliminating contamination and reducing the effect of unwanted electron interactions[10]. The consequences of inelastic scattering comes from the destruction to the sample itself. When the high energy electron beam interacts with the sample, a substantial amount of energy is deposited. This energy not only results in heat generation which can destroy sensitive samples, it also creates free radicals that can break bonds, creating more radicals than can propagate through the sample destroying its structural integrity and rapidly degrading data quality[11]. These problems were addressed by early investigators by cooling the sample to very low temperatures in attempts to reduce the effects of heating and slow the propagation of radicals, allowing researchers to begin studying biological materials with an electron microscope[12,13]. These experiments were also performed using a low dose beam to reduce the damaging effects, but this resulted in micrographs with a low signal-to-noise ratio (SNR), complicating further attempts to process the data.
In 1974, Taylor and Glaeser published work describing how these cryogenic temperatures, as well as the technique of vitrification, could be successfully employed to collect diffraction

patterns from frozen-hydrated protein crystals of ferritin, rather than having to dry them or embed them in a heavy metal stain and introducing artifacts as was required with previous methods[14]. This also enabled for longer collection times before the sample was destroyed beyond usability. Although Taylor and Glaessar were the first to demonstrate the use of frozen-hydrated samples in biological transmission electron microscopy (TEM), it is Jacques Dubochet that gets most of the credit for championing this process[15]. Vitrification, the process of converting an aqueous solution into a glass-like amorphous solid without the formation of ice crystals, eventually allowed researchers to visualize biological samples at near-atomic resolution. The rapid cooling of a sample to avoid ordered ice formation is essential because the crystalline water lattice can interact with the electron beam and distort the sample[16].

Richard Henderson is credited with some of the first successful applications of these advances in the new field of cryo-EM when he and colleagues were able to solve the structure of bacteriorhodopsin to 3.5 Å resolution from 2D crystals of the protein[17]. This propelled the technique from the depths of “blobology” to a plausible structural biology technique for studying near-atomic interactions inside biological materials. From here the technique exploded in popularity, being applied to larger and more complex samples. In fact cryo-EM became synonymous with virus structures for a large portion of its history, with the highest resolutions being obtained in these samples with high symmetry and many repeated copies of the asymmetric unit[18,19].

The resolution limit continued to be improved by parallel hardware and software advancements. One major leap came from the switch from charge-coupled detector (CCD) cameras to direct electron detectors (DDE). Previously with CCD cameras, the incident electron strikes a scintillator on the top layer of the sensor which releases a photon upon scattering. This resulting photon is captured by a fiber optic cable and brought down to pixels where the charge is

counted and stored[20]. Because of the stochastic nature of the scattering, the resulting photons may produce counts in pixels far away from where the electron struck the detector, or even produce counts in multiple pixels from one electron scattering event. All of this contributes to CCD-recorded micrographs having a substantial amount of noise. The DDE camera solved many of these problems by removing the intermediate step of producing photons with a scintillator, and uses a metal oxide detector (CMOS) to count the individual electrons and convert each count into a charge[21]. This results in a much more localized signal with a much enhanced signal-to-noise ratio (SNR). Additionally, direct electron detectors have a far faster readout speed, allowing for faster data acquisition with a much shorter exposures[22].

This rapid readout rate enabled researchers to make another great resolution-improving advance by tackling the issue of sample motion by “deblurring” the micrograph, revealing finer details lost in the original micrograph. When an electron beam interacts with a vitreous sample, the energy deposited in the form of heat creates a thermal expansion in the ice which propagates into global movements, or “drift”, in the sample[23]. This drift results in a blurring of the resulting micrograph and a loss of high resolution information. Due to the rapid readout rates of the direct electron detectors, computational and algorithmic advancements have further pushed the resolution boundaries for cryo-EM[24]. Among the many contributions, the implementation of a process known as motion correction has enabled researchers to directly combat the effects of beam-induced motion. The most prominent of such algorithms is called MotionCorr developed by David Agard and Yifen Chen of UCSF[25]. This program tracks the motion of the sample in a multi-frame movie and movement vectors can be created between frames in the stack. These frames are then all aligned and summed, revealing features that were priorly lost and enable higher resolution reconstructions to be obtained[26].

In addition to the advances described above, there has been an explosion in software development aimed at addressing every possible problem that might arise during your structural

studies. Processing programs such as RELION[27] and cryoSPARC[28] have emerged as resources to convert your raw data into solved structures using user-friendly interfaces. Tremendous efforts have been invested into solving problems such as how to accurately refine helically-symmetric samples[29,30], or those with high degrees of flexibility or heterogeneity[31,32].

What was described above pertains to electron microscopy in general, but was mainly written in the context of single particle electron microscopy, in which your vitreous sample is imaged in an electron microscope, where hundreds of micrographs containing thousands and thousands of copies of your particle of interest are recorded. The individual particles are extracted from the micrographs and aligned in a way such that every angle of the sample has been captured, allowing for a high resolution density map of the entire sample to be created[33]. However, there are two branches of this technique that should briefly be mentioned. First is electron diffraction, or MicroED, where instead of an aqueous solution being vitrified, your sample is first crystallized akin to traditional crystallography, before being interrogated with an electron microscope[34]. This technique has been able to achieve resolutions on par or higher than its x-ray cousins, but requires crystals much smaller than what could be diffracted with using x-rays. This can be a great benefit in situations where large crystals suitable for x-ray diffraction experiments could not be obtained[35].

The other branch of electron microscopy which has just begun to enter its golden age is called Cryo-Electron Tomography, or CryoET for short. This technique has become a powerful tool to study proteins and cells in their native environments[36]. The sample is embedded in vitreous ice just like a single particle experiment, but in cryoET, the sample may also be an entire cell that was grown on the surface of the grid. Where ET and EM diverge however, is in cryoET, the sample is rotated as the data is collected producing an image stack consisting of projections of your sample as various angles along the tilt series. These images are then aligned and

processed to produce a final 3D volume of your sample[37]. Still in its infancy, cryoET is experiencing the rapid improvements mirroring those seen with cryo-EM with avenues of research into areas such as tackling the missing wedge problem in data aquisition[38,39], or the challenge of sample thickness by creating new ways to thin your sample using an ion beam[40, 41].

1.2: Designing Symmetry

Symmetry is the one of the most important concepts in nature where we frequently find proteins with oligomeric nature, ranging from simple dimers and trimers to the complex architectures of viruses[42, 43]. It's known that roughly half of all proteins form an oligomeric complex to some extent and are almost universally all symmetric in their assembly[44]. The reasons for such oligomerization have been extensively studied throughout the decades and common themes have been identified as to why these assemblies are the way they are. It was found that protein oligomerization can enable cooperative binding, functional regulation, as well as impart structural function and enhanced protein stability[45,46]. Hemoglobin is an ideal example, whereby four hemoglobin monomers evolved to associate as a tetramer and the interaction of the subunits gives rise to allosteric regulation and cooperative binding of the system[47, 48]. It is also not surprising that so many of these homo-oligomeric complexes are seen throughout nature, as a symmetrically-assembled complex requires fewer distinct interfaces or contact points and thus is more likely to have evolved naturally than complexes requiring multiple changes to occur[49]. As mentioned previously, symmetry also plays an important role in cryo-EM data processing. To reiterate, the power of cryo-EM comes from the ability to average thousands- or millions or particles, drastically boosting the signal obtained from the low-dose imaging, and revealing the high-resolution features of your object[50]. Viruses, the poster children of high resolution electron microscopy, are composed of numerous repeating copies of one or a few protein building blocks, and in the case of icosahedral viruses, each particle contains 60 copies of the asymmetric unit - the smallest unit of the structure that can completely recreate the entire

structure by application of only translational and rotational symmetry operations - greatly improving the quality of the data collected. In fact, the iron-storing protein Apoferritin due to its high symmetry and stability, is often used as a benchmark for new microscopes and equipment trying to push the resolution boundaries[51].

Efforts to try to recreate these assemblies in the lab have their roots in the 1970's when Anfensen first proposed that a protein's three dimensional structure was dictated by its primary amino acid sequence, laying the foundation for the field that would become protein design[52]. From here, research advancements paralleled computational advances leading to work starting in the early 1990's, which would eventually become the branch of biology known as computational protein design, where researchers sought to design proteins and assemblies completely *de novo*[53].

Researcher's postulated that when protein molecules interact with one another, it could be in a myriad of ways- either in a stochastic fashion resulting in some sort of amorphous material, or in geometrically-determined ways resulting in complexes like those seen in nature[54]. They took these observations and began to develop rules that could be followed during their design processes that would reliably result in their intended architecture. An example of these geometric rules is shown in Table 1.1[55]. In this table, Yeates and colleagues have outlined the possible types of designed materials that are possible given the symmetry of the underlying building blocks. The symmetries of the assemblies range from the finite, like the platonic solids, to infinitely repeating materials like 1d filaments, 2d arrays or 3d crystals. As an example of how to read such a table, given two trimers of C3 symmetry, the only possible material is a tetrahedral assembly, whereas by changing one trimeric building block into a C4 tetramer opens up the possibility of creating a lot more geometric materials, such as octahedrons, p4 layers or I432 crystals. It was in the Yeates lab using these principles that the first finite protein "cage"

was created by Padilla et al, marking a substantial milestone in the field[56]. The method employed used an alpha-helical extension to rigidly connect two protein domains such that they were oriented in a proper way to enable assembly into a tetrahedral nanocage. Unfortunately, a high resolution structure was unable to be obtained until further improvements were made by Ting Lai et al. who then went on to generate cages with platonic geometries[57-60].

Although the earliest cages were designed as fusions with rigid helical linkers, the field quickly evolved towards using computational tools to completely design new interfaces of interaction between design components. Among the pioneering work stood out the efforts between a former Yeates protege, Neil King and David Baker, who's fruitful collaborations broke important and long-lasting barriers in the field by demonstrating the use of these computational tools at generating entire suites of nanocages with unique assemblies and properties[61,62]. In fact, one particular software suite created in the Baker lab, Rosetta, quickly became the industry standard[63]. The program was designed to be a wide-spread tool for protein structure prediction, docking, and design. The core of the program is the Rosetta scoring function, which is used as a key metric when analyzing the results of prediction or design jobs. In brief, the energy function is a weighted sum of empirically derived values that researchers have identified are important for structure and stability. Some of the terms within the function represent phenomena such as van der Waals interactions, hydrogen bond distances, solvation energy, backbone torsion angles and side chain rotamers[64]. Each new edition of the software brings noticeable improvements. For example, early versions tended to favor interfaces that were highly enriched in hydrophobic residues needed to drive association via the hydrophobic effect[65]. However, this often led to unintended assemblies as the designs lacked the binding specificity to accurately assemble into the intended architectures. Implementation of designed hydrogen bond networks intended to create interfaces resembling those found in nature results in a marked improvement in design results[66,67].

The recent advances in machine learning and artificial intelligence have contributed to great progress in the field of protein design as well. Every day the number of structures in the Protein Data Bank (PDB) increases, giving researchers a larger and larger pool of structures in which to train their machine learning models on. This is enabling researchers to design proteins that more and more resemble the interactions seen in nature. The program that has garnered the most acclaim has been DeepMind's AlphaFold[68,69], which has been smashing records at the annual Critical Assessment of Protein Structure Prediction (CASP) competition, a competition that tests the ability of researchers to accurately predict protein structure using their developed algorithms, and has been heralded as a technology that will change the future of structural biology[70,71].

In addition to the algorithms described above, many other groups are attempting to implement their own strategies to solve these complex challenges facing the design community. Some researchers are choosing to develop new machine learning models to more accurately recapitulate nature, as seen with Protein MPNN, who has demonstrated a marked improvement in the design outputs when compared to Rosetta[72]. Even more interesting still, members of the Yeates group have merged machine learning models for sequence design with a fragment-based approach to determining optimal interaction orientations between subunits by taking known interactions seen among protein pairs in the PDB[73,74].

1.3: Overview

The work laid out in this dissertation builds upon the lessons learned by previous researchers and shines light on the challenges still facing the structural biology field. Chapter 2 begins with a description of limitations of cryo-EM in determining the structure of small biological molecules. This review highlights the efforts of researchers over the years to develop imaging scaffolds to circumvent this size barrier by artificially enlarging the target cargo. Chapter 3 expands on this idea of scaffolding by describing the structure of an Apo crystal structure of an imaging scaffold

designed against important cancer-related proteins. This scaffold utilizes DARPin proteins raised against the cancer protein BARD1 to display multiple copies of the cargo on its exterior. The large solvent channels seen in the crystal may be useful for soaking ligands. In chapters 4 and 5, the development of cryo-EM imaging scaffolds specifically designed for RNA biolmolecules is explored. Chapter 4 describes the efforts to genetically fuse an RNA-binding protein to a tetrahedral protein nanocage via extension of their terminal helices. Chapter 5 describes fragment-based methods to computationally design interfaces between RNA-binding proteins and naturally occurring D3 assemblies.

In addition, an appendix describes the incorporation of an AlphaFold model into the microED and X-ray processing pipelines that led to the structure of a novel and unknown bacterial protein

1.4: References

[1] Ruska, E., & Knoll, M. (1931). Das Elektronenmikroskop. Zeitschrift für Physik, 78(5-6), 318-339.

[2] Hirsch, P. B. (1965). Electron Microscopy of Metals and Alloys (2nd ed.). Butterworth-Heinemann.

[3] Bancroft, J. D., & Gamble, M. (2007). *Theory and Practice of Histological Techniques* (6th ed.). Churchill Livingstone.

[4] Hoppe, W. (2011). *"50 Years of X-ray Diffraction and Protein Crystallography."* Biological Crystallography, 67(Pt 4), 227-241

[5] Branden, C., & Tooze, J. (1999). Introduction to Protein Structure (2nd ed.). Garland Science.

[6] McPherson, A., & Gavira, J. A. (2014). Introduction to protein crystallization. Methods, 34(3), 254-265.

[7] Evans, P. R. (2011). *An introduction to biological crystallography*. Oxford University Press.

[8] Werner Kühlbrandt, *The Resolution Revolution*. *Science* **343:** 1443-1444 (2014)

[9] Saha, A., Nia, S. S. & Rodríguez, J. A. *Electron Diffraction of 3D Molecular Crystals*. Chem. Rev. 122, 13883–13914 (2022).

[10] Bozzola, John J., and Lonnie D. Russell. *Introduction to Electron Microscopy*. 3rd ed., Jones & Bartlett Learning, 2016

[11] Egelman, Edward H. "The current revolution in cryo-EM." *Biophysical Journal* 110.5 (2016): 1008-1012

[12] Bragg, W. H., and W. N. Haworth. "The Structure of Ice." Nature, vol. 111, no. 2794, 1923, pp. 52-53

[13] Cohn, Z. A. "The Movement of Small Molecules Across Cell Membranes." Journal of Experimental Medicine, vol. 81, no. 3, 1945, pp. 233-246

[14] Taylor, K. A. & Glaeser, R. M. Electron diffraction of frozen, hydrated protein crystals. Science 186, 1036–1037 (1974)

[15] Dubochet, J., et al. "Cryo-electron microscopy of vitrified specimens." Journal of Structural Biology, vol. 100, no. 1, 1988, pp. 123-135

[16] Dubochet, J., Lepault, J., Freeman, R., Berriman, J. A. & Homo, J. -C. Electron microscopy of frozen water and aqueous solutions. J. Microsc. 128, 219–237 (1982)

[17] Henderson, R. et al. Model for the structure of bacteriorhodopsin based on high-

resolution electron cryo-microscopy. J. Mol. Biol. 213, 899–929 (1990)

[18] Schoehn, G., Moss, S.R., Nuttall, PA., and Hewat, EA. *Structure of Broadhaven virus by cryoelectron microscopy: correlation of structural and antigenic properties of Broadhaven virus and bluetongue virus outer capsid proteins.* Virology. **235(2):** 191-200 (1997)

[19] Laurinmaki, PA., Huiskonen, JT., Bamford, DH., and Butcher, SJ. *Membrane proteins modulate the bilayer curvature in the bacterial virus Bam35.* Structure. **13(12):** 1819-1828 (2005)

[20] Downing et al. Ultramicroscopy. **75:** 215 (1999)

[21] McMullan, G.; Clark, A. T.; Turchetta, R.; Faruqi, A. R. Enhanced Imaging in Low Dose Electron Microscopy Using Electron Counting. Ultramicroscopy 2009, 109 (12), 1411–1416

[22] McMullan, G., Faruqi, A. R. & Henderson, R. Direct Electron Detectors. Methods Enzymol. 579, 1–17 (2016)

[23] Brilot, A., Chen, J., Cheng, A., Pan, J., Harrison, S., Potter, C., Carragher, B., Henderson, R., and Grigorieff, N. Beam-Induced Motion of Vitrified Specimen on Holey Carbon Film. *J Struct Biol.* **177**(3): 630-637 (2012)

[24] Rawson, S., Iadanza, MG., Ranson, NA., and Muench, SP. Methods to account for movement and flexibility in cryo-EM data processing. *Methods* **100**: 35-41 (2016)

[25] Zheng, S.Q., Palovcak, E., Armache, J-P., Verba, K., Cheng, Y., Agard, D. MotionCor2 - anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods* **14**(4): 331-332 (2017)

[26] Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic resolution single particle cryoEM. Nat. Methods 10, 584 (2013)

[27] Scheres, S. H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology, 180(3), 519-530

[28] Punjani, A., Rubinstein, J. L., Fleet, D. J., & Brubaker, M. A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods, 14(3), 290-296

[29] Fitzpatrick, AWP., Falcon, B., He, S., Murzin, AG., Murshudov, G., Garringer, HJ., Crowther, RA., Ghetti, B., Goedert, M., Scheres, SHW. *Cryo-EM Structures of Tau Filaments from Alzheimer's Disease.* Nature: 547 (7662): 185-190 (2017)

[30] He, S., Scheres, SHW. *Helical Reconstruction in RELION*. Journal of Structural Biology. 198 (3): 163-176 (2017)

[31] Rawson S, Iadanza MG, Ranson NA, Muench SP. *Methods to account for movement and flexibility in cryo-EM data processing*. Methods. 100:35-41.(2016)

[32] Scheres, S. H. W. Processing of Structurally Heterogeneous Cryo-EM Data in RELION. Methods Enzymol. 579, 125–157 (2016)

[33] Cheng Y, Grigorieff N, Penczek PA, Walz T. A primer to single-particle cryo-electron microscopy. Cell. 2015 Apr 23;161(3):438-449

[34] de la Cruz MJ, Hattne J, Shi D, Seidler P, Rodriguez J, Reyes FE, Sawaya MR, Cascio D, Weiss SC, Kim SK, Hinck CS, Hinck AP, Calero G, Eisenberg D, Gonen T. Atomic-resolution structures from fragmented protein crystals with the cryoEM method MicroED. Nat Methods. 2017 Feb 13;14(4):399-402

[35] Bergfors, T. (2003). Protein Crystallization: Techniques, Strategies, and Tips. Oxford University Press

[36] Gan, L. & Jensen, G. J. Electron tomography of cells. Q. Rev. Biophys. 45, 27–56 (2012)

[37] Tocheva, E. I., Li, Z. & Jensen, G. J. Electron Cryotomography. Cold Spring Harb. Perspect. Biol. 2, (2010)

[38] Mastronarde, D. N. (1997). Dual-axis tomography: An approach with alignment methods that preserve resolution. Journal of Structural Biology, 120(3), 343-352

[39] Hagen, W. J., Wan, W., & Briggs, J. A. (2017). Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. Journal of Structural Biology, 197(2), 191-198.

[40] Marko, M., Hsieh, C., Schalek, R., Frank, J. & Mannella, C. Focused-ion-beam thinning of frozen-hydrated biological specimens for cryo-electron microscopy. Nat. Methods 2007 43 4, 215–217 (2007).

[41] Schaffer, M. et al. Cryo-focused Ion Beam Sample Preparation for Imaging Vitreous Cells by Cryo-electron Tomography. BIO-PROTOCOL 5, (2015).

[42] Ahnert SE, Marsh JA, Hernández H, Robinson CV, Teichmann SA: Principles of assembly reveal a periodic table of protein complexes. Science 2015, 350.

[43]Crick, FHC, Watson, JD: Structure of Small Viruses. Nature 1956, 177:473–475.

[44] Marsh JA, Teichmann SA: Structure, Dynamics, Assembly, and Evolution of Protein Complexes. Annual Review of Biochemistry 2015, 84:551–575.

[45] Monod, J: On symmetry and function in biological systems. In Nobel Symp. Symmetry Funct. Biol. Syst. Macromol. Lev., 11th, Stockholm. . Wiley; 1968:15–27.

[46] Goodsell DS, Olson AJ: Structural Symmetry and Protein Function. Annual Review of Biophysics and Biomolecular Structure 2000, 29:105–153.

[47] Perutz, M. F. (1970). Stereochemistry of cooperative effects in haemoglobin. Nature, 228(5273), 726-739.

[48] Baldwin, J. M., Chothia, C., & Lesk, A. M. (1986). Hemoglobin: structure, function, evolution, and pathology. In Hemoglobin (pp. 1-55). Springer, Boston, MA.

[49] Cannon, K., Ochoa, J.M., and Yeates, T.O. High-symmetry protein assemblies: patterns and emerging applications. *Current Opinion in Structural Biology* **55:** 77–84 (2019)

[50] Frank, J. (2016). Advances in Cryo-Electron Microscopy for Structural Biology. Academic Press

[51] Kayama, Y., Burton-Smith, RN., Song, C., Terahara, N., Kato, T., & Murata, K. *Below 3* Å *structure of apoferritin using a multipurpose TEM with a side entry cryoholder.* Scientific Reports. **11(8395)** (2021)

[52] Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), 223–230

[53] Dahiyat, B. I., & Mayo, S. L. (1997). De Novo Protein Design: Fully Automated Sequence Selection. *Science*, 278(5335), 82–87

[54] Yeates, TO. *Geometric Principles For Designing Highly Symmetric Self-Assembling Protein Nanomaterials*. Annual Rev. Biophysics. **47:** 23-42 (2017)

[55] Yeates, TO., Liu, Y., Laniado, J. *The design of symmetric protein nanomaterials comes of age in theory and practice.* Current Opinion in Structural Biology. **39:** 134-143 (2016)

[56] Padilla, JE., Colovos, C., Yeates, TO. *Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments.* Proc. Natl. Acad. Sci. U.S.A. **98(5):** 2217-21 (2001)

[57] Lai, YT., Cascio, D., Yeates, TO. *Structure of a 16-nm cage designed by using protein oligomers*. Science. **336(6085):** 1129 (2012)

[58] Lai. YT., Reading, E., Hura, GL., Tsai, KL., Laganowsky, A., Asturias, FJ., Trainer, JA., Robinson, CV., Yeates, TO. *Structure of a designed protein cage that self-assembles into a highly porous cube*. Nat. Chem. **6(12):** 1065-71 (2014)

[59] Hsia, Y., Bale, JB., Gonen, S., Shi, D., Sheffler, W., Fong, KK., Nattermann, U.,Chunfu, X., Huang, PS., Ravichandran, R., Yi, S., Davis, TN., Gonen, T., King, NP., Baker, D. *Design of a hyperstable 60-subunit protein icosahedron*. Nature. **535(7610):** 136-139 (2016).

[60] Cannon, KA., Nguyen, VN., Morgan, C., Yeates, TO. *Design and Characterization of an Icosahedral Protein Cage Formed by a Double-Fusion Protein Containing Three Distinct Symmetry Elements.* ACS Synth Biol. **9(3):** 517-524 (2020)

[61] King NP., Sheffler W., Sawaya MR., Vollmar BS., Sumida JP., André I., Gonen T., Yeates TO., Baker D. *Computational design of self-assembling protein nanomaterials with atomic level accuracy.* Science. **336(6085):** 1171-4 (2012)

[62] King NP., Bale JB., Sheffler W., McNamara DE., Gonen S., Gonen T., Yeates TO., Baker D. *Accurate design of co-assembling multi-component protein nanomaterials.* Nature. **510(7503):** 103-8 (2014)

[63] Rohl, CA., Strauss, CE., Misura, KM., and Baker, D. *Protein Structure Prediction Using Rosetta*. Methods in Enzymology. **383:** 66-93 (2004)

[64] Leaver-Fay, A., Tyka, M., Lewis, SM., Lange, OF., Thompson, J., Jacak, R., Kaufman, K., Renfrew, PD., Smith, CA., Sheffler, W., Davis, IW., Cooper, S., Treuille, W., Mandell, DJ., Richter, F., Ban, YEA., Fleishman, SJ., Corn, JE., Kim, DE., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, JJ., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, JJ., Kuhlman, B., Baker, D., Bradley, P. *ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules.* Methods in Enzymology. **487:** 545-574 (2011)

[65] Dill, K. A., & MacCallum, J. L. *The Protein-Folding Problem, 50 Years On*. Science, **338(6110):** 1042–1046 (2012)

[66] Boyken, SE., et al. *De novo design of tunable, pH-driven conformational changes*. *Science.* **364:**658-664 (2019)

[67] Cannon, KA., Park, RU., Boyken, SE., Nattermann, U., Yi, S., Baker, D., King, NP., Yeates, TO. *Design and Structure of two new protein cages illustrate successes and ongoing challenges in protein engineering.* Protein Sci. **29(4):** 919-929 (2019)

[68] Senior, AW., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, AWR., Bridgland, A., Penedones, H., Peterson, S., Simonyan, K., Crossan, S., Kohli, P., Jones, DT., Silver, D., Kavukcuoglu, K., Hassabis, D. *Protein Structure Prediction Using Multiple Deep Neural Networks in the 13 Critical Assessment of Protein Structure Prediction (CASP13).* Proteins: Structure, Function, and Bioinformatics. **87(12):** 1141-1148 (2019)

[69] 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, SAA., Ballard, AJ., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholaska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, AW., Kavukcuoglu, K., Kohli, P., Hassabis, D. *Highly accurate protein structure prediction with AlphaFold.* Nature. **596:** 583-589 (2021)

[70] Marcu, SB., Tăbîrcă, S., Tangeny, M. *An Overview of of Alphafold's Breakthrough.* Front. Artif. Intell. **5** (2022)

[71] Service, RF. *'The game has changed.' AI triumphs at protein folding.* Science **370(6521):** 1144 (2020)

[72] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, RJ., Milles, LF., Wicky, BIM., Courbet, A., de Haas, RJ., Bethel, N., Leung, PJY., Huddy, TF., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, AK., King, NP., Bake,r D.. *Robust deep learning-based protein sequence design using ProteinMPNN.* Science. **378(6615):** 49-56 (2022)

[73] Laniado, J., Meador, K., Yeates, TO. *A fragment-based protein interface design algorithm for symmetric assemblies.* Protein Eng Des Sel. **34** (2021)

[74] Meador, K., Castells-Graells, R., Aguirre, R., Sawaya, MR., Arbing, MA., Sherman, T., Senaranthne, C., Yeates, TO. *A suite of designed protein cages using machine learning and protein fragment-based protocols.* Structure (2024)

x	C2	C3	C4	C6	D2	D3	D4	D6	T	O
C2	b	D3, T, O, I p6, p321 $I2_13$, $P4_132$	D4, O p4, $p42_12$ I432	D6 p6, p622	c222, p422, p622 $I4_122$, $P6_222$, I432, $I4_132$	p312, p622 R32, $P6_322$, $F4_132$, $I4_132$, I432, $P4_132$	p422 I422, P432, I432	p622 P622	P23, F23, $F4_132$	P432, F432, I432
C3		T p3 $P2_13$	O F432	p6	p622 P23, F432, $I4_132$	p321, p312 $P4_132$	P432	p622	F23	F432
C4			p4 P432		p422, $p42_12$ I432, F432	I432	p422 P432		F432	P432
C6					p622	p622				
D2					p222, p622 F222, $P4_222$, $P6_222$, $P4_232$, $I4_132$	p622 P622, $P4_232$, $I4_132$	p422 P422, I422, I432	p622 P622	P23, F432, $P4_232$	F432, I432
D3						p321 P312, $P6_322$, $P4_232$, $P4_132$	I432	p622 P622	$F4_132$	I432
D4							p422 P422, P432			P432
D6										
T									F23	F432
O										P432, F432

Table 1.1: Multiplication table for designing self-assembling protein nanomaterials using combinations of simpler, symmetric components. Finite assemblies (point group symmetries) are indicated in the blue font. 2-D layers are indicated in red, 3-D crystalline arrays in purple. Gray boxes indicate symmetry combinations that are disallowed mathematically. Table adapted from Yeates et al[54].

CHAPTER TWO: Development of Imaging Scaffolds for Cryo-Electron Microscopy

The following is a reprint of a review article from:

Current Opinion in Structural Biology

60: 142-149 (2020)

DOI: 10.1016/j.sbi.2020.01.012

Development of Imaging Scaffolds for Cryo-Electron Microscopy

Todd O. Yeates, Matthew Agdanowski, Yuxi Liu

Abstract

Following recent hardware and software developments, single particle cryo-electron microscopy (cryo-EM) has become one of the most popular structural biology tools. Many targets, such as viruses, large protein complexes and oligomeric membrane proteins, have been resolved to atomic resolution using single-particle cryo-EM, which relies on the accurate assignment of particle location and orientation from intrinsically noisy projection images. The same image processing procedures are more challenging for smaller proteins due to their lower signal-to-noise ratios. Consequently, though most cellular proteins are less than 50 kDa, so far it has been possible to solve cryo-EM structures near that size range for only a few favorable cases. Here we highlight some of the challenges and recent efforts to break through this lower size limit by engineering large scaffolds to rigidly display multiple small proteins for imaging. Future design efforts are noted.

Introduction

The broad field of structural biology has transformed our understanding of biology at the atomic level. Methods including X-ray crystallography, multi-dimensional NMR, and electron microscopy (EM) have all played key roles, with different methods offering their own advantages and challenges. Regardless of method, obstacles and uncertainties still challenge structural

biology efforts, with a common theme that the difficulties often relate to the suitability of the macromolecular sample [1]. Is the macromolecule under investigation large enough, small enough, sufficiently pure, sufficiently abundant, and so on. As a consequence, the heaviest exertions typically go towards modifying and optimizing the macromolecule under study to make it more amenable to the structural biology technique at hand.

Towards improving the properties of a protein or nucleic acid molecule for structure determination, all manners of modification have been explored. Optimizing the stability and homogeneity of the structural target are common goals. Certain strategies for optimization are method-specific. For instance, because forming lattice contacts is such a critical limiting event for x-ray crystallography applications, diverse ideas for modifying proteins or nucleic acids to improve their chances of forming well-ordered intermolecular contacts have been exploited [2-7]. Another category of modifications relates to principles of ‘marking’ the molecule for comparative purposes in downstream analysis. In x-ray crystallography applications, the introduction of seleno-methionine residues for anomalous phasing is an example [8]. For EM applications, in cases where resolution is insufficient to unambiguously trace the molecular backbone, antibodies to specific regions of the target macromolecule or macromolecular complex have been used to aid in structural interpretation [9-11].

Yet other efforts to engineer macromolecules for structural biology purposes do not concern any innate defect in the target molecule itself, but relate instead to intrinsic methodological limits. The parameters of the molecule in question may fall outside the application range of the method being employed. Such is the case for single-particle cryo-EM, which is most amenable to large macromolecular complexes, but faces major challenges for small proteins or nucleic acid molecules (Fig. 1) [12]. In this review we discuss recent advances on the particular challenge of how to make macromolecules amenable to cryo-EM when they are otherwise below the suitable

size range, essentially by engineering the small macromolecule of interest to become part of a larger imaging scaffold.

Motivation for Cryo-EM Scaffolds and their Challenges

Recent technical advances in cryo-EM have revolutionized the field, making it possible to reach atomic resolution for diverse macromolecular systems [13-16]. Structure determination has been possible for complexes of extraordinary size and complexity, but smaller proteins or nucleic acids remain largely outside the scope of cryo-EM. The key challenges concern low signal-to-noise in single particle imaging, making it difficult to identify and establish the correct 3-dimensional orientations of noisy 2-D particle projections, which is a prerequisite to reconstructing a 3-dimensional view of the molecule. The signal to noise problem is most severe for smaller macromolecules [17]. The importance of the problem is illustrated by the size distribution of macromolecular complexes that have been successfully determined at atomic resolution. Only a few (about 2%) fall below the 100 kDa size (Fig. 1A). This contrasts with the size of typical cellular proteins; the median bacterial protein is roughly 30 kDa while eukaryotic proteins are slightly larger, with a median size of around 45 kDa (Fig. 1B). Recent studies to push the lower size limit, by improvements in instrumentation, or sample preparation and data processing, have succeeded in elucidating the structures of a few protein assemblies in the 40 to 70 kDa range [18-21], but the general difficulty of resolving the structures of proteins of typical cellular size by routine cryo-EM applications remains. Breaking through this size barrier could have high impact as it would be an important step towards making cryo-EM a near-universal approach for atomic level structure determination of cellular proteins and possibly nucleic acids.

One strategy for circumventing the size limitation in cryo-EM is to attach a smaller imaging target to a larger macromolecular structure known to be amenable to imaging; the former can be described as the ‘cargo’ and the latter as the ‘scaffold’. Besides being a direct attack on the

problem of size, introducing a separate binding component opens up additional possibilities. For example, some proteins become better ordered upon (or are only ordered upon) binding to another protein [22-25]. Laboratory evolution studies have further shown that some proteins that have multiple relevant conformational states can sometimes be bound or trapped in distinct conformations by binding to different partners [26-30]. Therefore, for cryo-EM applications, the binding aspect of scaffolding approaches offers broad and potentially important prospects for exploring alternative structural forms of proteins with dynamic behavior.

While the idea of attaching a smaller protein to a larger scaffold seems straightforward, the obviousness of the idea belies serious technical challenges and failure risks in the context of imaging. Flexibility is the primary concern. If the cargo is rigidly connected to the scaffold, then all the favorable imaging advantages of the scaffold are acquired by the cargo; in a 3-dimensional reconstruction, the latter simply appears as an added bonus with the former. But if the attachment between the cargo and the scaffold is completely flexible, then the presence of the scaffold provides little help in narrowing down the position and orientation of the cargo, as required for its image reconstruction [31]. Even moderate degrees of flexibility can have major confounding effects, as emphasized below [32]. Absent careful design considerations, genetically fusing one protein to another generally results in highly flexible arrangements, owing to effectively free rotation about the phi and psi backbone torsion angles at the point of fusion [31, 33-35]. Similarly, covalent attachment between proteins by way of chemical linkers generally introduces single-bonds between the components, and this also allows potentially problematic degrees of rotation [36, 37]. A further concern for scaffolding approaches is the molecular engineering effort required. If it is necessary, for each and every cargo protein, to pursue laborious mutation and evaluation experiments anew in order to obtain a working scaffold, then the practical utility of the approach is limited. To provide the most utility, an ideal scaffold would provide a facile route for rigidly attaching diverse cargo molecules for imaging,

without extensive re-engineering.

Beyond the most essential features noted above, useful scaffolds might confer additional advantages. As an example, symmetry is often a favorable feature for imaging studies. For cryo-EM reconstruction, high symmetry can produce favorable results from fewer particle images, as each particle image effectively provides views from multiple different directions of projection [38, 39]. It is notable that icosahedrally symmetric viruses have been particularly rich subjects for cryo-EM elucidation [40-42]. A further and possibly more critical advantage is that highly symmetric molecular assemblies largely mitigate the relatively common and sometimes insurmountable problem of preferred particle orientation [43]. If particles tend to lie on an EM grid in certain orientations, then 3-dimensional reconstruction has lower resolution along the direction of preferred projection. This particle orientation problem can persist for symmetries up to dihedral, but they are mitigated by cubic (tetrahedral or octahedral) and icosahedral symmetries.

Recent Successes for Cryo-EM Scaffolds

A few early studies on attaching small proteins to larger assemblies provided impetus for developing general scaffolds. An early attempt to use a symmetric complex as an EM scaffold came in 1999. Kratz et al. attempted to resolve the structure of green fluorescent protein (GFP, ~26 KDa) by inserting it into a flexible loop in the hepatitis B virus capsid protein [31]. While the researchers were unable to solve the structure of the fused GFP, a shell of density was clearly visible on the exterior of the viral capsid. The blurring of density for the GFP was attributed to the flexible, poly-glycine linker, suggesting that higher resolutions might be achieved by using a more rigid connection. More than a decade later, in a more systematic study by Coscia et al., maltose binding protein (MBP) was genetically fused to a dodecameric glutamine synthetase (D6 symmetry) by joining the α-helical termini of the two proteins. This was an application of the

idea developed by Padilla et al. [44], and expanded upon by others [45–49], for controlling the relative orientation of two proteins with compatible termini by a continuous helical fusion between them. The authors tested different lengths for the continuous α-helix, and found an optimal helical linker length that enabled the 40 kDa MBP to be reconstructed at 6–10 Å local resolution, albeit at a somewhat different orientation than modeled due to steric hindrance (Fig. 2A) [50]. This study demonstrated the potential prospects for using a continuous α-helix to connect cargos to a scaffold for cryo-EM imaging purposes. However, the scaffold scheme explored in that work is limited to cargo proteins with α-helical termini, and would require laborious testing of the linker length for each new cargo. As noted above, for an imaging scaffold to be broadly useful, more modular and generally applicable schemes are required.

Parallel efforts have been undertaken to develop nucleic acids and nucleic acid complexes as scaffolds for either proteins or RNA molecules. Using DNA, Martin et al. created a scaffold in which DNA double-helices are patterned out along a hexagonal grid [51]. A central cavity in the designed pattern is then spanned by a double-helix containing the DNA binding sequence recognized by the tumor-suppressor protein p53, which is thereby anchored in the middle of the cavity. This DNA-patterning approach addressed two important issues in cryoEM structural analysis. The height of the DNA support promotes formation of a uniform ice thickness across the grid while also protecting the cargo protein from denaturation at the air-water interface. By changing the register of the p53 binding sequence, the authors were able to control the orientation of p53, allowing them to view it in multiple orientations. Using this DNA support system, a final reconstruction obtained for the 53 kDa p53 protein reached 15 Å resolution. In another 2019 study focusing on cryo-EM refinement methods, Zhang et al. fused the small HIV-1 Transactivation Response (TAR) element RNA into a double-helix of the bacterial large 23S and 70S ribosomes, which essentially served as the scaffolds. Using a focused classification scheme, the authors were able to resolve the ribosome to an overall resolution 3.0

Å. It was only possible to refine the important TAR section of the fusion to an intermediate resolution, but still high enough to identify most of the A form RNA helix [52]. The authors describe their fusion construct as a lever pivoting around a fulcrum, with greater displacements further from the point of fusion leading to poorer resolution.

In 2018, Liu et al. designed a new protein scaffold to simultaneously address the key issues of flexibility and modularity that had limited previous studies. In this work, a modular adaptor protein, Designed Ankyrin Repeat Proteins (DARPin, see also review by Mittl, et al., in this issue), was fused to an engineered tetrahedrally symmetric (n=12) protein complex, using a continuous α-helical connection to promote rigidity [44]. Through selection experiments on libraries of DAPRins bearing sequence variation in their loops, DARPins can be obtained to bind diverse cargo proteins with high affinity and specificity, and with retention of their structural integrity in the bound state [28, 53, 54]. Thus, it was postulated that this DARPin-tetrahedron scaffold could be modular – i.e. that it could be made to bind different cargo proteins by inserting the appropriate cargo-binding sequences into the loops of the DARPin adaptor (Figure 2B). In a first cryo-EM study on this scaffold prior to binding any cargo protein, it was found that the critical helical connection between the tetrahedral core and the DARPin adaptor is rigid enough that the 17 kDa DARPin could be visualized at 3.5 – 5 Å local resolution [55]. The resulting structure furthermore revealed an additional small, fortuitous interface between the DARPin and the tetrahedral core of the scaffold. A presumptive stabilizing effect of this interaction could be partly responsible for the favorable resolution obtained with this scaffold. It is notable that by fusing to a tetrahedral complex, each particle contained 12 copies of the DARPin adaptor. As with the earlier Coscia study, the polyvalent nature of the scaffold provides multiple sites of attachment, and therefore multiple independent views of the cargo from a single particle. High symmetry, to the extent that it is preserved in the assembled complex, provides further imaging advantages through symmetry-averaging protocols. It should be noted however that polyvalent

scaffolds are not well-suited for homo-oligomeric cargo attachment, as the binding of such cargo to polyvalent scaffolds tends to produce extended network-type assembly and aggregation.

Two recent studies have tested DARPin-based helical-fusion scaffolds for their ability to resolve the structures of bound cargo proteins. Following the original work described above, Liu et al. (2019) used GFP as the first cargo for their DARPin-tetrahedron scaffold. Some smearing effect of the density for the GFP was evident, caused by minor flexibility around the shared alpha helix. Exploiting the benefit of having multiple copies of GFP per scaffold, the smearing effect could be partly mitigated by separating different orientations of the GFP using focused classification protocols. The 3-D reconstruction reached a local median resolution of 3.8 Å for the GFP, where notable atomic details were evident (Fig. 2B) [56]. Qing et al. [57] independently explored a series of different symmetric scaffolds as cores for fusion to a DARPin adaptor, and found the best results using D2 tetrameric aldolase. The DARPin was fused again through a continuous α-helix. The authors showed that the continuous α-helix scheme led to more rigid and better behaving scaffolds than their experiments where adaptor proteins were inserted into flexible loop regions of scaffold core proteins. Qing et al. also used GFP as their first test cargo, and reached 5–8 Å local resolution (Fig. 2C) [57]. Coincidently, the two groups of researchers both chose GFP as the first test cargo, so the true modularity of such systems awaits experimental verification from the cryo-EM community.

Next Steps

The scaffolds described above show considerable promise as novel strategies for cryo-EM imaging, but critical steps remain to be addressed, especially for reaching atomic resolution. First, further engineering efforts are needed to realize better rigidity of the adaptor proteins relative to the scaffold core. The experimental studies show, not surprisingly, that even moderate rotational freedom can limit the resolution of the cargo to ~4 Å, or considerably worse.

As opposed to the fortunate and small secondary contact point in the Liu 2018 study or the generic interaction brought by steric hindrance between MBP and glutamate synthase in the Coscia study, engineering multiple stable contact points between the adaptor and the scaffold may be required in order to reach atomic resolution. Additionally, researchers should keep in mind that adaptor proteins might in some cases bind to their cargo in a range of different conformations or orientations, introducing another layer of flexibility.

Currently, symmetric protein scaffolds that have shown potential prospects for cryo-EM imaging have been tested using well-behaved proteins -- GFP or MBP -- as cargos (Fig. 2) [50, 56, 57]. For broader applications to important cellular proteins, the cargo molecules are likely to be more delicate, and some may require very specific solution environments for stability. Accordingly, one important future engineering goal is to design working scaffolds that are stable under wide-ranging conditions related to: protein concentration, pH, ionic strength, metal additives, and so on. Of course, convenient scaffolds should also be easy to purify to high quantity and purity, preferably from the simplest protein expression systems.

This promising new technology is not limited to soluble proteins. Designed scaffolding approaches could in principle be extended to membrane proteins, which are of high biological interests and have been historically difficult to characterize via traditional structural techniques. Lipid nanodisc technology has enabled the cryo-EM structure determination of several membrane proteins, especially those of larger size and/or oligomeric composition [58. 59] (also see review by Nasr et al., in this issue). For smaller monomeric membrane proteins, symmetric scaffolds of the type described here could be used if space permits binding of the membrane protein in its detergent micelle environment. Uniquely challenging targets such as membrane proteins may call for unique scaffolds. RNA macromolecules are another category of great biological interest but high recalcitrance to structure determination. The limited chemical

diversity, significant flexibility, and highly charged backbone of RNA combine to make it challenging to study by standard structural biology techniques [60–63]. With suitable strategic variations, scaffolding methods being developed for proteins could be applied to RNA as well. Small RNA-binding domains such as U1A have been exploited as modular RNA binders in previous x-ray crystallography work (64, 65), which suggests that RNA-binding scaffolds based on such domains could be developed for cryo-EM in the future.

While recent developments on designing cryo-EM imaging scaffolds are largely in the proof-of-principle stages, the work described in this review should provide encouragement – along with guiding principles and challenges to be addressed – on the way to establishing their routine use for structural biology studies.

Acknowledgements

This work was supported by NIH grant R01 GM129854 (to TOY).

The authors declare no conflicts of interest.

Scaffold	Cargo	Display Method	Resolution	Authors
Hepatitis B Viral Capsid Protein	Green Fluorescent Protein	Poly-G Fusion	NA	Kratz et al. 1999
2-D Hexagonal DNA Array	p53 Transcription Factor	DNA-protein interaction	~15 Å	Martin et al. 2016
23S and 70S Ribosomes	HIV-I TAR RNA	A-Form Helical Fusion	6-10 Å	Zhang et al. 2019
D6 Glutamine Synthetase	Maltose Binding Protein	α-Helical Fusion to Cargo	6-10 Å	Coscia et al. 2016
D2 Aldolase	Green Fluorescent Protein	α-Helical Fusion to DARPin adaptor	5-8 Å	Qing et al. 2019
Tetrahedral Nanocage	Green Fluorescent Protein	α-Helical Fusion to DARPin adaptor	3.8 Å	Liu et al. 2018, 2019

Table 2.1 Summary of selected cryo-EM scaffolding efforts

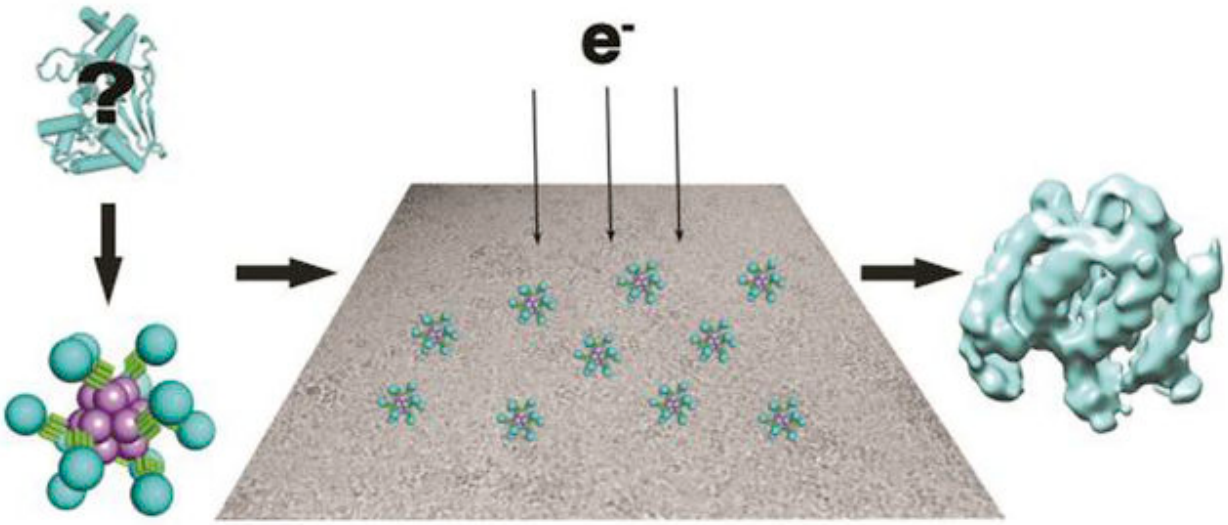


Figure 2.1: Graphical Abstract

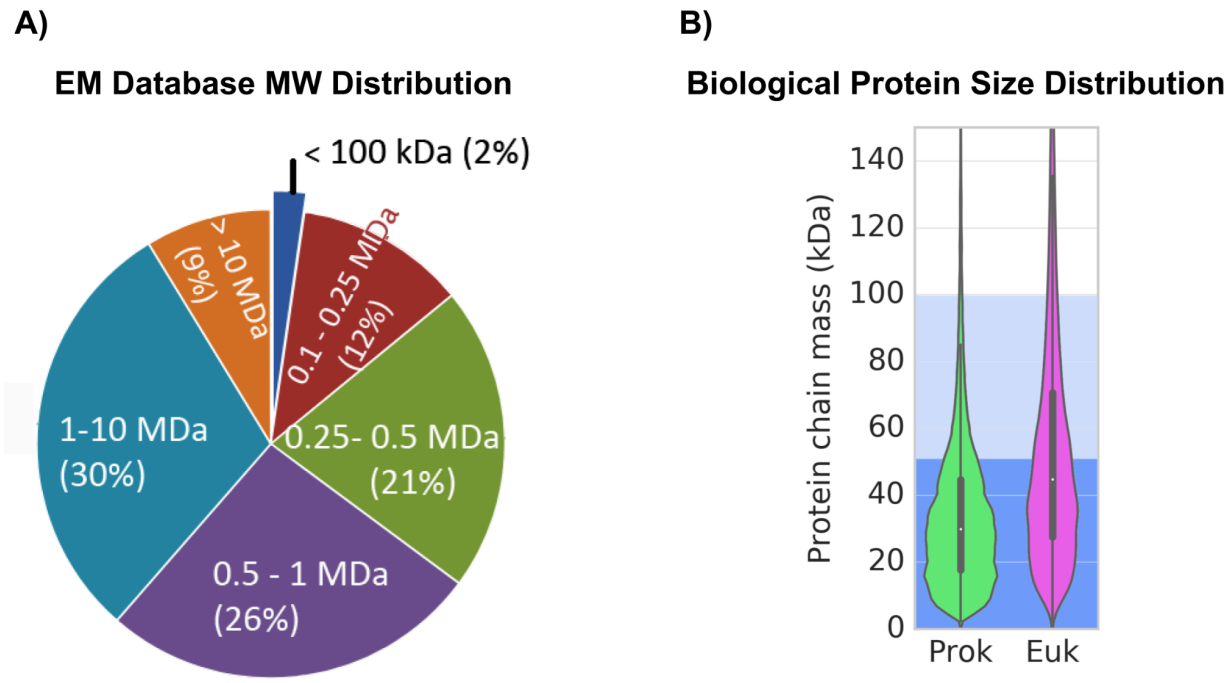


Figure 2.2: Size comparison between natural cellular proteins and protein structures elucidated by cryo- EM. Panel A shows a statistical breakdown by size of all solved structures deposited in the EM Database (EMDB) as of 2018. Panel B shows violin plots of the natural protein chain size distributions for prokaryotes ([green] bacteria and archaea, based on 1501 complete genomes) and eukaryotes ([magenta] based on the genomes of 7 model organisms: yeast, Chlamydomonas, Arabidopsis, Drosophila, zebrafish, mouse, and human). For eukaryotic genomes, in cases where multiple protein isoforms of a single gene are known, only one was included. The median molecular mass values are 29.9 kDa for prokaryotes and 45.6 kDa for eukaryotes. Blue shading is used to highlight molecular weights below which atomic resolution is difficult to achieve using standard (non-scaffolded) cryo-EM approaches (light blue), or where it has not been possible so far (medium blue). Note that the data presented refers to individual protein chain sizes and so does not account for the important effect of increasing size that comes with oligomeric assemblies, which are common.

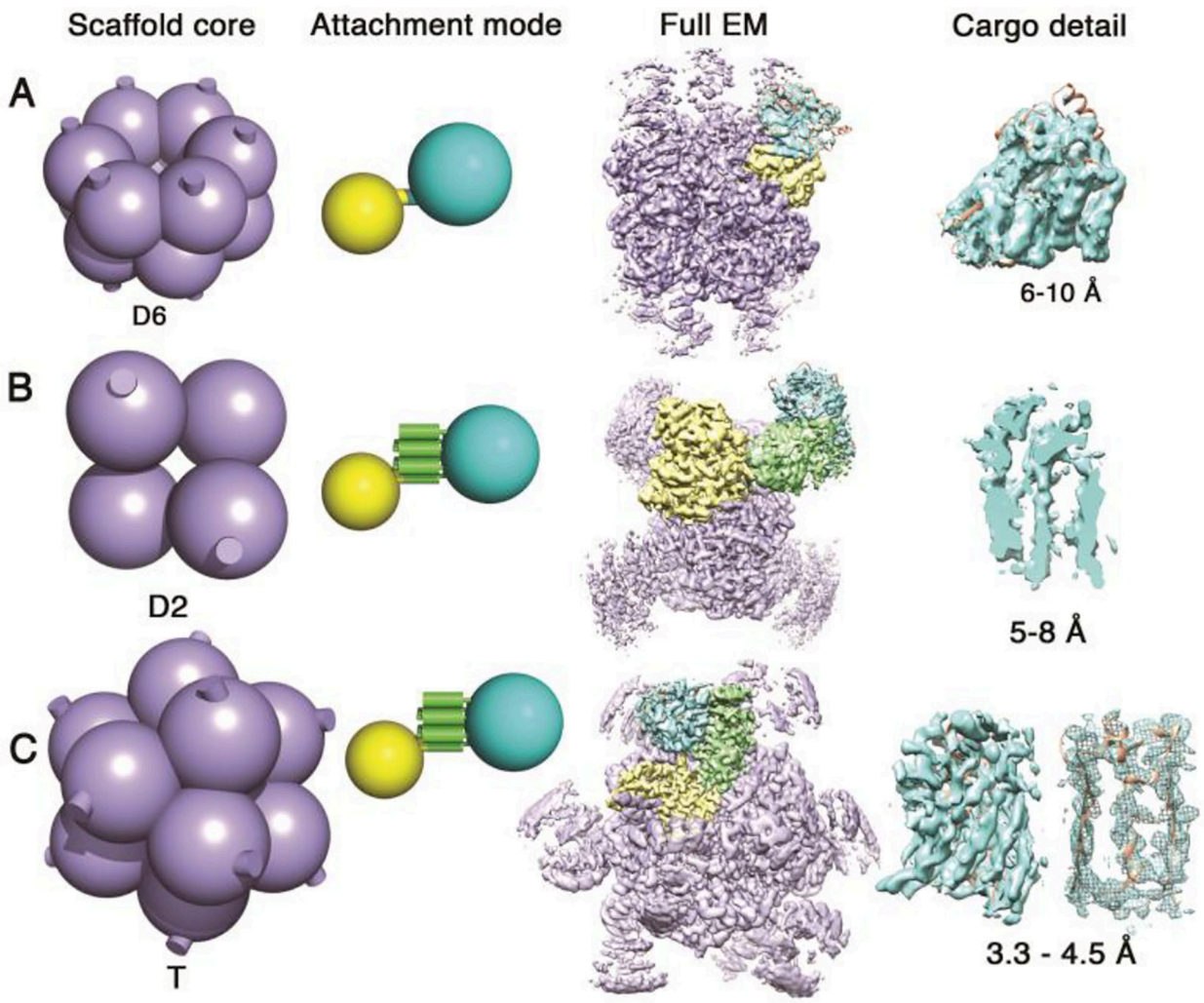


Figure 2.3: Recent progress in imaging small proteins on symmetric scaffolds. Experiments are ordered from top to bottom by resolution achieved for a bound/attached cargo protein, based on data from (A) Coscia et al, 2016; (B) Qing et al, 2019; (C) Liu et al, 2019. The geometry of the scaffolding core is shown in lavender on the left, with the symmetry type indicated; the cylindrical stubs denote the protruding terminal alpha helices present on the scaffold core subunits in all the cases described here. The mode of attachment for each scaffold is diagrammed to the side, with one protein subunit from the scaffold core shown highlighted in yellow. The middle (B) and bottom (C) scaffolds are modular, with the adaptor protein (DARPin) in green fused to the core subunit through a continuous alpha

helix. The cyan sphere indicates the cargo protein for imaging. Density for a complete EM reconstruction is shown, colored according to the same scheme as on the left, with a focus on the density around the cargo proteins (cyan) shown on the right. The underlying atomic model is shown in orange in select cases. For scaffold B, and the right-most image for scaffold C, the cargo EM density is shown as a slice through the beta barrel of GFP to emphasize structural details evident there.

