De Novo Design of Pseudosymmetric Protein Hetero-Oligomers

Ryan D. Kibler

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

David Baker, Chair

Frank Dimaio

Jesse Zalatan

Program Authorized to Offer Degree:

Biochemistry

©Copyright 2022

Ryan D. Kibler

University of Washington

Abstract

De Novo Design of Pseudosymmetric Protein Hetero-Oligomers

Ryan D. Kibler

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

Protein design success is considerably higher now than it has been in the recent past, but the simultaneous design of many complex interactions, such as hetero-oligomers, remains difficult. We developed a strategy for efficiently designing pseudosymmetric hetero-oligomers by completing a longer series of simpler tasks and created specifically assembling hetero-oligomers. The property of psuedosymmetry could also be leveraged to create protein nanocages with T=4 triangulation numbers, which are prevalent in nature but difficult to design. We start by designing T=1 nanocages with tetrahedral, octahedral, and icosahedral symmetries using symmetric homotrimers, then expand the cages to T=4 through pseudosymmetrization of the component trimers and design of additional interfaces. These approaches highlight the benefits of pseudosymmetry and stepwise approaches to protein material design and represent general avenues for the design of higher complexity structures.

Table of Contents

Preamble

I remember wondering when I first started graduate school why anyone would bother with designing symmetric protein complexes? Symmetry sounded too easy – lots of designs assembled like this by accident. You weren't going to solve any real problems by using symmetry. The title of this dissertation is "*De Novo* Design of Pseudosymmetric Protein Hetero-Oligomers", so clearly something changed. It's often illuminating to recall your journey through life and reflect on the decisions you made, so in service to providing context on my dissertation, which is mostly concerned with the design and application of (pseudo)symmetric proteins, I would like to spend some time summarizing the major milestones of my journey.

I rotated in the Baker Lab in the fall of 2017, around the time that Bobby Langan's LOCKR[1] was really taking off and Marc Lajoie and Robin Kirkpatrick were adapting LOCKR to be activated by co-localization[2,3]. The original LOCKR switch was a large 6-helix bundle protein which tended to aggregate if mistreated or stored at 4°C (yes, seriously). My rotation project, supervised by Zibo Chen, was to create smaller, more compact, and better behaved switches. While I did successfully design some new switches, they had high rates of background activation and weren't very useful.

This was my first of three rotations at the University of Washington in the Biological Physics, Structure, and Design (BPSD) program, so while I was rotating in the labs of Frank DiMaio and Georg Seelig[4], Zibo was in the process of developing the CIPHR logic based on his DHDs[5,6]. I joined the Baker Lab officially in the summer of 2018 and began working with Zibo on protein logic, during which I became interested in exploring how to redirect the output of the CIPHR logic gates to have forms other than activation of transcription. The LOCKR system's large array of output types (fluorescence, bioluminescence[7], peptide binding, protein binding,

chemical conjugation, etc) was attractive, and since the direct output of CIPHR logic operation and the input to co-LOCKR were both co-localization events, it seemed like an obvious choice.

After some time, I discovered a problem: the "simplest" CIPHR logic gate, the two-input "AND", failed to activate co-LOCKR. The CIPHR AND gate forms a fairly large, disordered complex in the activated state. The activity of co-LOCKR depends on the lengths of the linkers between the LOCKR and its co-localizer, so I hypothesized that the distance between the switch and key of the activated CIPHR complex was the issue. I attempted to solve this by creating a structurally rigid version of the AND gate by using WORMS[8] to link all the required DHD hetero-dimers together. The structure was intended to position the ends of the activated horseshoe-shaped (or tiara-shaped, hence the path to the directory storing the bulk of my thesis work: `/home/rdkibler/projects/tiara_gen2/`) complex much closer together and increase the local concentration of the switch and key to the point of activation. This was a bad idea, in retrospect, not only because the logic gates require cooperative assembly to work (the "C" in CIPHR is "cooperative") and this setup could not be cooperative, but also because I was greedy and tried to design ~9 fusion junctions at once – a very large number. The tiaras never assembled correctly.

David Baker and I reasoned that, because assembly of a ring-shaped protein complex is a cooperative event, we could re-implement the CIPHR AND gate using a cyclic heter-tetramer where two chains which are too weak to interact by themselves are fused to the switch and key and the other two chains serve as the inputs which, when assembled, bring the two weakly interacting chains close together and enhance their interaction strength. While there were many heterodimers at the time[6] and Sherry Bermo was designing heterotrimers[9], nobody had designed a cyclic hetero-tetramer with four unique chains (ABCD), so I made that my main project.

I knew from my folly with the tiaras that the less which is changed about a working protein, the more likely it is to continue to work. Cyclically symmetric ring-shaped homo-oligomers with the same numbers of chains as the desired hetero-oligomer already assemble into roughly the correct structure, so they would make good starting points. All we would need to do is redesign the interfaces to be orthogonal to one another which is possible by incorporating hydrogen bond networks (HBNets) at the interface[6].

Multiple interfaces designed simultaneously in the same complex each have a chance of failing and ruining the entire assembly, meaning three interfaces could be good and correct but you would never know because of the one that failed. Nick Woodall and I realized that this problem could be mitigated if each new interface was independently validated as a separate symmetric homo-oligomer, which verifies that the redesigned interface is still capable of forming the correctly sized ring, and then the pre-validated individual interfaces could be combined into one structure to arrive at the desired hetero-oligomer. The final hetero-oligomers produced by this process are pseudosymmetric because they retain the appearance of symmetry from their parent, despite having different sequences at symmetrically equivalent positions. This property ends up being the secret weapon of these hetero-oligomers, as it allows the easy reduction of symmetry of structures built out of the symmetric parents via modular interface replacement and enables larger protein complexes to increase in complexity and function in a few simple and manageable steps.

In Chapter 1 of this dissertation, I will cover my work on redesigning two families of homo-oligomers, BGLs (coloquially known as “churros”) and RTRs, using the above technique to make pseudosymmetric ABC, A2B2, A3B3, and A2B2C2 hetero-oligomers. I carried out all of the design except for the N-terminal miniprotein RTR fusions (courtesy of Yang Hsia), all of

the initial purifications, the initial biochemical characterization of the designs, and nsEM. X-ray crystallography was performed by Madison Kennedy and Barry Stoddard at the Fred Hutch, SAXS data were collected by SIBYLS at Lawrence Berkeley National Lab, nMS was performed at the Wysocki Lab at Ohio State University, and SEC-MALS was performed by Lauren Carter at the IPD. Samples for X-ray crystallography were expressed and purified by the IPD's General Protein Production Core under the direction of Cameron Chow. Basile Wicky wrote the code which I used to find the helical parameters for the helix rebuilding designs. During this work, my amazing postdoc, Sangmin Lee, who also hails from the University of Michigan[10,11], joined the lab and assisted me in designing and characterizing "Set 2" of the hetBGLs.

In Chapter 2, I will discuss how Sangmin and I took advantage of the property of pseudosymmetry and its symmetry reduction capabilities to design large T=4 nanocages in a stepwise manner. Erin Yang found that the BGLs didn't have shapes which were amenable to docking, so Yang Hsia and I added designed helical repeat (DHR) arms to the BGLs to diversify their shapes. I characterized the initial set of these with some help from Sangmin, then he, David, and I conceptualized the stepwise design and testing strategy. Sangmin designed, expressed, and fully characterized the T=1 nanocages, intermediate crowns, and T=4 nanocages with my input and occasional assistance. CryoEM of the nanocages was performed by Annika Philomin and Andrew Borst and X-ray crystallography of an armed BGL was performed by Madison Kennedy.

Chapter 1:

Stepwise design of pseudosymmetric protein hetero-oligomers

Abstract

Pseudosymmetric hetero-oligomers with three or more unique subunits with overall structural (but not sequence) symmetry play key roles in biology, and systematic approaches for generating such proteins *de novo* would provide new routes to controlling cell signaling and designing complex protein materials. However, the de novo design of protein hetero-oligomers with three or more unique chains is a challenging problem because it requires the accurate design of multiple protein-protein interfaces simultaneously. Here, we describe a divide-and-conquer approach which breaks the multiple interface design challenge into a series of more tractable symmetric design problems, and then combines the validated interfaces to form pseudosymmetric hetero-oligomers. Starting from *de novo* designed circular homo-oligomers composed of 9 or 24 tandemly repeated units, we redesigned the subunit-subunit interfaces to generate 15 new homo-oligomers and recombined them to make 17 new hetero-oligomers, including ABC heterotrimers, A2B2 hetero-tetramers, and A3B3 and A2B2C2 heterohexamers which assemble with high structural specificity. The symmetric homo-oligomers and pseudo-symmetric hetero-oligomers generated for each system share a common backbone, and hence are ideal building blocks for generating and functionalizing larger symmetric assemblies.

Introduction

Pseudosymmetric oligomers in Nature likely evolved from homo-oligomers composed of single subunits which, following gene duplication, underwent divergence and specialization leading to structures where each subunit is structurally similar but can have different specialized functions[12,13] (Fig. 1.1a). These functions are often concerned with the recruitment of a more specific set of proteins than is possible to assemble on a purely symmetric homo-oligomeric scaffold, as it is with the eukaryotic pseudosymmetric hetero-heptameric and -hexameric complexes of Sm and LSm proteins, which mediate a wide network of RNA-protein interactions via specific combinations of protomers[14,15] and developed from a shared common ancestor with the bacterial homohexamer Hfq[14,16]. Another example, PCNA, acts as a DNA-modifying protein recruitment hub and forms homotrimers in eukaryotes, yet pseudosymmetric heterotrimers in archaea with unique protein binding sites on each subunit and a different assembly pathway[17]. These examples of pseudosymmetric hetero-oligomers which have symmetric homomeric paralogs that perform similar but often simpler functions illustrates one way which Nature increases complexity in a fairly economical way – rather than inventing an entirely new subunit, it makes copies of a gene which is almost working for a new purpose and evolves into a product with greater functions.

As in Nature, pseudosymmetric oligomers can enable increases in complexity in the design of protein assemblies, for example extending de novo nanocages to higher triangulation number assemblies[18,19]. However, while there has been considerable progress with de novo design of protein homo and hetero oligomers[6,9,20–28], with the exception of heterodimers and simple coiled coil heterotrimers, there are no systematic routes to pseudosymmetric structures with nearly identical protomer backbones but different sequences. This is a challenging problem

because, for a pseudosymmetric cyclic n-mer with n distinct chains in the target ground state oligomer, there are n^n-1 off target assemblies (any of the n distinct chains at each of the n positions), all with similar or identical backbone geometry; the specificity for the intended heteromeric complex must therefore come from the amino acid sidechain identities.

We set out to develop a general de novo design approach for creating pseudosymmetric hetero-oligomers (Fig. 1.1b) inspired by the likely evolutionary route from symmetric homo-oligomers to hetero-oligomers (Fig. 1.1a). We focus on architectures in which the interfaces between different subunits are structurally disjoint; in the cyclic case these correspond to circular ring-like assemblies with each subunit forming part of the arc of the ring (see Fig. 1.1c,d). Such architectures have the advantage that each interface can be validated independently in the context of a homo-oligomer that can be experimentally characterized.

We developed a two step strategy (Fig. 1.1b). In the first step, multiple distinct homo-oligomers are generated containing distinct hydrogen bond networks (HBNets) at the subunit-subunit interfaces. Because there is a large energetic penalty disfavoring burial of polar groups, interfaces with extensive hydrogen bond networks have higher specificity than purely hydrophobic interfaces, with reduced off-target interactions[22]. HBNets have been used previously to create a large number of orthogonal interfaces on similarly-shaped helical backbones[6]. We increase the variety of HBNets that can be incorporated by introducing limited backbone diversity in the secondary structure elements at the subunit-subunit interface. In the second step, validated homo-oligomers are recombined in a manner analogous to DNA recombination to generate pesudosymmetric hetero-oligomers. This “divide and conquer” approach avoids the need to successfully design multiple interfaces simultaneously (“one-shot design”) and enables

the generation of a large number of hetero-oligomers through the recombination of a small base set of homo-oligomers (Supplemental Fig. 1.1).

Psuedosymmetrizing a 9 repeat toroid

We chose as a first model system a donut-shaped toroid composed of three identical subunits forming a ring (Fig. 1.1c, tcTRP9_sub3 from ref 29 referred to here as BGL0). Because Rosetta HBNet placement is very sensitive to small backbone displacements[30], we enhanced HBNet sampling without drastically altering the interface by generating small backbone variations at the interface using normal modes (NM) relaxation[31] restricted to the interface, forming the NM design set, or by complete backbone resampling of the C-terminal helix, which is on the outer ring of helices and forms the HBNet interface in conjunction with two inner ring helices. In the latter resampling, we varied both the internal geometry of the helix using the Crick alpha-helix parameters[32] starting from the closest fit parameters to the C-terminal helix and varying r_0, ω_0, ω_1, φ_0, φ_1, and Δz, and then rejoined the helix to the C-terminus (HR-C design set) or, alternatively, attached it anew to the N-terminus (HR-N design set). The HR-N design set is expected to be orthogonal to the HR-C and NM designs due to steric hindrance imposed by mismatching terminal helices (clashing in one case, a large gap with reduced hydrophobic contacts in the other).

We obtained synthetic genes encoding 20 redesigns of BGL0 (named BGL01-20, see Supplemental Table 1.1) and produced the proteins in E *coli*. Out of the 20 tested, 10 (NM: 3 of 4; HR-N: 3 of 8; HR-C: 4 of 8) were trimeric according to size exclusion chromatography (SEC) (Fig. 1.2b, Supplemental Fig. 1.2), native mass spectrometry (nMS)[33] (Fig. 1.2c, Supplemental Fig. 1.3), and SEC-Multi-Angle Light Scattering (SEC-MALS) (Supplemental Fig. 1.4) and had the correct shape by small-angle X-ray scattering (SAXS)[34,35] (Fig. 1.2d, Supplemental Fig. 1.5)

and negative-stain electron microscopy (nsEM) (Fig. 1.2e); the most common failure mode being polydisperse or aggregated SEC traces (Supplemental Fig. 1.2). These 10 designs have SAXS profiles closely consistent with the design models ($\chi^2 < 3.8$, Supplemental Fig. 1.5) and nsEM reveals the expected ring-like structures, confirming their structures at low resolution, an important prerequisite to recombination which ensures that the structures are similar enough to merge.

We obtained crystal structures of BGL06 from the HR-C set (Fig. 1.3a), BGL14 and BGL15 from the HR-N set (Fig. 1.3b,c), and BGL18 from the NM set (Fig. 1.3d). On the backbone level, the crystal structures show good agreement with the design models (1.2 Å, 1.3 Å, 0.7 Å, and 1.6 Å CA-RMSD, respectively) and to the parental BGL0 crystal structure (TM-scores 0.94, 0.82, 0.81, and 0.97, respectively). Sidechains at the interfaces of BGL06, BGL14, and BGL18 were well enough resolved to determine that HBNets in all three structures formed correctly. BGL06 showed three copies of its interchain bidentate HBNet involving Gln14 and Gln146, each backed up with second shell interactions from Ser18 and Ser125, respectively (Fig. 1.3a). The string of hydrogen bonds in BGL14's interface alternates between chains with a fully satisfied Asn13 in the very middle which donates a hydrogen bond to His153. BGL18 has two copies of the ring in the unit cell and all 6 interfaces are resolved and highly similar to each other. It has a generally polar interface with a meandering HBNet spanning Asn10, Thr17, Asn126 (which may be donating a hydrogen bond to the backbone oxygen of Asn10), Thr152, Ser155, and Ser159.

We construct pseudosymmetric oligomers from the 11 independently validated homotrimers (the 10 redesigns + the original) in a computational manner analogous to DNA homologous recombination. We superimpose the subunits of two parental homo-oligomers over

their conserved regions and select one interface from each of the parents to form a hybrid of the two parents (the recombination point relative to the BGL0 scaffold is between Arg82 and Gly83 in the loop closest to the middle of the subunit). The correct combinations of hybrid chains should form a specific hetero-oligomer. By using only interfaces which are known experimentally to form into the expected shape with the intended geometry, the probabilities of the interfaces working at all are decoupled from the probabilities that the interfaces are compatible with each other, thus likely increasing the overall success rate. The number of possible combinations of interfaces into hetero-n-mers in the order needed for assembly grows very quickly with the number of validated homo-n-mers (Supplemental Fig. 1.1). There are:

$$\frac{k!\,(n-1)!}{n!\,(k-n)!}$$

ways to combine *k* interfaces into a hetero-n-mer with n different interfaces. With k=11 interfaces to choose from for a heterotrimer (n=3), there are 330 possible different heterotrimers which could be assembled. It is worth noting that three interfaces can be combined in two different orders to yield two unique designs which could each behave differently.

We tested two different strategies for choosing interfaces to recombine. First, in order to attempt to maximize binding specificity, we constructed a set of 6 putative heterotrimers (called "Set 1") where each design includes one interface from each of HR-N, HR-C, and NM to attempt to maximize geometric specificity. Heterotrimers in this set appear globally symmetric but each subunit has different numbers of helices due to the combination HR-N with members of the other two sets, resulting in rings where one subunit receives two terminal outer helices and one receives none. We also constructed heterotrimers with "pure" pseudosymmetry where each subunit has the same backbone shape and rely completely on sidechain-level specificity from HBNets. The best candidate interfaces for this set come from the original design BGL0 and

validated designs of the NM set (BGL17, BGL18, and BGL19). With only 4 interfaces to choose from we could enumerate all 8 possible combinations (called "Set 2").

Both sets of designed heterotrimers (total 14, see Supplemental Table 1.3) were ordered as tricistronic constructs with TEVp-cleavable GFP or miniprotein[36] tags on two of the chains to give the chains significantly different masses for analysis by SDS-polyacrylamide gel electrophoresis (PAGE) and nMS, with one chain having a His6-tag for affinity co-purification (Fig. 1.4a). We found that 11 copurified as three bands, 9 of which at roughly stoichiometric ratios (Fig. 1.4b, Supplemental Fig. 1.6). The SEC profiles of 8 of the 14 designs showed mostly monodisperse peaks around the expected retention volume (Fig. 1.4c, Supplemental Fig. 1.7). Similarly sized non-specific combinations may look similar by SEC so we used nMS to unambiguously distinguish the correct "ABC" heterotrimer. The 8 designs with clean SEC profiles were found to be the intended ABC species without any major off-target contaminants (Fig. 1.4d,f, Supplemental Fig. 1.8). In addition, low resolution structural analysis via nsEM and SAXS of a selection of samples confirmed that the heterotrimers retain their original toroid shape (Supplemental Fig. 1.9, Supplemental Fig. 1.10).

From the 6 designs from Set 1, hetBGL03-15-18 and hetBGL08-09-19 were experimentally confirmed to form heterotrimers, along with 6 of 8 from Set 2 (Fig. 1.4f, Supplemental Fig. 1.8). The two failures from Set 2, hetBGL00-19-18 and hetBGL17-19-18, which pulled down incorrect stoichiometries of bands by SDS-PAGE (Supplemental Fig. 1.6) and had polydisperse traces on SEC (Supplemental Fig. 1.7), are the only two designs which have a subunit containing the N-terminal half of BGL19 and the C-terminal half of BGL18, suggesting that this is a combination of interfaces that may interfere with each other.

