Title
Design, Structural Characterization and Application of High Symmetry Protein Nanocages

Author
Cannon, Kevin Alexander

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Design, Structural Characterization and Application of High Symmetry Protein Nanocages

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Chemistry

by

Kevin Alexander Cannon

2019

ABSTRACT OF THE DISSERTATION

Design, Structural Characterization and Application

of High Symmetry Protein Nanocages

by

Kevin Alexander Cannon

Doctor of Philosophy in Chemistry

University of California, Los Angeles, 2019

Professor Todd O. Yeates, Chair

In nature, it is extremely common to find proteins that assemble into homo-oligomeric complexes from multiple copies of themselves. Almost half of known proteins form such complexes, most of which are cyclically or dihedrally symmetric. In some exceptional cases, however, protein molecules will self-assemble into much larger, closed three-dimensional geometries resembling the Platonic solids. Examples include icosahedral viral capsids, bacterial microcompartment shells, and octahedral ferritin assemblies. Protein scientists have studied and marveled at these exquisite protein cage structures for decades, and some have even ventured to produce novel types of protein cage assemblies unseen in nature through their own engineering efforts. In recent years, the field of protein design has seen striking progress in the development of design methodologies for taking proteins found in nature and modifying them to self-assemble into cages of tetrahedral, octahedral, or icosahedral point group symmetry, and these unique new

types of protein assemblies are even beginning to find use in medicine, imaging, and biomaterials applications. My thesis work addresses both the design and application areas of the field of symmetric protein cage design.

In Chapter 1, I include a recent review article on high symmetry protein assemblies, both natural and designed. A survey of all known structures in the Protein Data Bank that self-assemble into unique complexes with tetrahedral, octahedral, or icosahedral symmetry gives context for the types of biological functionality that seem to necessitate or benefit from such higher-order symmetries, although some intriguing mysteries remain unsolved. Our comparison of natural protein assemblies to the recent types of designed protein cages also emphasizes some unique properties of designed cages that remain unseen in natural assemblies.

Next, I go on to describe some recent efforts to improve cage design methods to make cages that more reliably self-assemble into desired architectures when produced in the laboratory. In Chapter 2, we describe the design and characterization of two tetrahedral protein cage assemblies which were engineered to have hydrogen bonding networks at the interface between their two oligomeric components. These cages exhibit exceptionally high levels of soluble expression compared to most previous designed cages, but atomic structures solved by X-ray crystallography reveal some surprising deviations from the designed models.

In Chapter 3, I describe efforts to design and characterize a protein icosahedron that self-assembles from 60 copies of a single designed protein subunit. To date, a designed icosahedral protein assembly formed from genetically fused protein oligomers (as opposed to multiple proteins self-assembling with a computationally designed interface) has yet to be validated in atomic detail. Challenges in achieving this goal have made it clear that novel, alternative design strategies are necessary. We describe the creation of a double-fusion protein containing dimer-,

trimer-, and pentamer-forming protein domains in a single protein construct, which forms an icosahedral assembly when overexpressed in bacteria. The cage assembly is characterized by electron microscopy, small angle X-ray scattering, and other solution-state methods.

I then go on to describe a project which applies a previously characterized tetrahedral protein cage scaffold as a platform for the multivalent display of cellulase enzymes. In Chapter 4, we describe the utilization of the T33-21 cage scaffold as a platform for covalently fusing other proteins to the exterior of the cage post-translationally using a sortase ligation method. In this work, we attach two different cellulase enzymes simultaneously to the cage scaffold and demonstrate increased synergy between the two enzymes in cellulose degradation assays.

In conclusion, the work described in this thesis contributes to the ongoing development of novel design methodologies for engineering high symmetry protein cages, including some important lessons learned along the way, and goes on to describe the application of a designed cage scaffold as a multi-enzyme display platform for potential use in biofuels and other biomaterials technologies.

The dissertation of Kevin Alexander Cannon is approved.

David S. Eisenberg

Bruce Dunn

Todd O. Yeates, Committee Chair

University of California, Los Angeles

2019

For Mom, Dad, Grandma, and Nicole.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

Graduate school is a long and rocky road, an endeavor that cannot possibly be taken on alone. I would not have made it past day one without the help and support of many mentors, friends, and family. First and foremost, I must thank my thesis advisor, Dr. Todd Yeates, for giving me the opportunity to learn and grow as a scientist and as a person for the last five and a half years in his laboratory. My experience there has had a positive impact on my personal and professional growth in more ways than he could know.

There are many other mentors that I have had the great fortune of working with in my time at UCLA. In the Yeates Lab, I must thank several previous graduate students who trained me in everything I now know about biochemical lab techniques: Dr. Yen-Ting Lai, Dr. Nicole Wheatley, Dr. Sunny Chang, and especially Dr. Yuxi Liu, without whom I would be completely and utterly lost. I also thank other members of the Yeates Lab for their friendship as well as constant helpful discussions and advice: Joshua Laniado, Jessica Ochoa, Justin Miller, Kyle Meador, Matthew Agdanowski, Eric Lee, and our lab manager Inna Pashkov. I also thank Dr. Duilio Cascio and Dr. Michael Sawaya for their incredible teaching, assistance, and encouragement in all things related to structural biology and life. Further, I would like to acknowledge the people who opened the door to Electron Microscopy for me, especially David Boyer, Dr. Peng Ge, and other members of the UCLA EICN staff. I should also thank the amazing scientists from my undergraduate research lab at NYU that inspired me to pursue research and graduate school in the first place, especially Dr. Ned Seeman, Dr. Yoel Ohayon, Dr. Nam Nguyen, Dr. Arun Chandrasekaran, Dr. Jennifer Padilla, and Dr. Ploi Udomprasert.

I have also had the privilege of mentoring two incredible undergraduate researchers who contributed significantly to much of the work presented in this dissertation: Christian Morgan

and Vy Nguyen. Without their enthusiasm, dedication, and hard work, I could not have completed the research presented here.

Outside of the laboratory, there are many friends near and far that kept me sane during the most trying times of grad school, whether they knew it or not. To name just a few: Andrew, Anthony, Beccah, Chris, Collin, Eleni, Frances, Gio, Janice, Jayhee, Jeff, Kanishk, Kelsey, Linda, Nick H., Nick K., Nick W., Nikie, Robin, Ryan, Shaun, Sherry, and Zach.

I also thank my family: Mom, Dad, Jeff, Paul, and Grandma. Words could never express how lucky and grateful I am to have you all as role models for everything I do.

And last, but not least, I would like to thank my partner, Nicole. To borrow a phrase from Yeates Lab members past: she is the greatest discovery I made in grad school. Without her support, I'm quite sure I would have quit a thousand times before making it to this point.

In this dissertation, Chapter 1 is an adapted version of Cannon, K. A., Ochoa, J. M. & Yeates, T. O. High-symmetry protein assemblies: patterns and emerging applications. Current Opinion in Structural Biology 55, 77–84 (2019). DOI: 10.1016/j.sbi.2019.03.008. The list of proteins that self-assemble into cubic symmetries and their properties was prepared by KAC. Analysis of the assembly properties of proteins in the list was performed by KAC and TOY. Figures were prepared by JMO and TOY. The written manuscript was prepared by KAC, JMO and TOY. TOY was the principal investigator.

Chapter 2 is a manuscript in preparation for submission. KAC carried out X-ray crystallography experiments. X-ray diffraction data was analyzed by KAC and TOY. Protein design, expression, and purification were performed by RUP, SEB, SY, UN and NPK. The

written manuscript was prepared by KAC, NPK, and TOY. DB, NPK, and TOY were the principal investigators.

Chapter 3 is a manuscript in preparation for submission. Protein design was done by KAC. Cloning, expression, and purification of protein constructs were carried out by KAC, VN, and CM. Electron microscopy was performed by KAC. Analysis of small angle X-ray scattering data was performed by TOY. The written manuscript was prepared by KAC and TOY. TOY was the principal investigator.

Chapter 4 is a manuscript in preparation for submission. Project design was carried out by SAM, KAC, BRA, TOY, and RTC. Preparation of cellulase and sortase enzyme constructs was carried out by SAM and BRA. Preparation of protein cage constructs was performed by KAC and CM. Sortase ligation experiments and biochemical assays were carried out by SAM. Electron microscopy was performed by KAC. The written manuscript was prepared by SAM, KAC, TOY, and RTC. TOY and RTC were the principal investigators.

Vita

Education

2013, B.S. Chemistry, New York University

2015, M.S. Chemistry – Materials, University of California, Los Angeles

Publication

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

*These authors contributed equally.

Chapter 1. High-Symmetry Protein Assemblies: Patterns and Emerging Applications

Kevin A. Cannon[1,2], Jessica M. Ochoa[1,3], Todd O. Yeates[1,2,3]

[1]UCLA Department of Chemistry and Biochemistry

[2]UCLA-DOE Institute for Genomics and Proteomics

[3]UCLA Molecular Biology Institute

KAC and JMO contributed equally.

Keywords: protein assembly; symmetry; biotechnology, protein design; encapsulation

ABSTRACT

The accelerated elucidation of three-dimensional structures of protein complexes, both natural and designed, is providing many new examples of large supramolecular assemblies with intriguing shapes. Those with high symmetry – based on the geometries of the Platonic solids – are particularly notable as their innately closed forms create interior spaces with varying degrees of enclosure. We survey known protein assemblies of this type and discuss their geometric features. The results bear on issues of protein function and evolution, while also guiding novel bioengineering applications. Recent successes using high-symmetry protein assemblies for applications in interior encapsulation and exterior display are highlighted.

INTRODUCTION

Nearly half of all known natural proteins form homo-oligomeric complexes through the assembly of multiple copies of the same (or homologous) subunits [1–3]. Such structures have evolved for diverse purposes, ranging from cooperative binding behavior (as exemplified by hemoglobin) to architectural function (as exemplified by microtubules and many other large cellular structures). Yet specific functional advantages are known for only a fraction of the vast number of homo-oligomeric protein and enzyme assemblies seen in nature. The reasons for protein oligomerization have been discussed as far back as Monod [4] and expertly discussed by Goodsell [5]. The wide-ranging advantages that have been offered as explanation – enhanced stability, mitigation of cellular crowding, functional regulation, to name a few – suggest the possibility of multifunctional advantages. Another perspective is that the abundance and diversity of explanations for homomeric assemblies belies an incomplete understanding of the

phenomenon, suggesting that continued investigations of trends and outlying cases might be informative.

A nearly universal observation is that the subunits in homo-oligomeric assemblies are arranged in symmetric ways. Compared to the more open question of why so many proteins form homo-oligomers in the first place, the explanation of symmetry is clearer [5]. Symmetric arrangements require the fewest distinct kinds of interfaces between equivalent subunits. This makes symmetric arrangements more likely to occur through natural evolutionary events, as articulated by Crick and Watson in their prescient 1956 prediction that viral capsids would assemble according to cubic symmetries [6]. The idea of minimum contact types has also guided developments in the area of designing highly symmetric protein assemblies.

Within the natural hierarchy of symmetry types in three dimensions, from cyclic to dihedral to cubic, the latter category offers special features for investigation and exploitation. By their nature, cubic symmetries take the forms of the Platonic solids. They are therefore closed assemblies, like a cage or shell, with defined interiors. As limiting cases or outliers on the symmetry spectrum, cubic protein assemblies offer intriguing case studies. As with other symmetric homo-oligomeric assemblies, the functional purpose for cubic symmetry is clear in some cases (i.e. viral capsids) and less clear in others. In addition to their potential biological implications, cubic assemblies offer unique advantages in various bioengineering applications. Recent studies have begun to explore a range of novel uses, including interior encapsulation and exterior multivalent display. Below we survey the known cubic protein assemblies, natural and designed, both as interesting cases for biological insight and starting points for diverse applications in medicine, biomaterials, and synthetic biology.

HIGH-SYMMETRY ASSEMBLIES IN THE PDB

Known Cubic Structures

The Protein Data Bank (PDB) provides a rich source of information about oligomeric protein structures. However, discerning whether observed subunit arrangements in crystals represent biologically relevant assembly states remains a long-standing and challenging problem [7–10]. A recent computational analysis by Levy and colleagues (available online at www.qsbio.org as the QSbio database) provides arguably the most robust assignments to date, with an estimated error rate of 15% for predicting true biological assembly forms [11]. The occurrence of cubic assemblies among natural proteins is rare enough that manual curation of the literature is feasible. Thus, we were able to augment computational inferences in the QSbio database, which are based largely on geometric analysis of interfaces and their preservation across structures of homologous proteins. Literature analyses showed the confidence assignments from QSbio to be generally conservative; several prospective cubic assemblies bearing 'low' or 'very low' confidence ratings could be validated by solution data. Excluding viral capsids, and counting only structurally unique representatives, the set of cubic protein assemblies – tetrahedral, octahedral, and icosahedral – includes 46 unique natural protein assemblies, 12 designed protein cages, and 5 other assemblies we designate as "semi-synthetic" (see supplemental materials).

The known cubic assemblies span a wide range of sizes with distinct geometric properties, which may be relevant for function and critically important for prospective engineering applications. For each structure, Figure 1 shows the size of the interior cavity as well as the size of the largest opening or 'window' from the exterior. The former quantity bears

on encapsulation capacity while the latter relates to porosity and access to the interior by other molecules. Suitable ranges for these parameters will vary depending on the application. Applications related to molecular containment and delivery will benefit from large cavity size and small window size. This has been articulated in recent studies (discussed below) where the goal was to encapsulate large nucleic acid molecules without permitting access to nucleases of varying sizes [12,13].

Connections to Function and Evolution

Relatively rare cases where high-symmetry assemblies have arisen in nature prompt questions about function. In many instances, a closed shape is clearly necessary for a protein's native function. Viruses require a capsid to protect their encapsulated genetic material; ferritins form a shell to store iron; chaperonins encapsulate misfolded proteins to promote re-folding; and so on. However, numerous proteins form closed, highly symmetric structures without any obvious functional explanation [14,15]. It has been suggested that protein cage formation could have evolved in some enzymes to protect them against environmental stressors [16–18], such as in the case of ornithine carbamoyltransferase in thermophilic bacteria (PDB: 1A1S). There the catalytically competent homotrimeric unit seen in other species assembles further into a 12-mer tetrahedron (i.e. a tetramer of trimers), with the increased subunit interactions presumed to contribute to stability [19]. A somewhat similar argument has been presented for the 2-hydroxypentadienoic acid hydratase enzyme from *E. coli* (PDB: 2WQT). Its formation of a tightly packed 60-mer icosahedral cage from 12 copies of a pentameric unit seems to be restricted to extreme conditions (e.g. low pH and high phosphate buffer concentrations). Most proteins of the same family carry out their function as simple C5 pentamers, and it remains

unclear what environmental conditions in the cell, if any, might trigger cage formation in a biological setting [15].

Some cubic assemblies exploit geometric advantages that arise from spatial clustering of enzymatic active sites. This can improve the flux through pathways that involve multiple sequentially-acting enzymes. The pyruvate dehydrogenase complex (PDC) is a well-studied case of a multi-enzyme system built around a highly symmetric core. Early work by Hol et al. demonstrated that the porous cage-like structure of the complex can take the form of either a 24-mer cube (PDB: 1DPB) or a 60-mer dodecahedron (PDB: 1B5S) depending on the species of origin [20]. Long, flexible poly-peptide tails on the exterior of the assembly recruit other enzymes in the complex to perform their functions. The pyruvate substrate is shuttled via a ping-pong mechanism from one exterior enzyme to the octahedral or icosahedral core's active site and back out to the third enzyme in the sequence to complete the conversion to acetyl-CoA**)** [21,22]. A recent 3.1 Å structure obtained by cryo-electron microscopy demonstrates the first high-resolution structure of a mammalian PDC dodecahedral core and provides new insights into its properties in solution [23].

The integrity of cubic assemblies requires multiple distinct types of interactions between subunits, which naturally suggests evolutionary routes from simpler assembly forms. Previous studies have detailed likely pathways for the evolution of higher-order symmetric structures, with lower symmetry forms of the protein assemblies presumed to represent evolutionary intermediates [24]. Understanding how simpler symmetries can combine together in order to give rise to higher symmetry cubic structures has also provided a foundation for recent developments in designing novel protein cages and other geometrically ordered materials.

Designed Assemblies

Highly symmetric natural protein assemblies have motivated design efforts in the creation of novel, symmetric structures. Symmetry-based design principles and the earliest designed protein cage were described by Padilla et al. [25]. There a tetrahedral cage was formed by the genetic fusion of simple oligomeric protein components (i.e. dimers and trimers) by continuous alpha helical linkers. Tetrahedral and octahedral designs were later demonstrated using this approach [26–28]. Larger collections of symmetric cages, generally less porous and more rigid than those created by the fusion approach, were created by King, Bale, Baker et al. based on computational amino acid sequence design of new interfaces between oligomeric components [29–32]. As described below, some of the cages created by interface design are finding utility in novel applications.

Designing novel protein assemblies remains challenging, as engineered proteins often present difficulties in protein folding, expression and proper assembly. To date, 12 designed cages have been validated in atomic detail by x-ray crystallography and have been deposited in the PDB. These include seven tetrahedral structures, two octahedral structures, and three icosahedral structures. The geometries of these successful designs are described in Fig. 1. Seven additional icosahedral assemblies based on interface design have been shown to form symmetric particles by electron microscopy [31,32]. Two additional cages – one tetrahedral and one octahedral – have been designed by more flexible fusions between a trimeric protein and various coiled coil segments. The designs were confirmed at the level of low resolution EM [33–35]. Designs in these latter groups are presumed to represent successful designs but have not been validated in atomic detail.

Different design strategies have led to geometric protein structures of somewhat less regular forms, including some that are smaller [36,37] and some that are larger [38] than the highly symmetric ones discussed here. Moreover, introducing metal-binding sites into simpler oligomeric building blocks has also fortuitously led to cage-like structures in other design studies [39].

Augmenting the designed novel structures above, a few recent studies have re-engineered proteins already known to form cages or shells, in order to create intriguing alternative assembly forms. We refer to these structures as 'semi-synthetic' (Fig. 2c). In these studies, exploring sequence mutations or variations in subunit composition have led to novel structures whose detailed forms, which could not be divined in advance, were illuminated by x-ray crystallography and cryo-EM. Jorda et al. fortuitously obtained a T=1 60-mer icosahedral cage from a circular permutation of a bacterial microcompartment (BMC) shell protein [40]. Kerfeld and coworkers demonstrated a larger T=9 icosahedral shell built from a subset of BMC shell proteins present in a native bacterial microcompartment [41]. In another remarkable set of structures, Hilvert and coworkers have shown the scalability of lumazine synthase assemblies, demonstrating that by adding negatively charged residues to the interior of the icosahedral cage, they could generate expanded cage forms of either 180 or 360 subunits [42].

The growing suite of natural, designed, and semi-synthetic protein assemblies is providing new opportunities for diverse applications.

RECENT APPLICATIONS AND FUTURE DIRECTIONS

Protein cages or capsids have been explored for numerous applications. Viral capsids have been widely exploited in prior work, and a few non-viral proteins, including ferritin [43]

and the thermophilic heat shock protein Hsp have been similarly investigated [44–47]. Those systems have been widely reviewed [48,49], including recently by Orner et al [50]. Among large but non-cubic assemblies, eukaryotic vaults [51,52] and chaperonins have also been explored [53,54]. Here we focus on some of the most recent studies and highlight a few newly emerging protein systems.

Interior Encapsulation

The closed nature of cubic assemblies naturally leads to prospective applications in containment and delivery. Ferritin, a ubiquitous and well-characterized natural octahedral protein assembly (PDB: 3BVE, Fig 2), has unique physical and chemical properties that have made it a work-horse in previous applications in drug delivery, vaccine development, bioassays and molecular imaging (reviewed by López-Sagaseta et al. and He and Marles-Wright) [55,56]. The rigid ferritin cage encloses an 8-nm interior cavity, is capable of reversibly disassembling under acidic conditions, and is biocompatible. Prior studies have shown ferritin's capacity to encapsulate and deliver anti-tumor drugs [57–60]. Recent studies have demonstrated ferritin's ability ferritin to encapsulate doxorubicin and cross the blood-brain barrier [61], and new work by Fan et al. showed that ferritin could selectively target glioma cells and release its drug payload to kill tumor cells *in vivo* [62].

Additional cases of non-viral protein cages provide new prospective applications, with each system offering specific advantages in terms of geometry, amenability to mutagenesis and heterologous expression, and chemical and physical stability (as explored recently by Heinze et al. [63]). Taking a designed icosahedral cage as a starting point, Butterfield et al. showed that mutated versions of this cage (modified by introduction of interior positive charge or RNA

binding motifs) could encapsulate the RNA molecules encoding the capsid shell proteins, mimicking the way a virus contains its genome within its own capsid [13]. Subsequent rounds of optimization were employed, with roughly 9% of the cages successfully encapsulating mRNA. The interior RNA binding was non-sequence specific, but packaging was strongly correlated with expression level such that 74% of encapsulated RNA was found to encode the capsid. Capsids were stable in blood for up to 6 hours and capable of circulating *in vivo* for 4.5 hours.

Recent studies have similarly exploited the lumazine synthase system as a framework for evolving a nucleocapsid [64]. Mutants bearing a designed RNA-binding peptide tag fused to the lumenal side of the capsid achieved comparable levels of mRNA encapsulation, with ~10% of capsids containing the full-length cage genome [65]. In another case, Azuma et al. were able to engineer a distinct icosahedral lumazine synthase variant to encapsulate smaller mRNA sequences (up to 300-nt) with tunable size-selectivity via different lengths of poly-arginine tags on the lumenal surface of the cage [66]. Controlling the length of encapsulated nucleic acids provides an element of partial selectivity. In another study, a smaller (octahedral) designed cage was mutated to encapsulate short nucleic acid segments (approximately 21 nucleotides), again by addition of interior positive charge [12]. Because of the large 35 Å windows in this cage, small guest nucleic acids could be added after protein purification in order to incorporate specific RNA molecules. However, these openings were large enough for nucleases, such as RNase A, to also gain access. The cages were taken up by mammalian cells via endocytosis and cargo RNA was released inside the cytoplasm upon competitive binding of native tRNA to the cage. Loading the cages with siRNA led to successful gene knockdown of GFP and showed low levels of cytotoxicity. These novel cage systems are opening new avenues in drug delivery applications,

with concomitant challenges related to nucleic acid sequence specificity, immunogenicity, and susceptibility to cellular nucleases and proteases.

Exterior Display

The high copy number of cubic assemblies offers prospects for polyvalent external attachments for varied purposes (Fig 3). Virial capsids, ferritin and other diverse assemblies have been popular choices in pioneering efforts to functionalize the outer surfaces of protein cages, including for vaccine design and therapeutic and biomaterials applications [43,67–72]. New work by Dostalova and coworkers showed that decorating the surface of Dox-loaded apoferritin cages with mouse antibodies that target a prostate-specific membrane antigen leads to more reliably targeted payload release and lower off-target toxicity [73].

Protein cages of the designed variety have also started to find applications for exterior display. Votteler et al. adapted designed icosahedra to form extracellular vesicles *in vivo via* genetic fusion of a membrane-binding peptide motif on the cage's exterior. These cages were able to achieve cellular escape by recruiting endosomal sorting complexes required for transport (ESCRT) machinery, thus demonstrating a viable technology for the transfer of molecular cargo from one cell to another [74].

Inspired by complexes that bring the active sites of sequentially-acting enzymes into spatial proximity to improve pathway flux [75], a recent study has demonstrated the use of a designed cage as a scaffold for the multi-copy display of enzymes that break down plant-derived cellulosic material to glucose. Attaching an endocellulase and an exocellulase to the exterior of a designed cage increased enzymatic activity by more than 5-fold compared to free enzymes in solution (McConnell, Cannon et al., unpublished). The power of high valency was also

emphasized in a different kind of application in a recent study by Phippen et al., in which genetically fusing antifreeze protein motifs to the exposed C-terminus of a designed cage resulted in a greater than 50-fold increase in the freezing point depression of water compared to free protein [76].

Recent work has also demonstrated the powerful utility of cubic protein cages as rigid scaffolding for imaging small proteins by cryo-electron microscopy (cryo-EM). Despite a recent technological boom in the cryo-EM field, it has been impossible to gain atomic-level structural detail from proteins that are smaller than ~50kDa, due to a low signal-to-noise ratio in this size class. Designed Ankyrin Repeat Proteins (DARPins) can be obtained by *in vitro* selection methods to wide-ranging target proteins with high affinity. Exploiting their modularity, Liu et al. used DARPins as an adaptor component to overcome the CryoEM size barrier by rigidly fusing DARPins to a tetrahedral cage. By design, the 13-nm particle symmetrically displays 12 copies of the target cargo protein bound to DARPins on the cage's exterior. An 18kDa DARPin itself, not bound to a target, could be visualized by single particle cryo-EM at 3.5 Å to 5 Å resolution [77], and the structure of the first bound cargo protein, super-folder GFP (27 kDa), has been imaged now at a resolution of 3.7 Å by single particle cryo-EM (Liu 2019, in preparation).

Future directions

Future studies will benefit from further examples of highly symmetric protein assemblies and their applications for novel purposes. Broader choices for protein frameworks will enable more-tailored design features to be realized. Relevant properties include: overall shape, interior accessibility, charge distribution, and disposition of chain termini for applications involving genetic fusion to novel components. New strategies for selecting cages with favorable

properties, e.g. robust assembly, from large libraries of mutants will be important for improving the pipeline for new designs. A recent foray by Orner at al. showed how successful cage formation could be connected to a fluorescent readout [78]. As further successes emerge, more complex design schemes will be enabled, including active systems where specific phenomena (e.g. release) are triggered by particular cellular events or components, and where multiple signals can be integrated into systems that exhibit complex digital logic for modern synthetic biology applications.

ACKNOWLEDGEMENTS

The authors thank Joshua Laniado, Justin Miller, and other members of the Yeates lab for ideas and helpful discussions. This work was supported by DOE BER Office of Science award # DE-FC02-02ER63421 and National Science Foundation grant CHE-1629214.