REFERENCES

[1] -. Slabinski L, Jaroszewski L, Rodrigues APC, Rychlewski L, Wilson IA, Lesley SA, and Godzik A The challenge of protein structure prediction - lessons from structural genomics. Protein Sci. 16(11): 2472–2482 (2007). [PubMed: 17962404]

[2] -. Leibly DJ, Arbing MA, Pashkov I, DeVore N, Waldo GS, Terwilliger TC, and Yeates TO A Suite of Engineered GFP Molecules for Oligomeric Scaffolding. Structure. 23(9): 1754–1768 (2015). [PubMed: 26278175]

[3] -. Derewenda ZS Application of protein engineering to enhance crystallizability and improve crystal properties. Acta Crystallogr. D Biol. Crystallogr 66(5): 604–15 (2010). [PubMed: 20445236]

[4] -. Derewenda ZS Rational protein crystallization by mutational surface engineering. Structure. 12(4): 529–35 (2004). [PubMed: 15062076]

[5] -. Moon AF, Mueller GA, Zhong X, and Pedersen LC A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. Protein Sci. 19(5): 901–13 (2010). [PubMed: 20196072]

[6] -. Ferré-D'Amaré AR, Zhou K, and Doudna JA A general module for RNA crystallization. J. Mol Biol 279, 621–631 (1998). [PubMed: 9641982]

[7] -. Yamada H, Tamada T, Kosaka M, Fujiki S, Tano M, Yamanishi M, Honjo E, Tada H, Ino T, Yamaguchi H, Futami J, Seno M, Nomoto T, Hirata T, Yoshimura M, and Kuroki R 'Crystal Lattice Engineering,' an approach to engineer protein crystal contacts by creating intermolecular symmetry: crystallization and structure determination of a mutant human RNase 1 with a hydrophobic interface of leucines. Protein Science, 16(7): 1389–97 (2007). [PubMed: 17586772]

[8] -. Hendrickson WA, Horton JR, and LeMaster DM Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J. 9(5): 1665–72 (1990). [PubMed: 2184035]

[9] -. Wu S, Avila-Sakar A, Kim JM, Booth DS, Greenberg CH, Rossi A, Liao, Li, Alian, Griner SL, Narinobu J, Yu Y, Mergel CM, Chaparro-Riggers J, Strop P, Tampé R, Edwards RH, Stroud RM, Craik CS, and Cheng Y Fabs enable single particle cryoEM studies of small proteins. Structure 20(4): 582–592 (2012). [PubMed: 22483106]

[10] -. Kim JM, Wu S, Tomasiak TM, Mergel C, Winter MB, Stiller SB, Robles-Colmanares Y, Stroud RM, Tampe R, Craik CS, and Cheng Y Subnanometre-resolution electron cryomicroscopy structure of a heterodimeric ABC exporter. Nature 517, 396–400 (2015). [PubMed: 25363761]

[11] -. Jiang J, Miracco EJ, Hong K, Eckert B, Chan H, Cash DD, Min B, Zhou HZ, Collins K, and Feigon J The architecture of Tetrahymena telomerase holoenzyme. Nature. 496(7444): 187–92 (2013). [PubMed: 23552895]

[12] -. Glaeser RM How good can cryo-EM become? What Remains Before It Approaches Its Physical Limits Annu. Rev. Biophys 48: 45–61 (2019). [PubMed: 30786229]

[13] -. Zhang X, Jin L, Fang Q, Hui WH, and Zhou ZH 3.3 Å cryo-EM structure of a non-enveloped virus reveals a priming mechanism for cell entry. Cell. 141(3): 472–82 (2010). [PubMed: 20398923]

[14] -. Alushin GM, Lander GC, Kellogg EH, Zhang R, Baker D, and Nogales E High-resolution microtubule structures reveal the structural transitions in αβ-tubulin upon GTP hydrolysis. Cell. 157(5): 1117–29 (2014). [PubMed: 24855948]

[15] -. Allegretti M, Mills DJ, McMullan G, Kühlbrandt W, and Vonck J Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. Elife. 3: 01963 (2014).

[16] -. Ognjenović J, Grisshammer R, and Subramaniam S Frontiers in Cryo Electron Microscopy of Complex Macromolecular Assemblies. Annu. Rev. Biomed. Eng 21: 395–415 (2019) [PubMed: 30892930]

[17] -. Henderson R The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Quart. Reviews of Biophysics 28: 171–193 (1995).

[18] -. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MJ, Pragani R, Boxer MB, Earl LA, Milne JLS, and Subramaniam S Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. Cell. 167(7): 1698–1707 (2016).

[19] -. Khosouei M, Radjainia M, Baumeister W, and Danev R Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. Nature Comm. 8:16099 (2017). [PubMed: 28665412]• The authors push the size and cryo-EM resolution limit in this study, not by molecular engineering, but by using a phase plate to enhance the signal-to-noise of haemoglobin in near-focus images.

[20] -. Herzik MA Jr., Wum M, and Lander GC High-Resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. Nature Comm. 10(1032) (2019). [PubMed: 30833564]• The authors used conventional defocused-based cryo-EM to show that, by using a 200 keV microscope, one is still able to solve the structures of small (less than 100 kDa) proteins at sub-nanometer resolution.

[21] -. Fan X, Wang J, Zhang X, Yang Z, Zhang JC, Zhao L, Peng HL, Lei J, and Wang HW Single Particle cryo-EM reconstruction of 52kDa steptavidin at 3.3 Angstrom resolution. Nature Comm. 10(2386) (2019). [PubMed: 31160591]• By using a Cs-corrected microscope with a Volta phase plate and graphene-coated grids, the authors are able to solve the structure of streptavidin to high resolution.

[22] -. Dyson HJ, and Wright PE Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol 12(1): 54–60 (2002). [PubMed: 11839490]

[23] -. Oyen D, Torres JL, Cottrell CA, King CR, Wilson IA, and Ward AB Cryo-EM structure of P. falciparum circumsporozoite protein with a vaccine-elicited antibody is stabilized by somatically mutated inter-Fab contacts. Science Advances. 4(10) (2018).

[24] -. Bonetti D, Troilo F, Brunori M, Longhi S, Gianni S How Robust Is the Mechanism of FoldingUpon-Binding for an Intrinsically Disordered Protein? Biophys J. 114(8): 1889–1894 (2018). [PubMed: 29694866]

[25] -. Rogers JM, Oleinikovas V, Shammas SL, Wong CT, De Sancho D and Clarke J Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. PNAS 111(43): 15420–5 (2014). [PubMed: 25313042]

[26] -. Borgnia MJ, Banerjee S, Merk A, Matthies D, Bartesaghi A, Rao P, Pierson J, Earl LA, Falconieri V, Subramaniam S, and Milne JLS Using Cryo-EM to Map Small Ligands on Dynamic Metabolic Enzymes: Studies with Glutamate Dehydrogenase. Mol. Pharmacol 89(6): 645–651 (2016). [PubMed: 27036132]

[27] -. Luo BH, Karanicolas J, Harmacek LD, Baker D, and Springer TA Rationally Designed Integrin β3 Mutants Stabilized in the High Affinity Conformation. J. Biol. Chem 284(6): 3917–3924 (2009). [PubMed: 19019827] [PubMed: 19019827]

[28] -. Plückthun A Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy. Annu. Rev. Pharmacol. Toxicol 55: 489–511 (2015). [PubMed: 25562645]

[29] -. Guillard S, Kolasinska-Zwierz P, Debreczeni J, Breed J, Zhang J, Bery N, Marwood R, Tart J, Overman R, Stocki P, Phillips C, Rabbitts T, Jackson R, and Minter R Structure and functional characterization of a DARPin which inhibits Ras nucleotide exchange. Nature Comm. 8(16111) (2017).

[30] -. Mohan K, Ueda G, Kim AR., Jude KM, Guo Y, Hafer M, Miao Y, Saxton RA, Piehler J, Sankaran VG, Baker D, and Garcia KC Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. Science. 364(6442). (2019).

[31] -. Kratz PA, Böttcher B, and Nassal M, Native display of complete foreign protein domains on the surface of hepatitis B virus capsids. PNAS. 96(5): 1915–20 (1999). [PubMed: 10051569]

[32] -. Lyumkis D Challenges and opportunities in cryo-EM single-particle analysis. J. Biol. Chem 294(13): 5181–5197 (2019). [PubMed: 30804214]

[33] -. Lai YT, Jiang L, Chen W, Yeates TO. On the predictability of the orientation of protein domains joined by a spanning alpha-helical linker. Prot. Eng. Des. Sel 28(11): 491–499 (2015).

[34] -. Jose J, Tang J, Taylor AB, Baker TS, and Kuhn RJ Fluorescent Protein-Tagged Sindbis Virus E2 Glycoprotein Allows Single Particle Analysis of Virus Budding from Live Cells. Viruses. 7(12): 6182–99 (2015). [26633461] [PubMed: 26633461]

[35] -. McGonigle R, Yap WB, Ong ST, Gatherer D, Bakker SE, Tan WS, and Bhella D An N-terminal extension to the hepatitis B virus core protein forms a poorly ordered trimeric spike in assembled virus-like particles. J. Struct. Biol 189(2): 73–80 (2015). [PubMed: 25557498]

[36] -. Fleissner MR, Cascio D, and Hubbell WL Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. Protein Sci. 18(5): 893–908 (2009). [PubMed: 19384990]

[37] -. Huber TR, McPherson EC, Keating CE, and Snow CD Installing Guest Molecules at Specific Sites within Scaffold Protein Crystals. Bioconjug. Chem 29(1): 17–22 (2018). [PubMed: 29232505]

[38] -. Penczek P Chapter One - Fundamentals of Three-Dimensional Reconstruction from Projections. Methods in Enzymology. 482: 1–33 (2010). [PubMed: 20888956]

[39] -. Drulyte I, Johnson RM, Hesketh EL, Hurdiss DL, Scarff CA, Porav SA, Randon NA, Muench SP, and Thompson RF Approaches to altering particle distributions in cryo-electron microscopy sample preparation. Acta Crystallogr. D Struct. Biol 74(6): 560–571 (2018). [PubMed: 29872006]

[40] -. Liu YT, Jih J, Dai X, Bi GQ, and Zhou ZH Cryo-EM structures of herpes simplex virus type 1 portal vertex and packaged genome. Nature. 570: 257–261 (2019). [PubMed: 31142842]

[41] -. Jung J, Grand T, Thomas DR, Diehnelt CW, Grigorieff N, and Joshua-Tor L High-resolution cryo-EM structures of outbreak strain human norovirus shells reveal size variations. PNAS. 116(26): 12828–12832 (2019). [PubMed: 31182604]

[42] -. Baggen J, Liu Y, Lyoo H, van Vliet ALW, Wahedi M, de Bruin JW, Roberts RW, Overduin P, Meijer A, Rossmann MG, Thibaut HJ, and van Kuppeveld FJM Bypassing pan-enterovirus host factor PLA2G16. Nature Comm. 10(1): 3171 (2019).

[43] -. Tan YZ, Baldwin PR, Davis JH, Williamson JR, Potter CS, Carragher B, and Lyumkis D Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nature Methods. 14: 793–796 (2017). [PubMed: 28671674] [PubMed: 28671674] • Tilting the sample during data collection allows for acquisition of projections from multiple angles, providing information normally lost due to the preferred orientations of particles on the grid.

[44] -. Padilla J, Colovos C, and Yeates TO Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. PNAS 98(5): 2217–2221 (2001). [PubMed: 11226219]

[45] -. Lai YT, Cascio D, and Yeates TO Structure of a 16-nm Cage Designed by Using Protein Oligomers. Science. 336(6085): 1129 (2012). [PubMed: 22654051]

[46] -. Lai YT, Reading E, Hura GL, Tsai KL, Laganowsky A, Asturias F, Trainer JA, Robinson CV, Yeates TO Structure of a designed protein cage that self-assembles into a highly porous cube. Nature Chemistry 6: 1065–1071 (2014).

[47] -. Jeong WH, Lee H, Song DH, Eom JH, Kim SC, Lee HS, Lee H, and Lee JO Connecting two proteins using a fusion alpha helix stabilized by a chemical crosslinker. Nature Comm. 7(11031) (2016).

[48] -. Youn SJ, Kwon NY, Lee JH, Kim JH, Choi J, Lee H, and Lee JO Construction of novel repeat proteins with rigid and predictable structures using a shared alpha helix. Sci. Rep 7: 2595 (2017). [PubMed: 28572639]

[49] -. Wu Y, Batyuk A, Honegger A, Brandl F, Mittl PRE, and Plückthun A Rigidly connected multispecific artificial binders with adjustable geometries. Sci. Rep 7(1): 11217 (2017). [PubMed: 28894181] [PubMed: 28894181] • This paper describes a method of adjoining multiple DARPins

in geometrically constrained orientations by fusing them through continuous α-helical linkers. The authors were able to create multivalent constructs with specific orientation between the DARPins while retaining the high specificity and affinity of each DARPin component.

[50] -. Coscia F, Estrozi LF, Hans F, Malet H, Noirclerc-Savoye M, Schoehnm G, and Petosa C Fusion to a homo-oligomeric scaffold allows cryo-EM analysis of a small protein. Sci. Rep 6: 30909 (2016). [PubMed: 27485862] [PubMed: 27485862] • Maltose Binding Protein was genetically fused to a dodecameric glutamine synthetase scaffold by a continuous alpha helical connection. The MBP was held rigid enough to resolve its features at intermediate resolution by cryo-EM.

[51] -. Martin TG, Bharat TA, Joerger AC, Bai XC, Praetorius F, Fersht AR, Dietz H, and Scheres SH Design of a molecular support for cryo-EM structure determination. PNAS. 113(47): 7456–7463 (2016).

[52] -. Zhang C, Cantara W, Musier-Forsyth K, Grigorieff N, and Lyumkis Analysis of discrete local variability and structural covariance in macromolecular assemblies using Cryo-EM and focuses classification. Ultramicroscopy. 203: 170–180 (2019). [PubMed: 30528101] [PubMed: 30528101] • The authors genetically fused the HIV-I TAR RNA to a double-helical junction of the 23S ribosome. Using focused classification, the target RNA was visible at intermediate resolution via cryo-EM.

[53] -. Veesler D, Dreier B, Lichière J, Tremblay D, Moineau S, Spinelli S, Tegoni M, Plückthun A, Campanacci V, and Cambilau C Crystal structure and function of a DAPRin neutralizing inhibitor of lactococcal phage TP901–1: comparison of DARPin and camelid VHH binding mode. J. Biol. Chem 284(44): 30718–26 (2009). [PubMed: 19740746]

[54] -. Amstutz P, Koch H, Binz HK, Deuber SA, and Plückthun A Rapid selection of specific MAP kinase-binders from designed ankyrin repeat protein libraries. Protein Eng. Des. Sel 19(5): 219– 29 (2006). [PubMed: 16551653]

[55] -. Liu Y, Gonen S, Gonen T, and Yeates TO Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. PNAS. 115(13): 3362–3367 (2018). [PubMed: 29507202] [PubMed: 29507202] • This paper demonstrated the successful design of a modular cryo-EM scaffold by genetically fusing a DARPin adaptor to a high symmetry (tetrahedral) protein core by way of a continuous α-helical linker. The DARPin adaptor was held rigidly enough to resolve its structure at near-atomic resolution.

[56] -. Liu Y, Huynh DT, and Yeates TO A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. Nature Comm. 10(1): 1864 (2019). [PubMed: 31015551] •• The protein GFP was bound to the scaffold developed in (Liu et al, 2018), with an anti-GFP DARPin genetically fused to a tetrahedral protein cage. The scaffold facilitated cryo-EM structure determination of the bound cargo, with some atomic features of GFP visible.

[57] -. Yao Q, Weaver SJ, Mock JY, and Jensen GJ Fusion of DARPin to Aldolase Enables Visualization of Small Protein by Cryo-EM. Structure. 27(7): 1148–1155 (2019). [PubMed: 31080120] [PubMed: 31080120] •• The protein GFP was bound to an anti-GFP DARPin that was genetically fused to an aldolase enzyme exhibiting D2 symmetry, among other scaffolding choices tested. The bound cargo was then imaged on the symmetric scaffold by cryo-EM.

[58] -. Matthies D, Bae C, Toombes GE, Fox T, Bartesaghi A, Subramaniam S, and Swartz KJ Singleparticle cryo-EM structure of a voltage-activated potassium channel in lipid nanodiscs. Elife 5: 37558 (2018).

[59] -. Laverty D, Desai R, Uchański T, Masiulis S, Stec WJ, Malinauskas T, Zivanov J, Pardon E, Steyaert J, Miller KW, and Aricescu AR Cryo-EM structure of the human α1β3γ2 GABAA receptor in a lipid bilayer. Nature. 565(7740): 516–520 (2019). [PubMed: 30602789]

[60] -. Barnwal RP, Yang F, Varani G Applications of NMR to structural determination of RNAs large and small. Arch. Biochem. Biophys 628:42–56 (2017). [PubMed: 28600200]

[61] -. Reyes F, Garst A, and Bately R Chapter 6 - Strategies in RNA Crystallography. Methods in Enzymology 496: 119–139 (2009).

[62] -. Ferré-D'Amaré AR, Zhou K, Doudna JA A General Module for RNA Crystallization. J. Mol. Biol 279: 621–631 (1998). [PubMed: 9641982]

[63] -. Zhang K, Keane SC, Su Z, Irobalieva RN, Chen M, Van V, Sciandra CA, Marchant J, Heng X, Schmid MF, Case DA, Ludtke SJ, Summers MF, and Chiu W Structure of the 30 kDa HIV-1 RNA Dimerization Signal by a Hybrid Cro-EM, NMR and Molecular Dynamics Approach. Structure. 26(3): 490–498 (2018). [PubMed: 29398526]

[64] -. Ferré-D'Amaré AR, and Doudna JA Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. J. Mol. Biol 295(3): 541–56 (2000). [PubMed: 10623545]

[65] -. Ferré-D'Amaré AR. Use of the spliceosomal protein U1A to facilitate crystallization and structure determination of complex RNAs. Nature Methods. 52(2): 159–167 (2010).

CHAPTER THREE: X-ray structure of a designed rigidified imaging scaffold engineered to bind the therapeutic protein target BARD1

The following is a modified manuscript of an accepted research article in:

Acta Crystallographica Section F

80 (5): 107-115 (2024)

DOI: 10.1107/S2053230X2400414X

X-ray crystal structure of a designed rigidified imaging scaffold in the ligand-free conformation

Matthew P. Agdanowski, Roger Castells-Graells, Michael R. Sawaya, Duilio Cascio, Todd O. Yeates and Mark A. Arbing

Synopsis: An imaging scaffold engineered to bind and study therapeutic protein targets has been crystallized at 3.8 Å resolution. Cargo protein binding DARPins are positioned within the large solvent channels of an unusually porous crystal lattice suggesting that it may be possible to soak crystals with small target proteins to determine their structures.

Abstract

Imaging scaffolds composed of designed protein cages fused to Designed Ankyrin Repeat Proteins (DARPins) have enabled structure determination of small proteins by cryo-electron microscopy (cryo-EM). One particularly well-characterized scaffold type is a symmetric tetrahedral assembly comprised of 24 subunits, 12 A and 12 B, which has three cargo-binding DARPins positioned on each vertex. Here, we report the X-ray crystal structure of a representative tetrahedral scaffold at 3.8 Å resolution in the apo state. The X-ray crystal structure complements recent cryo-EM findings on a closely related scaffold, while also suggesting potential utility for crystallographic investigations. As observed in our crystal

structure, one of the three DARPins, which serve as modular adaptors for binding diverse “cargo” proteins, present on each of the vertices is oriented towards a large solvent channel. The crystal lattice is unusually porous suggesting that it may be possible to soak crystals of the scaffold with small (≤30 kDa) protein cargo ligands and subsequently determine cage-cargo structures via X-ray crystallography. Our results suggest the possibility that cryo-EM scaffolds may be repurposed for structure determination by X-ray crystallography thus extending the utility of EM scaffold designs for alternative structural biology applications.

Keywords: DARPin; protein cage; protein design; imaging scaffold

Introduction

Imaging scaffolds composed of protein cages fused to Designed Ankyrin Repeat Proteins (DARPins) have emerged as a powerful technology for determining high-resolution structures of small proteins using single particle cryogenic electron microscopy (cryo-EM) (Liu *et al.*, 2019, 2018; Castells-Graells *et al.*, 2023; Yeates *et al.*, 2020). Binding small (~30 kDa) protein targets to the modular DARPin domains of large, half-megadalton, symmetric scaffolds increases the size of the target into the range amenable to single particle cryo-EM image processing. Using this approach, recent studies have achieved near atomic resolution for small proteins including the oncogenic protein KRAS (Castells-Graells *et al.*, 2023) in apo and ligand bound forms. While initial development has focused on cryo-EM we have also pursued a parallel approach to scaffold-facilitated structure determination using X-ray crystallography.

Structure determination by X-ray crystallography is a laborious process requiring extensive screening to identify conditions that produce crystals suitable for structure determination.

Experimental data from high throughput crystallization screening facilities shows that approximately 21% of protein targets subjected to screening ultimately result in crystallographic models (Lynch *et al.*, 2023). Given this relatively low success rate there has been a strong focus on “salvage” pathways to obtain structures of proteins of interest. Successful strategies include modification of protein surface properties (e.g. pI, hydropathy, and surface entropy) by chemical modification (Kim *et al.*, 2008) and site directed mutagenesis (Derewenda, 2004) or the use of crystallization chaperones to promote lattice formation. The latter technique is roughly divided into two approaches: protein fusions or complexation with non-covalently bound epitope-specific protein binders. Examples of the former approach are fusion with maltose binding protein via flexible or rigid linkers (Waugh, 2016) or incorporation of T4 lysozyme into loops of membrane proteins (Thorsen *et al.*, 2014) to increase solvent-accessible surface area amenable to forming crystal contacts. Examples of the latter technique are the use of protein-specific binders such as nanobodies, FAb fragments, and related derivatives to generate protein complexes more amenable to crystallization (Koide, 2009).

Designed Ankyrin repeat proteins (DARPins), synthetic protein binding proteins derived from naturally occurring protein binding motifs, have also been used as crystallization chaperones (Mittl *et al.*, 2020) and, more recently, as “adapters” to bind small proteins to imaging scaffolds for cryo-EM structure determination (Liu *et al.*, 2019, 2018; Castells-Graells *et al.*, 2023). As part of a project targeting oncogenic protein targets we generated DARPins against the C-terminal domain of the oncogenic protein BARD1 using a yeast display system and subsequently fused the anti-BARD1 DARPins to a previously characterized tetrahedral protein cage. To investigate whether EM imaging scaffolds with rigid DARPin fusions can act as crystallization chaperones, thus extending their utility for structural studies, we determined the X-ray structure of one of these imaging scaffolds in the ligand-free state. Our X-ray crystal structure suggests that

cryo-EM scaffolds may have multiple applications in the elucidation of structures of small proteins.

Materials and methods

Macromolecule production

BARD1 expression and DARPin selection:

A construct encoding the BARD1 tandem BRCT domains (amino acids 423-777) with an N-terminal SUMO fusion protein followed by a HRV 3c protease site, an AVI tag, and a TEV protease site was synthesized in pET29b (Twist Bioscience). The BARD1 BRCT construct was expressed in *E. coli* BL21-Gold (DE3) using Terrific Broth and overnight induction at 18°C with 0.5 mM IPTG; biotinylated protein was produced *in vivo* by co-expression of BirA (Addgene plasmid #102962) and the addition of biotin (final concentration 50 μM) to the media at the time of induction (Fairhead & Howarth, 2015). Cells were harvested by centrifugation, resuspended in Buffer A (25 mM Tris-HCl pH 8.0, 500 mM NaCl, 5% glycerol, 20 mM imidazole, 5 mM β-ME) supplemented with 1 mM EDTA, 1 mM PMSF, and Complete protease inhibitor (Roche). Cells were lysed with three passes through an Emulsiflex C-3 (Avestin) at 15K PSI and the lysate subsequently clarified by centrifugation. The BARD1 SUMO fusion was purified from the clarified supernatant using a 5 ml HisTrap Crude FF (Cytiva) column with bound protein eluted with Buffer B (Buffer A with 300 mM imidazole). TEV (for removal of all N-terminal tags) or 3c protease (for removal of the SUMO moiety but retention of the AVI tag and TEV protease site) was added to the eluted protein and the digestion mix was dialyzed against 2L Buffer A overnight at 4°C. The following day SDS-PAGE was used to determine that digestion was complete and subsequently the reaction mix was loaded on a 5 ml HisTrap with the flow through collected and further purified by size exclusion chromatography using a Superdex 75 (Cytiva)

column equilibrated with Buffer C (25 mM HEPES pH 7.5, 300mM NaCl, 5% glycerol, 1mM DTT). Fractions containing BARD1 were pooled, concentrated, flash-frozen with liquid nitrogen, and stored at -80°C.

DARPins that bind BARD1 were identified using a yeast DARPin surface display system (Morselli *et al.*, 2024). A cell population displaying BARD1 binders was enriched using two rounds of magnetic activated cell sorting (MACS) followed by five rounds of fluorescence activated cell sorting (FACS) using a Bio-Rad S3 cell sorter. The selections were carried out using previously described methods (Chao *et al.*, 2006; McMahon *et al.*, 2018). Briefly, the MACS experiments were performed using Dynabeads MyOne Streptavidin T1 beads (Invitrogen) while FACS experiments used an AlexaFluor488-conjugated anti-HA monoclonal antibody (Invitrogen) to select DARPin-displaying cells while cells that bound biotinylated BARD1 were selected by alternating fluorescent anti-biotin conjugates, Streptavidin R-Phycoerythrin or NeutrAvidin Rhodamine Red-X (both from Invitrogen). Target protein concentration was decreased in each selection round to isolate higher affinity binders with the initial MACS experiment carried out using 1.0 μM protein and the final FACS selection with 30 nM protein. Enriched cell populations were grown in non-inducing media and a 50 μl cell sample was centrifuged, washed with water, and lysed by addition of an equivalent volume of 40 mM NaOH and heated for 45 minutes at 95°C. This cell lysate served as the template for PCR-amplification of enriched DARPin sequences; PCR amplification was carried out using primers (DARP.pYDS.Amp.For., 5'-GATGAAGTTCGTATTCTGATGGCAAATGG-3'; DARP.pYDS.Amp.Rev., 5'-CGGTGTTTTACCAAATTTATCCTGGGC-3') that bind conserved sequences in the N- and C-caps of the DARPin. The PCR reaction used PrimeStar GXL polymerase (Takara) with a 30 second extension and 20 amplification cycles. PCR products were purified by gel extraction and subjected to Next Generation Sequencing (Genewiz, NJ).

Forward and reverse reads were merged with NGMerge (Gaspar, 2018) and sequence abundance and characteristics analyzed using the MAMETS program (Morselli *et al.*, 2024).

The most abundant DNA sequences encoding putative anti-BARD1 DARPins were synthesized and cloned into pET29b (Twist Bioscience) with an N-terminal His6 tag for expression and purification. DARPins were expressed and purified using a similar procedure as for BARD1 with the substitution of 50 mM Tris pH 8.0, 300 mM NaCl, 5% glycerol, 5 mM β-mercaptoethanol, 20/300 mM imidazole as the affinity chromatography buffers and 20 mM Tris pH 7.5, 150 mM NaCl as the size exclusion chromatography buffer. Screening for DARPins that formed a stable complex with BARD1 was performed with biolayer interferometry (BLI) and subsequently confirmed using analytical size exclusion chromatography (AnSEC). BLI experiments were carried out with an Octet Red 96e (Sartorius) and NTA Biosensors. The His-tagged DARPins were diluted to 25 µg/mL in kinetic buffer (PBS with 0.1% BSA and 0.02% tween 20) and loaded on NTA Biosensors by dipping the biosensors into a 96-well plate (Greiner 655209) with 200 µL/well DARPin for five minutes. Biosensors were then dipped in fresh kinetic buffer to establish baseline (three minutes) and subsequently were dipped in BARD1 (10 µg/ml) for five minutes (association step), and then transferred to fresh buffer for five minutes (dissociation step). Each experiment was doubly reference subtracted using biosensors with zero analyte (BARD1) or that were not loaded with DARPins. Lead candidates were confirmed to bind BARD1 by adding a 3-fold molar excess of the DARPin to BARD1 and injecting the protein mixture onto an analytical SEC70 column (Bio-Rad Laboratories) equilibrated in 20 mM Tris pH 7.5, 150 mM NaCl. Fractions were collected and samples from elution peaks were electrophoresed on SDS-PAGE to identify the protein constituents.

Design, expression, and purification of the imaging scaffolds:

Anti-BARD1 DARPin sequences identified by yeast display were genetically fused to a tetrahedral nanocage via helical extension with the N-terminus of the DARPin sequence fused to the C-terminus of the cage component (Liu *et al.*, 2018). Stabilizing mutations (Castells-Graells *et al.*, 2023) were incorporated to rigidify the trimer interface. DNA sequences were synthesized (Twist Bioscience) and incorporated in bacterial expression vectors: pSAM (Liu *et al.*, 2018) for subunit A and pET22b for the subunit B-DARPin fusion.

The plasmids containing both components of the imaging scaffold were co-transformed into *E. coli* BL21-Gold (DE3) and expression and solubility of the two cage components was evaluated at 18 and 37°C. Designs where both components were solubly expressed and could be affinity purified using NiNTA beads were chosen for large scale purification. Imaging scaffolds were grown in 1L LB, supplemented with ampicillin and kanamycin, to an OD_{600} of ~0.6 and protein expression induced with 0.5 mM IPTG. Proteins were expressed at 18°C overnight (~18 hours) and harvested by centrifugation. Cell pellets were resuspended in buffer D (50 mM Tris pH 8.0, 300 mM NaCl, 20 mM imidazole) and lysed using the same conditions as for SUMO-BARD1 but the protein was purified by affinity chromatography using a gravity column and Buffer E (50 mM Tris pH 8.0, 300 mM NaCl, 500 mM Imidazole) as the elution buffer. Fractions were assessed with SDS-PAGE and those containing both scaffold components were concentrated with a 100 kDa Amicon Ultra-15 concentrator (Millipore Sigma) and further purified by size exclusion chromatography using a 16/600 Suprose6 column (Cytiva) equilibrated with 20 mM Tris pH 8.0, 100 mM NaCl. Peak fractions were analyzed by SDS-PAGE and fractions containing both components were pooled, concentrated with a 100 kDa Amicon Ultra-15 concentrator, and the purified protein stored at 4°C pending subsequent X-ray and electron microscopy experiments.

For scaffold analysis via negative stain electron microscopy, a 5 µl sample of concentrated protein adjusted to ~50 µg/mL was applied to a glow-discharged Formvar/Carbon 300 mesh (Ted Pella Inc) for 1 minute and blotted to remove any excess liquid. After blotting, the grid was washed 3 times with sterile MilliQ water before being stained with a 2% uranyl acetate solution for 1 minute. Micrographs were taken on Tecnai T12 and Talos F200C electron microscopes. Negative stain micrographs were converted to .MRC format and imported into cryoSPARC for processing. Micrographs were CTF corrected using patch CTF correction and ~3,000 particles were manually picked for further analysis. Two rounds of 2D classification resulted in rough averages that were used to assess scaffold assembly. The best 2D classes containing roughly 2,000 particles were used to create a low resolution *ab initio* 3D map with T symmetry enforced in which the X-ray structure was docked.

Crystallization

Crystallization screening of BARD1-specific imaging scaffolds using the hanging drop vapor diffusion method were conducted at the UCLA-DOE Crystallization Core. Imaging scaffolds (16 mg/mL) and BARD1 (3 mg/mL) were mixed at a 1:1 ratio (v/v) and five 96-well screens were set up using 1:1, 2:1, and 1:2 ratios of protein to reservoir solution (final drop volume of 210 nL) for each condition using a TTP Labtech Mosquito. Screens were incubated at room temperature (~20°C). Crystals of the DARP3 scaffold were grown by mixing protein solution 1:1 with reservoir solution (JCSG+ condition D11: 0.14 M calcium chloride, 0.07 M sodium acetate, pH 4.6, 14% v/v isopropanol, 30% v/v glycerol). Prismatic crystals (approximately 70 microns thick) appeared after nine days and were mounted in loops, flash frozen in liquid nitrogen, and stored in liquid nitrogen until data collection.

Data collection and processing

X-ray diffraction data were collected at the microfocus beamline 17-ID-2 of the National Synchrotron Light Source II located at Brookhaven National Laboratory. Data collection was at a temperature of 100 K with 0.2 degree oscillation (1800 frames collected) and an X-ray wavelength of 0.9793 Å. Diffraction data were indexed, integrated, scaled, and merged using the programs XDS and XSCALE (Kabsch, 2010). Data collection statistics are reported in Table 3.1.

Structure solution and refinement

The structure was solved by molecular replacement using the program Phaser (McCoy *et al.*, 2007) and a search model consisting of subunit B lacking the DARPin domain (PDBid 5CY5). The molecular replacement solution was unambiguous, exhibiting a high positive log likelihood gain (LLG) of 2533. Difference maps revealed positive residual density for the DARPin domains. A second round of molecular replacement, keeping the cage core fixed, was performed searching for three copies of the DARPin domain using a GFP-specific DARPin (PDBid 5MA6; 77% sequence identity to BARD1-specific DARPin) as the search model. The molecular replacement solution further improved the atomic model as evidenced by an increase in LLG to 3348 and decrease in R-factors (R_{work}=0.299 R_{free}=0.327). Manual model building was performed using the graphics program Coot (Emsley *et al.*, 2010). Atomic refinement was performed with the program Phenix (Liebschner *et al.*, 2019). To minimize overfitting to the 3.8 Å data, non-crystallographic symmetry restraints and conformational restraints to a reference model consisting of PDB entries 8G3K (cage core cryoEM structure at 2.2 Å resolution) and 5MA6 (GFP-specific DARPin cryo-EM structure at 2.3 Å resolution). No residual density was observed near the DARPin cargo-binding loops, indicating that BARD1 was not bound in this crystal form.

Final atomic refinement statistics are reported in Table 3.1. Structure illustrations were created using PyMOL (Schrödinger, LLC).

Results

Selection and characterization of DARPins against BARD1

BARD1 (BRCA1-associated RING domain protein 1) is an important oncogenic protein that forms a heterodimeric complex with BRCA1 (Breast cancer gene 1); the complex has E3 ubiquitin activity associated with DNA damage repair and tumor suppression (Brzovic *et al.*, 2001; Ruffner *et al.*, 2001; Wu *et al.*, 1996) and mutations in both BRCA1 and BARD1 are associated with breast, ovarian and pancreatic cancers (De Brakeleer *et al.*, 2016; Foulkes, 2008). A yeast DARPin display system was used to generate DARPins against the ligand-binding C-terminal BRCT and ankyrin domain of BARD1 (Watters *et al.*, 2020). After magnetic- and fluorescence-activated cell sorting DARPin sequences were isolated from the enriched cell population by PCR and the sequence abundance and diversity determined by Next Generation Sequencing (NGS) of PCR amplicons. The ten most abundant sequences ranged between 0.75 to 12% of the total number of sequences (353K) obtained from NGS sequencing. Five of these sequences were cloned into bacterial expression vectors and were subsequently expressed and purified by affinity chromatography. Interaction with BARD1 was confirmed by biolayer interferometry, and analytical size exclusion chromatography and SDS-PAGE analysis (Figure 3.1).

Design of the imaging scaffold and biochemical characterization

The helical N-termini of evolved anti-BARD1 DARPins were genetically fused to the helical C-terminus of the B subunit of a two-component tetrahedral protein nanocage (Cannon *et al.*, 2020) using recently described stabilizing “staple” mutations at the subunit B trimer interface (Castells-Graells *et al.*, 2023); subunit A of the tetrahedral assembly is invariant and is the same for all designs. In total three subunit B-DARPin fusion constructs were made. Together, both components co-assemble into a discrete particle obeying tetrahedral symmetry containing 12 copies of the DARPin-fusion subunit and 12 copies of the non-fusion component (four sets of each trimeric protein). The total assembly has a predicted mass of ~660 kDa and a diameter of approximately 19 nm.

The plasmids containing the two subunits were co-transformed into *E. coli* and protein cages expressed and purified by affinity and size exclusion chromatography (SEC). Of the three designs that were investigated only one, DARP3, formed a soluble assembly as assessed by analytical SEC (Fig. 3.2A) and SDS-PAGE (Fig. 3.2B). Negative stain electron microscopy (Fig. 3.2C, D) analysis showed particles with the expected tetrahedral geometry and size of approximately 19 nm, with a preferred orientation displaying its 2-fold axis of symmetry.

Protein crystallization and structure determination

The DARP3 assembly was subjected to crystallization screening in the apo and ligand-bound state. In mixing studies it was determined that the DARP3 assembly could tolerate only four BARD1 molecules per cage with BARD1 amounts in stoichiometric ratios above 4 cargo molecules per cage (or one BARD1 per DARPin trimer at each vertex) resulting in immediate and severe aggregation as indicated by an increase in opacity of solution upon mixing; this

suggests some degree of steric clashing between BARD1 proteins at cage vertices when more than one BARD1 was bound to a DARPin trimer. As a result the sample was set up with a 1:3 ratio of cargo:DARPin trimer for the ligand bound state.

No crystals were found in the crystallization screens for the apo DARP3 assembly however crystals in space group I222 that diffracted to 3.81 Å were identified in one condition for the screens of the BARD1-DARP3 assembly. The structure was solved by molecular replacement using a single component of the cage (subunit B) and an isolated DARPin molecule as search models. Three copies of Subunit A were subsequently fit to the electron density manually in Coot (Emsley *et al.*, 2010). There was no electron density for the BARD1 cargo protein indicating that we had crystallized and solved the structure of the apo state of our scaffold. The asymmetric unit contains three copies of subunit A (chains A-C in the PDB file) and three copies of the subunit B-DARPin fusion (chains D-F in the PDB file) with the tetrahedral assembly generated via symmetry operations (Fig. 3.3A). The structure of the core assembly was first crystallized without DARPin fusions (Cannon *et al.*, 2020) and there is excellent agreement between the structures of the conserved cage core chains with an average rmsd of 0.47 +/- 0.03 Å for the superposition 141 Cα of chains D-F of the DARP3 assembly with chain B of the T33-51 assembly; a structure-based superposition, using the Coot SSM tool, of chains A-C of the DARP3 assembly with chain A of T33-51 had an rmsd of 0.36 Å for all three comparisons with alignment of 137, 134, 136 amino acids for DARP3 chains A, B, and C, respectively.

The DARP3 assembly crystals have a very high solvent content of 71.47% and a Matthew's coefficient of 4.31. As a result the lattice has large solvent filled channels with an approximate cross-section of 120 x 180 Å that is periodically restricted by the protruding of the DARPin moiety of chain E into the solvent channel (Fig. 3.3B). The DARPin moieties of chains D and F

are involved in mediating crystal contacts in the crystal lattice and are thus unavailable for cargo binding. The substrate binding face of the chain E DARPin is oriented such that substrate binding is possible without creating steric clashes with other components of the lattice. Superposition of the anti-GFP DARPin in GFP-bound state (Hansen *et al.*, 2017) on the anti-BARD1 DARPin in our structure (PDBid 5MA6, chain B residues Lys16-Ala168; DARP3 assembly, PDBid 8V9O, chain E residues Lys169-Ala321) gives an rmsd of 0.57 Å for the superposition of 153 Cα atoms with a sequence identity of 77% and supports the ability of the lattice to support cargo binding as the GFP barrel, with dimensions of 24 x 42 Å (Ormö *et al.*, 1996), is oriented in such a way that it does not interfere with the cage core structure (Fig. 3.3C). Likewise, superposition of the structure of the anti-KRAS DARPin bound to KRAS (Guillard *et al.*, 2017)(PDBid 5O2S) on DARP3 chain E (Fig. 3.3D; rmsd of 0.96 Å for the superposition of 155 Cα atoms with a sequence identity of 75.3%) also shows that binding of a small globular protein cargo within the solvent channel is also possible without physically clashing with cage core components.

Discussion

We sought to validate our newly developed DARPin display system (Morselli *et al.*, 2024) and to use the selected DARPins in conjunction with our suite of designed protein cages to structurally characterize an important cancer-related protein, BARD1. Using yeast display we identified a number of candidate anti-BARD1 DARPins and four of these were found, via analytical size exclusion chromatography, to form stable complexes with BARD1. Three of these candidate DARPins were fused to our improved imaging scaffold using an established protein fusion strategy (Castells-Graells *et al.*, 2023) and one of the DARPin-cage fusions was expressed and purified to high yields. SEC and SDS-PAGE analysis showed that the cage fusion eluted as a high molecular weight species that contained both subunits in a roughly 1:1 stoichiometric ratio.

Negative stain EM analysis confirmed that we had successfully purified a homogeneous assembly of the expected size and shape.

The primary objective of our protein cage design projects has been to design imaging scaffolds for structural characterization of small proteins by cryo-EM. If the designed cage and cargo proteins are available in sufficient quantities we have also pursued structural characterization of our designs, in apo and ligand-bound forms, by X-ray crystallography. In this project a single design was expressed in quantities sufficient for crystallization screening. Interestingly, during solution binding studies, it was observed that rapid aggregation would occur when the cargo and cage were mixed at ratios corresponding to one BARD1 per DARPin binding site. This result is not totally surprising given the BARD1 construct used in this study consists of two domains that adopt an extended structure (Dai *et al.*, 2021) and the orientation of BARD1 binding to the DARPin is unknown. We hypothesize this elongated structure may be positioned such that a substantial part of the BARD1 cargo crosses the threefold axis and causes steric clashes with symmetrically related cargo copies. This, compounded with the high affinities that DARPins possess for their cognate ligand, likely leads to rapid association between the two causing cage dissociation and aggregation of dissociated cage subunits. We believe this aggregation will not occur once the scaffold is locked into the crystal lattice and only one DARPin is left available for ligand binding. In the crystallization trials in this study we loaded the cage with cargo at a 1:1 ratio of ligand to trimeric DARPin binding site to avoid scaffold dissociation.

The DARP3 scaffold with BARD1 cargo grew multiple prismatic crystals of approximately 70 microns in length which diffracted to 3.81 Å. While we have determined structures of similar DARPin-displaying scaffolds by electron microscopy, this is the first instance in which we have

determined the crystal structure of a designed cage with cargo-binding DARPin fusions. Unfortunately the structure is of the apo cage with no electron density seen for the BARD1 cargo. The most likely explanation for ligand dissociation is the composition of the crystallization solution which has a low pH (0.07 M sodium acetate pH 4.6) and contains a not insignificant concentration of a non-polar solution (14% isopropanol) which may interfere with protein-protein interactions and/or protein solubility.

Protein design efforts focused on creating self-assembling protein cages have been an active area of research since the early 2000's (Padilla *et al.*, 2001) and a significant number of designed cages have been crystallized and their structures determined (Table 3.2). The resolution of crystal structures for protein cages ranges from 2.1 – 7.08 Å with an average resolution of 3.62 +/- 1.68 Å and a median resolution of 3.5 Å for this set of 15 structures including the DARP3 scaffold from this study which is a derivation of T33-51H (PDB accession code 5CY5); if the current structure is excluded the set of cage structures has an average resolution of 3.61 +/- 1.34 Å with a median resolution of 3.45 Å. The resolution of the current structure (3.81 Å) is similar to the naked T33-51H cage (3.5 Å) and to the median resolution for crystallized protein cages. Higher resolution may be possible through optimization of our existing crystallization conditions or by finding alternative crystal forms via additional crystallization screening. This particular cage assembly has already benefited from strategically engineering staple mutations that stabilize the DARPin near the point of helical extension from the scaffold core (Castells-Graells *et al.*, 2023) and this new structure will facilitate ongoing protein engineering to further rigidify the scaffold for high resolution structural studies.

During processing and refinement, it was noted that the crystal contained a high solvent content (71.5%) resulting in large solvent-filled channels throughout the crystal. This agrees with our

experience that proteins of high symmetry tend to have fairly high solvent content as they require fewer unique contacts to generate the lattice. Interestingly one of the three DARPins present at a cage vertex is positioned within the channel formed by the lattice such that it is available for cargo binding. The other two DARPins (chains D and F) present on the vertex are involved in mediating crystal contacts with adjacent tetrahedral assemblies. With the exception of a single hydrogen bond (2.88 Å; between the carbonyl oxygen of Leu167 chain C and the CZ2 atom of Trp209 chain F) the variable cargo binding surfaces of the DARPins (chains D and F) are not involved in lattice contacts and protein-protein interactions occur through conserved invariant residues in the DARPin moieties.

The large solvent channels suggest the possibility that cage crystals could be soaked with protein substrates which could bind to the free DARPin binding sites, similar to techniques in which crystals are soaked in solutions of small ligands, allowing cargo protein structures to be determined. This would be a valuable addition to the structural biologists toolbox as an additional salvage pathway through which to determine the crystal structures of protein recalcitrant to crystallization. There are a number of possible complicating factors including that the solvent channels may not be big enough to allow proteins to freely diffuse throughout the lattice in the same way a small molecule can, or that penetration of the protein ligand is incomplete, leading to outer shell DARPin occupancy, but leaving the innermost lattice DARPins in their apo state. However, there is a significant upside in that the ligand binding loops from other DARPins could be grafted on the DARP3 scaffold allowing for easy soaking experiments and structure solution via molecular replacement. These ideas await future studies.

Acknowledgements

The authors would like to thank Genesis Falcon of the UCLA-DOE Institute X-ray and EM Structure Determination Core for assistance with crystallization screening. This research used beamline 17-ID-2 at the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). The authors declare no competing financial interests. MPA, RCG, and MAA conceived the project; MPA, MRS, DC, and MAA performed the experiments; MPA, RCG, MRS, TOY, and MAA analysed the data; and all were involved in writing the manuscript.

Author contributions

M.P.A., R.C.-G., and M.A.A. designed research; M.P.A., M.R.S., D.C., and M.A.A. performed research; M.P.A., R.C.-G., M.R.S., and M.A.A. analyzed data; and M.P.A., R.C.-G., M.R.S., T.O.Y. and M.A.A. wrote the paper.

References

Bale, J. B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T. O., Gonen, T., King, N. P. & Baker, D. (2016). *Science* **353**, 389–394.

Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M.-C. & Klevit, R. E. (2001). *Nat. Struct. Biol.* **8**, 833–837.

Cannon, K. A., Park, R. U., Boyken, S. E., Nattermann, U., Yi, S., Baker, D., King, N. P. & Yeates, T. O. (2020). *Protein Sci. Publ. Protein Soc.* **29**, 919–929.

Castells-Graells, R., Meador, K., Arbing, M. A., Sawaya, M. R., Gee, M., Cascio, D., Gleave, E., Debreczeni, J. É., Breed, J., Leopold, K., Patel, A., Jahagirdar, D., Lyons, B., Subramaniam, S., Phillips, C. & Yeates, T. O. (2023). *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2305494120.

Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M. & Wittrup, K. D. (2006). *Nat. Protoc.* **1**, 755–768.

Dai, L., Dai, Y., Han, J., Huang, Y., Wang, L., Huang, J. & Zhou, Z. (2021). *Mol. Cell* **81**, 2765-2777.e6.

De Brakeleer, S., De Grève, J., Desmedt, C., Joris, S., Sotiriou, C., Piccart, M., Pauwels, I. & Teugels, E. (2016). *Clin. Genet.* **89**, 336–340.

Derewenda, Z. S. (2004). *Struct. Lond. Engl. 1993* **12**, 529–535.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501.

Fairhead, M. & Howarth, M. (2015). Vol. *Site-Specific Protein Labeling: Methods and Protocols*, edited by A. Gautier & M. J. Hinner. pp. 171–184. New York, NY: Springer.

Foulkes, W. D. (2008). *N. Engl. J. Med.* **359**, 2143–2153.

Gaspar, J. M. (2018). *BMC Bioinformatics* **19**, 536.

Guillard, S., Kolasinska-Zwierz, P., Debreczeni, J., Breed, J., Zhang, J., Bery, N., Marwood, R., Tart, J., Overman, R., Stocki, P., Mistry, B., Phillips, C., Rabbitts, T., Jackson, R. & Minter, R. (2017). *Nat. Commun.* **8**, 16111.

Hansen, S., Stüber, J. C., Ernst, P., Koch, A., Bojar, D., Batyuk, A. & Plückthun, A. (2017). *Sci. Rep.* **7**, 16292.

Jorda, J., Leibly, D. J., Thompson, M. C. & Yeates, T. O. (2016). *Chem. Commun.* **52**, 5041–5044.

Kabsch, W. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132.

Kim, Y., Quartey, P., Li, H., Volkart, L., Hatzos, C., Chang, C., Nocek, B., Cuff, M., Osipiuk, J., Tan, K., Fan, Y., Bigelow, L., Maltseva, N., Wu, R., Borovilos, M., Duggan, E., Zhou, M., Binkowski, T. A., Zhang, R. & Joachimiak, A. (2008). *Nat. Methods* **5**, 853–854.

King, N. P., Bale, J. B., Sheffler, W., McNamara, D. E., Gonen, S., Gonen, T., Yeates, T. O. & Baker, D. (2014). *Nature* **510**, 103–108.

King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., André, I., Gonen, T., Yeates, T. O. & Baker, D. (2012). *Science* **336**, 1171–1174.

Koide, S. (2009). *Curr. Opin. Struct. Biol.* **19**, 449.

Lai, Y.-T., Cascio, D. & Yeates, T. O. (2012). *Science* **336**, 1129–1129.

Lai, Y.-T., Hura, G. L., Dyer, K. N., Tang, H. Y. H., Tainer, J. A. & Yeates, T. O. (2016). *Sci. Adv.* **2**, e1501855.

Lai, Y.-T., Reading, E., Hura, G. L., Tsai, K.-L., Laganowsky, A., Asturias, F. J., Tainer, J. A., Robinson, C. V. & Yeates, T. O. (2014). *Nat. Chem.* **6**, 1065–1071.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877.

Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. (2018). *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3362–3367.

Liu, Y., Huynh, D. T. & Yeates, T. O. (2019). *Nat. Commun.* **10**, 1864.

Lynch, M. L., Snell, M. E., Potter, S. A., Snell, E. H. & Bowman, S. E. J. (2023). *Acta Crystallogr. Sect. Struct. Biol.* **79**, 198–205.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Crystallogr.* **40**, 658–674.

McMahon, C., Baier, A. S., Pascolutti, R., Wegrecki, M., Zheng, S., Ong, J. X., Erlandson, S. C., Hilger, D., Rasmussen, S. G. F., Ring, A. M., Manglik, A. & Kruse, A. C. (2018). *Nat. Struct. Mol. Biol.* **25**, 289–296.

Mittl, P. R., Ernst, P. & Plückthun, A. (2020). *Curr. Opin. Struct. Biol.* **60**, 93–100.

Morselli, M., Holton, T. R., Pellegrini, M., Yeates, T. O. & Arbing, M. A. (2024). *Curr. Protoc.* **4**, e960.

Ormö, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y. & Remington, S. J. (1996). *Science* **273**, 1392–1395.

Ruffner, H., Joazeiro, C. A., Hemmati, D., Hunter, T. & Verma, I. M. (2001). *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5134–5139.

Schrödinger, LLC The PyMOL Molecular Graphics System, Version 1.2r3pre.

Thorsen, T. S., Matt, R., Weis, W. I. & Kobilka, B. (2014). *Struct. Lond. Engl. 1993* **22**, 1657–1664.

Watters, A. K., Seltzer, E. S., MacKenzie, D., Young, M., Muratori, J., Hussein, R., Sodoma, A. M., To, J., Singh, M. & Zhang, D. (2020). *Genes* **11**, 829.

Waugh, D. S. (2016). *Protein Sci. Publ. Protein Soc.* **25**, 559–571.

Wu, L. C., Wang, Z. W., Tsan, J. T., Spillman, M. A., Phung, A., Xu, X. L., Yang, M. C., Hwang, L. Y., Bowcock, A. M. & Baer, R. (1996). *Nat. Genet.* **14**, 430–440.