The Set 1 heterotrimers are topologically non-pseudosymmetric due to their uneven number of helices in each protomer. We explored whether we could convert the Set 1 heterotrimers into a more pseudosymmetric form by deleting the loop of the N-terminal outer helix and building a new loop to connect it as a C-terminal outer helix (essentially converting the HR-N interface to HR-C). We tested such rearrangements of hetBGL03-15-18, hetBGL03-18-15, and hetBGL08-09-19 and found that hetBGL03-15-18_reloop and hetBGL08-09-19_reloop remained as clean heterotrimers (Supplemental Fig. 1.8), bringing the number of validated hetBGLs to 10.

We succeeded in solving the crystal hetBGL03-15-18, a heterotrimer from Set 1 composed of the interfaces extracted from BGL03, BGL15, and BGL18, at 2.1 Å resolution (Fig. 1.4e). The structure superimposes with the design model with 1.2Å CA-RMSD. We used US-align[37] to assess the degree of pseudosymmetry compared to the crystal structure of BGL0 (PDB: 6XR2) and found that, at a sequence identity of 62.3%, this design achieved a high TM-score of 0.89 with an aligned CA-RMSD of 1.45 Å, showing that this method is capable of producing highly (pseudo)symmetric hetero-oligomers. This design re-used the interfaces of the previously crystallized BGL15 and BLG18. The BGL18 HBNet is in the same conformation as observed in the homotrimer crystal structure, except the amide nitrogen of Asn126 (homotrimer numbering) is no longer positioned to donate a hydrogen bond to the backbone oxygen of Asn10. The HBNet of BGL15, which was not resolved in the homotrimer crystal structure, is confirmed to be in the correct confirmation, with the exceptions of an added hydrogen bond between Ser140 and a crystallographic water and there being two solutions for the sidechain of His16. The structure of the BGL03 interface was solved for the first time in this structure and found to be mostly incorrect because of fraying of the C-terminal outer helix, which was intended to

protect the HBNet from solvent, and subsequent exposure of the HBNet. This may have happened because two serines on the outer helix were within 7 aa of the end and thus there was little hydrophobic clamping to hold the terminus close to the interface. This was avoided in future designs. Remarkably, this disorder did not seem to impact the specificity of the assembly.

Psuedosymmetrizing a 24 repeat toroid

We next explored whether our hierarchical pseudo-symmetrization method was sufficiently general to be applied to another homo-oligomeric system. We selected a designed tetrameric toroid with 24 internal repeats, divided to have 6 repeats within each protomer (Fig. 1.1d, C4_1 from ref 38 referred to here as RTR0). Because of the multiple ways of factoring the 24 repeats into protomers, this system can be readily redesigned to generate C3, C4, and C6 assemblies (Supplemental Fig. 1.11a), the most useful oligomeric states for building larger symmetric materials. A hetero-oligomeric version of RTR0 could potentially be modularly convertable into a diverse set of oligomeric structures (trimers: ABC; tetramers: A2B2, ABCD; hexamers: A3B3, A2B2C2, ABCDEF, etc). Based on the success of the NM set of the BGL family, we only used NM perturbation with the RTR family. HBNets and the surrounding sequences at the interface were designed as for BGL0, but we took an additional step to install disulfide bonds at diverse positions[39], since interface disulfides were required for the original toroid to form tetramers and above[40]; the disulfides could also provide an additional level of specificity (Fig. 1.5a).

We tested 24 redesigns of RTR0, called RTR01 through RTR24, as homotetramers (Supplemental Table 1.4). For 10 of the 24, SEC and nMS suggested single homogenous species with the target subunit composition (Supplemental Fig. 1.12, Supplemental Table 1.7), and for 5 of these, SAXS spectra were consistent with the design model (Supplemental Fig. 1.14) (Fig.

1.5b-d). The interface cysteines of RTR11 and RTR24 were found not to be required for assembly so their cysteins were mutated to alanines in some later designs involving these interfaces to reduce the risk of disulfide-locked off-target species. To facilitate characterization, designs were tagged with designed helical repeat protein (DHR) extensions which protrude from the C-terminus of each subunit (Fig. 1.1d). These features are easily identified by nsEM and are used to further confirm the shapes and subunit stoichiometries for these 5 designs (Fig. 1.5e, Supplemental Figure 1.15). The 2D class averages of RTR0 show a nearly perfectly circular ring with four studs spaced equally around the circle, consistent with the view of the ring through the symmetry axis. In contrast, all of the redesigns show some degree of structural imperfection, with distorted rings that appear to be pinched into oval shapes or have bulges or sharp corners at one interface. Additionally, RTR05 contains a high proportion of pentamers with larger rings and five studs (Fig. 1.5e, blue/right), which was not apparent by nMS.

We assessed the pairwise interface compatibility of the 5 redesigns by building the full set of 10 simple C2-symmetric A2B2 hetero-tetramers via the same recombination approach used for the BGLs, using Gly97 and Leu98 as the cut point. The compositions of these structurally pseudosymmetric A2B2-hetRTRs would be indistinguishable by nsEM so we extended the existing DHR arms or attached DHR-DHR junctions with diverse shapes[41] to the C-terminus or used HFuse[8] to attach miniproteins (Credit: Brian Coventry) into the lumen on the N-terminus. We characterized these first as homo-oligomeric variants and found them to be sufficiently distinguishable (Supplemental Fig. 1.17), so we designed the A2B2 hetero-tetramers to have each chain display a unique C-terminal extension and bear either GFP or a histag. Out of the 10 designs (Supplemental Table 1.5), many had four peaks on SEC, likely corresponding to aggregate, tetramer, dimer, and monomer, and the dimer and monomer peaks were off-target

his-tagged chains according to lack of GFP absorbance (Supplemental Fig. 1.18a). Performing nMS on the tetramer peaks showed that four designs, A2B2-hetRTR05-11, A2B2-hetRTR05-11$_{noCys}$, A2B2-hetRTR05-18$_{noCys}$, and A2B2-hetRTR18$_{noCys}$-24$_{noCys}$, were specific A2B2s (Supplemental Table 1.7), and this was confirmed by nsEM for A2B2-hetRTR05-11 where the distinct arm extensions can be seen with C2 symmetry (Fig. 1.5 upper left). Notably, A2B2 hetero-tetramers containing the RTR05 interface, which as a homo-oligomer partially forms pentamers, do not contain any off-target pentameric species, which is consistent with the observation that the heteromeric assemblies of LSm proteins have reduced variation in ring size compared to related homomeric assemblies[16]. We reasoned that the A2B2 complex should assemble successfully as long as the final assembly contained 24 total repeats in the ring, so we modified the protomers to have one more or one fewer ring repeat. 3 out of 6 tested (Supplemental Table 1.5) retained assembly (Supplemental Figure 1.18b), and $A_5$2$B_7$2-hetRTR05-11$_{noCys}$, which we validated using nsEM, shows the studs shifted by the expected amount (Fig. 1.5 upper right).

To take this a step further, by varying the number of repeats in the protomer, we were able to design homomeric C2, C3, C6, C8, and C12 versions of the RTR toroid (Supplemental Table 1.4). The C2 and C3 versions assembled as desired (Supplemental Table 1.7), but the C6 complexes and above often assembled into a range of complexes with different numbers of subunits (Supplemental Fig. 1.11). We reasoned that the heteromeric versions of these assemblies would be able to assemble with the correct stoichiometries, so we created hetero-hexameric rings via our computational recombination approach, which combine either two different interfaces into C3-symmetric A3B3 hexamers or three interfaces into C2-symmetric A2B2C2 hexamers, and we focused on using the interfaces found to be successful in the A2B2 designs. The same

hybrid protomers could be used in either A3B3 or A2B2C2 hexamers as the combination of interfaces programs the final assembly. When expressed separately and mixed (Supplemental Table 1.6), 3 out of 12 tested A3B3 combinations assembled and 2 out of 4 tested A2B2C2 combinations assembled according to SEC (Supplemental Figure 1.19). A3B3-hetRTR05-11$_{noCys}$ was verified by nMS (Supplemental Table 1.7) and the C3 symmetric particle was observed by nsEM (Fig. 1.5f lower left). While the yield of the A2B2C2 mixtures was too low to confirm assembly by nMS (Supplemental Figure 1.19), we were able to confirm one assembly by nsEM where C2 symmetric classes can be found (Fig. 1.5f lower right). In either case, the samples taken from the assembly peak of SEC contained no variation in ring size.

Discussion

De novo hetero-oligomer design is a challenging endeavour when more than one interface has to be designed simultaneously, and previously required testing hundreds of designs[6,27]. By enabling evaluation of each interface separately, our hierarchical homologous recombination-inspired design strategy yielded 17 experimentally confirmed hetero-oligomers in testing of only 49 designs in total. The BGL family was particularly amenable to interface redesign, and we were able to generate 10 new robust and specific ABC heterotrimers (of 17 tested) which considerably expand the possibilities for pseudosymmetric nanomaterials (next paragraph). The RTR family yielded 4 (of 10) A2B2s, of which the combination of interfaces utilizing RTR05 and RTR11$_{noCys}$ were particularly robust and were used as the basis for the fully validated A3B3 (of 12 tested) and the A2B2C2 (of 4 tested). The structural versatility of the RTR family, with their ability to independently and modularly change the stoichiometries, chain identities, shapes, and spacings of their DHR extensions/termini, will provide diverse hubs for domain clustering in cell signaling and other applications[21,42]. The progenitor of this family has been previously used

as a component of designed molecular machines[38], and the RTRs should be readily incorporated into more complex molecular machines. In this work, we applied our pseudosymmetrization strategy to de novo designed assemblies to make pseudosymmetric hetero-oligomers with three unique chains, but it would be straightforward to generalize the approach to natural homo-meric proteins of interest, such as TNF ligands which might induce novel signalling responses[43], or to create hetero-oligomers with four or more unique subunits.

Pseudosymmetry provides a means to easily transition between simple homo-oligomeric systems and more complicated hetero-oligomeric systems. Fusions or other modifications to hetero-oligomers can first be prototyped using their homo-oligomeric paralogs which are often easier and cheaper to produce, as demonstrated by the diversification of the RTR arms by fusion of DHR junctions (Supplemental Fig. 1.17). Homo-oligomers with pseudosymmetric variants can similarly be used for material design by initially building higher order symmetric assemblies out of homo-oligomers and, as a second step, breaking symmetry by substituting the homo-oligomers for their pseudosymmetric hetero-oligomeric variant,s which enables the customization of some interactions while preserving interaction geometry between the pseudosymmetric building blocks. To illustrate this ability, we have used the BGL homotrimers to design simple tetrahedral, octahedral, and icosahedral nanocages which we then transformed into more complex T=4 nanocages by substituting hetBGL trimers in place to desymmetrize the nanocage, preserving the interface geometries of two of the interfaces and modifying the third to interact with a new homotrimer which increases the T number and physical size of the nanocage[19].

Figures

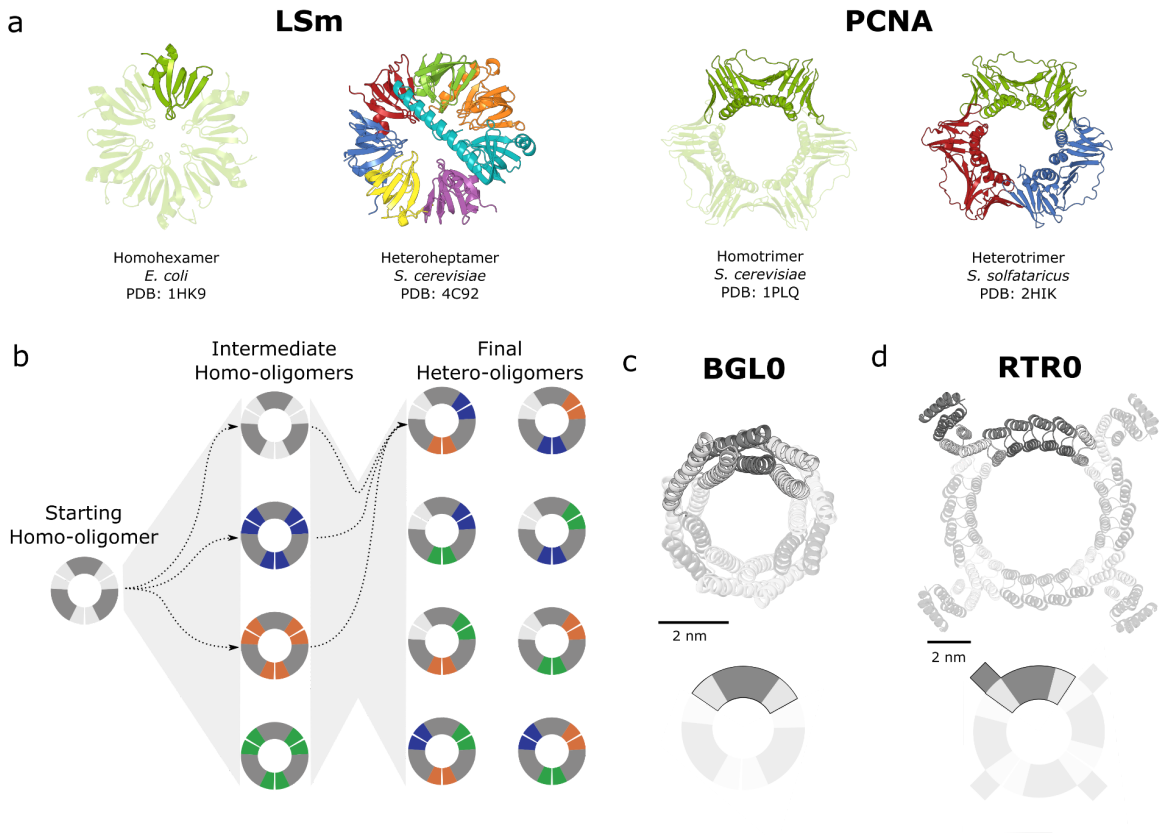


Figure 1.1: Overview of stepwise hetero-oligomer design.

(a) Two examples of natural pseudosymmetric hetero-oligomers (each pair, right) which likely arose from symmetric ancestors, resembling the paired symmetric paralogs (each pair, left), through gene duplication and diversification. Throughout this figure, chains in each complex which have identical sequences are shown as the same color with duplicates faded to white and chains with different sequences are shown in different colors. Left, LSm complexes from *Escherichia coli* (Hfq) and *Saccharomyces cerevisiae*. Right, PCNA from *S. cerevisiae* and *Sulfolbus solfataricus*. (b) Diagram of a two-step parallel testing strategy for efficiently validating new interfaces as homo-oligomers before assembling them into hetero-oligomers.

Dashed lines show an example path taken through the diagram to result in a hetero-oligomeric design. An existing homo-oligomer (left; a homotrimer in this example) is selected and redesigned symmetrically. The "homology" regions in dark gray are not changed during design. Different interfaces are represented as different colors. Homo-oligomers that assemble correctly (middle) are superimposed over their shared homology regions and different interfaces are combined onto the same backbone to make heterotrimers (right). It is possible to create two different heterotrimers with the same interfaces arranged differently. (c,d) (top) The two test scaffolds used in this work, BGL0 (c) and RTR0 (d). One chain in each is highlighted with the interfaces colored light gray and the approximate location of the homology region is shown in dark gray. (bottom) The schematic cartoons used to represent the respective scaffolds in other figures.

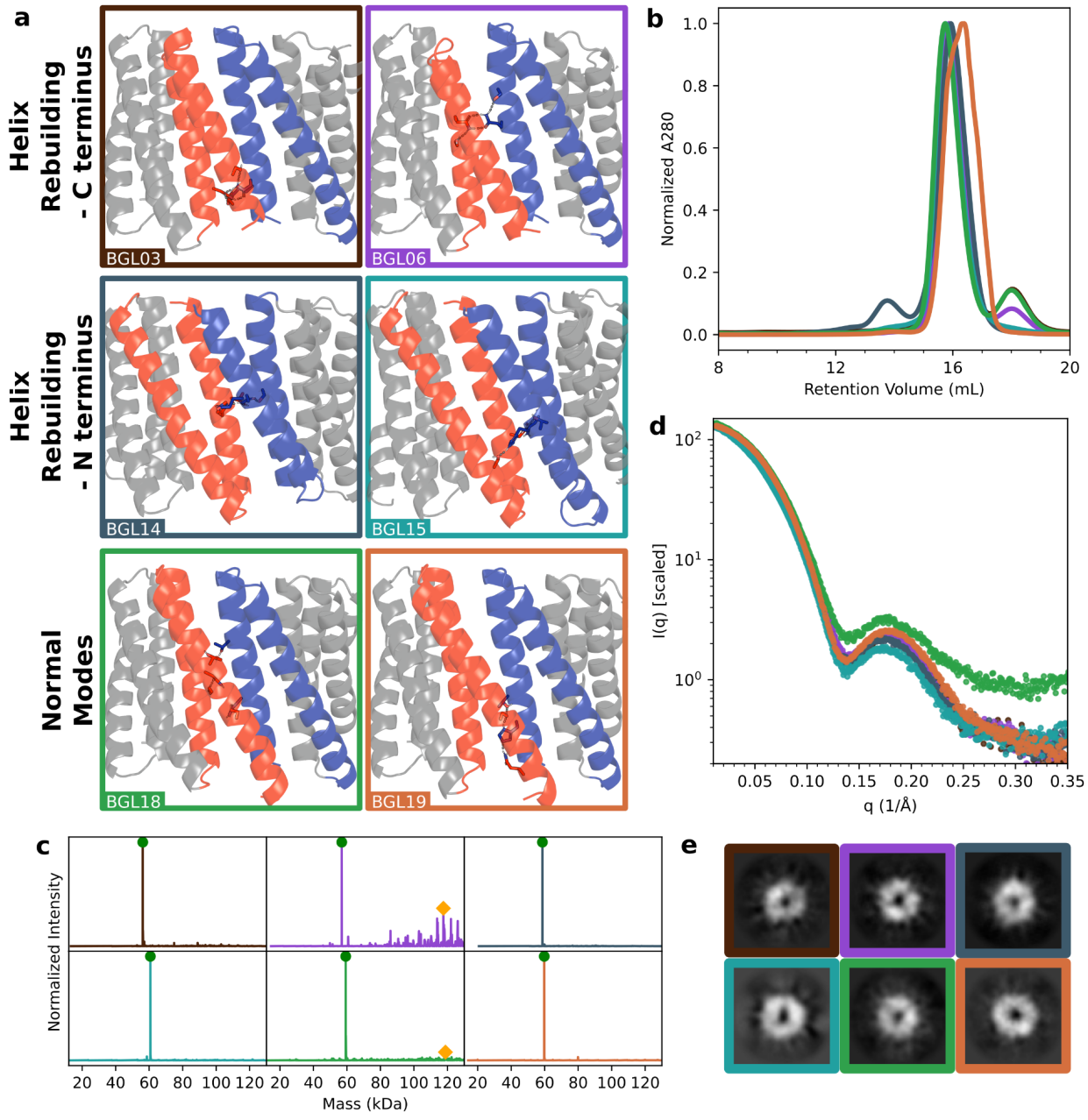


Figure 1.2: Characterization of BGL homotrimers

(a) Cartoon-style design models of 6 representative successful redesigns of BGL0, with two from each design set: Helix Rebuilding + C terminal attachment (HR-C), Helix Rebuilding + N terminal attachment (HR-N) and Normal Modes (NM) relaxation. Blue and red colored

backbones highlight N- and C-terminal two-helix segments comprising the interface which was redesigned, and newly designed hydrogen bond networks (HBNets) are shown in sticks with the hydrogen bonds in dashed lines. Gray backbones are supporting structures which were not greatly modified. Design names in (bottom left corner of each box) and box outline colors are used consistently throughout. Characterization of the six examples from (a) by SEC (b), nMS (c), SAXS (d) and nsEM (e). (b) Overlay of A280 normalized SEC traces. Samples were previously purified by SEC to remove any soluble aggregate. SEC traces of all 20 designs are provided in Supplemental Figure 1.2. (c) Individual mass deconvoluted nMS data of cleaved samples with all peaks that match the expected mass for each homotrimer are annotated with green circles and homohexamers (likely to be trimer-dimers) are annotated with orange diamonds. Trace color corresponds to sample ID. See Supplemental Figure 1.3 for complete data, including m/z spectra. (d) Overlay of SAXS data of TEV-cleaved samples. Data were scaled appropriately to align at q=0.01 for the purpose of displaying similarity of profile. Radii of gyration calculated from the SAXS profiles match the design models within 5.7Å. See Supplemental Figure 1.5. (e) Individual 2D class averages of nsEM data collected on each sample. Outline color corresponds to sample ID.