Figure 1

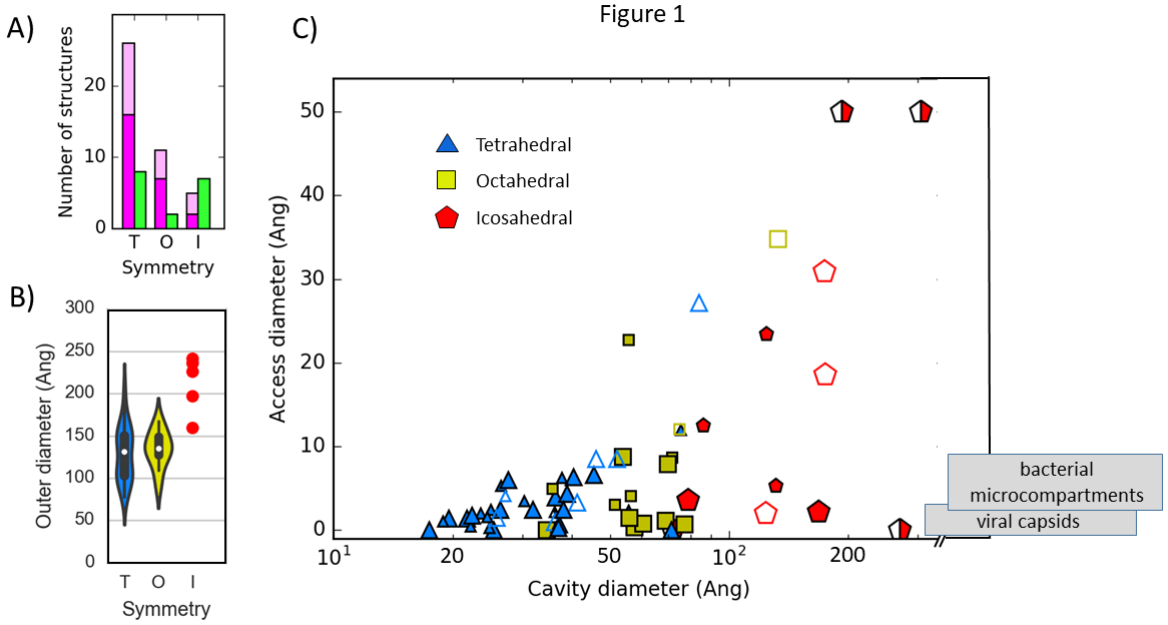


Figure 1.1. Geometric properties of known cubic and icosahedral protein assemblies. (A) Abundance of symmetry types T (tetrahedral), O (octahedral), and I (icosahedral) among natural structures in the PDB. (B) Distribution of sizes (outer diameter). (C) Sizes of interior cavities and window openings in known structures. Symmetry types are indicated by shape and color. Solid symbols indicate natural structures. Empty symbols indicate designed assemblies. Half-filled symbols indicate re-engineered forms of natural cage or shell structures.

FIGURE 2

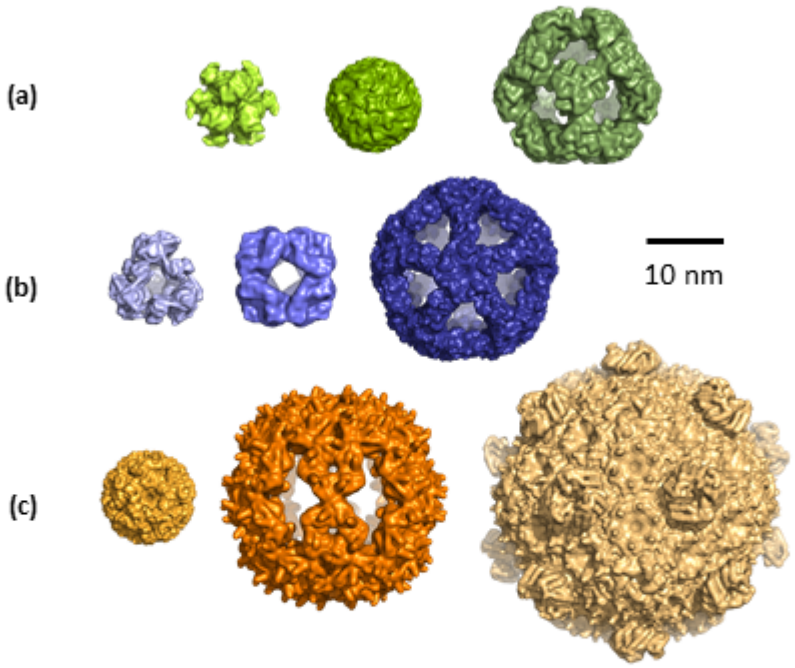


Figure 1.2. Representative structures of cubic assemblies from diverse categories highlighting the diversity in symmetry, shape, size and porosity. (A) Naturally evolved structures. (B) Designed assemblies. (C) Semisynthetic structures obtained as variations on natural cages or shells. PDB: 1A1S, ornithine carbamoyltransferase; PDB:3BVE - *Helicobacter pylori* ferritin; PDB:1B5S - dihydrolipoyl transacetylase; PDB:4QES – Designed tetrahedral cage, 3+2 symmetry type; PDB:3VCD – Designed octahedral cage, 3-fold plus 2-fold interface symmetry type; PDB:5IM4 - 2-component icosahedral cage, 5+2 symmetry type; PDB:5V74 – Icosahedral shell formed by a subset of bacterial microcompartment proteins from *Haliangium ochraceumi*; PDB:5MQ3 – Expanded icosahedral cage based on lumazine synthetase; and PDB: 5HPN - Icosahedral cage from circularly permuted bacterial microcompartment protein PduA.

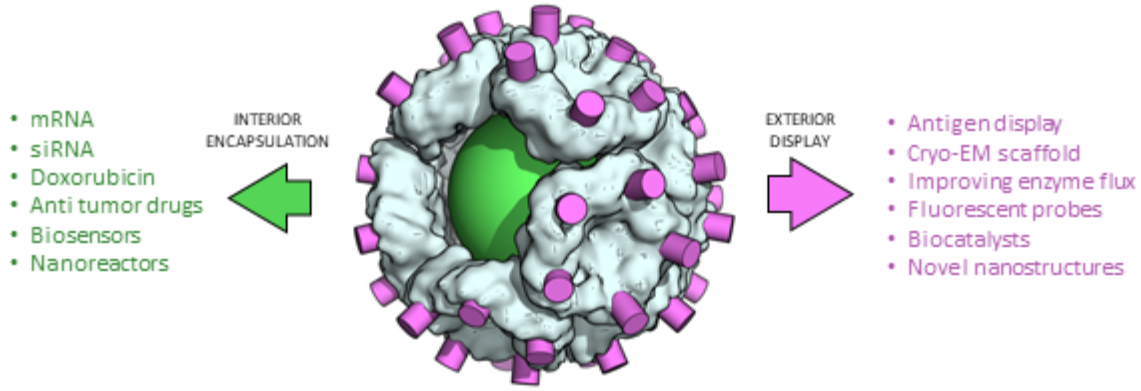


Figure 1.3. Current and future applications for symmetric cubic assemblies. Shells or cages with large interiors allow for encapsulation of other proteins, nucleic acids, metals, fluorophores, and drug molecules. The multivalent nature of cubic assemblies enables exterior high-copy display of antigens, cell-targeting molecule, bioactive motifs, and small proteins for cryoEM imaging.

REFERENCES

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

2. André I, Strauss CEM, Kaplan DB, Bradley P, Baker D: **Emergence of symmetry in homooligomeric biological assemblies**. *PNAS* 2008, **105**:16148–16152.

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

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

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

6. Crick, FHC, Watson, JD: **Structure of Small Viruses**. *Nature* 1956, **177**:473–475.

7. Janin J, Rodier F: **Protein–protein interaction at crystal contacts**. *Proteins: Structure, Function, and Bioinformatics* 1995, **23**:580–587.

8. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends in Biochemical Sciences* 1998, **23**:358–361.

9. Ponstingl H, Kabir T, Thornton JM: **Automatic inference of protein quaternary structure from crystals**. *Journal of Applied Crystallography* 2003, **36**:1116–1122.

10. Krissinel E, Henrick K: **Inference of Macromolecular Assemblies from Crystalline State**. *Journal of Molecular Biology* 2007, **372**:774–797.

11. Dey S, Ritchie DW, Levy ED: **PDB-wide identification of biological assemblies from conserved quaternary structure geometry**. *Nature Methods* 2018, **15**:67–72.

12. Edwardson TGW, Mori T, Hilvert D: **Rational Engineering of a Designed Protein Cage for siRNA Delivery**. *Journal of the American Chemical Society* 2018, **140**:10439–10442.

13. Butterfield GL, Lajoie MJ, Gustafson HH, Sellers DL, Nattermann U, Ellis D, Bale JB, Ke S, Lenz GH, Yehdego A, et al.: **Evolution of a designed protein assembly encapsulating its own RNA genome**. *Nature* 2017, **552**:415–420.

14. McNamara DE, Cascio D, Jorda J, Bustos C, Wang T-C, Rasche ME, Yeates TO, Bobik TA: **Structure of Dihydromethanopterin Reductase, a Cubic Protein Cage for Redox Transfer**. *J Biol Chem* 2014, **289**:8852–8864.

15. Montgomery MG, Coker AR, Taylor IA, Wood SP: **Assembly of a 20-nm Protein Cage by Escherichia coli 2-Hydroxypentadienoic Acid Hydratase**. *Journal of Molecular Biology* 2010, **396**:1379–1391.

16. Ha N-C, Oh S-T, Sung JY, Cha KA, Lee MH, Oh B-H: **Supramolecular assembly and acid resistance of *Helicobacter pylori* urease**. *Nature Structural & Molecular Biology* 2001, **8**:505–509.

17. D'Abusco AS, Casadio R, Tasco G, Giangiacomo L, Giartosio A, Calamia V, Di Marco S, Chiaraluce R, Consalvi V, Scandurra R, et al.: **Oligomerization of Sulfolobus solfataricus signature amidase is promoted by acidic pH and high temperature**. *Archaea* 2005, **1**:411–423.

18. Carter EL, Tronrud DE, Taber SR, Karplus PA, Hausinger RP: **Iron-containing urease in a pathogenic bacterium**. *PNAS* 2011, **108**:13095–13099.

19. Villeret V, Clantin B, Tricot C, Legrain C, Roovers M, Stalon V, Glansdorff N, Beeumen JV: **The crystal structure of Pyrococcus furiosus ornithine carbamoyltransferase**

reveals a key role for oligomerization in enzyme stability at extremely high temperatures. *PNAS* 1998, **95**:2801–2806.

20. Izard T, Aevarsson A, Allen M, Westphal A, Perham R, de Kok A, Hol WGJ: **Principles of quasi-equivalence and Euclidean geometry govern the assembly of cubic and dodecahedral cores of pyruvate dehydrogenase complexes.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**:1240–1245.

21. Tsai CS, Burgett MW, Reed LJ: **α-Keto Acid Dehydrogenase Complexes XX. A KINETIC STUDY OF THE PYRUVATE DEHYDROGENASE COMPLEX FROM BOVINE KIDNEY**. *J Biol Chem* 1973, **248**:8348–8352.

22. Zhou ZH, McCarthy DB, O'Connor CM, Reed LJ, Stoops JK: **The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes**. *PNAS* 2001, **98**:14802–14807.

23. Jiang J, Baiesc FL, Hiromasa Y, Yu X, Hui WH, Dai X, Roche TE, Zhou ZH: **Atomic Structure of the E2 Inner Core of Human Pyruvate Dehydrogenase Complex**. *Biochemistry* 2018, **57**:2325–2334.

24. Levy ED, Erba EB, Robinson CV, Teichmann SA: **Assembly reflects evolution of protein complexes**. *Nature* 2008, **453**:1262–1265.

25. Padilla JE, Colovos C, Yeates TO: **Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:2217–2221.

26. Lai YT, Cascio D, Yeates TO: **Structure of a 16-nm cage designed by using protein oligomers**. *Science* 2012, **336**:1129–1129.

27. Lai YT, Tsai KL, Sawaya MR, Asturias FJ, Yeates TO: **Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly**. *Journal of the American Chemical Society* 2013, **135**:7738–7743.

28. Lai Y-T, Reading E, Hura GL, Tsai K-L, Laganowsky A, Asturias FJ, Tainer J a., Robinson CV, Yeates TO: **Structure of a designed protein cage that self-assembles into a highly porous cube**. *Nature Chemistry* 2014, **6**:1065–1071.

29. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I, Gonen T, Yeates TO, Baker D: **Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy**. *Science* 2012, **336**:1171–1174.

30. 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* 2014, **510**:103–8.

31. Hsia Y, Bale JB, Gonen S, Shi D, Sheffler W, Fong KK, Nattermann U, Xu C, Huang P-S, Ravichandran R, et al.: **Design of a hyperstable 60-subunit protein icosahedron**. *Nature* 2016, **535**:136–139.

32. Bale JB, Gonen S, Liu Y, Sheffler W, Ellis D, Thomas C, Cascio D, Yeates TO, Gonen T, King NP, et al.: **Accurate design of megadalton-scale two-component icosahedral protein complexes**. *Science* 2016, **353**:389–395.

33. Cristie-David AS, Koldewey P, Meinen BA, Bardwell JCA, Marsh ENG: **Elaborating a coiled-coil-assembled octahedral protein cage with additional protein domains**. *Protein Science* 2018, **27**:1893–1900.

34. Sciore A, Su M, Koldewey P, Eschweiler JD, Diffley KA, Linhares BM, Ruotolo BT, Bardwell JCA, Skiniotis G, Marsh ENG: **Flexible, symmetry-directed approach to assembling protein cages**. *Proc Natl Acad Sci USA* 2016, **113**:8681–8686.

35. Badieyan S, Sciore A, Eschweiler JD, Koldewey P, Cristie-David AS, Ruotolo BT, Bardwell JCA, Su M, Marsh ENG: **Symmetry-Directed Self-Assembly of a Tetrahedral Protein Cage Mediated by de Novo-Designed Coiled Coils.** *Chembiochem* 2017, **18**:1888–1892.

36. Lapenta F, Aupič J, Strmšek Ž, Jerala R: **Coiled coil protein origami: from modular design principles towards biotechnological applications.** *Chem Soc Rev* 2018, **47**:3530–3542.

37. Gradišar H, Božič S, Doles T, Vengust D, Hafner-Bratkovič I, Mertelj A, Webb B, Šali A, Klavžar S, Jerala R: **Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments.** *Nat Chem Biol* 2013, **9**:362–366.

38. Fletcher JM, Harniman RL, Barnes FR, Boyle AL, Collins A, Mantell J, Sharp TH, Antognozzi M, Booth PJ, Linden N, et al.: **Self-assembling cages from coiled-coil peptide modules.** *Science* 2013, **340**:595–599.

39. Ni TW, Tezcan FA: **Structural characterization of a microperoxidase inside a metal-directed protein cage**. *Angewandte Chemie - International Edition* 2010, **49**:7014–7018.

40. Jorda J, Leibly DJ, Thompson MC, Yeates TO: **Structure of a novel 13 nm dodecahedral nanocage assembled from a redesigned bacterial microcompartment shell protein**. *Chemical Communications* 2016, **52**:5041–5044.

41. Sutter M, Greber B, Aussignargues C, Kerfeld CA: **Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell**. *Science* 2017, **356**:1293–1297.

42. Sasaki E, Böhringer D, Van De Waterbeemd M, Leibundgut M, Zschoche R, Heck AJR, Ban N, Hilvert D: **Structure and assembly of scalable porous protein cages**. *Nature Communications* 2017, **8**:1–10.

43. Wang Z, Gao H, Zhang Y, Liu G, Niu G, Chen X: **Functional ferritin nanoparticles for biomedical applications**. *Front Chem Sci Eng* 2017, **11**:633–646.

44. Murata M, Narahara S, Umezaki K, Toita R, Tabata S, Piao JS, Abe K, Kang JH, Ohuchida K, Cui L, et al.: **Liver cell specific targeting by the preS1 domain of hepatitis B virus surface antigen displayed on protein nanocages.** *Int J Nanomedicine* 2012, **7**:4353–4362.

45. Toita R, Murata M, Abe K, Narahara S, Piao JS, Kang JH, Hashizume M: **A nanocarrier based on a genetically engineered protein cage to deliver doxorubicin to human hepatocellular carcinoma cells.** *Chem Commun (Camb)* 2013, **49**:7442–7444.

46. Uchida M, Kosuge H, Terashima M, Willits DA, Liepold LO, Young MJ, McConnell MV, Douglas T: **Protein cage nanoparticles bearing the LyP-1 peptide for enhanced imaging of macrophage-rich vascular lesions.** *ACS Nano* 2011, **5**:2493–2502.

47. Abedin MJ, Liepold L, Suci P, Young M, Douglas T: **Synthesis of a cross-linked branched polymer network in the interior of a protein cage.** *J Am Chem Soc* 2009, **131**:4346–4354.

48. Flenniken ML, Uchida M, Liepold LO, Kang S, Young MJ, Douglas T: **A Library of Protein Cage Architectures as Nanomaterials**. *Viruses and Nanotechnology* 2009, doi:10.1007/978-3-540-69379-6_4.

49. Czapar AE, Steinmetz NF: **Plant viruses and bacteriophages for drug delivery in medicine and biotechnology**. *Current Opinion in Chemical Biology* 2017, **38**:108–116.

50. Zhang Y, Ardejani MS, Orner BP: **Design and Applications of Protein-Cage-Based Nanomaterials**. *Chemistry – An Asian Journal* 2016, **11**:2814–2828.

51. Yu K, Yau YH, Sinha A, Tan T, Kickhoefer VA, Rome LH, Lee H, Shochat SG, Lim S: **Modulation of the Vault Protein-Protein Interaction for Tuning of Molecular Release**. *Sci Rep* 2017, **7**:14816.

52. Rome LH, Kickhoefer VA: **Development of the vault particle as a platform technology**. *ACS Nano* 2013, **7**:889–902.

53. Yuan Y, Du C, Sun C, Zhu J, Wu S, Zhang Y, Ji T, Lei J, Yang Y, Gao N, et al.: **Chaperonin-GroEL as a Smart Hydrophobic Drug Delivery and Tumor Targeting Molecular Machine for Tumor Therapy**. *Nano Lett* 2018, **18**:921–928.

54. Hoersch D, Roh S-H, Chiu W, Kortemme T: **Reprogramming an ATP-driven protein machine into a light-gated nanocage**. *Nat Nanotechnol* 2013, **8**:928–932.

55. López-Sagaseta J, Malito E, Rappuoli R, Bottomley MJ: **Self-assembling protein nanoparticles in the design of vaccines**. *Computational and Structural Biotechnology Journal* 2016, **14**:58–68.

56. He D, Marles-Wright J: **Ferritin family proteins and their use in bionanotechnology**. *New Biotechnology* 2015, **32**:651–657.

57. Falvo E, Tremante E, Arcovito A, Papi M, Elad N, Boffi A, Morea V, Conti G, Toffoli G, Fracasso G, et al.: **Improved Doxorubicin Encapsulation and Pharmacokinetics of Ferritin-Fusion Protein Nanocarriers Bearing Proline, Serine, and Alanine Elements**. *Biomacromolecules* 2016, **17**:514–522.

58. Zhen Z, Tang W, Chen H, Lin X, Todd T, Wang G, Cowger T, Chen X, Xie J: **RGD-Modified Apoferritin Nanoparticles for Efficient Drug Delivery to Tumors**. *ACS Nano* 2013, **7**:4830–4837.

59. Zhen Z, Tang W, Guo C, Chen H, Lin X, Liu G, Fei B, Chen X, Xu B, Xie J: **Ferritin Nanocages To Encapsulate and Deliver Photosensitizers for Efficient Photodynamic Therapy against Cancer**. *ACS Nano* 2013, **7**.

60. Uchida M, Flenniken ML, Allen M, Willits DA, Crowley BE, Brumfield S, Willis AF, Jackiw L, Jutila M, Young MJ, et al.: **Targeting of Cancer Cells with Ferrimagnetic Ferritin Cage Nanoparticles**. *J Am Chem Soc* 2006, **128**:16626–16633.

61. Chen Z, Zhai M, Xie X, Zhang Y, Ma S, Li Z, Yu F, Zhao B, Zhang M, Yang Y, et al.: **Apoferritin Nanocage for Brain Targeted Doxorubicin Delivery**. *Molecular Pharmaceutics* 2017, **14**:3087–3097.

62. Fan K, Jia X, Zhou M, Wang K, Conde J, He J, Tian J, Yan X: **Ferritin Nanocarrier Traverses the Blood Brain Barrier and Kills Glioma**. *ACS Nano* 2018, **12**:4105–4115.

63. Heinze K, Sasaki E, King NP, Baker D, Hilvert D, Wuite GJL, Roos WH: **Protein nanocontainers from nonviral origin: Testing the mechanics of artificial and natural protein cages by AFM**. *Journal of Physical Chemistry B* 2016, **120**:5945–5952.

64. Wörsdörfer B, Woycechowsky KJ, Hilvert D: **Directed Evolution of a Protein Container**. *Science* 2011, **331**:589–592.

65. Terasaka N, Azuma Y, Hilvert D: **Laboratory evolution of virus-like nucleocapsids from nonviral protein cages**. *PNAS* 2018, **115**:5432–5437.

66. Azuma Y, Edwardson TGW, Terasaka N, Hilvert D: **Modular Protein Cages for Size-Selective RNA Packaging in Vivo**. *J Am Chem Soc* 2018, **140**:566–569.

67. Karch CP, Burkhard P: **Vaccine technologies: From whole organisms to rationally designed protein assemblies**. *Biochem Pharmacol* 2016, **120**:1–14.

68. Kanekiyo M, Bu W, Joyce MG, Meng G, Whittle JRR, Baxa U, Yamamoto T, Narpala S, Todd J-P, Rao SS, et al.: **Rational Design of an Epstein-Barr Virus Vaccine Targeting the Receptor-Binding Site**. *Cell* 2015, **162**:1090–1100.

69. Kanekiyo M, Wei C-J, Yassine HM, McTamney PM, Boyington JC, Whittle JRR, Rao SS, Kong W-P, Wang L, Nabel GJ: **Self-assembling influenza nanoparticle vaccines elicit broadly neutralizing H1N1 antibodies**. *Nature* 2013, **499**:102–106.

70. Negahdaripour M, Golkar N, Hajighahramani N, Kianpour S, Nezafat N, Ghasemi Y: **Harnessing self-assembled peptide nanoparticles in epitope vaccine design**. *Biotechnol Adv* 2017, **35**:575–596.

71. Sliepen K, Ozorowski G, Burger JA, van Montfort T, Stunnenberg M, LaBranche C, Montefiori DC, Moore JP, Ward AB, Sanders RW: **Presenting native-like HIV-1 envelope trimers on ferritin nanoparticles improves their immunogenicity**. *Retrovirology* 2015, **12**:82.

72. Zhao Q, Li S, Yu H, Xia N, Modis Y: **Virus-like particle-based human vaccines: quality assessment based on structural and functional properties**. *Trends Biotechnol* 2013, **31**:654–663.

73. Dostalova S, Polanska H, Svobodova M, Balvan J, Krystofova O, Haddad Y, Krizkova S, Masarik M, Eckschlager T, Stiborova M, et al.: **Prostate-Specific Membrane Antigen-Targeted Site-Directed Antibody-Conjugated Apoferritin Nanovehicle Favorably Influences in Vivo Side Effects of Doxorubicin**. *Scientific Reports* 2018, **8**:1–13.

74. Votteler J, Ogohara C, Yi S, Hsia Y, Nattermann U, Belnap DM, King NP, Sundquist WI: **Designed proteins induce the formation of nanocage-containing extracellular vesicles**. *Nature* 2016, **540**:292–295.

75. Moon TS, Dueber JE, Shiue E, Prather KL: **Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered E. coli.** *Metab Eng* 2010, **12**:298–305.

76. Phippen SW, Stevens CA, Vance TDR, King NP, Baker D, Davies PL: **Multivalent Display of Antifreeze Proteins by Fusion to Self-Assembling Protein Cages Enhances Ice-Binding Activities**. *Biochemistry* 2016, **55**:6811–6820.

77. Liu Y, Gonen S, Gonen T, Yeates TO: **Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system**. *PNAS* 2018, **115**:3362–3367.

78. Cornell TA, Fu J, Newland SH, Orner BP: **Detection of specific protein-protein interactions in nanocages by engineering bipartite FlAsH binding sites.** *J Am Chem Soc* 2013, **135**:16618–16624.

Chapter 2. Structural studies of two designed protein cages illustrate progress and ongoing challenges in protein engineering

Kevin A. Cannon[1,2,*], Rachel U. Park[3,*], Scott E. Boyken[3,4], Sue Yi[3,4], Una Nattermann[3,4], David Baker[3,4,5], Neil P. King[3,4], and Todd O. Yeates[1,2,6]

[1]UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA

[2]UCLA Department of Chemistry and Biochemistry, Los Angeles, CA

[3]University of Washington Institute for Protein Design, Seattle, WA

[4]University of Washington Department of Biochemistry, Seattle, WA

[5]Howard Hughes Medical Institute, Seattle, WA

[6]UCLA Molecular Biology Institute, Los Angeles, CA

*These authors contributed equally.

Abstract

The field of engineering novel self-assembling protein nanomaterials has seen rapid advances in recent years. More than a dozen symmetric protein cages designed by various methods have now been validated by X-ray crystallography to form closed assemblies that match their design models with near-atomic accuracy. However, many protein cage designs that are tested in the lab do not form the desired assembly, often showing low soluble expression instead or failing to express, and improving the success rate of design has been a point of recent emphasis. Here we present two protein structures solved by X-ray crystallography of designed protein oligomers that form two-component cages with tetrahedral symmetry. To improve on the past tendency towards

poor soluble expression in designed protein cages, we used a computational protocol that specifically emphasized the formation of hydrogen bonding networks (rather than almost exclusively hydrophobic interactions) to stabilize the designed protein-protein interfaces. Preliminary characterization showed highly soluble expression, and solution studies indicated successful cage formation by both designed proteins. For one of the designs, a crystal structure confirmed at high resolution that the intended tetrahedral cage was formed, though several flipped amino acid side chain rotamers resulted in an interface that deviates from the detailed hydrogen bonding pattern that was intended. A structure of the other designed cage showed that under the conditions where crystals were obtained, a non-cage structure was formed wherein a porous 3D protein network in space group $I2_13$ is generated by an off-target 2-fold homomeric interface instead of the designed heteromeric interface. These results illustrate some of the ongoing challenges of developing computational methods for polar interface design, and add two potentially valuable new entries to the growing list of engineered protein materials that can be used in downstream applications.

1. Introduction

A long-standing goal in the field of protein engineering has been to develop reliable methods to create designed protein nanocages with high symmetry, which could find applications in wide-ranging fields such as drug delivery, imaging, energy, and nanotechnology (Cannon *et al.*, 2019; Howorka, 2011; Aumiller *et al.*, 2018; Zhang *et al.*, 2016). Since 2012, more than a dozen examples of successful cage designs have been validated in atomic detail by X-ray crystallography (Lai *et al.*, 2012; King *et al.*, 2012, 2014; Bale *et al.*, 2016), and several others have been verified at lower resolution by electron microscopy (Hsia *et al.*, 2016; Bale *et*

al., 2016). Recent studies are beginning to generate novel cages with specific applications in mind, including the ability to encapsulate other molecules such as nucleic acids (Butterfield *et al.*, 2017; Azuma *et al.*, 2018; Edwardson *et al.*, 2018; Terasaka *et al.*, 2018), present viral antigens (Marcandalli *et al.*, 2019) or rigidly bind other proteins for cryo-EM imaging (Liu *et al.*, 2018, 2019).