Yeates, T. O., Agdanowski, M. P. & Liu, Y. (2020). *Curr. Opin. Struct. Biol.* **60**, 142–149.

Table 3.1: Data collection and refinement statistics

	DARP3
Data Collection	
Beamline	NSLS-II 17-ID-2
Space group	I222
Resolution (Å)	3.81 (3.91-3.81)*
Unit cell dimensions: a,b,c (Å)	128.0, 195.6, 228.4
Unit cell angles: α,β,γ (°)	90, 90, 90
Measured reflections	191827 (12243)
Unique reflections	28155 (1980)
Overall completeness (%)	98.9 (96.7)
Overall redundancy	6.8 (6.2)
Overall R_{merge}	0.129 (2.05)
$CC_{1/2}$	99.9 (48.6)
Overall I/δ	11.1 (1.1)
Refinement	
R_{work} / R_{free}	0.188 / 0.225
RMSD bond length (Å)	0.003
RMSD angle (°)	0.6
Number of protein atoms**	10638
Number of water atoms	0
Number of other solvent atoms	1
Average B-factor of protein ($Å^2$)	190
Average B-factor of water ($Å^2$)	N/A
Average B-factor other solvent ($Å^2$)	159
PDB ID code	8V9O

Table 3.2: Structures of designed protein cages solved by X-ray crystallography

Protein Cage	Symmetry	Resolution (Å)	PDB accession code	Reference
DARP3 T33-51H	Tetrahedral	3.81	8V9O	This study
T33-51H	Tetrahedral	3.4	5CY5	(Cannon et al., 2020)
I52-32	Icosahedral	3.5	5IM4	(Bale et al., 2016)
I53-40	Icosahedral	3.7	5IM5	(Bale et al., 2016)
I32-28	Icosahedral	5.59	5IM6	(Bale et al., 2016)
13 nm cpPduA	Icosahedral	2.51	5HPN	(Jorda et al., 2016)
16 nm protein cage	Tetrahedral	4.19	4QES	(Lai et al., 2016)
Cube-shaped cage	Octahedral	7.08	4QCC	(Lai et al., 2014)
T32-28	Tetrahedral	4.50	4NWN	(King et al., 2014)
T33-15	Tetrahedral	2.80	4NWO	(King et al., 2014)

T33-21	Tetrahedral	2.10	4NWP	(King et al., 2014)
T33-28	Tetrahedral	3.50	4NWR	(King et al., 2014)
16 nm Cage	Tetrahedral	3.0	3VDX	(Lai et al., 2012)
T3-10	Tetrahedral	2.25	4EGG	(King et al., 2012)
O3-33	Octahedral	2.35	3VCD	(King et al., 2012)

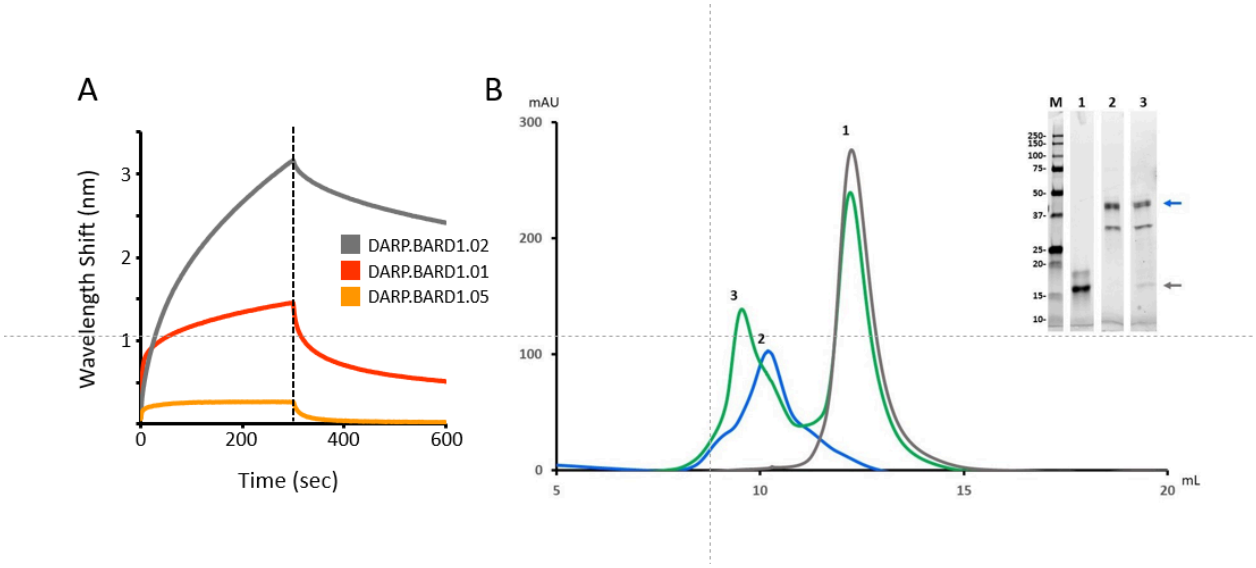


Figure 3.1: Verification of DARPin-BARD1 binding by biolayer interferometry (BLI) and analytical size exclusion chromatography (AnSEC). A, Putative anti-BARD1 DARPin molecules were screened for BARD1 binding by loading His-tagged DARPins on NTA biosensors and then incubating with BARD1 for five minutes and then analyte-free buffer for five minutes. Large wavelength shifts for DARP.BARD1.01 and DARP.BARD1.02 are indicative of strong antigen binding. B, The size exclusion profile shows a DARPin-BARD1 mixture (peak 3) has an altered retention time relative to BARD1 (peak 2) or the anti-BARD1 DARPin (DARP.BARD1.02; peak 1) alone. SDS-PAGE analysis (inset) of the peaks from AnSEC purification. Lane M, Broad-range molecular weight marker; lanes 1 to 3 correspond to peaks 1 to 3, respectively. Blue and gray arrows indicate the positions of BARD1 and DARP.BARD1.02 DARPin, respectively.

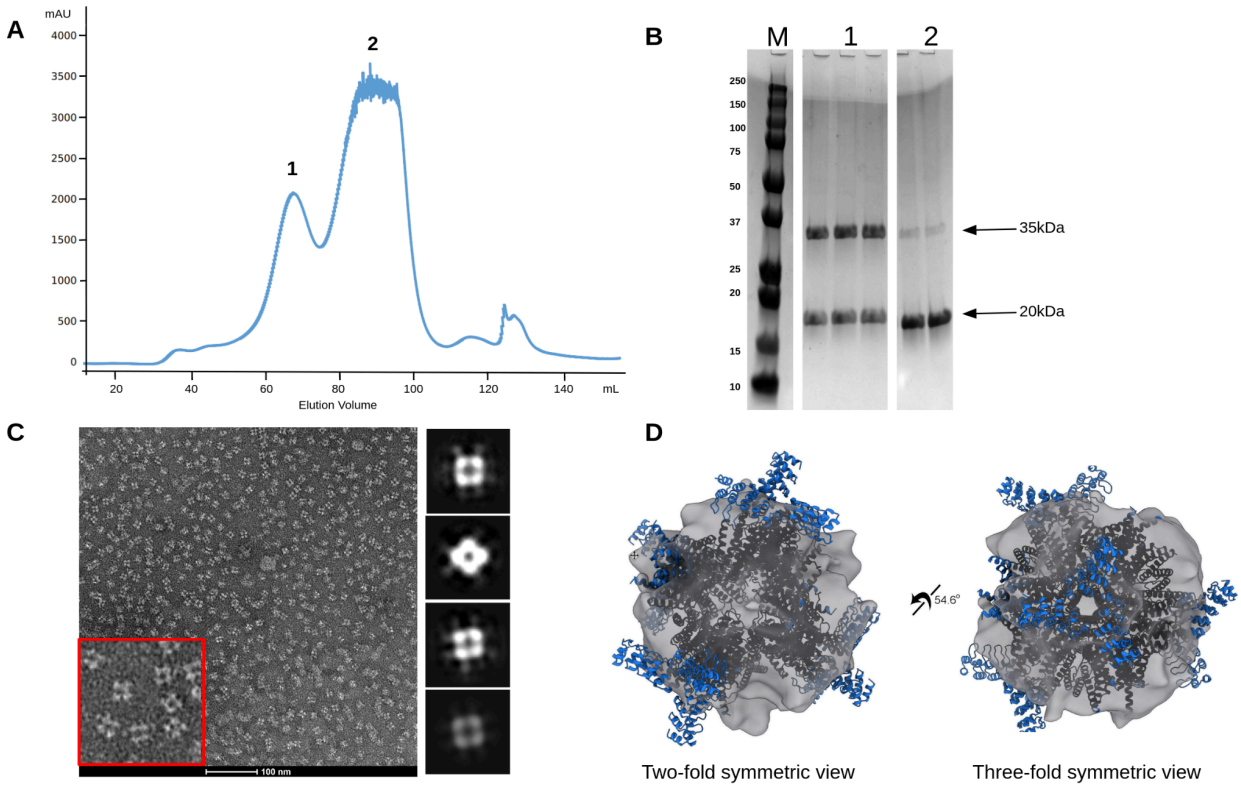


Figure 3.2: Validation of scaffold assembly. A, The size exclusion profile shows a peak corresponding to an assembled scaffold (peak 1), and is able to be separated from the unassembled or partially assembled cage components (peak 2). B, SDS-PAGE analysis of the peak fractions shows the presence of both scaffold components at their correct molecular weights, denoted by black arrows. C, Higher order assembly was verified by negative stain electron microscopy identifying particles of the proper size and symmetry. A blown up view of the micrograph (red box) shows particles with an estimated diameter of approximately 19 nm, matching the dimensions of the X-ray structure of the DARP3 scaffold (Figure 3; PDBid: 8V9O). Particles had a tendency for a preferred orientation along the 2-fold viewing axis. To the right of the micrograph are rough 2D averages processed from negative stain data. D, 2D classes were used to generate coarse *ab initio* 3D models.

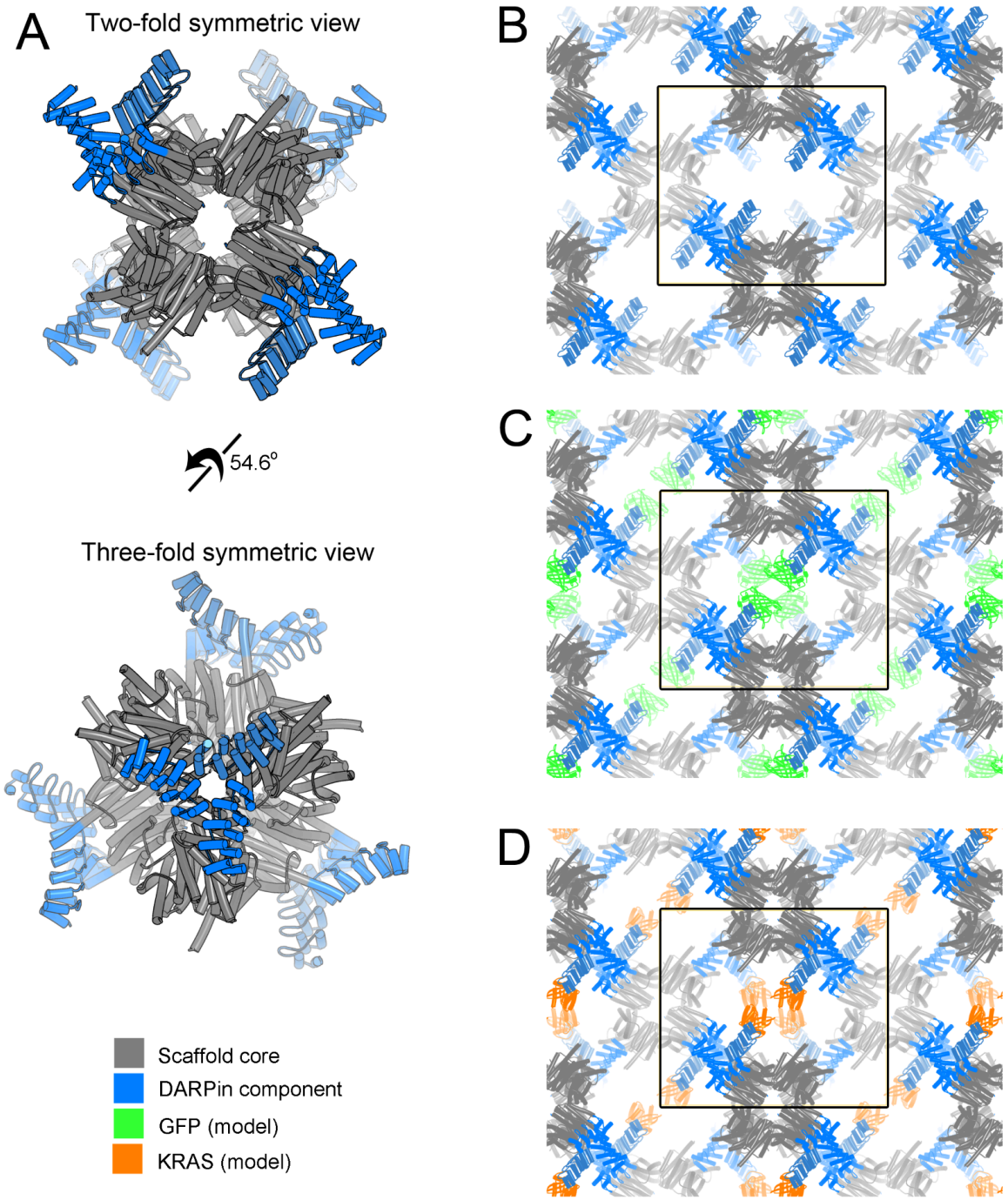


Figure 3.3: Structure of the DARP3 scaffold and its crystal packing. A, Views of the fully assembled 24-subunit DARP3 scaffold along the two-fold (top) and three-fold (bottom) axes of symmetry. The crystal structure closely resembles the structure of the KRAS-binding DARPin

scaffold (Castells-Graells *et al.*, 2023). The asymmetric unit consists of a trimer of the DARPin-fused component and the un-fused native cage component. B, Crystal packing of the DARP3 assembly. A large solvent channel is present between four copies of the DARP3 scaffold. One DARPin from each scaffold points into the cavity allowing for cargo binding at one of the available DARPins. C, A model illustrating that GFP molecules could theoretically fit without steric clash in the solvent channel when bound to one of the three DARPin modules. D, A model illustrating the same is true for KRAS. The black outline denotes the unit cell.

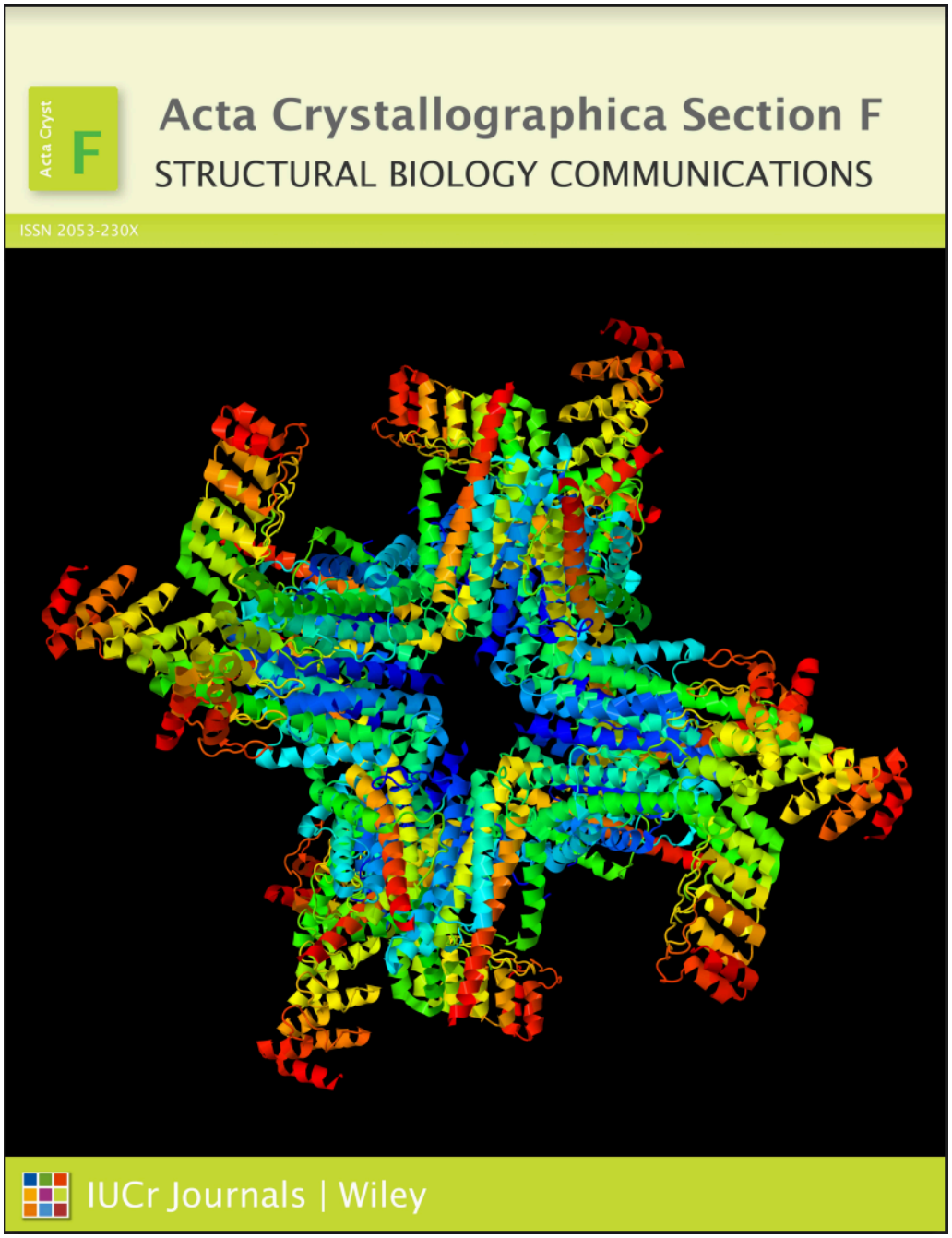


Figure 3.4: Cover photo for Acta Crystallographica Section F. The structure of our DARP3 scaffold made the journal cover. ISSN 2053-230X (2024).

CHAPTER FOUR: Design and Characterization of RNA Imaging Scaffolds Using Helical Fusion

4.1: Background and Significance

Explosion of novel RNA molecules

Although proteins and DNA have historically garnered most of the scientific community's attention, it has been known for some time that RNA also plays important roles in a variety of cellular processes. The best known examples involve gene expression and protein translation. In human cells, as a gene is expressed, the protein RNA Polymerase slides along the DNA building the corresponding chain of RNA in a process known as transcription. This pre-messenger RNA is then modified, by addition of a poly-adenine sequence to the 3' end and a 5'-cap to the transcript for stability, as well as potential slicing events of introns and exons to form the final mature mRNA[1]. This mRNA eventually will become protein as it meets up with two other well-known types of RNA; ribosomal RNA, which constitutes the ribosome along with a variety of protein components, and transfer RNA, which is responsible for bringing the specified amino acids into the growing polypeptide chain.

In humans alone it is estimated that 85% of the genome gets transcribed into RNA, but only 1.5% of it results in proteins[2,3]. In addition to the well-known types of RNA, in recent years there has been an explosion in the discovery of RNA molecules with interesting properties and functions once thought to be junk. An example of The idea of catalytic RNA challenged the central dogma of biology and eventually led to the discoverers, Thomas Cech and Sidney Altman, being awarded the 1989 Nobel Prize in Chemistry[4]. On the other spectrum of RNA size comes small RNA molecules that still pack a big biological punch. One of the most intriguing

new molecules discovered is a family of small non-coding RNA termed MicroRNA (miRNA). These RNA have been shown to regulate a wide array of biological processes, but are most famous in their regulation of gene expression by their binding to the transcripts 3'-Untranslated Region (UTR). This regulation is accomplished through a process involving a biomolecular superstructure known as the RNA-Induced Silencing Complex (RISC)[5,6]. Precursors known as pri-miRNA are processed by series of cleavage events by the proteins Dicer and Drosha to produce mature miRNA which is then loaded into the RISC complex to carry out its regulating functions, a schematic of which is shown in Figure 4.1[7]. What is even more important to human health is that dysregulated miRNAs have been shown to affect the hallmarks of cancer, including sustaining proliferative signaling, evading growth suppressors, resisting cell death, activating invasion and metastasis, and inducing angiogenesis. An increasing number of studies have identified miRNAs as potential biomarkers for human cancer diagnosis, prognosis and therapeutic targets or tools, which needs further investigation and validation[8].

It is commonly understood that when it comes to biological macromolecules, the structure of the molecule often dictates its function, a term colloquially known as the "Structure-Function Relationship". To better understand the function of many of these RNA molecules, researchers have long sought to better understand their structures. The pioneering work that sparked the RNA structural biology field was Rich et al.'s determination of the structure of the yeast tRNAphe to 3 Å resolution[9]. Using X-ray crystallography, this demonstrated the first time the 3D structure of an RNA molecule was solved to near-atomic resolution. Since that work more than 40 years ago, a little over 5700 structures containing RNA have been solved and deposited into the PDB, and of those, only ~1500 are protein-free[10,11]. Representing only 6% of all depositions, this highlights the work still to be undertaken in this field.

Difficulty studying RNA via traditional methods

For years structural studies on RNA have been primarily reserved to the use of X-ray crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy. Both techniques, while providing an excellent means of studying biomacromolecules, are not without their shortcomings. This is particularly the case with RNA compared to proteins and as evident by the total number of RNA structures deposited into the PDB accounting for only ~1% of the total PDB depositions, while the human genome codes for many more functional RNAs than proteins[12].

In terms of crystallography, we think of a crystal as an ordered array of molecules aligned in a lattice and held together by non-covalent interactions that we term as “crystal contacts”[13]. Unlike proteins which can utilize a large number of structural and chemical features on their surface, stemming from the 20 canonical amino acids, the density of negatively charged phosphate groups on the RNA backbone makes crystal packing extremely difficult[14,15]. This, coupled with significant flexibility from weak tertiary interactions, often results in loosely packed, poorly ordered crystals. These crystals diffract to low resolution which makes RNA structure determination extremely challenging[16]. Additionally, in most cases, in order to obtain crystals large enough to obtain good quality diffraction data, a wide range of precipitants and buffer additives are required. The addition of these chemicals may alter the native shape of the macromolecule and could lead to incorrect conclusions being drawn because of the generation of an incorrect structure.

The second pillar of structural biology techniques, NMR spectroscopy has troubles with RNA compared to proteins as well. Firstly it should be noted that unlike crystallography which gives a static representation of a molecule, NMR provides an ensemble structure with the accuracy of the ensemble being dependent on the number and type of constraints used[17]. Structural determination is more difficult for RNA molecules since they often form extended structures,

which yield only a limited number of long-range restraints compared to the more globular, compactly folded proteins. Furthermore, the number of restraints for RNAs is also often smaller than in proteins of similar molecular weight[18]. This is because RNA consists of only 4 monomeric units and its predominant secondary structural element is the A-form helix, which can become very difficult to discern chemical shifts from different bases that experience similar chemical environments[19]. This poor dispersion and substantial overlapping makes it difficult to obtain unambiguous resonance assignments in RNA especially in the ribose proton region, as displayed in Figure 4.2

There has been substantial efforts devoted to studying RNA molecules using cryo-EM, albeit to moderate, yet promising results. Some of the earliest breakthroughs date back to efforts of trying to understand the ribosome, the biological machinery responsible for protein synthesis, in atomic detail. The researchers generated a map to 4.5 Å of the small 30S ribosomal subunit from *Thermus thermophilus* using cryo-EM. They were able to unambiguously place previously solved protein segments of the complex into the density. Since then, efforts have expanded to smaller and smaller RNA molecules. The results have been modest to this point with only a handful of structures being solved to atomic-resolution. At the time of writing, the highest resolution RNA-only structure solved by cryo-EM is the *Tetrahymena* group I intron, which Liu and colleagues solved to sub-3 Å [20]. This rather larger RNA required substantial engineering efforts to create a cyclic, homomeric complex in order to stabilize the construct for high resolution collection. While this work demonstrated a substantial technological advancement, the amount of engineering required significantly limits its potential uses, especially when targeting some of these more interesting RNA molecules. Since then, researchers have been focusing on smaller and smaller target molecules. The smallest RNA molecule solved so far is a 40kDa SAM IV riboswitch by Wa Chiu's group at Stanford in 2019[21]. They were able to solve the entire structure directly by a combination of cryo-EM and computational modeling and

refinement programs designed for RNA structure building[22]. While this represents a target almost three times smaller than the previously discussed group I intron, it is far from the size of many RNA molecules implicated in diseases[23].

The Yeates has recently developed a method to circumvent the resolution limit problem plaguing the field of small-molecule cryo-EM and image single proteins using single particle cryo-EM by engineering a symmetrical protein scaffold onto which various imaging targets can be docked, which is discussed in length in chapter 2 of this dissertation. Taking advantage of the cages defined symmetry and rigidity, the first generation of this scaffold was used to solve the structure of a small 26kDa GFP protein at 3.8Å[24]. Further engineering improvements were able to push the resolution to 2.9Å[25], feats that were previously unattainable by cryo-EM. My mission for this project was to extend this technology by replacing the use of DARPins for RNA-binding proteins to studying the ever-increasing catalog of novel RNA molecules at atomic detail.

4.2: Results and Discussion

Design and Biochemical Characterization

Taking previous scaffolding attempts regarding proteins as a starting point, initial RNA-imaging scaffolds were generated between a tetrahedral nanocage[26,27],T33-21, and the RNA binding protein YbxF. The underlying cage core was chosen due to its demonstrated rigidity and effectiveness as a scaffold backbone. YbxF, an 8.3kDa protein implicated in streptomycin resistance, was chosen due to it having a few beneficial characteristics we sought to exploit[28]. Firstly, it contains an alpha-helical secondary structure in its N-terminal domain. This is critical as the design methodology employed relies on fusion between components by genetic extension between their terminal helices. Secondly, YbxF binds to a particular RNA motif, the

Kink-Turn, with a high degree of specificity and affinity, allowing for a universal motif to be engineered into targets of interest[29]. A diagram of the Kink-turn motif is shown in Figure 4.3[30,31]. Finally, YbxF has been demonstrated to be successful as a crystallization chaperone, enabling the structure determination of hard to crystallize RNA targets. Multiple alignments were performed between cage core and binding protein resulting, adjusting the alignment registry to tune the angle or display of YbxF. We used a co-crystal structure of YbxF complexed with a SAM I riboswitch for design creation. The large bound RNA aided us in curating designs as obvious clashing among cargo was obvious. Since the goal of these scaffolds will ve to eventually study small biological RNA, we figured that using the large riboswitch would provide an upper bound for the possible target cargo using these particular designs. Additionally, a large, bulky RNA with a known structure will greatly aid in downstream alignment and processing efforts. Manual generation of linker sequences was performed using information taken from literature about alpha-helical rigidity[32,33], resulting in 11 designs for experimental testing. A list of the designed sequences with linker amino acids can be found in Table 4.1. A consensus model of the T33-21-based designs is depicted in Figure 4.4. The scaffolds consist of a 24-subunit cage core, consisting of two trimeric proteins (hereby referred to as subunit A and B). Subunit A, a 18kDa of alpha-helical nature was chosen as the fusion subunit due to its terminal C-terminal helix resulting in a 27kDa RNA-binding subunit. Subunit B is a 14kDa protein that is composed of mostly beta-sheets and serves a structural role in the scaffold. Subunit B is the subunit that contains the his6 tag for protein purification. Tags were only placed on one component to enrich the pulldown in assembled complexes during purification. The entire assembly is 496kDa and approximately 160 Å in diameter in its apo state. Designs were ordered from Integrated DNA Technologies (IDT) and Twist Biosciences and cloned into pET22b(+) plasmids for protein expression.

Plasmids were transformed into *E. coli* BL21 Gold (DE3) cells for plasmid propagation and initial biochemical characterization. Small scale cultures of 50mL LB or TB supplemented with 100ug/mL of ampicillin. Cultures were grown at 37°C until OD_{600} reached 0.6 before induction. At this stage, growth and expression conditions were also varied. Cultures were grown in duplicate with one being induced with 1mM IPTG and expressed at 37°C for 4 hours and the other being induced with 0.5mM IPTG and expressed at 18°C overnight. Cultures were pelleted by centrifugation and aliquoted into eppendorf tubes for expression and solubility testing. A full list of buffers tested is shown in Table 4.2 Pellets were resuspended in 2mL of buffer and lysed via sonication. Lysates were clarified by centrifugation before incubating the supernatant with NiNTA resin. Proteins were eluted after a wash step and subjected to SDS-PAGE for expression and solubility. Representative gel images are shown in Figure 4.5.

After buffer screening, the design T33-21-YbxF was chosen for further characterization, as it showed the highest levels of expression among variants tested. In addition to the biochemical evidence, T33-21-YbxF did not require addition of amino acids to bridge the linker distance between cage core and RNA binding adaptor protein. Upon extensive characterization of T33-21-YbxF and other designs, the fusion subunit A frequently suffered from low expression and poor solubility, resulting in an imbalance in stoichiometry between subunits and thus an overall poor yield for purified scaffold, as evidences by SEC and EM data (Figure 4.6). This instability was a noted problem that marked previous scaffolding attempts with this cage as well. In the previous work, the challenge was overcome by brute force purifications of large volumes of cultures until enough scaffold could be homogeneously obtained from cryo-EM studies. Rather than simply outwork the underlying design problem, I set out to try to improve the T33-21 cage to make it a more robust scaffold. Various stabilization efforts were attempted to rescue the expression of fusion subunit A, some of which included PROSS[34] mutational landscape

exploration, circular permutation, and chaperone co-expression, none of which had any measurable effect (data not shown).

An interesting contaminant was always seen during purification. An unknown 75kDa protein always co-eluted with my scaffold, even at high imidazole concentrations. Additionally the contaminant appeared to be roughly the same size and shape as my particle as it also came out in the same fractions under SEC (Figure 4.7A). Negative stain analysis showed particles of ~16nm in diameter existing in clumps resembling dimers and trimers (Figure 4.7B). A literature search revealed that the contaminant was a polymyxin resistance protein named ArnA, with a recent cryo-EM structure available to compare against my micrographs (Figure 4.7C)[35,36]. This protein has long been shown to be a common *E. coli* protein contaminant during affinity chromatography owing to the presence of six histidine amino acids clustered together on the surface of the protein. This causes a high affinity of NiNTA columns, often eluting at imidazole concentrations in excess of 140mM imidazole[36]. Optimization of purification conditions allowed for successful elimination of the contaminant prior to the scaffold elution step during affinity purification (Figure 4.7D).

Around this time a new cage was published by Cannon et al. that described successful efforts to rescue the expression and solubility of a previously failed cage[37]. The manuscript outlined using Rosetta's improved HBNet protocol to redesign the interface between cage subunits by adding extended hydrogen bond networks to mimic interfaces seen in nature. Upon inspection of the new cage, termed T33-51, showed the presence of N-terminal helices on both the A and B subunit of the cage, potentially doubling the designability of the system. Another notable feature of this new cage is the similarity between subunits. While the previous T33-21 cage was composed of two quite distinct subunits (Subunit A PDB: 1WY1, Subunit B PDB: 3E6Q), this new cage is composed of two trimers that are homologs of one another, sharing 38% sequence

identity and having almost identical folds. (Subunit A PDB: 1WY1, Subunit B PDB: 1NOG). Alignments were generated using the same method and the initial round of designs, this time being performed on both subunits, resulting in 4 unique alignments between cage and RNA binder. Design of linker sequences based on the same criteria as the T33-21-based designs resulted in 17 sequences for biochemical characterization. The identity of the designs and properties of the alignments are listed in Tables 4.3 and 4.4. A design model showing views down the 2-fold and 3-fold axes of symmetry is shown in Figure 4.8. The second round of designs using the new cage core showed improved solubility and stability, indicated by the intense bands present in the SDS-PAGE gel, a tall, sharp peak under SEC, and abundance of monodisperse particles as observed by negative stain EM (Figure 4.9). The drastic increase in yield is especially evident from the negative stain micrographs (Fig. 4.9D), where taking an aliquot directly from the SEC peak resulted in a grid so packed with homogeneous particles it resembled a monolayer, with local regions of order as particles packed together. The most promising design, AA1.01, hereby referred to as T33-51-AA1-YbxF, or simply AA1, was chosen to move forward with for high resolution structural studies (Figure 4.9). In addition to AA1 being the most homogeneous and abundant design, similar to the most promising T33-21 designs, it was chosen due to the lack of additional linker amino acids to bridge the two domains. Whenever possible, I ruled in favor of designs that had the shortest bridging alpha helices because even with the best engineering efforts, a tremendous amount of flexibility is still inherent and would limit the final resolution of the target cargo, as observed with previous scaffolding attempts[24].

Electron Microscopy Characterization of RNA Scaffolds

Before moving forward with cryo-EM studies of the scaffold, a brief analysis was performed via negative stain. Size exclusion-purified particles were stained and a small data set was collected at the TF20 at CNSI. Micrographs were processed in the cryoSPARC[38] and RELION[39] software

suites to create low resolution models that were inspected by docking a model of the scaffold's core into the density (Figure 4.10). The resulting map was quoted at ~18 Å and was able to accommodate the entirety of the scaffold's core within it and distinct pores were identifiable corresponding to the axes of symmetry of the design. No density was resolved that could be attributed to the YbxF proteins, which wasn't totally surprising as they are expected to have some degree of flexibility and the resolution attainable from negative stain is inherently limiting. To verify proper genetic fusion had occurred between the RNA-binding protein and the cage subunit, mass spec analysis also was performed on SDS-PAGE bands from the size exclusion purification to ensure presence of all components (Figure 4.10C). The resulting map was quoted at ~18 Å and was able to accommodate the entirety of the scaffold's core within it, and distinct pores were identifiable corresponding to the axes of symmetry of the design.

After verification that we have successfully purified an assembled nanocage scaffold, I moved forward with cryo-EM analysis. The first stage of the cryo process involves optimization of a number of parameters that influence particle distribution and ice thickness. These parameters can be separated into two categories: sample-based conditions, and vitrobot settings. On the sample side, parameters that were optimized include such things as: sample concentration and purity, as well as were any additives added to aid in sample prep that may affect ice thickness. On the vitrobot side of freezing optimization comes a whole host of parameters that play an important role. Two of the most important factors for getting the best ice for high resolution data collection are blot time- the amount of time the filter papers are removing the excess buffer, and blot force- how much force is being applied to your grid by the filter paper paddles, optimization of these two parameters along could be a majority of the freezing condition process. Some minor factors on the vitrobot that also play important roles are the humidity and temperature of the interior blotting compartment. In addition to all the considerations above, optimizing the amount of sample applied to your grid as well as the manner in which your position the grid on

the tweezers can have noticeable effects on your ice conditions. When screening for ice conditions, you're looking for the right thickness of ice that allows your particles to be evenly distributed throughout, oriented in all directions, which allows for every angle to be captured and included in your 3D reconstruction. If the ice becomes too thin, then your particles might be excluded from the holes, or clustered along the edges. Thin ice may also exacerbate any preferred orientation bias your sample may have, as ice becomes too thin to accommodate certain orientations. Even more dramatic still, your ice may become so thin that portions of our protein may congregate at the air-water interface causing aggregation, unfolding and degrading the quality of your data. An example of the types of ice seen during optimization may be seen in the low magnification images in Figure 4.11. After identifying freezing conditions that would result in the best ice conditions and particle distribution, a small cryo dataset was collected at CNSI on the FEI TF20 in preparation for a high resolution imaging session on the Krios. A total of 75 cryo micrographs were collected at ~75,000x magnification and a defocus range of 1.0μm - -2.0μm, converted to .MRC format where they were first imported and CTF-corrected using RELION[38] and then transferred to cryoSPARC[39] for further processing and analysis. The result was a low-resolution map in which a model of the cage core could be docked (Figure 4.12).

Satisfied with the preliminary results we were seeing, we moved forward with high resolution data collection. Over the course of 2 days, almost 3500 movies were collected at the Titan Krios at CNSI at a magnification of 81,000x corresponding to a pixel size of 1.1 Å/pix on a Gatan K3 camera with a defocus range of -1.0μm to -2.2μm. Using the software Leginon[40], data collection could continue for the full 48 hours being monitored and altered remotely. An image of the remote data collection session, showing our particles and what holes in the grid the images derived from is shown in Figure 4.13. After the session was complete, the data was transferred to our file server for processing.

Cryo-EM Data Processing

After micrographs were successfully transferred to our local storage system, I began the task of trying to some the structure of my apo scaffold. Early on it became evident that the processing of this dataset may not be completely straightforward. The movies collected required minimal drift correction and passed CTF estimation and automated picking jobs without issue, with cryoSPARC estimating the CTF fit all the way out to 2.9 Å. The first complication I encountered stemmed from the drastic preferred orientation problem we were seeing in the micrographs. My scaffold, possibly due to the ice thickness, tended to like sitting on the grid such that you were looking at the particles down their two-fold axes of symmetry. Initial classification yielded only 2D classes containing these distinct, box-like views. Using these particles in both *ab initio* and template-based 3D reconstructions yielded low-resolution, featureless volumes akin to the results you would see in the era of “blobology” (Figure 4.14). The number of particles extracted from my session numbered over three million, so I was confident that even with my extreme preferred orientation problem, there should be at least a small subset of views of my particles in the various orientations somewhere in the dataset that I could uncover and use for further processing. Being more rigorous, I was able to pull out multiple views of my scaffold through iterative rounds of 2D classification and manual curation. Initially, I began asking cryoSPARC for a large number of classes, totalling over 250, so that every unique view could have its own classes, and then decreasing the number of classes as the classification jobs continued. During the first round of processing, I was only asking for the standard number of classes, usually between 50-100 that work fine for most projects, which was causing the under-represented views to be shoved into incorrect classes and lost (Figure 4.15). By asking for more classes at the beginning, I was able to weed out the bad particles and classes while still maintaining the necessary views for accurate 3D reconstructions.

After overcoming the first processing hurdle, I was able to continue until I generated a 3D volume of my scaffold which is where my second set of complications arose. Analysis of the resulting volume showed some red flags in our designs. After rounds of homogeneous and heterogeneous refinements I got a density map that looks roughly like my desired particle but was lacking any high resolution features (Figure 4.16). Secondary structural elements are present and alpha helices are identifiable but the overall structure looks “blown-out.” My hypothesis is that our issues are inherent to the design and the choices I made at the very beginning. When making these new T33-51-based designs, the fact that both subunits shared the same fold was taken as a benefit - it doubled the designable space because both subunits now had helical termini for fusion (Figure 4.17A). Now this similarity was stalling processing efforts because during the alignment and classification stages of the processing, the algorithms in both RELION and cryoSPARC were failing to distinguish between subunit A and B and mixing their orientations together (Figure 4.17B). We would hope that the fusion of YbxF onto subunit A would have caused enough of an asymmetry to be able to break this ambiguity but that was not the case. Unfortunately, it looks like the ~8kDa RNA-binding protein we had chosen for this pilot study was not large enough to distinguish between the fusion and non-fusion subunit. Additionally, we like to think of our alpha-helical extensions as a rigid fusion that holds the desired fusion partner in a fixed, predictable orientation, but we know from many studies now, that these helices have a tremendous amount of flexibility, with motions that are often described as “swaying” or “breathing”[24-26, 41-43]. It is very likely that compounding on top of our extremely small RNA-binding protein is also a considerate amount of motion going on. This movement of YbxF in relation to the cage is washing away any signal that might have been present and making high resolution structure determination impossible.

In a desperate search for information to help overcome my problem, I found a recently published paper that was a collaboration between Hong Zhou and Bill Gelbart at UCLA where

they were able to use a particle subtraction strategy to solve a low resolution structure of the RNA genome of a brome mosaic virus[45]. The researchers in that study subtracted the virus capsid signal from the raw cryo-EM micrographs and then used those subtracted micrographs for processing in which the software was able to latch onto the RNA signal and allowed the researchers to solve a blob-like structure, which unfortunately was determined to be disordered and lacked distinguishable features. We postulated that this technique might be applicable to our data in that if we could subtract the signal corresponding to the cage from the raw micrographs, the algorithms might be able to latch onto the remaining density, that should correspond to YbxF, and we would be able to perform and RNA-binding protein-focused asymmetric reconstruction and further processing on those features may be help us break our symmetry problem. With these ideas in mind we tried particle subtraction routines in RELION using a wide variety of masks, from the entire cage core and entire monomers and trimeric units, all the way down to various helices, in attempts to try to break the symmetry but our attempts yielded no useful results. In the cases where we masked out the cage, there was not enough signal remaining in the micrographs to result in meaningful averages that could be used. In the situation where we removed smaller regions of the scaffold, such as a monomer of subunit A and an entire adjacent trimer of subunit B, the resulting reconstructions turned out to be volumes that looked somewhat like our particle, probably due to the enforcement of T symmetry we were imposing, but with a lot more noise and worse resolution than when I was processing the unaltered micrographs (Figure 4.18).

A literature review of nucleic acid-binding proteins and their properties uncovered some interesting facts that helped shed some light on why our efforts in processing the T33-51-AA1-YbxF Krio data was becoming so problematic. It's known that many proteins that bind to nucleic acids as their functions exist in a disordered state and do not land on a distinct structure until interaction with their cognate ligands[46]. Many of these disordered proteins are

transcription factors that need to mediate the interaction between potentially many DNA sequences and their downstream responses. Outside of a handful of articles, including the co-crystal structure of YbxF with a segment of riboswitch RNA, there is not much known about our chosen RNA-binding protein. Analysis of the crystal structure doesn't provide any additional information to me as YbxF is expected to be in its stable folded state when crystallized with its target RNA. There just so happened to be a paper from 2014 on an archaeal homolog of YbxF, the ribosomal protein L7Ae that provided a possible explanation for the phenomena we were experiencing. In the study, researchers solved the structure of the ~13kDa L7Ae protein by solution-state NMR in both the Apo and bound states[47]. The bound state contained a 25nt long stretch of K-turn RNA. They found that a stretch of the protein is disordered but becomes an ordered alpha helix that makes contact with the RNA molecule after binding. My hypothesis about YbxF is that it also shares a similar degree of disorder to that of its homologue. Either all or a significant portion of YbxF is probably disordered and doesn't adopt the structure seen in the PDB until after it binds its appropriate RNA. So between the disorder inherent in YbxF itself, plus the flexibility introduced into our design by the choice of genetic helical fusion caused the only difference between subunit A and B, a small 8kDa protein, to have such a degree of motion that the symmetry between components of the cage could not be broken.

In Vitro Transcription of RNA Cargo

Despite our challenges with solving the apo structure, we thought it would still be worth pursuing cargo binding efforts in parallel to our processing efforts. The driving force behind this idea was the fact that RNA molecules contain a highly electron-dense phosphate backbone that interacts strongly with the electron beam, producing a large signal. Additionally, based on my findings from homologous RNA-binding proteins, the interaction of YbxF with its target RNA should cause the protein to leave its disordered state and land on a stable, folded state that we can solve. We postulated that this rigidification and the added signal from the RNA molecules may

provide the orientational information that our cage was lacking from YbxF alone, allowing us to solve the structure of not only the cage, but also the RNA cargo at the same time.

For the T33-51-AA1-YbxF design, the target RNA cargo was chosen to be an S-Adenosylmethionine riboswitch mRNA regulatory element (PDB: 2GIS). This particular piece of RNA was chosen for several reasons. Firstly, it has a known crystal structure, which will aid in downstream model building endeavors. Secondly, its structure is that of a riboswitch - a large, ~30kDa, 94 nucleotide, knotted piece of RNA whose features we believed would give us a visual marker when refining and interpreting the resulting density maps. Lastly, the most important reason why this RNA was chosen is that the structure of the molecule is a co-crystal structure of the RNA complexed with YbxF. A quick literature search has indicated this interaction is quite strong, being somewhere on the order of ~400nM[48.] This proven association will alleviate any complications stemming from first demonstrating an interaction between components that will be making up our scaffold. Figure 4.19 depicts the structure of the 2GIS riboswitch co-crystallized with the YbxF RNA-binding protein, as well as a model of how the RNA would be displayed on the surface of our imaging scaffold.

The first method I used to generate the RNA molecules for future binding studies was to take advantage of the self-cleaving ability of the HDV ribozyme. This ribozyme has been long used in the RNA biology field due to its reliability and demonstrated improvement on the target transcripts 3'-end homogeneity. A diagram of the process is shown in Figure 4.20 but briefly, a target construct is designed in the layout of T7 promoter - Target RNA sequence - HDV ribozyme and production of RNA is done in vitro by run-off transcription. When a polymerase is added to the reaction mixture, the polymerase begins transcribing a nascent RNA chain until it reaches the end of the transcript where it then falls off the template and is able to repeat the process. When this new RNA chain is released from the polymerase, it begins to fold. The HDV

ribozyme, in its folded state is able to cleave the transcript directly on the 5' end of the HDV ribozyme, resulting in 2 fragments of RNA - your desired product, and the HDV ribozyme. These RNA's can then be purified from each other leaving just your RNA construct for binding studies. With the help of Yan Li from the Guo lab, I began optimizing transcription conditions by varying parameters such the magnesium ion, polymerase and NTP concentrations to try to obtain the highest yield and most homogeneous RNA sample. A list of the transcription reaction conditions tested is shown in Table 4.5. For these initial tests, reactions were carried out in volumes of 50uL and incubated at 37 °C for 4 hours. Samples were collected and quenched after 2 hours for analysis on denaturing acrylamide gels. The next stage after transcription is the separation of products by tube gel purification.

Unfortunately, during my experiments I noticed that this HDV-assisted tube gel method of RNA production was causing a lot of contamination and inconsistency in my purification results. My collaborator, Feng Guo, suggested switching to an alternative method for RNA cargo production that they favored in their lab. Instead of using a HDV ribozyme, this second method utilizes a biotinylated forward 5' and 3' 2-O-methoxy reverse primers. The biotinylated forward primer is used for later removal of the DNA template, whereas the methylation of the 3' reverse primer has been shown to increase product stability by increasing nuclease resistance[49],as well as increase product homogeneity by improving run-off efficiency. After proper optimization of transcription conditions using this new method, in vitro transcription reactions were scaled up to 10mL volume to increase RNA yields. A comprehensive outline of the purification procedure can be found in the materials and methods section. After transcription, the target 2GIS was purified by ion exchange chromatography over a linear gradient. Resulting fractions were analyzed by denaturing PAGE and fractions containing the 2GIS riboswitch were pooled, buffer exchanged, and concentrated. Figure 4.21 shows a workflow of the general purification process with representative HiTrap purification run images as well as the corresponding denaturing PAGE

gel. Following proper purification of my target 2GIS RNA, refolding conditions were optimized with the folded state being tracked by native PAGE. The proper fold of our riboswitch is essential for proper function, but more importantly for our binding studies, the three-dimensional K-turn motif needs to be folded and accessible by YbxF in order for binding to occur. To ensure the proper structure of 2GIS, a variety of refolding procedures from snap cooling to slow cooling of the RNA after melting, were tested. Table 4.6 outlines the various refolding parameters tested. The decision to include the purification buffers was made to ensure that both our scaffold and RNA would be happy during subsequent binding and chromatography experiments. Of the refolding methods tested, there was no noticeable difference as assessed by native PAGE (Figure 4.22), so a default starting point in the binding studies the refolding procedure used was the same one used in the crystal structure paper and involve heating the dilute RNA sample at 85°C for 2 minutes before slow cooling on the benchtop for 8 minutes to begin the folding process. After the initial refolding had begun, the sample was supplemented with a final concentration of 17.9mM $MgCl_2$ and allowed to continue folding for 15 minutes at 37°C. Addition of SAM was forgone as the stock I had access to was fairly impure when analyzed by gas chromatography and discussions from the 2GIS's crystal paper states that addition of ligand showed no noticeable change in the RNA's structure.

Binding Assays of RNA Imaging Scaffold and RNA Cargo

After properly producing my target 2GIS RNA riboswitch and refolding it to its final three-dimensional structure, it was time to proceed with binding experiments. Before continuing, I first wanted to see where the 2GIS RNA alone eluted on our Superose 6 Increase SEC column. I observed that even after refolding, the 2GIS riboswitch appeared to exist in two distinct populations, as evidenced by discrete peak at around 17mL in elution volume, and an adjacent shoulder peak when looking at the SEC chromatogram trace (Figure 4.23A). This peak doublet was a cause for concern, especially when both peaks looked identical under native

PAGE (Figure 4.23B). My only explanation for this phenomenon is that the shoulder coming out earlier is most likely a folding intermediate that hadn't come to completion or got stuck in a local minima during my refolding. If the RNA is partially unfolded, one would expect it to be in a less compact structure that in its full folded state and since size exclusion chromatography actually separates based on radial diameter and not purely mass, then the elongated species would come out earlier in the elution profile, which is exactly what I observe. An explanation as to why the gel bands look so similar could be that the second peak is much more intense than the shoulder and the thicker band could be hiding subtle differences that might be observed if the sample was diluted before running on the gel. Binding studies were conducted using the discrete, more intense peak as I would expect the fully folded RNA to exist in a single smaller species for the reasons described above. As a control, the apo T33-51-AA1-YbxF scaffold were also run on the same column (Figure 4.23C) and a clean peak was observed around the expected elution volume of ~14mL and verified to contain both scaffold components by SDS-PAGE (Figure 4.23D).

In my initial binding tests, I opted for RNA ratios that were two to three times higher than that of the scaffold to ensure full occupancy. Because each one of our scaffolds contains 12 binding sites for the target RNA, when calculating the concentrations for mixing that adjusted stoichiometry must be taken into account. For my calculations, I treated one chain from the subunit A-YbxF fusion trimer and one chain from the subunit B trimer as a single unit. That way, I can do my mixing experiments on a per binding site basis instead of the full cages molarity. Unfortunately, doing so led to a signal from the RNA that completely drowned out any signal that was coming from the scaffold (Figure 4.24A). This is because, although the maximum absorbance for nucleic acids exists at a wavelength of 260nm, they exhibit a substantial amount of absorbance at 280nm as well. The A280 signal from the RNA was still so strong as to obfuscate any information about the scaffold or complex. To remedy this, I attempted to lower

the RNA concentration, first starting at a stoichiometry of 1:1 RNA:AA1, but even lowering it to as low as 0.25:1. At higher stoichiometries, the drowning effect from the RNA was still significant, but even at lower RNA concentrations, no interaction was observed as would be indicated by a clear shifting in the scaffold peak and reduction in the free-RNA peak (Figure 4.24B-D). Slight bumps in the SEC trace were observed for higher molecular weight species, elution around 12mL, but only showed to be aggregates of the T33-51-AA1-YbxF scaffold when analyzed by SDS-PAGE and negative stain EM and no RNA signal was observed when stained with RNA-specific dyes (Figure 4.24E,F).

In an attempt to try to remove any excess unbound RNA, after incubation, I began spinning the cage-cargo mixture briefly in a 100kDa Amicon concentrator at low speeds. Use of such a large molecular weight cutoff was chosen to retain the large, almost 1MDa complex if fully occupied, and allowed the small 30kDa RNA and 20-25kDa cage components to flow through. Unfortunately, even after this modification to my protocol, I was still observing large peaks corresponding to free RNA on my size exclusion chromatograms (Figure 4.25A). After the spin step, it was noticed that the small aggregation peak was more pronounced than in previous experiments, perhaps because of the removal of some free RNA (Figure 4.25A,B). When these fractions were analyzed under native PAGE, no evidence of RNA signal was seen (Figure 4.25C,D).

A great deal of time and effort was exhausted in trying various methods and parameters for my binding experiments. From changing the refolding conditions, to the addition of magnesium into my SEC buffers to prevent dissociation of cargo from scaffold during SEC, all manner of suggestions was tried to no avail. At times I convinced myself that I saw slight evidence of binding but upon discussions with my collaborators in the Guo lab, discovered I was misled; the A260 absorbance from RNA molecules is so intense that even a small degree of association

would result in a noticeable shift in the absorbance peak - but in all my experiments, I still only saw peaks for free RNA. In addition to my gel filtration-based binding readouts, I also experimented with isothermal titration calorimetry (ITC) and bilayer interferometry, which yielded either negative or inconclusive results. Additionally, rough negative stain and cryo-EM analysis was preformed on fractions from various mixing experiments, none of which was able to resolve either the fused YbxF, or the bound 2GIS RNA. After much heartache, I was forced to come to the conclusion that this combination of T33-51-AA1-YbxF scaffold and 2GIS RNA were not able to associate.

Concluding Remarks

Although ultimately this project did not reach the intended results, a lot of work has gone into the design, execution and analysis of the experiments presented here in this chapter. This section is meant to try to address some of the lessons we have learned during our endeavor.

Firstly, my choice in using the T33-51 cage as the scaffold core ended up complicating the project more than it helped. Originally, the switch in cores was necessary as the expression and solubility levels of subunit B in the T33-21 cage core made further use in my designs impossible. While this cage is much more robust, and in fact has been adopted by many lab members in their projects and has become a real workhorse in some of our protein design efforts[25, 50-52]. When comparing our processing results to that from Liu et al's T33-21-based scaffold, the utilization of two distinctly-structured components allowed for unambiguous identification of proper orientations. Additionally, DARPins are roughly twice the mass of YbxF and thus provided added signal to help with symmetry breaking. Castells-Graells et al utilized the same cage core as I, but the larger DARPin's plus stabilizing mutations introduced allowed them to distinguish between subunits A and B. If I were to restart this project again, I would

either try the original T33-21 cage core again and brute force the yield problem by growing vast amounts of bacteria for protein production, or start with a different cage core completely. At the time of project initiation, there were a limited number of designed tetrahedral cages to choose from, but since then the library has expanded. Additionally, work in the Yeates lab has resulted in new protein design algorithms intended to generate natural interfaces has shown tremendous promise[53]. Future design efforts to generate imaging scaffolds can take advantage of these new cages to expand the geometric library of potential designs. Careful design considerations must be taken into account when expanding designs into larger symmetries, as more components can cause confounding effects when it comes to biochemical characterization, and potential steric clashes must be investigated thoroughly.