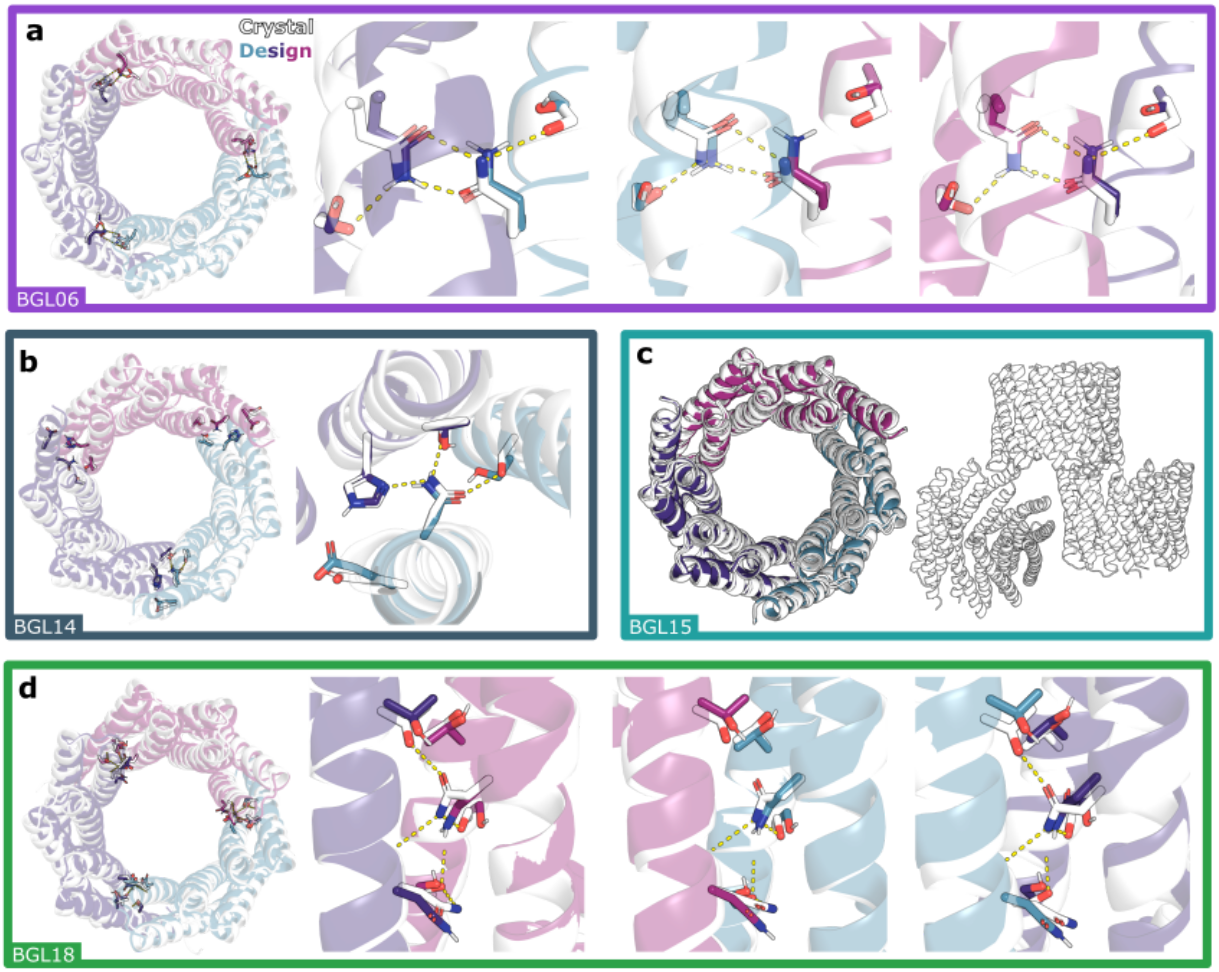


Figure 1.3: Crystal structures of BGL homotrimers

(a,b,d) Crystal structures (white) of four different BGL homotrimers are superimposed onto the design models which are colored in teal, purple, and magenta. (left) The top-down view of the full ring with HBNets highlighted and visible through transparent cartoon models. (Right) close up view(s) of each HBNet which is well enough resolved to build full side chains. (a) BGL06 (2.1 Å resolution; RMSD = 1.2 Å over all Ca atoms; PDB: 8E0L) and (d) BGL18 (3.0 Å resolution; RMSD = 1.6/1.4 Å over all Ca atoms; PDB: 8E0N) were both well enough resolved to model all three interfaces. BGL18 contained two copies of the full trimer in the crystal unit cell, which are structurally similar to each other (RMSD = 0.3 Å over all Ca atoms between

copies). The interfaces of (b) BGL14 (3.0 Å resolution; RMSD = 1.3 Å over all Ca atoms; PDB: 8E12) were not as well resolved and a key sidechain Asn14 could only be modeled into the one which is shown (right). (c) BGL15 (PDB: 8E0M) only has enough resolution to model the backbone (4.0 Å), which is shown overlaid with the design model on the left. (right) All three copies of the trimer in the asymmetric unit. See Supplemental Table 1.2 for crystallographic details.

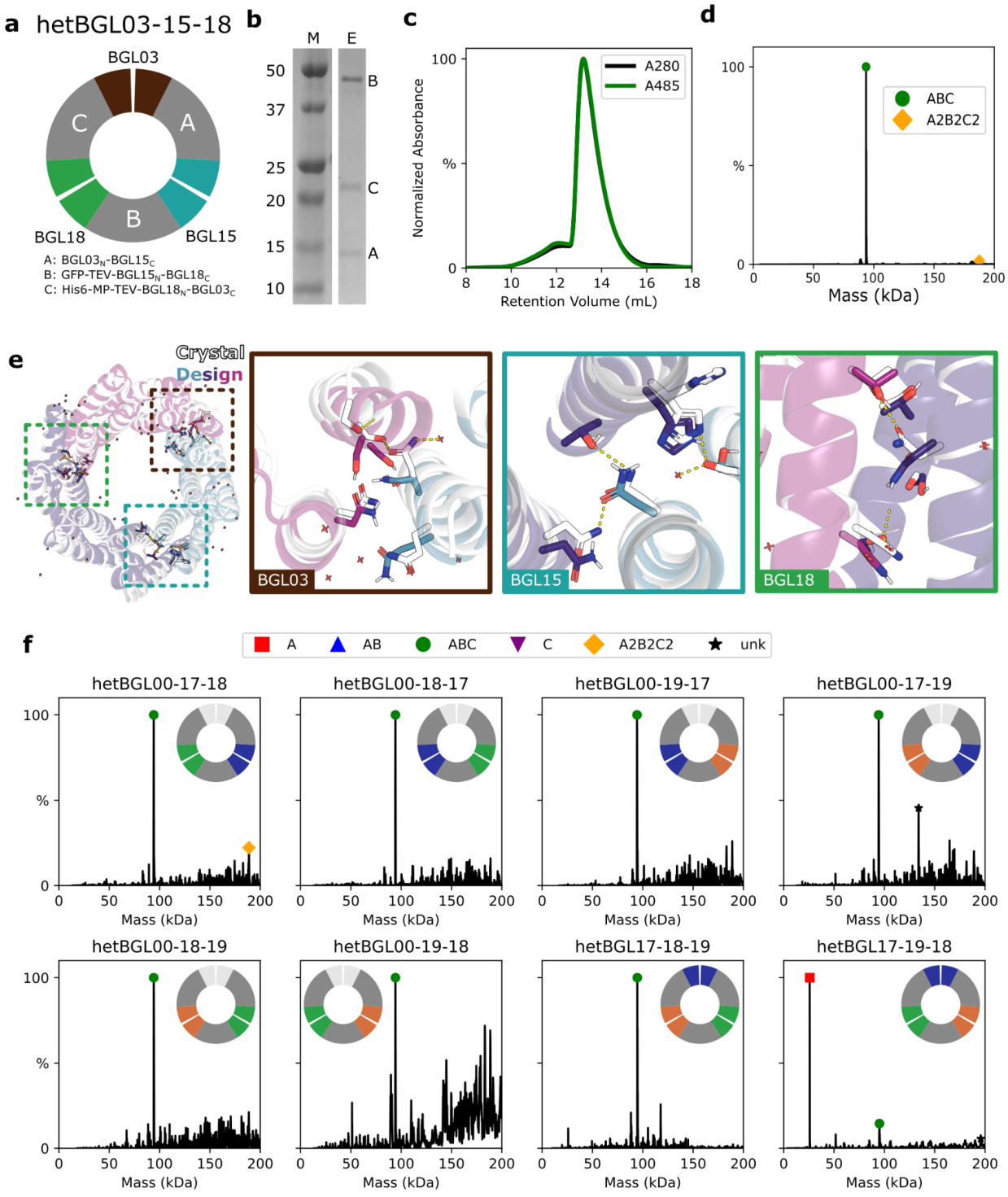


Figure 1.4: Characterization of hetBGL heterotrimers

(a-e) Results of hetBGL03-15-18. (a) (top) Cartoon schematic showing the result of recombination between BGL03, BGL15, and BGL18. (bottom) The expression constructs for each chain as expressed tricistronically. GFP is superfolderGFP, included to add a large amount of mass and provide an additional spectroscopic identification, and "MP" is EHEE_rd2_0005[36], a stable and small protein included to add a small amount of mass. (b) SDS-PAGE of IMAC elution. M: dual xtra protein ladder. E: IMAC elution. Protein size affects band staining, so it is difficult to judge stoichiometry from the band darkness/size, but the correct trend in stain density is observed for equal stoichiometry. (c) SEC trace overlaying normalized absorbances at A280 and A485. (d) Mass deconvoluted nMS data showing the intended major species (green circle, ABC) and a minor species with twice the mass, likely to be trimer-dimer (orange diamond, A2B2C2). See Supplemental Figure 1.8 for complete data, including m/z spectra. (e) (left) Crystal structure in white (2.1 Å resolution, PDB: 8E0O) and design model in teal, purple, and magenta overlaid to show global agreement (CA-RMAD: 1.2 Å, TM-score: 0.96). Dashed boxes indicate approximate locations of zoomed-in views of the interfaces derived from BGL03, BGL15, and BGL18, showing sidechains of HBNet residues and their inferred hydrogen bonds as yellow dashes. (f) Individual mass deconvoluted nMS data for all hetBGL Set 2 samples. See Supplemental Figure 1.8 for complete data, including m/z spectra.

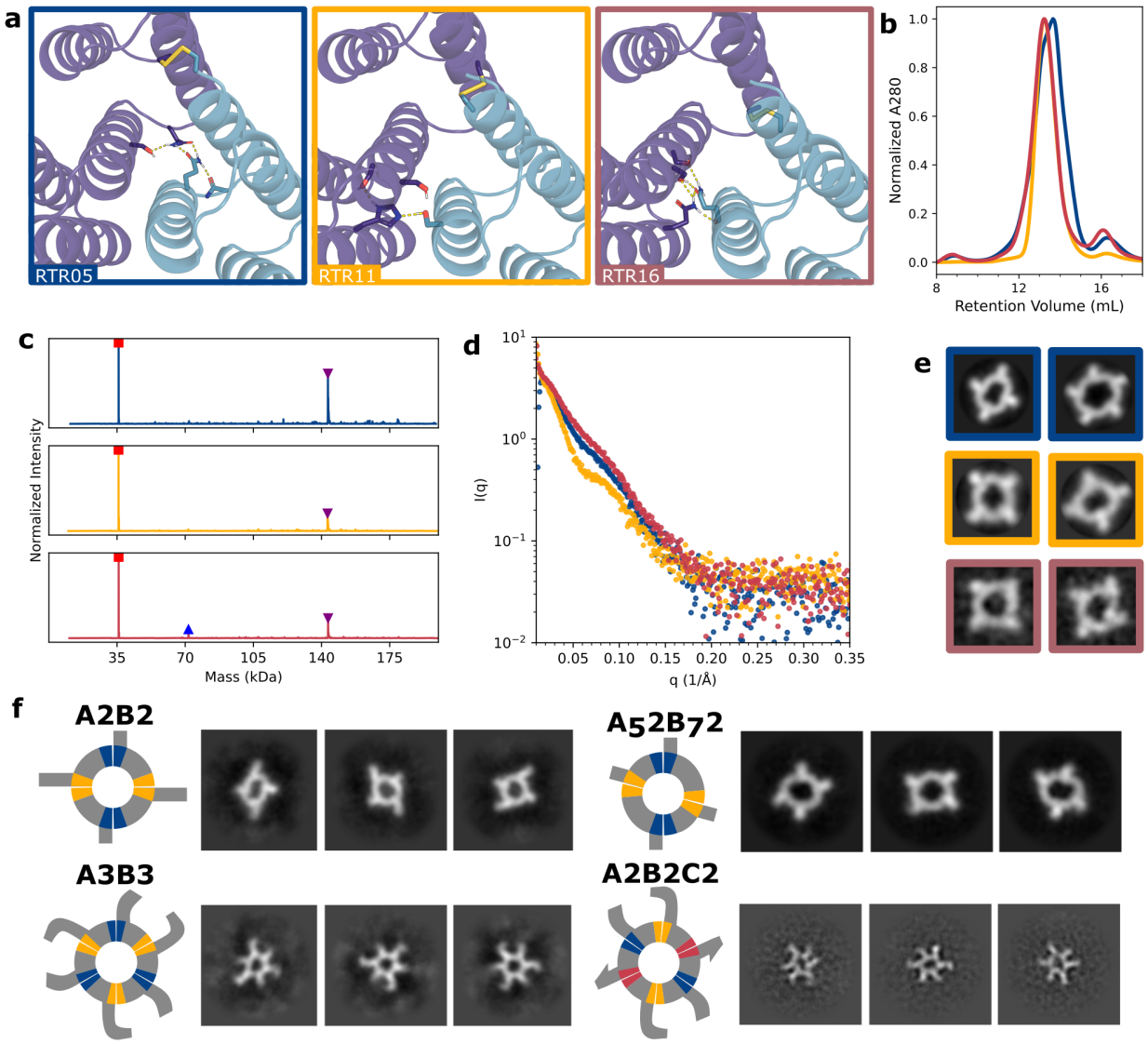
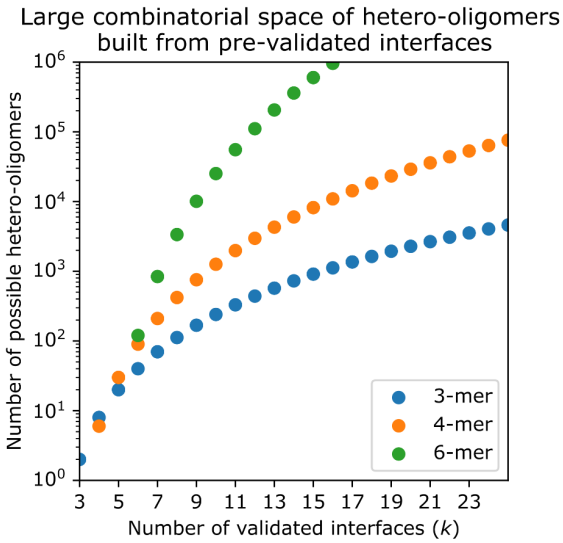


Figure 1.5: Characterization of RTR homotetramers and hetero-oligomers

(a) Cartoon-style design models of 3 representative successful redesigns of RTR0. Blue and purple colored backbones represent N- and C-terminal interfaces, respectively. HBNets and disulfide bonds are shown in sticks with hydrogen bonds in dashed lines. Design names in (bottom left corner of each box) and box outline colors are used consistently throughout to represent the same interfaces. Characterization of the 3 examples from (a) by SEC (b), SAXS (c),

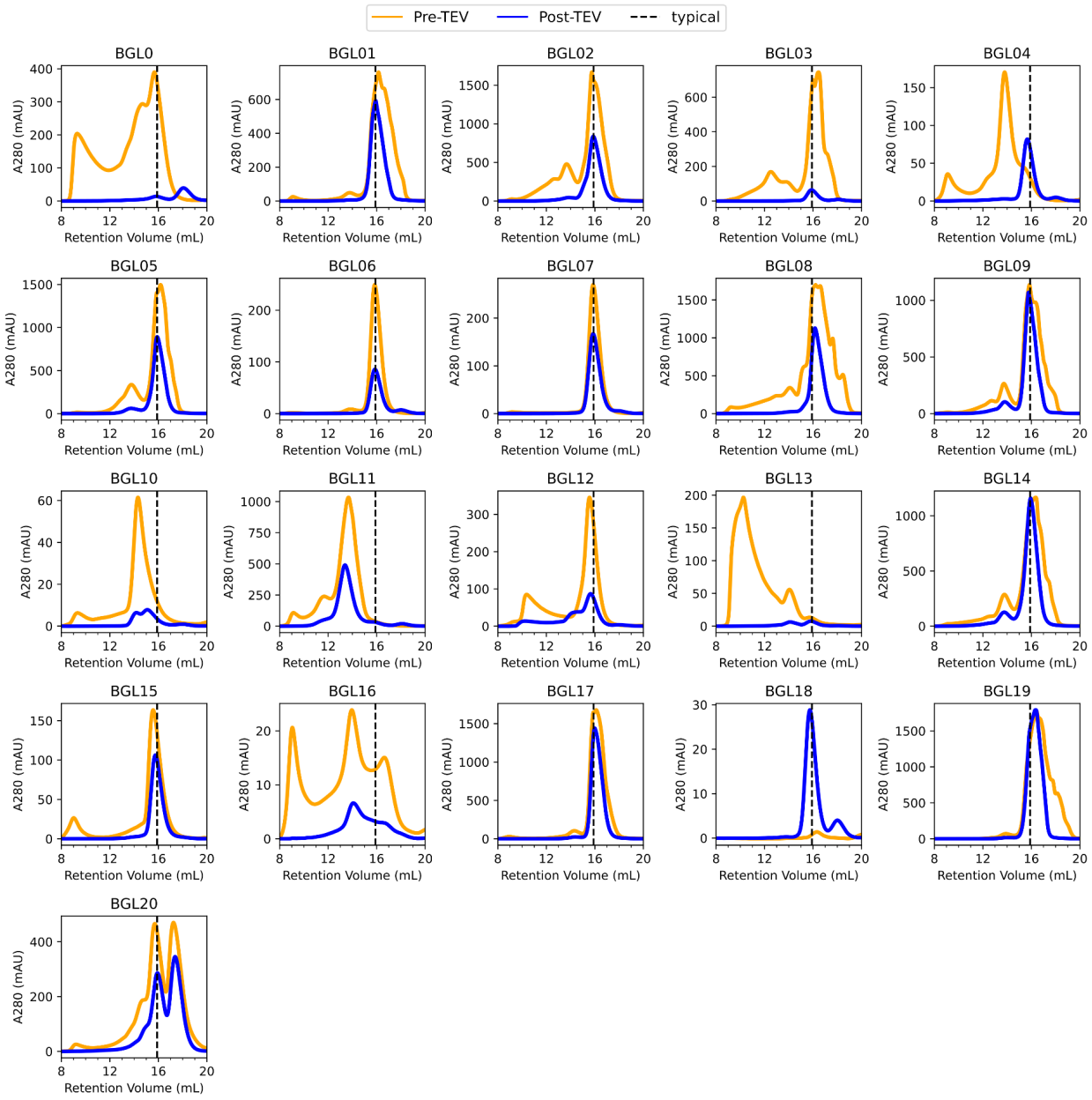
nsEM (d), and nMS (e). (b) Overlay of A280 normalized SEC traces on an S200 column. SEC traces of all 24 designs are provided in Supplemental Figure 1.12. (c) Individual mass deconvoluted nMS data. Peaks corresponding to monomer are annotated with red squares, dimer are annotated with blue triangles, and tetramer are annotated with inverted purple triangles. No other species were observed. Trace color corresponds to sample ID. All nMS data of RTRs, including m/z spectra, are provided in Supplemental Table 1.7. (d) Overlay of SAXS data of cleaved samples. Radii of gyration calculated from the SAXS profiles match the design models within 7.2 Å. (e) Individual 2D class averages of nsEM data collected on each sample. Outline color corresponds to sample ID. (f) Schematics of hetRTRs with interfaces colored corresponding to interface source next to nsEM data collected on each of the samples. Upper left: A2B2-hetRTR05-11; upper right: $A_5$2$B_7$2-hetRTR05-11$_{noCys}$; lower left: A3B3-hetRTR05-11$_{noCys}$; lower right: A2B2C2-hetRTR11$_{noCys}$,05-16.