Notwithstanding the impressive successes that have been reported for designed protein cages, considerable experimental challenges remain. Literature data suggest that success rates for designed cages are in the range of around 10 percent (King *et al.*, 2014; Bale *et al.*, 2016; Yeates, 2017). Different strategies for designing cages present distinct challenges. Methods based on genetic fusion between different oligomeric components face difficulties in flexibility (Padilla *et al.*, 2001; Lai *et al.*, 2012, 2013, 2014; Sciore *et al.*, 2016; Badieyan *et al.*, 2017; Cristie-David *et al.*, 2019). Methods based on the computational design of novel interfaces between oligomeric units have led to numerous successes, but they have revealed a tendency for proteins with computationally designed interfaces to fail during protein expression trials or to aggregate (e.g. by misfolding or assembling in indefinite fashion) upon expression. This has motivated efforts to improve solubility. In one study, increasing the net charge on the solvent-exposed surfaces of the designed protein subunits had a positive effect (Bale *et al.*, 2015).

Further efforts to improve the success rates for designed assemblies are needed, including to improve their solubility. One approach that would provide a general solution would be to design more polar surfaces at the protein-protein interfaces that drive assembly of the material. Hydrogen bonding interactions across the interface could offset the cost of desolvating the polar groups upon interface formation. This approach could reduce solubility problems, but at the expense of more demanding design, as hydrogen bonding is geometrically more exacting than

hydrophobic burial. Here we report crystal structures of two protein cages that were designed with a greater emphasis on interfacial hydrogen bonding, with some surprising findings.

2. Materials and Methods

2.1 Protein scaffold design

Computational design of two-component tetrahedral cages was performed using Rosetta as described earlier (King *et al.*, 2014) with some modifications. Briefly, two different trimeric components of known structure were chosen as building blocks for a given candidate cage. For each pair of components, four copies of each trimer were placed at alternating vertices of a cube of sufficient size to assure that the components did not collide. Then the single rotational degree of freedom and the single translational degree of freedom for each trimer type (i.e., four degrees of freedom in total) were sampled to identify docked configurations featuring high numbers of Cβ-Cβ contacts in proximity between the two components, without interpenetration. These docked configurations were then subjected to computational amino acid sequence design at the novel trimer-trimer interface. The sequence design procedure employed an early prototype of the recently developed HBNet protocol that favors the formation of extended hydrogen bonding networks (Boyken *et al.*, 2016). Two candidate designs with low calculated Rosetta energies were chosen to test experimentally. These two test cases were based on the same component trimers, but with slightly different rigid body degrees of freedom and different amino acid sequences at the designed interfaces.

2.2 Protein expression and purification

Genes encoding the designed protein components of each cage design were cloned into the pETDuet™-1 expression vector (Novagen) as reported previously (King *et al.*, 2014). The two components for each cage were co-expressed in *E. coli* from the same plasmid. The B component in each design contains a C-terminal His_6-tag, which was used for Ni-affinity purification. Pooled and concentrated nickel elution fractions were further purified by size exclusion chromatography (SEC) using a Superdex 200 10/300 gel filtration column (GE Life Sciences) with 25 mM TRIS pH 8.0, 150 mM NaCl, 1 mM DTT as running buffer. SEC fractions containing pure protein in the desired assembly state were pooled, concentrated, and stored at 4 °C for subsequent use in analytical SEC, electron microscopy, and X-ray crystallography.

2.3 Analytical size exclusion chromatography

Analytical SEC was performed on a Superdex 200 10/300 gel filtration column using the running buffer described above. The designed materials were each loaded onto the column at a concentration of 50 μM (of each subunit). The apparent molecular weights of the designed assemblies were estimated by comparison to previously determined nanocage standards (King *et al.*, 2014).

2.4 Negative stain electron microscopy

6 μl of purified T33-51 and T33-53 samples at concentrations ranging from 0.006 mg/mL to 0.02 mg/mL were applied to glow discharged, carbon coated 400-mesh copper grids (Ted Pella, Inc.), washed with Milli-Q water and stained with 0.075% uranyl formate based on

methods described previously (Nannenga *et al.*, 2013). Grids were visualized for assembly validation and optimized for data collection. Screening and sample optimization were performed on a 100kV Morgagni M268 transmission electron microscope (FEI, Hillsboro, OR) equipped with a Gatan Orius CCD camera (Gatan, Pleasanton, CA). Data collection was performed on a 120 kV Tecnai G2 Spirit transmission electron microscope (FEI, Hillsboro, OR). All images were recorded using a Gatan Ultrascan 4000 4k x 4k CCD camera (Gatan, Pleasanton, CA) at 52,000x magnification at the specimen level. The contrast of all micrographs was enhanced in ImageJ (Schneider *et al.*, 2012) for clarity.

Coordinates for 20,388 (T33-51) and 37,425 (T33-53) unique particles were obtained for averaging using EMAN2 (Tang *et al.*, 2007). Boxed particles were used to obtain 2D class averages by refinement in EMAN2. Back-projection images at 20 Å were computed in EMAN2 based on coordinates of the design models.

2.5 Crystallization, data collection and processing

Crystals of the T33-51 protein assembly were obtained using the hanging drop vapor diffusion crystallization method with a reservoir of 0.1 mM HEPES, pH 8.8 and 15% PEG 3350. Initial crystals were obtained at the UCLA Crystallization Facility using a Mosquito liquid handling device (TTP LabTech). Optimized crystals were obtained using a microbatch crystallization method in which a 600 nL drop containing 12 mg/mL protein and the previous reservoir solution in a 2:1 ratio was overlaid with 20 μL of a 6:5 ratio of paraffin oil to silicone oil on top of the drop to slow down evaporation. Cubic crystals of about 150 μm in length were observed after seven days at room temperature. T33-51 crystals were cryo-protected with 33% glycerol and flash frozen in liquid nitrogen prior to data collection.

T33-53 crystals were obtained using the hanging drop vapor diffusion crystallization method with a protein concentration of 12 mg/mL and a reservoir of 0.1 M Na citrate, pH 5.6, 2.5 M 1,6-hexanediol, and 0.01 M manganese chloride. Cubic crystals ranging from 100-200 μm in length were obtained from initial crystallization screens. Data were collected from these crystals without further optimization.

All datasets were collected at the Advanced Photon Source (APS), beamline 24-ID-C (NE-CAT). Data were indexed, integrated, and scaled using XDS/XSCALE (Kabsch, 2010). Structures were determined by molecular replacement using the known trimeric structures as the search models. A non-standard molecular replacement analysis was required in one case as described in Results.

2.6 Structure solution and refinement

The T33-51 model was refined to 3.40 Å with iterative rounds of model building and refinement carried out using Coot (Emsley & Cowtan, 2004) and PHENIX (Adams *et al.*, 2002). After each cycle of model rebuilding, reciprocal space refinement, including refinement of coordinates and atomic displacement parameters, was carried out using *phenix.refine*. Subsequent cycles of refinement were performed using Buster (Bricogne *et al.*, 2011) and included TLS refinement (Painter & Merritt, 2006), with the final cycle being done using *phenix.refine*, leading to a model with R_{work}/R_{free} of 0.156/0.197 at a resolution of 3.40 Å. These crystals were partially hemihedrally twinned, as noted subsequently, and the statistical averaging effects of twinning presumably contribute in part to the favorable refinement R-value obtained. Note that residues 98-103 of chain B could not be reliably modeled and were therefore omitted

from the final model. Coordinates and structure factors for the T33-51 protein cage have been deposited with the PDB code 5CY5.

The T33-53 model was refined to 4.10 Å in a similar fashion to T33-51, except that Buster was used for all refinement cycles. Symmetry considerations (discussed in Results) made it clear that the crystal contained only one of two co-expressed protein components (and did not comprise the two-component cage structure), so two initial models were generated by molecular replacement based on each of the distinct components and refined separately. Note that the two proteins A and B are homologous and structurally similar to each other, so either could have represented the crystallized component. Refinement against the B component of the design model (PDB: 1NOG) resulted in a significantly lower R-factor (23% vs. 35%), so further refinement was pursued based on this. The R_{work}/R_{free} of the final model was 0.210/0.264. Coordinates and structure factors for the T33-53 non-cage assembly have been deposited with the PDB code 5VL4.

3. RESULTS AND DISCUSSION

3.1 Protein components and design principles

Using a preliminary version of HBNet, a Rosetta protocol that designs extensive hydrogen-bonding networks at protein-protein interfaces (Boyken *et al.*, 2016), two novel two-component assemblies with tetrahedral symmetry were designed, T33-51 and T33-53 (Figure 1A; see Methods). In each case, the cage is intended to assemble from two different trimeric protein building blocks. Four trimers of each type occupy alternating vertex positions of a cube. A computationally designed interface holds the two different trimeric components together at the

edges, resulting in a cage with tetrahedral symmetry and $A_{12}B_{12}$ stoichiometry. The diameters of the designed cages are 13 nm, with molecular masses of about 480 kDa.

The two designs studied here were based on the same pair of trimeric building blocks (PDB 1NOG and PDB 1WY1). These two naturally trimeric protein components are homologs of each other, sharing 38% sequence identity and similar folds. The design of a tetrahedral cage (called T33-08) constructed from these components had been attempted before, prior to development of the HBNet protocol, but the resulting construct proved insoluble (King *et al.*, 2014). The new protocol led to designed constructs with different sequences compared to the earlier design effort. Being based on the same underlying components, the two constructs examined in the present study were similar to each other, differing only in the amino acid sequences at the designed interfaces intended to drive cage assembly.

3.2 Expression and Preliminary Structural Characterization

For production of the designed cages, both subunit types were expressed together in the same *E. coli* cells so that assembly of the intact $A_{12}B_{12}$ cages could occur *in vivo*. Both pairs of proteins exhibited high levels of soluble expression and were easily purified by Ni-affinity chromatography. For both designs, the two co-expressed proteins could be concentrated after Ni-affinity purification up to about 60 mg/mL in aqueous buffer without noticeable aggregation. SEC analysis showed a strong peak for an assembly with an apparent molecular weight of about 500 kDa (elution volume of ~12 mL on a Superdex 200 10/300 column) for both constructs, consistent with the expected tetrahedral assembly state (Figure 1B and C). After purification, the two designed assemblies were examined by negative stain electron microscopy to see whether the formation of geometrically regular structures could be confirmed visually. Protein assemblies

resembling the ~13 nm designed structures were readily observed for both constructs, and 2D averages of each particle matched low-resolution projections calculated from the computational design model of T33-51 (Figure 1B and C). Together, these data indicate that the computationally designed interfaces of T33-51 and T33-53 successfully drive assembly to the intended tetrahedral complexes.

3.3 Crystal structure of the T33-51 Tetrahedral Cage

In an effort to assess the designed materials in atomic detail, both cage constructs were crystallized by hanging drop vapor diffusion and examined by X-ray diffraction.

The diffraction data for T33-51 were processed in space group P23 with unit cell dimensions $a=b=c$=106.7Å, with data extending to 3.40 Å resolution. Data collection and processing statistics are shown in Table 1. Because the two oligomeric protein components of the cage are designed based on proteins that are homologous to each other, there exists a pseudo 4-fold axis along the tetrahedral 2-fold axis, which caused the data to also process reasonably well in space group P432. If this were the true space group, then only one of the components could be present in the crystal in order for there to be 4-fold symmetry. Fortunately, this space group was ruled out because the value for R_{merge} was 12.8% in P432, compared to 9.2% in P23, indicating that the four-fold axis was imperfect or pseudosymmetric.

A hypothetical atomic model for the tetrahedral cage was available at the outset, since the design protocol outputs a full-atom description of the intended assembly in PDB format. The unit cell of the crystal we obtained was only large enough to contain a single copy of the designed cage, giving a solvent content of 51%. Therefore, according to the P23 crystal symmetry, the cage had to sit at the origin of the unit cell (i.e., at the point of tetrahedral symmetry), with the

cage oriented so that its symmetry axes were aligned with the symmetry axes of the crystal space group. The molecular replacement problem was therefore greatly simplified. There were, however, two distinct orientations of the cage that had to be considered, related to each other by 90° rotation about a principal axis; this is because the rotational point group symmetry of the cubic lattice is octahedral or 432 while the crystal point group is only tetrahedral or 23. Of the two possible orientations for the cage in the unit cell, one gave a lower starting R-value between calculated and observed structure factor amplitudes (33% compared to 38%) and was used as the starting point for restrained refinement. The asymmetric unit, which comprised the atomic components to be refined, consisted of one copy of each subunit type. In the later stages it was discovered that the crystal specimen was partially hemihedrally twinned, which was consistent with an earlier observation that the diffraction data could be reduced in P432 with a value for R_{merge} that was only moderately worse than for P23 (12.8% vs 9.2%). The presence of twinning was also supported by a statistical analysis of the intensity data according to the L-test, which gave a value of 0.41 for data between 8 Å and 4 Å, whereas the theoretically expected values for untwinned and perfectly twinned data are 0.5 and 0.375 respectively (Padilla & Yeates, 2003). The value for the twin fraction refined to 33% under an appropriate twin law (l,-k,h).

The final model for this first cage construct, T33-51, could be refined to R_{work}/R_{free} of 0.156/0.197 at 3.40 Å resolution. The arrangement of subunits in the cage closely matched the design: the rmsd between the refined model and the design was only 0.69 Å on C-alpha atoms for the asymmetric unit (Figure 2A and B). Agreement at the amino acid side chain level at the designed interface was mixed (Figure 2C and D). Overall, many of the designed atomic contacts at the interface were observed in the intended configuration: for 27 out of 36 amino acids within 6 Å of the other subunit, the deviation at chi1 between the design model and refined structure

was ≤ 25°. However, the agreement was lower for the residues intended to form the hydrogen bonding network. For example, residue S90 of component A and E81 of component B form an unexpected hydrogen bond network with residue K94 of component A, whereas by design residue K151 of component A was intended to form a hydrogen bond network with S90 and E81. In the crystal structure, K151 is flipped such that it no longer comes close to the other residues with which it was designed to form hydrogen bonds (Figure 2D). This suggested that the designed hydrogen bonds were not making critical contributions to the stability of the tetrahedral material, though the polarity of the designed interface residues may have helped mitigate aggregation problems during expression.

3.4 Crystal structure of the T33-53 Accidental Minimum Contact Lattice

The X-ray diffraction data for the second design, T33-53, initially appeared to process well in space group I23. Similar to the situation for the structure of the first cage (in P23), space group I23 also supports a position at the origin with 23 (T) point symmetry where a cage with symmetry T could sit. However, the body-centered unit cell dimensions of $a=b=c$=138.4Å in this case appeared to be slightly too small to accommodate the designed cage situated at the origin. We considered that the trimeric components had perhaps rotated relative to the design, which then would have allowed a slightly deformed cage to pack in the observed unit cell. However, molecular replacement with the trimeric components as search models failed to identify any alternate solutions with rotated trimers. We considered that our observed diffraction was in fact consistent with either I23 or I213; this is a rare case where systematic absences do not distinguish between space groups. Our reasons for favoring I23 at the outset were noted above; I213 does not possess any point of 23 (T) symmetry and therefore could not support a tetrahedral cage (given that the

unit cell was not large enough to fit multiple copies of a subunit with non-crystallographic symmetry). When we repeated the molecular replacement calculations in space group I213, a non-cage solution was immediately apparent.

The correct structure of the T33-53 crystal is shown in Figure 3, and the data collection and processing statistics are listed in Table 2. Structure determination in the correct space group resulted in a porous, interconnected 3D crystal lattice. Remarkably, the structure is made up of only one component from the original design; the asymmetric unit consists of just one copy of subunit B (see Methods). The natural trimeric unit for component B is correctly formed, and sits on a crystallographic 3-fold axis, but there is an accidental crystallographic 2-fold interaction that arises where two trimers make contact using regions of their surfaces that were intended to form the designed heterotypic interface with subunit A, which is absent from the crystal (Figure 3A). Nearly all of the 11 residues mutated from the native protein sequence of 1NOG lie at or near the unintended dimeric interface, where they create a large and relatively hydrophobic patch on the protein surface which sticks to the same region on an adjacent protein subunit and gives rise to the unintended 2-fold axis (Figure 3B). Additionally, a loop region of the native protein sequence (residues 76-81), which should have been solvent-exposed by design, appears to form new electrostatic interactions that help hold the 2-fold interface together as well.

In principle, trimers of subunit B might conceivably have come together to make a cage structure, but the orientation of the trimers with respect to one another in the 3D lattice places their three-fold symmetry axes in incompatible positions. Instead, the arrangement is a rather remarkable network structure (Figure 3C). The natural 3-fold symmetry axis of the original protein trimer and the fortuitous 2-fold axis between trimers are non-intersecting, with an angle of 54° between them. Those two interactions – the (natural) 3-fold interface and the fortuitous 2-fold –

are the only molecular contacts present in the crystal. The solvent content for the crystal is 76.6%. Interestingly, this situation of a connected 3D lattice in space group I2(1)3 formed by non-intersecting 3-fold and 2-fold axes was discussed in 2001 by Padilla et al. in the context of the minimum contact requirements for designing three-dimensional crystalline materials from simpler oligomeric units (Padilla *et al.*, 2001).

The observed crystal structure stands in contrast to the solution (SEC) data and electron microscopy results for the T33-53 construct, which supported a cage structure. A likely explanation is that the computationally designed interface was not stable to the crystallization conditions employed, and that an alternate homotypic interface involving drastically rotated trimers (of component B) was favored and selected by crystal growth. It is perhaps notable that the pH of the crystallization buffer was 5.6, and unexpected protonation (e.g., of histidines) could have impacted the energetics. Although the unusual network structure seen in the crystal structure was not the intended outcome for this redesigned protein oligomer, engineered porous 3D protein crystals such as this could find diverse applications, such as in creating catalytic reaction vessels for preparing inorganic materials (Ueno, 2013) or in immobilizing enzymes to create biosensors or microbioreactors (Kowalski *et al.*, 2019).

4. Conclusion

This study was motivated by a common problem of low solubility that often afflicts designed proteins bearing novel protein-protein interfaces; such interfaces tend to be somewhat hydrophobic. More hydrophilic interfaces based on geometrically specific hydrogen bonding networks offer a potential advantage, but those interactions require highly accurate design. A designed tetrahedral cage (called T33-08) from an earlier study (King *et al.*, 2014), which

expressed insolubly, was taken a starting point for improving the design using a new Rosetta protocol (a preliminary version of HBNet) that emphasizes favorable hydrogen bonding. Two computationally designed sequences were produced experimentally. Both formed the intended cages in solution, which led to crystallographic work to investigate their atomic details.

The crystal structure data on the two designed variants revealed two new interesting results for designed protein assemblies, one matching closely to its intended computational design (a tetrahedrally symmetric cage) and one forming an unexpected kind of extended 3D lattice. The first case, which crystallized in the form of a cage, was surprising at the atomic level. Owing to side chain differences compared to the design, some of the intended hydrogen bonding interactions do not occur in the interface, yet the proteins were indeed more soluble than an earlier design based on the same building blocks. One interpretation of this finding is that the designed sequences from the HBNet protocol were indeed more hydrophilic in the interfaces, and that this mitigated the previous solubility problem, which enabled the intended cage structures to form since most of the specifically intended atomic interactions were made. The second designed construct presented other surprises. In this case, the cage structure evident in solution was apparently not stable to crystallization conditions (i.e., pH 5.6). Nonetheless, the features introduced into the trimeric protein surface introduced a general tendency towards association. As a result, under specific crystallization conditions, one of the two trimeric components that was intended to assemble into the cage instead *self*-associated to make a porous and highly unusual 3D crystalline network held together by the natural 3-fold trimeric contacts and a single fortuitous 2-fold contact between trimeric units.

In broad terms, the computational design efforts here were successful. Using a new protocol, two new protein cages were designed, and solution experiments indicated assembly to

the intended configurations; a previous design attempt using the same proteins as building blocks had yielded only insoluble protein. Yet finer considerations show that current design methods still leave room for improvements, including in designing hydrogen bonding interactions. We can identify two factors that likely contributed to the failure to design a hydrogen bonding network in T33-51 that was recapitulated in the crystal structure (it is unknown whether the hydrogen bonding network in T33-53 is present as intended in the tetrahedral complex). First, we used a preliminary version of the HBNet protocol that was incomplete. Subsequent refinement of the protocol has led to the design of numerous structures with extended hydrogen bonding networks that have been faithfully recapitulated in high-resolution crystal structures, including pH-responsive hydrogen bonding networks rooted by histidine residues (Boyken *et al.*, 2016; Chen *et al.*, 2019; Boyken *et al.*, 2019). Second, we applied the preliminary design protocol to an extremely limited design space. We searched only the docked configurations that we experimentally characterized in our previous report on two-component tetrahedral complexes (King *et al.*, 2014), and we did not allow backbone movement at any point during the HBNet or RosettaDesign portions of our design protocol. It is therefore somewhat unsurprising that we were unable to identify hydrogen bonding networks of exceptional quality or stability. Future efforts to design more polar protein-protein interfaces that drive assembly of protein nanomaterials should search as broad a design space as possible, and should incorporate the lessons learned from successful instances of HBNet-based design mentioned above. Among others, these include requiring the absence of buried unsatisfied hydrogen bonding groups and constraining the possible conformations of the hydrogen bonding residues through burial. Achieving stability across experimental conditions – or controlled responsiveness to changes in conditions – is another challenge for future efforts. In practical terms, the present work adds two

new two-component protein nanomaterials to the growing repertoire available for potential applications in medicine and nanotechnology.

Acknowledgements

This work was supported by NSF grant CHE 1629214 (T.O.Y, N.P.K., D.B.), a grant from Takeda Pharmaceuticals (N.P.K.), and the Howard Hughes Medical Institute (D.B.). Initial screening and X-ray diffraction analysis was performed at UCLA core facilities supported by grant DE-FC02-02ER63421 from the BER program of the US Department of Energy Office of Science. We thank Mike Collazo, Mike Sawaya and Duilio Cascio at the UCLA-DOE X-ray Crystallization and Crystallography Core Facilities, (supported by Department of Energy Grant DE-FC02-02ER63421) for assistance with crystallization screening, model building and refinement. This work is based upon research conducted at the Northeastern Collaborative Access Team beamlines, which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant (S10 RR029205). 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.

Protein Sequences:

Residues mutated from native sequence are underlined.

Residues added to native sequence are in **bold**.

A components are based on PDB 1WY1 and B components are based on PDB 1NOG.

T33-51A (13 mutations):

MRITTKVGDKGSTRLFGGEEVWKDDPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQ

NDIYKIMGEIGSKGKIEGISEERIKWLAGLIERYSEMVNKLSFVLPGGTLESAKLDVCRTI

ARRAERKVATVLREFGIGTLAAIYLALLSRLLFLLARVIEIEKNKLKEVRS

T33-51B (13 mutations + C-terminal His_6-tag):

MFTRRGDQGETDLANRARVGKDSPVVEVQGTIDELNSFIGYALVLSRWDDIRNDLFRIQ

NDLFVLGEDVSTGGKGRTVTMDMIIYLIKRSVEMKAEIGKIELFVVPGGSVESASLHMA

RAVSRRLERRIKAASELTEINANVLLYANMLSNILFMHALISNKRLNIPEKIWSIHRVS**LE**

HHHHHH

T33-53A (11 mutations):

MRITTKVGDKGSTRLFGGEEVWKDDPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQ

NDIYKIMGEIGSKGKIEGISSERIKWLAGLISRYEEMVNKLSFVLPGGTLESAKLDVCRTI

ARRAERKVATVLREFGIGTNAAIYLAALSDLLFLLARVIEIEKNKLKEVRS

T33-53B (11 mutations + C-terminal His_6-tag):

MFTRRGDQGETDLANRARVGKDSPVVEVQGTIDELNSFIGYALVLSRWDDIRNDLFRIQ

NDLFVLGEDVSTGGKGRTVTLEMILYLVERVTEMKAEIGKIELFVVPGGSVESASLHMA

RAVSRRLERRIKAASRLTEINDNVLLYAAMLSSILFMHALISNKRLNIPEKIWSIHRVS**LE**

HHHHHH

Table 1 – T33-51 crystallographic data

T33-51	
Data Collection	
PDB accession code	5CY5
Beamline	APS-NECAT-24-ID-C
Space group	P23
Unit cell dimensions	
a=*b*=*c* (Å)	106.7
α=β=γ (°)	90.0
Reflections observed	154895
Unique reflections	5827
Wavelength (Å)	0.9795
Resolution (Å)	75.45-3.40
Highest Resolution Shell (Å)	3.48-3.40
R_{sym} (%)	9.2 (135.9)
CC(1/2)	100.0 (70.4)
I/σ	29.0 (2.8)
Completeness (%)	99.9 (99.5)
Wilson B value ($Å^2$)	107.1
Refinement	
Resolution (Å)	75.45-3.40
Resolution (Å) (last shell)	3.74-3.40

Reflections Used	5825
R_{work}(%)/R_{free}(%)	15.6 (17.8)/19.7(23.0)
Protein Molecules in Asymmetric Unit	2 (1 of each subunit type)
Number of non-H atoms	2255
RMS deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.00
Average *B*-factor (Å^2)	81.1
Ramachandran plot regions	
Favored (%)	95.4
Allowed (%)	4.6
Outliers (%)	0.00

Table 2 – T33-53 crystallographic data

T33-53	
Data Collection	
PDB accession code	5VL4
Beamline	APS-NECAT-24-ID-C
Space group	I2(1)3

Unit cell dimensions	
a=*b*=*c* (Å)	138.4
α=β=γ (°)	90.0
Reflections observed	21295
Unique reflections	3554
Wavelength (Å)	0.9795
Resolution (Å)	97.85-4.10
Highest Resolution Shell (Å)	4.20-4.10
R_{sym} (%)	14.7 (77.6)
CC(1/2)	100.0 (74.2)
I/σ	6.57 (2.31)
Completeness (%)	98.9 (94.9)
Wilson B value ($Å^2$)	135.5
Refinement	
Resolution (Å)	97.85-4.10
Resolution (Å) (last shell)	4.58-4.10
Reflections Used	3551
R_{work}(%)/R_{free}(%)	21.0(27.9)/26.4(32.4)
Protein Molecules in Asymmetric Unit	1
Number of non-H atoms	1179

RMS deviations	
Bond lengths (Å)	0.011
Bond angles (°)	1.02
Average *B*-factor (Å^2)	267.2
Ramachandran plot regions	
Favored (%)	97.9
Allowed (%)	1.4
Outliers (%)	0.7

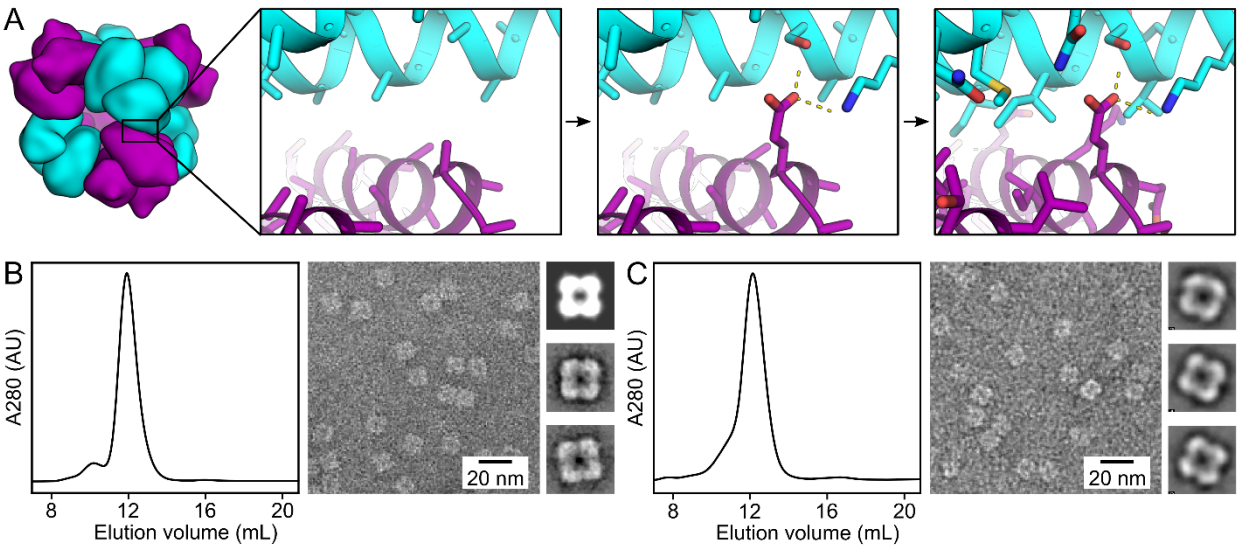


Figure 2.1. Design and characterization of T33-51 and T33-53. A) Graphical depiction of the hydrogen bond network-focused design protocol. (*Left*) Docked configurations for T33-51 and T33-53, based on the previously unsuccessful design T33-08, (*second from left*) featured contacts between the trimeric building blocks based on well-anchored elements of secondary structure.