Next, our choice in RNA-binding protein additionally contributed to the complications faced in this project. We were first suggested this protein by our collaborators because it has been shown to bind the K-turn motifs[29], and has potential to be used as crystallization chaperones in the same manner other RNA binding proteins have been[54]. As shown in this chapter, as well as work done on this protein in the Guo lab has failed to show adequate binding. On top of this, the fact that it was such a small protein (~8kDa) complicated our efforts in breaking the symmetry of the cage. Literature from homologous proteins have shown that several alpha helices are disordered until binding of its cognate RNA. It is possible that the fusion between YbxF and our scaffold core has inhibited YbxF's ability to either bind RNA or restricts its motion such that rearrangement of key structural components becomes impossible. If future efforts are to continue, a wider search of potential RNA-binding proteins will be required to increase our odds of success.

In terms of processing efforts, we exhausted all our knowledge and resources when trying to break the symmetry of the Krios dataset. Since these initial efforts, there has been substantial

improvement in the cryo-EM processing programs; cryoSPARC alone has gone through two entire version updates since we initially began processing this data. A lot of the improvements are focused around heterogeneity and flexibility in our particle data. It is very possible that if we spent the time to reprocess this data with the new resources we might be able to break the asymmetry in our cage design. However, due to the nature of helical fusions, the attachment between scaffold core and YbxF might be too flexible to enable structure determination, as the motion of YbxF causes any signal to be averaged away.

Finally, while advances in cryo-EM have provided great strides in breaking resolution boundaries once thought inaccessible, the need to imaging scaffolds is not going away. Although in recent years, microscopists have taken on the challenge of solving RNA via cryo-EM, the current results are mixed and only achieving moderate resolutions with large RNA[55,56]. While these will improve with further advancements, there will be many biologically relevant RNA that will never be able to be imaged directly because they are far below the resolution limit and in those cases, I believe imaging scaffolds will shine[57].

4.3: Materials and Methods

Design and Sequence Generation of RNA Imaging Scaffolds

For both T33-21 and T33-51-based designs, alignments were performed using in-house alignment scripts in python. A 10-residue window in the C-terminal helix of the cage core was aligned to a 10-residue window on the N-terminal side of an idealized alpha-helix composed of 25 alanine residues. A 10-residue window of the N-terminus of YbxF was aligned to the C-terminal 10 residues of the idealized helix. By sliding the window of the YbxF alignment, fine tunes the orientation of the display for the RNA-binding protein. Alignments were performed at every register along the idealized helix and outputs were manually analyzed to avoid clashing between cage and adapter, but also between symmetry-related copies of each subunit. Manual

addition of some bridging linker sequences was done according to published literature about the rigidity and stability of helices[32,33]. A hexahistidine (His6) tag was added only to subunit B. This is intended to pulldown only full-assembled complexes. Genes for both subunits were combined into one multicistronic construct with both subunits separated by an intergenic region derived from the pETDUET-1 vector as described previously[58]. Flanking sequences containing HindIII and NdeI restriction sites were added to each side of the construct to aid in cloning. Codon-optimized genes were synthesized and delivered by Integrated DNA Technologies or Twist Biosciences and cloned into pET22b(+) vectors via Gibson assembly[59]. Linker mutations genes were either ordered directly from or point mutations were generated by quick change and blunt-end ligation PCR reactions.

Expression and Characterization of Ordered Designs

Expression plasmids were transformed into BL21 Gold(DE3) *E. coli* strains for both plasmid propagation and protein production. Initial designs were tested for expression and solubility in a small scale buffer screen. Cells were grown in either LB or TB media supplemented with 100µg/mL of ampicillin at 37 °C in a shaker for 4 hours before being induced with 1mM isopropyl-thio-β-D-galactopyranoside (IPTG) and allowed to express for 4 hours, or induced with 0.5mM IPTG and incubated overnight at 18°C. Cells were pelleted in eppendorf tubes and resuspended in the lysis buffers listed in Table 4.2. Resuspended cells were lysed by sonication and clarified by centrifugation at 15kxg for 10 minutes. Clarified lysate was decanted and incubated with 50uL NiNTA beads. Pulldown purification was performed by one round of pelleting by centrifugation and washing with lysis buffer followed by elution with 100uL elution buffer listed in Table 4.2. Expression and solubility was assessed by SDS-PAGE and negative stain electron microscopy.

Successful designs were scaled up in 1 liter flasks of LB or TB medium supplemented with 100 mgμg/mL of ampicillin at 37 °C until an OD_{600} of 0.5-0.6 for LB or 0.8-0.9 for TB was reached. Protein production was induced by addition of 0.5mM IPTG and allowed to proceed overnight (~16 hours) at 18 °C before cells were harvested by centrifugation. Cell pellets were resuspended in an affinity buffer containing 50mM Tris 8.0, 250mM NaCl, and 20mM imidazole at a ratio of 3-4mL buffer per gram of cell pellet. Buffers were either supplemented with 1% glycerol, 1mM DTT or both during purification of various constructs. The resuspended cell pellet was lysed by 2-3 passages through an Elmusiflex until cell disruption was complete. Proteins were purified either by gravity column chromatography using NiNTA-conjugated resin or using a 5mL GE HiTrap and eluted with a linear gradient. Optimization of gravity elution discussed in Figure 4.7 consisted of performing stepwise elutions of imidazole concentrations of 20mM, 62.66mM, 84mM, and 148mM, before elution of 500mM. Contamination of the purification with ArnA was not as prevalent with the HiTrap due to the linear gradient allowing for better separation of scaffold and contaminant.

Mass Spectrometry Analysis of SDS-PAGE bands

Purified protein was run on an any Kd SDS-PAGE gel (BioRad) and stained with coomassie brilliant blue for visualization of protein bands. Target bands were excised with a clean razor blade and placed into a sterile eppendorf tube under flame to minimize contamination. Excised bands were delivered to Janine Fu in the lab of Joseph Loo at UCLA where they were de-stained and recovered from the gel and subjected to digestion by the protease trypsin. Bottom-up LC-MS was performed on the trypsin-digested protein samples. The resulting spectra was collected and analyzed by Janine before providing me the data. These data included information such as sequence identity, number of peptides detected in the spectra, and the overlap of the identified peptides to target sequence

Negative Stain Electron Microscopy

5uL of 0.05mg/mL purified cages were deposited on a formvar supported carbon film on 300-mesh copper grid that has been negatively glow discharged for 30secs. The excessive sample was blotted away with filter paper after 1 minute, washed twice with nanopure water and stained with 2% uranyl acetate for 30 sec. Grids were allowed to air dry before being imaged at room temperature with FEI Tecnai T12, FEI Tecnai TF20 and Talos F200C electron microscopes.

Negative Stain Data Processing

Micrographs were converted from .tif to .mrc format and imported into cryoSPARC[60] for processing. CTF estimation was performed using patch CTF. A small subset of ~150 particles was manually picked and averaged to create an initial model used for automated particle picking. Successive rounds of 2D averaging on extracted particles reduced the particle count from 104,000 to 24,000. *Ab initio* models were created on the final 2D classes in both C1 and T symmetry followed by homogeneous refinement in both C1 and T. Final refinement maps were analyzed in UCSF Chimera[61].

Cryo-Electron Microscopy

3uL of 1.0mg/mL purified cages were deposited on Quantifoil 1.2/1.3 300mesh copper grids (Ted Pella Inc) that had been negatively glow discharged for 30secs. Freezing was performed by blotting for 3 seconds at -15N force using a Vitrobot Mark IV (ThermoFischer Scientific) at 4°C and 100% humidity. Grids were transferred into dewars before ice screening using FEI TF20 and high resolution data collection with a Titan Krios at 300KeV equipped with a Gatan K3 camera.

Cryo-EM Data Processing

For the small TF20 dataset, the processing is as follows: A total of 75 micrographs were first collected at varying defocus values and then batch converted into .MRC format using in-house scripts. The micrographs were first imported into RELION3.0.8 for CTF estimation. The resulting output files were fed into cryoSPARC where particles were picked and extracted using a template derived from ~200 manually picked particles. Iterative rounds of 2D classification resulted in ~15,000 particles encompassing multiple orientations. These particles were used to create an *ab initio* 3D volume, created using C1 symmetry and later refined using homogeneous refinement with T symmetry enforced. The resulting map was viewed in UCSF Chimera[61] where a model of the T33-51 cage core was docked and analyzed.

For Krios dataset, the processing is as follows: A total of 3501 movies were collected at a range of defocus values over 2 days on the Titan Krios in CNSI at dose rate of 50.122 electrons/movie. Both all processing steps, a combination of RELION3.0.8 and cryoSPARC v3.0 was implemented as the procedure deemed fit. Movies were motion corrected using USCF's MotionCor2[62] and showed little drift over the course of each exposure. Patch CTF estimation algorithms were used to successfully calculate a CTF estimation of between 2.9 Å and 3.5 Å for the dataset. Initial particle picking was carried out by manually picking ~350 particles, and using the resulting model for template-guided autopicking protocols. Final particle picking to achieve more uniform particle views was accomplished by using the previously published DARP14 anti-GFP scaffold (PDB: 6C9K) as a template as it lacked the problematic symmetry our cage has. A box size of 168 Å was used for final particle extraction. Initial 2D classification jobs of 50 classes showed an extreme orientation bias of our particle along the two-fold axis of symmetry. Iterative 2D classification jobs starting asking for 250+ classes were performed and the number of classes was gradually reduced until classes representing a diverse set of particle views were obtained. Subsequent *ab initio* 3D reconstruction and 3D refinement (both heterogeneous and

homogeneous) jobs were run with both T symmetry enforced and not enforced but resulting maps were unable to achieve high resolution. Various masks were created in attempts to give the processing programs a template in which breaking the particle's symmetry might be possible. The masks used included but are not limited to fusion subunit A, the entire A trimeric unit, combinations of subunit B chains and subunit A chains, removal of entire secondary structural elements from the cage core, and the entire YbxF binding protein; mask optimization was also carried out on a wide range of map resolutions generated in UCSF Chimera. All particle subtraction jobs were performed in RELION3.0.8 following the steps outlined in the materials and methods of Beren et al. and the results were analyzed in UCSF Chimera. For subtraction, the general command: *relion project --i [density to remove.mrc] --subtract_exp --angpix [pixel value] --ctf --ang [particle star file.star] --o [subtracted particle micrographs.mrc]* was used.

In-Vitro RNA Synthesis

Genes encoding the SAM-I riboswitch (PDB: 2GIS) were designed on Benchling by attaching a T7 RNA polymerase polymerase sequence to the start of 2GIS's coding region and appended with an HDV ribozyme sequence on its 3' end. Flanking overlap regions containing HindIII and NdeI restriction sites were added to each side of the construct to aid in cloning. Genes were ordered by Twist Biosciences and cloned into pET22b by Gibson assembly. Plasmids were linearized on their 3' end by incubation with NdeI for 4 hours at 37°C to enable run-off transcription. For later transcription methods, biotinylated forward (5'-/5Biosg/TAATACGACTCACTATAGGCTTATCAAGAGA-3')and 2'-methoxy reverse (5'-mTmGGCTCATCTTTCAACGTTTCCGC-3') primers complementary to the T7 promoter and 2GIS coding sequence were ordered from IDT and used to PCR amplify the DNA template. Following PCR, DNA templates were purified by anion exchange chromatography (AEX) on a 5mL HiTrap Q column (Sigma-Aldrich) at a flow rate of 5mL/min and a gradient with elution

occurring from 20% and 100% Buffer B. HiTrap Buffer A is composed of 10mM Tris pH 7.5, 10mM NaCl; HiTrap Buffer B is composed of 10mM Tris pH 7.5, 2M NaCl. Fractions were analyzed on an agarose gel supplemented with syber SAFE DNA gel (ThermoFischer Scientific), pooled, concentrated, and stored for in vitro transcription reactions.

In vitro RNA transcription was performed in 10mL reaction volumes containing 1mL of 10X transcription buffer (400mM Tris 7.5, 250mM $MgCl_2$, 40mM DTT, 20mM spermidine), 1.5mL 20mM NTP mixture, 700uL T7 RNA Polymerase (6.4mg/mL), 0.5mL 1M $MgCl_2$, and 200pmol DNA template and nanopure sterile water. Reactions were allowed to proceed for 4 hours shaking at 110 rpm in a 37°C shaker. After 4 hours, further purification was performed or the reaction was quenched and saved for future use by storing samples in -20°C freezer. Removal of DNA template was accomplished by addition of 6µL of streptavidin-conjugated agarose beads (ThermoFischer) and allowed to incubate at room temperature for 15 minutes on a plate rotator. Beads were pelleted by centrifugation and the supernatant containing purified RNA was decanted and saved for further purification. Volume was either concentrated down to 5mL for HiTrap purification or multiple rounds of 5mL purifications were performed.

RNA transcription product was injected into a 5mL HiTrap Q (Sigma-Aldrich) for anion exchange purification and eluted using the same gradient protocol as used for purifying the prior DNA template. Fractions matching A260 peaks were assessed for yield and purity by denaturing PAGE and fractions containing 2GIS were pooled, buffer exchanged into 10mM HEPES 7.0 and concentrated with a 10kDa Amicon tube (MilliporeSigma), aliquoted into smaller fractions and stored at 4°C for immediate use or -20°C for long-term storage.

RNA-Scaffold Binding Studies

RNA was first refolded by various methods outlined in Table 4.6. For all binding experiments after initial refolding tests, one method was utilized. RNA was diluted to 1μM in a refolding buffer (26mM HEPES pH 7.5, 53mM KCl) and melted for 2 minutes at 85℃. The solution was then removed from the heating block and allowed to slow cool for 8 minutes at room temperature before being supplemented with 5mM $MgCl_2$, concentrated and stored at -20°C until ready for binding experiments. To first begin binding experiments, an accurate mixing stoichiometry must first be calculated. Because there are 12 copies of the RNA-binding protein per scaffold, concentrations and component mixing was done on a per-binding-site basis by combining one chain from subunit A with one chain of subunit B and using their molecular weight and extinction coefficients to assess concentrations. Initial binding experiments were conducted using 2:1 molar excess of RNA cargo to cage, but were reduced all the way down to 0.25:1 in subsequent experiments. Both AA1 scaffold and 2GIS RNA were diluted to corresponding concentrations, mixed at equal volumes and left on ice for 30 minutes. After incubation, excess RNA was removed by centrifugation at 14,000xg for 5 minutes using a 100kDa cutoff Amicon concentrator tube (Millipore). The resulting mixture was injected onto a Superose 6 Increase (Cytiva) and eluted with SEC buffer (50mM Tris pH 8.0, 150mM NaCl) either with or without addition of 5mM $MgCl_2$. Subsequent fractions were analyzed by native PAGE and stained with either a protein-specific dye (coomassie brilliant blue) or various RNA specific dyes (xylene cyanol-bromophenol blue, toluidine blue[63]).

4.4: References

[1] Manning, K., and Cooper, T. The role of RNA processing in translating genotype to phenotype. *Nature Rev. Mol. Cell* **18**, 102-114 (2017)

[2] Djebali, S. et al. *Landscape of transcription in human cells.* Nature. **489:** 101-108 (2012)

[3] The ENCODE Project Consortium. *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature **447:** 799–816 (2007)

[4] Walter, N., and Engelke, D. Ribozymes: Catalytic RNAs that cut things, make things, and do odd and useful jobs. *Biologist (London, England)* **49**(5): 199-203 (2002)

[5] MacFarlane L., and Murphy, P., MicroRNA: Biogenesis, Function and Role in Cancer. *Current Genomics* **11**: 537-561 (2010)

[6] Shukla, G., Singh, J., and Barik, S. MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Mol. Cell Pharmacol.* **3**(3): 83-92 (2011)

[7] Ha, M., and Kim, V.N. Regulation of microRNA biogenesis. *Nature Reviews* **15:** 509-524 (2014)

[8] Peng, Y., and Croce, C. The role of MicroRNAs in human cancer. *Nature Signal Transduction* **1**: 1-9 (2016)

[9] Thornton, J., Zimmerman, K., Templeton, M. N., Howard, D. L., & Rich, A. *Structure of yeast phenylalanine transfer RNA at 3 A resolution.* Nature. **274(5669):** 168-174 (1978)

[10] Adamczyk, B., Antczak, M., Szachniuk, M. *RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. Bioinformatics.* **38(14):** 3668–3670 (2022)

[11] Ma, H., Jia, X., Zhang, K., Su, Z. *Cryo-EM advances in RNA structure determination.* Sig. Transduct. Target. Ther. **7:** 58 (2022)

[12] Barnwal, R.P., Yang, F., Varani, G. Applications of NMR to structural determination of RNAs large and small. *Arch. Biochem. Biophys*. **628**:42–56 (2017)

[13] Capitani, G., Duarte, J., Baskaran, K., Bliven, S., and Somody, J. Understanding the fabric of protein crystals: computational of biological interfaces and crystal contacts. *Bioinformatics* **34**(4): 481-489 (2016)

[14] Ke, A., and Doudna, J.A. Crystallization of RNA and RNA-Protein Complexes. *Methods* **34**, 408-414 (2004)

[15] Reyes, F., Garst, A., and Bately, R. Chapter 6 - Strategies in RNA Crystallography. *Methods in Enzymology* **496**: 119-139 (2009)

[16] Ferré-D'Amaré, A.R., Zhou, K., Doudna, J.A. A General Module for RNA Crystallization. J. Mol. Biol. **279:** 621-631 (1998)

[17] Allain, F.H-., and Varani, G. How Accurately and Precisely Can RNA Structure be Determined by NMR? *J. Mol. Biol.* **267:** 338-351 (1997)

[18] Lukavsky, P, and Puglisi, J.D. Structure Determination of Large Biological RNAs *Methods in Enzymology* **394:** 399-416 (2005)

[19] Dance, HE., Stonehouse, NJ., and Bingham, RJ. *"NMR Methods for Studying RNA Dynamics and Structure."* Progress in Nuclear Magnetic Resonance Spectroscopy. **(95):** 31-51 (2016)

[20] Liu, D., Thélot, FA., Piccirilli, JA., Liao, M., Yin, P. *Sub-3-Å cryo-EM structure of RNA enabled by engineered homomeric self-assembly.* Nat Methods **19:** 576–585 (2022)

[21] Zhang, K., Li, S., Kappel, K., Pintilie, G., Su, Z., Mou, TC., Schmid, MF., Das, R., Chiu, W. *Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution.* Nat. Commun. **10:** 5511 (2019)

[22] Kappel, K., Zhang, K., Su, Z., Watkins, AM., Kladwang, W., Li, S., Pintilie, G., Topkar, VV., Rangan, R., Zheludev, IN., Yesselman, JD., Chiu, W, Das, R. *Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures.* Nat. Methods. **17(7):** 699-707 (2020)

[23] Xiong, Q., Zhang, Y., Li, J., Zhu, Q. *Small Non-Coding RNAs in Human Cancer.* Genes. **13(11):** 2072 (2022)

[24] Liu, Y., Huyng, D., Yeates, TO. *A 3.8*Å *resolution cryo-EM structure of a small protein bound to an imaging scaffold.* Nature Comm. **10(1):** 1864 (2019)

[25] Castells-Graells R., Meador K., Arbing MA., Sawaya MR., Gee M., Cascio D., Gleave E., Debreczeni JÉ., Breed J., Leopold K., Patel A., Jahagirdar D., Lyons B., Subramaniam S., Phillips C., Yeates TO. *Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold.* Proc Natl Acad Sci USA. **120(37)** (2023)

[26] Liu Y, Gonen S, Gonen T, and Yeates TO Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. PNAS. 115(13): 3362–3367 (2018)

[27] King NP., Bale JB., Sheffler W., McNamara DE., Gonen S., Gonen T., Yeates TO., Baker D. *Accurate design of co-assembling multi-component protein nanomaterials.* Nature. **510(7503):** 103-8 (2014)

[28] Sojka L., Fučík V., Krásný L., Barvík I., Jonák J. *YbxF, a Protein Associated with Exponential-Phase Ribosomes in Bacillus subtilis.* J Bacteriol **189** (2007)

[29] Baird, NJ., Zhang, J., Hamma, T., and Ferré-D'Amaré, AR. *YbxF and YlxQ are bacterial homologs of L7Ae and bind to K-turns but not K-loops*. RNA. **18(4):** 759-770 (2012)

[30] Schroeder, KT., McPhee, SA., Ouellet, J., and Lilley, DMJ. *A structural database for k-turn motifs in RNA.* RNA. **16(8):** 1463-1468 (2010)

[31] Tiedge, H. *K-turn motifs in spatial RNA coding.* RNA Biology. **3(4):** 133-139 (2006)

[32] Sivaramakrishnan, S., Spink, BJ., Sim, AYL., Doniach, S., and Spudich, JA. *Dynamic charge interactions create surprising rigidity in ER/K α-helical protein motif.* Biophysics and Computational Biology. **105(36):** 13356-13361 (2008)

[33] Wolny, M., Batchelor, M., Bartlett, GJ., Baker, EG., Kurzawa, M., Knight, PJ., Dougan, L., Woolfson, DN., Paci, E., Peckham, M. *Characterization of long and stable de novo single alpha-helix domains provides novel insights into their stability.* Scientific Reports. **7(44341)** (2017)

[34] Goldenzweig, A. et al. *Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability*. Mol. Cell. **63(2):** 337–346 (2016).

[35] Yang, M., Chen, YS., Ichikawa, M., Calles-Garcia, D., Basu, K., Fakih, R., Bui, KH., Gehring, K. *Cryo-electron microscopy structures of ArnA, a key enzyme for polymyxin resistance, revealed unexpected oligomerizations and domain movements.* J Struct Biol. **208(1):** 43-50 (2019).

[36] Robichon, C., Luo, J., Causey, TB., Benner, JS., and Samuelson, JC. *Engineering Escherichia coli BL21(DE3) Derivative Strains To Minimize E. coli Protein Contamination after Purification by Immobilized Metal Affinity Chromatography.* Appl. Environ. Microbiol. **77(13):** 4634-4646 (2011).

[37] Cannon, KA., Park, RU., Boyken, SE., Nattermann, U., Yi, S., Baker, D., King, NP., Yeates, TO. *Design and Structure of two new protein cages illustrate successes and ongoing challenges in protein engineering.* Protein Sci. **29(4):** 919-929 (2019)

[38] Punjani, A., Rubinstein, JL., Fleet, DJ., and Brubaker, MA. *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.* Nature Methods. **14:** 290-296 (2017)

[39] Scheres, SHW. *RELION: Implementation of a Baysian approach to cryo-EM structure determination.* Journal Struct. Biol. **180:** 519-530 (2012)

[40] Suloway, C., Pulokas. J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, CS., Carragher, B. *Automated molecular microscopy: The new Leginon system.* Journal of Structural Biology. **151:** 41-60 (2005)

[41] Yao, Q., Weaver, SJ., Mock, JY., Jensen, GJ. *Fusion of DARPin to Aldolase Enables Visualization of Small Protein by Cryo-EM*. Structure. **27(7):** 1148-1155 (2019)

[42] Emberly, EG., Mukhopadhyay, R., Wingreen, NS., Tang, C. *Flexibility of alpha-helices: results of a statistical analysis of database protein structures.* J Mol Biol. **327(1):** 229-237 (2003)

[43] Kung, JE., Johnson, MC., Jao, CC., Arthur CP. *Disfulfide constrained Fabs overcome target size limitation for high-resolution single-particle cryo-EM.* Biorxiv doi.org/10.1101/2024.05.10.593593. (2024)

[44] Schrödinger, L., & DeLano, W. (2020). *PyMOL*. Retrieved from http://www.pymol.org/pymol

[45] Beren, C., Cui, Y., Chakravarty, A., Yang, X., Rao, ALN., Knobler, CM., Zhou, ZH., Gelbart, WM. *Genome organization and interaction with capsid protein in a multipartite RNA virus*. Proc Natl Acad Sci USA **117(20):** 10673-10680 (2020)

[46] Wang, X., Bigman, LS., Greenblatt, HM., Yu, B., Levy, Y., Iwahara, J. *Negatively charged, intrinsically disordered regions can accelerate target search by DNA-binding proteins.* Nucleic Acids Research. **51(10):** 4701-4712 (2023)

[47] Moschen, T., Wunderlich, C., Kreutz, C., Tollinger, M. *NMR resonance assignments of the archaeal ribosomal protein L7Ae in the apo form and bound to a 25 nt RNA.* Biomol NMR Assign. **9(1):** 177-180 (2015).

[48] Montange, RK., Batey, RT. *Structure of the S-adenosylmethionine riboswitch regulatory element*. Nature. **441:** 1172-1175 (2006)

[49] Wan, WB., Migawa, MT., Vasquez, G., Murray, HM., Nichols, JG., Gaus, H., Berdeja, A., Lee, S., Hart, CE., Lima, W., Swayze, EE., Seth, PP. *Synthesis, biophysical properties and biological activity of second generation antisense oligonucleotides containing chiral phosphorothioate linkages*. Nucleic Acids Res. **42(22):** 13456-13468 (2016)

[50] Gladkov, N., Scott, EA., Meador, K., Lee, EJ., LAganowsky, AD., Yeates, TO., Castells-Graells, R. *Design of a symmetry-broken tetrahedral protein cage by a method of internal steric occlusion.* Protein Sci. **33(4)** (2024)

[51] Lee, EJ., Gladkov, N., Miller, JE., Yeates, TO. *Design of Ligand-Operable Protein-Cages That Open Upon Specific Protein Binding*. ACS Synth Biol. **13(1):** 157-167 (2024)

[52] Agdanowski, MP., Castells-Graells, R., Sawaya, MR., Cascio, D., Yeates, TO., Arbing, MA. *X-ray crystal structure of a designed rigidified imaging scaffold in the ligand-free conformation.* Acta Crysta F **80:** 107-115 (2024)

[53] Meador, K., Castells-Graells, R., Aguirre, R., Sawaya, MR., Arbing, MA., Sherman, T., Senarathne, C., Yeates, TO. *A suite of designed protein cages using machine learning and protein fragment-based protocols.* Structure. **24** (2024)

[54] Banna, HA., Das, NK., Ojha, M., Koirala, D. *Advances in chaperone-assisted RNA crystallography using synthetic antibodies.* BBA Adv. **4** (2023)

[55] Zhang, K., Li, S., Kappel, K., Pintilie, G., Su, Z., Mou, TC., Schmid, MF., Das, R., Chiu, W. *Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution.* Nature Comms. **10(5511)** (2019)

[56] Ma, H., Jia, X., Zhang, K., Su, Z. *Cryo-EM advances in RNA structure determination.* Signal Transduct Target Ther. **7(1):** 58 (2022)

[57] Glaeser, RM., and Hall, RJ. *Reaching the Information Limit in Cryo-EM of Biological Macromolecules: Experimental Aspects.* Biophys J. **100(10):** 2331-2337 (2011)

[58] King NP., Bale JB., Sheffler W., McNamara DE., Gonen S., Gonen T., Yeates TO., Baker D. *Accurate design of co-assembling multi-component protein nanomaterials.* Nature. **510:**103-108. (2014)

[59] Gibson, DG., Young, L., Chuang, RY., Venter, JC., Hutchison III, CA., Smith, HO. *Enzymatic assembly of DNA molecules up to several hundred kilobases.* Nature Methods. **6:** 343-345 (2009)

[60] Punjani, A., Rubinstein, JL., Fleet, DJ., Brubaker, MA. *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.* Nature Methods. **14:** 290-296 (2017)

[61] Pettersen, EF., Goddard, TD., Huang, CC., Couch, GS., Greenblatt, DM., Meng, EC., and Ferrin, TE. *UCSF Chimera - A Visualization System for Exploratory Research and Analysis.* J. Comput. Chem. **25(13):** 1605-1612 (2004)

[62] Zheng, SQ., Palovcak, E., Armache, JP., Verba, KA., Cheng, Y., Agard, DA. *MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy.* Nature Methods. **14:** 331-332 (2017)

[63] Sridharan, G., Shankar, AA. *Toluidine blue: A review of its chemistry and clinical utility*. J Oral Maxillofac Pathol. **16(2):** 251-255 (2012)

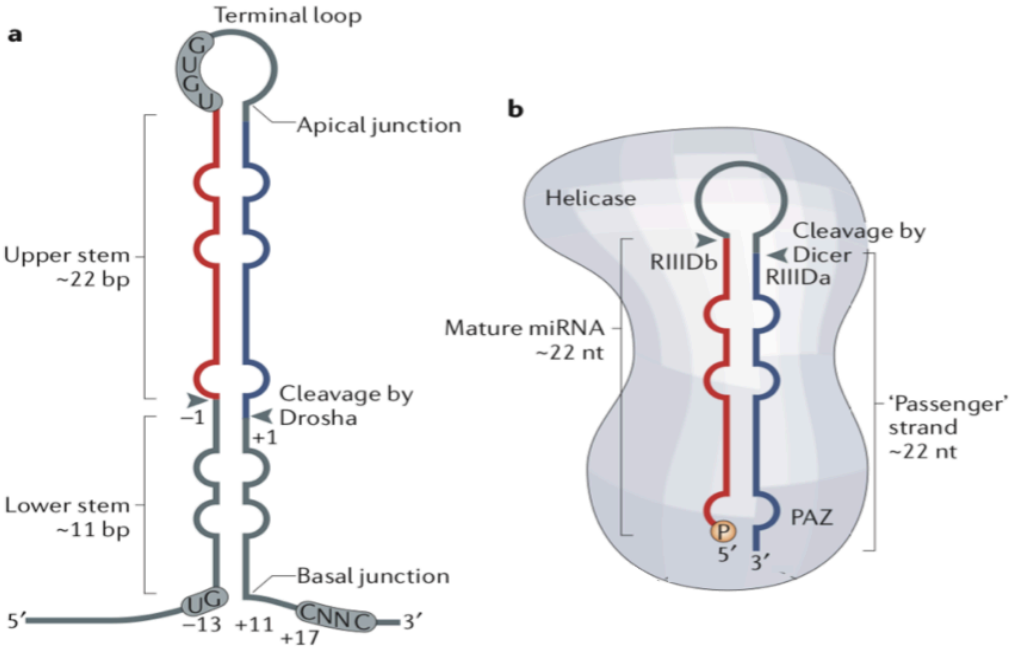


Figure 4.1: Example of microRNA molecule. a) A schematic of a typical pri-microRNA molecule's 2D diagram with Drosha cleavage sites labeled. b) Cartoon representation of the same pri-microRNA, post-Drosha cleavage, loaded into the Dicer complex; Dicer cleavage sites labeled. Figure adapted from Ha et al[5].

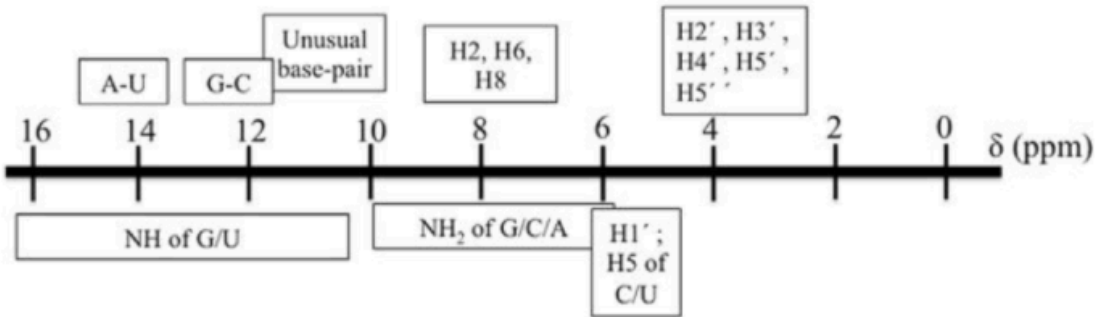


Figure 4.2: RNA Proton Chemical Shifts. This is a typical distribution of chemical shifts resulting from RNA molecules in an NMR experiment[12]. A particular region of interest that causes complications during peak assignment is seen around 4ppm in the spectrum, where there is significant overlap among the protons from the ribose sugars in the H2' and H5' region. This overlap becomes more pronounced as the size of the RNA molecule studied increases, further complicating the analysis and limiting the uses of NMR for particular samples. Figure adapted from Barnwal et al[12].

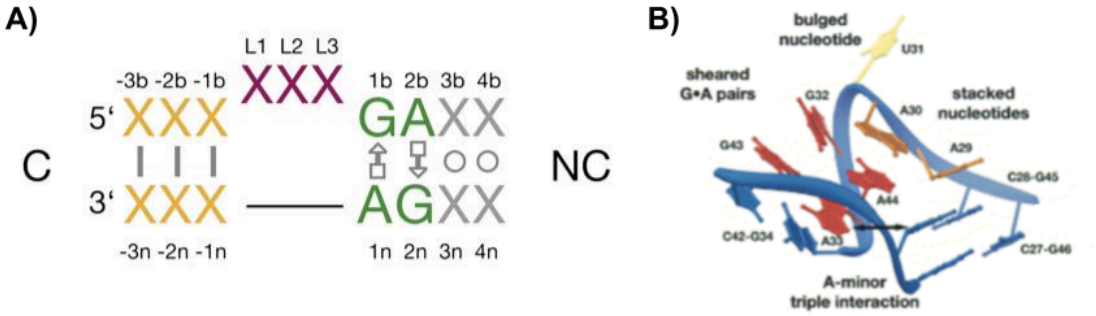
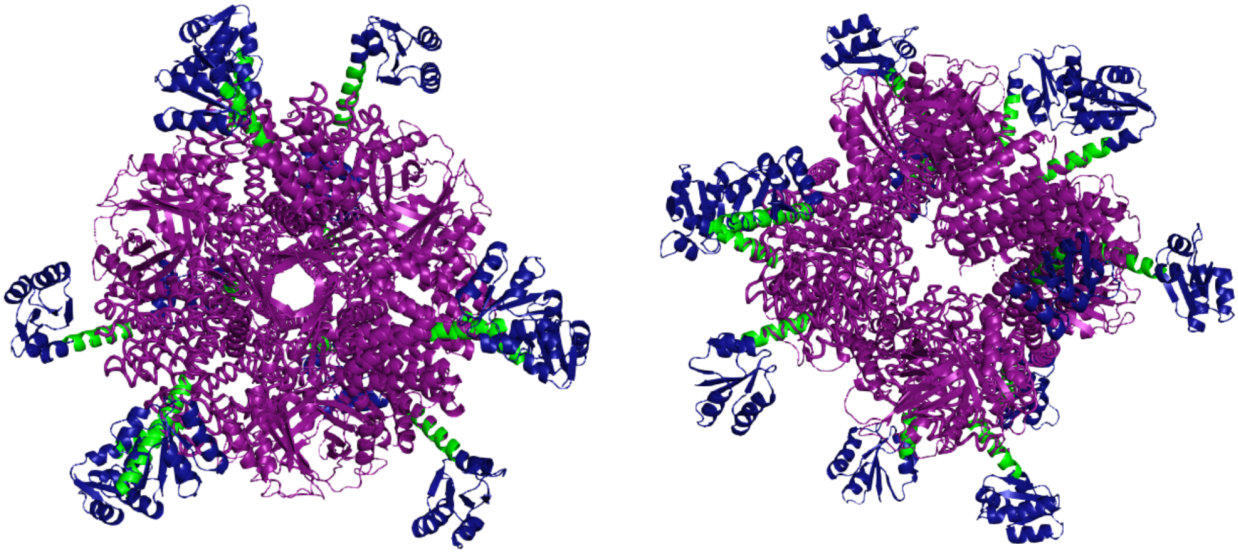


Figure 4.3: Structural and Sequence layout of a K-turn motif. A) Secondary structure motifs shown for typical K-turn sequences. Each motif contains a canonical (C) helix where traditional Watson-Crick pairing occurs, and a non-canonical (NC) helix where non-Watson-Crick pairing occurs. The NC helix is typically anchored by two sets of GA base pairs. Between the C and NC helices is a 3-nucleotide stretch of bulging bases that causes a 60-degree kink in the backbone, giving the motif its name. B) A 3D diagram of an example K-turn with key features highlighted. The drastic curvature, "kink", in the backbone is shown in blue. Non-canonical GA-pairs shown in red, with the characteristic bulging nucleotide shown in yellow. The 3-nucleotide stretch shown in A) can be visualized as A29, A30, and AU31 shown in orange and yellow in B). Figure is adapted from Schroeder et al 2012 and Tiedge 2006[30,31].

Table of Construct Sequences	
Construct Name	Amino Acid Sequence
T33-21-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKL TVTREFGIGAEAAAYLLALSDLLFLLARVIEIEK**EAQK**SYDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-Q166A-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEK**EAAK**SYDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-YbxF-mut1	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEIE**REGERV**YDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-YbxF-mut2	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEI**SREGERVL**DKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-EAER-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEK**EAER**SYDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-EAKR-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEKEAKRSYDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-NAQK-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKV LTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEK**NAQK**SYDKVSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-E4K4E4-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKL TVTREFGIGAEAAAYLLALSDLLFLLARVI**EEEEKKKKEEEE**VSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-K4E4K4-YbxF	RITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKI MGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKLTV TREFGIGAEAAAYLLALSDLLFLLARVI**EEEEKKKKEEEE**VSQAKSIIIGTKQTVKALKRGSVKE VVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-21-E4R4E4-YbxF	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIY KIMGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKL TVTREFGIGAEAAAYLLALSDLLFLLARVI**EEEERRRREEEE**VSQAKSIIIGTKQTVKALKRGSV KEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL
T33-21-R4E4R4-YbxF	RITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKI MGEIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKLTV TREFGIGAEAAAYLLALSDLLFLLARVI**RRRREEEERRRR**VSQAKSIIIGTKQTVKALKRGSVKE VVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKACGIEVGAAAVAIIL*
T33-2_Subunit_B	MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLH ACLSILDGRDIATRTLLGASLCAVLAEAVAGGGEEGVQVSVEVREMERLSYAKRVVARQRLEH HHHHH*

Table 4.1: List of amino acid sequences used for the first round of designs. Table includes all 11 of the initial design names and their corresponding amino acid sequences. Linker amino acids are in bold. The amino acids to the left of the bolded linker correspond to the T33-21 cage core. Amino acids to the right of the bolded linker correspond to the YbxF RNA binding protein. First 11 entries correspond to the fusion subunit A, whereas the last entry, T33-21_Subunit_B corresponds to the second, his-tagged component of the scaffold. The sequence for subunit B stayed invariant throughout all designs. Amino acid sequence was codon optimized for *E. coli* production prior to synthesis



Three-fold symmetric view

Two-fold symmetric view

Figure 4.4: Model of T33-21-YbF designs. Left: view down the three-fold axis of symmetry. Right: View down the two-fold axis of symmetry. Underlying T33-21 cage core shown in purple is composed of two trimeric proteins of C3 symmetry. Bridging linker amino acids shown in green. YbxF RNA binding protein shown in blue. All binding pockets for YbxF are positioned such that bound cargo would be displayed outward from the cage, avoiding clashing with cage or symmetry-related cargo.

Table of Lysis Buffer Conditions	
Buffer Name	Buffer Composition
SB1_Lysis	10mM HEPES 7.2 ; 100mM NaCl ; 10mM imidazole
SB1_Elution	10mM HEPES 7.2 ; 100mM NaCl ; 500mM imidazole
SB2_Lysis	20mM Sodium phosphate 8.0 ; 300mM NaCl ; 10mM Imidazole
SB2_Elution	20mM Sodium phosphate 8.0 ; 300mM NaCl ; 500mM Imidazole
SB3_Lysis	50mM Sodium phosphate 8.0 ; 300mM NaCl ; 10mM Imidazole ; 1% glycerol
SB3_Elution	50mM Sodium phosphate 8.0 ; 300mM NaCl ; 500mM Imidazole ; 1% glycerol
SB4_Lysis	50mM Tris 8.0 ; 250mM NaCl ; 10mM Imidazole
SB4_Elution	50mM Tris 8.0 ; 250mM NaCl ; 500mM Imidazole
SB5_Lysis	50mM Tris 8.0 ; 250mM NaCl ; 10mM Imidazole ; 5% glycerol
SB5_Elution	50mM Tris 8.0 ; 250mM NaCl ; 500mM Imidazole ; 5% glycerol

Table 4.2: List of buffers used in initial expression screen. Buffers were selected based on successful usage in previous scaffolding projects in the lab as well as from literature from the protein design field.

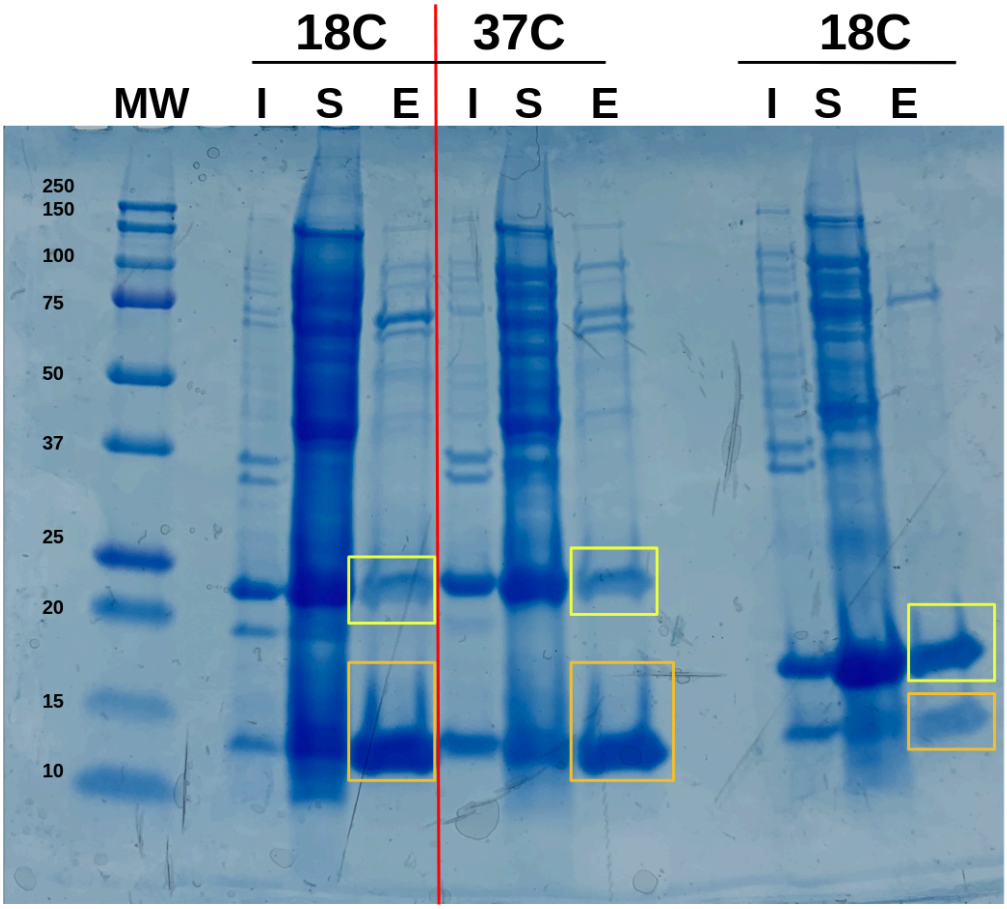


Figure 4.5: SDS-PAGE of small-scale expression screen. Depicted is a gel from the small-scale screen for the T33-21-YbxF scaffolding design. Lanes are as follows: molecular weight marker, insoluble fraction, soluble fraction, elution. Tests were performed at both 18C and 37C as described in Materials and Methods. The fusion subunit A is depicted in yellow boxes - the band migrates slightly smaller than the actual size of 27kDa. His-tagged subunit B is shown in orange boxes and runs at approximately the correct weight. The last 3 lanes in the gel correspond to the T33-21 cage core itself as a control. All experimental testing was carried out the same among scaffold designs and control. Control subunits are color coded in the same manner as T33-21-YbxF scaffold lanes.

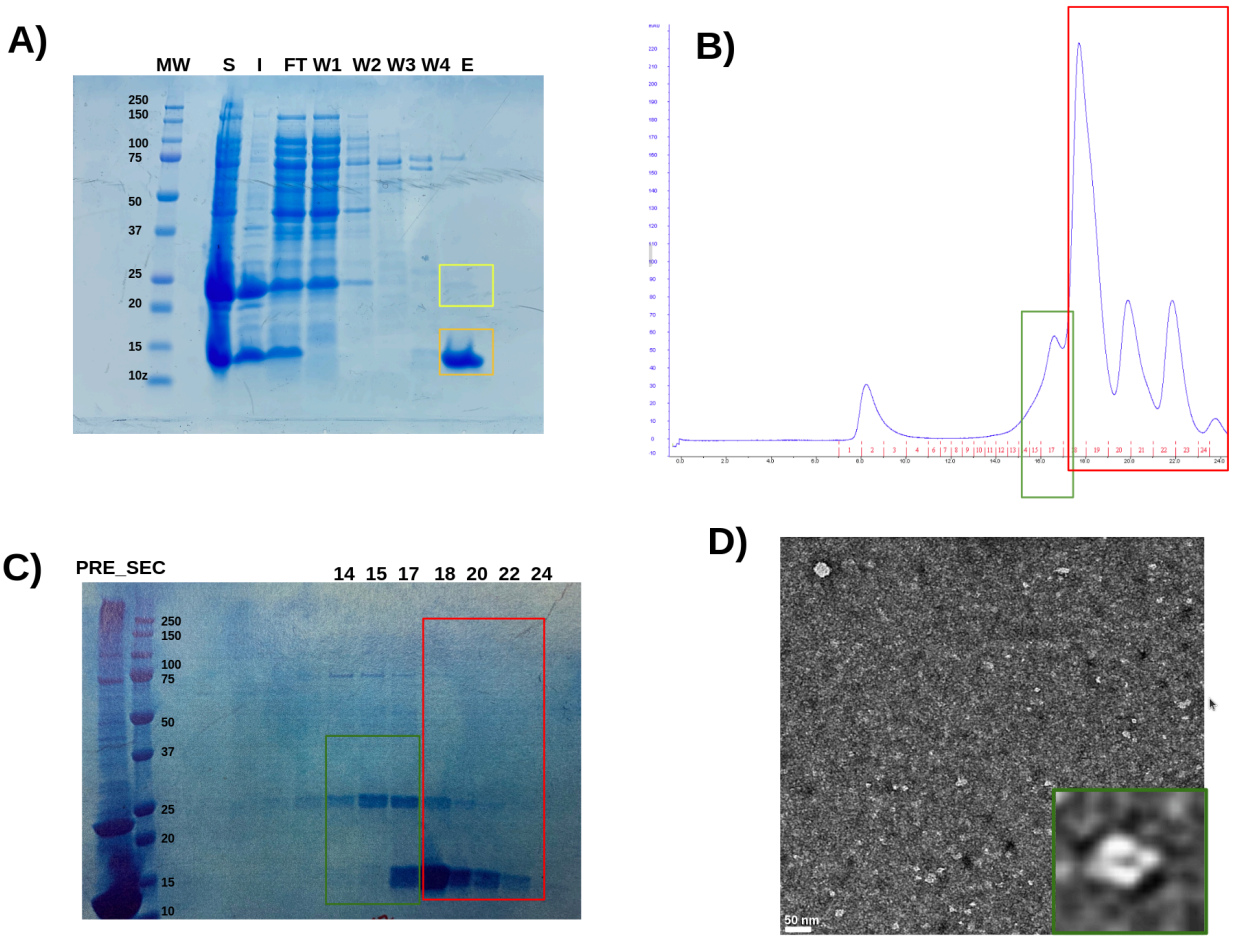


Figure 4.6: Expression inconsistency of fusion subunit. The YbxF-fused subunit A has been shown to experience issues with expression and proper folding. A) SDS-PAGE from gravity-fed affinity elution. The band corresponding to subunit B (bottom band, orange box) is drastically more intense than subunit A (top band, yellow box) which is barely noticeable after imaging. B) concentrated affinity elution loaded onto Superose6 Increase column. Most of the sample contained disassembled particles as evidenced by the large peak in fraction 18 compared to the region corresponding to fully assembled scaffolds (fractions 14-17, ~16.5mL). The imbalance in stoichiometry is seen clearer from the SDS-PAGE in C). Later fractions contain mainly subunit B with fraction 18 containing a mixture of unassembled subunits. D) Low yield is seen by the lack of assembled particles in negative stain micrographs. Inset: zoomed in view of an assembled scaffold. Lack of abundant monodisperse particles indicative of assembly issues.

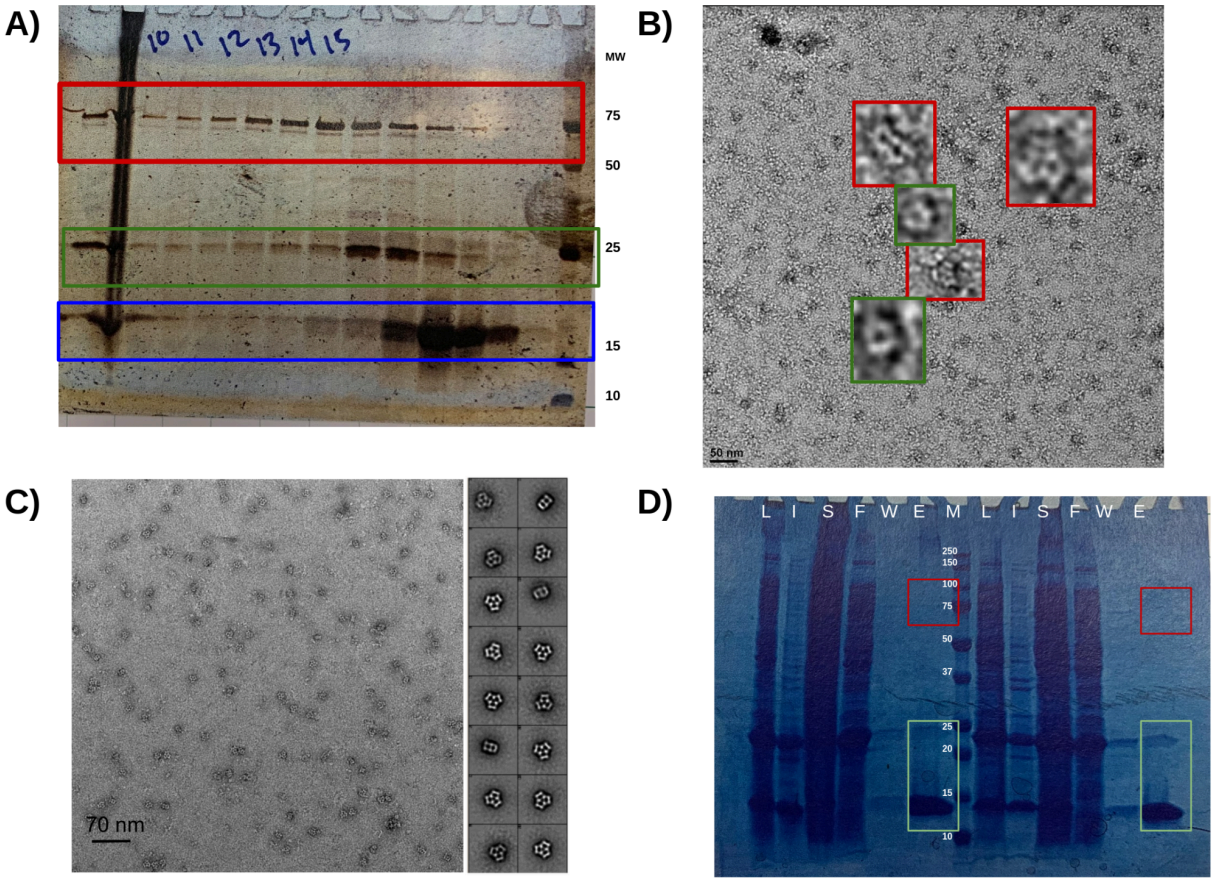


Figure 4.7: Investigation of unknown contaminant. A large unknown contaminant was always present alongside the scaffold during affinity and size exclusion chromatography purification and visualized by electron microscopy. A) A large, 75kDa protein (red box) was co-purifying with scaffold (fusion component in green, non-fusion tagged component in red) in both affinity elution and size exclusion peak fractions. B) Unknown proteins formed large assembly of ~160 Å and appears in both dimers and trimers (red boxes) dispersed among assembled cages (green boxes). C) Literature figure showing micrograph and 2D classes for unknown protein - ArnA, a polymyxin-resistance protein. Clusters of histidine molecules on the protein's surface allow binding of NiNTA columns at high affinities[35]. ArnA has been identified as a common *E. coli* purification contaminant[36]. D) Optimization of elution gradient during affinity purification was able to eliminate ArnA in the final elution step, as indicated by absence of a band outlined in the red boxes in the elution (E) lanes. Success isolation of scaffold alone was achieved (green boxes). Images shown in figure C adapted from Yang et al, 2019.