Supplemental Data



Supplemental Figure 1.1: Combinatorics of homo-oligomer recombination

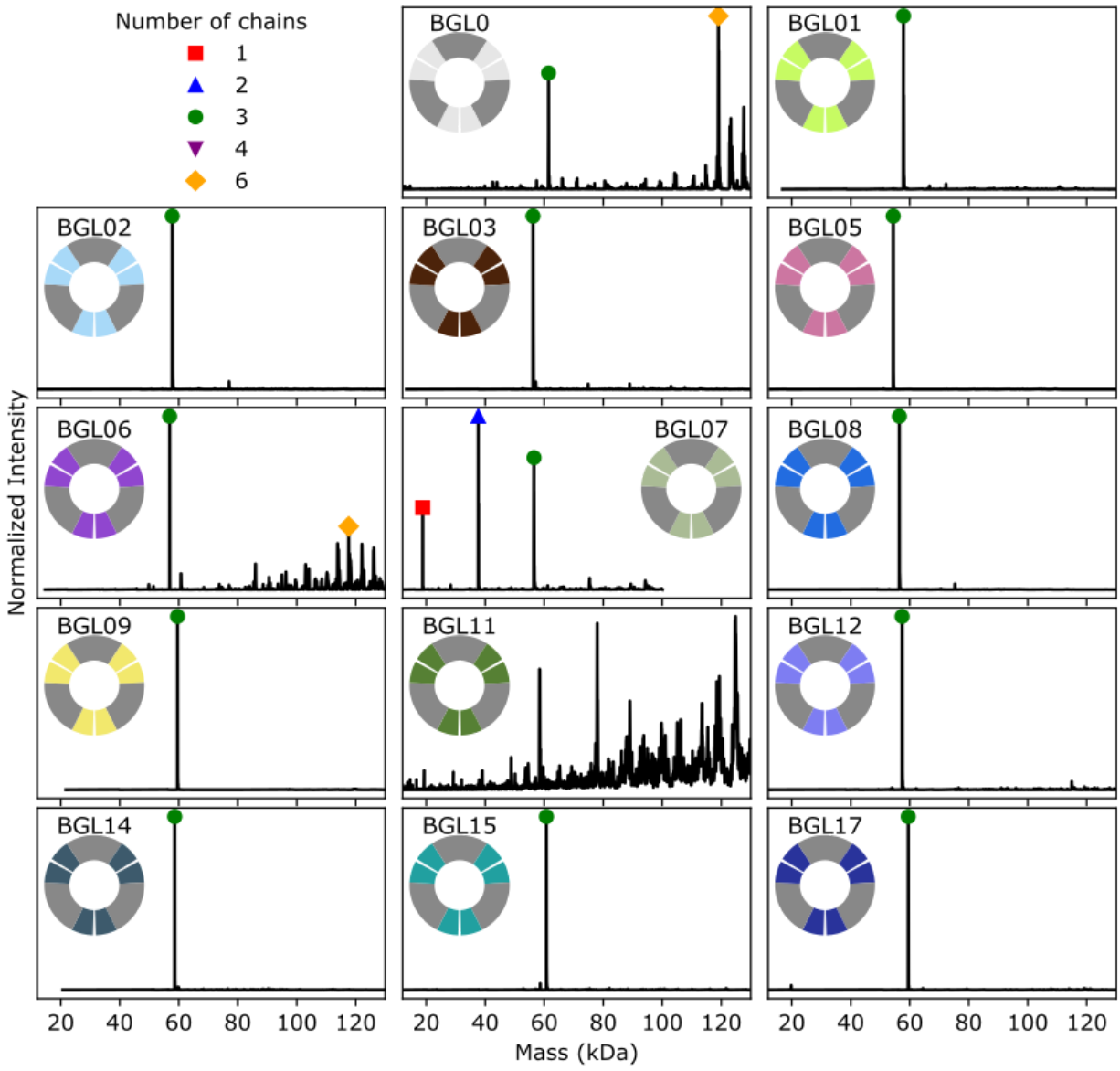
Plot of the equation (k! * (n-1)!) / (n! * (k-n)!) for a few n, where n is the number of different chains in the final heter-n-mer (e.g. heterotrimer = heter-3-mer) and k is the number of interfaces or homo-n-mers which have been validated to assemble similarly. Shown are plots of n=3, 4, or 6, some of the most commonly useful hetero-oligomers, to illustrate the large number of options for combining a few interfaces.



Supplemental Figure 1.2: SEC of BGL homotrimers

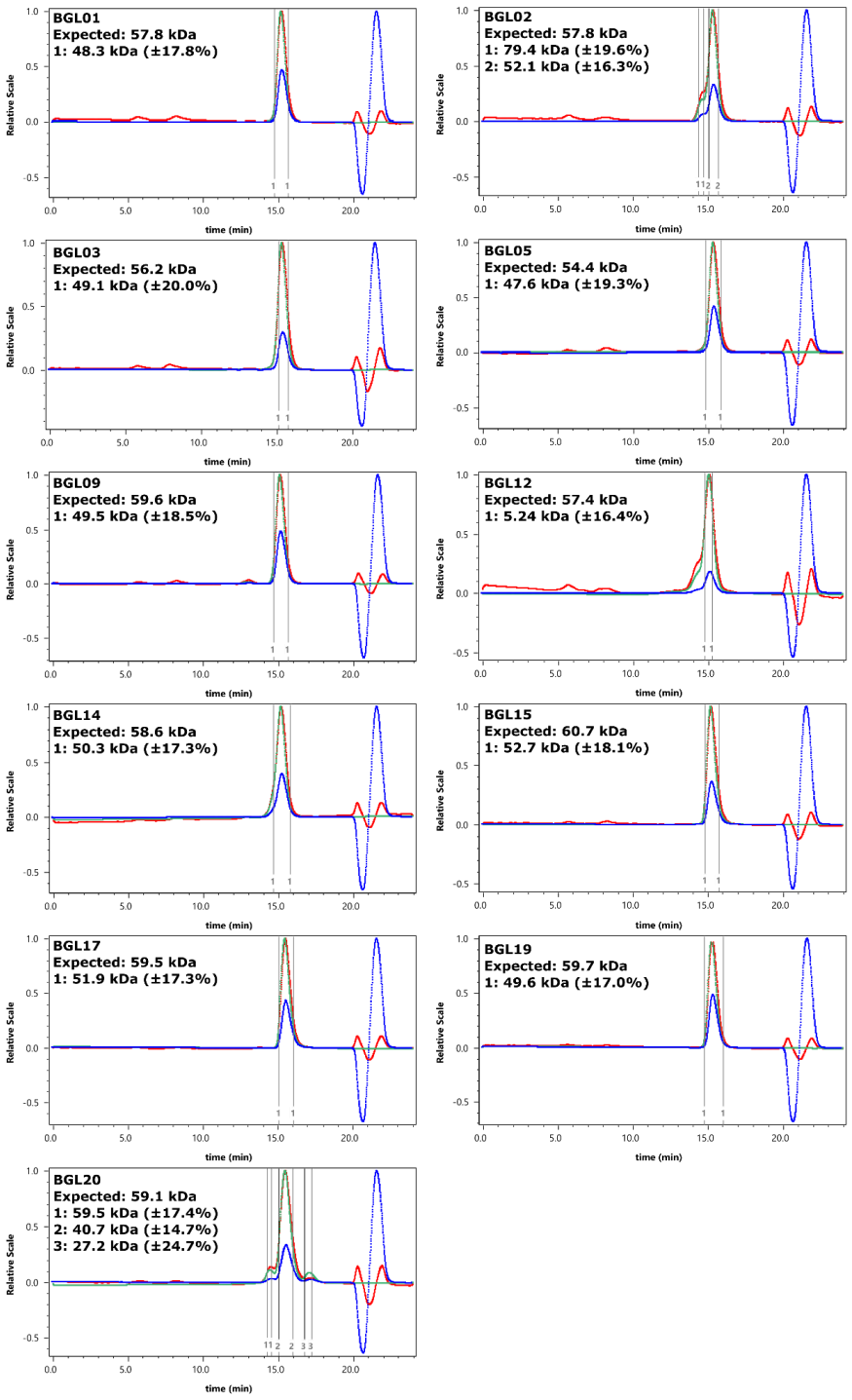
Plots of all designs before (orange) and after (blue) TEV cleavage, run on S200 Increase 10/300 GL. Note that some samples were split and run separately so absolute abundance is not comparable in all cases. See Supplemental Table 1.1 for yield information. The right-shift pre- to

post-TEV in BGL04 can be explained by removal of interactions via mediated his-tags, which are a relatively common occurrence in cyclic homo-oligomers. BGL0 was not efficiently cleaved, causing low yield of possibly mis-assembled protein in post-TEV SEC.



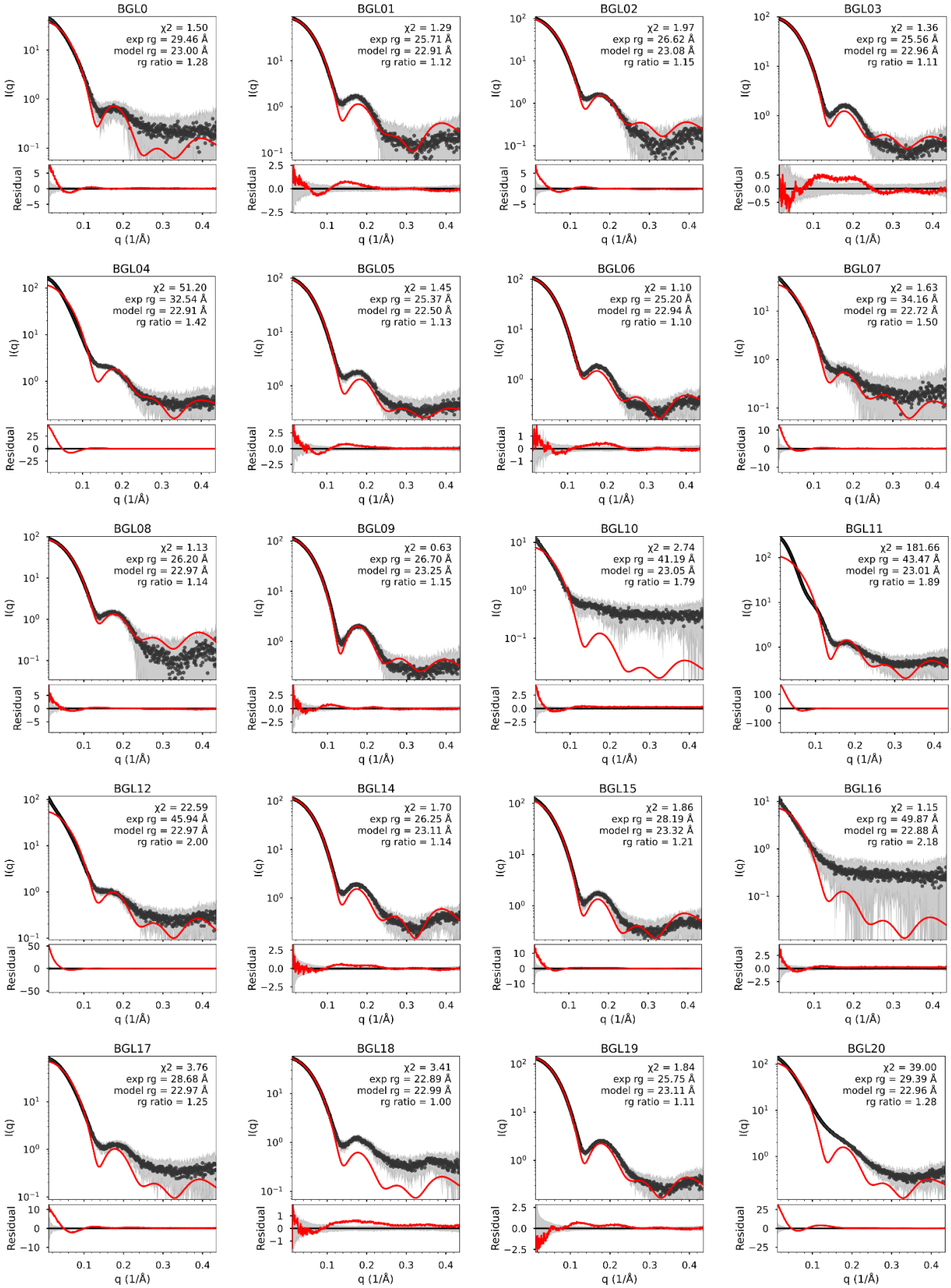
Supplemental Figure 1.3: nMS of BGL homotrimers

Plots of all designs with nMS data. Black, deconvoluted masses obtained from m/z spectra (see Supplemental Table 1.7). Identified species corresponding to monomer (red square), dimer (blue triangle), trimer (green circle), tetramer (purple upside-down triangle), and hexamer (yellow diamond) are annotated; other un-annotated peaks are unidentified species.



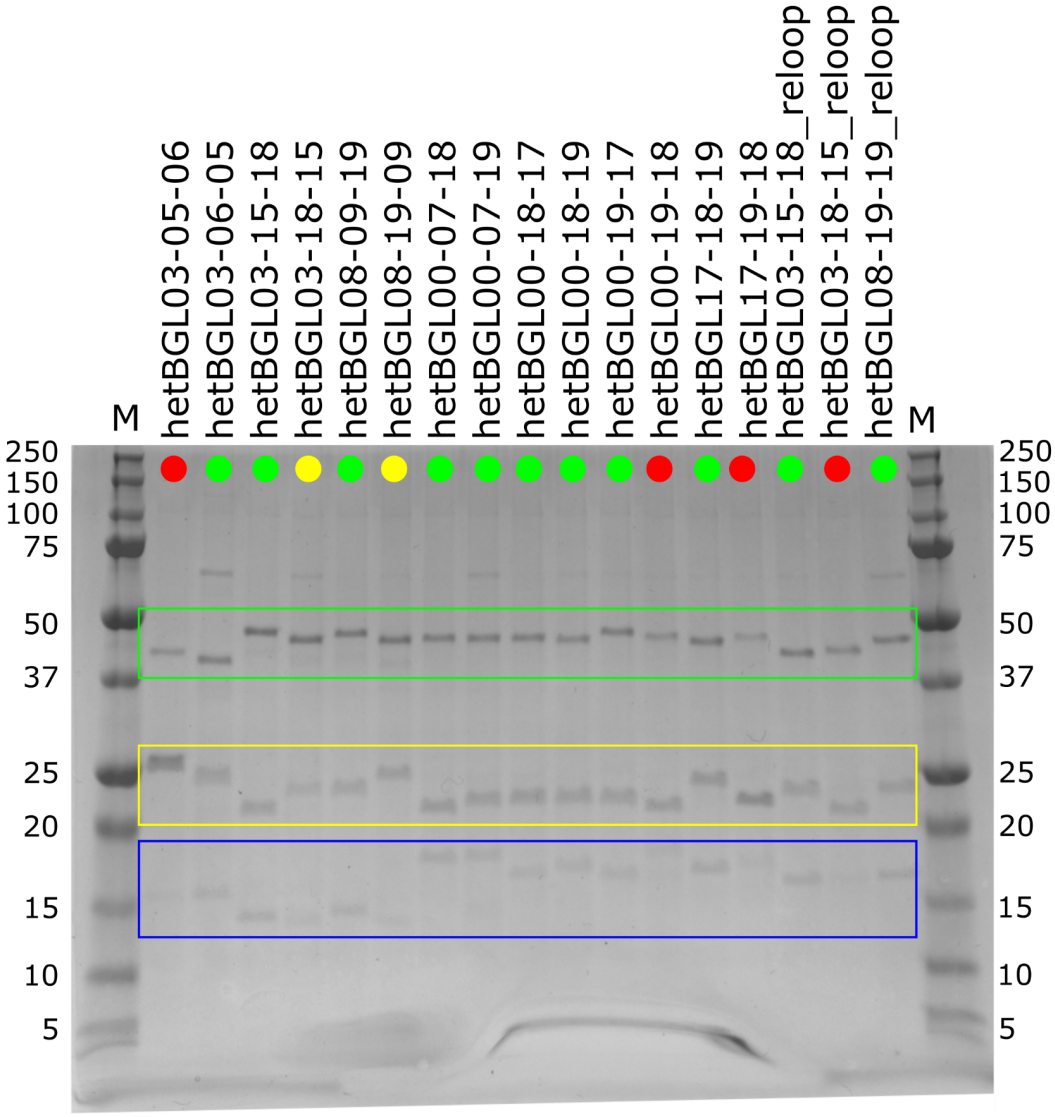
Supplemental Figure 1.4: SEC-MALS of homotrimers

Plots of all designs with SEC-MALS data. Red, light scattering signal. Green, UV absorbance. Blue, differential refractive index (dRI). Peaks are labeled by goalposts. Molecular weight estimates for each peak are shown in the top right corner of each plot.