(*Second from right*) In an initial sequence design step, hydrogen bond networks compatible with the docked backbone were identified and placed. (*Right*) Subsequently, the full protein-protein interface was designed, keeping the identities of the hydrogen bond network residues fixed. B) SEC and EM of T33-51. C) SEC and EM of T33-53. Both SEC profiles show strong single peaks at the expected elution volume for the ~500 kDa tetrahedral assemblies. Representative negative stain EM micrographs for each assembly are shown. Uniform particles of ~13 nm were observed for both T33-51 and T33-53. Insets: four-lobed, square-shaped particle averages closely resemble a projection of T33-51 calculated from the computational design model along its 2-fold symmetry axis (top inset in B).

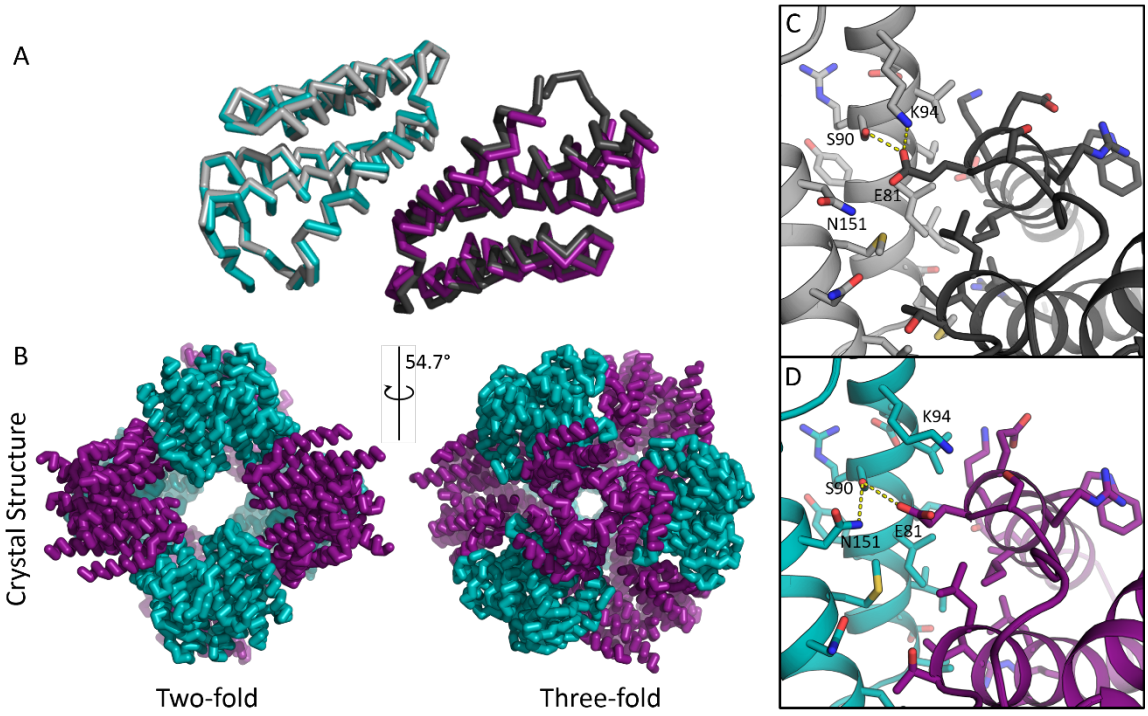


Figure 2.2. Crystal structure and comparison with design for cage T33-51. A) Alignment of

the crystal structure of T33-51 (violet and cyan) with the design model (gray). The asymmetric unit consisted of one monomer of each subunit type: T33-51A (violet and dark gray) and T33-51B (cyan and light gray). B) Views of the fully assembled 24-subunit T33-51 cage along the 2-fold/pseudo-4-fold axis (left) and along one of the 3-fold symmetry axes (right). C) and D) Both similarities and differences between the designed (C) and experimental (D) interfaces are seen at the sidechain level. Flipped rotamers in the crystal structure compared to the design led to unanticipated hydrogen bond networks.

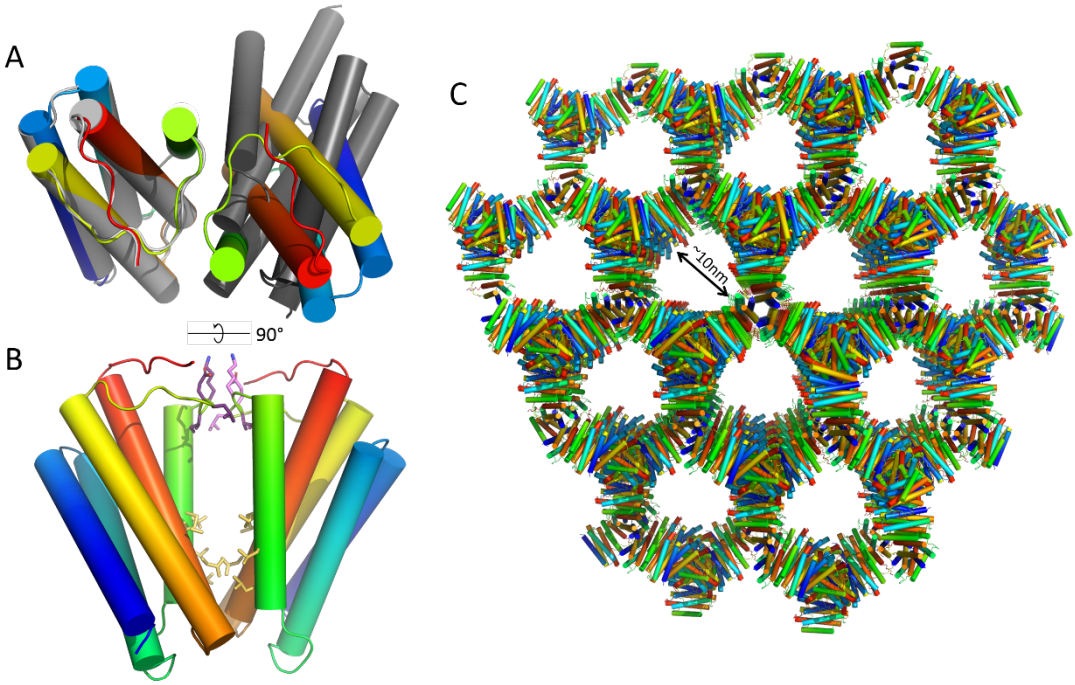


Figure 2.3. Structure of the T33-53 crystal lattice. A) Ribbon models of the design vs. crystal structure, viewed down the 2-fold axis of the unanticipated homotypic interface. Design model is shown in gray, and crystal structure is shown in rainbow coloring (blue at the N-terminus to red at the C-terminus). The orientation of the adjacent protein subunit in the observed crystal structure is significantly rotated from that in the design model. B) View perpendicular to 2-fold

axis. Hydrophobic sidechains of residues that were mutated from the native sequence of 1NOG are shown in yellow. A hydrophobic patch on T33-53B at the designed interface forms an unintended 2-fold homotypic interaction. Another interaction surface, held together by electrostatic interactions between sidechains, occurs at the C2 symmetry axis between residues 76-81 of each chain (sidechains colored in purple). This segment is a loop region, which contains no mutations from the native protein sequence and was predicted to remain solvent-exposed in the design model. C) Lattice structure of the $I2_13$ crystal of T33-53B oriented along the C3 symmetry axis.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst D*. **58**, 1948–1954.

Aumiller, W. M., Uchida, M. & Douglas, T. (2018). *Chemical Society Reviews*. **47**, 3433–3469.

Azuma, Y., Edwardson, T. G. W., Terasaka, N. & Hilvert, D. (2018). *J. Am. Chem. Soc.* **140**, 566–569.

Badieyan, S., Sciore, A., Eschweiler, J. D., Koldewey, P., Cristie-David, A. S., Ruotolo, B. T., Bardwell, J. C. A., Su, M. & Marsh, E. N. G. (2017). *ChemBioChem*. **18**, 1888–1892.

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**, 389–395.

Bale, J. B., Park, R. U., Liu, Y., Gonen, S., Gonen, T., Cascio, D., King, N. P., Yeates, T. O. & Baker, D. (2015). *Protein Science*. **24**, 1695–1701.

Boyken, S. E., Benhaim, M. A., Busch, F., Jia, M., Bick, M. J., Choi, H., Klima, J. C., Chen, Z., Walkey, C., Mileant, A., Sahasrabuddhe, A., Wei, K. Y., Hodge, E. A., Byron, S., Quijano-Rubio, A., Sankaran, B., King, N. P., Lippincott-Schwartz, J., Wysocki, V. H., Lee, K. K. & Baker, D. (2019). *Science*. **364**, 658–664.

Boyken, S. E., Chen, Z., Groves, B., Langan, R. A., Oberdorfer, G., Ford, A., Gilmore, J. M., Xu, C., DiMaio, F., Pereira, J. H., Sankaran, B., Seelig, G., Zwart, P. H. & Baker, D. (2016). *Science*. **352**, 680–687.

Bricogne G., Blanc E., Brandl M., Flensburg C., Keller P., Paciorek W., Roversi P, Sharff A., Smart O.S., Vonrhein C., Womack T.O. (2011). BUSTER Cambridge, United Kingdom: Global Phasing Ltd.

Butterfield, G. L., Lajoie, M. J., Gustafson, H. H., Sellers, D. L., Nattermann, U., Ellis, D., Bale, J. B., Ke, S., Lenz, G. H., Yehdego, A., Ravichandran, R., Pun, S. H., King, N. P. & Baker, D. (2017). *Nature*. **552**, 415–420.

Cannon, K. A., Ochoa, J. M. & Yeates, T. O. (2019). *Current Opinion in Structural Biology*. **55**, 77–84.

Chen, Z., Boyken, S. E., Jia, M., Busch, F., Flores-Solis, D., Bick, M. J., Lu, P., VanAernum, Z. L., Sahasrabuddhe, A., Langan, R. A., Bermeo, S., Brunette, T. J., Mulligan, V. K., Carter, L. P., DiMaio, F., Sgourakis, N. G., Wysocki, V. H. & Baker, D. (2019). *Nature*. **565**, 106–111.

Cristie-David, A. S., Chen, J., Nowak, D. B., Bondy, A. L., Sun, K., Park, S. I., Banaszak Holl, M. M., Su, M. & Marsh, E. N. G. (2019). *J. Am. Chem. Soc.* **141**, 9207–9216.

Edwardson, T. G. W., Mori, T. & Hilvert, D. (2018). *Journal of the American Chemical Society*. **140**, 10439–10442.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst D*. **60**, 2126–2132.

Howorka, S. (2011). *Current Opinion in Biotechnology*. **22**, 485–491.

Hsia, Y., Bale, J. B., Gonen, S., Shi, D., Sheffler, W., Fong, K. K., Nattermann, U., Xu, C., Huang, P.-S., Ravichandran, R., Yi, S., Davis, T. N., Gonen, T., King, N. P. & Baker, D. (2016). *Nature*. **535**, 136–139.

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

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., Andre, I., Gonen, T., Yeates, T. O. & Baker, D. (2012). *Science*. **336**, 1171–1174.

Kowalski, A. E., Johnson, L. B., Dierl, H. K., Park, S., Huber, T. R. & Snow, C. D. (2019). *Biomater. Sci.* **7**, 1898–1904.

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

Lai, Y. T., Tsai, K. L., Sawaya, M. R., Asturias, F. J. & Yeates, T. O. (2013). *Journal of the American Chemical Society*. **135**, 7738–7743.

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). *Nature Chemistry*. **6**, 1065–1071.

Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. (2018). *PNAS*. **115**, 3362–3367.

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

Marcandalli, J., Fiala, B., Ols, S., Perotti, M., de van der Schueren, W., Snijder, J., Hodge, E., Benhaim, M., Ravichandran, R., Carter, L., Sheffler, W., Brunner, L., Lawrenz, M., Dubois, P., Lanzavecchia, A., Sallusto, F., Lee, K. K., Veesler, D., Correnti, C. E., Stewart, L. J., Baker, D., Loré, K., Perez, L. & King, N. P. (2019). *Cell*. **176**, 1420-1431.e17.

Nannenga, B. L., Iadanza, M. G., Vollmar, B. S. & Gonen, T. (2013). *Current Protocols in Protein Science*. **72**, 17.15.1-17.15.11.

Padilla, J. E., Colovos, C. & Yeates, T. O. (2001). *Proceedings of the National Academy of Sciences of the United States of America*. **98**, 2217–2221.

Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst D*. **59**, 1124–1130.

Painter, J. & Merritt, E. A. (2006). *Acta Cryst D*. **62**, 439–450.

Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012). *Nat Methods*. **9**, 671–675.

Sciore, A., Su, M., Koldewey, P., Eschweiler, J. D., Diffley, K. A., Linhares, B. M., Ruotolo, B. T., Bardwell, J. C. A., Skiniotis, G. & Marsh, E. N. G. (2016). *PNAS*. **113**, 8681–8686.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. (2007). *Journal of Structural Biology*. **157**, 38–46.

Terasaka, N., Azuma, Y. & Hilvert, D. (2018). *PNAS*. **115**, 5432–5437.

Ueno, T. (2013). *Chemistry – A European Journal*. **19**, 9096–9102.

Yeates, T. O. (2017). *Annual Review of Biophysics*. **46**, 23–42.

Zhang, Y., Ardejani, M. S. & Orner, B. P. (2016). *Chemistry – An Asian Journal*. **11**, 2814–2828.

Chapter 3. Design and Characterization of an Icosahedral Protein Cage Formed by a Double-Fusion Protein Containing Three Distinct Symmetry Elements

Kevin A. Cannon[1,2], Vy Nguyen[1,2], Christian Morgan[3], and Todd O. Yeates[1,2,4]

[1]UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA; [2]UCLA Department of Chemistry and Biochemistry, Los Angeles, CA; [3]UCLA Department of Ecology and Evolutionary Biology, Los Angeles, CA; [4]UCLA Molecular Biology Institute, Los Angeles, CA

Abstract

Exploiting the symmetry common to many natural protein oligomers as a starting point, several recent studies have succeeded in engineering novel self-assembling protein architectures reminiscent but distinct from those evolved in the natural world. Designing symmetric protein cages with a wide range of properties has been of particular interest for potential applications in the fields of medicine, energy, imaging and more. Using a genetically encoded alpha-helical protein fusion approach, which holds two natural oligomeric protein components in a specific orientation without requiring difficult computational interface design, we designed and characterized an icosahedral protein cage that self-assembles from 60 identical subunits. In addition to fusing dimeric and pentameric protein components with an alpha-helical linker, to further encourage icosahedral cage assembly we incorporated a flexibly linked trimeric coiled-

coil to create a double fusion protein that contains all three types of rotational symmetry elements (C2, C3, and C5) present in icosahedral symmetry. Negative stain electron microscopy and solution-state methods indicated successful formation of an approximately icosahedral cage. However, diverse experimental studies indicated substantial degrees of flexibility, expansion, and asymmetric deformation of the assembled particle. The results add further insight into strategies and challenges in designing atomically precise protein materials.

Introduction

In the field of protein engineering, much recent effort has focused on developing reliable computational design methods to generate novel proteins with structures predictable at atomic level accuracy[1]. One area that has seen rapid advancement in recent years is the design of highly symmetric protein nanocages that assemble into closed three-dimensional structures, taking the shapes of the Platonic solids[2–5]. To date, more than a dozen designed protein cage structures have been characterized in near-atomic detail by X-ray crystallography[6–11]. Numerous current efforts aim to develop novel uses for these designed cages, for example as containers for encapsulated cargo[12,13] or as scaffolds for multivalent display of other protein molecules[14–17]. Other recent work has even suggested that the design space of engineered protein cage assemblies could be expanded to include more complex symmetric architectures such as Archimedean geometries[18].

Despite some remarkable design successes and exciting application prospects, the majority of protein cage designs lead to experimental failure. It is estimated that only about 1 in 10 designs result in cages that form properly when overexpressed in bacteria[19]. Designs often suffer from low protein expression levels, insolubility or unintended and heterogeneous assembly states[2,8,9,11].

Currently there are two primary methods of protein design that have led to success in generating proteins that correctly self-assemble into their desired symmetric architectures: computational protein interface design *via* Rosetta[7,9] and genetic fusion of natural protein oligomers[2,6,20,21]. The majority of currently validated designed protein cages are the result of the former method. Two-component cages of tetrahedral, octahedral, and icosahedral symmetry designed using this method have all been characterized in atomic detail by X-ray crystallography[7,9–11]. Notwithstanding these exciting cases, application of interface design methods is extremely challenging in terms of computational complexity and the experimental trials often required. Computationally designed interfaces often have relatively hydrophobic surface patches, which tend to limit solubility and/or cause protein misfolding. Successful results are typically achieved by testing a large number of designs experimentally[9,11].

In the other category of cage design, that is *via* oligomeric fusion, which was introduced first by Padilla et al.[2], crystal structures have validated designed assemblies with tetrahedral (12-subunit) and octahedral (24-subunit) symmetry[6,8]. The oligomeric fusion design method benefits from requiring less intensive computational methods. Few (or no) mutations need to be made to the native sequences of the component protein oligomers whose combined symmetry properties drive the desired cage assembly. Instead of creating a new interface between two oligomers by engineering new sidechain interactions between them, the component oligomers are covalently connected to each other *via* genetic fusion; i.e., the termini of the two components are joined by a short linking sequence. In the cases where it has been possible to elucidate the atomic details of assembled cages resulting from such genetically fused proteins, the linker was an alpha-helix that extended between the two proteins, which both also had alpha-helices at their termini[2], giving rise to a continuous or shared helix extending into both parts. For a precisely defined outcome,

this alpha-helical linkage should be rigid enough to hold the protein oligomers in the desired orientation, as dictated by the geometric properties of the Platonic solids. It has been shown in several experimental studies, however, that this kind of continuous alpha helical connection generally admits a substantial degree of flexibility[22–24]. In one case it was found that multiple alternative assembly states of lower symmetry than the design model were also observed in solution[8]. This issue of flexibility in the linker region has challenged the oligomer fusion design method, despite its occasional successes.

More recently, other design efforts have shown that a rigid linkage between two protein oligomers is not strictly necessary to create designed protein assemblies resembling Platonic solids. By linking globular protein oligomers to smaller coiled-coil assemblies *via* a simple flexible loop region, Marsh et al. have demonstrated the design of tetrahedral, octahedral, and icosahedral cage assemblies[20,25,26]. Woolfson et al. also showed the creation of large particles using flexible linkages between simple oligomeric components[21]. These assemblies have interesting properties that could be useful in various bioengineering applications, though the flexible linkages between oligomeric components have made them elusive to atomic level structure determination.

To date, an atomic structure of a designed icosahedral protein cage formed from 60 copies of a protein construct based on oligomeric fusions has yet to be elucidated. Owing to the larger size and stoichiometry of icosahedral assemblies, issues of flexibility and heterogeneity that have arisen in designing 12-subunit tetrahedra and 24-subunit octahedra are expected to be heightened in attempts to design 60-subunit icosahedra. One potential way to mitigate this challenge is to use a combination of the two highest rotational symmetries present in icosahedral point group symmetry (i.e., C3 + C5) for the two oligomeric protein components. In contrast to

the case for lower symmetry combinations such as C2 + C3, which can feasibly lead to the formation of either tetrahedral, octahedral or icosahedral assemblies if the linker is not rigid, icosahedral symmetry is the only symmetric outcome consistent with both C3 and C5. This principle was exploited by Cristie-David et al.[26] in their flexible fusion of a trimeric protein to a pentameric coiled-coil motif. Unfortunately, in our present study we were unable to design a continuous alpha helical fusion between a known pentamer and trimer owing in part to the relative paucity of those choices in combination. [N.B. Pentamers are about 70 times less common than dimers among known proteins with at least one alpha-helical terminus.] Instead, we undertook an alternative fusion-type strategy for assembling a protein icosahedron from 60 copies of a single protein molecule, namely by incorporating all three elements of icosahedral symmetry into a single protein construct by two fusions, one based on the continuous alpha helical method and one being flexible. To our knowledge, this is the first designed protein cage to have three different rotational symmetry elements represented by protein oligomers at once. Successes and challenges in characterizing the detailed structure of the designed assembly are described.

Materials & Methods

Design of the 'kico-03' protein fusion construct

Design of the initial alpha-helical fusion scaffold was carried out in a similar fashion as described previously[2,8]. Briefly, all homodimeric and homopentameric proteins with C2 or C5 symmetry were first downloaded from the Protein Data Bank. The secondary structures of these proteins were then analyzed using Stride[27] to determine which homo-oligomers contained alpha-helical regions near their N-terminal or C-terminal ends. Then, all possible pairs of dimers with

C-terminal helices and pentamers with N-terminal helices were aligned pairwise on opposite ends of a generic alpha-helix using locally written computer scripts. Linker lengths from 0 to 15 residues were tested for each pair, and the angle and distance between the cyclic symmetry axes of the two oligomeric domains were calculated for each possibility. Any alignments that resulted in symmetry axes that came within 3Å of intersecting, and having an angle of intersection within a 2° deviation from the ideal icosahedral value of 31.7°[3,28] were then manually curated to remove designs that contained obvious steric clashes, membrane proteins, or other features likely to hinder their self-assembly into the desired icosahedral shape. One potential design (named hereafter kico-03) contains a 6-residue alpha-helical linker region with amino acid sequence EEEQRR between the extant terminal helices of the dimeric and pentameric component domains, resulting in a continuous helix length of 29 residues (about 8 turns). One additional residue was also mutated from its native sequence on each side of the linker. The kico-03 construct in particular was chosen for this study because there was a large window around the 3-fold axes of the icosahedron and because the helical fusion protein's free N-terminus was pointing directly into that window, suggesting that fusion of an additional small protein domain would be feasible.

The trimeric coiled-coil was genetically fused to the N-terminus of the helical fusion design. This was determined to be possible by manually placing 20 copies of the known crystal structure of the trimeric coiled-coil region approximately at the 3-fold symmetric icosahedral faces of the design model. As the modeling of a flexible linker does not define a singular unique solution, multiple linker lengths were tested experimentally to determine the ideal number of residues to place between the trimeric and dimeric domains of the fusion protein, as described later (see Results).

The three fused components of the designed construct are based on proteins with sequences and crystal structures that have been previously deposited in the PDB. The dimer domain is a protein from the DUF1048 family in *Bacillus halodurans* (PDB code 2O4T); the pentamer is a chlorite dismutase from *Nitrospira defluvii* (PDB code 3NN1)[29]; and the trimer is a *de novo* designed coiled-coil (PDB code 4DZL)[30]. All three components had been successfully overexpressed in *Escherichia coli* for their previous structure determination.

Plasmid preparation and protein expression

Genes encoding the various protein constructs used in this study were purchased from IDT and cloned into the pET-22b expression vector (Novagen) *via* Gibson Assembly. Plasmids were subsequently transformed into BL21 (DE3) *E. coli* cells (New England Biolabs) and overexpressed in LB supplemented with ampicillin at 100 μg/ml. Cell cultures were grown with shaking at 37°C until reaching OD_{600} of 0.6. Overexpression of the designed protein constructs was then induced with 1mM IPTG at 18°C overnight. Cells were then harvested by centrifugation at 5,000 x *g* for 10 minutes.

Initial purification of designed protein constructs

Cells containing overexpressed kico-03 variants were lysed using an Emulsiflex C3 (Avestin) in a buffer of 50 mM Tris, pH 8.0, 250 mM NaCl, and 20mM imidazole. The recombinant proteins were then purified by HisTrap Ni-affinity purification using gradient elution in the range of 20-500mM imidazole. The presence of recombinant protein in the elution fractions was evaluated by SDS-PAGE. Fractions containing the protein of interest were then pooled and concentrated before being subjected to size exclusion chromatography (SEC) for

further purification of the cage assembly. SEC was carried out on an Akta FPLC system (GE Life Sciences) using a Superose 6 Increase 10/300 column (GE Life Sciences) in a running buffer of 25mM Tris, pH 8.0 and 150mM NaCl. Fractions judged to fall near the boundary of the void volume contained the kico-03 protein and were pooled for subsequent characterization.

Sucrose gradient purification

For experiments requiring a higher level of cage purity (discussed in Results), a sucrose gradient centrifugation purification step was performed after SEC. A 10mL continuous 10-50% sucrose gradient was formed using a Gradient Master (Biocomp Instruments). SEC-purified protein cage sample was then placed at the top of the gradient and subjected to ultracentrifugation at 25,000 rpm for 18 hours using an L8-70 ultracentrifuge equipped with an SW40 rotor (Beckman Coulter). The gradient was subsequently separated into 500µL fractions, which were then subjected to dialysis in order to remove the sucrose and return the cages to the previous buffer conditions. Relatively homogeneous protein cages were found in fractions containing roughly 30% sucrose, as determined by negative stain electron microscopy (EM).