Table of Design Amino Acid Sequences						
	Subunit A			Subunit B		
Design Name	Fusion	His6	Sequence	Fusion	His6	Sequence
AA1	Yes	No	MRITTKVGDKGSTRLFGGEEVWK DDPIIEANGTLDELTSFIGEAKHYV DEEMKGILEEIQNDIYKIMGEIGSK GKIEGISEERIKWLAGLIERYSEMV NKLSFVLPGGTLESAKLDVCRTIA RRAERKVATVLREFGIGTLAAIYLA LLSRLLFLLARVIEIEKNYDKVSQA KSIIIGTKQTVKALKRGSVKEVVVA KDADPILTSSVVSLAEDQGISVSM VESMKKLGKACGIEVGAAAVAIIL*	No	Yes	MFTRRGDQGETDLANRARV GKDSPVVEVQGTIDELNSFIG YALVLSRWDDIRNDLFRIQND LFVLGEDVSTGGKGRTVTMD MIIYLIKRSVEMKAEIGKIELFV VPGGSVESASLHMARAVSRR LERRIKAASELTEINANVLLYA NMLSNILFMHALISNKRLNIPE KIWSIHRVSLEHHHHHH*
AA2	Yes	No	MRITTKVGDKGSTRLFGGEEVWK DDPIIEANGTLDELTSFIGEAKHYV DEEMKGILEEIQNDIYKIMGEIGSK GKIEGISEERIKWLAGLIERYSEMV NKLSFVLPGGTLESAKLDVCRTIA RRAERKVATVLREFGIGTLAAIYLA LLSRLLFLLARVIEIEKN**AKRA**YDK VSQAKSIIIGTKQTVKALKRGSVKE VVVAKDADPILTSSVVSLAEDQGIS VSMVESMKKLGKACGIEVGAAAV AIIL*	No	Yes	MFTRRGDQGETDLANRARV GKDSPVVEVQGTIDELNSFIG YALVLSRWDDIRNDLFRIQND LFVLGEDVSTGGKGRTVTMD MIIYLIKRSVEMKAEIGKIELFV VPGGSVESASLHMARAVSRR LERRIKAASELTEINANVLLYA NMLSNILFMHALISNKRLNIPE KIWSIHRVSLEHHHHHH*
BA1	No	Yes	MRITTKVGDKGSTRLFGGEEVWK DDPIIEANGTLDELTSFIGEAKHYV DEEMKGILEEIQNDIYKIMGEIGSK GKIEGISEERIKWLAGLIERYSEMV NKLSFVLPGGTLESAKLDVCRTIA RRAERKVATVLREFGIGTLAAIYLA LLSRLLFLLARVIEIEKNKLKEVRSL EHHHHHH*	Yes	No	MFTRRGDQGETDLANRARV GKDSPVVEVQGTIDELNSFIG YALVLSRWDDIRNDLFRIQND LFVLGEDVSTGGKGRTVTMD MIIYLIKRSVEMKAEIGKIELFV VPGGSVESASLHMARAVSRR LERRIKAASELTEINANVLLYA NMLSNILFMHALISNKRL**AKE A**YDKVSQAKSIIIGTKQTVKAL KRGSVKEVVVAKDADPILTSS VVSLAEDQGISVSMVESMKK LGKACGIEVGAAAVAIIL*
BA2	No	Yes	MRITTKVGDKGSTRLFGGEEVWK DDPIIEANGTLDELTSFIGEAKHYV DEEMKGILEEIQNDIYKIMGEIGSK GKIEGISEERIKWLAGLIERYSEMV NKLSFVLPGGTLESAKLDVCRTIA RRAERKVATVLREFGIGTLAAIYLA LLSRLLFLLARVIEIEKNKLKEVRSL EHHHHHH*	Yes	No	MFTRRGDQGETDLANRARV GKDSPVVEVQGTIDELNSFIG YALVLSRWDDIRNDLFRIQND LFVLGEDVSTGGKGRTVTMD MIIYLIKRSVEMKAEIGKIELFV VPGGSVESASLHMARAVSRR LERRIKAASELTEINANVLLYA NMLSNILFMHALISNKRL**RAK EA**YDKVSQAKSIIIGTKQTVK ALKRGSVKEVVVAKDADPILT SSVVSLAEDQGISVSMVESM KKLGKACGIEVGAAAVAIIL*

Table 4.3: List of sequences for T33-51-based alignments. For the new cage core, helical alignments were possible for both subunits. Design name indicated base name for alignment (final name concatenation between Table 4.3 alignments and Table 4.4 linker sequence identities). Fusion column indicated whether helical alignment to the RNA-binding protein was done on this subunit. His6 identifies which subunit contains the His6 tag for purification. Underlined sequences correspond to the RNA-binding portion of the construct. Linker amino acids if present in initial alignment denoted in bold.

List of Linker Amino Acids		
Construct Name	Subunit Fusion	Linker Identity
AA2.01	A	AKRA
AA2.02	A	EAQK
AA2.03	A	EAKR
AA2.04	A	NAQK
AA2.05	A	EAER
BA1.01	B	AKEA
BA1.02	B	EAQK
BA1.03	B	EAKR
BA1.04	B	NAQK
BA1.05	B	EAER
BA2.01	B	RAKEA
BA2.02	B	REAQK
BA2.03	B	REAKR
BA2.04	B	RNAQK
BA2.05	B	REAER
BA2.06	B	AEKER
BA2.07	B	KEAER

Table 4.4: List of bridging linker sequences for T33-51 designs. For designs that required additional amino acids, a multi-amino acid linker sequence was added following similar guidelines as T33-21 designs (Table 4.1). Designs were visually inspected for clashes before ordering.

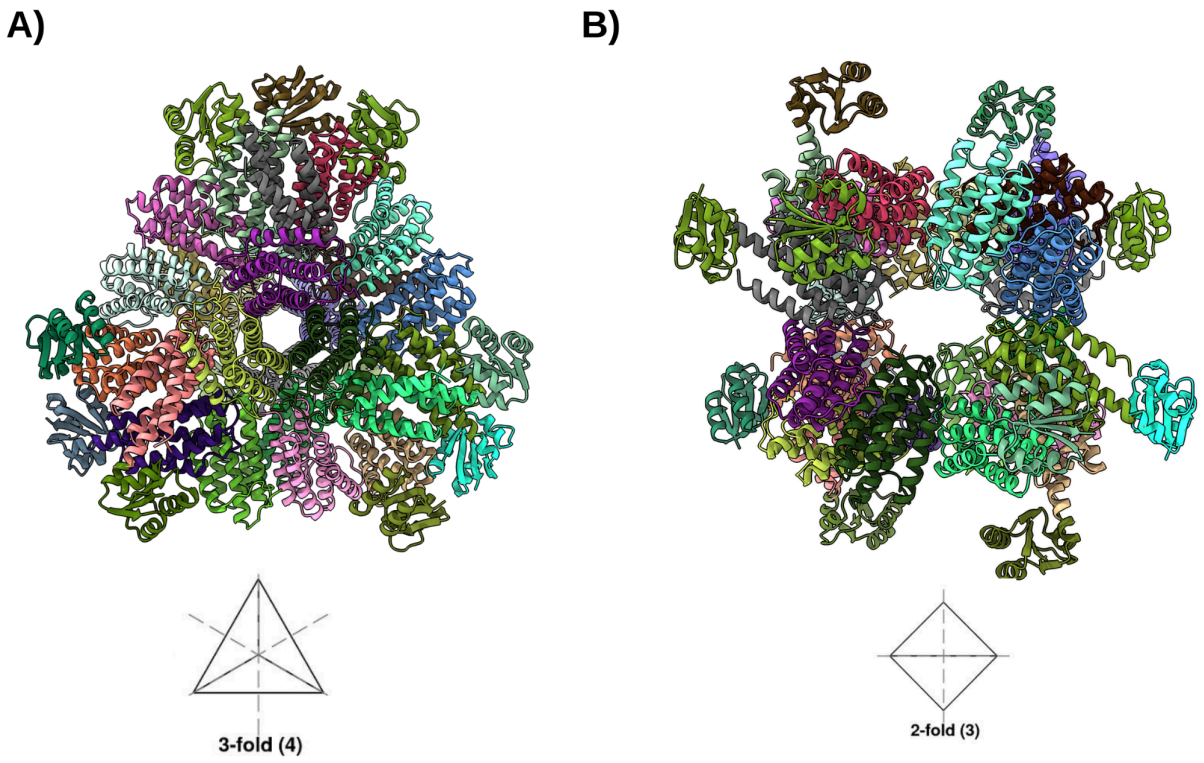


Figure 4.8: Representative model of T33-51-YbxF designs. Cartoon representation of the second round of symmetric scaffold designs. A) Looking down the three-fold axis of symmetry. B) Looking at the two-fold axis of symmetry. Below each cartoon model is a diagram of the tetrahedron that each design's architecture mimics, oriented in the same direction as the scaffold.

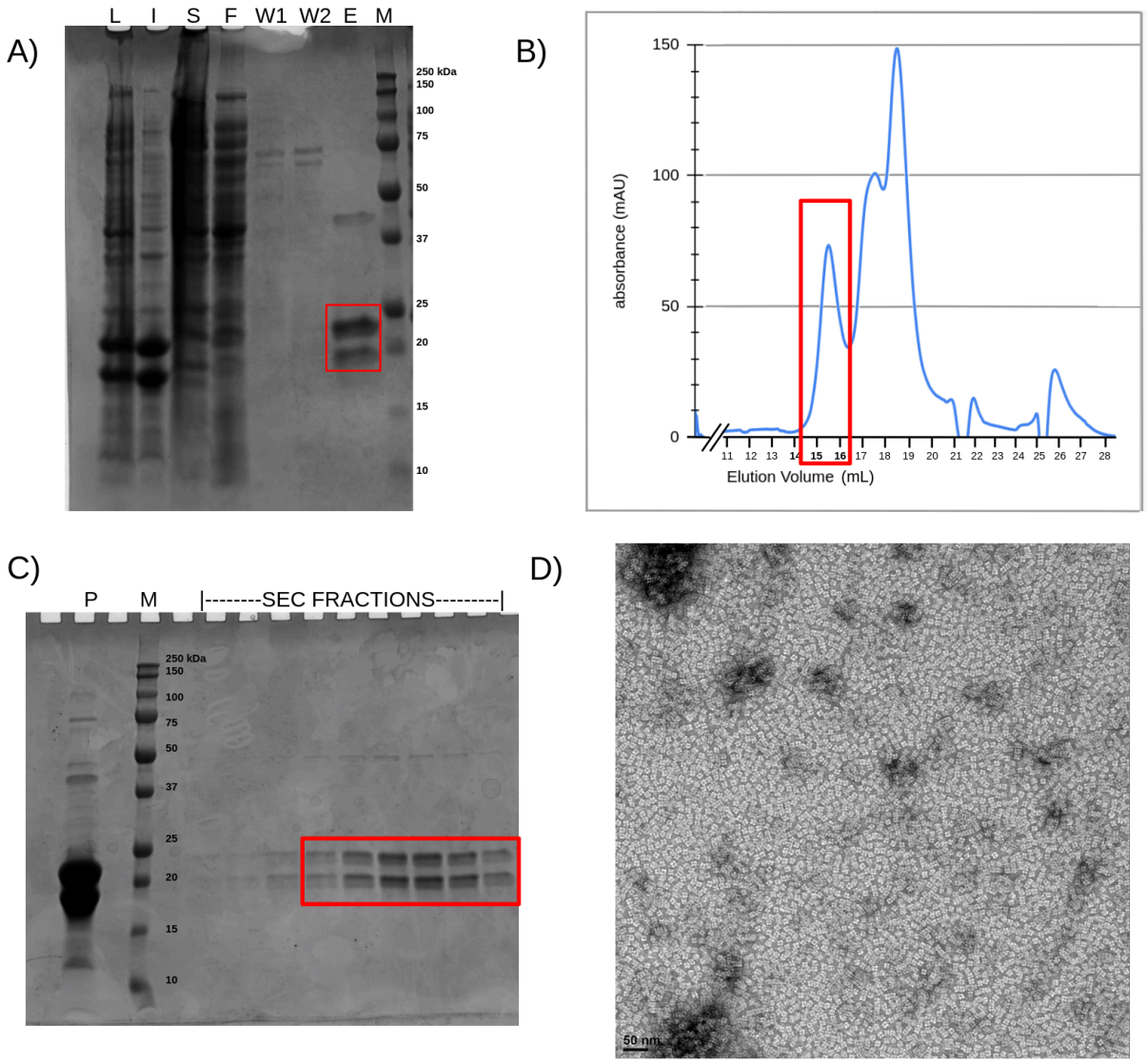
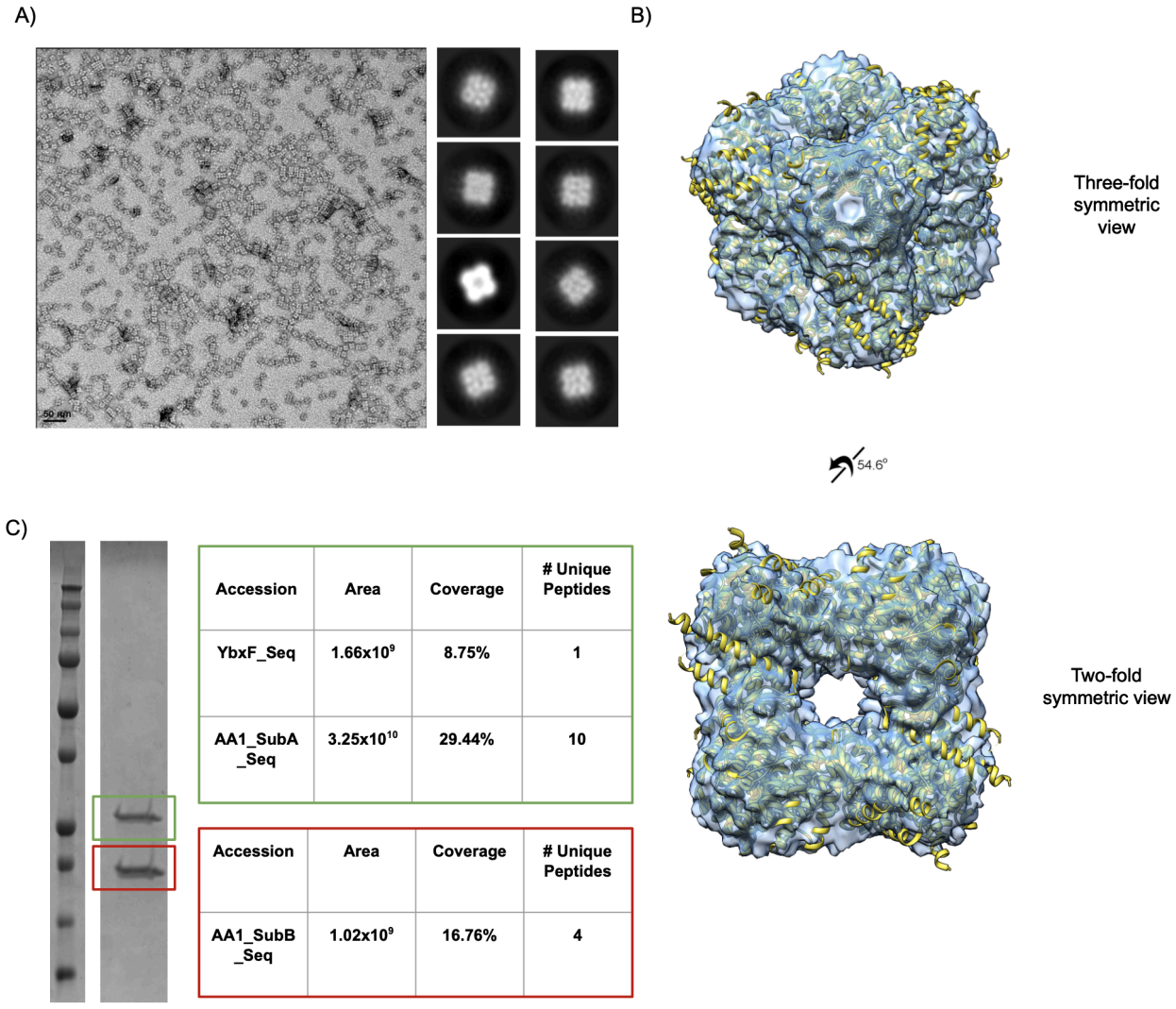


Figure 4.9: Characterization of T33-51-based designs. A) SDS-PAGE from an affinity purification of T33-51-AA1-YbxF. Lanes are as follows: Lysate (L), Insoluble (I), Soluble (S), Flowthrough (F), Wash 1 (W1), Wash 2 (W2), Elution (E), Molecular Weight marker (M). The bands corresponding to both components are seen at their appropriate sizes boxed in red in the elution lane. B, Size Exclusion profile of T33-51-AA1-YbxF concentrated affinity elution. 300uL was loaded onto a Superose6 Increase column and assembled scaffolds were seen eluting starting at 14mL, boxed in red. C) resulting SDS-PAGE from SEC run. Lanes are as follows: Pre-SEC concentrated elution (P), Molecular Weight marker (M), SEC fractions. Fractions of the assembled cages are boxed in red and correspond to the peak boxed in B). D) Representative negative stain electron micrograph of T33-51-AA1-YbxF design. This particular micrograph was prepared directly from the elution fraction with no concentration or dilution, illustrating the high yields that are possible resulting from the new cage cores' increased stability.



Accession	Area	Coverage	# Unique Peptides
YbxF_Seq	1.66×10^{9}	8.75%	1
AA1_SubA_Seq	3.25×10^{10}	29.44%	10

Accession	Area	Coverage	# Unique Peptides
AA1_SubB_Seq	1.02×10^{9}	16.76%	4

Figure 4.10: Biophysical analysis of T33-51-based scaffolds. A) A representative negative stain micrograph from the TF20 data collection session. 71 micrographs in total were collected for processing. To the right of the micrograph are the best set of 2D classifications used in further processing. B) *Ab initio* 3D density maps created in cryoSPARC generated using particles from the 2D classes in 4.10A. Top is looking down the design's three-fold axis of symmetry, whereas the bottom image is being viewed down the two-fold axis. The entirety of the T33-51 cage core could be modeled within the density. No density corresponding to YbxF was seen. C) Bands for both components of the scaffold were excised and analyzed by bottom-up mass spectrometry. Analysis verified presence of YbxF attached to subunit A (top table, green box) despite not seeing density for it in the map.

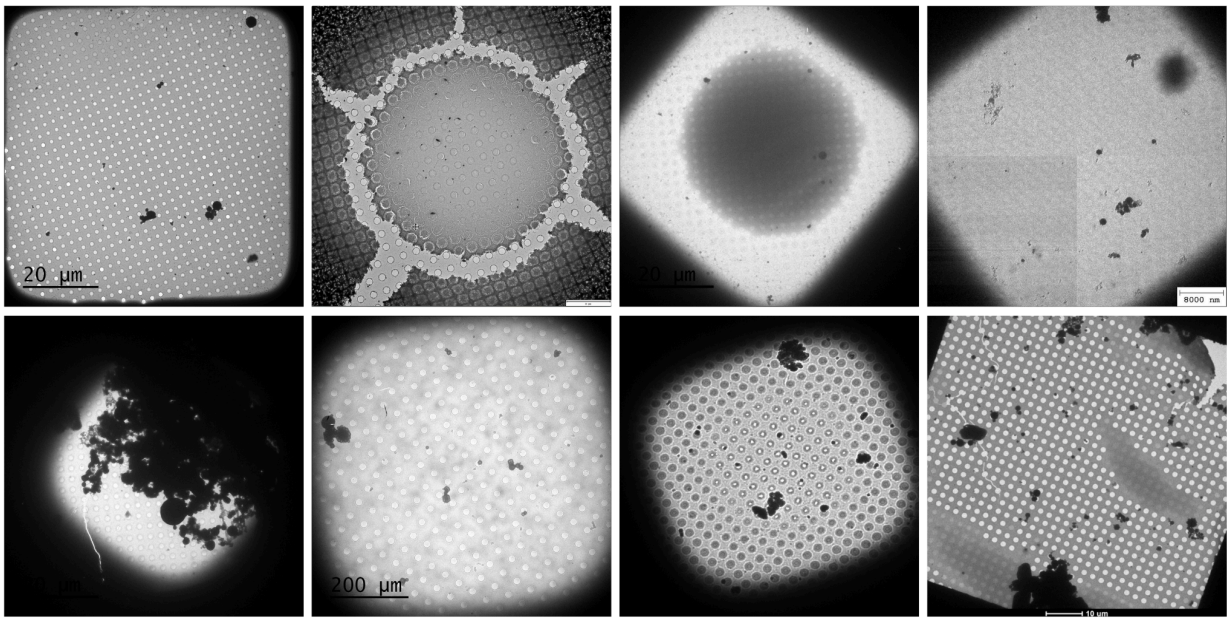


Figure 4.11: Low magnification cryo micrographs of ice conditions. When optimizing your sample for high resolution cryo data collection, you must first fine tune your freezing conditions and assess your efforts by screening for ice quality. Some of the kinds of ice you may encounter are shown above. Starting clockwise from left to right: Suitable ice for data collection (top left), holes are filled but still visible, not many empty. Severe cracking of ice (top, middle left), unsuitable for collection. Regions of extremely thick ice (top, middle right), some holes may be thin enough for collection. Ice too thick (top right), blocks visibility of holes. Particles within holes will lack contrast and not be visible, not suitable for collection. Empty holes (bottom right), ice was too thin, possibly due to blotting too long or too hard. Many holes will lack particles, limiting the amount of data you can collect, which may be suitable for collection in holes of ideal ice. Variable ice thickness within holes (bottom, middle right). Ice is thinner in the center of the hole, giving a white spot, and thicker along the edge of the hole. May be suitable for collection for some samples whose orientations tend to favor different thicknesses of ice. Splotchy (bottom, middle left) ice may be caused by additives such as glycerol and DMSO that change the properties of the solution. May be suitable for data collection as long as ice inside the hole is suitable. Severe contamination (bottom left), caused by numerous factors such as dirty nitrogen. Water molecules freeze on the grid or are deposited on the grid causing black objects impenetrable by the electrons. May be suitable for collection if sufficient holes are unobstructed.

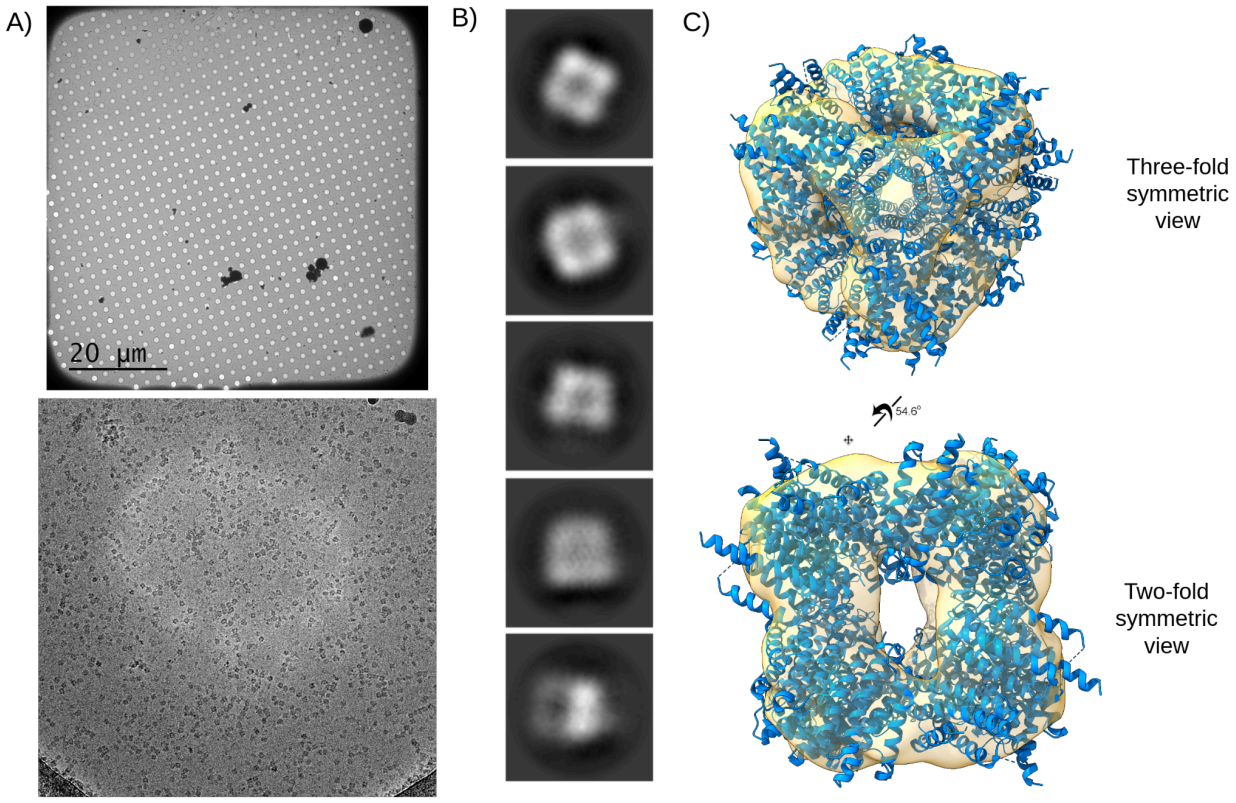


Figure 4.12: Preliminary cryo-EM characterization of T33-51-AA1-YbxF. Before data collection on the Titan Krios, a small dataset was taken on the FEI TF20 to assess cryo quality of the sample. A) top, low magnification image of the ice quality of a typical hole used in data collection. Bottom, hole-level view of grid, particles are clearly visible as dark, square-like objects. B) 2D classification images of scaffold. C) low-resolution 3D maps coming from cryo data. Model of the T33-51 core could be loosely fit into density.

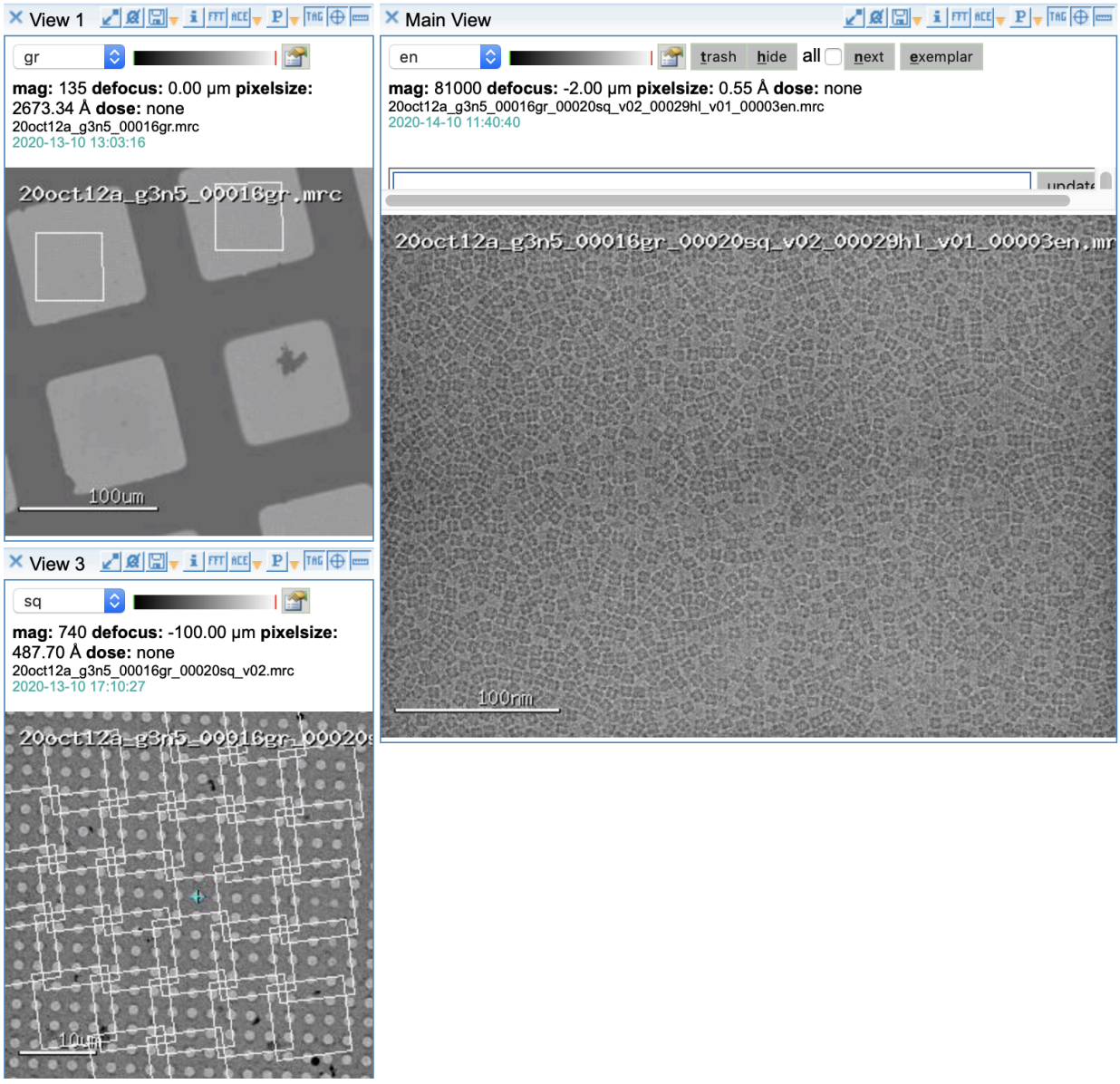


Figure 4.13: Example of Leginon remote data collection session. Using the remote Leginon GUI, users can track and make changes to their session in real-time. Important parameters such as magnification, defocus and pixel size are displayed above each image. Top left, low magnification, grid-level view of the grid. White boxes denote regions where data will be collected. Bottom left, hole-level view of grid. White boxes denote a 3x3 grid of holes to be used for collection. Right, high magnification micrograph of the holes selected. Monodisperse particles are clearly visible throughout the entire hole.

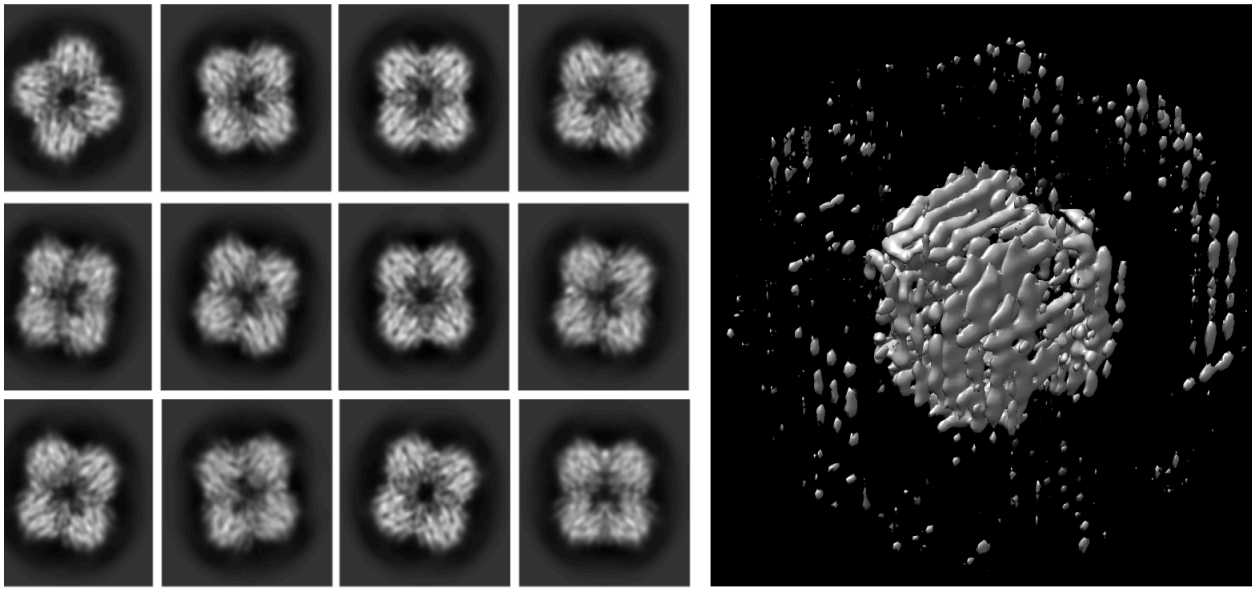


Figure 4.14: Early processing of AA1 Krios data. Initial processing attempts were complicated by the severe preferred orientation problem encountered by how my particles were positioned on the grid. Left, representative 2D classes from the initial rounds of 2D classification on roughly 3 million particles. Right, resulting 3D volume from an *ab initio* 3D refinement job, viewed from a screw position to illustrate the flatness caused by the preferred orientation of the particles.

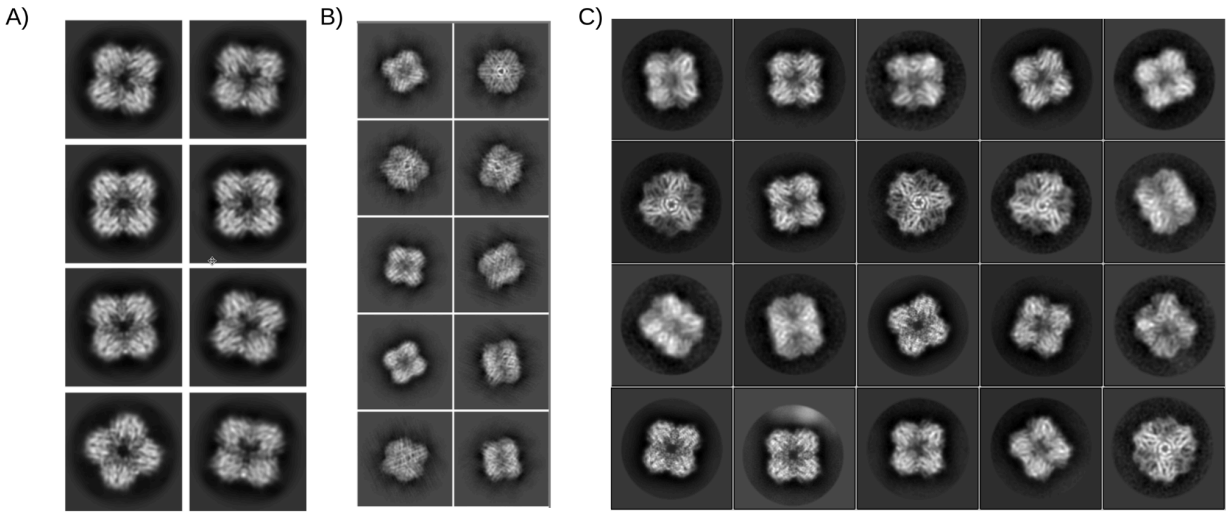


Figure 4.15: Recovering underrepresented particle orientations. Due to the preferred orientation of my scaffolds on the grids, early processing efforts were complicated by the lack of representative views needed for 3D reconstructions. Through iterative rounds of 2D classification, I was able to extract enough orientations for further refinement. Above shows the job progress as I start with only two-fold views (A), recover additional, but lower resolution views (B), and finally extract views of multiple angles of my particles with high enough resolution that secondary features are visible (C).

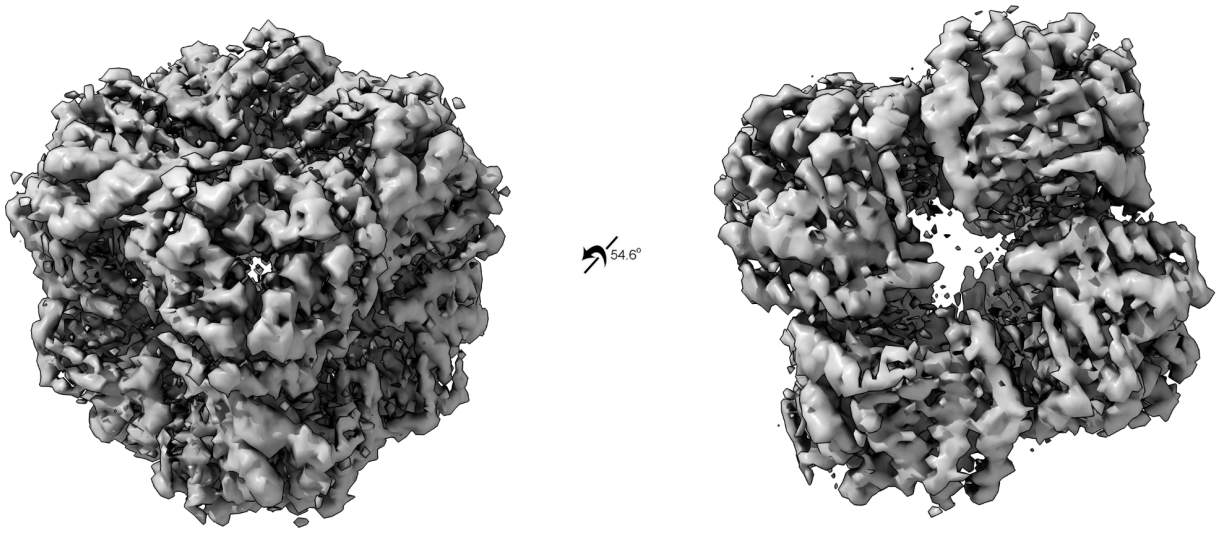


Figure 4.16: Poor 3D Refinement volumes of AA1 Krios data. Following homogeneous and heterogeneous refinement of particles stemming from our expanded viewset of 2D classes, we are left with a low resolution, blown-out volume. Rough secondary features are identifiable as both subunit A and B share the same fold, but high resolution information is smudged from the averaging of different sequences together.

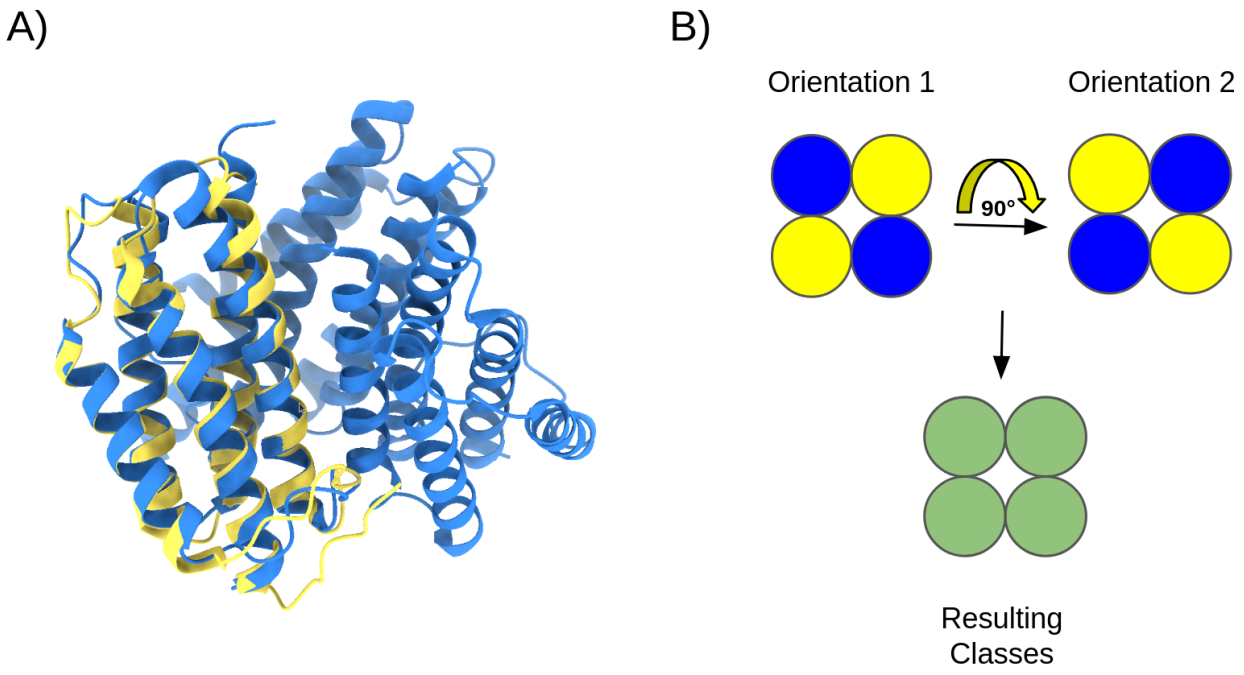


Figure 4.17: Orientation issues with AA1 Krios processing. Due to the similarity between the subunits and the lack of clear density stemming from YbxF to break the symmetry, the processing programs are failing to distinguish between subunit's A and B and are averaging the two orientations together. A) Alignment of a subunit B monomer (yellow, PDB: 1NOG) on top of a trimer of cage core subunit A (blue, PDB: 1WY1). The overlap between subunit structures is significant with an RMSD difference of only 0.85 Å as calculated by PYMOL's alignment protocol[44]. B) Cartoon representation of our hypothesis of what is occurring. Without YbxF to break the symmetry, processing algorithms see only the cage core in two possible orientations. Without the additional features, both orientations are being combined into the same bin, resulting in an average between the two orientations.

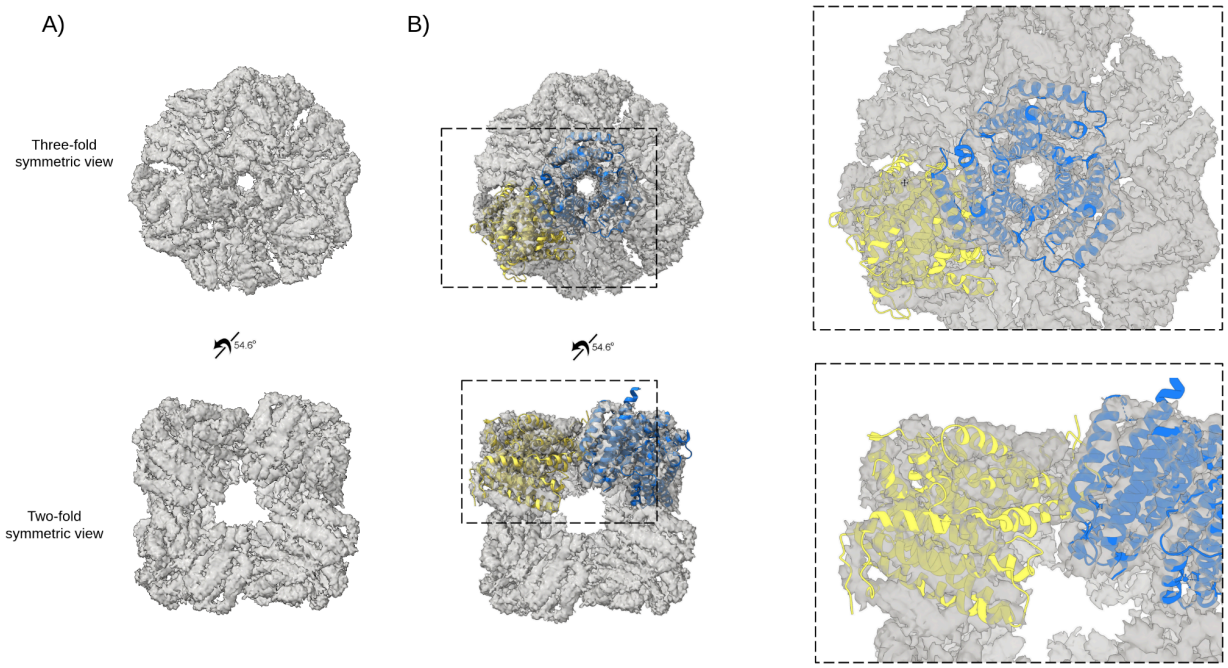


Figure 4.18: Resulting volumes from particle subtraction. Following our series of masking and particle subtraction attempts, the “best” resulting 3D volumes were still worse than the maps obtained before the particle subtraction attempts. Top is looking down the three-fold axis of symmetry and the bottom images are oriented down the two-fold axis. A) Representative 3D volume from a homogeneous refinement job on micrographs subtracted with an entire trimeric subunit B mask. Particle looks tetrahedral but not many distinct features are visible and a lot of noise is present. B) Both components of the T33-51 cage core were fit into the density. In dashed boxes to the right are zoomed in orientations showing the models fit into the density. Neither subunits fit particularly well into the density. Fusion component A (blue) fits a lot better than the non-fusion component B (yellow), as evidenced by fewer portions of the protein sticking out from the grey density.

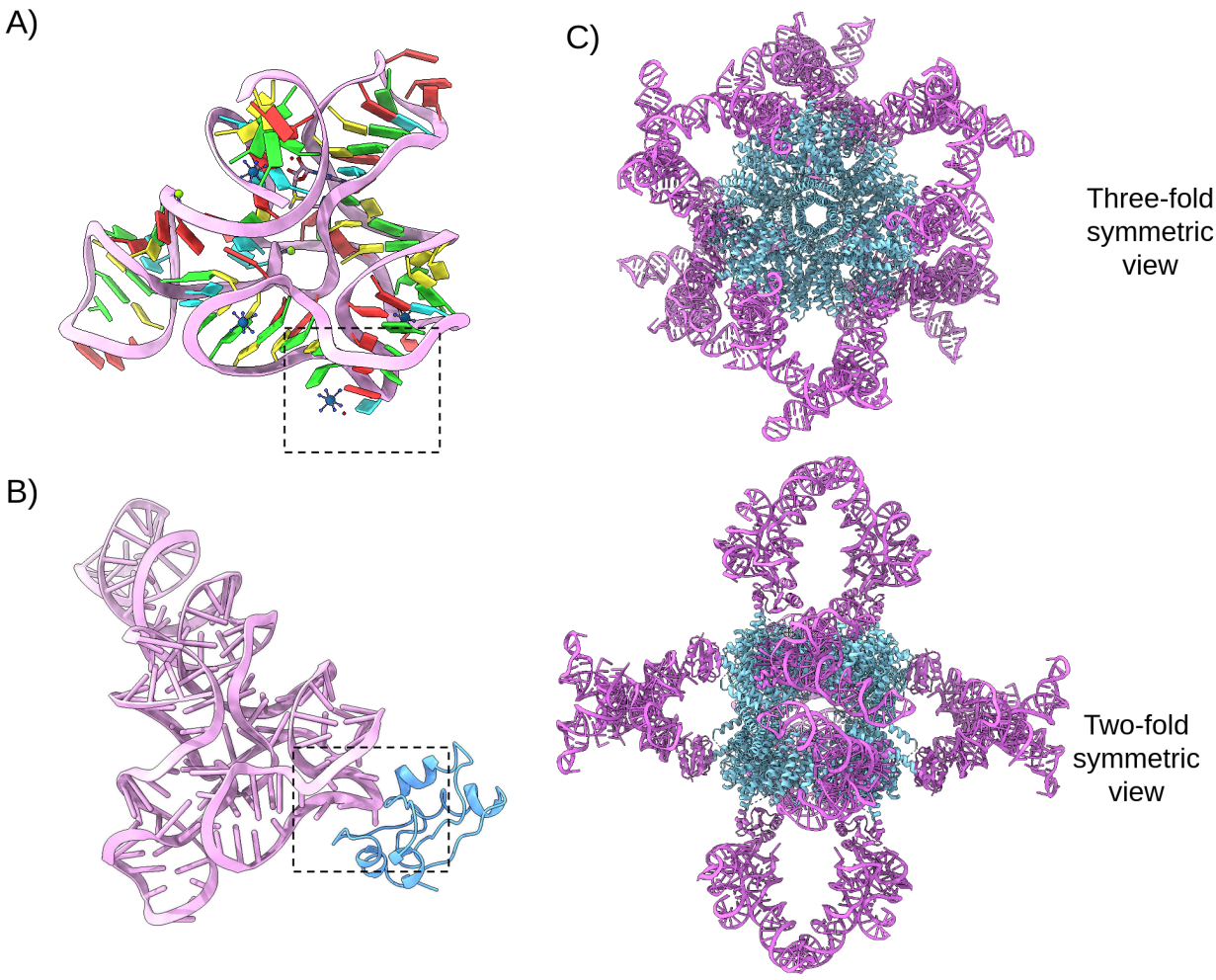


Figure 4.19: Free, complexed and displayed structure of RNA riboswitch cargo. Above is a series of structures of the 2GIS riboswitch chosen to be our initial RNA cargo. A) The free structure of the SAM-I riboswitch with K-turn motif used in binding outlined in the checkered box. B) Structure of the YbxF RNA-binding protein co-crystallized with a SAM-I riboswitch, with interacting surface outlined in the checkered box. This particular riboswitch construct has had the helix distal to the binding surface elongated to aid in crystal formation. C) Design model of T33-51-AA1-YbxF scaffold looking down the three-fold (top) and two-fold (bottom) axes of symmetry. RNA cargo is positioned outward, facing away from the scaffold to avoid clashing with the cage itself or with symmetry-related copies of the cargo. Visual inspection shows the distinct pore caused by the two-fold axis is occluded by two copies of 2GIS from separate trimers.

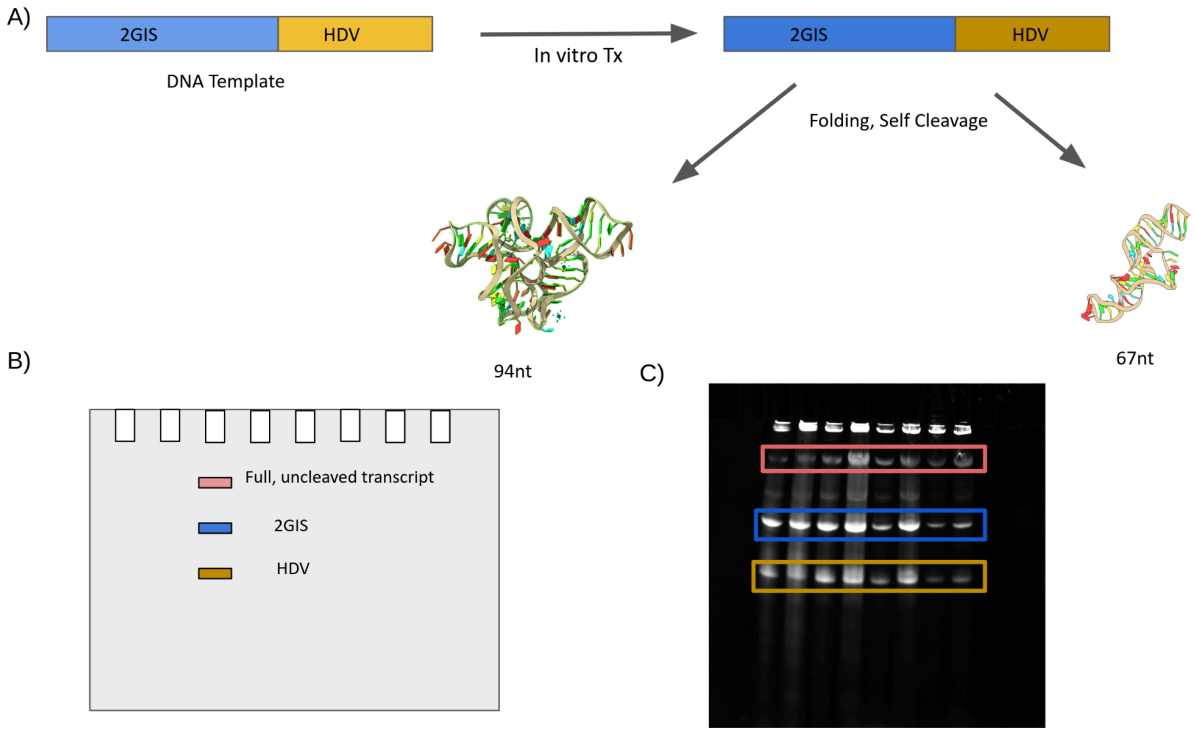


Figure 4.20: HDV ribozyme-assisted run-off transcription of 2GIS. To generate the RNA cargo for binding studies with the T33-51-AA1-YbxF cage, I utilized run-off transcription to generate a RNA fragment containing 2GIS and an HDV ribozyme. A) A cartoon schematic of the transcription process. The 2GIS gene contains its own T7 promoter. Once the construct is transcribed, the HDV ribozyme folds and performs a self-cleavage event just upstream of it on the 5' end, creating two RNA fragments which can be further purified to isolate your target RNA. B) A hypothetical gel of the finished transcription reaction. Some amount of misfolded, uncleaved, or inactive HDV will be present resulting in a full intact construct band (red, top). If the cleavage reaction is successful, two resulting bands should be present in the gel, a larger band corresponding to the ~100nt 2GIS target RNA (blue, middle), and a smaller band corresponding to the smaller ~70nt HDV ribozyme (orange, bottom). C) A representative acrylamide gel of a post-transcription reaction. All three species discussed in B) are identifiable, but also some degree of degradation or contamination present.

In Vitro Run-Off Tx Optimization - 50uL reactions					
Reaction number	Volume 10x Tx buffer	Volume 60nm template	Volume 6.4mg/ml T7 Polymerase	Volume 20mM NTPs	Volume nanopure water
1	2uL	1.67uL	0.6uL	3uL	12.73uL
2	2uL	1.67uL	1.2uL	3uL	12.13uL
3	2uL	3.3uL	0.6uL	3uL	11.1uL
4	2uL	3.3uL	1.2uL	3uL	10.5uL
5	2uL	5.0uL	0.6uL	3uL	9.4uL
6	2uL	5.0uL	1.2uL	3uL	8.8uL
7 (control)	2uL	1.0uL PCR product	1.2uL	3uL	12.8uL

Table 4.5: List of transcription conditions for 2GIS in vitro production. Reactions were run in 50uL reactions for 4 hours at 37°C, with half the reaction being harvested and quenched after 2 hours for PAGE gel analysis.