BGL0
χ2 = 1.50
exp rg = 29.46 Å
model rg = 23.00 Å
rg ratio = 1.28
BGL01
χ2 = 1.29
exp rg = 25.71 Å
model rg = 22.91 Å
rg ratio = 1.12
BGL02
χ2 = 1.97
exp rg = 26.62 Å
model rg = 23.08 Å
rg ratio = 1.15
BGL03
χ2 = 1.36
exp rg = 25.56 Å
model rg = 22.96 Å
rg ratio = 1.11
BGL04
χ2 = 51.20
exp rg = 32.54 Å
model rg = 22.91 Å
rg ratio = 1.42
BGL05
χ2 = 1.45
exp rg = 25.37 Å
model rg = 22.50 Å
rg ratio = 1.13
BGL06
χ2 = 1.10
exp rg = 25.20 Å
model rg = 22.94 Å
rg ratio = 1.10
BGL07
χ2 = 1.63
exp rg = 34.16 Å
model rg = 22.72 Å
rg ratio = 1.50
BGL08
χ2 = 1.13
exp rg = 26.20 Å
model rg = 22.97 Å
rg ratio = 1.14
BGL09
χ2 = 0.63
exp rg = 26.70 Å
model rg = 23.25 Å
rg ratio = 1.15
BGL10
χ2 = 2.74
exp rg = 41.19 Å
model rg = 23.05 Å
rg ratio = 1.79
BGL11
χ2 = 181.66
exp rg = 43.47 Å
model rg = 23.01 Å
rg ratio = 1.89
BGL12
χ2 = 22.59
exp rg = 45.94 Å
model rg = 22.97 Å
rg ratio = 2.00
BGL14
χ2 = 1.70
exp rg = 26.25 Å
model rg = 23.11 Å
rg ratio = 1.14
BGL15
χ2 = 1.86
exp rg = 28.19 Å
model rg = 23.32 Å
rg ratio = 1.21
BGL16
χ2 = 1.15
exp rg = 49.87 Å
model rg = 22.88 Å
rg ratio = 2.18
BGL17
χ2 = 3.76
exp rg = 28.68 Å
model rg = 22.97 Å
rg ratio = 1.25
BGL18
χ2 = 3.41
exp rg = 22.89 Å
model rg = 22.99 Å
rg ratio = 1.00
BGL19
χ2 = 1.84
exp rg = 25.75 Å
model rg = 23.11 Å
rg ratio = 1.11
BGL20
χ2 = 39.00
exp rg = 29.39 Å
model rg = 22.96 Å
rg ratio = 1.28
I(q)
Residual
q (1/Å)

Supplemental Figure 1.5: SAXS profiles of BGL homotrimers

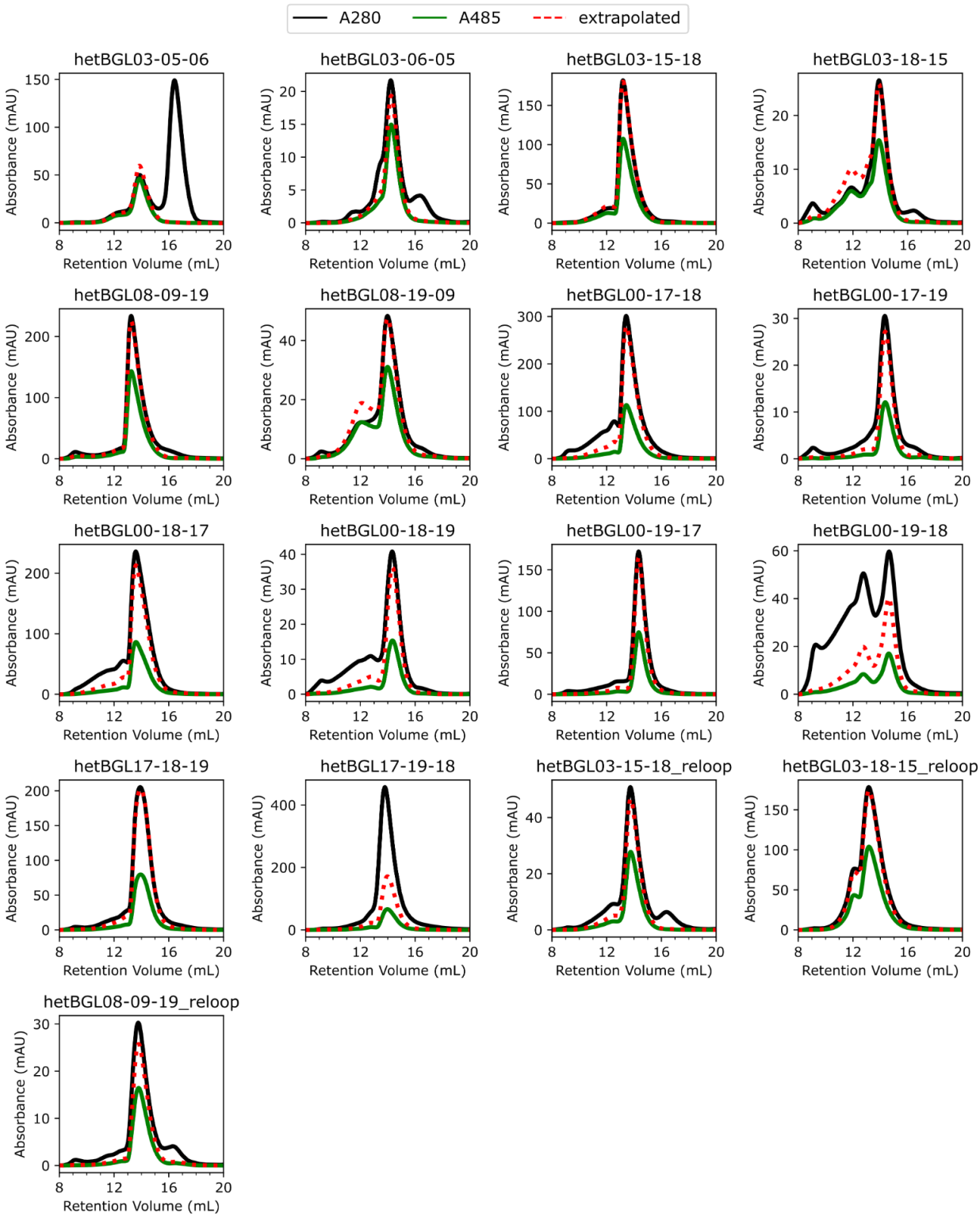
Plots of all designs with SAXS data. (Top of each panel) The scattering profiles plotting log intensity vs scattering angle. Black dots, averaged experimental data from many individual frames. Gray shading, the standard deviation of the averaged frames. Red, computed profiles generated from design models and scaled using FoXS to fit the experimental data with c1=1.05 and c2=2.0. (Bottom of each panel) The residual of the fit between experimental data and the computed profile, shown on a linear scale. Red, magnitude of the residual. Gray shading, same as top. Quality of fit (X^2), radius of gyration (rg) from experimental data and the design model, and the ratio between the two, are listed in the top right of each panel.



Supplemental Figure 1.6: SDS-PAGE of hetBGL co-purifications

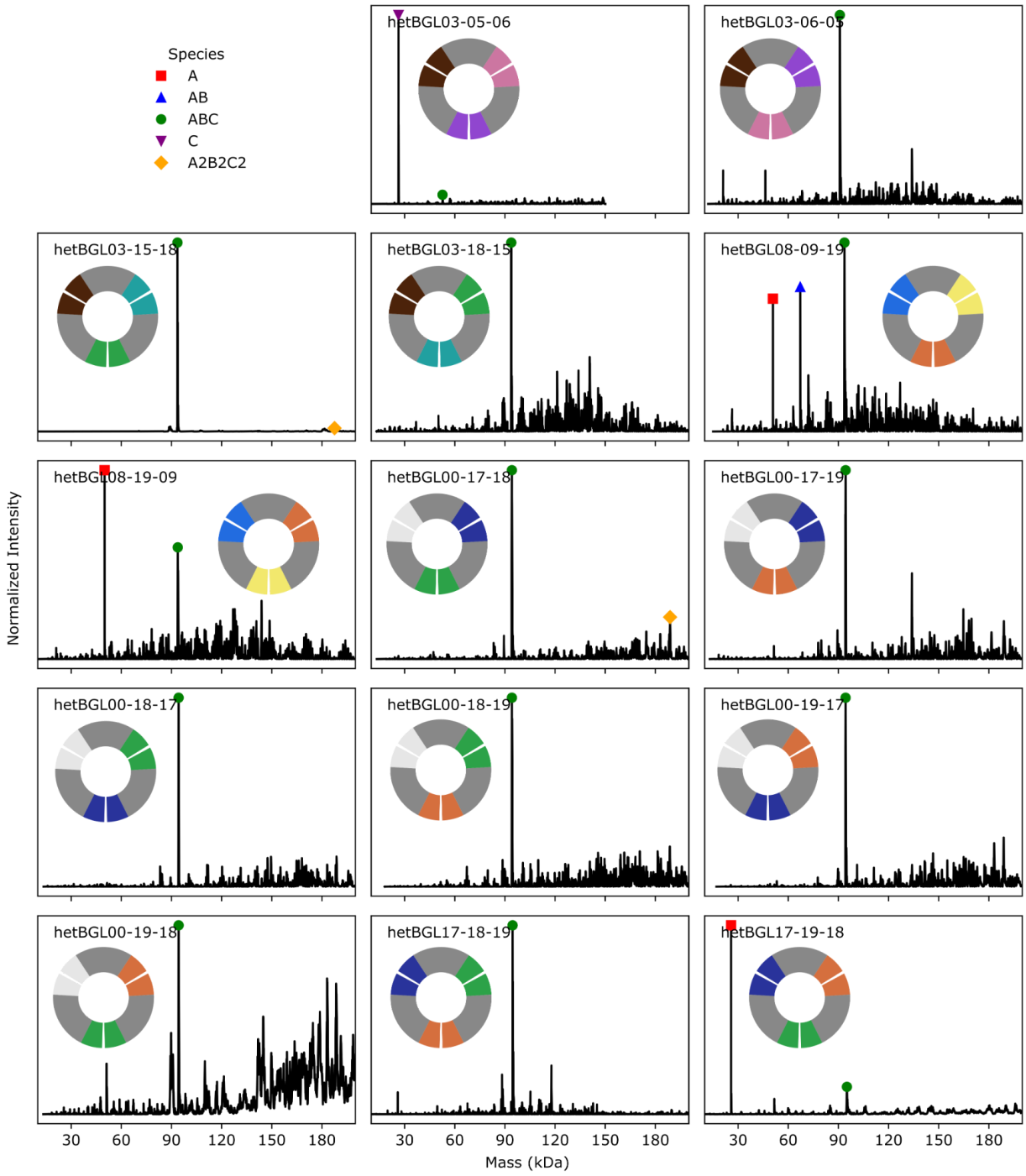
IMAC co-purified co-expressions, where each construct only contains a single histag. Refer to Supplemental Table 1.3 for histag placement. The three chains are fused to different mass-adding tags to differentiate them. Green box, GFP-tagged. Yellow box, EHEE-tagged. Blue box, no mass tag. M lanes are Precision Plus Protein Dual Xtra (Bio-Rad) and listed masses are in kDa.

Sample lanes are annotated with dots to indicate qualitative success. Red dot, missing chain. Yellow dot, all 3 chains present but uneven stoichiometries. Green dot, all 3 chains present in approximately equal stoichiometries.


A280
A485
extrapolated
hetBGL03-05-06
hetBGL03-06-05
hetBGL03-15-18
hetBGL03-18-15
hetBGL08-09-19
hetBGL08-19-09
hetBGL00-17-18
hetBGL00-17-19
hetBGL00-18-17
hetBGL00-18-19
hetBGL00-19-17
hetBGL00-19-18
hetBGL17-18-19
hetBGL17-19-18
hetBGL03-15-18_reloop
hetBGL03-18-15_reloop
hetBGL08-09-19_reloop
Absorbance (mAU)
Retention Volume (mL)

Supplemental Figure 1.7: SEC of hetBGLs

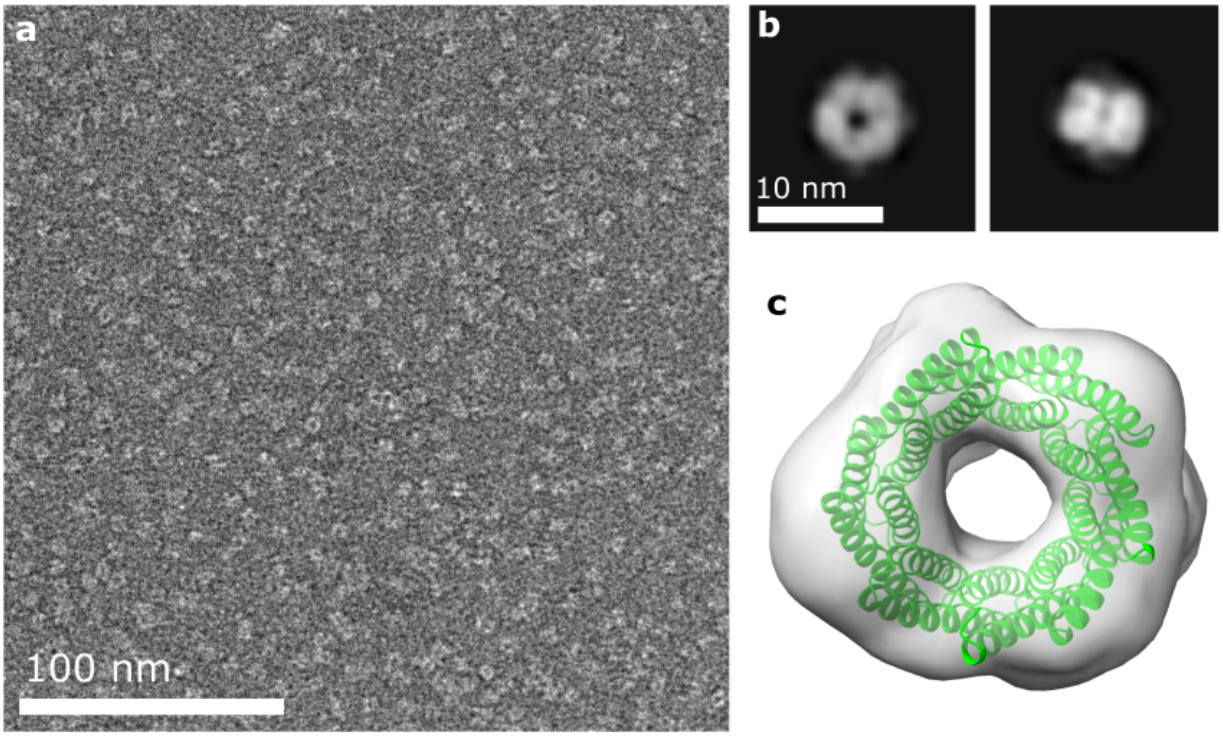
Each successfully assembling construct will have a single GFP tag associated with it, so tracking A280 (absorbance of aromatics in every chain) and A485 (absorbance specific to the GFP chromophore) separately allows for estimation of expected signal given another signal which implies GFP:complex stoichiometry. Black, measured A280. Green, measured A485. Red dashed, expected A280 signal given the A485 signal using an estimated EC485 of 36650 $M^{-1}cm^{-1}$ and estimated A280 specific to each construct.



Supplemental Figure 1.8: nMS of all hetBGLs

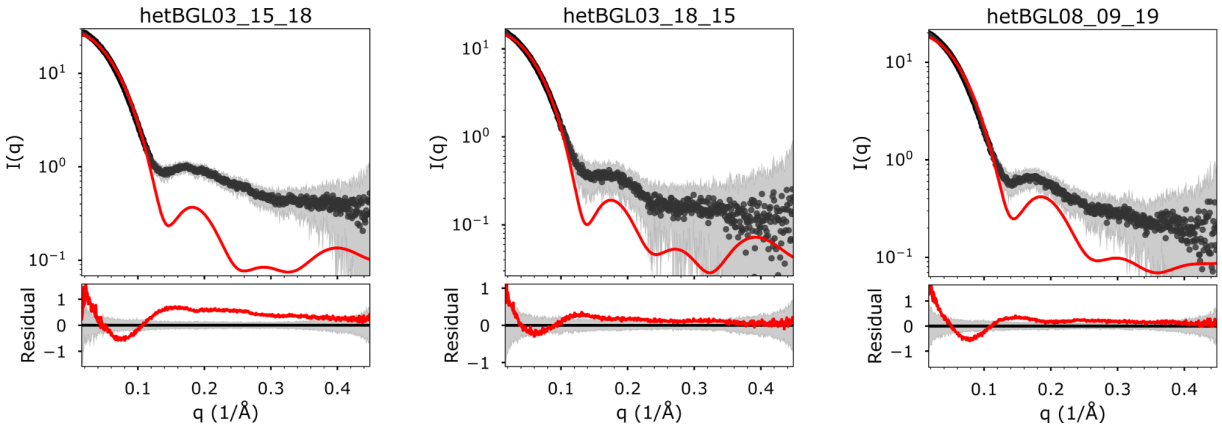
Plots of all designs with nMS data. Black, deconvoluted masses obtained from m/z spectra (see Supplemental Table 1.7). Identified species corresponding to chain A (red square), A+B (blue

triangle), A+B+C (green circle), C (purple upside-down triangle), and (A+B+C)x2 (yellow diamond) are annotated; other un-annotated peaks are unidentified species.



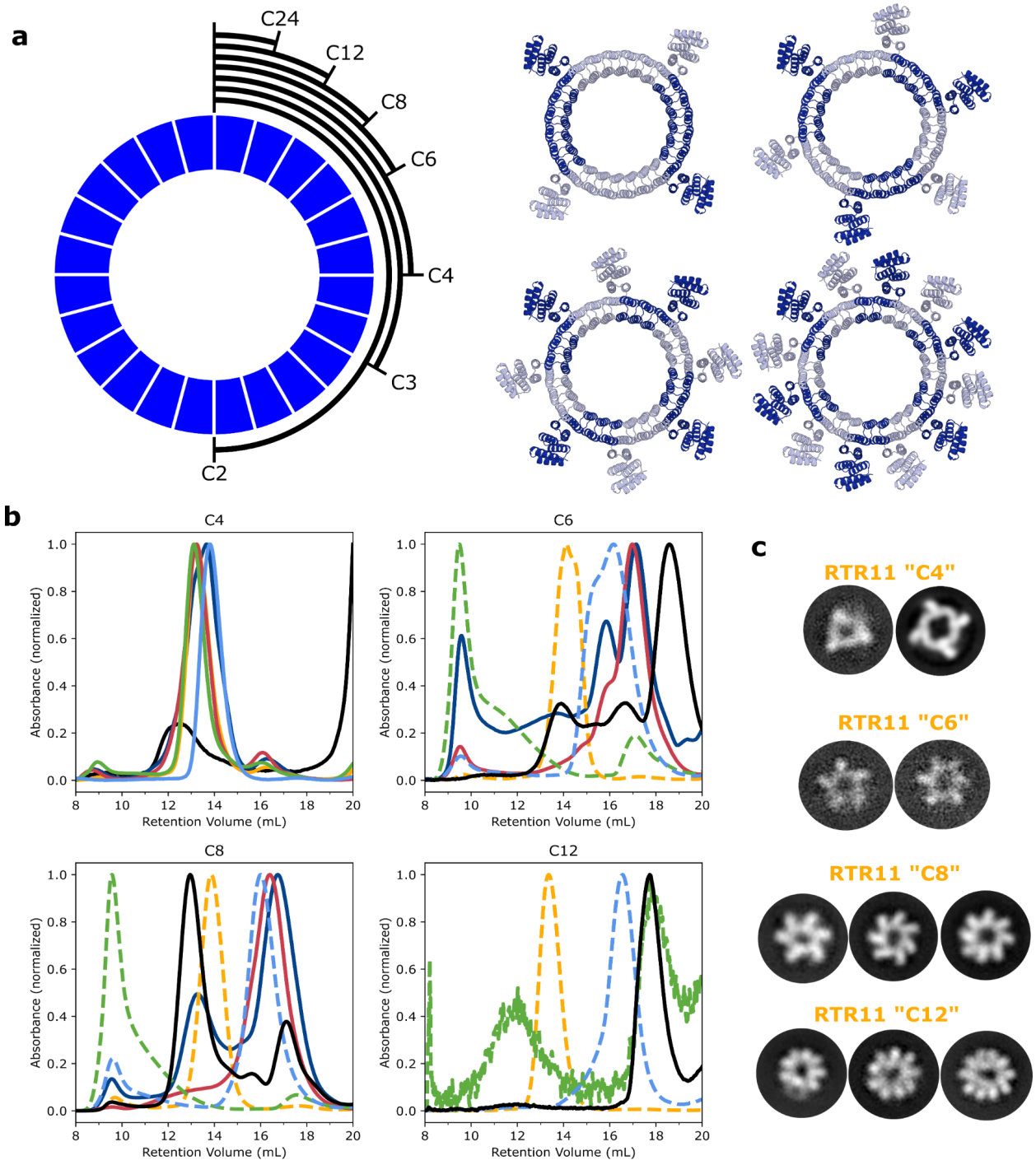
Supplemental Figure 1.9: nsEM of hetBGL03-15-18

(a) Randomly chosen example micrograph. (b) Two examples 2D class averages showing the top-down view and a probable side view. (c) Design model fit into 3D reconstruction.