Dynamic Light Scattering

Sucrose gradient-purified kico-03 cages were subjected to dynamic light scattering (DLS) on a DynaPro Plate Reader II system (Wyatt Technology) to determine their size and level of monodispersity in solution. DLS measurements were made with protein concentrations of approximately 1mg/mL in triplicate, with 5 acquisitions per replicate.

Negative stain electron microscopy

Negative stain EM was used to verify the assembly of kico-03 cage particles. 5 µL of SEC-purified cages in the concentration range of 5 µg/mL to 20 µg/mL were applied to glow discharged, formvar/carbon-coated 300-mesh copper grids (Ted Pella, Inc.) and stained with 2% uranyl acetate. Cages were imaged at 120kV on an FEI Tecnai T12 (Thermo Fisher) transmission electron microscope (TEM).

Cryo-electron microscopy

Samples of kico-03 with good particle density as assessed by negative stain EM were then investigated by cryo-EM. Protein samples in the concentration range of 0.5 to 0.8 mg/mL were flash frozen on glow-discharged Quantifoil 200 mesh 1.2/1.3 copper grids (Electron Microscopy Sciences) using a Vitrobot Mark IV (FEI). 3 µL of sample was applied to the grids immediately prior to plunge freezing in liquid ethane. Freezing conditions were screened by imaging grids at 200kV on a FEI Tecnai TF20 TEM (Thermo Fisher), and high-resolution data of selected grids were collected at 300kV on a FEI Titan Krios TEM (Thermo Fisher). Data were processed using the cryoSPARC software[31].

Small-angle X-ray scattering

Sucrose gradient-purified samples of the kico-03 cage were dialyzed into a buffer of 25mM Tris, pH 8.0 and 150mM NaCl and subsequently subjected to small-angle X-ray scattering (SAXS) experiments to analyze the assembly properties of the particle in solution. Data were collected at the Advanced Light Source (ALS) SIBYLS beamline at Lawrence Berkeley National Laboratory according to established protocols[32].

Results and Discussion

Design of a three-symmetry-component icosahedron

With the goal of creating a fusion protein that would be more stable in an icosahedral assembly form than any lower assembly states, we designed a protein made up of three distinct protein domains with different rotational symmetries. In an icosahedron, there are three types of cyclic symmetry axes present: C2 (edges), C3 (faces), and C5 (vertices). To the best of our knowledge, all previous protein cage designs to date have contained at most two symmetry components, one from each protein oligomer type. In previous icosahedral design studies, for example, combinations of C2+C3, C2+C5, and C3+C5, have all been demonstrated; in each case a gap is unfilled at the unused symmetry element.

While examining potential cage candidates that resulted from our initial design protocols, which generated a pool of potential alpha-helical fusions of pentamers linked to dimers, we noticed that some potential icosahedral scaffolds had large, open pores where the trimeric faces of the icosahedron would be. Of these, one design showed further promise because the free N-terminus of the initial fusion protein was pointing directly into the void around where the C3 axis of the icosahedron would be in our design model. As such, we determined that it would be feasible to fuse an additional N-terminal coiled-coil domain to that exposed protein terminus *via* a flexible peptide linker, thus creating a three-component protein based on two fusions, a dimer-pentamer alpha-helical fusion and a flexibly linked trimer-dimer fusion (Figure 1). This was done in the hopes that the presence of all three rotational symmetry elements would help force the designed protein to self-assemble with the correct stoichiometry required for a 60-subunit icosahedron, while significantly lowering the chances of unwanted assemblies of lower symmetry.

The final cage design was named kico-03, comprising 60 copies of a 42 kDa double-fusion protein, assembling to form a 2.5 MDa icosahedral cage roughly 30 nm in diameter. According to the design, its interior cavity would be 18 nm in diameter.

Comparison of two-component (single-fusion) and three-component (double-fusion) design constructs

Two versions of the kico-03 cage design were prepared in order to test the effect of having the additional trimeric coiled-coil motif present, as compared to a simpler single-fusion protein design comprised of only a dimer and pentamer fused *via* an alpha-helical linker. Although the single-fusion (pentamer-dimer) protein showed exceptionally high expression levels and solubility, SEC analysis of the protein construct revealed no peaks in the size range of a fully assembled icosahedron, even at high concentrations approaching 50 mg/mL of protein (Figure 1). EM imaging of the highest molecular weight SEC peak showed the presence of only smaller protein assemblies, possibly representing just one or a few joined oligomers of the designed construct.

By contrast, gel filtration of the three-component double-fusion construct, which contains the additional trimer coiled-coil motif, showed a prominent peak near the edge of the void volume of our Superose 6 Increase 10/300 column, wherein particles resembling a roughly spherical protein assembly of the correct size were readily observed by negative stain EM (Figure 1).

Although lower molecular weight peaks are still present in the SEC profile for the double-fusion construct, it appears that the formation of the icosahedral cage assembly is in fact strongly facilitated by the addition of the flexibly linked trimer coiled-coil domain.

Investigation of flexible linker length

Estimates based on manual inspection of the initial helical fusion design model suggested that a linker region ranging anywhere from one to three residues in length between the trimeric coiled-coil and the dimeric protein domain might allow the trimer enough space to assemble properly and lead to cage formation. In order to find the optimal linker length, we tested constructs containing one, two, or three glycine residues inserted between the C-terminus of the coiled-coil motif and the N-terminus of the dimeric protein domain. Somewhat surprisingly, all three of these constructs led to well-formed cages when assessed visually by negative stain EM. We therefore reasoned that the construct with the shortest linker length would likely be most optimal for further characterization, as this should allow for the least potential flexibility of the trimer motif.

Notably, the C-terminal amino acid of the trimer coiled-coil structure that was chosen for this design is also a glycine residue. As such, we wondered if having any additional linker residues between the trimer and dimer domains was necessary at all. Indeed, cages were still observed after removing the single additional glycine residue from the linker region, but these assemblies appeared less abundant and well-formed (Figure 3). Many particles appearing to be partial or collapsed shells were observed, suggesting the full icosahedral assembly state was no longer favored in this construct, likely due to conformational strain.

Imaging by electron microscopy

After optimizing the kico-03 sample preparation to the point where we were able to achieve negative stain EM grids with a high concentration of well-formed protein cages

compared to other contaminants, we attempted to further characterize the cage by cryo-EM. We identified freezing conditions that allowed for initial imaging of kico-03 cages in a layer of vitreous ice on holey carbon grids (see Methods). Unwanted large contaminants and assemblies that are heterogeneous in size were still observed on cryo-EM grids along with the desired kico-03 cages (Figure 4A). Unfortunately, we were unable to establish highly favorable freezing conditions for imaging sucrose gradient-purified cage particles by cryo-EM; frozen grids had much lower particle densities than we had achieved in the negative EM experiments.

Despite the challenges in identifying good sample freezing conditions for cryo-EM, we attempted image analysis based on data obtained on a Titan Krios electron microscope at 300kV. A dataset containing ~6000 cage particles was collected. Attempts to produce reference-free 2D class averages displaying surface features and symmetry elements of the cage were unsuccessful. Most 2D classification attempts in cryoSPARC led to templates with the appearance of a "spiky ball" without many additional features (Figure 4B). When a soft mask was applied to the particles from 0 to 280 Å in order to mask out the spikes, some surface features arose in the class averages (Figure 4C). However, rotational symmetry elements were not readily apparent in these averaged images, and there were not enough particles per class to (i.e., less than 100) to permit reliable interpretations. Three-dimensional reconstruction attempts were largely unsuccessful except for the production of relatively featureless shells of density. The inability to generate reliable and detailed structural information from this data implies that the kico-03 cage particles likely suffer from heterogeneity in conformation or possibly in assembly form, rendering them imperfectly symmetric. Heterogeneity and asymmetry would explain the challenges encountered in three-dimensional reconstruction. If much greater numbers of particles could be imaged, then in principle a detailed structure might be possible for specific conformational forms, but as noted

above this was not possible. A further possibility is that well-assembled cages, which were evident by negative stain EM, are disrupted by the freezing conditions employed.

Solution-state characterization of kico-03

Because the icosahedral cage assembly state of the designed protein elutes near the void volume in our SEC experiments, purifying the cages from the other large contaminants that also come out in the void – protein aggregates, liposomes, and other large unwanted species – proved to be a difficult task. In order to achieve a sample composed almost entirely of purified cages, further purification by sucrose gradient ultracentrifugation was performed after SEC. This method produced much cleaner samples as assessed by negative stain EM, SEC, and DLS (Figure 5). These more purified cage samples were suitable for subsequent characterization by more sensitive solution-state methods that would have been confounded by the less pure preparations. DLS experiments (performed in triplicate) showed a well-defined major particle species. The measured frictional coefficient of the particle would correspond to an outer radius of 20 nm for a smooth spherical particle. This is somewhat higher than expected based on the radius from our design (about 15 nm). Possible reasons for the somewhat larger than expected radius include: (1) the presence of larger (i.e. > 60 subunit) assembly species; (2) non-spherical deformation of the hollow assembly; and (3) non-ideal frictional properties of the complex-featured particle. Overall, the DLS data indicate a reasonably discrete particle preparation after purification by sucrose density centrifugation, notwithstanding possible perturbations in shape and radius.

Because the cage construct proved to be resistant to both crystal formation and structure determination by cryo-EM, we undertook small-angle X-ray scattering (SAXS) experiments to

gain more insight into the size and shape of the kico-03 icosahedron in solution. The low angle (Guinier) region of the SAXS profile was nearly straight, with minor curvature consistent with a degree of heterogeneity suggested by the other experimental methods discussed (Figure 6). A linear fit of the data in this region gives an estimated radius of gyration of 138 Å for the particle in solution. Again, this was somewhat higher than the calculated radius of gyration of the symmetric design model (118 Å), suggesting that the cage is likely expanded or elongated to some degree. Significant deviations are further revealed by the full SAXS profiles (Figure 6). Highly symmetric (e.g. spherical or perfectly icosahedral) objects give rise to characteristic oscillations in a SAXS (radial scattering) curve. The absence of prominent oscillations was noted in a previous study of a designed cage, whose flexibility also led to significant deviations from perfect symmetry[8]. A similar scenario is evident here.

To see if we could create a distorted version of our model that would agree reasonably well with the experimental data, custom computer scripts were written for distorting the design model to varying degrees while maintaining a degree of integrity in the cage (i.e. restraining distances between fused components). After allowing initial distortions (e.g. compression in some directions and expansion in others), agreement between observed and calculated (model) SAXS profiles was used to guide the model distortions. In this way we found that families of flattened or otherwise distorted cages could be produced that gave good agreement with the observed SAXS data (Figure 6). Compared to protein samples that showed relatively good geometric shape under negative stain EM, samples used for SAXS experiments were subject to shipping and extended dialysis (for ~ 3 days) prior to data collection. Further experiments would be required to test the possible effects of various solution conditions and sample handling protocols.

Conclusions

The main emphasis of the current study was to test the utility of an approach for designing novel icosahedral protein assemblies based on fusing three different oligomeric protein components together: a pentamer plus trimer plus dimer. Though the addition of the flexibly linked trimer coiled-coil to the initial pentamer-dimer helical fusion appeared to promote the correct stoichiometry for an icosahedral assembly to form, the flexibility issues seen in previously designed oligomer fusion cage assemblies persists in the current design. Observed heterogeneity and polydispersity of particle size across different techniques (EM, DLS, SAXS) suggest that the assembled cages are readily deformable, and presumably highly dynamic, in solution. With the ability to achieve a higher level of cage purity by sucrose gradient centrifugation, further experiments such as analytical ultracentrifugation or native mass spectrometry could potentially provide more insight into the assembly properties of the flexible structure obtained. Further optimization of purification protocols may also lead to improved sample homogeneity to improve subsequent experimental characterization. One example would be investigating alternative gradients (other than sucrose) that may be gentler (e.g,, to prevent degradation or deformation of assembled particles) or result in more robust separation of spherical cage particles from other unwanted assembly states. While dynamic properties could ultimately be important for future engineering goals, at present the flexibility issues present notable obstacles to detailed characterization. With respect to the goal of generating highly defined architectures, the findings emphasize that for fusion-based approaches to reach the levels of rigidity realized by some previous interface design studies, further strategic improvements would be required.

On a positive note, despite challenges related to flexibility and heterogeneity, the current study introduces another example of a novel self-assembling protein icosahedron based in this case on a three-component fusion design. Although details of its properties when fully assembled are partially unresolved, as a new entry to the growing list of designed protein cages it may show promise for future use in downstream applications. The interior cavity of kico-03 is about 18 nm in diameter by design, big enough to encapsulate substantial cargo such as nucleic acids or other small proteins. With the trimer coiled-coil domain sitting on the 3-fold axes of the cage, the remaining pores are small enough (~2 nm) to prevent the unwanted entry or escape of most biological macromolecules into or out of the cage interior. In addition to the possible value of further rigidification, additional efforts could go toward developing kico-03 as an encapsulation platform, from which release of interior cargo could be triggered, for example, by introducing a specific proteolytic cleavage site in the flexible linker region, so that (potentially regulated) proteolysis of the coiled-coil segment would trigger cage disassembly.

Acknowledgements

This work was supported by NSF grant CHE 1629214. Biophysical characterization experiments were based on shared instrumentation supported by the Department of Energy Office of Science grant DE-FC02-02ER63421. The authors thank Greg Hura for assistance in collecting and interpreting the SAXS data on the designed protein cage. SAXS data at the SIBYLS beamlines was supported by DOE-BER IDAT, NIGMS P30 ALS-ENABLE (P30 GM124169) and NIGMS S10OD018483. The authors thank Prof. Hong Zhou, Peng Ge, David Boyer and Yuxi Liu for guidance and advice on electron microscopy. Transmission electron microscopy was performed

at the Electron Imaging Center for NanoMachines at the California NanoSystems Institute. The authors thank Joshua Laniado for assistance with cage design programs.

Amino Acid Sequences

Underlined amino acids are linker regions between oligomeric protein domains.

Bolded amino acids have been mutated from native protein sequences.

kico-03

MHHHHHHGEIAAIKQEIAAIKKEIAAIKFEIAAIKQGYGGHVSRVEKLPKDYQIVYKEIQK
YLFKVGPVELNEGIGLLSEILGFFEEGAAAGKGVLDVTGTDVAAFCDALIGDSKTYADL
YQESIQQHV**D**EEEQRR**K**REKLLTESGVYGTFATFQMDHDWWDLPGESRVISVAEVKGL
VEQWSGKILVESYLLRGLSDHADLMFRVHARTLSDTQQFLSAFMGTRLGRHLTSGGLL
HGVSKKPTYVAGFPESMKTELQVNGESGSRPYAIVIPIKKDAEWWALDQEARTALMQE
HTQAALPYLKTVKRKLYHSTGLDDVDFITYFETERLEDFHNLVRALQQVKEFRHNRRFG
HPTLLGTMSPLDEILEKFAQ

kico-03-notrimer

MHHHHHHGAHVSRVEKLPKDYQIVYKEIQKYLFKVGPVELNEGIGLLSEILGFFEEGAA
AGKGVLDVTGTDVAAFCDALIGDSKTYADLYQESIQQHV**D**EEEQRR**K**REKLLTESGVY
GTFATFQMDHDWWDLPGESRVISVAEVKGLVEQWSGKILVESYLLRGLSDHADLMFRV
HARTLSDTQQFLSAFMGTRLGRHLTSGGLLHGVSKKPTYVAGFPESMKTELQVNGESG
SRPYAIVIPIKKDAEWWALDQEARTALMQEHTQAALPYLKTVKRKLYHSTGLDDVDFIT
YFETERLEDFHNLVRALQQVKEFRHNRRFGHPTLLGTMSPLDEILEKFAQ

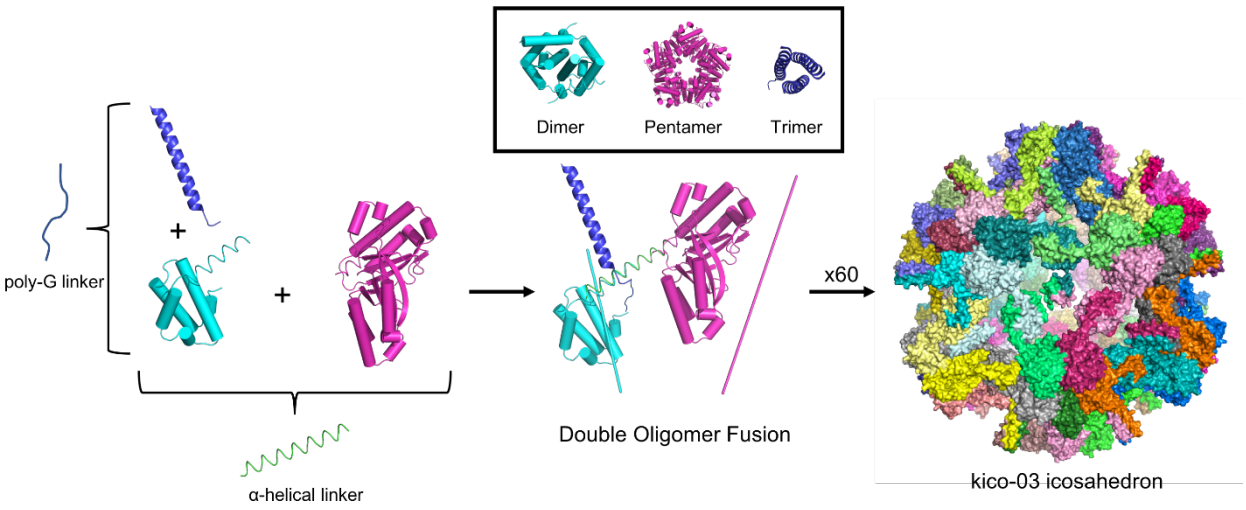


Figure 3.1. Design of a three-symmetry component, double oligomer fusion protein icosahedron. The kico-03 design combines an alpha-helical fusion of dimeric and pentameric protein domains with a flexibly linked trimer-dimer fusion (left) to create a single 42kDa fusion protein containing three distinct protein domains of different symmetries (middle). Upon proper self-assembly of this fusion protein, a 60-subunit icosahedral protein cage resembling the idealized design model (right) would form.

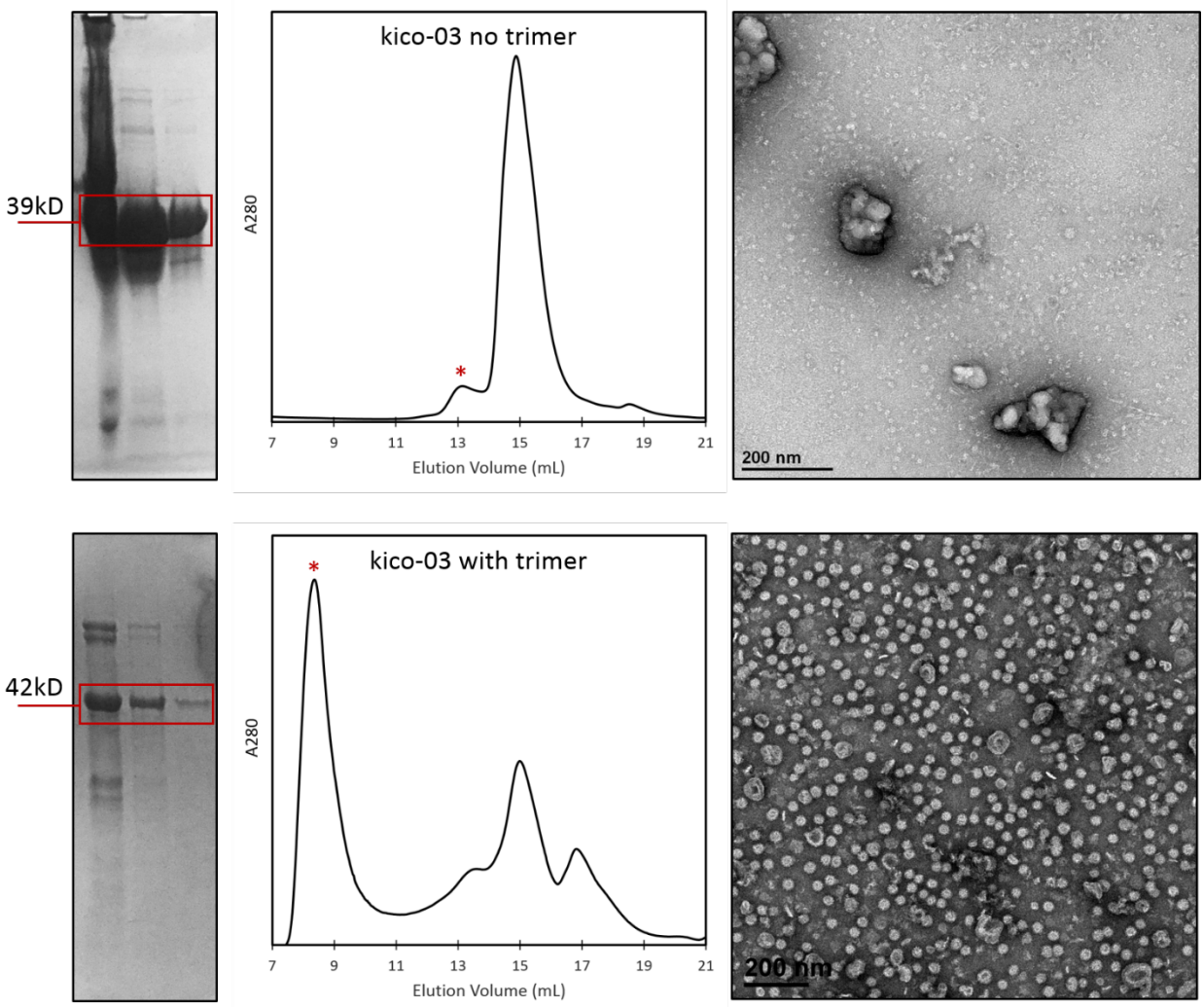


Figure 3.2. Purification of single- and double-fusion design variants. The kico-03 alpha-helical fusion construct without the flexibly linked trimer coiled-coil exhibits high soluble expression (top left), but no peak in the expected range of a 2.3MDa complex appeared on SEC (top middle). Other than some large aggregates, only small protein assemblies less than 10nm in diameter are observed on negative stain EM (top right). When the trimer coiled-coil is added to the construct, the three-symmetry double fusion protein exhibits more modest levels of soluble expression (bottom left), but a prominent peak appears near the edge of the void volume on SEC (bottom middle). 30-nm cage assemblies are readily observed from this peak on negative stain

EM grids (bottom right). Red asterisks indicate peaks that were visualized in corresponding EM micrographs.

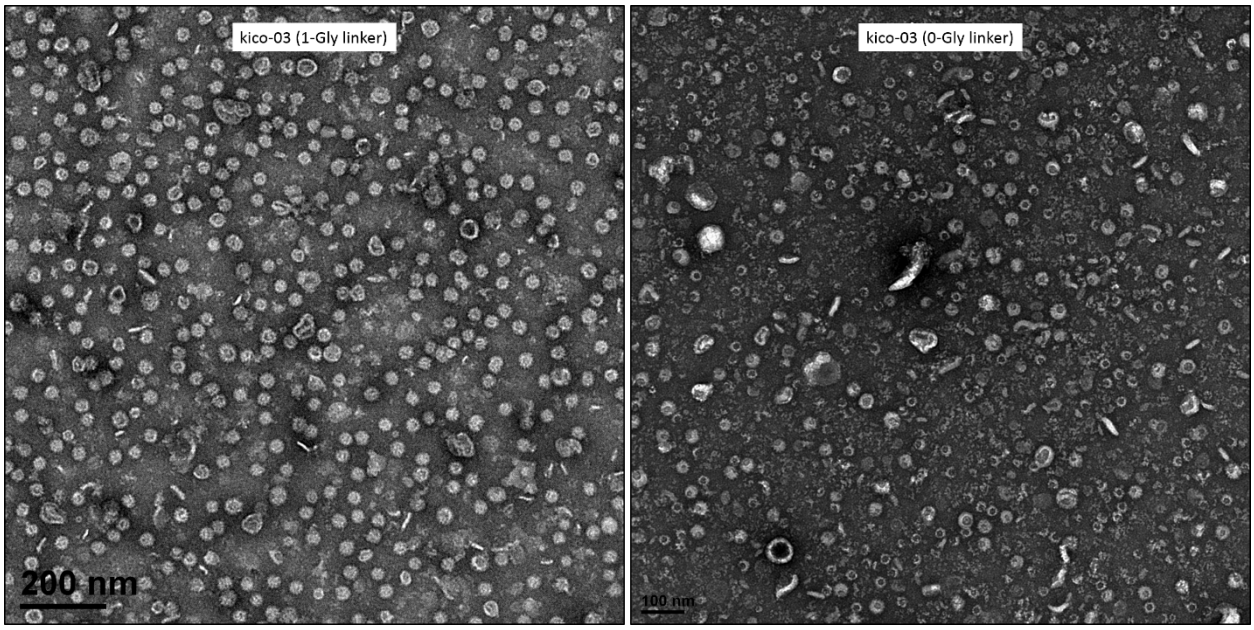


Figure 3.3. Investigation of flexible linker length. The initial dimer-pentamer alpha-helical fusion construct was fused at its N-terminus to a trimeric coiled-coil domain via a flexible glycine linker region. Cages are readily observed with 1 to 3 additional glycine residues added in between the dimer and trimer domains (left). However, when this linker region is removed entirely, kico-03 appears to primarily form partially-closed shells or smaller, distorted assemblies (right).

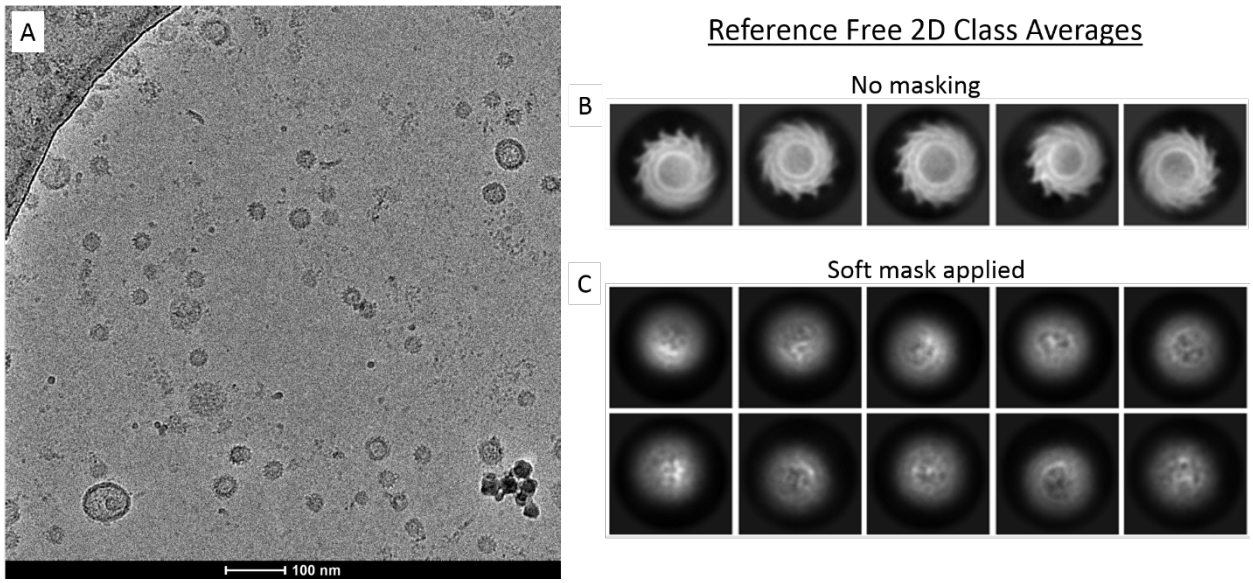


Figure 3.4. Imaging of kico-03 by cryo-EM. (A) Representative cryo-EM micrograph of a flash-frozen kico-03 cage sample. Particles of approximately 30 nm in diameter are readily observed, along with significant larger contaminants. (B) Reference free class averages of particles results in "spiky balls" with no apparent surface features, presumably due to significant heterogeneity of cage assemblies. (C) When a soft mask is applied in cryoSPARC to reduce fitting to the previously observed "spikes," some surface features appear, although icosahedral symmetry elements are still not readily apparent.