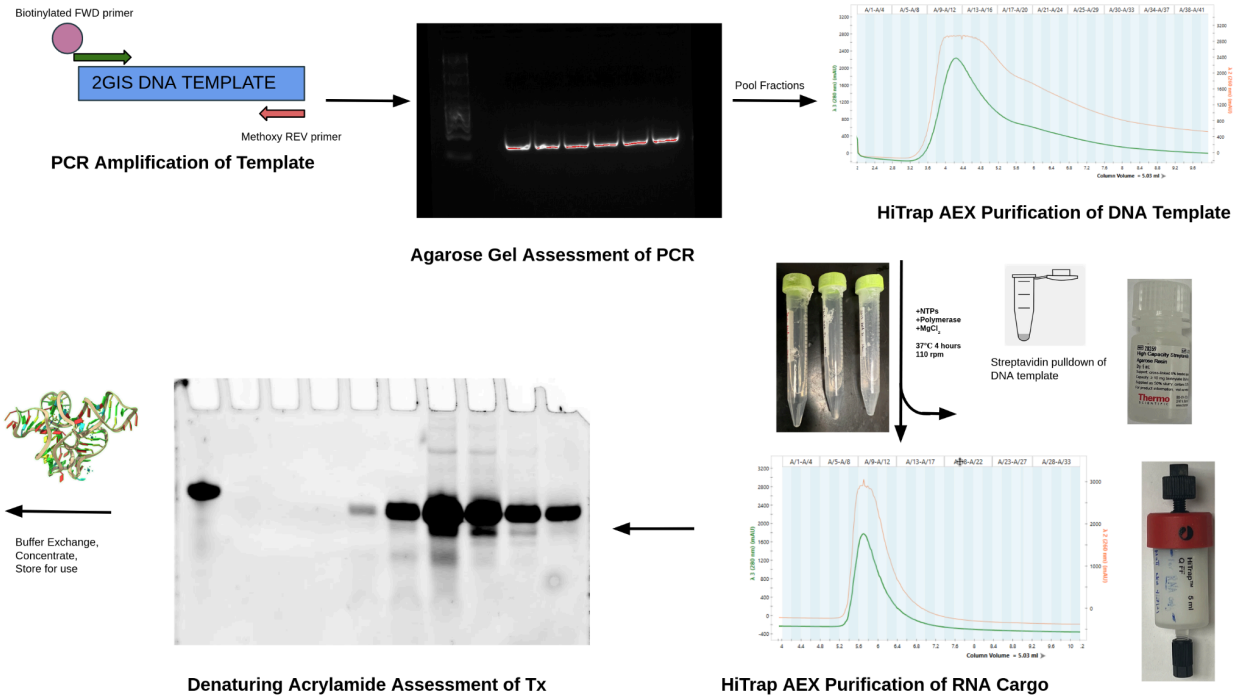


Figure 4.21: In Vitro RNA Transcription of 2GIS. Final workflow for the in vitro transcription of RNA riboswitch cargo. First step is a PCR amplification of the DNA and introduction of biotin for future pulldown by a 5' forward primer. PCR reaction products analyzed by agarose gel electrophoresis. Successful reaction products were pooled and purified by anion exchange chromatography. Elution fractions were pooled, concentrated and used for RNA production. In vitro transcription reactions were shaken and allowed to proceed for 4 hours before quenching followed by incubation with streptavidin-conjugated beads. A pulldown was performed to remove any DNA templates before subsequent purification. RNA cargo was purified by anion exchange chromatography where fractions were analyzed by denaturing PAGE and homogeneous fractions were buffer exchanged, concentrated and stored for future binding studies.

Table of Melting Conditions					
RNA concentration	Melting Temperature	Melting Duration	Snap or Slow Cool?	Additives?	Buffer
26uM	65 ℃	5 minutes	Slow, RT 10mins	No	Water
26uM	65 ℃	5 minutes	Snap, on Ice 10mins	No	Water
26uM	90 ℃	1 minute	Slow, RT 10mins	No	Water
26uM	90 ℃	1 minute	Snap, on Ice 10mins	No	Water
15uM	90 ℃	2 minutes	Slow, RT 8mins	Yes, 10mM $MgCl_2$	10mM HEPES 7.0
15uM	90 ℃	2 minutes	Slow, RT 8mins	Yes, 10mM $MgCl_2$	50mM Tris 8.0
2uM	95 ℃	2 minutes	Slow, RT 10mins	Yes, 5mM $MgCl_2$, 5mM SAM	10mM HEPES 7.0
4uM	95 ℃	2 minutes	Slow, RT 6mins	Yes, 5mM $MgCl_2$, 5mM SAM	10mM HEPES 7.0, 150mM NaCl
4uM	95 ℃	2 minutes	Slow, RT 6mins	Yes, 5mM $MgCl_2$, 5mM SAM	50mM Tris 8.0, 150mM NaCl
1uM	85 ℃	2 minutes	Slow, RT 8mins	No	50mM Tris 8.0, 150mM NaCl
1uM	85 ℃	2 minutes	Slow, RT 8mins	Yes, 5mM $MgCl_2$	50mM Tris 8.0, 150mM NaCl
1uM	85 ℃	2 minutes	Snap, on Ice 5mins	No	50mM Tris 8.0, 150mM NaCl
1uM	85 ℃	2 minutes	Snap, on Ice 5mins	Yes, 5mM $MgCl_2$	50mM Tris 8.0, 150mM

					NaCl
1uM	65 ℃	10 minutes	Slow, RT 5mins	No	50mM Tris 8.0, 150mM NaCl
1uM	65 ℃	10 minutes	Slow, RT 5mins	Yes, 5mM $MgCl_2$	50mM Tris 8.0, 150mM NaCl
1uM	65 ℃	10 minutes	Snap, on Ice 5mins	No	50mM Tris 8.0, 150mM NaCl
1uM	65 ℃	10 minutes	Snap, on Ice 5mins	Yes, 5mM $MgCl_2$	50mM Tris 8.0, 150mM NaCl
1uM	85 ℃	2 minutes	Slow, RT 10mins	No	25mM Tris 7.5, 40mM NaCl, 5mM $MgCl_2$
1uM	85 ℃	2 minutes	Snap, on Ice 5mins	No	25mM Tris 7.5, 40mM NaCl, 5mM $MgCl_2$
1uM	65 ℃	10 minutes	Slow, RT 10mins	No	25mM Tris 7.5, 40mM NaCl, 5mM $MgCl_2$
1uM	65 ℃	10 minutes	Snap, on Ice 5mins	No	25mM Tris 7.5, 40mM NaCl, 5mM $MgCl_2$
1uM	85 ℃	2 minutes	Slow, RT 10mins	Yes, 5mM $MgCl_2$	20mM HEPES 7.5, 100mM KOH
1uM	85 ℃	2 minutes	Snap, on Ice 5mins	Yes, 5mM $MgCl_2$	20mM HEPES 7.5, 100mM KOH
1uM	65 ℃	10 minutes	Slow, RT 10mins	Yes, 5mM $MgCl_2$	20mM HEPES 7.5, 100mM KOH
			Snap, on Ice	Yes, 5mM	20mM

1uM	65 ℃	10 minutes	5mins	$MgCl_2$	HEPES 7.5, 100mM KOH
1uM	85 ℃	2 minutes	Slow, RT 8mins	Yes, 17.9mM $MgCl_2$, 538μM SAM	26mM HEPES 7.5, 53mM KCl

* RT denotes room temperature, ~25℃
** SAM denotes S-AdenosylMethionine

Table 4.6: Summary of RNA refolding conditions tested. To ensure the accurate three-dimensional structure of our target cargo RNA, proper refolding conditions need to be identified. Taking protocols from the literature and advice from our collaborators, I tested a wide range of conditions. RNA was resuspended in various buffers and melted at a range of temperatures and times, followed by either a slow cooling on the benchtop or snap cooled via plunging the sample into a bucket of ice. As the RNA was allowed to start refolding, magnesium ions or SAM ligands were added and the RNA was allowed to continue folding. Various concentrations were tested as literature has shown better refolding can occur in more dilute conditions. Folding efficiency of each method was assessed by native PAGE.

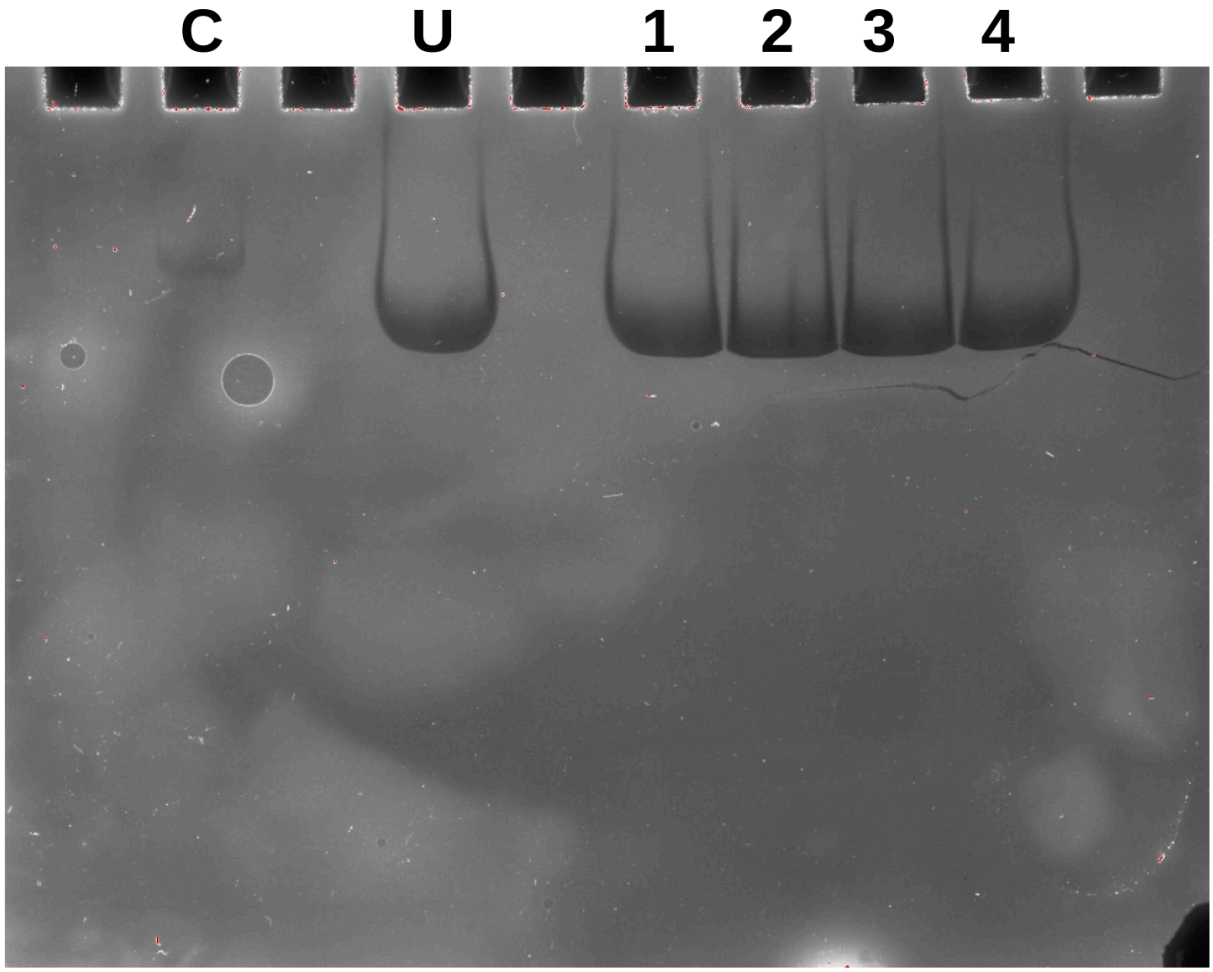


Figure 4.22: 2GIS refolding condition optimization. Representative native gel of RNA samples after refolding. Before binding experiments can be performed, proper folding of the 2GIS riboswitch must be accomplished. A wide variety of refolding conditions were tested (Table 4.6) and assessed by native PAGE. The above gel represents one such series of refolding experiments. In lane C, is a 100 nucleotide folded RNA from the Guo lab used as a control. Lane U shows the unmodified RNA, fresh from the purification process. Lanes 1 - 4 represent various conditions tested. 1) 65°C melt, slow cool. 2) 65°C, snap cool. 3) 90°C, slow cool. 4) 90°C, snap cool. No noticeable difference was observed by native PAGE between the unmodified 2GIS and the refolded 2GIS lanes.

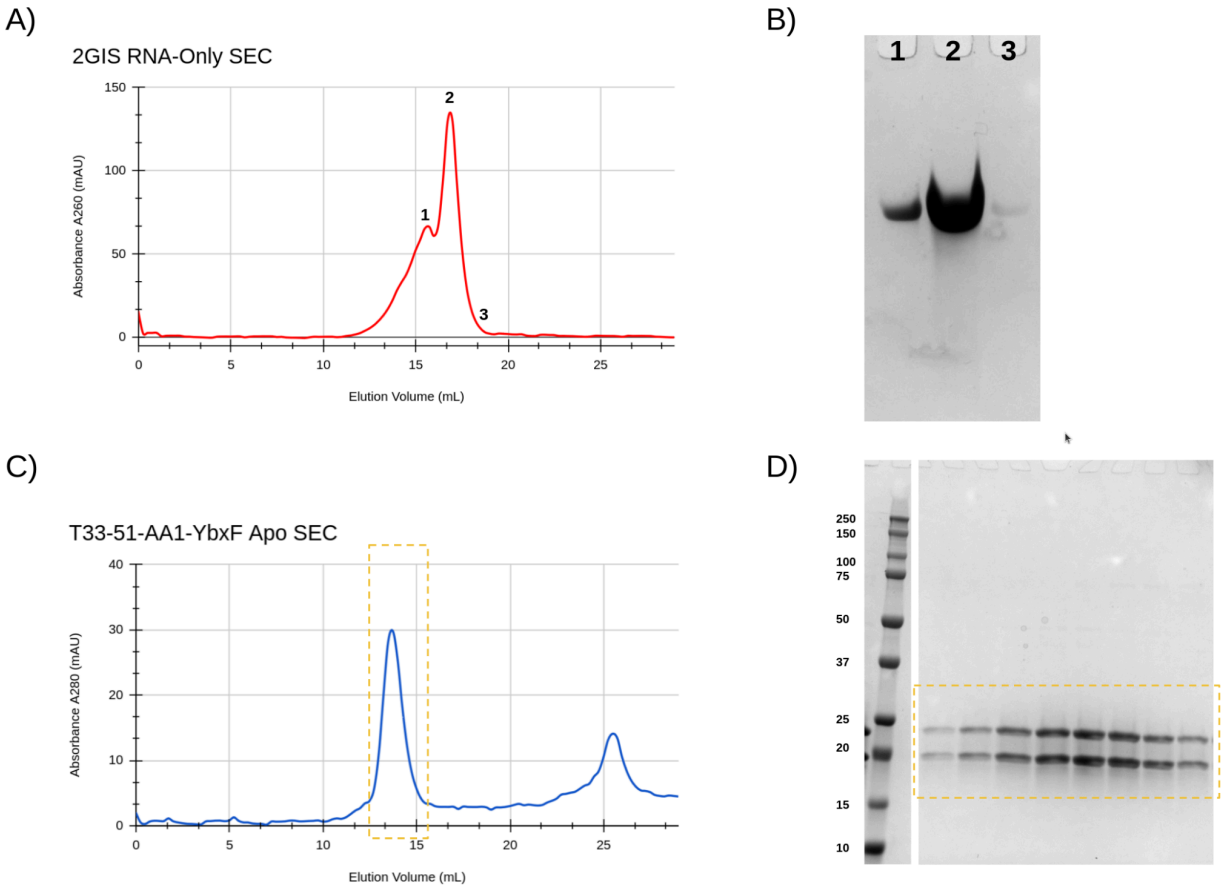


Figure 4.23: RNA and apo scaffold controls. Before mixing experiments, I wanted to verify the elution profiles of both the apo AA1 scaffold and 2GIS RNA on the particular column I would be performing the binding experiments on.

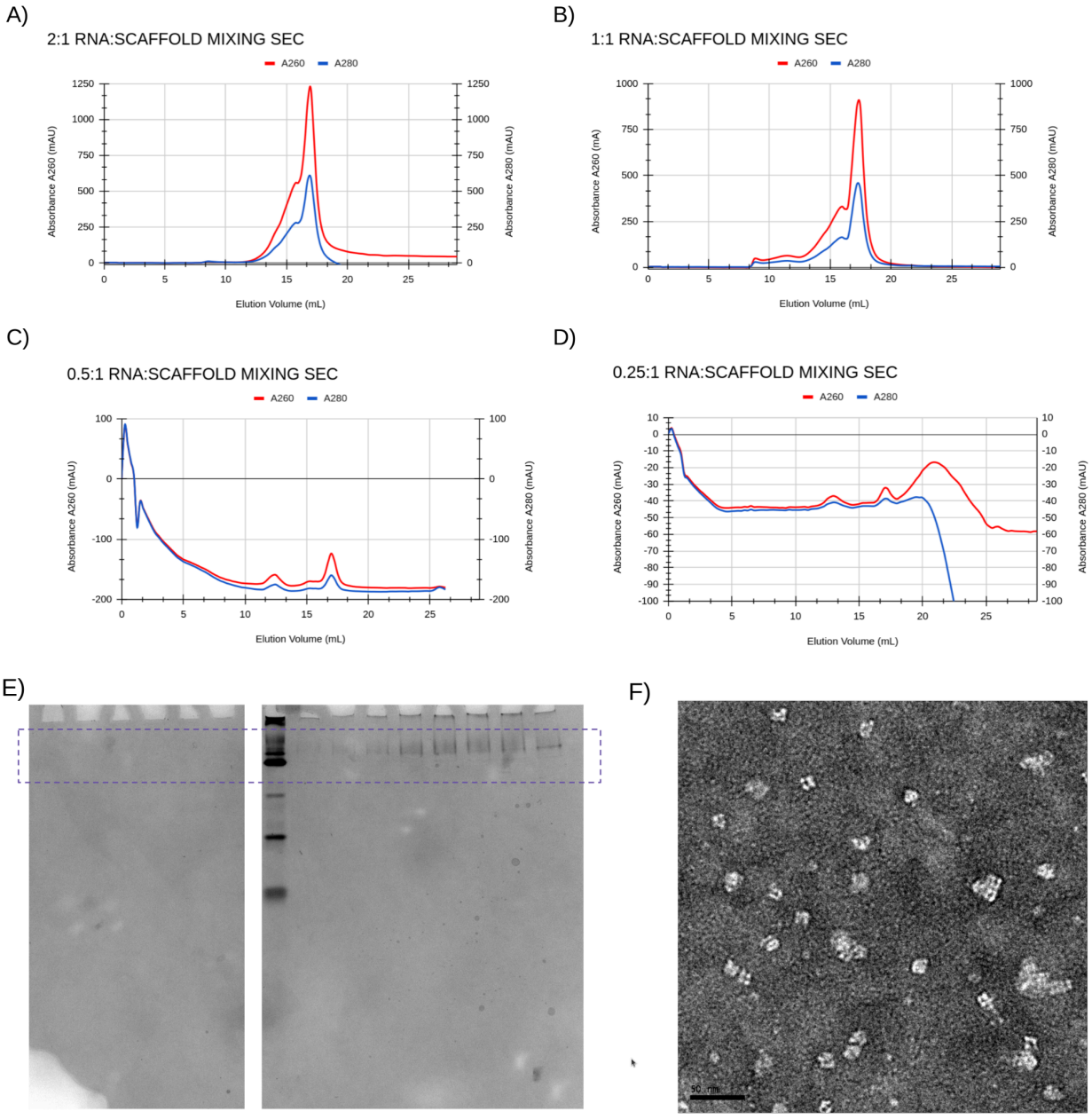


Figure 4.24: RNA-Scaffold binding experiments. To test interactions between the AA1 designed imaging scaffold and the 2GIS RNA cargo, mixing was performed and association was monitored by observing a peak shift of the A260 signal in the SEC chromatograph on a Superose 6 Increase column. A) A 2:1 stoichiometric mixture of RNA:Scaffold was tried at first. No signal was observed for the cage or cage-RNA as all information was obscured by the high A260 signal. B) A 1:1 stoichiometric mixture of RNA:Scaffold was next tried. As with the 2:1 experiment, no clear peak indicating association is observed. A slight void peak and shoulder is evident indicating some degree of aggregation is occurring. C,D) When low ratios of RNA:Scaffold is tried, no noticable shift in the A260 absorbance profile is seen. Some aggregation is observed in these lower concentrations, as evidenced by a larger molecular weight species. E) Fractions corresponding to these sooner-elution bumps were run in duplicate

on a Native PAGE gel where one set was stained with a toluidine blue RNA due (left gel slice) and the other set of fractions were stained with coomassie brilliant blue protein dye (right gel slice). High molecular weight bands thought to correspond to the T33-51-AA1-YbxF scaffold are seen in the protein-stained gel, but signal for the bound RNA is not seen in the same location on the native gel in the RNA-stained portion (checkered box). F) Negative stain EM micrographs from the same small bump showed sparse particles and other aggregates. Clustering of the scaffold was also observed.

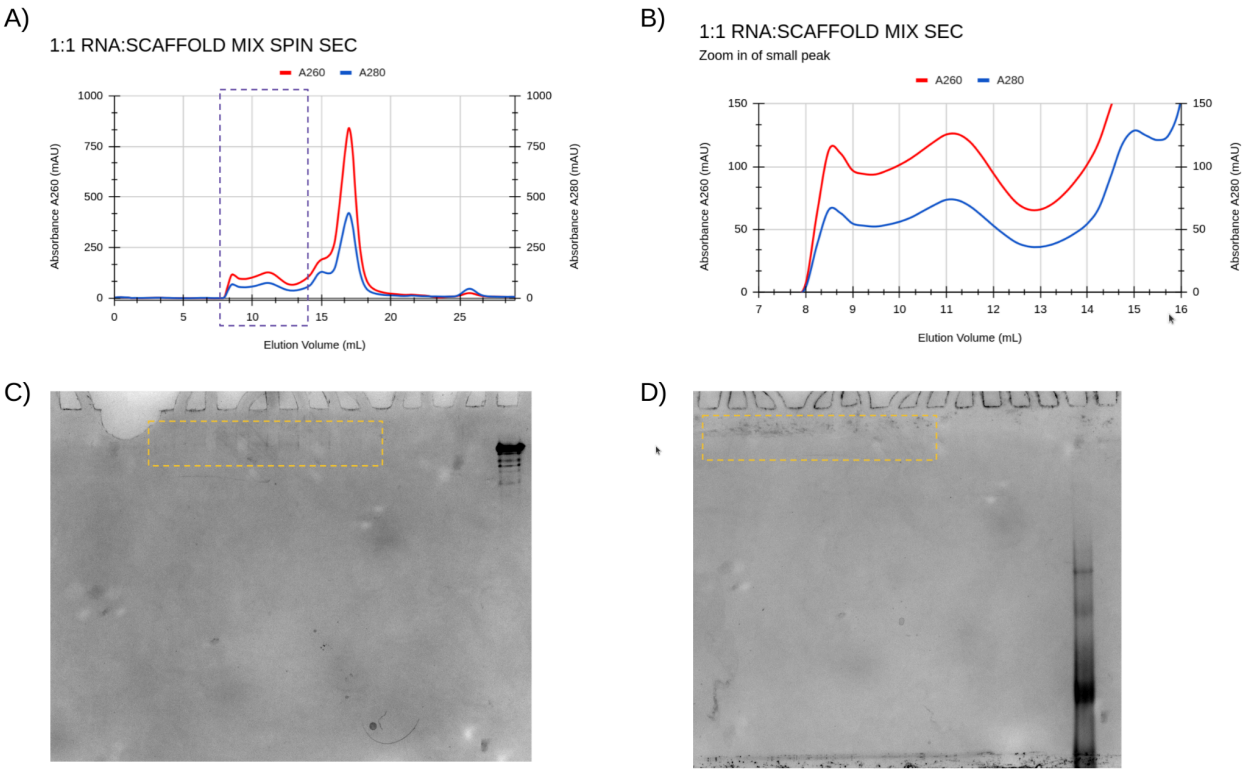


Figure 4.25: Analysis of post-spin SEC mixing experiments. To try to better resolve the RNA-scaffold complex peak, excess free RNA was removed by centrifugation of the mixture using a 100kDa concentrator tube. A) Resulting post-spin SEC chromatogram. Large free RNA peak is still seen. Aggregation is observed by presence of a void peak, followed by a large molecular weight shoulder early in the elution profile. B) Zoom in of the void peak and shoulder. C) Coomassie-stained native PAGE of the small bump shows a faint signal that is thought to correspond to aggregated T33-51-AA1-YbxF (yellow box). Scaffold standard from the original purification is seen in the last lane, some heterogeneity is observed. D) Toluidine blue-stained fractions from the same peak show no evidence of RNA signal (yellow box). Free RNA from the major peak seen in A) was run as a control. Some degree of heterogeneity or degradation is also observed.

CHAPTER FIVE: Design and Characterization of RNA Imaging Scaffold Using Fragment-Based Interface Algorithms

5.1: Background and Significance

The Need for RNA Scaffolds

The field of RNA structural biology has made great strides in recent years, in particular with several striking structures having been solved to moderate resolution in recent years with cryo-EM[1,2]. A more thorough description of some of the advances can be found in Chapter 4.1 of this dissertation. Briefly however, I want to emphasize the importance of the work still to be done. First, a search of the EMBD shows how very little RNA structures have been deposited by cryo-EM (Figure 5.1)[3]. Of the almost 18,000 total structures solved by cryo-EM, only four are of RNA-only samples that have reached high resolution. Second, a lot of the popular RNA's being investigated are the larger targets like riboswitches, segments of viral genomes- the low hanging fruit, or have been heavily engineered so as to not make them appealing targets for robust applications. This leaves the vast majority of biological RNA molecules unreachable by current methods of cryo-EM structure determination[4-6]. To complicate matters further, unlike protein prediction methods, which has been a maturing field dating back to the work of Pauling and colleagues, RNA structure prediction methods have lagged behind[7-11]. This is partially due the fact that RNA molecules are able to adopt multiple different conformations while folding, many of which are of comparable energy, leading to these molecules often getting stuck in kinetic traps which cause the algorithms to improperly predict the true structures[12-14]. On top of this, much of the three-dimensional structure of RNA is governed by long-range interactions between nucleotides far in sequence space, which can be problematic for algorithms that focus

on local interactions[15-17]. I believe that the scaffold-based imaging technology developed in the Yeates lab can still be leveraged to solve this complex problem[18-20].

Challenges with Helical Fusion

Probably the most well known secondary structural element in proteins is that of the alpha-helix. One might remember from undergraduate biochemistry that the basic motif of the helix consists of amino acids arranged in a helical fashion with 3.6 residues per helix turn and is stabilized by directional interactions between backbone carbonyl and amino groups[21]. Additionally, it has been known for some time that the rigidity of these helices is greatly influenced by their amino acid compositions[22]. Because of this, alpha helices are often modeled as cylinders, and are thought to be these solid objects with little sway. Experimental evidence coming from the realm of protein design directly contracts these assumptions. An analysis of cryo-EM structures shows that regions of proteins that exhibit high degrees of motion suffer tremendously in terms of resolution, and it is often domains that contain substantial alpha helices that display these stark motions[23,24]. We have directly observed this phenomenon with our own work regarding the design of imaging scaffolds for cryo-EM analysis[18,19,25]. As shown in Figure 5.2, when fused to the scaffold core, the alpha-helical fusion to the DARPin exhibits a wide range of motion, limiting the ability to achieve atomic detail. To directly combat this, Castells-Graells et al. introduced stabilizing mutations that were able to greatly improve the resolution of docked cargo[20]. A plethora of strategies have been employed over the years, a lot involving the addition of cross-linkers or stabilizing agents[26]. However, I believe that in order to achieve a rigid imaging scaffold capable of achieving atomic-level resolution, we must abandon the alpha-helical fusion approach to protein design in favor of methods that have proven to be more successful[27].

Recent Advances in Interface Design

Recent computational advances have shown tremendous potential as another avenue for designing RNA imaging scaffolds. While early work tended to focus on metrics like buried surface area in designing new interfaces, this caused a lot of unintended consequences from the lack of specificity, leading to aggregation and low success rates[28-30]. Since then, much research has been done into the nature of protein-protein interfaces and those lessons have been applied to new versions of these prediction packages with significant results[31-33]. Even more recently, advances in artificial intelligence and machine learning algorithms have been applied to both the protein folding and protein design problems[34,35]. Concurrently, fragment-based methods developed in the Yeates lab have successfully demonstrated the creation of new symmetric macromolecular assemblies[36,37].

5.2: Results and Discussion

Construct Design - Docking

To begin this new project's designs, a search of the PDB was done to identify suitable building blocks to use for our scaffold. We chose to use a simpler geometry - dihedral, specifically D3, than previous scaffolds as there are a vastly more dihedral structures in the PDB than structures of tetrahedral geometry or higher, and one of this projects goals was to alleviate the added complication of using a synthetic cage core in favor of assemblies found in nature that could be adapted to our purposes of being imaging scaffolds. The focus on the PDB search was to focus on thermophilic structures that could be expressed in *E. coli* because this would give us a robust and stable set of building blocks to start with. In addition the the thermophilic D3 protein's that will serve as the scaffold's core, two RNA-binding proteins, referred to in the design scheme as the C1 component as they lack any symmetry, were chosen as the second building block in the

scaffold that will be responsible for binding and orienting our RNA cargo for imaging. The RNA binding proteins are YbxF (PDB: 3V7E), the same binder used in the helical fusion designs, and U1A (PDB:1NU4), an RNA-binding protein widely used in structural biology studies as a chaperone to aid in crystallization. Following identification of design components, a modified python program was used to perform Nanohedra-style docking procedures to match interface fragments seen in nature to that of fragments on the surface of each of my design components[36,37]. A list of the dihedral assemblies docked are listed in Table 5.1. The 36 identified D3 assembles were docked against the two C1 RNA-binding protein which resulted in an output of the top 5 docked poses for each pair that I visually inspected for steric clashing and suitable designability (Figure 5.3A) before moving forward with the interface residue redesign that will allow the two components to associate when expressed. As part of the visual inspection, we wanted to confirm that the displayed cargo would not be clashing with either the scaffold or with symmetry-related copies of themselves so we generated a framework of dummy atoms were created to represent a series of positions of major parts of bound RNA molecules from a few reported structures (Figure 5.3B). Examples of bad designs that suffered from significant clashing or component overlap is shown in Figure 5.4. After filtering down the docking results, 27 poses passed the eye check.

The last step before moving forward with interface sequence design is to develop a construct scheme. A design component we decided to engineer into the assemblies is the addition of a flexible linker between the two termini of the scaffold subunits. The idea here is not to use a rigid linker to hold one particular component in place, but rather to keep both components in close proximity to each other in solution, therefore lowering the overall entropy and improving the thermodynamic favorability for interface association. Additionally, a hexahistidine and HA tag were added to separate components of the scaffold to enable affinity purification and aid in downstream complex analysis. A summary of the analysis results is shown in Table 5.2.

Interface Redesign - Rosetta

To begin, the pdb files corresponding to selected docked poses were fed into the SymDesign pipeline for sequence redesign[36]. To accomplish the interface residue, two separate protocols were implemented to increase the output sequence diversity. The first protocol is called “Rosetta structure_backbone” which was Rosetta’s original sequence design algorithm and tends to favor more hydrophobic, less natural interfaces[38]. The second protocol, HBNet, is a new software edition to the Rosetta Commons suite and is designed to generate hydrogen bond networks between amino acids along an interface[33]. This protocol has been shown to provide binding specificity, increase protein folding and solubility by creating interfaces that better emulate nature. After successfully running both protocols on my 27 poses, only 13 successfully passed SymDesign’s internal designability checks, but this still resulted in 468 unique sequences for curation. To select a reasonable number of sequences for experimental characterization, I used prior knowledge about interfaces as well as an exhaustive literature search to identify key parameters we thought might be important for a successful interface design. Table 5.3 is taken from Janin et al. and displays the statistics about interfaces seen in nature that was used to guide my selection[39]. The important interface metrics I chose to focus my efforts on were: shape complementarity, buried unsatisfied hydrogen bond density, interface area, and the percent area hydrophobic. High shape complementarity maximizes the contact surface area between interacting proteins, allowing for more extensive non-covalent interactions such as hydrogen bonds, van der Waals forces, and hydrophobic interactions[40]. This results in stronger and more stable protein complexes. Having a low hydrogen bond density is important as natural proteins rarely have unsatisfied hydrogen bonds. Total interface area was an important metric because since we are dealing with such small proteins (YbxF is 8.3kDa and U1A is ~12kDa), we want to select for interfaces with higher total surface area to provide enough binding energy to form a stable complex. Finally, for this project we leaned towards interfaces with a slightly lower

hydrophobic surface area because in our previous endeavors we found that when we leaned too heavily into this parameter, we were left with poor interfaces that lacked specificity to drive binding. Of the 468 sequences resulting from Rosetta redesign, 25 were chosen for the pilot screen. A list of the sequence names, abbreviations, and protocol used is shown in Table 5.4 and four representative designs are shown in Figure 5.5. For initial biochemical characterization, constructs were created without the flexible serine-glycine linker as we wanted to determine whether both of the scaffold components were soluble and stable after the mutations were introduced and if both components were fused from the beginning, it may not be clear if one component isn't folded properly or associating with the other component, or if the stability of one component is pulling the other along. Sequences were submitted to Twist Bioscience for synthesis and cloning into pET28a plasmids.

Biochemical Characterization

Plasmids containing my scaffold constructs were cloned into *E. coli* BL21 gold (DE3) cells for protein production. Expression test cultures were grown in 96-well plates in LB or TB media supplemented with 50µg/mL of kanamycin and expressed at either 37°C for 4 hours or 18°C overnight. Cells were lysed by sonication and expression and solubility levels were assessed by criterion SDS-PAGE gels. A Table of component and assembly molecular weights are shown in Table 5.5. Unfortunately for a vast majority of designs, either no expression or solubility was observed. For several designs, low expression of one subunit, the D3 component, was evident. But for no design, was the C1 RNA-binding component was observed (Figure 5.6). While 96-well screens often serve the purpose of a rapid readout of many protein constructs, a negative result at this stage does not necessarily mean the design is a failure. It is the case that cultures sometimes behave differently when grown in such small volumes, compared to when they are grown at a liter-scale or larger. Before making a judgment call about the designs, I grew

up each design in 50mL and 1L cultures of TB and purified them by NiNTA. Upon analysis of the purification by SDS-PAGE, I was left with the same results (Figure 5.7).

These results sparked me into looking into the designs a little more closely, particularly at the C1 RNA-binding components, and why none of them expressed or were soluble. To focus my efforts, I first began by looking at one particular design, MA003-1nu4_2brx_1_0_hbnet_0081 (abbreviated design C1), as it was one of the designs that showed significant expression in the scaffold core component, but no expression in the RNA-binding component (Figure 5.8A,B). Analysis of the mutations introduced into U1A shows that almost 20% of the entire protein had been mutated over the course of our redesign (Figure 5.8C). This could very well explain why we were seeing no soluble yield of this component. Changing so many residues could very well be affecting how this protein is folding. I tested this by running the sequences through AlphaFold to see if the mutations we introduced are predicted to affect the structure[11]. Figure 5.9 shows analysis of the AlphaFold job run for design C1. As seen from the alignment (Figure 5.9A) and lDDT scores (Figure 5.9B), the mutations introduced are not predicted to affect the global structure of U1A. The failure to associate with the D3 scaffold core must be down to the identity and chemical properties of the amino acids and not the folding process of the protein.

Interface Redesign - MPNN

Around the time that I was getting the results described above, a new machine-learning algorithm, Protein MPNN, was released that had been demonstrated to greatly improve the solubility of proteins compared to traditional design methods[42]. Starting from the same set of docked poses as was done with the Rosetta-based interface design, I fed all 27 PDB files into the new SymDesign-MPNN pipeline. Of the 27 manually-inspected poses, the same 13 passed the internal clashing checks and were green lit for interface redesign. Using MPNN's improved machine learning models, 9 sequences per pose were designed resulting in a total of 117

unique protein sequences for curation. The same interface criteria as described previously (buried surface area, interface hydrophobicity percent, buried unsatisfied hydrogen bonds, shape complementarity, etc) to narrow the list down to 17 sequences for biochemical characterization. A full list of the sequences and properties of the designs is found in Table 5.6. Interestingly, of the 13 unique poses, our selection criteria was only pulling out sequences stemming from the same five D3 components - 1ej2, 1v9l, 1vmd, 1vlh, and 4i4z - which were the same components who showed at least some degree of expression and solubility in the previous round of designs, possibly hunting that these proteins are more amenable to mutation that the others chosen. For this round of testing, I opted to order the designs with their poly-serine-glycine linkers, in hopes that the more robust D3 component might force the less soluble C1 component to be happy in the same way we utilize solubility tags like MBP. Constructs were assembled such that the correct termini could be linked and tagged for purification, and cloned by Twist Biosciences.

Biochemical Characterization

The plasmids containing my designs were cloned into *E. coli* BL21 gold (DE3) cells for protein production. This time around I decided to skip the small scale expression and spend the time to test each construct at a 1 liter scale right off the bat. Cells were seeded with overnight cultures and grown in TB media supplemented with 50µg/mL of kanamycin and grown at 37℃ until OD_{600} reached ~0.8. Protein expression was induced by addition of 1mM IPTG and grown overnight at 18℃ before being pelleted by centrifugation. Cells were lysed, purified by NiNTA affinity chromatography and assessed by SDS-PAGE. Of the 17 constructs tested, one, MPNN5-Link, seemed to show the most promise as indicated by a single band on the gel around the correct molecular weight of ~32kDa (Figure 5.10A) which eluted as a sharp peak as visualized by size exclusion chromatography (Figure 5.10B). Around this time I came to the realization that as the current stage I had no means of assessing whether the two components

were truly interacting via their designed interface or if they were both free floating in solution tethered to each other by this flexible linker. To create a rapid way to test my design hypothesis, I used PCR to add TEV-cleavage sites in the middle of the poly-sg linkers. That way I can take purified protein, cleave it, and run the resulting mixture back over a gel filtration column to test whether the two components are associating. I performed the mutations, grew new TEV-linked versions of MPNN5 and performed the digestion experiments. To my dismay, after incubation of my scaffold with TEV, I saw an appearance of a small peak around where U1A by itself elutes on our Superose 6 Increase column (Figure 5.11). Running both peaks on an SDS-PAGE confirmed by hypothesis - the secondary peak corresponds to the U1A RNA-binding component alone (Figure 5.11D). The larger peak, when analyzed on the gel shows only a single species, which is significantly smaller than the fully-linked construct and is closer to the 19 kDa of the D3 component alone.

Concluding Remarks

Unfortunately as with a vast majority of design projects, it is all about the number of sequences tested and sometimes required brute forcing your way to an answer. While ultimately we were unable to develop an imaging scaffold based on fragment-based design of naturally-occurring components, a few insights can be gleaned from these experiments to be built upon in the future.

I attribute much of the lack of success I was seeing in this project's designs to the low number of starting components. At the onset of the project, we started with 36 thermophilic D3 assemblies and two RNA-binding proteins, all of which got reduced to 13 clash-free poses after input into SymDesign. In all the rounds of interface design, the majority of sequences came from a combination of 4-5 D3 cores. This inherently limits the number of possible designs that could come out of this exercise. Future design effort should expand the number of design components

considered, both for the cores and the RNA-binding proteins. We chose D3 because there is a plethora of structures to choose from, and our choice of 36 was somewhat arbitrary at the start. If I were to perform the docking again, I would increase this pool of candidates first. Additionally, other symmetries could be explored, from other dihedral assemblies - like the D2 aldolase used in Yao et al[25] - to those of higher order (D4, D6, etc.). It has been shown with various other scaffolds that symmetry plays an important role in achieving high resolution reconstructions[18-20]. Our choice of RNA-binding proteins also limited the number of successful designs, and was one of convenience. I chose to start with two known proteins, but there are far more to choose from. A parallel project I was also working on was utilizing RNA-binding peptide segments to try to develop an imaging scaffold[43]. A more thorough dive into the literature on these types of biomolecules may provide more options to create designs from.

While it was clear that MPNN was able to produce designs that expressed better, it ultimately failed to produce successful designs in my case. This might not be totally surprising considering the limited number of poses we started with; design projects often begin with thousands or tens-of-thousands of sequences/poses[36,37,42]. Additionally, while protein MPNN is a great tool for protein design, its purpose was to help make proteins more soluble and increase their yield by making them more stable. In this aim, it was successful by dramatically increasing the solubility of at least one component of the scaffold. Future design efforts should rely on machine learning models like protein MPNN over traditional methods like those in Rosetta; AI will be a powerful tool in the future of protein design and we are only at the beginning.

Finally, while science never ends, one's funding and time in graduate school does. This project was developed towards the end of my PhD training and as such, there are many more avenues of exploration still available for future researchers.

5.3: Materials and Methods

Computational Docking of Natural Assemblies

A modified python program of the Nanohedra docking program described in Laniado et al. and Meador et al. (https://github.com/nanohedra/nanohedra)[36,37] was run between 36 naturally-occurring D3 thermophilic protein assemblies and two RNA-binding proteins, U1A (PDB: 1NU4) and YbxF (PDB: 3V7E). A brief description of Nanohedra is as follows. As an input, the algorithm takes two symmetric oligomers (C, D, etc.), oligomers A and B. The surface of oligomer A is subdivided into three amino acid residue fragments. These fragments are then searched against a library fragments and fragment pairs mined from the PBD of known interfaces seen in nature. Matching fragments then aligned to the surface of oligomer A such that it is now decorated with “ghost fragments” from natural interfaces. These ghost fragments are then used to guide docking and pose generation between oligomers A and B, where metrics are generated and subsequently analyzed and curated. The modification done to Nanohedra involves the capability of docking a symmetric oligomer against a non-symmetric protein (C1) since the RNA-binding proteins lack symmetry. The program outputs the five highest scoring poses. Some of the metrics influencing the score of each pose are the number of fragment pairs, R.M.S.D. overlap between fragments, and C-alpha backbone clashes between components. Manual curation of poses involved checking for potential clashes between symmetrically related copies of the RNA-binding proteins or the resulting RNA cargo when displayed in PyMOL[44].

Computational Interface Redesign

The following is in regards to the Rosetta-based interface design. 27 PDB files corresponding to the selected D3-C1 docked poses were moved to a separate directory for interface redesign

protocols. Using the SymDesign environment, two Rosetta design protocols were used to generate new sequences, Structutre_background (command: python $SymDesign interface_design --symmetry D3:{D3}{C1} --directory /input/pdb/pose/directory --structure_background) and HBNet (command: python $SymDesign interface_design --symmetry D3:{D3}{C1} --directory /input/pdb/pose/directory --hbnet). Of the 27 selected poses, only 13 passed the internal SymDesign metric checks, but still was able to produce 468 unique protein sequences. Using information about interfaces in the literature, 25 sequences were chosen for biochemical characterization. Analysis of each component's termini was conducted to allow addition of tag's and linkers downstream. Genes were ordered and cloned from Twist Biosciences into pET28a plasmids for protein production.

The following is in regards to the MPNN-based interface design. The same 27 pose PDB files used in the Rosetta-based designs were used in this second round of protein design. The new sequences were generated in the SymDesign environment that has been adapted to incorporate MPNN machine learning algorithms, 117 unique sequences were generated (command: python $SymDesign interface-design -d /input/pdb/pose/directory --structures --temperatures 0.1 0.2 0.5 --number-of-trajectories 3 --sym-entry 161 --symmetry D3:{D3}{C1} --no-evolution-constraint --preprocessed). Using the same criteria as the Rosetta-based designs, 17 new sequences were chosen for biochemical analysis. Genes encoding the constructs were ordered and cloned from Twist Biosciences into pET28a plasmids.

Biochemical Characterization

Plasmids containing our designs were transformed into BL21 Gold(DE3) *E. coli* strains for protein production. Small scale expression screens were carried out in 1mL cultures at either 37°C or 18°C following the protocol outlined in Knaust et al[45] in LB media supplemented with 50µg/mL kanamycin. Larger volume cultures were either grown in 50mL or 1L volumes of TB

supplemented with 50μg/mL kanamycin and grown to saturation at 37°C before lowering the temperature to 18°C. A final concentration of 0.5mM IPTG was added and protein production was allowed to continue overnight. Cells were harvested by centrifugation and resuspended in lysis buffer (50mM Tris 8.0, 250mM NaCl, 20mM Imidazole) at a ratio of ~3-4mL of buffer per gram of cell pellet. Cells were lysed by 2-3 passages through an Elmusiflex until cell disruption was complete. Proteins were purified either by gravity column chromatography using NiNTA-conjugated resin or using a 5mL GE HiTrap and eluted with a linear gradient. Resulting fractions were analyzed by SDS-PAGE and native PAGE.

Negative Stain Electron Microscopy

5uL of 0.05mg/mL purified cages were deposited on a formvar supported carbon film on 300-mesh copper grid that has been negatively glow discharged for 30secs. The excessive sample was blotted away with filter paper after 1 minute, washed twice with nanopure water and stained with 2% uranyl acetate for 30 sec. Grids were allowed to air dry before being imaged at room temperature with FEI Tecnai T12, FEI Tecnai TF20 and Talos F200C electron microscopes.

5.4: References

[1] Liu, D., Thélot, FA., Piccirilli, JA., Liao, M., Yin, P. *Sub-3-Å cryo-EM structure of RNA enabled by engineered homomeric self-assembly.* Nat Methods **19:** 576–585 (2022)

[2] Zhang, K., Li, S., Kappel, K., Pintilie, G., Su, Z., Mou, TC., Schmid, MF., Das, R., Chiu, W. *Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution.* Nat. Commun. **10:** 5511 (2019)

[3] Ma, H., Jia, X., Zhang, K., Su, Z. *Cryo-EM advances in RNA structure determination.* Sig. Transduct. Target. Ther. **7:** 58 (2022)

[4] Kiss, T. *Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions*. Cell. **109(2):** 145-148 (2002)

[5] Bartel, DP., *MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.* Cell. **116(2):** 281-297 (2004)

[6] Ponting, CP., Oliver, PL., Reik, W. *Evolution and Functions of Long Noncoding RNAs*. Cell. **136(4):** 629-641 (2009)

[7] Pauling, L., Corey, RB., & Branson, HR. *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain.* Proc Natl Acad Sci. **37(4):** 205-2011 (1951)

[8] Simons, KT., Kooperberg, C., Huang, E., Baker, D. *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol. **268(1):** 209-225 (1997)

[9] Jones, DT., Taylor, WR., Thornton, JM. *A new approach to protein fold recognition.* Nature. **258(6381):** 886-89 (1992)

[10] Bradley, P., Misura, KMS., Baker, D. *Toward high-resolution de novo structure prediction for small proteins.* Science. **309(5742):** 1868-1871 (2005)

[11] 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, SAA., Ballard, AJ., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholaska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, AW., Kavukcuoglu, K., Kohli, P., Hassabis, D. *Highly accurate protein structure prediction with AlphaFold.* Nature. **596:** 583-589 (2021)

[12] Treiber, DK., Williamson, JR. *Exposing the kinetic traps in RNA folding.* Curr Opin Struct Biol. **9(3):** 339-345 (1999)

[13] Chen, SJ., *RNA Folding: Conformational Statistics, Folding Kinetics, and Ion Electrostatics*. Annu Rev Biophys. **37:** 197-214 (2008)

[14] Woodson, SA., *Metal ions and RNA folding: a highly charged topic with a dynamic future.* Curr Opin Chem Biol. **9(2):** 104-109 (2005)

[15] Ding, F., Sharma, S., Chalasani, P., Demidov, VV., Broude, NE., and Dokholyan, NV. *Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanics.* RNA. **14(6):** 1164-1173 (2008)

[16] Bonilla, SL., Vicens, Q., Kieft, JS. *Cryo-EM reveals an entangled kinetic trap in the folding of a catalytic RNA.* Sci Adv. **8(34)** (2022)

[17] Watkins, AM., Rangan, R., Das, R. *FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds.* Structure. **28(8):** 963-976 (2020)

[18] Liu Y, Gonen S, Gonen T, and Yeates TO Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. PNAS. 115(13): 3362–3367 (2018)

[19]] Liu, Y., Huyng, D., Yeates, TO. *A 3.8*Å *resolution cryo-EM structure of a small protein bound to an imaging scaffold.* Nature Comm. **10(1):** 1864 (2019)

[20] Castells-Graells R., Meador K., Arbing MA., Sawaya MR., Gee M., Cascio D., Gleave E., Debreczeni JÉ., Breed J., Leopold K., Patel A., Jahagirdar D., Lyons B., Subramaniam S., Phillips C., Yeates TO. *Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold.* Proc Natl Acad Sci USA. **120(37)** (2023)

[21] Eisenberg, D. *The discovery of the α-helix and □-sheet, the principal structural features of proteins.* Proc Natl Acad Sci USA. **100(20):** 11207-11210 (2003)

[22] Sivaramakrishnan, S., Spink, BJ., Sim, AYL., Doniach, S., Spudich, JA. *Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif*. Proc Natl Acad Sci USA. **105(36):** 13356-61 (2008)

[23] Dou, H., Burrows, DW., Baker, ML., Ju, T. *Flexible Fitting of Atomic Models into Cryo-EM Density Maps Guided by Helix Correspondences*. Biophys J. **112(12):** 2479-2493 (2017)

[24] *Mapping the motion and structure of flexible proteins from cryo-EM data.* Nature Methods. **20(6):** 797-798 (2023)

[25] Yao Q., Weaver SJ., Mock JY., and Jensen GJ. *Fusion of DARPin to Aldolase Enables Visualization of Small Protein by Cryo-EM.* Structure. **27(7):** 1148–1155 (2019)

[26] Jeong, WH., Lee, H., Song, DH., Eom, JH., Kim, SC., Lee, HS., Lee, H., and Lee, JO. *Connecting two proteins using a fusion alpha helix stabilized by a chemical crosslinker*. Nat Commun. **7:** 11031 (2016).

[27] Kung, JE., Johnson, MC., Jao, CC., Arthur CP. *Disfulfide constrained Fabs overcome target size limitation for high-resolution single-particle cryo-EM.* Biorxiv doi.org/10.1101/2024.05.10.593593. (2024)

[28] Leaver-Fay, A., Tyka, M., Lewis, SM., Lange, OF., Thompson, J., Jacak, R., Kaufman, K., Renfrew, PD., Smith, CA., Sheffler, W., Davis, IW., Cooper, S., Treuille, W., Mandell, DJ., Richter, F., Ban, YEA., Fleishman, SJ., Corn, JE., Kim, DE., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, JJ., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, JJ.,

Kuhlman, B., Baker, D., Bradley, P. *ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules.* Methods in Enzymology. **487:** 545-574 (2011)

[29] Sinclair, JC., Davies, KM., Vénien-Bryan, C. & Noble, ME. *Generation of protein lattices by fusing proteins with matching rotational symmetry.* Nature Nanotechnology **6(558)** (2011)

[30] Shen, H., Fallas, JA., Lynch, E., Sheffler, W., Parry, B., Jannetty, N., Decarreau, J., Wagenbach, M., Vicente, JJ., Chen, J., Wang, L., Dowling, Q., Oberdorfer, G., Stewart, L., Wordeman, L., De Yoreo, J., Jacobs-Wagner, C., Kollman, J., and Baker, D. *De novo design of self-assembling helical protein filaments.* Science. **362(6415):** 705-709 (2018)

[31] Levy, ED. *A simple definition of structural regions in proteins and its use in analyzing interface evolution.* J Mol Biol. **403(4):** 660-670 (2010)

[32] Larsen, TA., Olson, AJ., Goodsell, DS. *Morphology of protein-protein interfaces.* Structure. **6(4):** 421-427 (1998)

[33] Cannon, KA., Park, RU., Boyken, SE., Natterman, U., Yi, S., Baker, D., King, NP., Yeates, TO. *Design and structure of two new protein cages illustrate successes and ongoing challenges in protein engineering.* Protein Sci. **29(4):** 919-929 (2020)

[34] Watson, JL., Juergens, D., Bennett, NR. et al. *De novo design of protein structure and function with RFdiffusion.* Nature. **620:** 1089–1100 (2023)

[35] Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Taniguchi, R., Turonova, B., Zimmerli, CE., Buczak, K., Schmidt, FH., Margiotta, E., Mackmull, MT., Hagen, WJH., Hummer, G., Kosinski, J., Beck, M. *AI-based structure prediction empowers integrative structural analysis of human nucleopores.* Science. **376(6598)** (2022)

[36] Laniado, J., Meador, K., Yeates, TO. *A fragment-based protein interface design algorithm for symmetric assemblies*. Protein Eng Des Sel. **34** (2021)

[37] Meador, K., Castells-Graells, R., Aguirre, R., Sawaya, MR., Arbing, MA., Sherman, T., Senarathne, C., Yeates, TO. *A suite of designed protein cages using machine learning and protein fragment-based protocols.* Structure. **24** (2024)

[38] Kuhlman, B. *Designing protein structures and complexes with the molecular modeling program Rosetta*. J Biol Chem. **294(50):** 19436-19443 (2019)

[39 Janin, J., Bahadur, RP., Chakrabarti, P. *Protein-protein interaction and quaternary structure.* Q Rev Biophys. **41(2):** 133-180 (2008)

[40] Jones, S., & Thornton, JM. *Principles of protein-protein interactions.* Proceedings of the National Academy of Sciences. **93(1):** 13-20 (1996).

[41] Mariani, V., Biasini, M., Barbato, A., and Schwede, T. *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.* Bioinformatics. **29(21):** 2722-2728 (2013)

[42] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, RJ., Milles, LF., Wicky, BIM., Courbet, A., de Haas, RJ., Bethel, N., Leung, PJY., Huddy, TF., Pellock, S., Tischer, D., Chan, F., Keopnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, AK., King, NP., and Baker,D. *Robust deep learning based protein design using ProteinMPNN*. Science. **378(6615):** 49-65 (2022)

[43] Das, C., and Frankel, AD. *Sequence and structure space of RNA-binding peptides.* Biopolymers. **70(1):** 80-85 (2003)

[44] Schrödinger, L., & DeLano, W. (2020). *PyMOL*. Retrieved from http://www.pymol.org/pymol

[45] Knaust, RK., Nordlund, P. *Screening for soluble expression of recombinant proteins in a 96-well format.* Annual Rev Biochem. **297(1):** 79-85 (2001).