Supplemental Figure 1.10: SAXS of hetBGLs

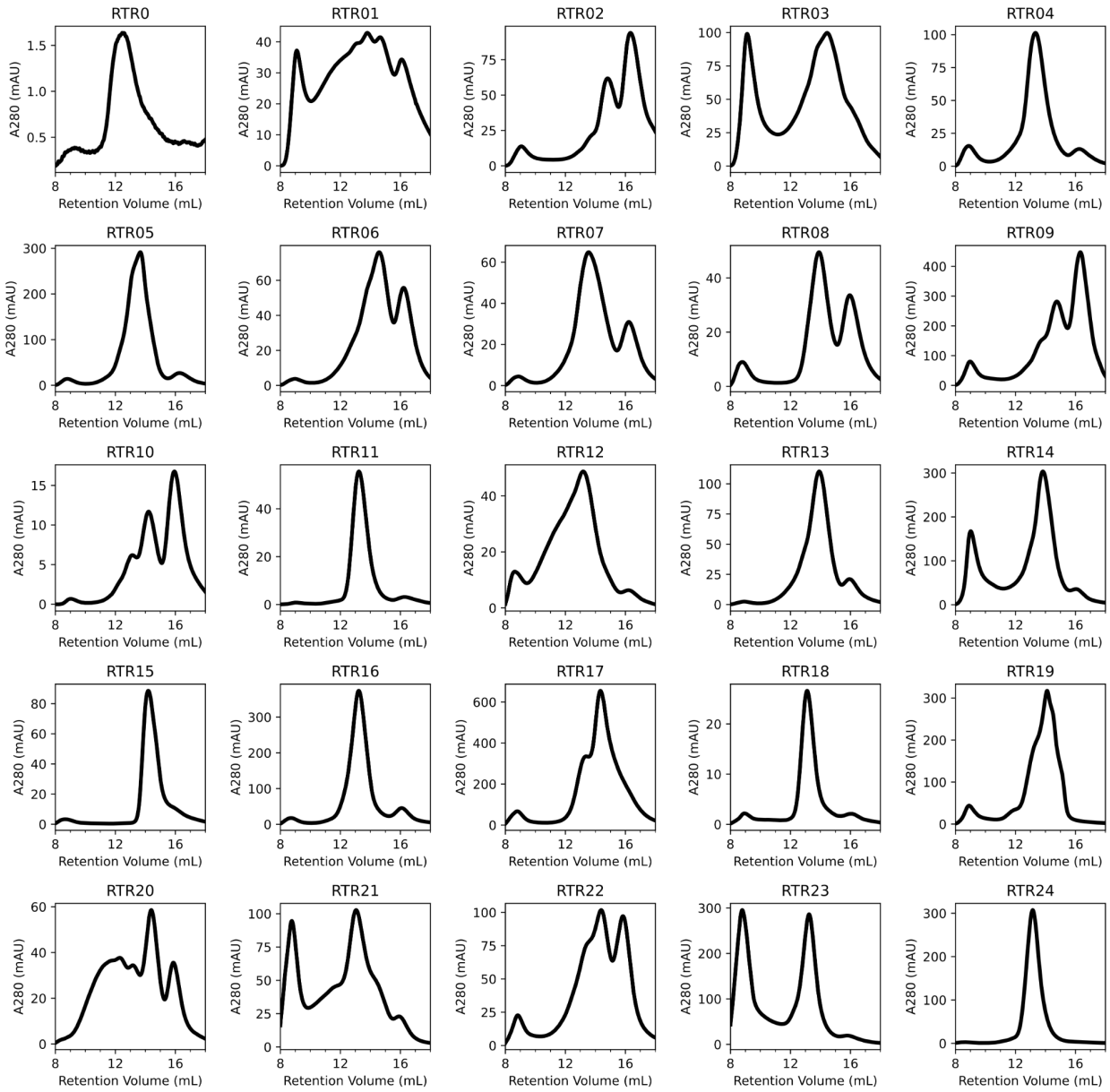
Plots of all designs with SAXS data. (Top of each panel) The scattering profiles plotting log intensity vs scattering angle. Black dots, averaged experimental data from many individual frames. Gray shading, the standard deviation of the averaged frames. Red, computed profiles generated from design models and scaled using FoXS to fit the experimental data. (Bottom of each panel) The residual of the fit between experimental data and the computed profile, shown on a linear scale. Red, magnitude of the residual. Gray shading, same as top. Quality of fit (X^2), radius of gyration (rg) from experimental data and the design model, and the ratio between the two, are listed in the top right of each panel.



Supplemental Figure 1.11: Variable oligomeric state of RTR homo-oligomers

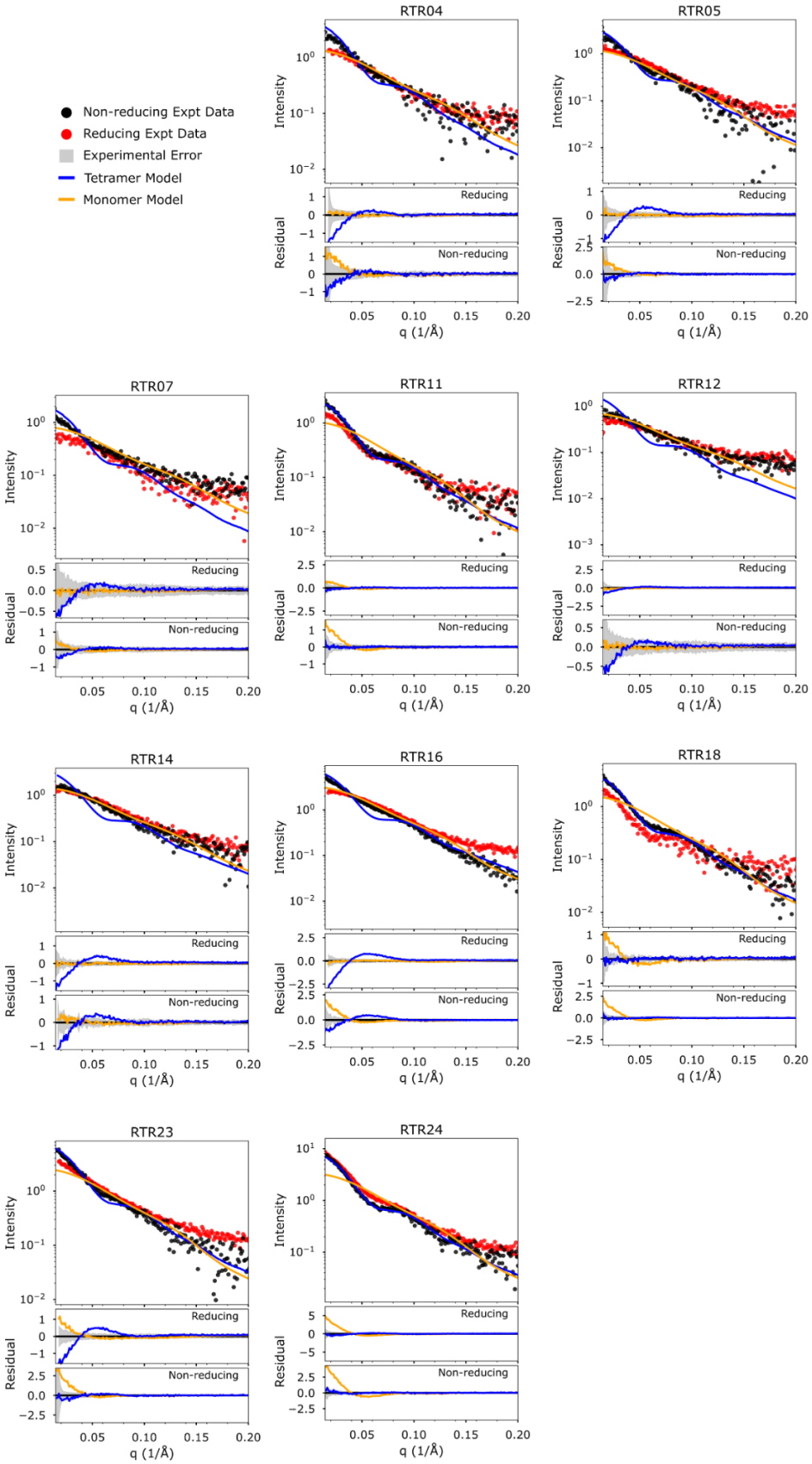
(a, left) Schematic representation of how a cyclic (24x) repeat protein (toroid) can be divided into different homo-oligomers which assemble into symmetries depending on the number of

repeats combined into a single chain. The final assemblies all have the same number of total repeats. For example, a C4 assembly is made from a single chain containing 6 repeats. (right) Cartoon models of C4, C6, C8, and C12 versions of the RTR backbone, constructed by removing repeats distant from the interface repeats. (b) SEC traces of RTR variants with different numbers of repeats per chain. Solid traces, interfaces with cysteines. Dashed traces, interfaces without cysteines. Dark blue, RTR05. Yellow, RTR11. Red, RTR16. Green, RTR18. Light blue, RTR24. (c) Negative stain EM 2D class averages of RTR11 with different numbers of repeats per chain. Each class represents a different observed oligomeric state.



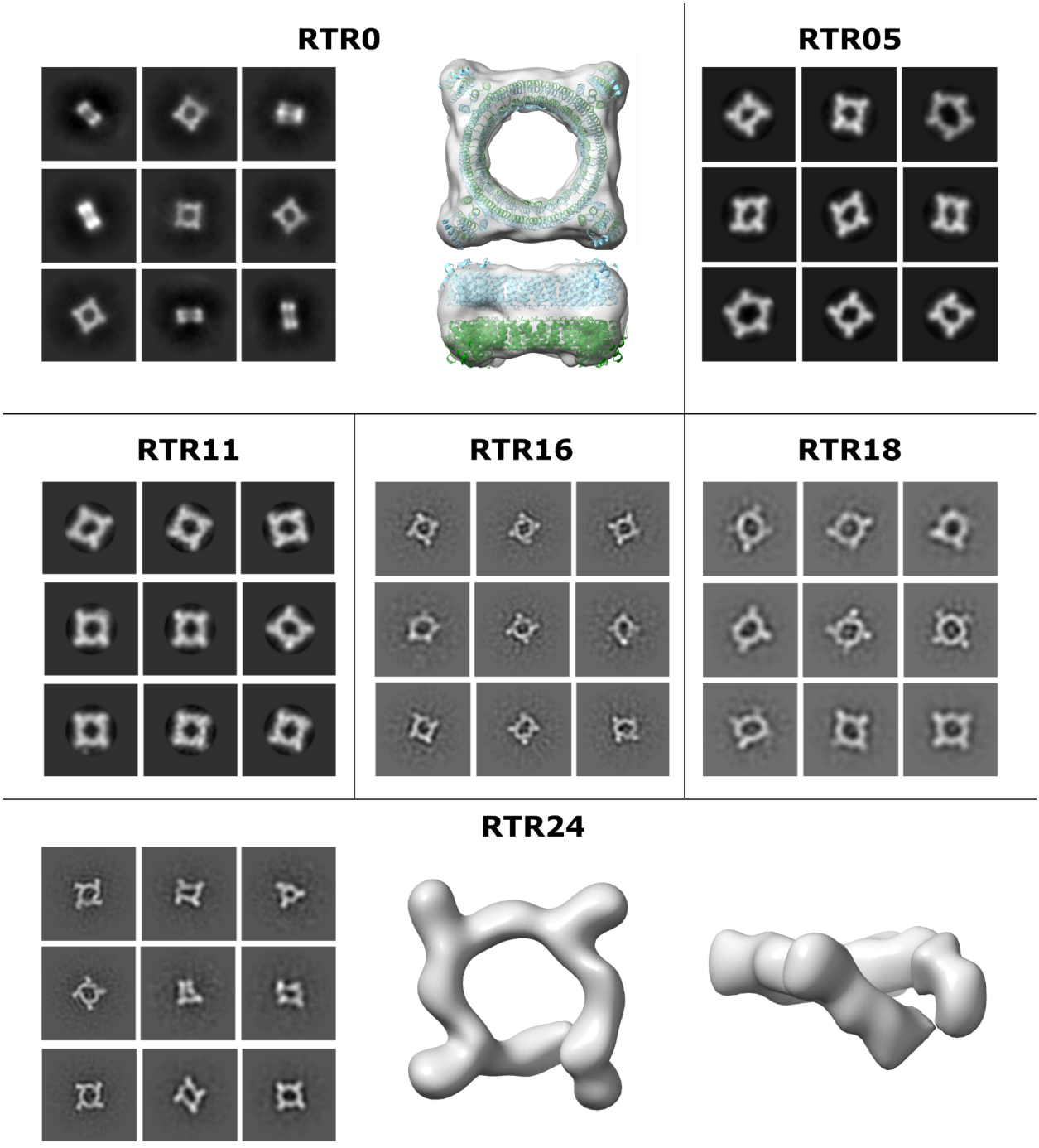
Supplemental Figure 1.12: SEC of RTR homo-tetramers

SEC traces of RTR samples on S200 column.


Non-reducing Expt Data
Reducing Expt Data
Experimental Error
Tetramer Model
Monomer Model
RTR04
RTR05
RTR07
RTR11
RTR12
RTR14
RTR16
RTR18
RTR23
RTR24
Intensity
Residual
Reducing
Non-reducing
q (1/Å)

Supplemental Figure 1.14: SAXS of RTR homo-tetramers

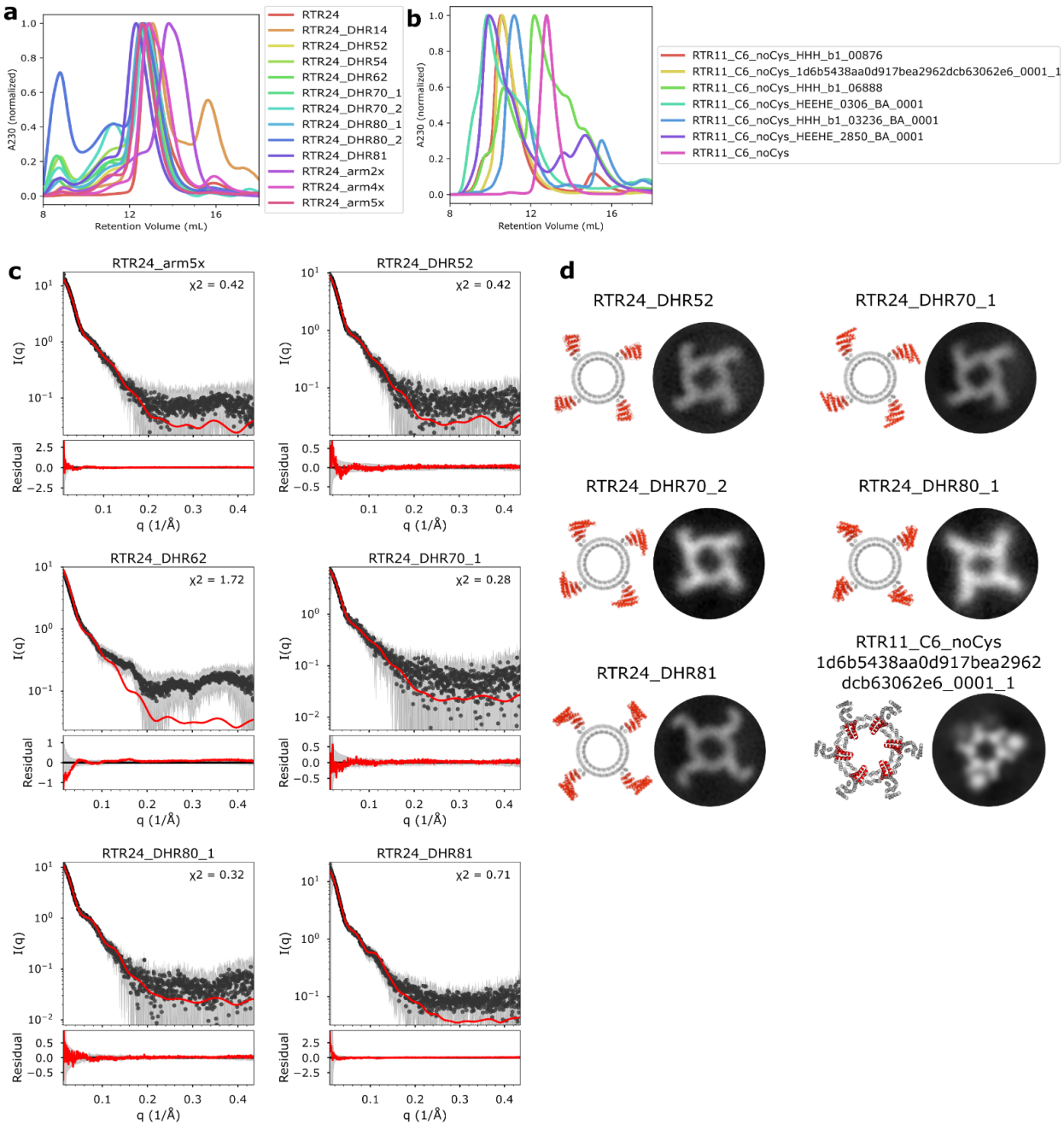
Plots of all designs with SAXS data. (Top of each panel) The scattering profiles plotting log intensity vs scattering angle. Black dots, averaged experimental data from many individual frames of non-reduced samples. Red dots, same as black dots but of reduced samples. Gray shading, the standard deviation of the averaged frames. Blue line, computed profiles generated from tetrameric design models and scaled using FoXS to fit the experimental data for the non-reduced data. Orange line, same as blue but with monomeric design models (Bottom of each panel) The residual of the fit between experimental data and the computed profile for reducing or non-reducing data, shown on a linear scale.