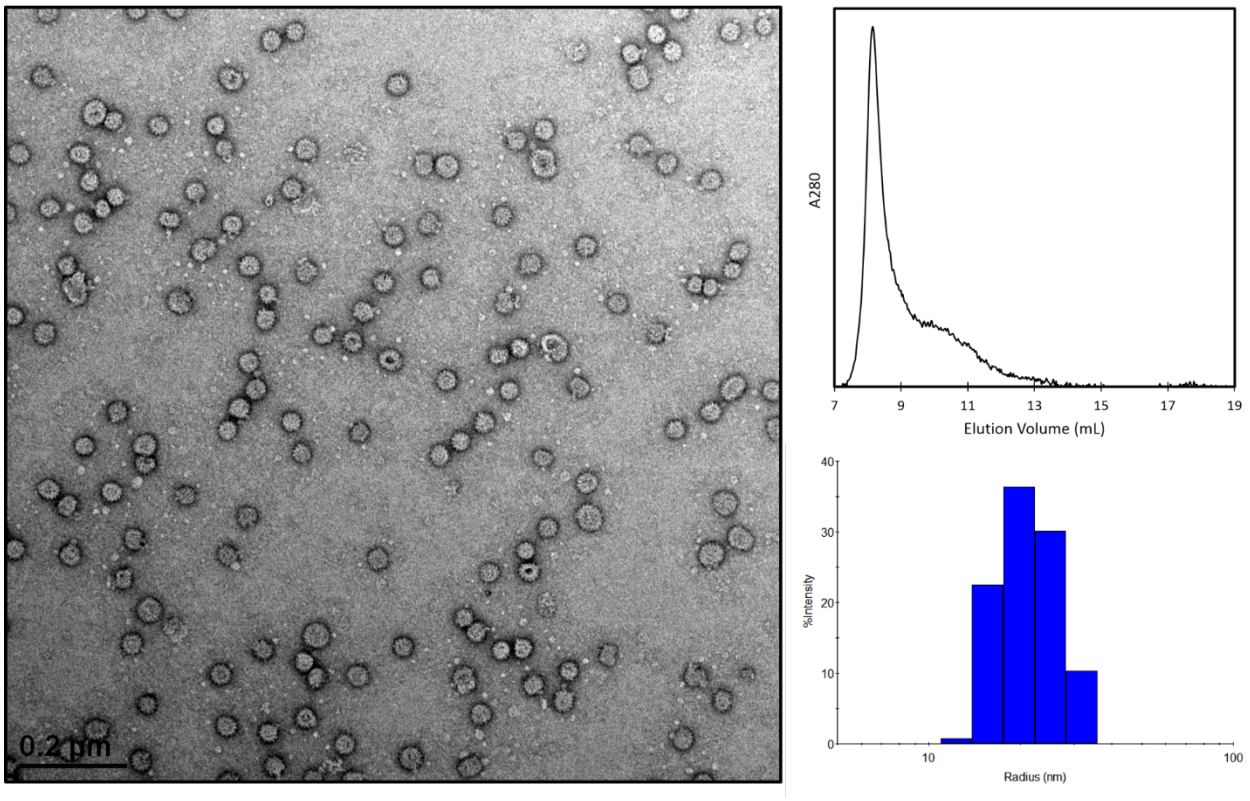


Figure 3.5. Characterization of sucrose gradient-purified kico-03 cages. Purification of the kico-03 cage construct by sucrose gradient centrifugation after SEC leads to cage particles that are much more monodisperse in size as observed by negative stain EM (left), although apparent heterogeneity and fluctuations of particle diameter persist. SEC of the cages post-sucrose gradient (top right) shows that all previous low molecular weight peaks disappear, and a strong single peak remains in the void volume. DLS measurements (representative size distribution plot shown, bottom right) show a major peak at 20 nm radius, with some potentially remaining size heterogeneity.

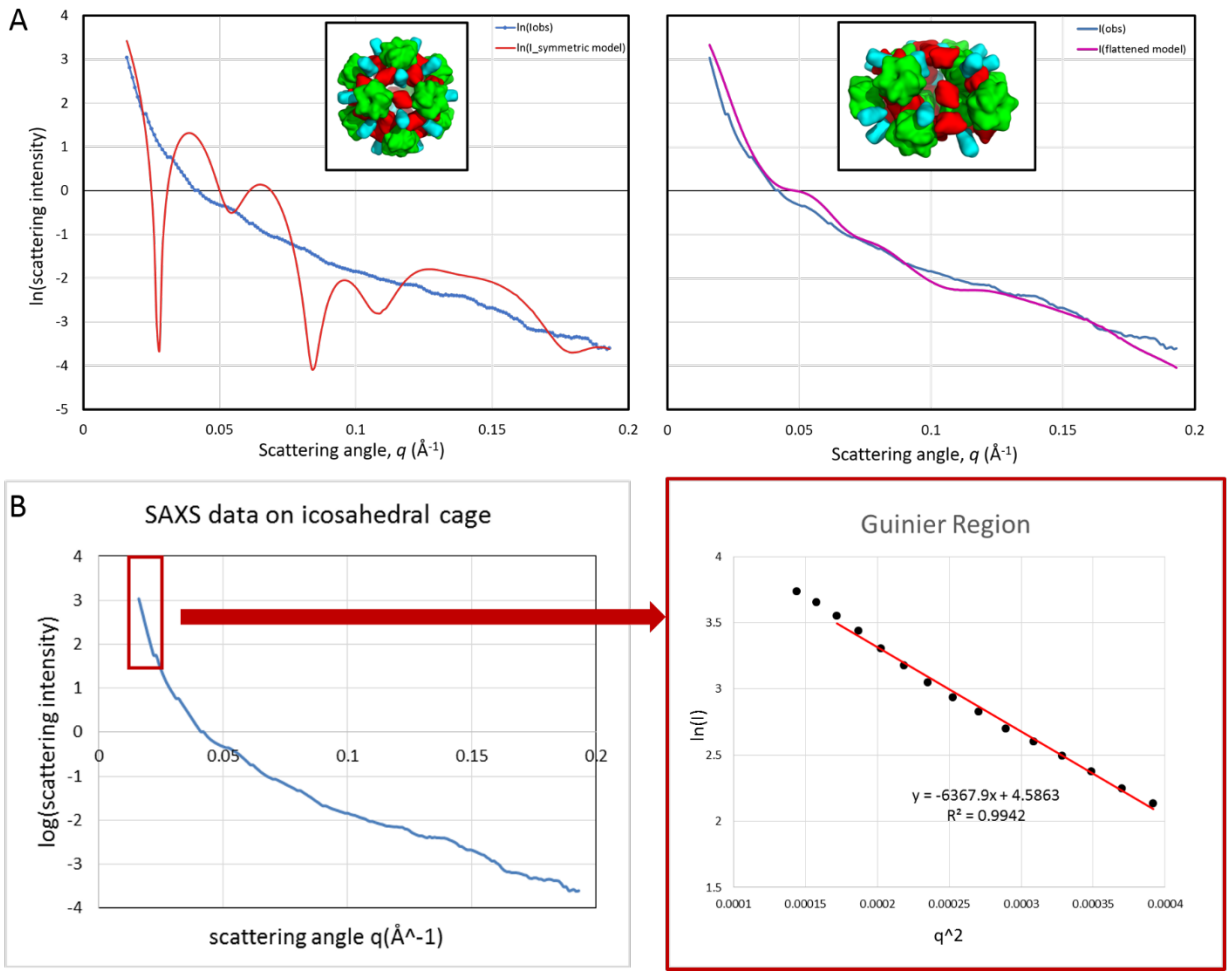


Figure 3.6. SAXS characterization of kico-03. SAXS characterization of kico-03. Sucrose gradient-purified kico-03 cage sample was examined by SAXS to investigate the size and shape of cage particles in solution. The experimental SAXS profile was first compared to the idealized icosahedrally symmetric kico-03 design model (A, left), however, the calculated profile for the model was a poor fit. By allowing significant distortions in the design model, significantly flattened versions of the cage could be generated that fit the observed scattering intensities very closely (see Methods) (A, right). Examination of the Guinier region of the SAXS profile (B) reveals a nearly straight line, with slight curvature indicating limited heterogeneity. A linear fit of the data in the Guinier region indicates a radius of gyration of about 138Å, approximately 17% higher than the calculated radius of gyration for the theoretical model (118Å).

References

1. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of *de novo* protein design. *Nature* **537**, 320–327 (2016).
2. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2217–2221 (2001).
3. Yeates, T. O., 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).
4. Zhang, Y., Ardejani, M. S. & Orner, B. P. Design and Applications of Protein-Cage-Based Nanomaterials. *Chemistry – An Asian Journal* **11**, 2814–2828 (2016).
5. Cannon, K. A., Ochoa, J. M. & Yeates, T. O. High-symmetry protein assemblies: patterns and emerging applications. *Current Opinion in Structural Biology* **55**, 77–84 (2019).
6. Lai, Y. T., Cascio, D. & Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129–1129 (2012).
7. King, N. P. *et al.* Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336**, 1171–1174 (2012).
8. Lai, Y.-T. *et al.* Structure of a designed protein cage that self-assembles into a highly porous cube. *Nature Chemistry* **6**, 1065–1071 (2014).
9. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–8 (2014).
10. Bale, J. B. *et al.* Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. *Protein Science* **24**, 1695–1701 (2015).

11. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–395 (2016).

12. Butterfield, G. L. *et al.* Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).

13. Edwardson, T. G. W., Mori, T. & Hilvert, D. Rational Engineering of a Designed Protein Cage for siRNA Delivery. *Journal of the American Chemical Society* **140**, 10439–10442 (2018).

14. Phippen, S. W. *et al.* Multivalent Display of Antifreeze Proteins by Fusion to Self-Assembling Protein Cages Enhances Ice-Binding Activities. *Biochemistry* **55**, 6811–6820 (2016).

15. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *PNAS* **115**, 3362–3367 (2018).

16. Liu, Y., Huynh, D. T. & Yeates, T. O. A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nature Communications* **10**, 1864 (2019).

17. Marcandalli, J. *et al.* Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17 (2019).

18. Malay, A. D. *et al.* An ultra-stable gold-coordinated protein cage displaying reversible assembly. *Nature* **569**, 438–442 (2019).

19. Yeates, T. O. Geometric Principles for Designing Highly Symmetric Self-Assembling Protein Nanomaterials. *Annual Review of Biophysics* **46**, 23–42 (2017).

20. Sciore, A. *et al.* Flexible, symmetry-directed approach to assembling protein cages. *PNAS* **113**, 8681–8686 (2016).

21. Fletcher, J. M. *et al.* Self-Assembling Cages from Coiled-Coil Peptide Modules. *Science* **340**, 595–599 (2013).

22. Lai, Y. T., Tsai, K. L., Sawaya, M. R., Asturias, F. J. & Yeates, T. O. Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly. *Journal of the American Chemical Society* **135**, 7738–7743 (2013).

23. Lai, Y.-T., Jiang, L., Chen, W. & Yeates, T. O. On the predictability of the orientation of protein domains joined by a spanning alpha-helical linker. *Protein Eng Des Sel* **28**, 491–500 (2015).

24. Lai, Y. T. *et al.* Designing and defining dynamic protein cage nanoassemblies in solution. *Science Advances* **2**, (2016).

25. Badieyan, S. *et al.* Symmetry-Directed Self-Assembly of a Tetrahedral Protein Cage Mediated by de Novo-Designed Coiled Coils. *ChemBioChem* **18**, 1888–1892 (2017).

26. Cristie-David, A. S. *et al.* Coiled-Coil-Mediated Assembly of an Icosahedral Protein Cage with Extremely High Thermal and Chemical Stability. *J. Am. Chem. Soc.* **141**, 9207–9216 (2019).

27. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* **32**, W500–W502 (2004).

28. Lai, Y.-T., King, N. P. & Yeates, T. O. Principles for designing ordered protein assemblies. *Trends in Cell Biology* **22**, 653–661 (2012).

29. Kostan, J. *et al.* Structural and functional characterisation of the chlorite dismutase from the nitrite-oxidizing bacterium “Candidatus Nitrospira defluvii”: Identification of a catalytically important amino acid residue. *Journal of Structural Biology* **172**, 331–342 (2010).

30. Fletcher, J. M. *et al.* A Basis Set of de Novo Coiled-Coil Peptide Oligomers for Rational Protein Design and Synthetic Biology. *ACS Synth. Biol.* **1**, 240–250 (2012).

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

32. Dyer, K. N. *et al.* High-Throughput SAXS for the Characterization of Biomolecules in Solution: A Practical Approach. in *Structural Genomics: General Applications* (ed. Chen, Y. W.) 245–258 (Humana Press, 2014). doi:10.1007/978-1-62703-691-7_18

Chapter 4. Designed Protein Cages as Scaffolds for Building Multi-Enzyme Cellulase Materials

Scott A. McConnell[1,2,3*], Kevin A. Cannon[1,2*], Christian Morgan[4], Brendan R. Amer[1,2,3], Todd O. Yeates[1,2,3], and Robert T. Clubb[1,2,3]

[1]UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA; [2]UCLA Department of Chemistry and Biochemistry, Los Angeles, CA; [3]UCLA Molecular Biology Institute, Los Angeles, CA; [4]UCLA Department of Ecology and Evolutionary Biology, Los Angeles, CA

*These authors contributed equally.

1. INTRODUCTION

Recent advances in computational protein engineering have created specifically designed nanocages that self-assemble into precise geometric architectures[1–4]. Current efforts are now focused on developing these designer assemblies into useful materials that display repetitive protein arrays[5,6]. Recent examples of this work include the creation of nanocages displaying: viral antigens to increase neutralizing antibody responses to infections[7], antifreeze enzymes to increase ice-binding capacity[8] and small proteins for structure determination by cryo-electron microscopy[9,10]. In principle, designer protein nanocages could be ideal platforms for displaying enzymes with precisely defined spatial relationships. These structures would effectively mimic naturally occurring enzyme complexes that increase the efficiency of multi-step reactions by channeling intermediates between sequentially acting enzymes. While recent synthetic biology

efforts have developed platforms that co-localize enzymes to increase pathway flux[13–17], to the best of our knowledge, the exquisite spatial control afforded by designer nanocages has yet to be harnessed.

Lignocellulosic plant biomass is a highly abundant and attractive renewable feedstock for producing biofuels, chemicals and materials. However, its recalcitrance to hydrolysis limits its cost-effective usage on an industrial scale. A common approach to degrade lignocellulose into its component sugars is to employ a consortium of synergistically functioning cellulase enzymes that have distinct substrate specificities. Lignocellulose is comprised of varying amounts of cellulose (25-55%), hemicellulose (8-30%), and lignin (18-35%). Cellulose is a polymer of β-1,4-linked glucose molecules that can hydrogen-bond with other cellulose polymers to form both amorphous and crystalline regions. It is synergistically degraded by three types of cellulases: endoglucanases, exoglucanases, and β-glucosidases. Endoglucanases attack within a cellulose strand to hydrolyze the β-1,4-glucosidic bonds, producing new reducing and non-reducing ends that can be further broken down by exoglucanases. The shorter cellodextrin chains that are produced by these enzymes, including the disaccharide cellobiose, are then degraded into glucose monomers by β-glucosidases. Hemicellulose is a sugar polymer that is composed of a number of different types of pentose and hexose sugars. As compared to cellulose, it is more readily degraded by a range of enzymes, including among others: xylanases, arabinases, and mannases. Finally, lignin surrounds and blocks enzyme access to cellulose and hemicellulose, and is a complex polymer containing a mixture of phenolic compounds linked through radical coupling reactions. A large number of enzymes are needed to degrade it, including peroxidases and laccases.

Clostridium thermocellum and other species of anaerobic bacteria efficiently degrade lignocellulose using cellulosomes, large surface displayed protein complexes that house cellulases

with complementary activities. Cellulase co-localization within these structures promotes synergistic enzyme-enzyme and enzyme-proximity interactions, where the cellulolytic activities of the complexed enzymes are greater than that of individual enzymes due to their complementary activities and optimal enzyme spacing. The presence of both hemicellulases and cellulases within the cellulosome also enables hemicellulose and cellulose fibers to be removed simultaneously, thereby overcoming potential physical hindrances. The benefits of the *C. thermocellum* cellulosomal system have been quantified: its specific activity against crystalline cellulose is 15-fold higher than the secreted enzyme system from *T. reesei*. Moreover, placement of the cellulosome on the microbial surface increases the rate of hydrolysis by promoting cellulose-enzyme-microbe synergy. In this process, sugar uptake by the microbe presumably becomes more efficient by promoting import of the products into cells and by removing potential enzyme inhibitors such as glucose and cellobiose from the environment.

In this study, we sought to develop a modular platform to produce protein-coated nanocages. Previously described cage modification methods have attached proteins via non-covalent interactions or by expressing proteins as genetic fusions with a cage subunit. Our approach is unique, and harnesses the robust ligation activity of the *S. aureus* sortase A enzyme, a widely used cysteine transpeptidase. Here we show that the surface of designer T33-21 nanocages can be elaborated with cellulase enzymes using sortase. Nanocage labeling is efficient, yielding cellulolytic protein nanoparticles whose component enzymes function synergistically. This enzyme display system enables virtually any protein or peptide containing molecule to be grafted to the exterior of the T33-21 cage and is a step toward creating more elaborate cellulase coated materials that could be useful in producing renewable biofuels and chemicals.

2. MATERIALS AND METHODS

Recombinant protein expression and purification. T33-21 protein cages were prepared by co-expressing genes encoding its T33-21A and -21B subunits from a pET-22b plasmid (Novagen). Genes were purchased from IDT and inserted *via* Gibson Assembly into the vector. The amino acid sequences were the same as previously reported for the T33-21 cage[3], except that a 16-residue sortase recognition tag (QSKKSELPETGGEEST) was appended to the C-terminus of the A component. All cellulase proteins used in this study were expressed with pE-SUMO (LifeSensors) expression vector. The expression plasmids containing the individual cellulase enzymes were ligated into a pSUMO expression plasmid using Gibson Assembly. The assembly reaction was engineered to include a penta-glycine repeat at the junction between the SUMO and cellulase gene, such that cleavage by Ulp1 protease yields a cellulase with a functional G_5 nucleophile at its amino-terminus. This approach avoids potential problems associated with the incomplete removal of the N-terminal methionine that can arise when sortase substrates are expressed in bacteria with the pentaglycine nucleophile directly following the start codon. Additionally, each cellulase was appended at its carboxy-terminus with a peptide epitope to enable immunological identification of each of the cellulase components. An improved pentamutant variant of the sortase A gene from *Staphyloccocus aureus* was used in this study, bearing the following mutations: P94R/D160N/D165A/K190E/K196T[23]. This gene was cloned into the pSUMO expression vector.

For all constructs, *E. coli* BL21 (DE3) cells harboring the target expression plasmid were grown in LB supplemented with ampicillin at 100 µg/ml (T33-21-SR) or kanamycin at 500 µg/ml (cellulase constructs) at 37^{0}C until OD_{600} of ~0.6. Cells were induced with 1 mM IPTG and protein expression was allowed to proceed overnight at 17°C. Cells were then harvested by centrifugation (7,000 x g for 10 min). Cells containing the T33-21-SR construct were lysed by sonication in a

buffer of 50 mM Tris pH 8.0, 250 mM NaCl, and 20mM imidazole. The cage was then subjected to Ni-affinity purification using gradient elution in the range of 20-500mM imidazole. Elution fractions containing the protein cage components (confirmed by SDS-PAGE) were then dialyzed into a buffer of 25 mM Tris pH 8.0, 150 mM NaCl, and 1 mM DTT to be used in subsequent experiments. All cellulase proteins were purified as a His_6-SUMO- fusion using HisPure Co^{2+} IMAC resin (Thermo Scientific) per the manufacturer's instructions. Briefly, cell pellets were resuspended in 50 mM Tris-HCl pH 8.0, 300 mM NaCl (lysis buffer) and lysed by sonication. The cell lysate was then fractionated by centrifugation (15,000 x g for 40 min) and the supernatant was loaded onto HisPure Co^{2+} IMAC resin. Proteins were then eluted from the resin using lysis buffer supplemented with 200 mM Imidazole. The His6x-SUMO tag was removed by adding His_6-Ulp1 protease, and subsequent HisPure Co^{2+} purification. All proteins were then loaded onto a SuperDex75 size exclusion column as a final purification step. Protein purity was determined by SDS-PAGE analysis.

Cellulase labeling of Nano-cages using sortase. Sortase bioconjugation reactions to covalently ligate the engineered cellulase components to synthetic protein cages were performed as 500µL reactions at room temperature. All proteins were dissolved in SrtA modification buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 5 mM DTT, 5mM $CaCl_2$). In the reactions the following protein concentrations were used: 1 µM SrtA enzyme (improved variant '5M' from *Staphylococcus aureus*), 20 µM of assembled T33-21-SR protein cages and 200 µM G_5-Cel8A-FLAG and/or 200 µM G_5-Cel48S-Myc. Reactions were allowed to proceed for 2.5 hours while being dialyzed against SrtA modification buffer to remove the hydrolysis product and prevent reverse reaction. The reaction components were then separated using a pre-cast NuPAGE 4-12% gradient Bis-Tris protein gel (Thermo Scientific). Gels were stained with Coomassie Blue G-250 for analysis of

reaction progress. Assembled cages were separated from other reaction components, including monomeric cellulases and sortase, using size-exclusion chromatography with a SuperDex75 column (GE Life Sciences).

Functionalized nano-cages quantification. Enzyme concentrations on purified cages were quantified by western blot using α-myc and α-FLAG antibodies. Protein standards were created for each of the cellulase constructs by dilution to known concentrations. On the same SDS-PAGE gel, serial dilutions of each type of cellulase-modified cage were run alongside the standards. After the SDS-PAGE was run, the separated proteins were transferred to a PVDF membrane using the iBlot2 device (20W, 7min, Invitrogen). The membranes were then sequentially blocked, incubated with primary antibody (mouse α-myc or α-FLAG), washed, incubated with secondary antibody (rabbit α-mouse IgG-HRP), and washed again by capillary action using the iBind system (Invitrogen). Immunological detection of proteins with appropriate peptide tags was facilitated by incubation of the membranes with luminol substrate. Luminescent signal is detected by autoradiography film (Genessee Scientific, 30 second exposure). The intensity of signal arising from lanes with protein standards (free cellulases of known concentration) was analyzed to create a standard curve. Intensity arising from bands with serial dilutions of the stocks of cellulolytic cages were fit to the standard curve to yield a precise measurement of the amount of cellulase present on each of the modified cages.

Cellulase activity of the modified cellulases was determined using a cellulose degradation assay. Avicel PH101 (Sigma) was weighed and washed 3 times with ddH2O, then diluted to the desired cellulose concentration in cellulase assay buffer (50mM sodium acetate, pH 5.5, 2mM DTT, 2mM $CaCl_2$, 0.01% BSA). For the cellulose degradation assays, Cel8A (an endoglucanase)

and Cel48S (an excoglucanase) were selected on the basis of their well-understood established synergistic interactions. Cellulolytic degradative activity of the engineered cellulase constructs was confirmed by measurement of reducing sugars released after incubation with amorphous cellulose substrates (CMC). All proteins were produced heterologously in *E.* coli. 800 µL of enzyme mixtures were then added to 200 µL of Avicel substrate and the enzyme assays were incubated at 37°C with shaking at 225 rpm for specified incubation periods. For each time point, 100 µL aliquots were centrifuged to pellet the insoluble Avicel and 50 µL of the supernatant was mixed with 75 µL 3,5-Dinitrosalicylic acid (DNS) reagent for quantification of reducing sugars released during the reaction, using glucose as a standard[24].

Negative stain electron microscopy was used to characterize the structure and integrity of the modified nanocages. Cage constructs were imaged by electron microscopy to verify the correct assembly of cages before and after the sortase-mediated attachment of cellulase enzymes. Proteins were subjected to size exclusion chromatography using a Superose 6 Increase 10/300 (GE Life Sciences) column before application to EM grids. 5 µL of purified cages in the concentration range of 0.005 mg/mL to 0.02 mg/mL were applied to glow discharged, formvar/carbon-coated 300-mesh copper grids (Ted Pella, Inc.) and stained with 2% uranyl acetate. Cages were imaged on an FEI Tecnai T12 transmission electron microscope at 120kV.

RESULTS AND DISCUSSION

We developed a general method to create enzyme coated designer nanocages toward the goal of emulating naturally occurring multienzyme complexes that catalyze reactions with improved activity and fidelity. As a ‘proof of principle’, we coated pre-assembled T33-21 nanocages with cellulases, which are known to exhibit synergistic activity against crystalline

cellulose when clustered together within protein complexes such as cellulosomes. T33-21 nanocages are constructed from two types of proteins (T33-21A and -21B) that spontaneously assemble into a "dual tetrahedral architecture" that contains twelve copies of each subunit (**Fig. 1A**). In our procedure, enzymes are covalently attached to pre-assembled nanocages using the *S. aureus* sortase A transpeptidase, a powerful bioconjugation tool that joins biomolecules together via a backbone-backbone peptide bond. Sortase links proteins together when they contain LPXTG and oligoglycine sequences at their C- and N-termini, respectively. It catalyzes a transpeptidation in which the Thr-Gly bond within the LPXTG sequence is broken, and then replaced with a new Thr-Gly peptide linkage to the N-terminal amine group within the oligoglycine sequence (**Fig. 1B**). We reasoned that pre-assembled T33-21 nanocages would be good substrates for modification by sortase, as the C-termini of its subunits are solvent exposed and thus potential sites for sortase modification if they contained a LPXTG sequence. We therefore produced T33-21 nanocages in which the T33-21A subunit is recombinantly fused to a C-terminal LPXTG sequence ('sort tag'). Cages were elaborated with purified cellulases that were recombinantly fused to an N-terminal penta-glycine (Gly_5) peptide sequence. To maximize their reactivity, the enzyme fusions were designed to unmask the terminal glycine during the purification procedure (see methods). Short peptide epitope tags (myc and FLAG) were also appended to the C-termini of each cellulase for immunological probing. Cages were modified with the Cel48S exoglucanase and Cel8A endoglucanase from *C. thermocellum* because they are well characterized and known to synergistically degrade cellulose.