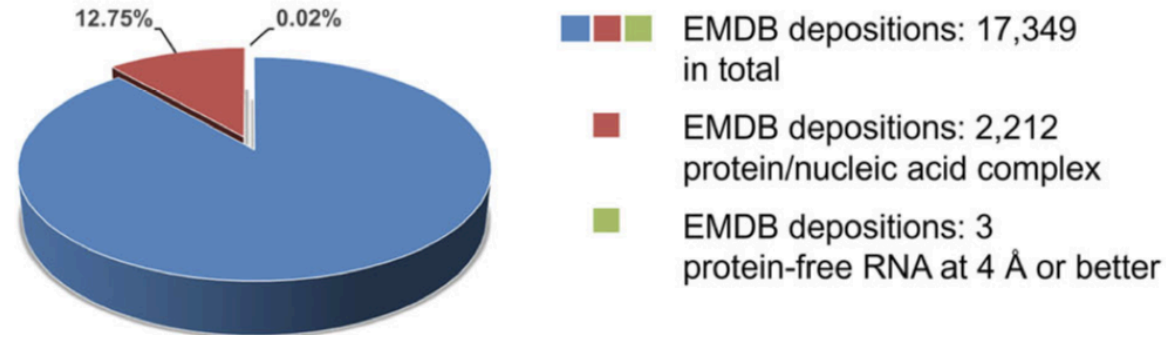


Figure 5.1: Number of solved RNA structures by cryo-EM. Figure adapted from Ma et al.

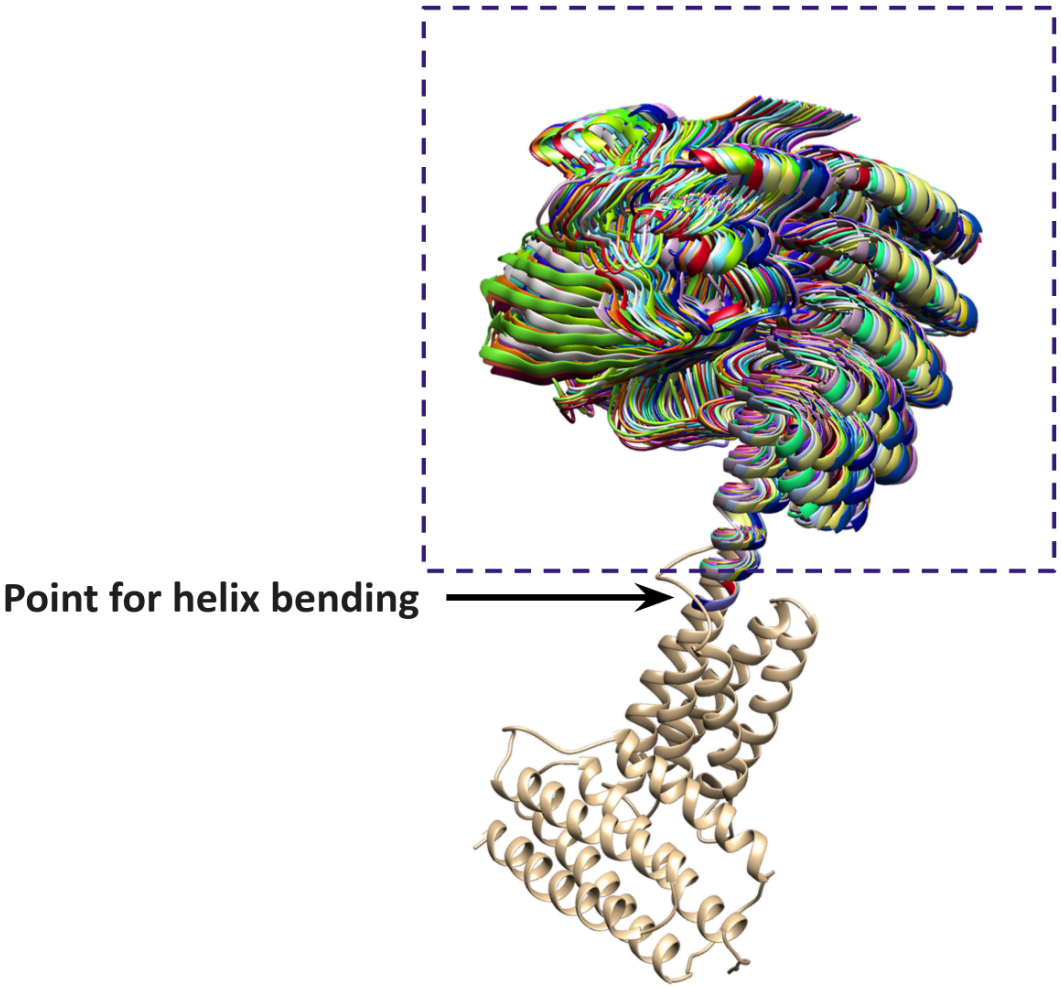


Figure 5.2: Analysis of helix flexibility in imaging scaffolds. To explain the stark drop-off in resolution observed at the periphery of the fused DARPin in terms of the helix flexibility, the design backbone was relaxed degrees of freedom of the helix were sampled in real space. The output files were aligned in relation to the scaffold core, allowing for the motion of the fused domain to be observed. Figure adapted from Castell-Graells et al. (unpublished).

List of D3 Scaffold Components and Description	
PDB Identifier	Protein Description
1ej2	*Methanobacterium thermoautotrophicum* nicotinamide mononucleotide adenylyltransferase
1od6	Phosphopantetheine adenylyltransferase from *Thermus thermophilus*
1odk	Purine nucleoside phosphorylase from *Thermus thermophilus*
1t57	Protein MTH1675 from *Methanobacterium thermoautotrophicum*
1twl	Inorganic pyrophosphatase from Pyrococcus furiosus
1v1a	2-Keto-3-Deoxygluconate Kinase from *Thermus thermophilus*
1v9l	L-glutamate dehydrogenase from *Pyrobaculum islandicum*
1vgg	Conserved hypothetical protein TTHA1091 from *Thermus thermophilus*
1vlh	Phosphopantetheine adenylyltransferase (TM0741) from *Thermotoga maritima*
1vlg	Ferritin (TM1128) from *Thermotoga maritima*
1vmd	Methylglyoxal synthase (TM1185) from *Thermotoga maritima*
1vrg	Propionyl-CoA carboxylase, beta subunit (TM0716) from *Thermotoga maritima*
1wo8	Methylglyoxal synthase from Thermus thermophilus
1wvq	Conserved hypothetical protein PAE2307 from *Pyrobaculum aerophilum*
1wz8	Probable Enoyl-CoA Dehydratase from *Thermus thermophilus*
1xqi	NDP kinase from Pyrobaculum aerophilum
1xx7	Hypothetical protein from *Pyrococcus furiosus*
2af7	Gamma-carboxymuconolactone decarboxylase from *Methanobacterium thermoautotrophicum*
2brx	UMP Kinase from *Pyrococcus furiosus*

2bty	Acetylglutamate kinase from *Thermotoga maritima*
2cwq	Conserved protein TTHA0727 from *Thermus thermophilus*
2d16	PH1918 protein from *Pyrococcus horikoshii* OT3
2ef4	Arginase from *Thermus thermophilus*
2eis	Acyl-CoA hydrolase-like protein, TT1379, from *Thermus thermophilus*
2i1o	Nicotinate Phosphoribosyltransferase from *Thermoplasma acidophilum*
2yyb	TTHA1606 from *Thermus thermophilus* HB8
3aog	Glutamate dehydrogenase (GdhB) from *Thermus thermophilus*
3bey	protein O27018 from *Methanobacterium thermoautotrophicum*
3fcy	Acetyl Xylan Esterase 1 from *Thermoanaerobacterium*
3pzl	agmatine ureohydrolase of *Thermoplasma volcanium*
3q46	Inorganic pyrophosphatase from *Thermococcus thioreducens*
3t3w	probable enoyl-COA hydratase from *Mycobacterium thermoresistibile*
3ug3	Alpha-L-arabinofuranosidase from *Thermotoga maritima*
3ubb	GlpG
4i4z	1,4-dihydroxy-2-naphthoyl-coenzyme A synthase (MenB)
4jyl	Enoyl-CoA hydratase from Thermoplasma volcanium

Table 5.1: List of Nature D3 assemblies used for docking. To begin design work, design components were chosen by an advance search through the PDB. The search parameters included things such as D3 assembly state, soluble expression in *E. coli*, and large enough size to be useful as an imaging scaffold core. All proteins chosen were originally from thermophilic organisms as they tend to be more stable, since they evolved to exist in environmental extremes, and better starting points from a design standpoint.

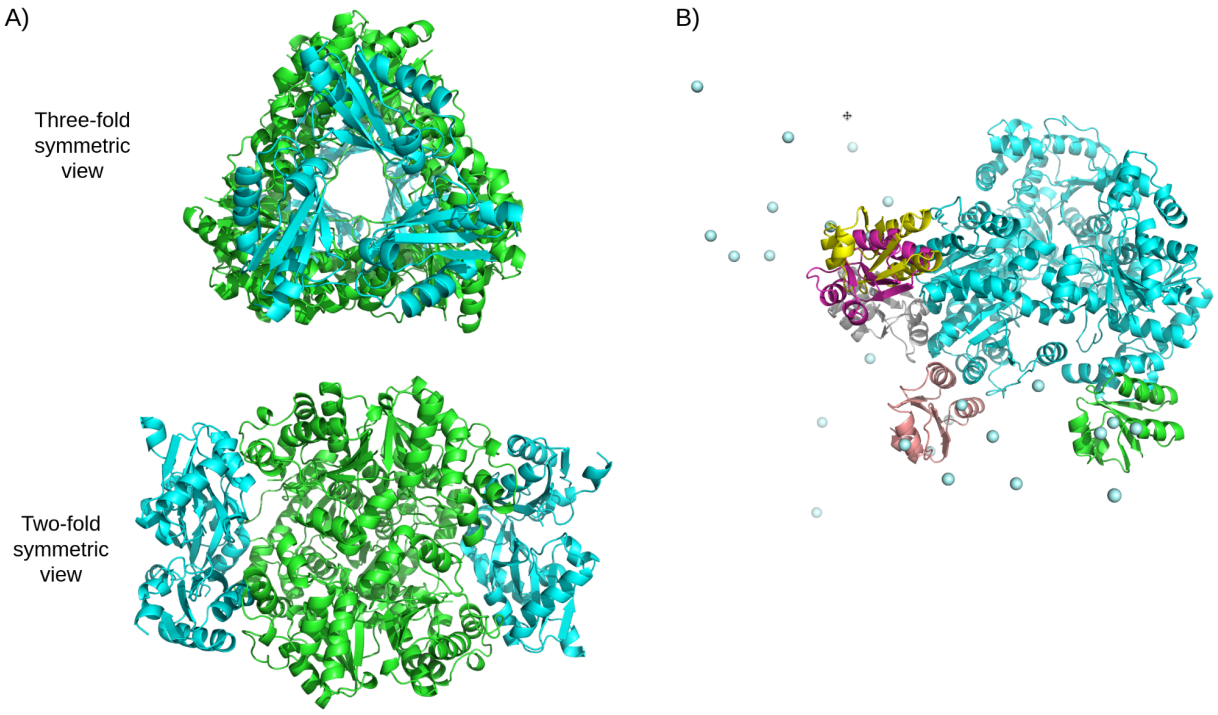


Figure 5.3: Visual inspection of docking results. After our python program finished its docking procedures, visual inspection of the top 5 scoring hits was performed to ensure no clashing is occurring and that all assemblies make sense. A) Expanded assembly of one particular docking combination between the D3 (PDB: 1EJ2) and C1 U1A (PDB: 1NU4). B) Composite of multiple docking outputs. The D3 scaffold core chains are shown in cyan, and the various docked poses of the U1A are shown in green, salmon, magenta and yellow respectively. Dummy atoms representing potential bound RNA orientations are shown as balls.

A)

B)

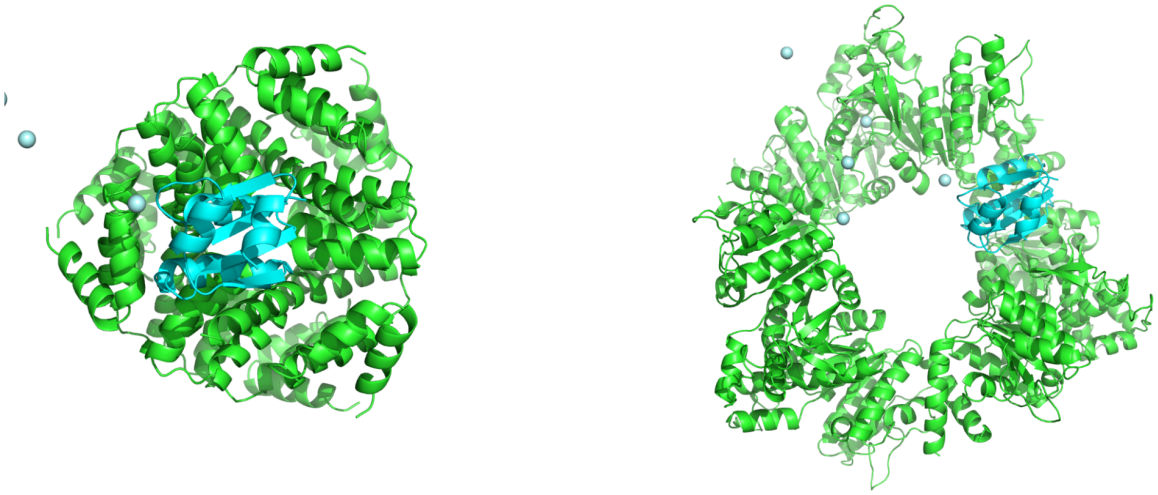


Figure 5.4: Examples of bad docking poses. Only a small subset of the docked poses will be suitable to design into imaging scaffolds. The ones that fail in the design stage typically fail due to clashes between protein subunits, or between RNA cargo. A) An example of a design that failed due to subunit clashes. The D3 scaffold core is shown in green. In cyan, is one C1 RNA binding protein docked onto the cage core. The orientation of docking places one C1 protein at the three-fold interface of the D3 component, which will cause clashes with the two symmetry-related components when the full design is expressed. B) An example of a design that failed due to RNA clashes. The hypothetical RNA cargo is depicted with the blue balls. One RNA-binding protein is shown in cyan docked to the D3 cage core (green). When fully expressed and loaded with RNA, the cargo will clash with the RNA-binding protein on the adjacent, symmetry-related scaffold chain.

Design_Name	D3 PDB	U1A PDB	D3 Chain Aligned	Termini Type	Termini Distance	Termini Type	Termini Distanc	# Fragments
1nu4_1vrg_1_0	1vrg_1	1nu4	A	U1A_C::D3_N	9.7	N/A	N/A	8
1nu4_1vrg_1_1	1vrg_1	1nu4	A	U1A_C::D3_N	9.7	NA	N/A	8
1nu4_1vrg_1_2	1vrg_1	1nu4	A	U1A_C::D3_N	10.1	N/A	N/A	8
1nu4_2af7_1_0	2af7_1	1nu4	B	U1A_C::D3_N	19.3	N/A	N/A	7
1nu4_2af7_1_1	2af7_1	1nu4	F	N/A	N/A	D3_C::U1A_N	13.1	6
1nu4_2af7_1_3	2af7_1	1nu4	B	U1A_C::D3_N	5.4	N/A	N/A	6
1nu4_2af7_1_4	2af7_1	1nu4	C	N/A	N/A	D3_C::U1A_N	16.7	6
1nu4_3bey_1_0	3bey_1	1nu4	A	U1A_C::D3_N	9.2	N/A	N/A	7
1nu4_3bey_1_3	3bey_1	1nu4	C	U1A_C::D3_N	18.9	N/A	N/A	6
1nu4_3bey_1_4	3bey_1	1nu4	F	U1A_C::D3_N	13.1	N/A	N/A	6
1nu4_1v9l_1_0	1v9l_1	1nu4	F	U1A_C::D3_N	19.2	N/A	N/A	7
1nu4_2ef4_1_0	2ef4_1	1nu4	D	U1A_C::D3_N	20	N/A	N/A	7
1nu4_2ef4_1_1	2ef4_1	1nu4	D	U1A_C::D3_N	19.7	N/A	N/A	7
1nu4_4jyl_1_0	4jyl_1	1nu4	A	U1A_C::D3_N	19.9	N/A	N/A	7
1nu4_4jyl_1_2	4jyl_1	1nu4	E	U1A_C::D3_N	14.3	N/A	N/A	5
1nu4_2brx_1_0	2brx_1	1nu4	A	U1A_C::D3_N	11.8	D3_C::U1A_N	18.6	7
1nu4_3t3w_1_2	3t3w_1	1nu4	B	U1A_C::D3_N	16.4	N/A	N/A	6
1nu4_ni4z_1_0	ni4z_1	1nu4	E	U1A_C::D3_N	16.6	N/A	N/A	6
1nu4_ni4z_1_2	ni4z_1	1nu4	D	U1A_C::D3_N	19.6	N/A	N/A	5
1nu4_1wvq_1_0	1wvq_1	1nu4	D	U1A_C::D3_N	15.6	N/A	N/A	6
1nu4_1ej2_1_1	1ej2	1nu4	E	N/A	N/A	D3_C::U1A_N	15.6	6
1nu4_2yyb_1_1	2yyb_1	1nu4	A	U1A_C::D3_N	17.2	N/A	N/A	6
1nu4_2yyb_1_4	2yyb_1	1nu4	A	N/A	N/A	D3_C::U1A_N	13.3	5
1nu4_1odk_1_4	1odk_1	1nu4	E	U1A_C::D3_N	19.2	N/A	N/A	6
1nu4_1vmd_1_1	1vmd_1	1nu4	A	U1A_C::D3_N	13.5	D3_C::U1A_N	19.2	6
1nu4_1vmd_1_2	1vmd_1	1nu4	F	U1A_C::D3_N	18.1	N/A	N/A	5
1nu4_1vmd_1_3	1vmd_1	1nu4	B	U1A_C::D3_N	19.9	N/A	N/A	5
1nu4_1vmd_1_4	1vmd_1	1nu4	A	U1A_C::D3_N	17.1	N/A	N/A	5
1nu4_1vlq_1_4	1vlq_1	1nu4	A	N/A	N/A	D3_C::U1A_N	4.8	6
1nu4_2bty_1_3	2bty_1	1nu4	D	N/A	N/A	D3_C::U1A_N	16.4	6
1nu4_2cwq_1_1	2cwq_1	1nu4	A	N/A	N/A	D3_C::U1A_N	18.9	6
1nu4_2cwq_1_2	2cwq_1	1nu4	E	U1A_C::D3_N	14.5	N/A	14.5	6
1nu4_2cwq_1_4	2cwq_1	1nu4	A	N/A	N/A	D3_C::U1A_N	18.5	5
1nu4_1xx7_1_4	1xx7_2	1nu4	D	U1A_C::D3_N	19.2	D3_C::U1A_N	17	5
1nu4_2eis_1_0	2eis_1	1nu4	A	U1A_C::D3_N	16.4	N/A	N/A	5
1nu4_2eis_1_1	2eis_1	1nu4	A	U1A_C::D3_N	17.5	N/A	N/A	5
1nu4_2eis_1_2	2eis_1	1nu4	D	U1A_C::D3_N	15.4	D3_C::U1A_N	18.8	5
1nu4_2eis_1_3	2eis_1	1nu4	C	U1A_C::D3_N	17.8	N/A	N/A	4
1nu4_2eis_1_4	2eis_1	1nu4	D	U1A_C::D3_N	18.2	D3_C::U1A_N	19.7	4

Output Score	Calpha Clashes	Symmetry Clashes	Cargo Clashes	Design Rating
12.25	0	No Obvious	No room for cargo	Bad
12.21	0	No Obvious	No room for cargo	Bad
11.35	0	No Obvious	No room for cargo	Bad
10.94	1	No Obvious	Potential clashes	Bad
9.03	1	Clashes	Potential clashes	Bad
8.99	2	No Obvious	No Obvious Clashes	Good
8.46	0	No Obvious	No Obvious Clashes	Good
10.81	2	No Obvious	Potential clashes	Bad
8.41	0	No Obvious	Potential clashes	Bad
8.28	1	Potential Stercs on inner side	Clashes	Bad
10.66	0	No Obvious	No obvious Clashes	Good
10.34	0	No Obvious	Potential clashes	Maybe
10.29	0	No Obvious	Potential clashes	Maybe
10.29	0	Potential Clashes subunits on othe	Potential clashes	Bad
7.64	0	Potential Clashes subunits on othe	Potential clashes	Bad
9.78	1	No Obvious	No obvious Clashes	Good
9.56	0	No Obvious	Potential clashes	Bad
9.49	1	Potential Clashes subunits on othe	Potential Clashes	Bad
7.64	0	Potential Clashes subunits on othe	Potential Clashes	Bad
9.13	0	No Obvious	Potential Clashes	Bad
8.71	2	No Obvious	No obvious Clashes	Good
8.64	1	No Obvious	Potential Clashes	Good
7.81	1	No Obvious	No obvious Clashes	Good
8.57	0	Potential Clashes	Potential Clashes	Bad
8.47	0	No Obvious	Potential Clashes	Bad
7.09	2	No Obvious	Potential Clashes	Bad
7	0	No Obvious	No obvious Clashes	Good
6.8	1	No Obvious	Clashes with cage	Bad
8.43	0	No Obvious	No Obvious Clashes	Good
8.39	0	No Obvious	No Obvious Clashes	Good
8.38	0	No Obvious	No obvious Clashes	Good
8.17	0	No Obvious	No obvious Clashes	Good
7.1	0	No Obvious	No obvious Clashes	Good
7.5	0	No Obvious	Cargo Clashes	Bad
7.3	0	No Obvious	Potential Clashes	Maybe
6.96	0	No Obvious	No obvious Clashes	Good
6.65	0	No Obvious	No obvious Clashes	Good
5.85	0	No Obvious	No obvious Clashes	Good
5.55	0	No Obvious	No obvious Clashes	Good

Table 5.2: Designability of docked poses. The output from the D3-C1 docking combinations were visually checked and summarized in the above table. Special consideration was given to which component termini could be used for purification tags and which could be utilized for the flexible linker.

Table 2. *Properties of protein–protein interfaces*

Parameter	Protein–protein complexes[a]	Homodimers[b]		Weak dimers[c]	Crystal packing[d]
		Bahadur	Dey		
Number in dataset	70	122	276	19	188
BSA (Å^2)	1910	3900	3700	1620	570, 1510
(S.D.)	(760)	(2200)	(2160)	(670)	(520)
Amino acids per interface	57	104	100	50	48
BSA (Å^2) per amino acid	34	38	37	32	32
Composition (BSA %)					
Non-polar	58	65	65	62	58
Neutral polar	28	23	22	25	25
Charged	14	12	13	13	17
Atomic packing					
f_{bu} (buried atoms %)	34	36	35	28	21
L_D packing index	42	45	43	34	32
S_c complementarity score	0·69	0·70			0·63
R_p propensity score[e]	0·9	4·3	2·1	0·5	−1·1
Chain segments[f]	5·6	3·4	3·2	5·8	6·3
H bonds					
n_{HB} (number per interface)	10	19	18	7	5
BSA per bond (Å^2)	190	210	209	230	280
Water molecules[g]					
Number per interface	20	44			23
Number per 1000 Å^2	10	11			15
Bridging H bonds	6	13			6
Residue conservation[h]					
% in core	55	60			40
s in core and rim	0·65 and 0·80	0·63 and 0·77			0·98 and 0·99

[a]Data from Chakrabarti & Janin (2002) on a subset of the complexes of Lo Conte *et al.* (1999).

[b]Data from Bahadur *et al.* (2003); the set of Dey *et al.* (unpublished data) was derived from the PiQSi database (Lévy, 2007) as described in the text.

[c]Homodimers described in PiQSi (Lévy, 2007) as being in equilibrium with the monomer, based on the literature.

[d]Pairwise interfaces in crystals of monomeric proteins. The first mean BSA value and the S.D. are for the 1320 interfaces in the 152 crystal forms analyzed by Janin & Rodier (1995). All other numbers are for 188 interfaces with BSA $>$ 800 Å^2 that were selected among those 1320 interfaces by Bahadur *et al.* (2004).

[e]Score obtained by summing over the whole interface the propensity of individual residues to occur at the interface of homodimers (Bahadur *et al.* 2004).

[f]Number of polypeptide segments per 1000 Å^2 of BSA. A separate dataset of 204 structures has been used for protein–protein complexes (Pal *et al.* 2007).

[g]Data from Rodier *et al.* (2005).

[h]Mean values in subsets that comprise 52 protein components of the complexes (excluding antigen–antibody complexes), 121 homodimers and 102 monomeric proteins in crystal contacts. s is the mean Shannon entropy in aligned sequences. Data from Guharoy & Chakrabarti (2005).

Table 5.3: List of natural protein-protein interface parameters. Table taken from Janin et al[39].

Table of Chosen Designs and Design Protocol - First Round				
Design Name	Abbreviation	D3 Component	C1 Component	Design Protocol
MA001-1nu4_1ej2_1_1_hbnet_0182	A1	1ej2	u1a	HBNet
MA009-1nu4_2cwq_1_2_SB_0005	A2	2cwq	u1a	Structure_backbone
MA017-1nu4_1vmd_1_3_SB_0007	A3	1vmd	u1a	Structure_backbone
MA002-1nu4_1ej2_1_1_SB_0009	B1	1ej2	u1a	Structure_backbone
MA010-1nu4_2cwq_1_2_SB_0012	B2	2cwq	u1a	Structure_backbone
MA018-3v7e_1vlh_1_3_hbnet_0077	B3	1vlh	YbxF	HBNet
MA003-1nu4_2brx_1_0_hbnet_0081	C1	2brx	u1a	HBNet
MA011-1nu4_2cwq_1_2_hbnet_0015	C2	2cwq	u1a	HBNet
MA019-3v7e_1vlh_1_3_hbnet_0100	C3	1vlh	YbxF	HBNet
MA004-1nu4_2brx_1_0_hbnet_0194	D1	2brx	u1a	HBNet
MA012-1nu4_2cwq_1_4_SB_0016	D2	2cwq	u1a	Structure_backbone
MA020-3v7e_1vlh_1_3_hbnet_0057	D3	1vlh	YbxF	HBNet
MA005-1nu4_2brx_1_0_hbnet_0147	E1	2brx	u1a	HBNet
MA013-1nu4_2cwq_1_4_SB_0009	E2	2cwq	u1a	Structure_backbone
MA021-3v7e_3aog_1_3_hbnet_0023	E3	3aog	YbxF	HBNet
MA006-1nu4_2brx_1_0_SB_0005	F1	2brx	u1a	Structure_backbone
MA014-1nu4_2cwq_1_4_SB_0012	F2	2cwq	u1a	Structure_backbone

MA023-3v7e_4i4z_1_2_hbnet_0045	F3	4i4z	YbxF	HBNet
MA07-1nu4_2brx_1_0_SB_0015	G1	2brx	u1a	Structure_backbone
MA015-1nu4_2cwq_1_4_SB_0013	G2	2cwq	u1a	Structure_backbone
MA024-3v7e_4i4z_1_2_SB_0005	G3	4i4z	YbxF	Structure_backbone
MA008-1nu4_2cwq_1_1_SB_0015	H1	2cwq	u1a	Structure_backbone
MA016-1nu4_1vmd_1_3_hbnet_0033	H2	1vmd	u1a	HBNet
MA025-3v7e_4i4z_1_3_hbnet_0095	H3	4i4z	YbxF	HBNet

Table 5.4: List of chosen designs and protocols used. The output of the two Rosetta interface redesign protocols resulted in 468 unique protein sequences. These were pruned down to 25 sequences for biochemical characterization.

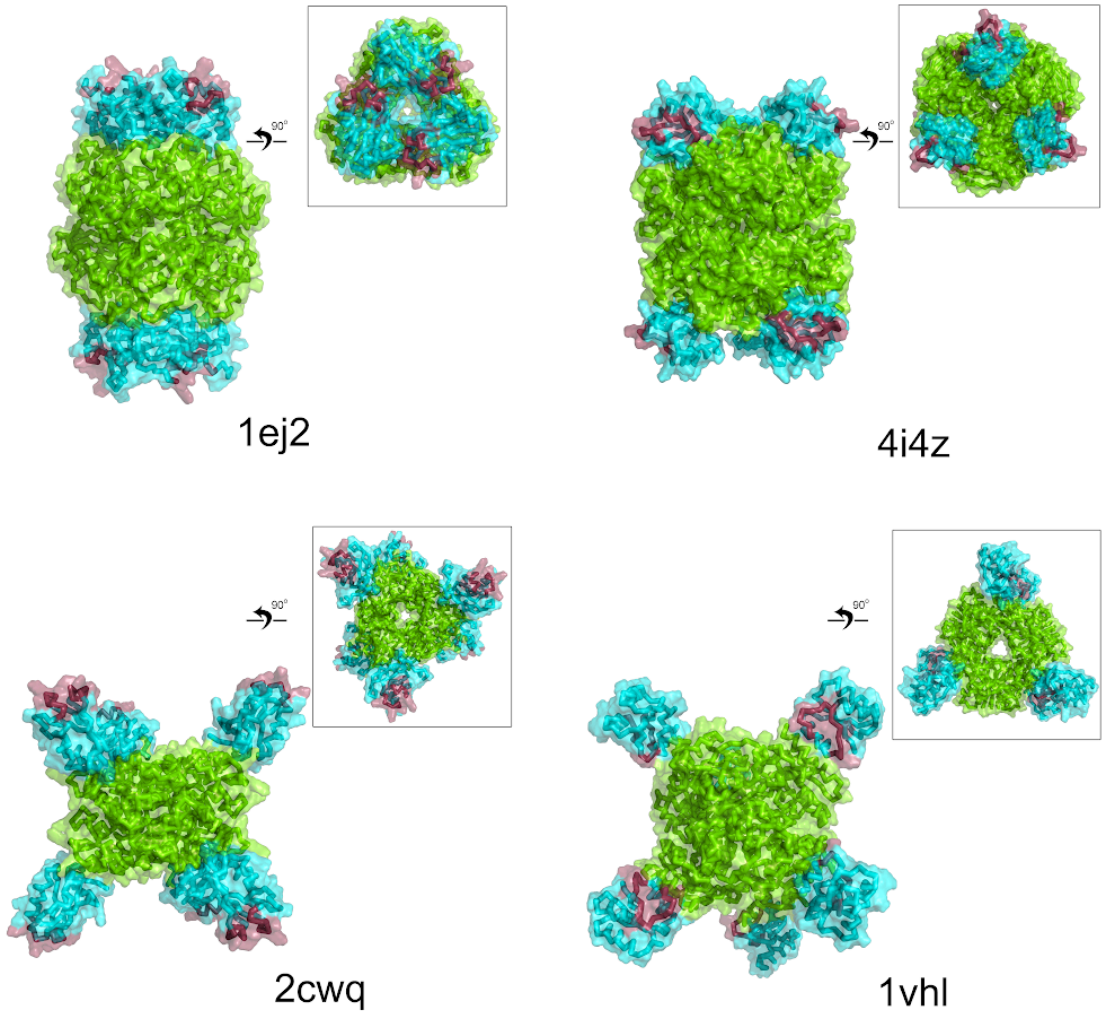


Figure 5.5: Representative D3-C1 full design assemblies. Expanded assemblies for four unique design poses are displayed along with their corresponding PDB cores. The D3 scaffold core is shown in green and the C1 RNA-binding proteins are shown in cyans. Highlighted in magenta are residues involved in RNA ligand binding.

Table of Component and Assembly Mass			
Design Abbreviation	D3 MW (kDa)	C1 MW (kDa)	Full Assembly MW (kDa)
A1	21.2	11.8	198
A2	13.31	12.30	153.66
A3	19.12	12.75	191.22
B1	21	11.9	197.4
B2	13.38	12.30	154.08
B3	18.22	9.42	165.84
C1	24.18	12.4	219.48
C2	13.30	12.25	153.3
C3	18.22	9.41	165.78
D1	25.02	12.24	223.56
D2	13.47	12.42	155.34
D3	18.22	9.46	166.08
E1	25.25	12.288	225.18
E2	13.51	12.43	155.64
E3	46.73	9.68	338.46
F1	25.04	12.41	224.7
F2	13.55	12.38	155.58
F3	30.67	9.74	242.46
G1	25.21	12.38	225.54
G2	13.60	12.40	156
G3	30.84	9.73	243.42
H1	13.51	12.42	155.58
H2	18.98	12.53	189.06
H3	30.65	9.59	241.44

Table 5.5: Table of component and assembly masses. The molecular weights of each of the design components are listed next to their abbreviated moniker (Table 5.4). A total of 6 copies of each D3 and C1 component assemble into the complete design assembly.

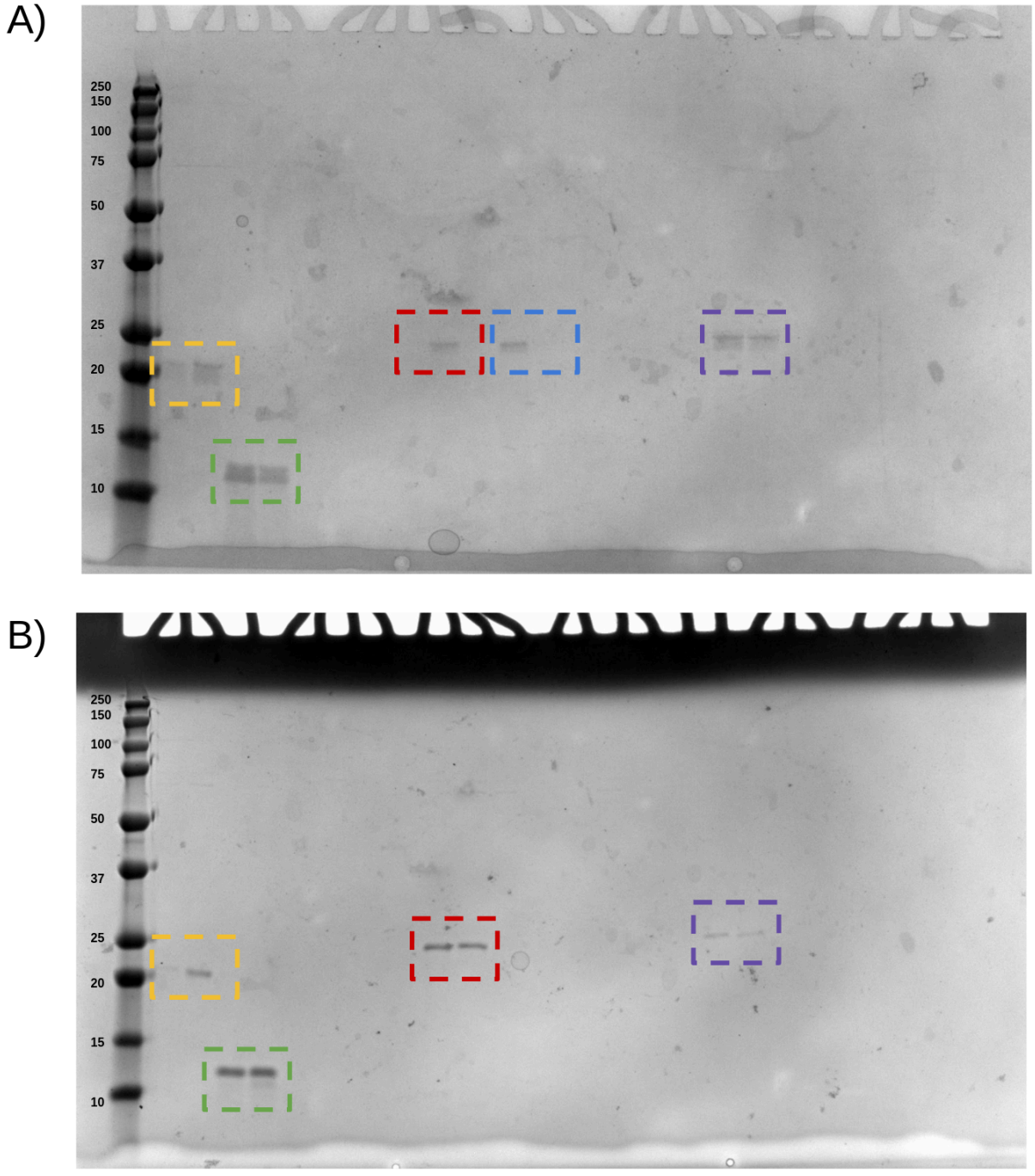


Figure 5.6: D3-C1 imaging scaffold design expression tests. To test so many constructs for expression and solubility, 96-well cultures were used to test growth media and expression temperature. A) Cultures grown and expressed in LB media. B) Cultures grown and expressed in TB media. On the gel there are two lanes that correspond to each design. The first lane is where 37°C expression was tested, the second lane corresponds to 18°C. Overall, designs preferred expression at lower temperature.

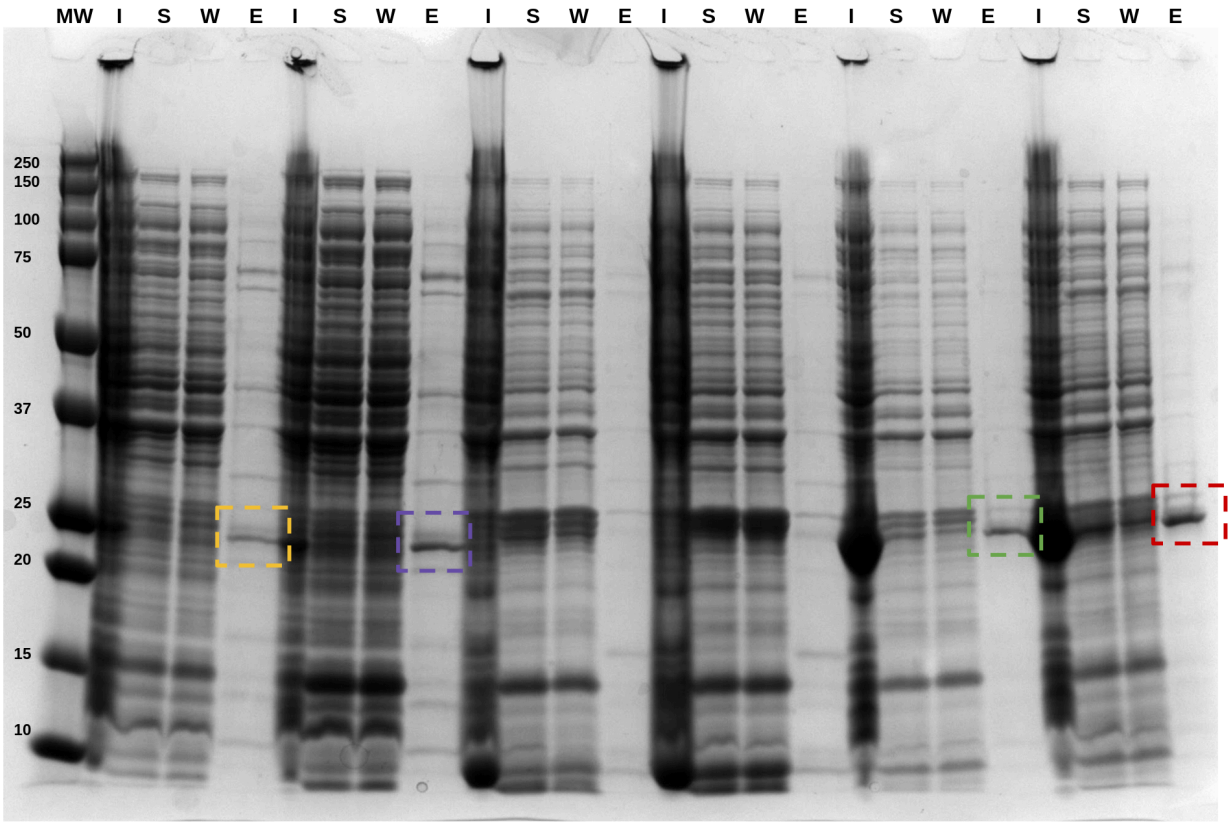
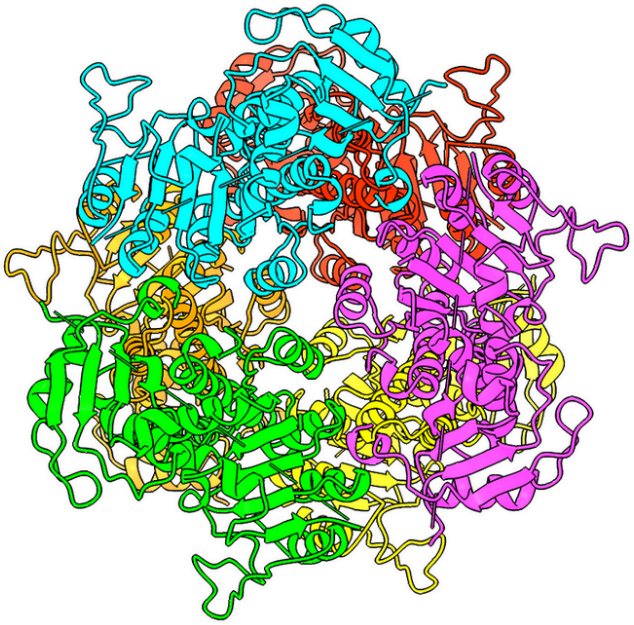
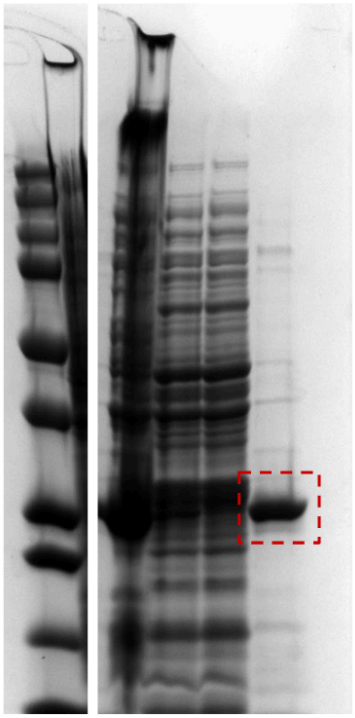


Figure 5.7: Large scale growths of Rosetta D3-C1 designs. Designs were grown at 1 liter scale and tested for solubility. A representative gel image is shown in this figure. The same trend seen with the small scale expression cultures is seen in these larger ones. If any expression or solubility was seen, it was only of the D3 component (colored boxes). In no case were bands for a soluble C1 component observed. Lane markers are as follows: molecular weight (MW), Insoluble (I), Soluble (S), Wash (W), and Elution (E).

A)



B)



C)

Analysis of Design C1 Sequence Mutations			
Design Component	Number of Mutations Introduced	Total Number of Residues	% of Protein Mutated
2brx (D3)	20	218	9.17%
U1A (C1)	18	91	19.78%

Figure 5.8: Analysis of mutations to design C1. The best behaving core was design C1 (MA003-1nu4_2brx_1_0_hbnet_0081), so it was chosen for further analysis. The redesign down to the D3 scaffold core was fairly mild with only ~10% of its amino acids being changed. The redesign of U1A was much more drastic with almost 1/5th of the entire protein being changed. Larger proteins are able to accommodate more changes than smaller ones, potentially explaining the discrepancy in expression and solubility observed.

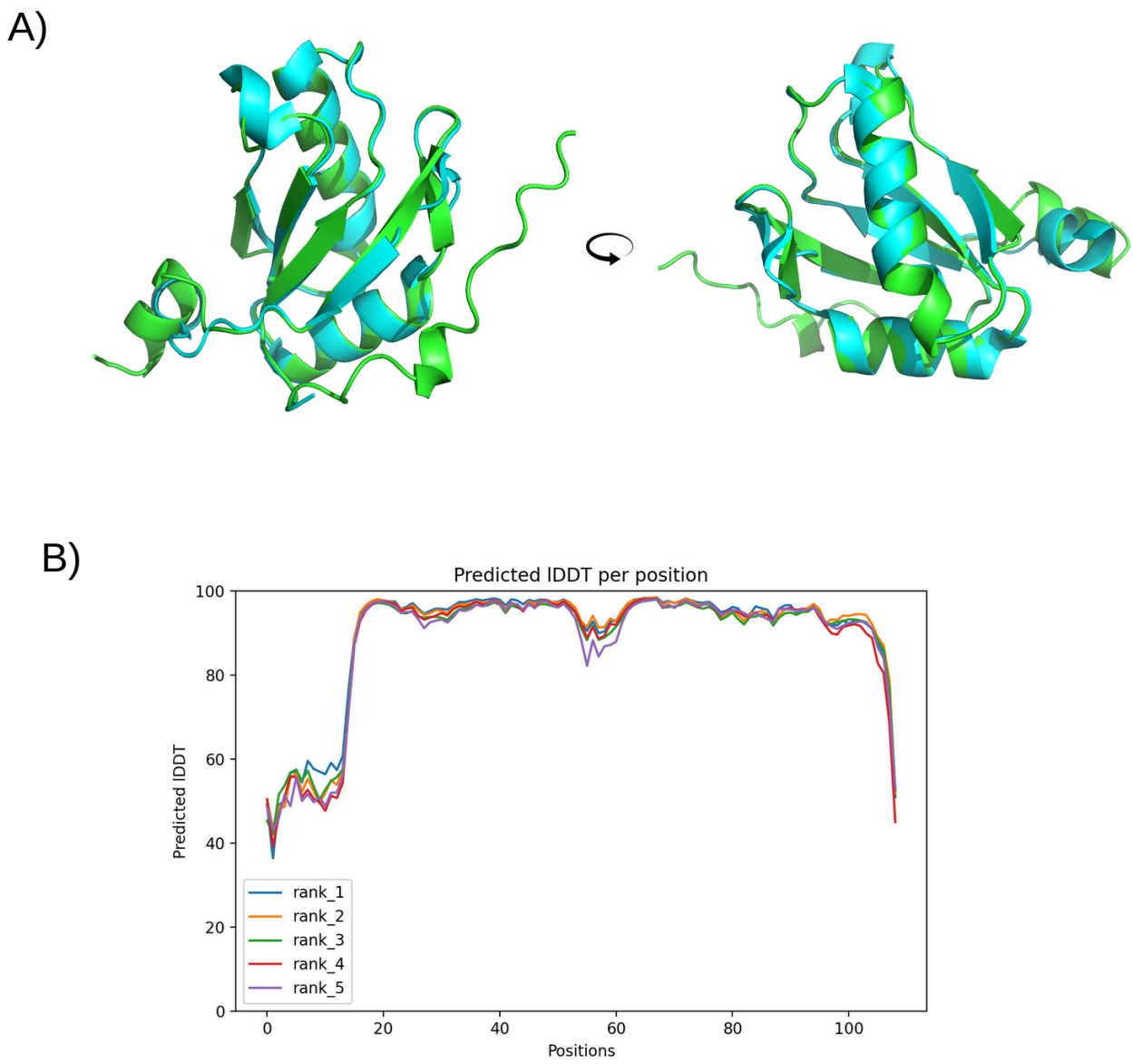


Figure 5.9: AlphaFold analysis of design C1. The sequence for both the D3 and C1 components were run through AlphaFold to determine if the mutations introduced are predicted to affect the three-dimensional structure of the designs. Shown in this figure is analysis on the U1A RNA-binding component. A) Overlay of the U1A structure (PDB: 1nu4, green) matches well with the predicted structure from AlphaFold (cyan). B) The Local Distance Difference Test (lDDT) metric is a measure of the local distance differences between atoms in the model and is used as a metric of assessing prediction outputs[41]. The lDDT plot shows a good prediction of the structure, indicating that the mutations we introduced do not affect the global fold of U1A to any significant degree.

List of D3-C1-Linked Designs Chosen					
Design Name	Abbreviation	D3 Component	C1 Component	Linker Identity	Scaffold Chain Size (MW)
1nu4_1ej2_1_1_MPNN-0008-By-Eye_Link	MPNN1-Link	1ej2	U1A	sggsggsgg	31.30 kDa
1nu4_2brx_1_0_MPNN-0008-By-Eye_Link	MPNN2-Link	2brx	U1A	ggsggs	26.81 kDa
1nu4_2cwq_1_2_MPNN-0007-By-Eye	MPNN3-Link	2cwq	U1A	ggsggsgsg	26.83 kDa
1nu4_2cwq_1_2_MPNN-0006-By-Eye_Link	MPNN4-Link	2cwq	U1A	ggsggsgsg	26.89 kDa
1nu4_1vmd_1_3_MPNN-0008-By-Eye_Link	MPNN5-Link	1vmd	U1A	ggsggsggsg	31.5 kDa
3v7e_1ej2_1_1_MPNN-0009-Link	MPNN6-Link	1ej2	YbxF	ggsggsgg	33.69 kDa
3v7e_4i4z_1_3_MPNN-0006-Link	MPNN7-Link	4i4z	YbxF	ggsggsg	40.76 kDa
1nu4_1v9l_1_0_MPNN-0005_Link	MPNN8-Link	1v9l	U1A	ggsggsggs	59.45 kDa
3v7e_1vlh_1_3_MPNN-0002_Link	MPNN9-Link	1vlh	YbxF	ggsggsggs	28.88 kDa

1nu4_1ej2_1_0005-Link	MPNN10-Link	1ej2	U1A	ggsggsgg	32.49 kDa
1nu4_1ej2_1_0009-Link	MPNN11-Link	1ej2	U1A	ggsggsgg	32.68 kDa
1nu4_1ej2_1_0003-Link-For-Compare	MPNN12-Link	1ej2	U1A	ggsggsgg	32.6 kDa
1nu4_1vmd_1_3_0010-Link	MPNN13-Link	1vmd	U1A	ggsggsggs	32.5 kDa
3v7e_1vlh_1_3_0004-Link	MPNN14-Link	1vlh	YbxF	ggsggsggs	29.04 kDa
3v7e_1vlh_1_3_0005-Link	MPNN15-Link	1vlh	YbxF	ggsggsggs	29.11 kDa
3v7e_4i4z_1_3_0010-Link	MPNN16-Link	4i4z	YbxF	ggsggsgg	40.79 kDa
3v7e_4i4z_1_3_0009-Link	MPNN17-Link	4i4z	YbxF	ggsggsgg	40.88 kDa

Table 5.6: List of MPNN selected designs. Of the 117 unique sequences that resulted from the MPNN interface redesign protocol, 17 were selected for biochemical characterization. Abbreviations for each design are given and are referred to as such in the main text of this chapter.

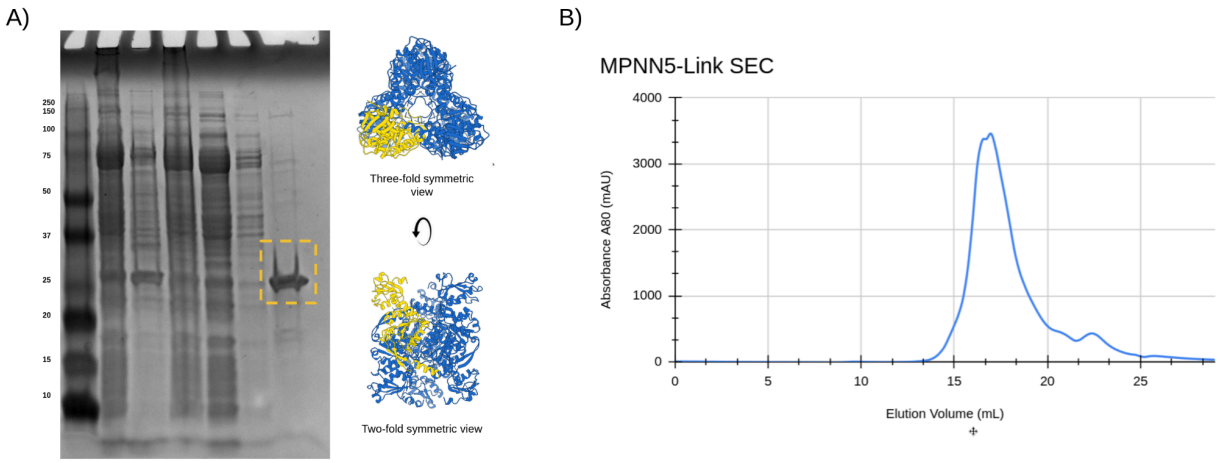


Figure 5.10: Biochemical characterization of MPNN5. The best preforming design from the pool of 17 MPNN-redesigned sequences was MPNN5, which expressed as a single, homogeneous band under NiNTA, as visualized by SDS-PAGE (A), and eluted as a monodisperse peak under SEC (B).