Supplemental Figure 1.15: nsEM of RTR homo-tetramers

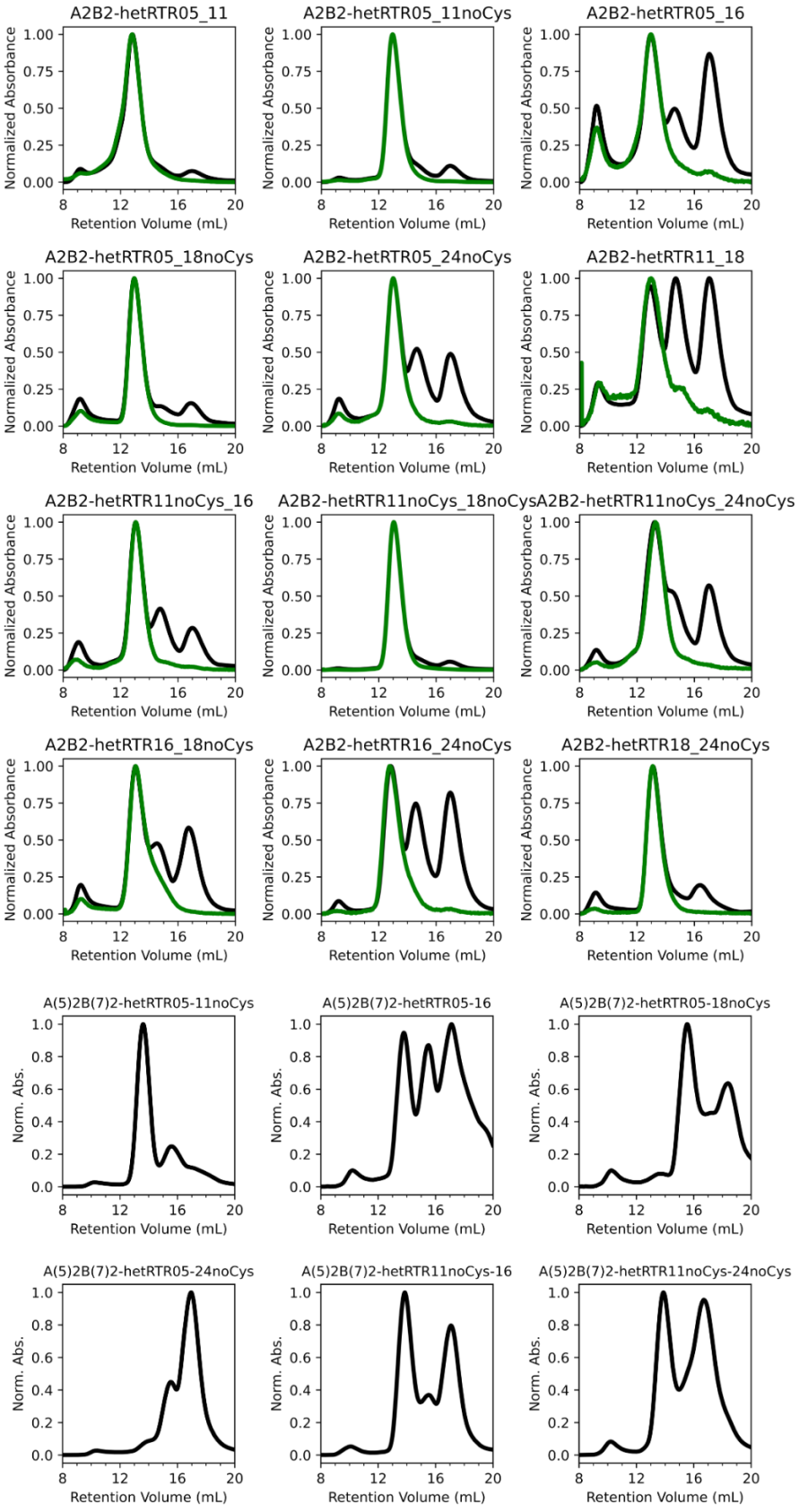
Negative stain 2D class averages and 3D reconstructions of RTR0 and RTR variants, if possible. RTR16, RTR18, and RTR24 were processed without CTF correction during 2D classification.

(RTR0) The map of RTR0 is an *ab inito* reconstruction with symmetry. Two copies of the design model of RTR0 were roughly fit into the map. (RTR24) The map of RTR24 is a heterogenous refinement without symmetry. Multiple classes of this map were observed with varying degrees of ring breakage.


a
A230 (normalized)
Retention Volume (mL)
RTR24
RTR24_DHR14
RTR24_DHR52
RTR24_DHR54
RTR24_DHR62
RTR24_DHR70_1
RTR24_DHR70_2
RTR24_DHR80_1
RTR24_DHR80_2
RTR24_DHR81
RTR24_arm2x
RTR24_arm4x
RTR24_arm5x
b
A230 (normalized)
Retention Volume (mL)
RTR11_C6_noCys_HHH_b1_00876
RTR11_C6_noCys_1d6b5438aa0d917bea2962dcb63062e6_0001_1
RTR11_C6_noCys_HHH_b1_06888
RTR11_C6_noCys_HEEHE_0306_BA_0001
RTR11_C6_noCys_HHH_b1_03236_BA_0001
RTR11_C6_noCys_HEEHE_2850_BA_0001
RTR11_C6_noCys
c
RTR24_arm5x
χ2 = 0.42
RTR24_DHR52
χ2 = 0.42
RTR24_DHR62
χ2 = 1.72
RTR24_DHR70_1
χ2 = 0.28
RTR24_DHR80_1
χ2 = 0.32
RTR24_DHR81
χ2 = 0.71
I(q)
Residual
q (1/Å)
d
RTR24_DHR52
RTR24_DHR70_1
RTR24_DHR70_2
RTR24_DHR80_1
RTR24_DHR81
RTR11_C6_noCys
1d6b5438aa0d917bea2962
dcb63062e6_0001_1

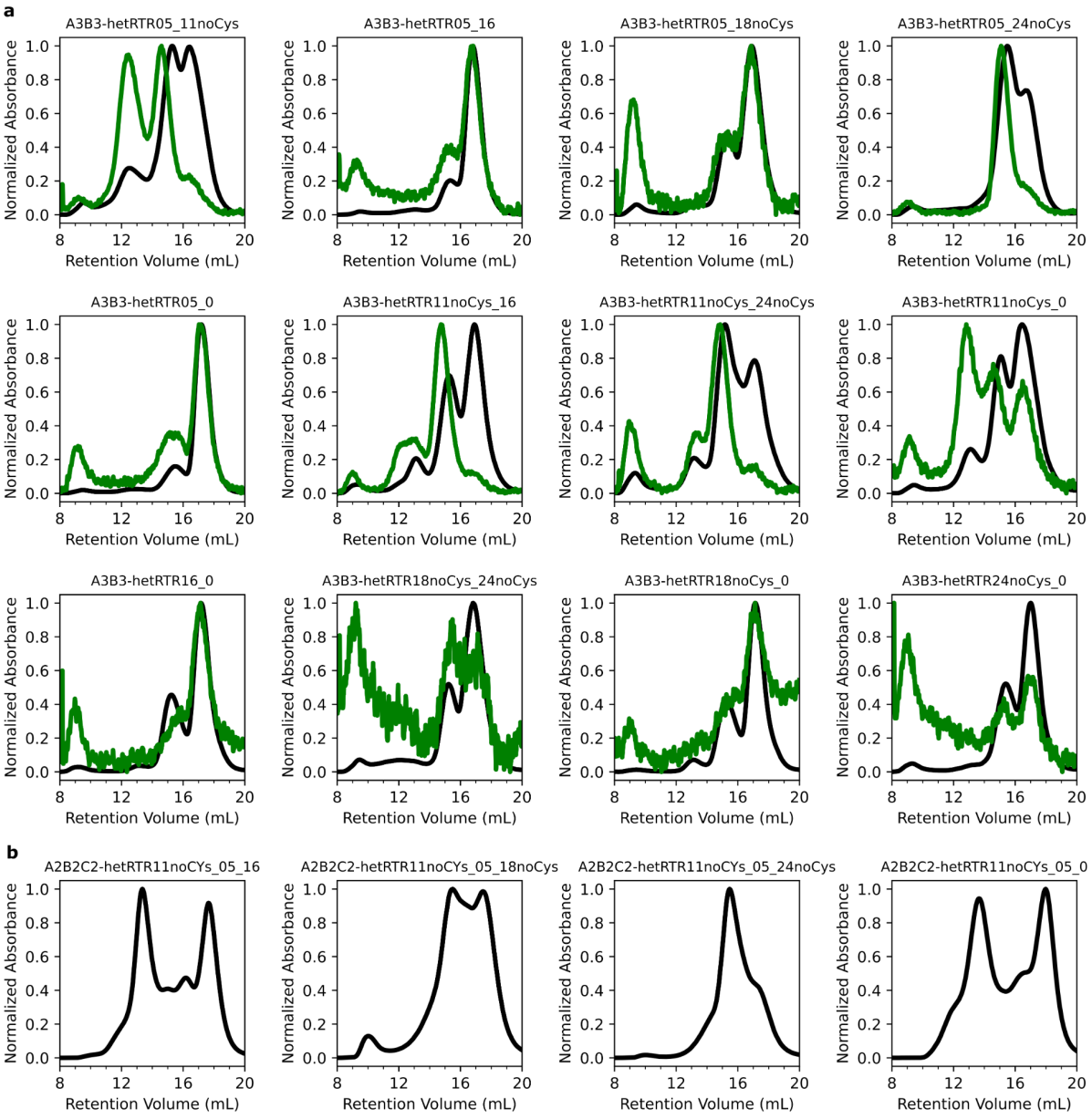
Supplemental Figure 1.17: Characterization of RTR extensions

(a) SEC traces of RTR24 with c-terminal extensions. RTR24 included as a size reference. (b) SEC traces of RTR11_C6 with n-terminal extensions. RTR11_C6 included as a size reference. (c) SAXS profiles of c-terminally extended designs. Black, experimental data. Red, theoretical profile based on design model fit to experimental data. Gray shading, standard deviation of averaged micrographs. (d) Negative stain EM 2D class averages of five c-terminally and one n-terminally extended RTR variants paired with cartoon models with the extended region shown in red. RTR11_C6_noCys_1d6b5438aa0d917bea2962dcb63062e6_0001_1 was tagged with GFP, which appear as ordered density between the outer arms.


A2B2-hetRTR05_11
A2B2-hetRTR05_11noCys
A2B2-hetRTR05_16
A2B2-hetRTR05_18noCys
A2B2-hetRTR05_24noCys
A2B2-hetRTR11_18
A2B2-hetRTR11noCys_16
A2B2-hetRTR11noCys_18noCys
A2B2-hetRTR11noCys_24noCys
A2B2-hetRTR16_18noCys
A2B2-hetRTR16_24noCys
A2B2-hetRTR18_24noCys
A(5)2B(7)2-hetRTR05-11noCys
A(5)2B(7)2-hetRTR05-16
A(5)2B(7)2-hetRTR05-18noCys
A(5)2B(7)2-hetRTR05-24noCys
A(5)2B(7)2-hetRTR11noCys-16
A(5)2B(7)2-hetRTR11noCys-24noCys
Normalized Absorbance
Norm. Abs.
Retention Volume (mL)

Supplemental Figure 1.18: SEC of A2B2-hetRTRs

(a) Black, normalized absorbance at 230nm. Green, normalized absorbance at 480nm, corresponding to the GFP-tagged chain. (b) Normalized absorbance at 230nm.



Supplemental Figure 1.19: SEC of A3B3/A2B2C2-hetRTRs

(a) Black, normalized absorbance at 230nm. Green, normalized absorbance at 480nm, corresponding to the GFP-tagged chain. (b) Normalized absorbance at 230nm.

Supplemental tables

included in Supplemental Files:

Supplemental Table 1.1: BGL homotrimer metadata

Supplemental Table 1.2: BGL crystallographic information

Supplemental Table 1.3: BGL heterotrimer metadata

Supplemental Table 1.4: RTR homo-oligomer metadata

Supplemental Table 1.5: A2B2-hetRTR metadata

Supplemental Table 1.6: A3B3/A2B2C2-hetRTR metadata

Supplemental Table 1.7: Native MS spectra of all designs

Computational Methods

All code is available at https://github.com/rdkibler/pseudosymmetric_hetero-oligomers.

While Rosetta is capable of performing every step of the interface redesign process using its symmetry machinery[44] and thus using a symmetric pose at every step, because we require the homology region to have identical sequence and coordinates between every design, we cannot make use of the concerted subunit rigid body transformation degrees of freedom that the symmetry machinery is, in part, intended to provide. Any full-chain movements away from the input structure in the homology region are not allowed. Additionally, we would not use the symmetry machinery's functions which make the packer and minimizer aware of the other chains in the structure due to the physical separation between the interfaces. Therefore, all steps until the final structure checks are performed on the structure asymmetrically using coordinate constraints to hold the homology regions in place. In the case of RTR0, we also got a large speed up by using a significantly truncated version of the interface region which included just the edge of the homology region.

1. BGL redesign

All BGL design was carried out on the full 18 helix structure of BGL0 (PDB 6XR1), since this is semantically equivalent to working on the fixed-size homotrimer using symmetry operations, and avoids the need for symmetry which increases the computational overhead.

1.1 Backbone perturbation

1.1.1 Helix Rebuilding (HR) set

Canonically coiled alpha-helical backbones can be described by their Crick helical parameters[32,45], that is, for a given helix coiling around the Z-axis, the values of r0, omega0,

omega1, phi0, phi1, detla_z, and whether the helix runs toward negative Z (inverted) or toward positive Z (not inverted). With the axis of structural symmetry lying on the Z-axis, the outer helices of BGL0 resemble the helix of a coiled coil with very large r0 (distance from the Z-axis). We sought to resample the backbone positions of the outer helix at the interface, and reasoned that replacement helices closeby in the parameter space of the original helix would also fit well physically while giving a large increase in backbone position (especially via sampling uniformly over phi1). To accomplish this, we wrote an algorithm which discovered the 7 helical parameters via least squares minimization. Most parameters were largely unconstrained save for r0 bounded between 1e-6 and 40 and phi0 bounded to >= 0. The parameters r0, omega0, omega1, phi0, phi1, and delta_z, were initialized to a generic left-handed supercoiled helix, with phi0 and r0 estimated from the position of the target helix's center of mass, and at every step the per-backbone-atom RMSD was computed between the target helix and a helix with the current parameters and a least squares algorithm was used to find the set of parameters which minimizes RMSD. The target in this campaign for BGL0 was the C-terminal helix, one of the outer ring of helices and the closest fit we achieved was 0.8A. Given this best description of the target helix as an ideal coiled coil helix, we deleted the target helix from the BGL0 structure and sampled 124,852 new helices with parameters r0 between 20 and 25, delta_z +/- 6 from the fitted value, omega0 +/- 1 degree from the fitted value, phi0 +/- 20 degrees from the fitted value, and phi1 +/- 154.3 degrees from the fitted value. Some solutions clashed with the rest of the BGL0 structure, so rather than fine-tuning parameter ranges which might have a complex relationship, we roughly evaluated clashing with a Rosetta scorefunction scoring only fa_rep, fa_atr, and fa_sol (delta REU from original structure < 9000) and buried surface area (at least 350 with either side of the interface).

The output of the helix rebuilding procedure is a disconnected helix roughly in the same position as the target helix and it needs to be re-connected to the rest of the structure via a loop. Because BGL0 is a toroid, the N- and C-termini are physically adjacent to each other and could be re-connected to either side. Therefore, for each output of helix rebuilding, we attempted to re-connect the helix as either the N-terminus (building a loop between the current N-terminus of the full structure and the new helix) or the C-terminus (building a loop between the current C-terminus of the full structure and the new helix) using ConnectChainsMover[41] with loop lengths between 2 and 5 residues and the option to add or remove five residues from either side of the loop. This resulted in the creation of 3524 outputs each for N-term closure and C-term closure.

Extra functionality

While not necessary for this work, we found that Crick helical parameters do not always generalize to helical segments which appear curved like coiled coils but, for example, whose superhelical curvature is on an axis which does not pass through the center of mass of the protein. To account for this, we have also developed a method for fitting a reduced set of Crick helical parameters and 6 additional parameters corresponding to a homogeneous transform (3 rotation and 3 translation) of the helix using least squares minimization. The parameters made redundant by this approach and which are thus never sampled are delta_z and invert because they can be easily described by the transform parameters. The algorithm proceeds in three phases.

In the first phase, the goal is to roughly locate the superhelical axis. The helix parameters are initialized to those of a generic left-handed superhelical helix (r0 = 5, phi1=0 degrees, phi0=100 degrees, omega0=-2.85 degrees, omega1=102.85 degrees) and the transform is

initialized to a rough initial guess based on the target helix's center of mass and orientation of the termini relative to the Z-axis. Phi1 and transform are fitted to the target helix to refine the initial guess of the transform of the superhelical axis. The signal from the phase of the minor helix (the atomic positions) is much stronger than the major helix (superhelical axis), so phi1 is allowed to vary to avoid biasing the transform by a bad guess of phi1.

In the second phase, which may not be necessary, the transform is fixed and omega1, omega0 (coupled to omega1 via the relationship `abs(abs(omega0+omega1)-radians(100)) < 10`, phi0, and phi1 are allowed to minimize in order to roughly shape the helix and get it close to its final form in the context of a new superhelical axis.

In the third phase, r0, omega0, omega1, phi1 (not phi0), and the transform are allowed to vary to simultaneously fine-tune the helical parameters and the effective location of the superhelical axis. The tolerance of the fit is set much tighter than for phases 1 or 2. Given a set of parameters that describe the new helix, a similar closeby parameter search and hash table clash check could be performed.

1.1.2 Normal Modes (NM) Relaxation

Separately from HR, we generated conformations very similar to the native structure but which may differ only via molecular breathing type motions by applying the NormalModeRelaxMover (NMRM)[31] to generate another set of designs (NM). We restricted backbone perturbation to the two and a half helices on either side of the interface (N-term Interface: n-term through the middle of the 3rd helix; C-term Interface: the middle of the 3rd-to-last helix through the c-term), and additionally used coordinate constraints to hold everything else in place. NMRM was run successively and separately on the C-term Interface

then the N-term interface, with the rationale that allowing the movement of the C-term Interface and its outer C-terminal helix first would create room for a greater range of movement by the N-term Interface. 500 backbones were generated with pertscale=2.0, nmodes=60, and mix_modes=true for both NMRM runs.

1.2 HBNet design

HBNet design was carried out slightly differently depending on the design set. Common to both were the region within which HBNets could be designed (the union of the N-term Interface and C-term interface), as well as a requirement to have at least 4 network residues in the core, have no unsatisfied heavy atoms or polar hydrogens, make at least 2 intermolecular hbonds, and have between 1 and 3 networks in the output structure. Hb_threshold was set to -0.6 and the MC sampling was turned on with seed_hbond_threshold -0.4 and 10,000 MC steps for HR and 100,000 for NM for deeper sampling. Core cutoff for HR was set to SASA < 20 (ball radius 2.5), but for HR was SASA < 6 (more stringent) to encourage deeper HBNets at a cost to sampling.

The HBNets created for the NM set were highly redundant due to the similarity of the backbones inputs and depth of sampling, so the 12954 outputs were grouped by the sequence of their buried HBNet residues (SASA < 20 with ball radius 2.5) and the best scoring output was kept for each group (3332 total).

The HBNets for HR could not be grouped as easily but were qualitatively more diverse, so the 112039 outputs were instead briefly fixed-backbone designed with HBNet's AtomPair distance constraints on and sidechain relaxed with constraints off to allow poorly positioned and/or poorly packed HBNets to move out of position. Then, designs were filtered on vbuns <= 2, # alanines <= 126, and total HBNet constraint score (indicating movement of HBNet residues)

< 1.0, yielding 2538 outputs. The new helix and new loops were cartesian minimized in preparation for interface design.

1.3 Interface design

In both cases, flexible backbone FastDesign was performed with the standard suite of TaskOperations (InitializeFromCommandline, DesignRestrictions, IncludeCurrent, LimitAromaChi2, and ExtraRotamersGeneric [ex1 & ex2]). Design was carried out with variable bondangle and bondlengths, and the approximate_buried_unsat_pentality and PruneBuriedUnsats[46] were used to aid in retaining or extending the HBNets. Flexible backbone movement was only allowed on the N-term Interface and C-term Interface regions, but sequence design was allowed at the interface between the Interface regions and in a 7A neighborhood around each HBNet residue (but not including the HBNet residues or proline or glycine). The residues within 6A of the designable region were allowed to repack and sidechain minimize and the HBNet residues had their AtomPair constraints enabled. FastDesign was carried out with a modified version of the MonomerDesign2019 relax script which does not ramp coordinate constraints.