T33-21 nanocages harboring a 'sort tag' are efficiently modified with cellulases by sortase. Initially we developed conditions to modify cages with a single type of enzyme, either Cel48S or Cel8A. In these reactions, assembled T33-21 nanocage containing a 'sort tag' is incubated with

the sortase enzyme (SrtA) and cellulase enzymes possessing an N-terminal Gly_5 sequence. In order to increase the rate of cage modification, we utilized a sortase variant that has 120-fold greater activity than the wild-type enzyme.[23] Sortase bioconjugation reactions are reversible, as the products contain N-terminal glycine and LPXTG sequences and can therefore react with sortase to regenerate the initial reactants. We mitigated this problem by conducting the reaction in dialysis conditions to remove the short peptide cleavage product and by contracting reaction times to 60 minutes[26–29]. SDS-PAGE analysis of the modification reactions confirms the cages are modified, as the T33-21A (21A-LPETG) protein component is progressively converted into higher molecular weight ligated species as the reaction proceeds (**Fig. 2A**). In particular, 21A-LPETG is converted to 21A-Cel8A and 21A-Cel48S polypeptides when reactions are performed with appropriate cellulases harboring an N-terminal Gly_5 sequence (left and middle panels of **Fig. 2A**, respectively). Based on quantification of the band intensities we estimate that ~80-90% T33-21A is modified after one hour.

We generated cages modified with both types of cellulases by repeating the modification reaction using Cel48S and Cel8A. As both enzymes contain a sortase reactive N-terminal Gly_5 sequence, they can be expected to be stochastically ligated to the T33-21A subunit of the cage. Prior studies have shown that maximal synergy is observed when the endoglucanase is present in excess relative to the exoglucanase (Yoav, 2017). Cages were therefore modified using a 2:1 ratio of Cel8A:Cel48S, leading to the expected ratio of the 21A-Cel8A and 21A-Cel48S polypeptides in the modified cages (**Fig. 2A**, right panel). To estimate the number cellulases attached to each cage they were purified using size exclusion chromatography (**Fig. 2B** and **Fig. S1**) and then analyzed using quantitative Western blotting. Purified singly (Cage-48S and Cage-8A) and doubly (Cage-8A/48S) modified nanocages were subjected to SDS-PAGE followed by immunoblot

detection of the myc or FLAG-epitopes present on Cel8A and Cel48S, respectively (**Fig. 3**). This analysis confirmed the presence of the appropriate cellulase on each type of nanocage; probing gels with an anti-myc antibody revealed the presence the 21A-Cel48S polypeptide in Cage:48S and Cage:8A/48S (**Fig. 3A**), whereas a band with a molecular weight consistent with the 21A-Cel8A polypeptide is detected with an anti-FLAG antibody only in the Cage:Cel8A and Cage:Cel8A/Cel48S nanocages (**Fig. 3B**). Absolute amounts of each type of enzyme attached to the cages were estimated by comparing band intensities of serial-diluted nanocages with corresponding standard curves obtained using known amounts of purified Cel8A or Cel48S (**Fig. S2**). This analysis reveals that the single modified T33-21:8A and T33-21:48S cages are coated with ~4.6 Cel8A and ~6.1 Cel48S enzymes, respectively; ~38% and ~50% of the available 21A subunits in the cages are modified with Cel8A and Cel48S, respectively. As expected, a greater fraction of the 21A submits are labeled with enzymes in the doubly modified Cage-8A/48S cages (~78% of the 21A subunits) because nearly twice as much cellulase substrate was used in labeling reactions. On the Cage-8A/48S structure, ~6.4 and ~3 molar equivalents of Cel8A and Cel48S are present.

Cage-8A/48S was examined by negative stain electron microscopy (EM) to confirm that they formed ordered assemblies **(Fig. 4).** Before cellulase attachment with sortase, images of unmodified cages correspond closely to previously published EM images of T33-21 nanocages **(Fig. 4A)**[3]. This is expected and demonstrates that adding the 'sort tag' to the solvent exposed C-terminus of the 21A subunit does not adversely affect cage assembly or stability. After fusing Cel8A and Cel48S to the cage via sortase, the symmetry of the cage assemblies is less readily apparent, but the scaffold core maintains its correct size and appropriate shape (**Fig. 4B**). Notably, the presence of the cellulase enzymes is observed despite their flexible linkage to the T33-21A

subunit, appearing as distinct puncta localized around the scaffold core. The EM results confirm that cellulases are attached to the exterior of the T33-21 nanocage.

Enzymes immobilized to the surface of the nanocage are active and function synergistically to degrade crystalline cellulose. The ability of the modified nanocages to degrade crystalline Avicel was determined using standard methods that monitor the production of reducing sugars (**Fig. 5**). Singly modified cages containing Cel8A are more active than cages that only contain Cel48S, an observation that is also observed when the activities of the isolated enzymes (non-cage attached) are compared. This finding is consistent with the distinct substrate specificities of these enzymes, since crystalline cellulose is expected to contain significantly fewer non-reducing ends that are substrates for the Cel48S exoglucanase as compared to internal sites that can be cleaved by Cel8A. Interestingly, statistically significant gains in activity are achieved by co-localizing both types of enzymes on Cage:Cel8A/Cel48S. The largest effects are observed within 8 hours, with Cage:Cel8A/Cel48S activity 1.9 fold greater than the summed activity of the singly modified cages (**Fig. 5A**). To quantify the effect of enzyme immobilization in the different types of cages we calculated the stimulation factor (SF), which is the fold increase of activity in the enzyme coated cage compared to the corresponding free enzyme. Strikingly, after 8 hours Cage:Cel8A/Cel48S exhibit a SF value of 2.2, whereas the corresponding mixture of singly modified cages has a SF value of 1.2. This indicates that co-localization on the same nanocage scaffold potentiates synergistic enzyme and is consistent with the complementary endo- and exo-glucanase activities of the displayed enzymes. Interestingly, the degree of synergy declines as incubation times approach 24 hours (**Fig. 5B**). The origin of this effect is unclear, but prolonged incubation could in principle cause the cages to dissociate leading the production of less active monomeric 21A-Cel8A and 21A-Cel48S polypeptides. To investigate this issue, we used EM to analyze enzyme

modified nanocages after they had been incubated at 37°C for 24 hours. A micrograph of these cages show that they are fully assembled, suggesting that they are thermostable or capable of reassembling upon transitioning to cryogenic temperatures.

Our system is distinct from previously reported protein complexes that have sought to harness cellulase synergies by co-localization. The most popular approach has been to construct miniaturized cellulomes ('mini'-cellulosomes) in which cellulases are bound to a central scaffoldin polypeptide via complementary dockerin-cohesin binding modules. While significant increases in activity have been achieved, these structures are generally less cellulolytic than more complex native cellulosomes[21,31–33,35] and their production can be laborious as the large scaffoldin is notoriously difficult to express and purify. Non-cellulosome based strategies have also been pursued, including functionalizing inorganic ferromagnetic nanoparticles with cellulases[31 36], and more recently, using naturally self-assembling protein scaffolds to display cellulases via dockerin-cohesin interactions[37 38]. In one such study, four different cellulases were engineered with corresponding dockerin modules and mixed in different ratios to create cellulolytic "Rosettazymes"[38]. Direct activity comparisons with these systems are not possible because distinct assay conditions and cellulose substrates have been employed. In addition, cellulose degradation is known to exhibit nonlinear kinetics, causing measured values to be heavily influenced by enzyme and substrate concentrations. Bearing this limitation in mind, our cellulolytic cages display a specific activity of 0.068 U/mg, which is comparable to related cellulolytic assemblies that contained two cellulases. Moreover, literature precedence suggests that straightforward additions to our platform will yield significant gains in cellulolytic activity. These include adding a β-galactosidase to eliminate the build-up cellobiose that can cause feedback inhibition, adding

cellulose binding modules to promote targeting to cellulose and increasing cellulase diversity to expand the range of substrates that can be degraded.

The labeling approach described here is general method that can be used to label T33-21A nanocages with any type of biomolecule as long as it contains an amino terminal glycine residue (e.g. lipids, peptides, fluorophores etc.). Indeed, we have also demonstrated robust T33-21A nanocage labeling with non-cellulolytic proteins **(Fig. S4)**. Our method can also be extended to decorate more complex cage designs as long as they possess surface exposed polypeptide termini. Importantly, these cages can contain either exposed N- or C-termini, which can be engineered to contain sortase reactive polyglycine and LPXTG sequences for modification, respectively. Protein scaffolds have previously been modified with functional groups by chemical reactions[39–41] to enable targeting to specific cells or tracking. Methods to encapsulate enzymes inside of protein cages to promote specialized reactions, or to provide controllable substrate release have also been reported[42–45]. The findings presented here represent the enzymatic functionalization of a designed protein scaffold using sortase technology and illustrate the enormous potential of this approach.

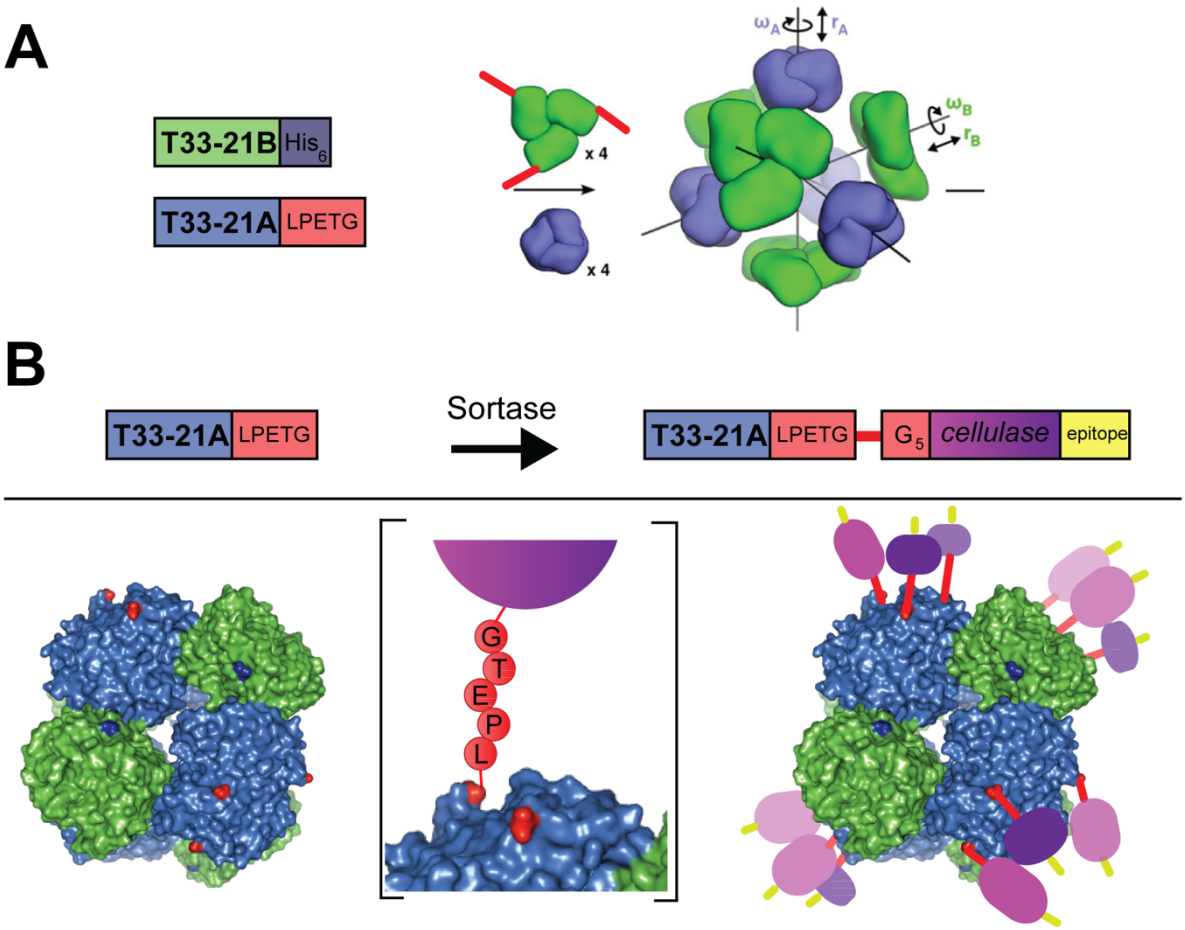


Figure 4.1. A) Schematic of self assembly architecture of designer T33-21-sort tag cage design. The self-assembling T33-21-SR protein nanocage consists of 4 sets of trimeric binding partners, T33-21A (blue) and T33-21B (green), which are engineered to include a C-terminal sort-tag or His_6 tag, respectively. B) Schematic diagram of the construction of cellulolytic nanocages. Incubation of designer cages with sortase and recombinant cellulase enzymes (purple) harboring nucleophilic pentaglycine (red) and peptide epitope (yellow) results in functionalized cage scaffold with cellulolytic enzymes of synergistic activities.

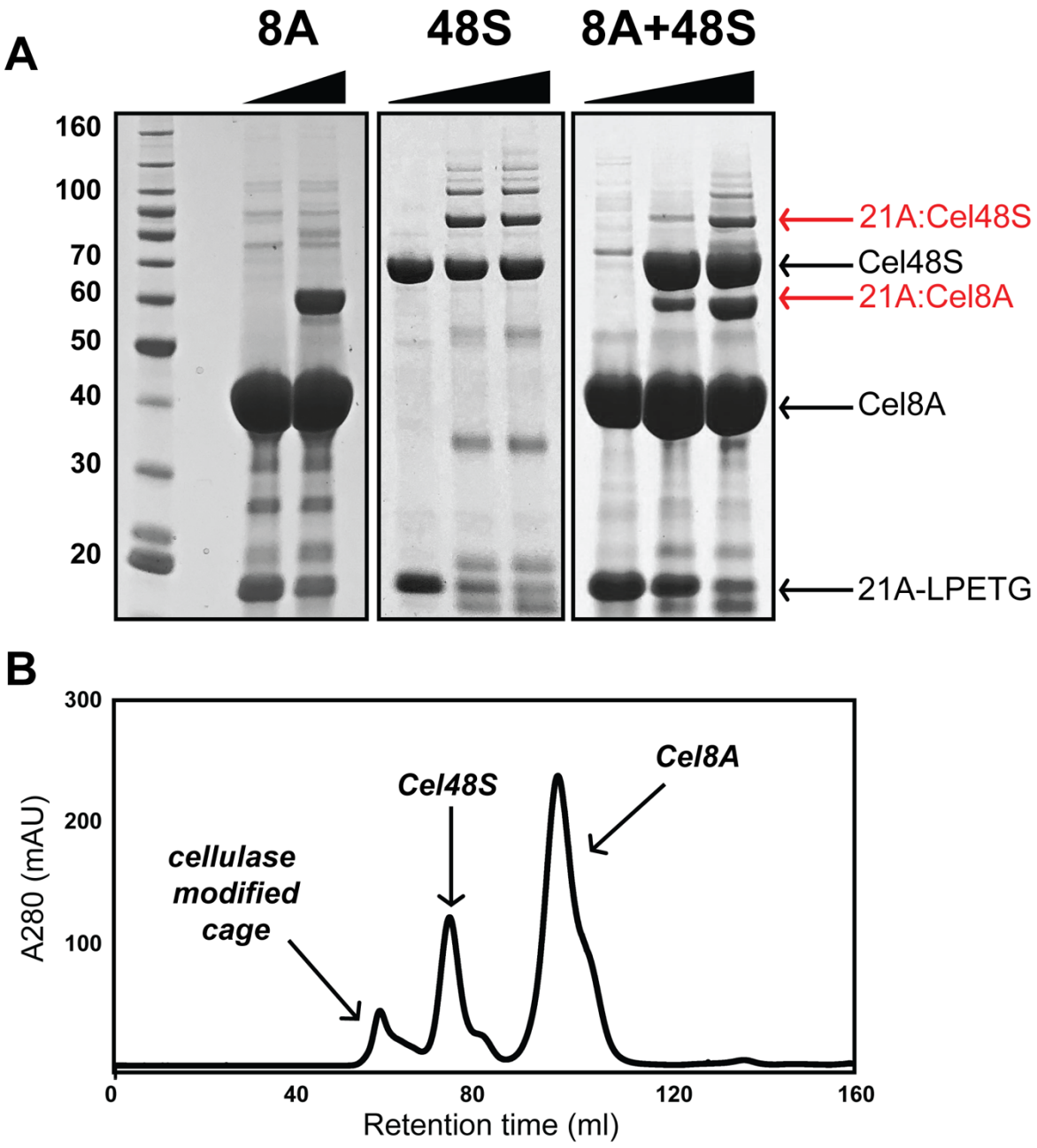


Figure 4.2. Sortase-catalyzed modification of cages is (A) tracked by SDS-PAGE separation of reaction components, (B) purified by size exclusion. (A) Reaction time points at 0, 60 min or 0, 30, 60 min are separated by SDS-PAGE. Expected molecular weights for the products and substrates are indicated. Left panel: monovalent modification with Cel8A, middle panel:

monovalent modification with Cel48S, right panel: divalent modification with both cellulases. For the divalent modification, a ratio Cel8A:Cel48S ratio of 2:1 was desired, so the cages were first reacted with Cel8A for 30 min, then Cel48S was added for an additional 30 min incubation. (B) Representative chromatogram from the divalent cage modification reaction from gel filtration purifications. Size exclusion was used to successful purify cellulolytic cages from monomeric cellulase components.

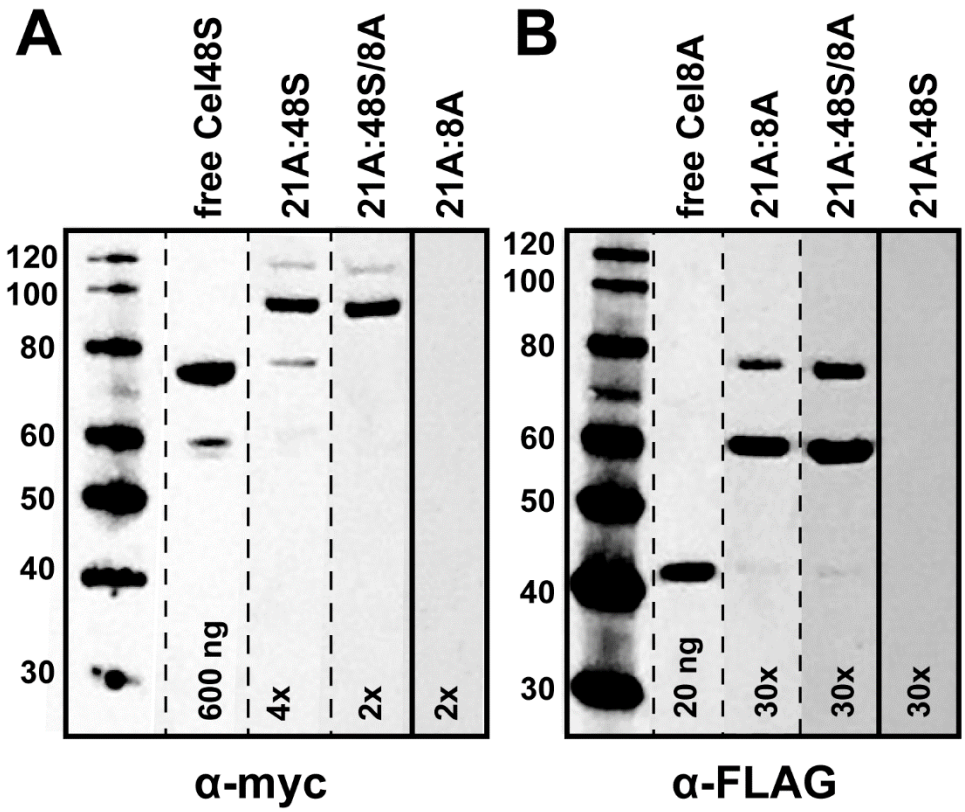


Figure 4.3. The presence of cellulase on each purified cellulolytic cage was probed by immunoblot. Detection by α-FLAG antibody for the endoglucanase Cel8A or α-myc for the exoglucanase Cel48S, and then detected using rabbit α-mouse IgG conjugated to horse radish peroxidase. For both immunoblots, a band corresponding to the ligated product T33-21A-

cellulase is observed when incubated with the corresponding immunoblot. Negligible amounts of monomeric cellulase are visible in each purified stock. No cross reactivity between the myc and FLAG epitopes is observed in monovalent cages lacking the epitope of interest (e.g. no signal is observed for mC-8A in the α-myc immunoblot.

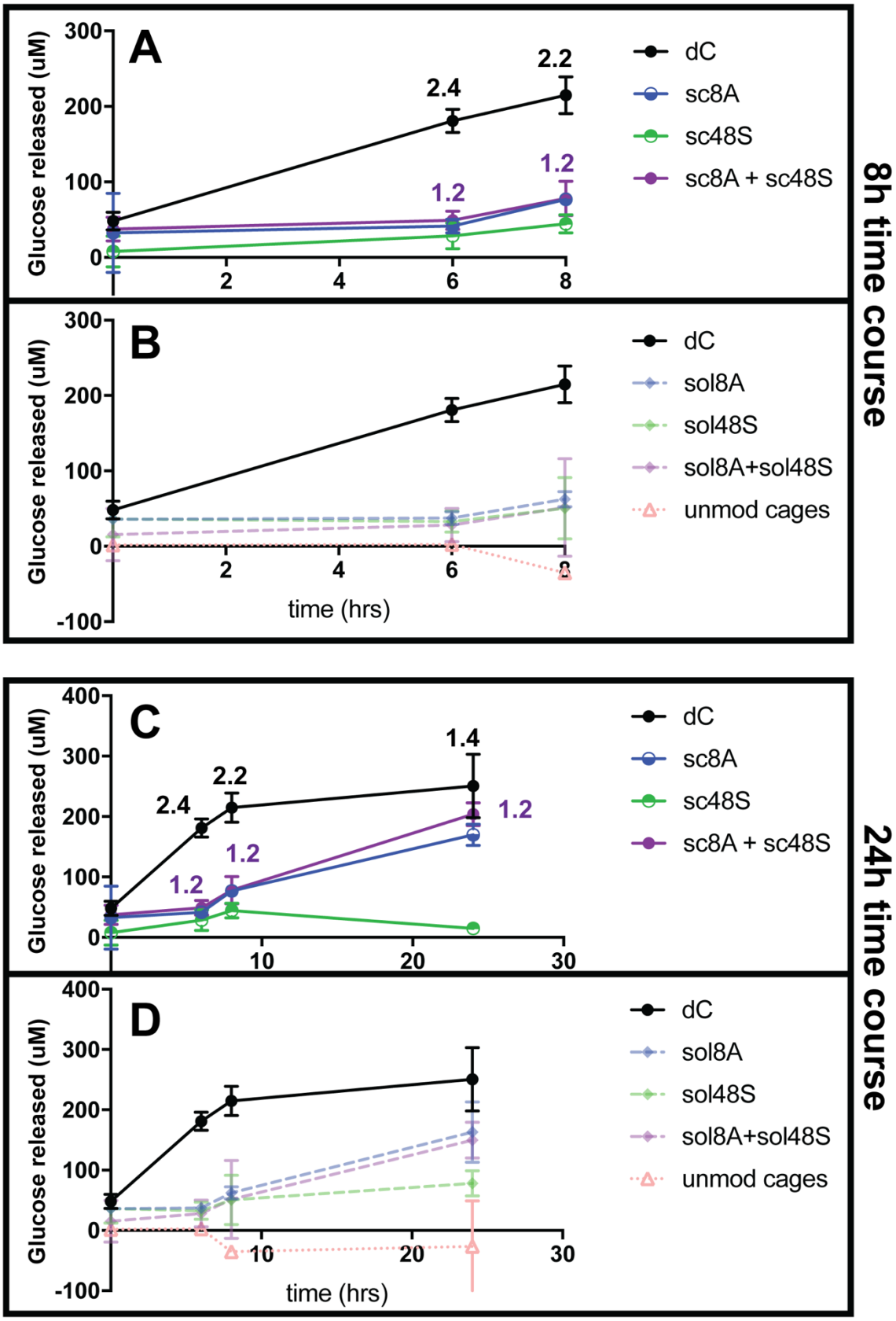

A
B
C
D
Glucose released (uM)
time (hrs)
dC
sc8A
sc48S
sc8A + sc48S
sol8A
sol48S
sol8A+sol48S
unmod cages
2.4
2.2
1.2
1.4
8h time course
24h time course

Figure 4.4. Avicel degradation by cellulolytic cages and corresponding free enzyme mixtures was assayed at several time points (6,8,24 hours). Cellulose degradation was quantified by the release of reducing sugars in each reaction which was detected by incubation of reaction aliquots with dinitrosalicylic acid using established methods. A/B) Full reaction time course (extended to 24 hour time point). A) Divalent cage cellulose degradation is compared at different time points to the activity of corresponding concentrations of the same enzymes on monovalent cages. B) Divalent cage cellulose degradation is compared at different time points to the activity of corresponding concentrations of the free enzymes.
C/D) 8 hour time course. C) Divalent cage cellulose degradation is compared at different time points to the activity of corresponding concentrations of the same enzymes on monovalent cages. D) Divalent cage cellulose degradation is compared at different time points to the activity of corresponding concentrations of the free enzymes.