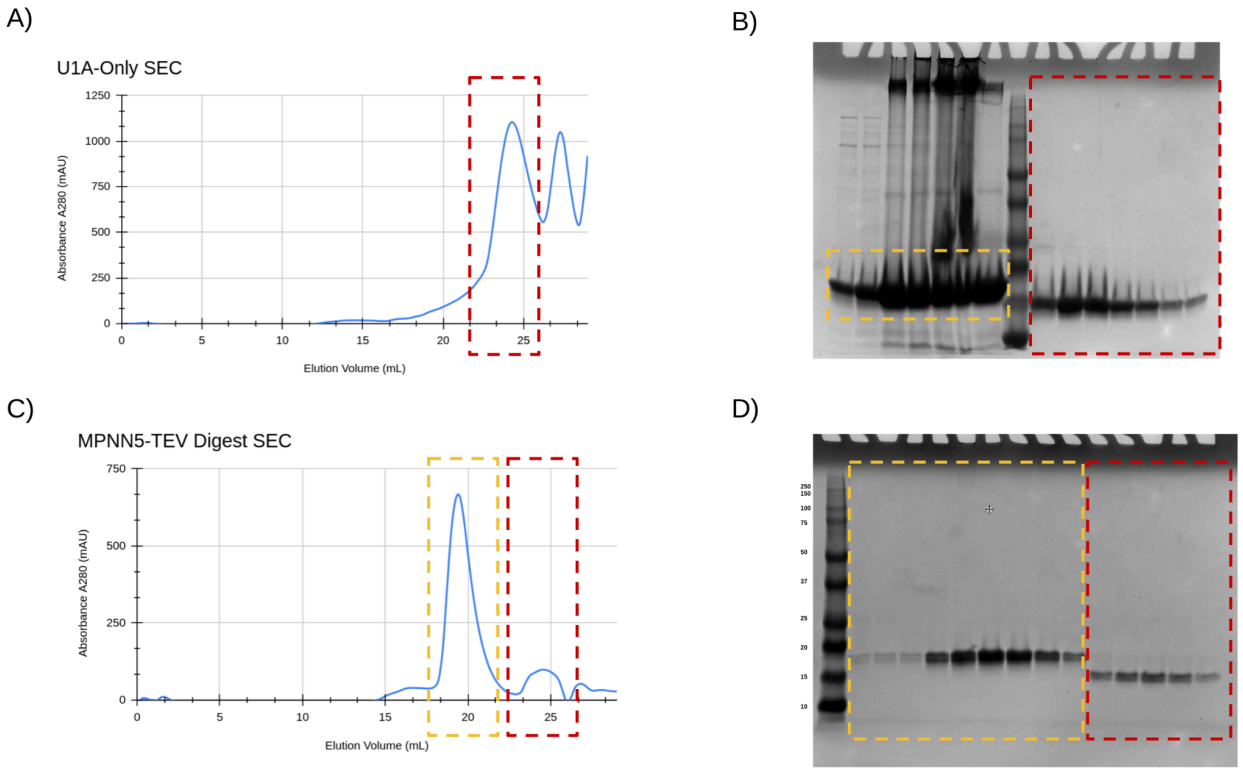


Figure 5.11: TEV Digestion of MPNN5-Link. In order to confirm that the two components of my scaffold are indeed associating, I PCR cloned TEV-digestion sites into the flexible linkers, purified the constructs, incubated them with TEV, and then repurified the mixture over SEC. A) The SEC profile for the U1A protein by itself. The RNA-binding protein elutes with a peak ~24mL on a Superose 6 Increase column (red box). B) The peak corresponding to U1A alone was run on an SDS-PAGE to see where our control migrates. Bands corresponding to U1A show a homogeneous sample of approximately the correct molecular weight of ~12 kDa (red box). As a second control, the D3 component of the MPNN5 scaffold (PDB: 1VMD) was also purified as a control and is shown to run around its molecular weight of approximately 19 kDa (yellow box). C) After incubation of the construct with TEV protease, the mixture was re-run over the Superose 6 Increase column and two peaks were observed (checkered boxes). D) SDS-PAGE analysis of each peak shows that upon TEV digestion, the construct dissociated into its constituent parts and failed to associate with each other under SEC (checkered boxes).

APPENDIX ONE: AlphaFold-Assisted Structure Determination of a Bacterial Protein of Unknown Function Using X-ray and Electron Crystallography

The following is a reprint of a research article from:

Acta Crystallographica Section D

80(4): 270-278 (2024)

DOI: 10.1107/S205979832400072X

AlphaFold-Assisted Structure Determination of a Bacterial Protein of Unknown Function Using X-ray and Electron Crystallography

Justin E. Miller , Matthew P. Agdanowski, Joshua L. Dolinsky, Michael R. Sawaya, Duilio Cascio, Jose A. Rodriguez, Todd O. Yeates

Keywords

Electron diffraction, protein structure prediction, bacterial proteins, molecular replacement

Abbreviations

MIR – multiple isomorphous replacement

Micro-ED – micro electron diffraction

Cryo-EM – cryo electron microscopy

LLG – log likelihood gain

pLDDT - per-residue model confidence score

RMSD – root mean square deviation

Abstract

Macromolecular crystallography generally requires the recovery of missing phase information from diffraction data to reconstruct an electron density map of the crystallized molecule. Most recent structures have been solved using molecular replacement as a phasing method, requiring an a priori structure that is closely related to the target protein to serve as a search model; when no such search model exists, molecular replacement is not possible. New advances in computational machine learning methods, however, have resulted in major advances in protein structure predictions from sequence information. Methods that generate predicted structural models of sufficient accuracy provide a powerful approach to molecular

replacement. Taking advantage of these advances, we applied AlphaFold predictions to enable structure determination of a bacterial protein of unknown function (UniprotKB Q63NT7, NCBI locus BPSS0212), based on diffraction data that had evaded phasing attempts by MIR and anomalous scattering methods. Using both X-ray and micro-electron (microED) diffraction data, we were able to solve the structure of the main fragment of the protein using a predicted model of that domain as a starting point. The use of predicted structural models importantly expands the promise of electron diffraction, where structure determination relies critically on molecular replacement.

Introduction

New variations on traditional x-ray crystallography are expanding the power of diffraction methods for macromolecular structure determination [1–6]. Two ongoing developments are notable for their potential scope. First, recent algorithmic advances in protein structure prediction have made it possible, in many cases, to generate three-dimensional models that are accurate enough for molecular replacement protocols [6–9]. Such cases ultimately allow for an experimental structure to be elucidated, without the need for experimental phasing (i.e. heavy atom or anomalous approaches), and without prior experimental knowledge of a similar protein structure. Second, on the side of experimental advance, electron-based diffraction is attracting attention as a potential approach suitable for very small crystals [10]. These two lines of exploration intersect. Heavy atom and anomalous scattering methods of phasing do not transfer readily to electron diffraction, elevating the importance of molecular replacement for that method, including with predicted models. More case studies are needed to demonstrate the utility, and the challenges, of these new structure determination approaches.

The subject of the present study is a bacterial protein of unknown structure and function, UniprotKB Q63NT7. It was chosen for structural investigation based on its unusual genomic

presentation. The tendency of the protein family PF08898 (proteins containing the domain: DUF1843) to be encoded as repeated paralogs within individual operons suggested that it might form part of a larger self-assembling protein complex, as proposed in an earlier bioinformatics study [11] (Figure 1A), but no structural data was available. Biochemical and structural studies were therefore undertaken to investigate the structure of this protein domain and to evaluate whether it might form a larger self-assembling complex. Difficulties in obtaining large crystals led to expanded efforts, including structure determination from small crystals by electron diffraction, and molecular replacement using predicted models.

Results

3.1 Protein Expression and Purification

The Q63NT7 protein from species Burkholderia pseudomallei is 212 amino acids long (MW 22.5kDa). It contains two predicted domains: an N-terminal domain of unknown function (14.5kDa) (DUF1842), and the aforementioned C-terminal domain (DUF1843, 5.4kDa) which tends to appear in multiple paralogous copies within individual bacterial operons. We ordered sequences encoding the Q63NT7 sequence with C-terminal 6xHistidine tags. We expressed the protein recombinantly in BL21 (DE3) E. coli cells (Supplemental Table S1). Biochemical characterization of this protein suggested the protein is monodisperse and likely monomeric in solution (Figure 2A).

3.2 Protein Crystallization and Crystal Forms

Encouraged by the purity of our protein sample, we attempted to solve the structure crystallographically. Q63NT7 presented a challenge for obtaining large, well-ordered crystals. This led us to explore multiple distinct crystal forms with the goal of improving diffraction quality and, as discussed later, to attempt to visualize a substantial region of the protein that could not be resolved in density maps.

Initial crystallization trials yielded abundant needles across many crystallization conditions, but attempts to obtain X-ray diffraction data were unsuccessful. We also observed inconsistent crystal formation across our replicated crystal trays. Even so, we were ultimately (after approximately 6-months) able to optimize these conditions and grow larger rectangular shaped crystals that diffracted beyond 3Å on a synchrotron microfocus beamline (Figure 2B). We collected datasets on these crystals, which we refer to as form 1. Diffraction data indexing revealed the space group as $P2_1$ (Supplemental Table S5). The highest resolution resulted from data collected on a single crystal specimen.

In parallel with efforts to phase data from form 1 crystals, we sought to achieve higher quality diffraction from Q63NT7 crystals. Anticipating that needle-shaped crystals might be especially suitable for micro-electron diffraction (Micro-ED) methods owing to their limited thickness, we used an electron microscope to investigate the order and diffraction quality of needle-shaped microcrystals that grew in showers in some of our drops). We first pipetted those drops onto formvar carbon electron microscopy grids, stained them with uranyl acetate and imaged them. The crystallinity of our sample was evident by the appearance of lattice lines in the sample (Supplemental Data Figure 1). We next investigated the diffraction quality of these crystals when frozen, so we proceeded to freeze microcrystals from similar conditions for cryo- EM Micro-ED. These microcrystals typically diffracted to 3 Å resolution in an electron microscope operating in diffraction mode (see Methods) (Figure 2B, bottom panel). We collected diffraction datasets from four microcrystals. The crystal unit cell dimensions were non-isomorphous with form 1 crystals (Supplemental Table S5), so we refer to these as form 2 crystals. Unfortunately, the crystals which appeared to be ribbon-shaped at high magnification suffered from preferred orientation problems. We were unable to collect diffraction at high tilt angles, and therefore achieved only 59% completeness in a merged diffraction data set. Furthermore, it was difficult to

confidently assign a space group due to a substantial missing cone of reflections (Figure 3). The quality of the individual datasets was poor, partly owing to weaker signal (e.g. unsatisfactory R sym values) from some specimens, merging multiple datasets did not substantially improve data quality, but improved completeness slightly. We therefore elected to proceed using a dataset obtained from merging diffraction from four crystals; unfortunately, since the regions of reciprocal space missing from distinct data sets were largely overlapping, the final dataset was only complete to 59%. Owing to the lack of data along the c* axis, systematic absences were difficult to discern from missing data, and determining the number of 2_1 screw axes was initially unclear. Attempts at molecular replacement with the form 2 electron diffraction data in space groups 16, 17, 18, and 19 ultimately confirmed that $P2_12_12_1$ was correct for form 2 microcrystals based on a much higher LLG value.

Lastly, we continued to optimize the crystallization conditions and identified another condition that grew well-diffracting needle-shaped crystals suitable for data collection on the synchrotron micro-focus beamline. We were able to collect a complete dataset from a single crystal that indexed in $P2_12_12_1$. We refer to this crystal as form 3, since its unit cell dimensions were distinct from forms 1 and 2 (Supplemental Table S5).

3.3 Molecular Replacement Using AlphaFold Models

Efforts to phase the highest quality dataset (form 1) with experimental techniques did not lead to immediate success; selenomethionine labeled protein crystals did not diffract, and we observed no heavy atom signal in the diffraction patterns of crystals soaked in $CsCl_2$ or KI. Inspired by studies that had used AlphaFold models to phase datasets with little a priori information, we used the software to generate a model of Q63NT7 (Figure 4). AlphaFold identified two domains in the protein, joined by a long linker. The N-terminal domain was predicted to fold into a β-barrel composed of 8 antiparallel strands. AlphaFold predicted this

domain with a high degree of confidence based on per-residue plDDT scores. The C-terminal domain was predicted to form a small helical bundle with modest plDDT confidence metrics. Applying existing molecular replacement methods to our AlphaFold-based molecular replacement efforts, we separated the coordinates of the two domains into independent files and removed extended loop segments, including the long linker between domains (Figure 4).

We used these two files as search models for molecular replacement with the program Phaser 13 . Remarkably, datasets from all three crystal forms gave solutions that passed Phaser's metrics for a correct solution using the N-terminal β-sheet rich domain. The solution was further validated using a test search model that excluded residue H125; maps phased from such a molecular replacement model produced positive density at the expected position in an Fo-Fc difference map (Supplemental Data Figure 2). All three crystals forms identified two copies of the N-terminal β-barrel domain in the asymmetric unit (Fig. 5). Form 1 crystals gave a combined LLG value of 719, Form 2 crystals gave an LLG value of 394, and Form 3 crystals gave an LLG value of 305. On the contrary, none of the crystals could be phased using the C-terminal alpha helical domain as a search model using similar program parameters. Given the small contribution of scattering attributed to this domain because of its small size, these negative results were not altogether surprising.

We next investigated whether similar structures in the PDB existed, and whether, in retrospect, they too could have served as search models for molecular replacement with our data. To do this, we submitted the structure obtained from the form 1 crystal dataset (after molecular replacement and preliminary refinement) to the DALI server and identified the top five closest matching protein folds (i.e. those with the highest Z-scores) in the PDB (Supplemental Data Figure 3) 14 . Interestingly, the structure that was identified as most similar to our own based on Z-score is an outer membrane protein from Pseudomonas aeruginosa. These attempts did not

produce plausible packing solutions using form 1 data. We went on to test whether these known models would produce solutions for the form 2 data with lower completeness, which might lend itself to incorrect solutions more so than the comparatively better form 1 data. For each of these trials, Phaser was unable to produce molecular replacement statistics indicative of a correct solution. LLGs for these trials were all below what would be expected for a correct solution, and all significantly lower than for the AlphaFold model: 129 (2erv), 57 (2f1v), 25 (4u8u), 112 (4rcl), and 119 (4bbo). We further tested whether FoldSeek's 15 search algorithm could be used to identify other molecular replacement search models, either from the pdb or amongst the vast number of predicted protein models. Performing molecular replacement using our form 2 data, the most similar structure identified using FoldSeek's search from the pdb: 6cd8 gave a Phaser LLG value of only 47. Our finding was that only the AlphaFold model predicted for the DUF1842 β-barrel domain was sufficiently close to the target structure to serve as a successful molecular replacement input.

3.4 Refinement of Atomic Structures

Because form 1 crystals gave the highest resolution diffraction data (from X-rays), a model for the β-sheet rich domain was refined against that data and then subsequently used as the starting point for model refinements in the other crystal forms (X-ray and electron). This strategy helped to prevent separation of R free and R work , especially in the case of the Micro-ED data, which suffered from low completeness and poor I/sigma and R_{merge} statistics.

Importantly, Phaser statistics for molecular replacement solutions using the refined form 1 crystal structure were much improved over those from AlphaFold predicted models; form 2 datasets gave Phaser LLG values of 624 while form 3 crystal datasets gave a phaser LLG value of 662 (compared to LLGs of 394 and 305). We therefore adopted those solutions as starting points for atomic refinements.

During refinement, we paid close attention to the C-terminal region of resulting density maps to observe whether density expected for the C-terminal domain would become visible. In all three forms, large solvent channels were noted adjacent to the C-terminus of the β-barrel domain (Supplemental Data Figure 4), which would have allowed for possible placement of the small C-terminal segment. Unfortunately, in all crystal forms, we observed no meaningful positive density in Fo-Fc difference maps in the regions that would have to be occupied by the C-terminal domain. We hypothesize this could be due to proteolysis, as we observed degradation products on SDS-PAGE gels, and subsequently in mass spectra from dissolved crystalline samples (Supplemental Data Figure 5 & unpublished data). Final refinement for the form 1 model gave an R factor of 25.1% and R_{free} of 28.7. The form 2 model had an R factor of 28.4% and R free of 30.7%. The form 3 model had an R factor of 27.3% and R free of 33.3%. The structure of the N-terminal β-rich domain was strongly conserved across all crystal forms and asymmetric units; no protein chain from any of the three crystal forms had a backbone RMSD above 0.6 Å compared to any other chain (Figure 5). There was also close agreement between the refined structures and the AlphaFold prediction. Backbone RMSD values between the experimental structures and the AlphaFold model were 0.35Å for form 1, 0.49Å for form 2, and 0.46Å for form 3.

Analyzing the non-crystallographic symmetry of the three crystal forms revealed molecular packing interfaces that were substantially different. As a result, no biologically relevant interfaces could be inferred with confidence. One potentially relevant exception is that the form 2 non-crystallographic interface is present as a crystallographic interface in form 3 crystals.

3.4 Structural Analysis

The overall structure of the Q63NT7 protein C-terminal domain forms an 8 stranded antiparallel β-barrel. Residues 67-77, corresponding to the amino acid sequence: GPPPRRDGSG, did not appear in any of the three electron density maps, and thus were left out from the structures

deposited in the pdb. Polar residues are found covering the exterior surface of the β-barrel, while the interior of the β-barrel is lined with mostly hydrophobic residues, without space for a channel through the barrel. Residues 88-97 form an unusual hydrophobic extended loop, with a conserved structure across crystal forms, interacting with strand 4 of the β-barrel.

Discussion and Conclusion

In several cases, microED has proven to be an important tool for structural biologists, enabling the extraction of high-resolution structural information from tiny crystals that are unusable using X-ray diffraction. The earliest demonstration of the method on protein crystals was seminal work on crystals of lysozyme[16] . Important early work from Rodriguez and colleagues advanced on these studies and demonstrated the utility of microED in solving the structures of small peptides[17]. Other work has demonstrated the method's utility in solving structures of proteins in cases where structures are already known for proteins that are closely or even distantly related, including ligand or drug-bound forms of proteins[4,18,19]. Nevertheless, experimental methods for phasing MicroED data have been elusive, limiting broader applications of the method. The work presented in this paper adds to the relatively small number of electron diffraction structures of novel proteins. Two recent studies have demonstrated success at phasing MicroED data using structures of distantly-related homologs as search models[4,5], while, as far as we are aware, the current study represents the first folded (globular) protein structure solved by MicroED whose structure could not be approximated in advance by virtue of a recognizably homologous known structure. We also note that collection of MicroED data was challenged by a strong tendency of crystals to adopt a preferred orientation on the EM grid, leading to an incomplete dataset, and to less than ideal statistics. This led to some initial uncertainties in assigning a space group and subsequent structure determination.

We also present the structure of a new small protein fold, and the first from protein family DUF1842. Notably, efforts to obtain structural information on the C-terminal domain from our maps were unsuccessful. Between the two domains, we note the presence of a ~25 amino acid long linker predicted to form a loop with low sequence conservation across homologues (Figure 1b). This could contribute to flexibility of the entire C-terminal region of the protein in the context of the crystal. We also observed several instances of proteolysis in our crystal trays, both with and without the sterilizing agent sodium azide added to the crystal drops (Figure 5). Degradation products appear to be composed of prominent fragments of 4-5kDa and 17-19kDa based on SDS-PAGE (Supplemental Data Figure 4). This could place the cut-site directly N-terminal to the C-terminal domain, which did not appear in our crystal structures. The tendency of the protein to undergo proteolysis also lends support to the hypothesis that some part of the C-terminal region of the protein was missing from all three crystal forms, explaining the absence of detectable density in all cases. Considering these data, there could still be unaccounted for scattering from up to ~60 amino acids based on the difference between the estimated molecular weight of abundant bands visible using SDS-PAGE (Supplemental Data Figure 4) and the molecular weight of our structures (~12kDa). We take this as possible explanation for higher-than-typical refinement R-values that we ultimately obtained in all three crystal forms.

Despite our initial predictions, based on genomic patterns, that Q63NT7 might be involved in oligomerization via its C-terminal domain, we were unable to observe any evidence of higher-order oligomer formation either in solution or in crystalline form. Our biochemical studies did not support that the protein of unknown function self-assembles into larger architectures under the conditions tested. Nonetheless, the appearance of a flexible linker to a terminal domain that was unresolved by crystallography is reminiscent of studies on bacterial microcompartment shell proteins[20] whose genomic patterns were the impetus for the original genomic investigation that identified the IPR014994 domain as a target in the current study[11].

Whether the architecture of the full protein molecule – i.e. with the small C-terminal domain intact – might be different remains unclear.

Materials and Methods

Gene Synthesis

Codon-optimized gene sequences were ordered from Integrated DNA Technologies or Twist Biosciences with overlapping sequences corresponding to flanking regions around the hindIII and ndeI restriction sites in the pET-22b expression vector. Intergenic sequences for two-component designs were taken from a pETDuet-1 expression plasmid and ordered as a single gene fragment.

Protein Expression and Purification

Designs were cloned into pET-22b expression vectors using Gibson assembly. Correct cloning of gene was verified using Sanger sequencing. Small-scale expression was performed in BL21(DE3) cells grown in 200 mL of cultures using auto-induction media grown for 24 h at 25 C. Cells were lysed in 50mm Tris-HCl pH 8.0, 250mm NaCl supplemented with 5 mM 2-mecaptoethanol and EDTA-free protease inhibitor tablets (Thermo Fisher Scientific) using an Emulsiflex C3 homogenizer and affinity purified using Ni-NTA agarose resin (Thermo Fisher Scientific) in a gravity flow column. Protein was washed with lysis buffer +100 mM imidazole and eluted in lysis buffer +500 mM imidazole. Eluted protein was dialyzed against imidazole overnight at 4 °C. Samples were run on SDS-PAGE to purity before SEC using a Superdex-75 column (Cytiva Life Sciences) attached to an Acta FPLC (Cytiva Life Sciences). Sodium Azide was then added to SEC elution fractions at a concentration of 0.05% as well as EDTA at a concentration of 5mM.

Crystallization

96 well crystal screens were set up using a Mosquito liquid handler (SPT Labtech) in hanging-drop vapor-diffusion format. Trays were allowed to incubate at 22°C until crystals were observed. For form 1 crystals, cubic crystals formed after ~6 months in conditions containing 100mM BisTris pH 5.5, 25% PEG 3350 with 20mg/ml of protein. Form 2 crystals formed within a week in conditions containing 100mM BisTris pH 5.5, 100mM Ammonium Acetate 17% PEG 10,000 with 20 mg/ml protein. Form 3 crystals were grown at a concentration of 100 mg/ml protein in 100mM TRIS HCl pH 8.5, 150mM MgCl, 12.5% PEG 8000.

X-Ray Data Collection and Processing

X-ray diffraction datasets were collected at the Advanced Photo Source on beamlines NE-CAT 24-ID-C equipped with an EIGER 16M detector and 24-ID-E equipped with Dectris PILATUS 6M-F detector. The XDS software package was used to index diffraction data 21 . Diffraction data statistics are provided in Supplemental Data Table S5.

Negatively-Stained Transmission Electron Microscopy (EM)

Crystal Drops containing crystals were diluted in 5uL distilled water and mixed using a pipette. 3uL was applied to glow-discharged Formvar/Carbon 300 mesh Cu grids (Ted Pella Inc.) for 60 seconds. Excess sample was wicked using filter paper, and the grid was immediately washed with distilled water two times. A 2% uranyl acetate solution was applied to the grid then immediately wicked using filter paper. A final incubation of the grid with 2% uranyl acetate was performed for 20 seconds and the grid was dried completely using filter paper. Imaging was performed on Technai T12, and Talos F200C microscopes (Thermo Fisher).

Micro-ED Data Collection

Crystal drops containing crystals were diluted in 5uL of mother-liquor from the crystal reservoir and mixed gently with a pipette. 5uL was applied to glow-discharged Quantifoil 300 mesh 2/2 copper grids (Electron Microscopy Sciences) and frozen using a Vitrobot Mark IV with pre-wet blotting paper (Thermo Fisher). Seven movies from unique crystals were collected on a Tecnai TF30 microscope (Thermo Fisher) fitted with a TVIPS TemCam-F416 and a single tilt cryo-transfer holder (Gatan Inc.) with a maximum employed tilt range of -60° - +60°. Continuous-rotation Micro-ED data was collected at a rotation rate of 0.085°/s. Diffraction data was indexed in the XDS software package[21] and scaled using XSCALE[21] . Diffraction data statistics are provided in Supplemental Data Table S5.

Molecular Replacement and Structure Refinement

The Phaser program[13] was used for molecular replacement. The AlphaFold program 7 was used to generate molecular replacement search models. After refining this initial AlphaFold search model on the basis of form 1 diffraction data, we used the refined structure to phase form 2 and form 3 crystals, as it gave the best statistics and resulted in the best electron density maps. The Coot program[22] was used for model building, and refinement was performed using Phenix[23]. Atomic refinement statistics are provided in Supplemental Data Table S6.

Acknowledgments

This work was funded by the U.S. Department of Energy Office of Science, award DE-FC02-02ER63421. X-ray diffraction data sets were collected at the Northeastern Collaborative Access Team (NECAT) beamlines, which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Eiger 16M detector on 24-ID-E is funded by a NIH-ORIP HEI grant (S10OD021527). This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User

Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. We thank the staff of NECAT for their help in data collection. This research used resources at the UCLA-DOE Institute's X-ray and EM structure Determination core which is supported by the U.S. Department of Energy (award DE-FC02-02ER63421). We thank Marcus Gallagher-Jones and Kevin Cannon for assistance with micro-ED data collection, and Michael Collazo and Genesis Falcon for assistance in crystallization.

Conflict Statement

The authors declare no competing interests.

References

1. Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomic resolution macromolecular structure determination. F1000Res 9, 667 (2020).
2. ohansson, L. C., Stauch, B., Ishchenko, A. & Cherezov, V. A Bright Future for Serial Femtosecond Crystallography with XFELs. Trends in Biochemical Sciences 42, 749–762 (2017).
3. Martynowycz, M. W. & Gonen, T. From electron crystallography of 2D crystals to MicroED of 3D crystals. Current Opinion in Colloid & Interface Science 34, 9–16 (2018).
4. Xu, H. et al. Solving a new R2lox protein structure by microcrystal electron diffraction. Sci. Adv. 5, eaax4621 (2019).
5. Clabbers, M. T. B. et al. MyD88 TIR domain higher-order assembly interactions revealed by microcrystal electron diffraction and serial femtosecond crystallography. Nat Commun 12, 2578 (2021).
6. Terwilliger, T. C. et al. Accelerating crystal structure determination with iterative AlphaFold prediction. Acta Crystallogr D Struct Biol 79, 234–244 (2023).
7. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
8. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021).
9. Giorgetti, A., Raimondo, D., Miele, A. E. & Tramontano, A. Evaluating the usefulness of protein structure models for molecular replacement. Bioinformatics 21, ii72–ii76 (2005).
10. Nannenga, B. L. & Gonen, T. MicroED: a versatile cryoEM method for structure determination. Emerging Topics in Life Sciences 2, 1–8 (2018).

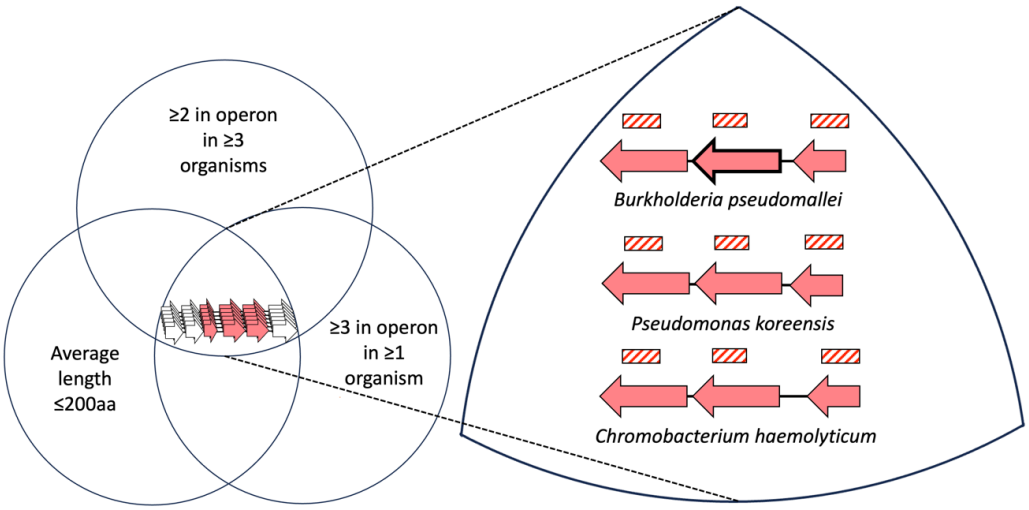
11. Beeby, M., Bobik, T. A. & Yeates, T. O. Exploiting genomic patterns to discover new supramolecular protein assemblies. Protein Science NA-NA (2008) doi:10.1002/pro.1.
12. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44, W344–W350 (2016).
13. McCoy, A. J. et al. Phaser crystallographic software. J Appl Crystallogr 40, 658–674 (2007).
14. Holm, L. & Laakso, L. M. Dali server update. Nucleic Acids Res 44, W351–W355 (2016).
15. Van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol (2023) doi:10.1038/s41587-023-01773-0.
16. Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. Three-dimensional electron crystallography of protein microcrystals. eLife 2, e01345 (2013).
17. Rodriguez, J. A. et al. Structure of the toxic core of α-synuclein from invisible crystals. Nature 525, 486–490 (2015).
18. Martynowycz, M. W. & Gonen, T. Ligand Incorporation into Protein Microcrystals for MicroED by On-Grid Soaking. Structure 29, 88-95.e2 (2021).
19. Martynowycz, M. W. et al. MicroED structure of the human adenosine receptor determined from a single nanocrystal in LCP. Proc. Natl. Acad. Sci. U.S.A. 118, e2106041118 (2021).
20. Thompson, M. C. & Yeates, T. O. A challenging interpretation of a hexagonally layered protein structure. Acta Crystallogr D Biol Crystallogr 70, 203–208 (2014).
21. Kabsch, W. XDS. Acta Crystallogr D Biol Crystallogr 66, 125–132 (2010).
22. Emsley, P. & Cowtan, K. Coot : model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60, 2126–2132 (2004).
23. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and

electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol 75, 861–877 (2019).

24. Studier, F. W. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41, 207–234 (2005).

Appendix Figure A1:

A



B

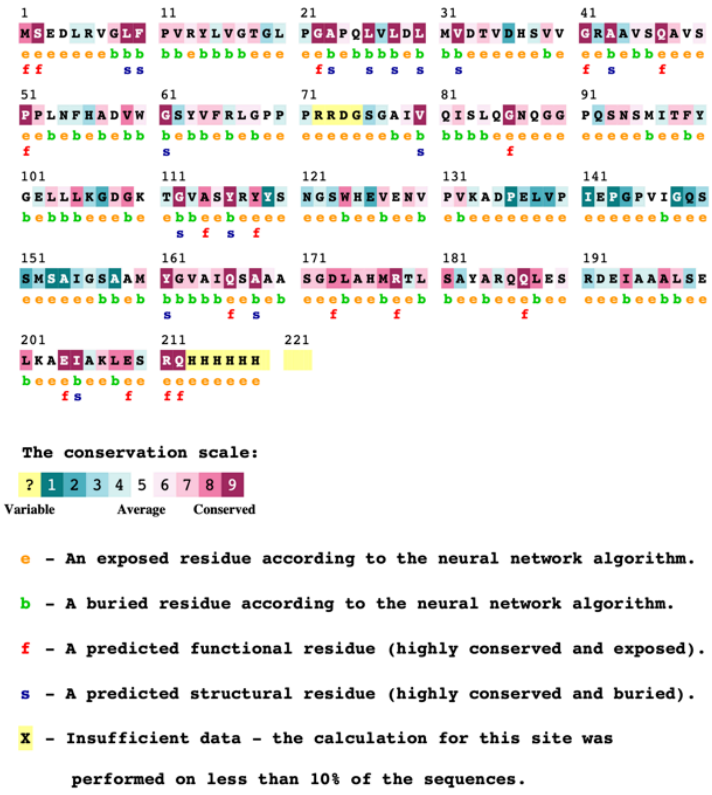


Figure 1: Representation of criteria used to select for genes encoding proteins with an elevated likelihood of self-assembly (including Q63NT7). (A) Graphical representation of the rationale for structural investigations on Q63NT7, where selection criteria are depicted as a Venn diagram as in Beeby *et al.* [11]. Several representative operons with respective organisms of origin obeying selection criteria are highlighted on the right. The Q63NT7 encoding gene is depicted by a bold arrow. DUF1843 containing genes are shown as red arrows, non-homologous genes shown as white arrows. (B) Consurf graphical representation of per-residue conservation of Q63NT7[12].

Appendix Figure A2:

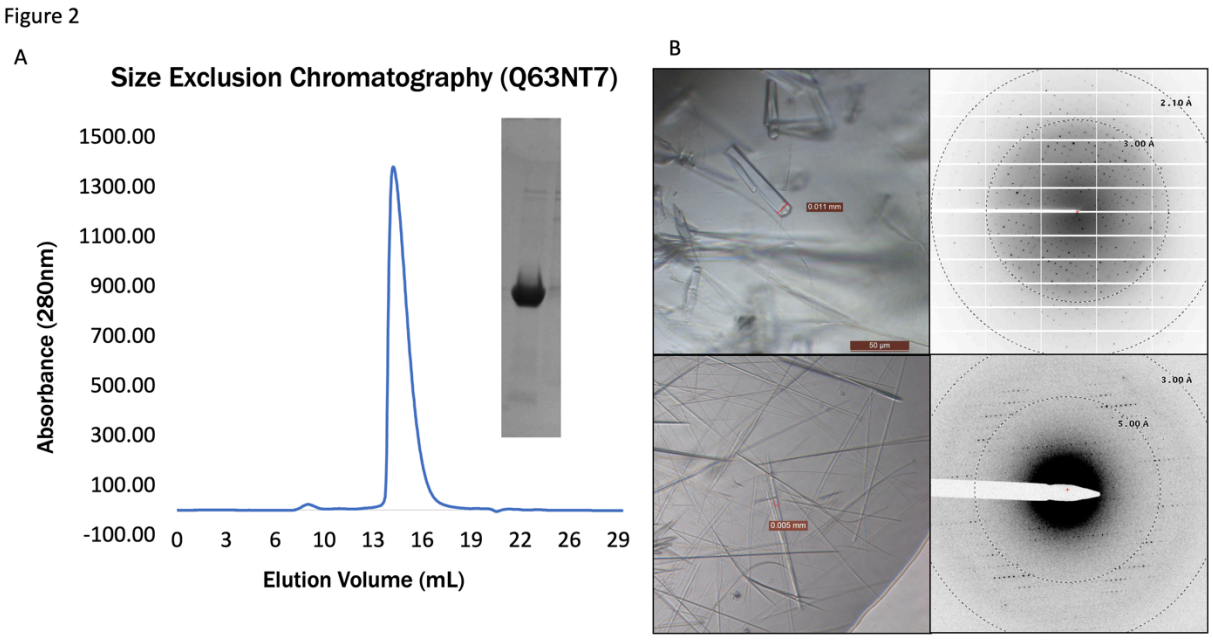


Figure 2: Biochemical characterization of the Q63NT7 protein: (A) SEC and SDS-Page reveal homogeneity and high purity of the Q63NT7 protein. (B) Form 1 (top) and Form 2 (bottom) crystals and representative diffraction data collected from an X-ray source or electron microscope respectively (see Methods).

Appendix Figure A3:

Figure 3

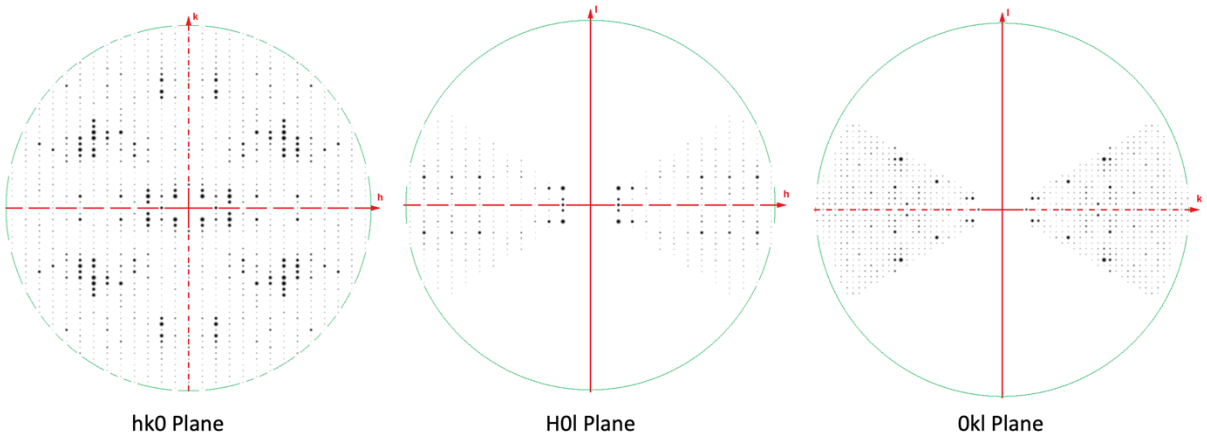


Figure 3. Slices through reciprocal space show the missing cone present in MicroED data collected from 2 crystals. Principal zones are shown to illustrate the missing cone of data due to preferred orientation of crystals on the grid.

Appendix Figure A4:

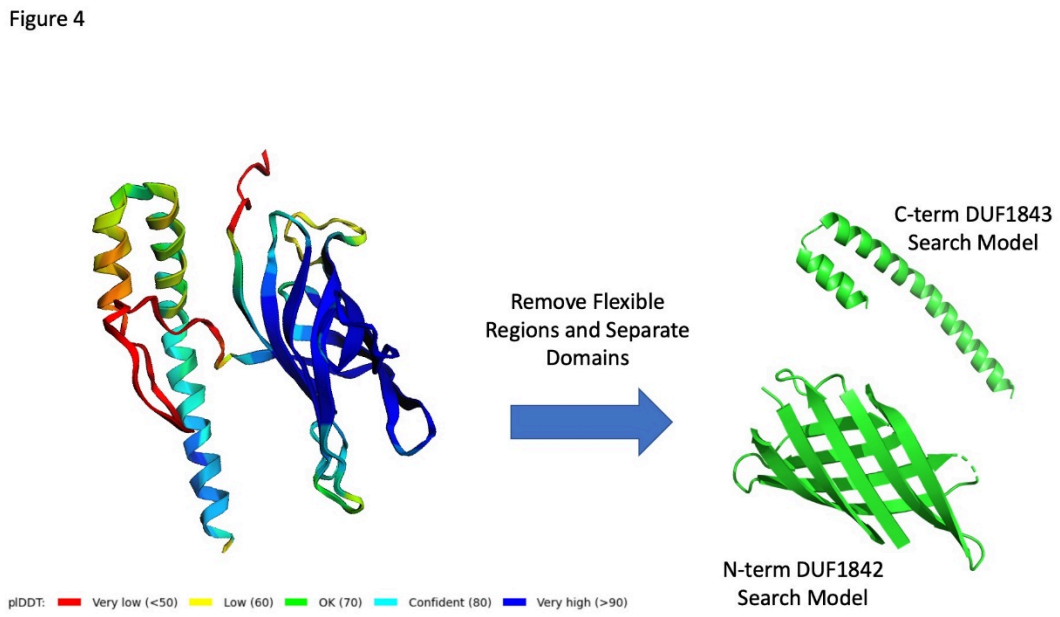


Figure 4: AlphaFold model of the Q63NT7 protein used as molecular replacement search model. plDDT gives a per-residue metric of confidence in model prediction.

Appendix Figure A5:

Figure 5

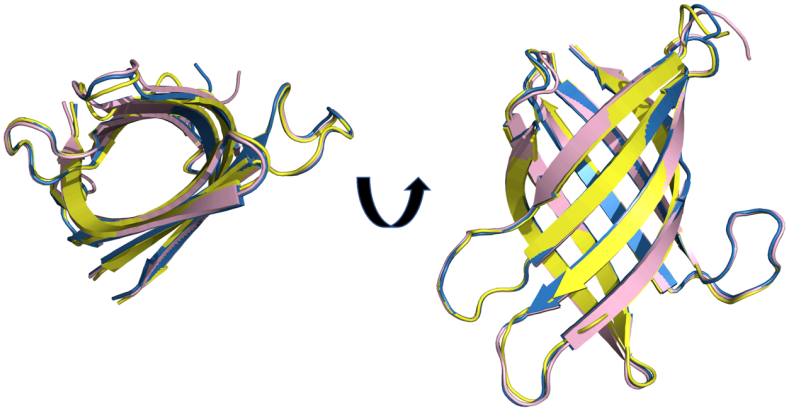
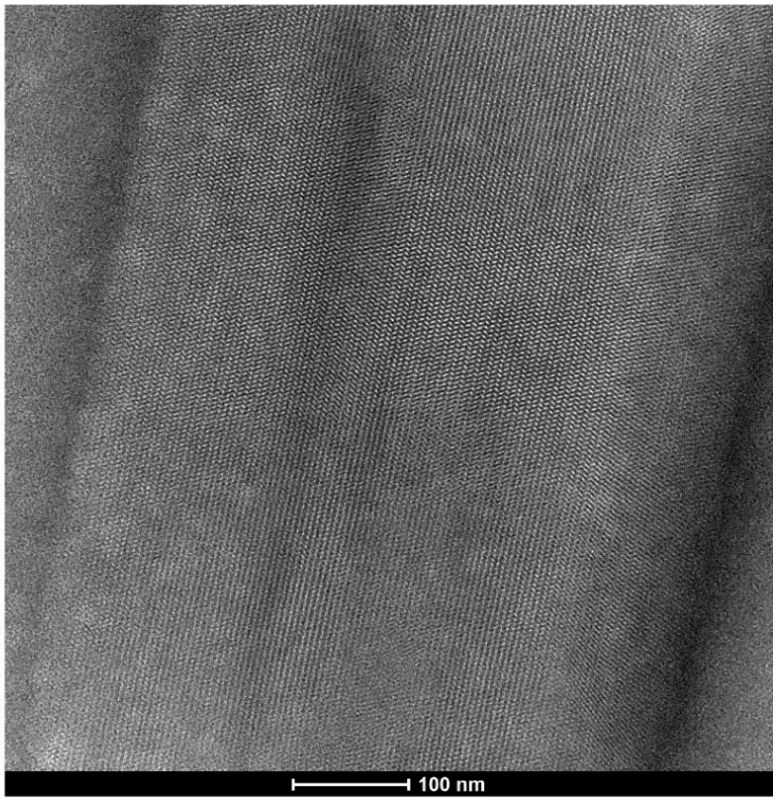


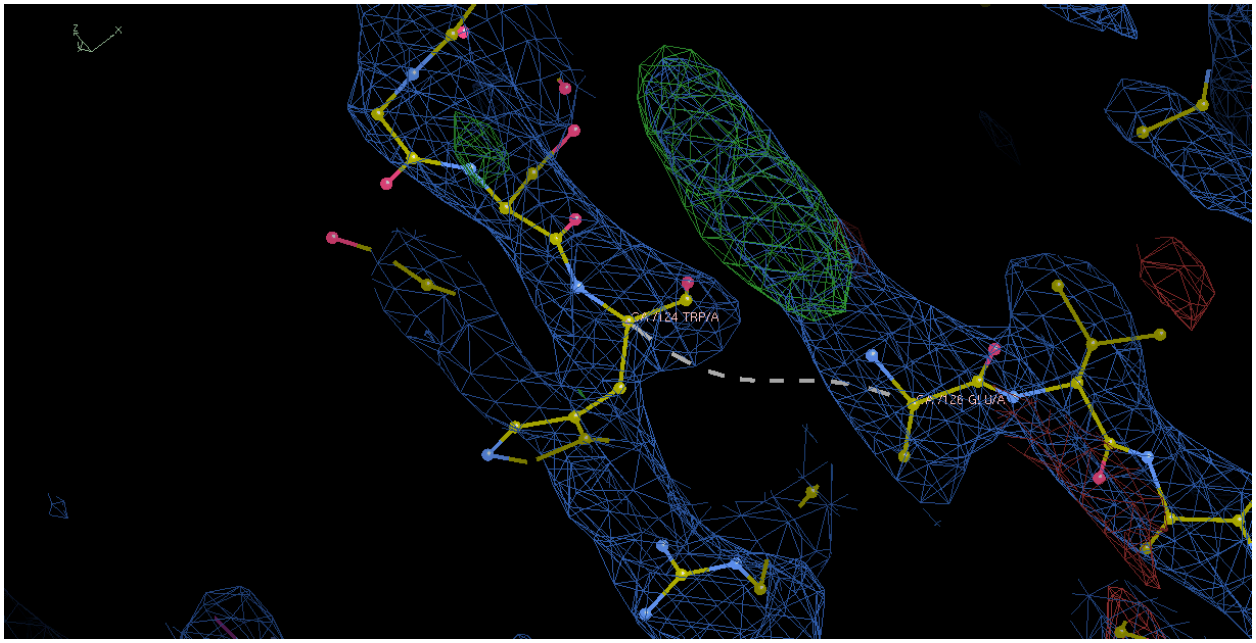
Figure 5. Structural comparison of monomers from three crystal forms: Cartoon representation of the structure solved from form 1 crystals (pink), form 2 crystals (yellow), and form 3 crystals (grey).

Appendix Figure A.S1:



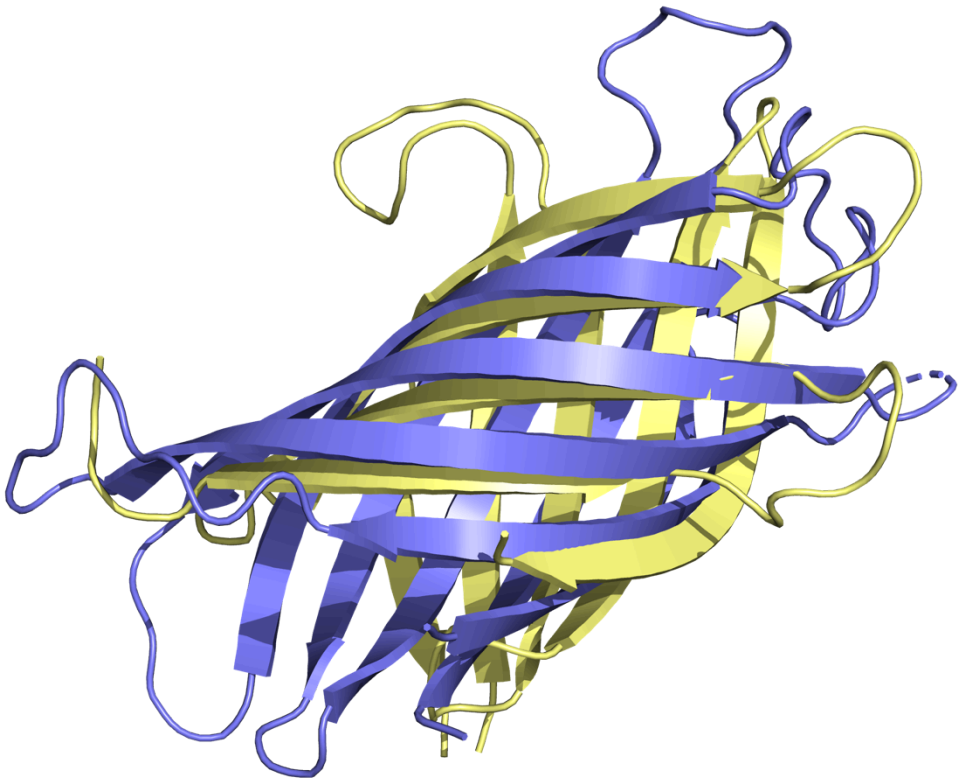
Supplemental Data Figure 1: Negatively stained Q63NT7 crystal visualized on a Talos F200C electron microscope.

Appendix Figure A.S2:



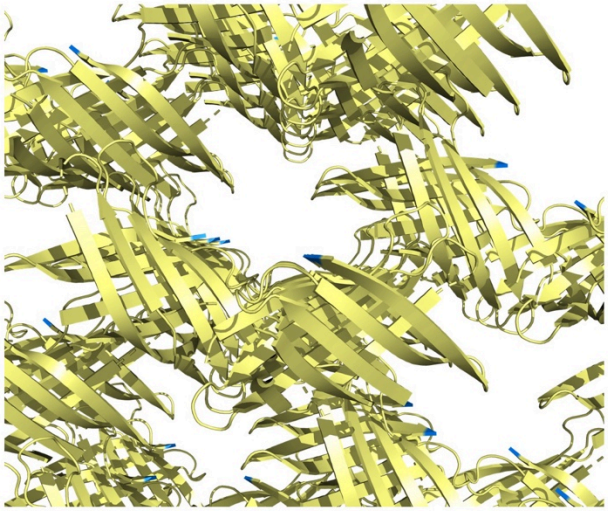
Supplemental Data Figure 2: A Micro-ED omit-map confirms the correct molecular replacement solution when using the AlphaFold search model on form 2 diffraction data. Histidine 125 was deleted from the search model, and density appears for this residue in an F_o-F_c map calculated using model phases. Molecular replacement search model shown as atomic model, green density corresponds to positive density in F_o-F_c map.

Appendix Figure A.S3:



Supplemental Data Figure 3: Comparison of the closest identifiable homolog of known structure (PDB 2erv) with the experimental structure of protein Q63NT7. The 2erv structure is shown as a cartoon in purple overlayed with the form 1 crystal structure of Q63NT7 (yellow).

Appendix Figure A.S4:



Supplemental Data Figure 4: Crystal packing for the form 1 crystal of the Q63NT7 protein reveals solvent channels at the C-terminus of the β-barrel domain.

Appendix Figure A.S5:

Supplemental Data Figure 5: SDS-PAGE analysis of Form 3 crystals reveals prominent degradation products for the Q63NT7 protein.

Appendix Table A1:

Supplemental Data Table S1: Macromolecule Production

Source Organism	*Burkholderia pseudomallei*
DNA Source	Synthetic
Expression vector	Pet 22b (+)
Plasmid Construction method	Gibson assembly
Expression host	*Escherichia coli* (BL21 (DE3))
Expression details	Autoinduction[24]
Complete amino-acid sequence of the protein produced: MSEDLRVGLFPVRYLVGTGLPGAPQLVLDLMVDTVDHSVVGRAAVSQAVSPPLNFHADVWGS YVFRLGPPPRRDGSGAIVQISLQGNQGGPQSNSMITFYGELLLKGDGKTGVASYRYYSNGSW HEVENVPVKADPELVPIEPGPVIGQSSMSAIGSAAMYGVAIQSAAASGDLAHMRTLSAYARQQL ESRDEIAAALSELKAEIAKLESRQHHHHHH	

Appendix Table A2:

Supplemental Data Table S2: Crystallization Form 1 Crystals

Method	Hanging drop
Plate type	96 well
Temperature (°C)	20
Protein Concentration	20 mg/ml
Buffer composition of protein	100mM BisTris pH 5.5, 25% PEG 3350
Volume and ratio of drop	2:1
Drop setting	SPT LabTech Mosquito
Seeding	No

Appendix Table A3:

Supplemental Data Table S3: Crystallization Form 2 Crystals

Method	Hanging drop
Plate type	96 well
Temperature (°C)	20
Protein Concentration	20 mg/ml
Buffer composition of protein	100mM BisTris pH 5.5, 100mM Ammonium Acetate 17% PEG 10,000
Volume and ratio of drop	2:1
Drop setting	SPT LabTech Mosquito
Seeding	No

Appendix Table A4:

Supplemental Data Table S4: Crystallization Form 3 Crystals

Method	Hanging drop
Plate type	96 well
Temperature (°C)	20
Protein Concentration	100 mg/ml
Buffer composition of protein	100mM TRIS HCl pH 8.5, 150mM MgCl, 12.5% PEG 8000
Volume and ratio of drop	1:1
Drop setting	SPT LabTech Mosquito
Seeding	No

Appendix Table A5

Supplemental Data Table 5: Data Collection and Processing

Crystal Form	Form 1	Form 2	Form 3
PDB code	8T0B	8T1N	8T1M
Diffraction source	APS 24-ID-C	Technai TF30	APS 24-ID-E
Wavelength	1.4586	0.01969	0.97918
Temperature (K)	100	100	100
Detector	DECTRIS PILATUS 6M-F	TVIPS TemCam-F416 (4k x 4k)	DECTRIS EIGER X 16M
Crystal to Detector distance (mm)	200	5280	400
Total Rotation Range (°)	180	70	70
Rotation per image (°)	0.5	0.85	0.5
Exposure time per image (s)	0.25	10	0.5
Space group	$P2_1$	$P2_12_12_1$	$P2_12_12_1$
a,b,c (Å)	39.5,40.4,78.5	40.6,95.0,101.5	40.11,70.82,94.5
α, β, γ (°)	90,97.01,90	90,90,90	90,90,90

Mosaicity (°)	0.184	0.356 (2,3,4) 0.157 (7)	0.176 (0-70) 0.188 (140-210)
Resolution Range (Å)	77.9-2.1 (2.15-2.10)	35.0-3.0 (3.10-3.02)	47.3-3.00 (3.08-3.00)
Total no. of reflections	45764	39900	13935
No. of unique reflections	25176	4846	5084
Completeness (%)	89.1 (82.9)	58.8(44.2)	87.9 (88.8)
Redundancy	1.8	11.7	2.74
$\langle I/\sigma(I)\rangle$	8.9 (1.4)	5.6(2.5)	4.4 (2.3)
$CC_{1/2}$	99.9 (76.6)	91.2 (14)	97.2 (29.9)
$R_{\text{r.i.m.}}$	0.059 (0.701)	0.386 (0.501)	0.272 (1.36)
Overall B factor from Wilson plot ($Å^2$)	45.7	24.0	56.3

Appendix Table A6

Supplemental Data Table S6: Refinement Statistics

Resolution range (Å)	77.9-2.10 (2.17-2.10)	35.0-3.02 (3.80-3.02)	47.25-3.00 (3.30-3.00)
Completeness (%)	95.0 (90)	58.8 (57)	87.7 (89)
No. of reflections, working set	12543	4581	4562
No. of reflections, test set	1395	242	508
R_{work}	25.1 (34.3)	28.3 (32.3)	27.4 (32.9)
R_{free}	28.7 (35.0)	30.7 (34.5)	33.3 (42.2)
No. of non-H atoms:			
Protein	1752	1740	1709
Ions	0	0	0
Ligands	0	0	0
Waters	11	0	0
Total	1763	1740	1709
RMSD Bond Lengths (Å)	0.008	0.014	0.011
RMSD Bond Angles (°)	0.95	1.59	1.44
Average B factors (Å^2):			
Protein	50.0	12.35	49.57
Waters	48.7	N/A	N/A

Ramachandran Outliers (%)	0	0	0
Ramachandra Favored (%)	96.1	95.7	96.9
Unmodelled/incomplete residues (%)	7.6	7.6	8.5