Table 4.1. Released sugars from Avicel after 8h at 37C by mono- or di-valent cages and corresponding free enzymes.

	dC8A,48S	**mC8A**	**mC48S**	**mC8A +mC48S**	**f8a**	**f48s**	**f8A + f48S**	**uC**	**BG**
released glucose (uM)	294.7 +/- 24	156.6 +/- 0.4	124.4 +/- 12	158.1 +/- 23	142.4 +/- 10	130.4 +/- 41	131.5 +/- 65	44.8 +/- 1	103.6 +/- 11
SF	2.24			1.20					
activity (U, x10^{-3})	0.61	0.33	0.26	0.33	0.30	0.27	0.27		
specific activity (U/mg)	0.068	0.065	0.065	0.037	0.059	0.068	0.030		

dC = divalent cages (T33-21:Cel8A:Cel48S), mC = monovalent cages (T33-21:Cel8A or T33-21-Cel48S), uC = unmodified cages, F = free enzyme, BG = beta-glucosidase only

SF = (released soluble sugars by cellulolytic cages)/(released soluble sugars by the corresponding free enzyme pairs)

1U= amt of enzyme producing 1umol of reducing sugar per minute

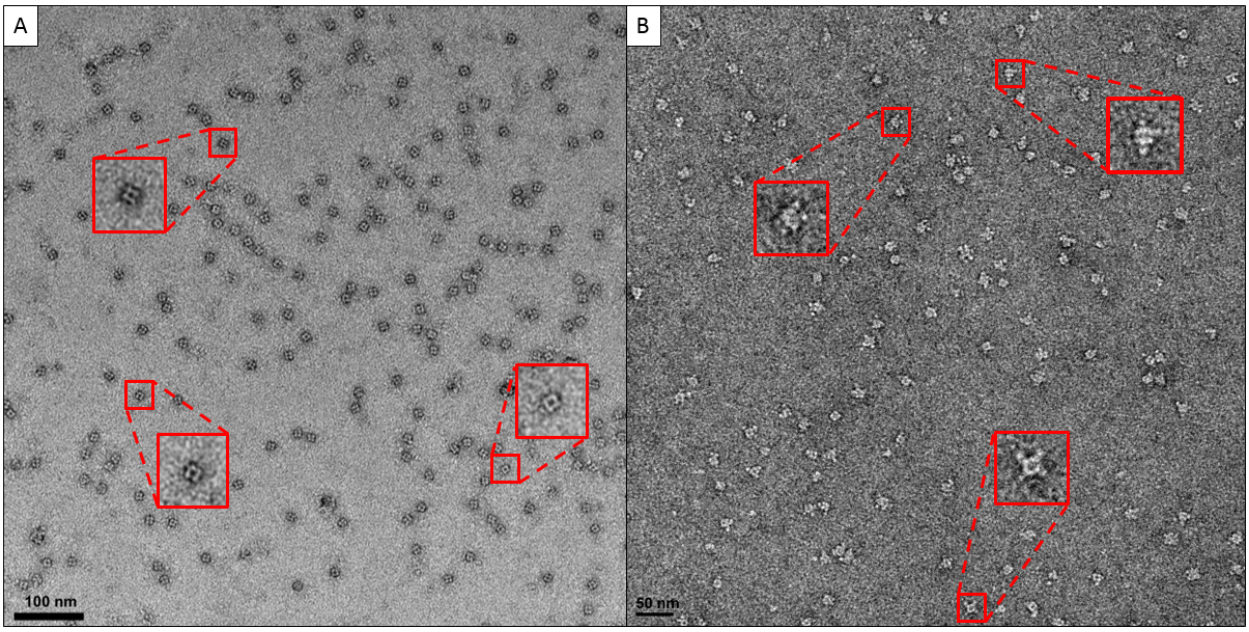
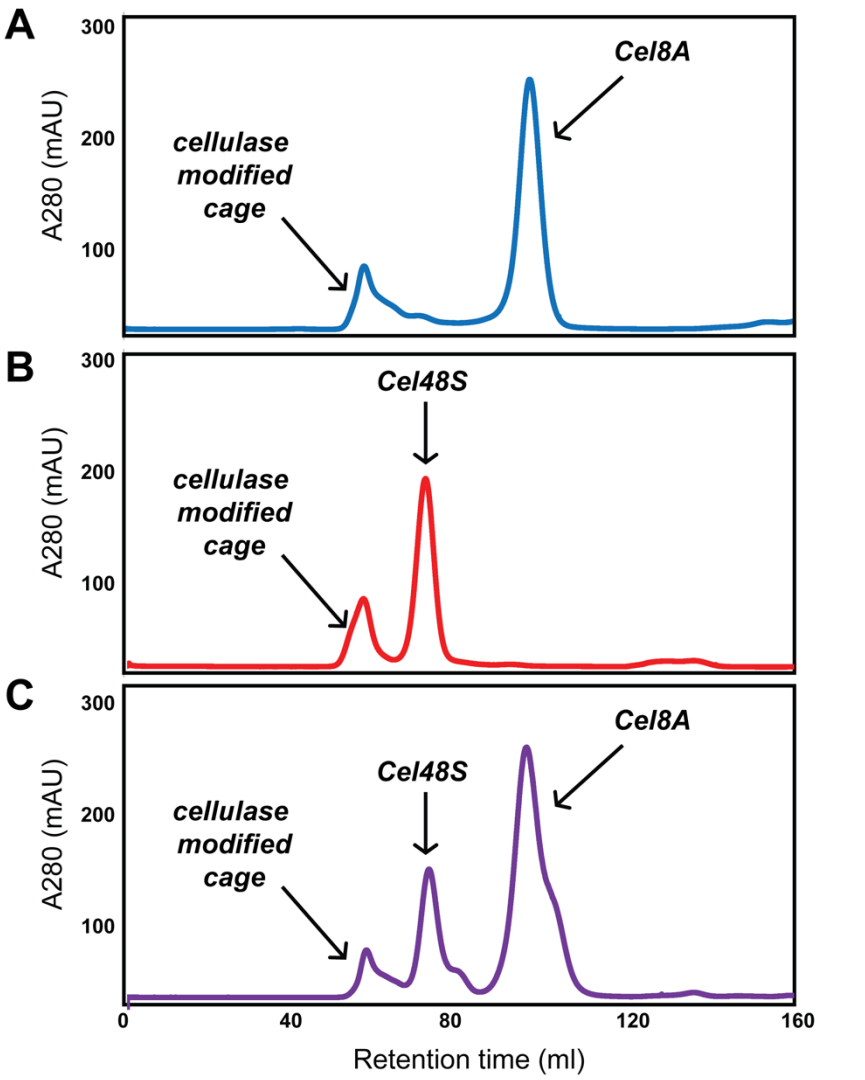
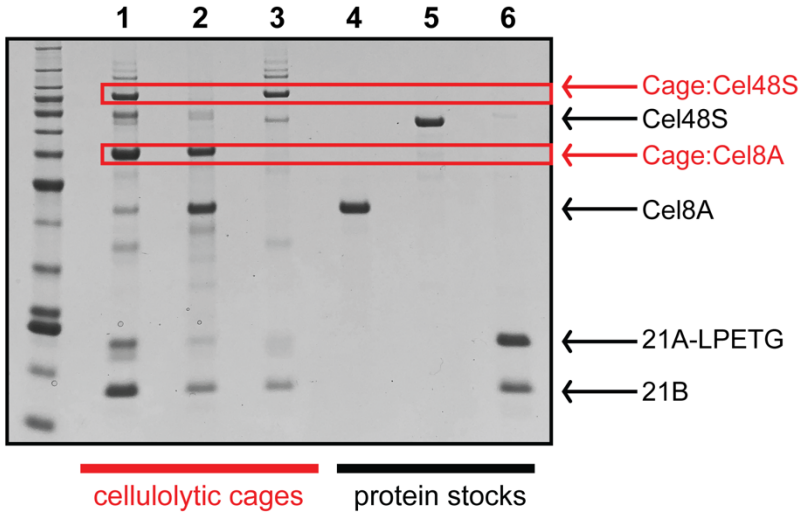


Figure 4.5. Representative negative stain EM images of the T33-21-SR cage before and after sortase ligation. (A) Individual T33-21-SR particles imaged before sortase ligation closely match previously published micrographs of the original cage construct. Highlighted particles show clear views along the 2-fold symmetry axis of the tetrahedron. (B) After fusing Cel8A and Cel48S to the cage *via* sortase ligation, distinct puncta localized around the cage particles, corresponding to the attached cellulase enzymes, are observed.

SUPPLEMENTAL DATA





1 - divalent cage:8A/48S, 2 - monovalent cage:8A, 3 - monovalent cage:48S
4 - free Cel8A, 5 - free Cel48S, 6 - unmodified cages

Figure 4.S1. Size-exclusion purification of cellulolytic cages. A) Representative chromatogram from gel filtration purification of the monovalent modification reaction with Cel8A. B) Representative chromatogram from gel filtration purification of the monovalent modification reaction with Cel48S. C) Representative chromatogram from gel filtration purification of the divalent modification reaction with Cel8A and Cel48S. D) Cellulolytic cage stocks after the above procedure were assessed for purity by SDS-PAGE. Each stock was diluted to the working enzyme concentrations used in the cellulose degradation assay.

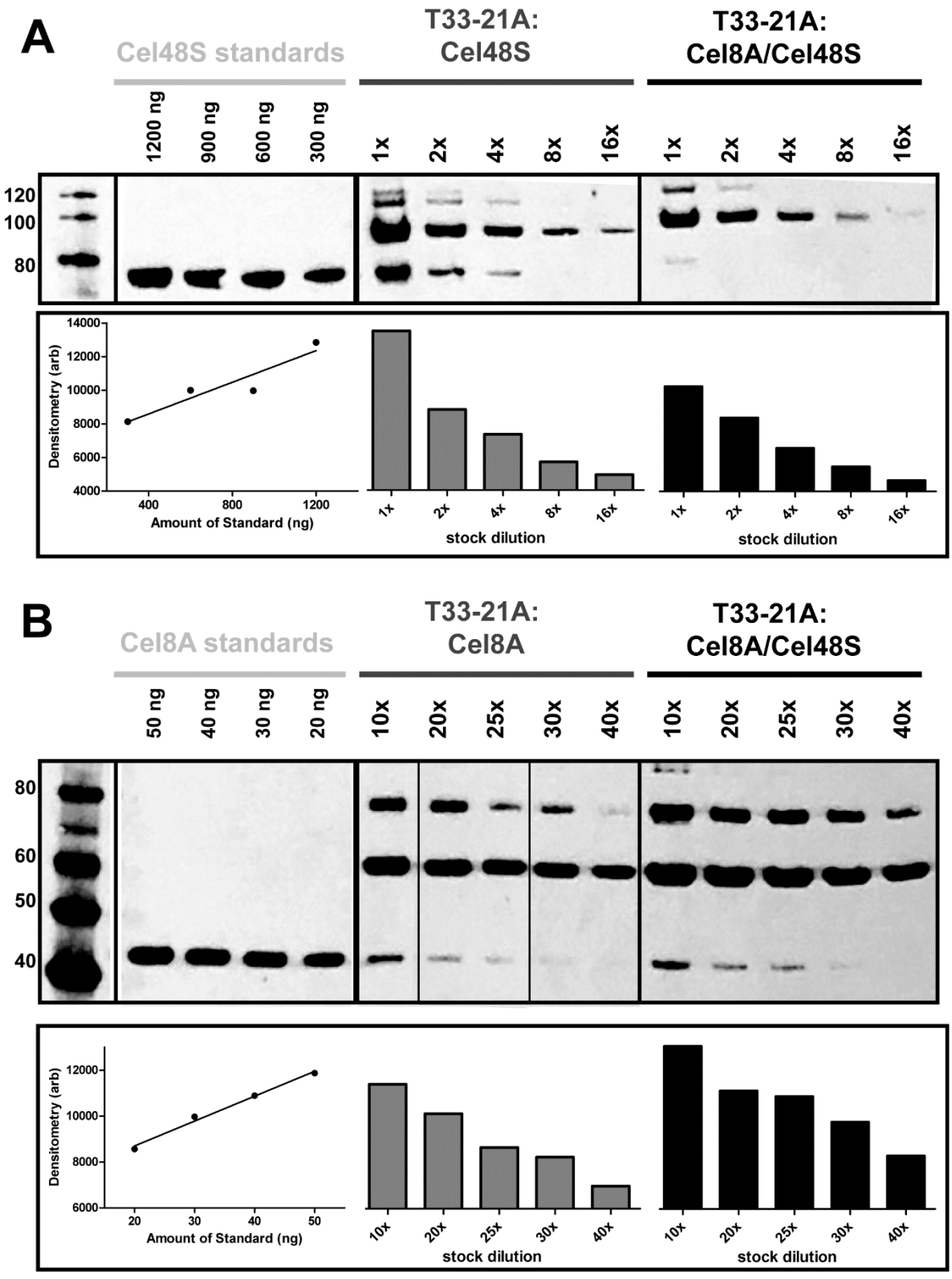

A
Cel48S standards
T33-21A: Cel48S
T33-21A: Cel8A/Cel48S
1200 ng
900 ng
600 ng
300 ng
1x
2x
4x
8x
16x
120
100
80
Densitometry (arb)
Amount of Standard (ng)
stock dilution
B
Cel8A standards
T33-21A: Cel8A
T33-21A: Cel8A/Cel48S
50 ng
40 ng
30 ng
20 ng
10x
20x
25x
30x
40x
80
60
50
40

Figure 4.S2. The presence of cellulase on each purified cellulolytic cage was probed with anti-FLAG for the endoglucanase Cel8A or anti-myc for the exoglucanase Cel48S, and then detected using goat anti-mouse IgG conjugated to horse radish peroxidase. The absolute amount of each cellulase on the cages was determined by comparison of the intensities of known concentrations of the corresponding monomeric cellulases. Serial dilutions of the cages were all in good agreement with the final calculated concentration of each cellulase on the cages. On the immunoblot, at higher concentrations, some cages remained associated into higher order species on the reducing gel. Those high molecular weight bands were summed in the total band intensity.

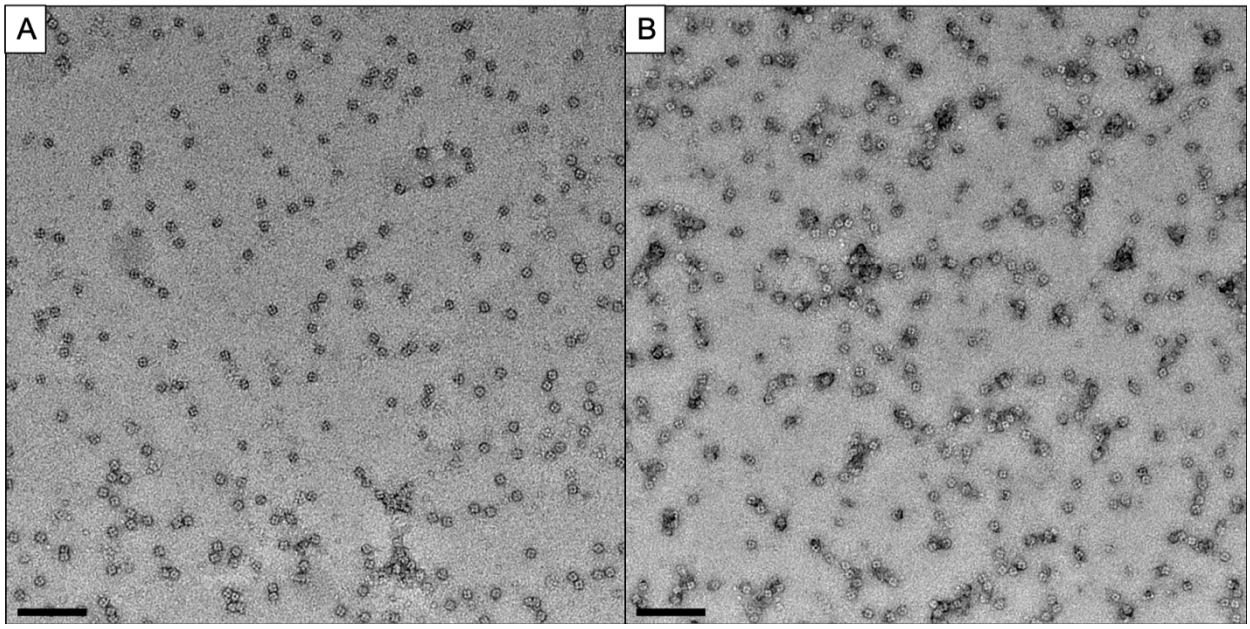


Figure 4.S3. Stability of T33-21-SR cages. Cages were observed by negative stain EM before (A) and after (B) incubation at 37°C for 24 hours to determine if subjecting them to elevated temperatures for an extended period of time under the Avicel degradation assay conditions would affect cage assembly. Properly assembled cages are still readily observed after 24 hours of incubation, suggesting they are adequately thermostable for this application. Scalebar = 100 nm.

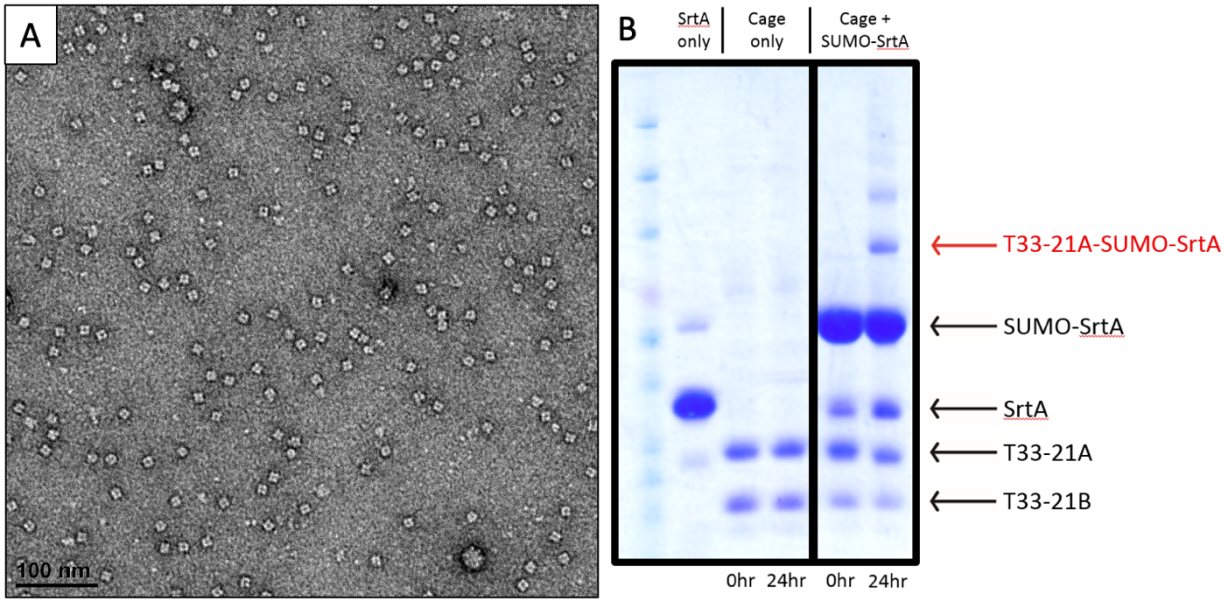


Figure 4.S4. Sortase-mediated fusion of SUMO to the T33-21-SR cage. (A) Cages are observed to properly assemble after SrtA-mediated fusion of SUMO to the T33-21A cage component. (B) SDS-PAGE analysis shows that a band corresponding to the molecular weight of T33-21A plus SUMO-SrtA appears after a 24-hour ligation period.

REFERENCES

1. King, N. P. *et al.* Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336**, 1171–1174 (2012).
2. Lai, Y. T., Cascio, D. & Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129–1129 (2012).
3. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).
4. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein

complexes. *Science* **353**, 389–395 (2016).

5. Aumiller, W. M., Uchida, M. & Douglas, T. Protein cage assembly across multiple length scales. *Chem. Soc. Rev.* **47**, 3433–3469 (2018).
6. Cannon, K. A., Ochoa, J. M. & Yeates, T. O. High-symmetry protein assemblies: patterns and emerging applications. *Curr. Opin. Struct. Biol.* **55**, 77–84 (2019).
7. Marcandalli, J. *et al.* Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17 (2019).
8. Phippen, S. W. *et al.* Multivalent Display of Antifreeze Proteins by Fusion to Self-Assembling Protein Cages Enhances Ice-Binding Activities. *Biochemistry* **55**, 6811–6820 (2016).
9. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *Proc. Natl. Acad. Sci.* **115**, 3362–3367 (2018).
10. Liu, Y., Huynh, D. T. & Yeates, T. O. A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nat. Commun.* **10**, 1864 (2019).
11. Izard, T. *et al.* Principles of quasi-equivalence and Euclidean geometry govern the assembly of cubic and dodecahedral cores of pyruvate dehydrogenase complexes. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 1240–1245 (1999).
12. Zhou, Z. H., McCarthy, D. B., O'Connor, C. M., Reed, L. J. & Stoops, J. K. The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes. *Proc. Natl. Acad. Sci.* **98**, 14802–14807 (2001).
13. Moon, T. S., Dueber, J. E., Shiue, E. & Prather, K. L. J. Use of modular, synthetic scaffolds

for improved production of glucaric acid in engineered E. coli. *Metab. Eng.* **12**, 298–305 (2010).

14. Conrado, R. J. *et al.* DNA-guided assembly of biosynthetic pathways promotes improved catalytic efficiency. *Nucleic Acids Res.* **40**, 1879–1889 (2012).
15. Price, J. V., Chen, L., Whitaker, W. B., Papoutsakis, E. & Chen, W. Scaffoldless engineered enzyme assembly for enhanced methanol utilization. *Proc. Natl. Acad. Sci.* **113**, 12691–12696 (2016).
16. Yang, Z. *et al.* Enhanced itaconic acid production by self-assembly of two biosynthetic enzymes in Escherichia coli. *Biotechnol. Bioeng.* **114**, 457–462 (2017).
17. Wang, Y., Heermann, R. & Jung, K. CipA and CipB as Scaffolds To Organize Proteins into Crystalline Inclusions. *ACS Synth. Biol.* **6**, 826–836 (2017).
18. Wheeldon, I. *et al.* Substrate channelling as an approach to cascade reactions. *Nat. Chem.* **8**, 299–309 (2016).
19. Zhao, X., Zhang, L. & Liu, D. Biomass recalcitrance. Part II: Fundamentals of different pre-treatments to increase the enzymatic digestibility of lignocellulose. *Biofuels Bioprod. Biorefining* **6**, 561–579 (2012).
20. Sethi, A. & Scharf, M. E. Biofuels: Fungal, Bacterial and Insect Degraders of Lignocellulose. in *eLS* (John Wiley & Sons, Ltd, 2013). doi:10.1002/9780470015902.a0020374
21. Hirano, K., Nihei, S., Hasegawa, H., Haruki, M. & Hirano, N. Stoichiometric assembly of the cellulosome generates maximum synergy for the degradation of crystalline cellulose, as revealed by in vitro reconstitution of the Clostridium thermocellum cellulosome. *Appl. Environ. Microbiol.* **81**, 4756–4766 (2015).

22. Hirano, K. *et al.* Enzymatic diversity of the Clostridium thermocellum cellulosome is crucial for the degradation of crystalline cellulose and plant biomass. *Sci. Rep.* **6**, 35709 (2016).
23. Chen, I., Dorr, B. M. & Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. doi:10.1073/pnas.1101046108/-/DCSupplemental
24. Miller, G. L. Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. *Anal. Chem.* **31**, 426–428 (1959).
25. Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annu. Rev. Biochem.* **79**, 655–681 (2010).
26. Proft, T. Sortase-mediated protein ligation: An emerging biotechnology tool for protein modification and immobilisation. *Biotechnol. Lett.* **32**, 1–10 (2009).
27. Kobashigawa, Y., Kumeta, H., Ogura, K. & Inagaki, F. Attachment of an NMR-invisible solubility enhancement tag using a sortase-mediated protein ligation method. *J. Biomol. NMR* **43**, 145–150 (2009).
28. Pritz, S. *et al.* Synthesis of Biologically Active Peptide Nucleic Acid−Peptide Conjugates by Sortase-Mediated Ligation. *J. Org. Chem.* **72**, 3909–3912 (2007).
29. Refaei, M. A. *et al.* Observing selected domains in multi-domain proteins via sortase-mediated ligation and NMR spectroscopy. *J. Biomol. NMR* **49**, 3–7 (2011).
30. Zhang, S., Wolfgang, D. E. & Wilson, D. B. Substrate heterogeneity causes the nonlinear kinetics of insoluble cellulose hydrolysis. *Biotechnol. Bioeng.* **66**, 35–41 (1999).
31. Fierobe, H.-P. *et al.* Degradation of Cellulose Substrates by Cellulosome Chimeras. *J. Biol. Chem.* **277**, 49621–49630 (2002).
32. Fierobe, H. P. *et al.* Design and production of active cellulosome chimeras. Selective

incorporation of dockerin-containing enzymes into defined functional complexes. *J. Biol. Chem.* **276**, 21257–21261 (2001).

33. Fierobe, H. P. *et al.* Action of designer cellulosomes on homogeneous Versus complex substrates: Controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin. *J. Biol. Chem.* **280**, 16325–16334 (2005).
34. Lytle, B., Myers, C., Kruus, K. & Wu, J. H. Interactions of the CelS binding ligand with various receptor domains of the Clostridium thermocellum cellulosomal scaffolding protein, CipA. *J. Bacteriol.* **178**, 1200–3 (1996).
35. Krauss, J., Zverlov, V. V. & Schwarz, W. H. In Vitro Reconstitution of the Complete Clostridium thermocellum Cellulosome and Synergistic Activity on Crystalline Cellulose. *Appl. Environ. Microbiol.* **78**, 4301–4307 (2012).
36. Blanchette, C., Lacayo, C. I., Fischer, N. O., Hwang, M. & Thelen, M. P. Enhanced cellulose degradation using cellulase-nanosphere complexes. *PLoS ONE* **7**, e42116 (2012).
37. Moraïs, S. *et al.* Enhanced cellulose degradation by nano-complexed enzymes: Synergism between a scaffold-linked exoglucanase and a free endoglucanase. *J. Biotechnol.* **147**, 205–211 (2010).
38. Mitsuzawa, S. *et al.* The rosettazyme: A synthetic cellulosome. *J. Biotechnol.* **143**, 139–144 (2009).
39. Carrico, Z. M., Romanini, D. W., Mehl, R. A. & Francis, M. B. Oxidative coupling of peptides to a virus capsid containing unnatural amino acids. *Chem. Commun.* 1205–1207 (2008). doi:10.1039/b717826c
40. Miller, R. A., Presley, A. D. & Francis, M. B. Self-assembling light-harvesting systems from synthetically modified tobacco mosaic virus coat proteins. *J. Am. Chem. Soc.* **129**, 3104–

3109 (2007).

41. Heddle, J. G., Chakraborti, S. & Iwasaki, K. Natural and artificial protein cages: design, structure and therapeutic applications. *Curr. Opin. Struct. Biol.* **43**, 148–155 (2017).
42. Patterson, D. P., Prevelige, P. E. & Douglas, T. Nanoreactors by programmed enzyme encapsulation inside the capsid of the bacteriophage P22. *ACS Nano* **6**, 5000–5009 (2012).
43. Glasgow, J. E., Asensio, M. A., Jakobson, C. M., Francis, M. B. & Tullman-Ercek, D. Influence of Electrostatics on Small Molecule Flux through a Protein Nanoreactor. *ACS Synth. Biol.* **4**, 1011–1019 (2015).
44. Patterson, D. P., Schwarz, B., Waters, R. S., Gedeon, T. & Douglas, T. Encapsulation of an enzyme cascade within the bacteriophage P22 virus-like particle. *ACS Chem. Biol.* **9**, 359–365 (2014).
45. Aanei, I. L. *et al.* Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjug. Chem.* **26**, 1590–1596 (2015).