1.3.1 HR specific

Calpha atoms were loosely coordinate constrained over the entire pose. The pose has a reasonable sequence already from the HBNet filtering step, so only minor changes to core packing were allowed for L, I, and V residues which were only allowed to change to L, I, or V. PruneBadRotamers and ConsensusLoopDesign TaskOperations were also employed. The surface was not designed in order to speed up design around the HBNets which were buried. Finally, because the new helix created by HR was given a “rosetta default” sequence, additional amino acid composition constraints were used to constrain the structure to have an amino acid sequence

profile roughly similar to the original design, namely requiring at most 1 methionine, 127 alanines, 2 phenylalanines, and 5 threonines, with some combination of linearly increasing penalties within a small range then increasing quadratically at higher counts. 7 separate design trajectories were run on each input, resulting in creation of 12256 outputs.

1.3.2 NM specific

Calpha atoms were tightly coordinate constrained only over the non-Interface region to allow more movement at the Interface. 7 separate design trajectories were run on each input, resulting in creation of 12879 outputs.

1.4 Filtering

Designs were filtered on a variety of criteria to reduce each design set to 100 for manual inspection. The criteria are the following:

HR:

Cst_after < 1.0

Mismatch_probability < 0.06

Network_holes < -1.4

Interface_holes < 0

dSASA_polar > 1500

vbuns5.5_heavy_ball_1.1D = 0

RotamerBoltzmannWeight (HBNet resis) < -0.51

lowest 100 by worst9mer

NM:

Cst_after <1.0

Network_holes < -1.4

Interface_holes < 0

P_aa_pp < -130

Mismatch_probability < 0.07

vbuns5.5_heavy_ball_1.1D = 0

RotamerBoltzmannWeight (HBNet resis) < -0.52

Highest 100 by dSASA_polar

The sets were then manually inspected, visually looking for hydrophobic interdigitation in the core, interesting and practical HBNets, and reasonable burial of HBNets. Finally, 16 HR designs and 4 NM designs from each set of 100 were rearranged to be (6 helix) homotrimers, ordered, and characterized.

1.2 RTR redesign

Design was carried out on a truncated sub-structure of RTR0 focusing on the interface region, comprising just 210 residues which represent four helices from the N-terminal side of the interface (two participating in the interface and two representing the fixed homology region) and 8 helices from the C-terminal side of the interface (4 which participate in the interface, two which reprepresent the fixed homology region of the ring, and two which represent the fixed part of the C-terminal extension).

1.1 Backbone perturbation

NMRM parameters were optimized to create a variety of pocket sizes and shapes in the interface (pertscale = 1.6, nmodes = 31, mix_modes=true). The regions allowed to move, Interface A and Interface B, were defined as regions 1-33 and 100-174 on the truncated

“mini”-pose RC4_20_mini.pdb. Before running NMRM, the sequence of the pose was changed to polyvaline and the beta_nov16_soft weights were used in order to allow the structures to slide past each other more easily. NMRM was also run in a single step, letting both interfaces move at once. After perturbation, the original sequence was restored. A total of 15500 outputs were produced.

To reduce the runtime of the protocol, the NMRM source code (source/src/protocols/normalmode/NormalModeRelaxMover.cc) was modified. NMRM stores all computed conformations it creates during the course of a given trajectory but only returns the best scoring one (throwing away the rest). We updated the mover to allow retrieval of all computed conformations via a MultiplePoseMover in either a random or best-score-sorted order.

1.2 HBNet design

The region where HBNets can be designed into the RTR interface is smaller and empirically more difficult to find extensive HBNets at diverse positions. Therefore, HBNet design in the case of RTR was modified to require fewer HBNet residues to be in the core (2) and filtering was set up to require that at least one buried HBNet residue be on either side of the interface. max_unsat_Hpol was also increased to two. See production_varset.json for a complete list of parameters used.

HBNet produced 15692 outputs and we found 3815 unique networks. We kept either the best 3 or the best 10% of structures with each network, whichever was higher, for a total of 12649.

1.3 Interface design

The interface between Chain A and Chain B and the region 7.0 A around the HBNet residues, not including the HBNet residues themselves, were allowed to redesign, excluding

serine 154 which seemed to be important. Flexible backbone movement was allowed for the n-terminal two helices and loops on chain A and the four helices which are closest to chain A on chain B with all other residues (essentially the two helices on chain A and chain B which connect the interface to the rest of the ring and the two helices which connect the interface to the rest of the extension on chain B) were fixed in place with coordinate constraints and prevented from backbone minimization. Then flexible backbone FastDesign was carried out in cartesian space with HBNet residue AtomPair distance constraints, allowing at most 2 methionines, allowing at most 2 boundary tryptophans, and requiring between 1 and 3 boundary tyrosines. The same task operations were used as 1.3 for NM designs.

Next, the surface was separately redesigned with sequence constraints which require a very negative net charge to ensure any combination of interfaces downstream has a very low isoelectric point (pI). Designs were filtered to have mismatch_probability, core_holes, interface_sc, and ss_sc not worse than the parent structure, resulting in 880 unique outputs.

The minimal structures were then converted back into the full (tetrameric) ring through superimposition of the homology region helices with the corresponding regions of the full structure and splicing of the redesigned interface into the full structure using PyMol (Schrödinger). The symmetric structures were then cartesian minimized with HBNet constraints on with Rosetta to fix any bad backbone geometry from the splicing procedure, and finally they were flexible backbone relaxed without HBNet constraints.

1.4 Filtering

DeepAccNet[47] was used to predict the lddt of the designs post-relaxation, and any designs with residues with lddt < 90 were discarded. Designs which did not increase much in energy after flexible backbone relaxation without HBNet constraints (REU after - REU before < 0.5)

were also discarded, resulting in 79 designs which were subsequently visually inspected and narrowed down to 24 designs.

All interface design was previously carried out while disallowing cystine, but an interface disulfide seemed to have been important for high order structure assembly[40], so we used stapler (https://github.com/atom-moyer/stapler) to consider potential positions for disulfide bonds to exist on the selected designs, then chose positions which were as different as possible from the others in the chosen set. These were then ordered and characterized.

2. Computational recombination

2.1 hetBGLs

The process of combining multiple homo-oligomers into one hetero-oligomeric structure was carried out using a python script, recombine_2.1.py. It uses PyMol to superimpose and splice together homo-oligomer structures at the specified recombination point in the desired order, controlled through commandline arguments. The recombined structure could then be directly outputted, or optimized using Rosetta cartesian backbone minimization to fix any bad backbone geometry. Additionally, because local interface pI was not considered during the initial BGL design, some naive combinations could have individual chains with pI >= 8, which would result in neutrally or positively charged chains which are more likely to have expression and stability issues. Therefore, the surfaces of hybrid chains could also be redesigned to enforce a small pI for each chain individually in the final hetero-oligomers. Finally, if not present before, tyrosines were added to the surfaces to act as crystal contacts to aid crystallography.

2.2 hetRTRs

Recombination of RTRs into hetRTRs was done either using a version of the recombinase script which does not perform surface design, or with a text editor to simply combine the amino acid sequences to form the appropriate hybrid chains. In the latter case, design models were constructed manually using PyMol.

Experimental Methods

Buffer and media recipes

All buffers and media were made using Milli-Q filtered water.

Autoinduction media (TBM-5052)

1.2% [wt/vol] tryptone, 2.4% [wt/vol] yeast extract, 0.5% [wt/vol] glycerol, 0.05% [wt/vol] D-glucose, 0.2% [wt/vol] D-lactose, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, 2 mM $MgSO_4$, 10 μM $FeCl_3$, 4 μM $CaCl_2$, 2 μM $MnCl_2$, 2 μM $ZnSO_4$, 400 nM $CoCl_2$, 400 nM $NiCl_2$, 400 nM $CuCl_2$, 400 nM Na_2MoO_4, 400 nM Na_2SeO_3, 400 nM H_3BO_3

Lysis buffer

25 mM Tris, 300 mM NaCl, 20 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Wash buffer

25 mM Tris, 300 mM NaCl, 40 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Elution buffer

25 mM Tris, 300 mM NaCl, 300 mM imidazole, 100 mM EDTA, 10% glycerol, pH 8.0 at room temperature

Het Lysis buffer

25 mM Tris, 300 mM NaCl, 10 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Het Wash buffer

25 mM Tris, 300 mM NaCl, 20 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Het Elution buffer

25 mM Tris, 300 mM NaCl, 50 mM imidazole, 10% glycerol, pH 8.0 at room temperature

SEC buffer

25 mM Tris, 300 mM NaCl, pH 8.0 at room temperature

SAXS buffer

25 mM Tris, 300 mM NaCl, 2% glycerols, pH 8.0 at room temperature

TEV buffer

25 mM Tris, 100 mM NaCl, 0.5 mM EDTA, 1mM DTT, pH 8.0 at room temperature

AEC buffer A

25 mM Tris, pH 8 at room temperature

AEC buffer B

25 mM Tris, 1M NaCl, pH 8 at room temperature

Construction of synthetic genes

All synthetic genes were ordered from either Integrated DNA Technologies Inc. (Coralville, IA, USA) (IDT) or Genscript Inc. (Piscataway, NJ, USA). Genes ordered from IDT were reverse translated and codon optimized using Domesticator (https://github.com/rdkibler/domesticator_3). Protein tags were added to aid in purification (hexahistidine tag/histag) or identification (EHEE_rd2_0005 (aka EHEE)[36], superfolderGFP, or mScarlet-I) were separated from the designed protein with a tobacco etch virus protease (TEVp) cleavage site (ENLYFQG). In many cases, stop codons were introduced at the end of the inserted gene to prevent incorporation of the vector's built-in tags. Cloning into pET29b+ or its derivatives is always done at the NdeI/NcoI site.

For expression of BGLs, protomer sequences were extracted and the n-terminal tag "MGHHHHHHGSENLYFQGWS" was added. Codon-optimized genes with a 3' stop codon were cloned into pET29b+ by IDT. See Supplemental Table 1.1 for full amino acid sequences.

For co-expression of hetBGLs, all three proteins were arranged sequentially off the same transcript, separated by ribosome binding sequences (RBSs) (first: TAAGAAGGAGATATCATCATG; second: TAAAGAAGGAGATATCATATG). One protomer received superfolderGFP, another received EHEE, and the third received nothing. One protomer (with either superfolderGFP for the EHEE) received a histag. See Supplemental Table 1.3 for full amino acid sequences. These were ordered from IDT.

For expression of RTRs, amino acid sequences from Supplemental Table 1.4 were reverse translated and inserted into pET29b+ with a stop codon added before the end of the insert. These were ordered from IDT.

For co-expression of tetrameric hetRTRs, amino acid sequences from Supplemental Table 1.5 corresponding with the A2B2-hetRTRs were reverse translated and inserted into a pET29b+ vector variant which expresses the n-terminal superfolderGFP tag such that chain A was tagged with TEV-cleavable sfGFP. Chain A and B were separated by an RBS (TAAGAAGGAGATATCATCATG), and chain B received the vector's c-terminal histag. The sequences corresponding to A(5)2B(7)2 were prepared in the same way but inserted into the

standard pET29b+ vector. All C4 sequences were ordered from IDT and everything else was ordered from Genscript.

For (separate) expression of hexameric hetRTRs, the amino acid sequences from Supplemental Table 1.6 were reverse translated and inserted into pET29b+ using the vector's c-terminal histag. These were ordered from Genscript.

Protein expression

For single plasmid expressions, plasmids (100ng) were transformed into chemically competent E. coli expression strain BL21(DE3)Star (Invitrogen) for protein expression following manufacturer's protocol, with the exception of using 10ul competent cells per reaction. Following transformation and recovery, the entire transformation products were used to inoculate 1 mL Luria-Bertani (LB) medium containing 100 ug/mL kanamycin and grown at 37°C with shaking at 225 rpm overnight. 500ul of overnight cultures were diluted into 50 mL TBM-5052 supplemented with 100 ug/mL kanamycin in 250 mL baffled flasks, and incubated at 37°C with shaking at 225 rpm for 18-24 hours.

Protein purification

Immobilized metal affinity chromatography (IMAC)

Cultures were harvested by centrifugation at 4000 rcf for 10 minutes, culture supernatant decanted, and pellets resuspended to 30 mL in Lysis buffer. 300ul PMSF (100mM in 100% EtOH) is added immediately prior to sonication at 70% power for 5 minutes. "Lysate" fractions are saved, and then lysates were clarified by ultracentrifugation at 14,000 rcf at 12°C for at least 30 minutes and applied to 1.5 mL Ni-NTA resin (Qiagen) pre-equilibrated with Lysis buffer and packed into Econo-Pac columns (Bio-Rad) for gravity chromatography. The columns were

washed twice with 15 mL Wash buffer and eluted with 10 mL Elution buffer. Co-expressed hetero-oligomers were purified according to a similar procedure, except they used the “Het” variants of the the Lysis, Wash, and Elution buffers as this was found to improve the yield of single-his-tagged complexes over non-specific multiple-his-tagged complexes which can arise at high concentrations and likely dominate binding to the Ni-NTA resin. Samples prepared for crystallization were treated similarly, except 500 mL cultures were used and lysate was divided among six gravity columns.

Size-exclusion chromatography (SEC)

Samples were concentrated using 10k MWCO spin concentrators and were purified using a Superdex 200 10/300 increase column (Cytiva) in SEC buffer using an ÄKTA pure system (Cytiva). SEC traces were also used to qualitatively determine homogeneity and quantitatively measure total yield by A280 absorbance integrated over the collected fractions using Unicorn (Cytiva).

TEVp cleavage

Purification and mass tags were buffer exchanged into TEV buffer and cleaved with TEVp at a ratio of 1 mg TEV per 100 mg substrate for 24-72h at room temperature. After TEV cleavage, samples were exchanged into Lysis buffer and passed over a Ni-NTA gravity column and washed with 10ml lysis buffer. Flowthrough was collected, concentrated using 10k MWCO spin concentrators, and purified once again by SEC.

Anion exchange chromatography (AEC)

TEVp-cleaved samples of hetBGL03-15-18 intended for crystallization remained contaminated with GFP due to an oversight that meant GFP did not have a histag. To remove the GFP, samples were exchanged into AEC buffer A and loaded onto a HiScreen Q FF (Cytiva) column pre-equilibrated in AEC buffer A using an ÄKTA pure system . A gradient of AEC buffer B was applied with pauses at 15% (150 mM NaCl) and 25% (250mM NaCl) to elute the GFP and hetBGL03-15-18, respectively. Separation was measured by differential absorbance at 480 nm (GFP absorbance) and 280 nm, and SDS-PAGE.

Sample analysis

SDS-PAGE

Samples were diluted 1:1 with 2x Laemmli Sample Buffer (Bio-Rad) without Beta-mercaptoethanol and 15ul were loaded onto AnykD™ Criterion™ TGX™ Precast Midi Protein Gels (Bio-Rad). Ladder was 10ul of Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad). Gels were run at 300V for 18 minutes, then stained using an eStain™ L1 Protein Staining System (Genscript). Stained gels were imaged using a Chemidoc XRS+ (Bio-Rad)

Liquid chromatography mass spectrometry (LC-MS)

To identify the molecular mass of each protein and thus verify sample identity and integrity, intact mass spectra was obtained via reverse-phase LC/MS on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column, and subsequently deconvoluted by way of Bioconfirm using a total entropy algorithm.

Native mass spectrometry (nMS)

Oligomeric state of SEC-purified and LC-MS verified samples was analyzed by on-line buffer exchange MS in 200 mM ammonium acetate using a Vanquish ultra-high performance LC system coupled to a Q Exactive ultra-high mass range Orbitrap mass spectrometer (Thermo Fisher Scientific). A self-packed buffer exchange column was used (P6 polyacrylamide gel, BioRad). The recorded mass spectra were deconvolved with UniDec version 4.2+.

Size Exclusion Chromatography - Multi Angle Light Scattering (SEC-MALS)

IMAC and SEC purified samples were were analyzed by SEC-MALS in 20 mM Tris, 150 mM NaCl, pH 8 on a Superdex 200 10/300 column in line with a Heleos multi-angle 9 static light scattering and an Optilab T-rEX detector (Wyatt Technology Corporation). Data was analyzed using ASTRA (Wyatt Technologies) to calculate the weighted average molar mass (Mw) of the selected species.

Negative stain electron microscopy (nsEM)

SEC purified samples were diluted to ~0.005 mg/ml using SEC buffer immediately before application for 45s to glow discharged thick carbon film-coated 400 mesh copper grids (CF400-CU TH) (Electron Microscopy Sciences). Grids were then stained and dried immediately twice using 2% uranyl formate. Dried grids were screened on a 120 kV Talos L120C transmission electron microscope. The *E. Pluribus Unum* (EPU) (FEI Thermo Scientific) software was used for automated data collection. Data processing was carried out in CryoSPARC™ (Structura Biotechnology Inc).

Small angle X-ray scattering (SAXS)

TEVp-cleaved (and optionally AEC purified) samples were re-purified by SEC in SAXS buffer and concentrated using thoroughly washed 10k MWCO small spin concentrators; the flowthrough of concentration was used as blanks for buffer subtraction. Scattering measurements were performed at the SIBYLS 12.3.1 beamline at the Advanced Light Source as part of the HT-SAXS program. The Xray wavelength (λ) was 1.27 Å, and the sample-to-detector distance was 1.5 m, corresponding to a scattering vector q ($q = 4\pi \sin\theta/\lambda$, where 2θ is the scattering angle) range of 0.01 to 0.3 $Å^{-1}$. A series of exposures, in equal subsecond time slices, were taken of each well: 0.3 second exposures for 10 seconds resulting in 32 frames per sample. For each sample, data was collected for two different concentrations to test for concentration dependent effects; "low" concentration samples were ~2.5 mg/mL and "high" concentration samples were ~ 5 mg/mL. Data was processed using the SAXS FrameSlice online server (https://bl1231.als.lbl.gov/ran). FoXS[48] was used to compare design models to experimental scattering profiles and calculate quality of fit (χ) values. The SAXS Similarity online server was used to compute the similarities of scattering profiles to each other and calculate quality of fit (χ) values.

X-ray crystallographic analyses

Crystals of BGL06, BGL14_styr, BGL15, BGL18, and hetBGL03-15-18 were grown using protein purified as described above and TEVp cleaved and optionally AEC purified. Protein samples dispensed in 1 uL drops at purification concentrations were mixed with equal volume of a crystallization solution and set in hanging drops (refer to Supplemental Table 1.2 for conditions). Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same

crystallization solution resulting in growth of crystals. The crystals were flash cooled in liquid nitrogen after transfer into a cryoprotective solution (refer to Supplemental Table 1.2 for conditions). Diffraction data were collected on a Pilatus areas detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.2 for BGL06, BGL14_styr, BGL15, and BGL18. Diffraction data were collected on a Rigaku HyPix-6000HE hybrid photon counting detector at the Fred Hutchinson Cancer Center (FHCC) for hetBGL03-15-18. The resulting data sets (Supplemental Table 1.2) extend to 2.1 Å, 3.0 Å, 3.3 Å, 3.0 Å, and 2.1 Å resolution for BGL06, BGL14_styr, BGL15, BGL18, and hetBGL03-15-18, respectively. Most data had complete trimers within the asymmetric unit (three copies of a protein subunit), with exceptions to BGL18 and BGL15 which had 2 and 4 trimers in the asymmetric unit, respectively. Data was processed using the program HKL2000[49] or Aimless[50]. The placement of subunits was determined using the molecular replacement algorithm in program PHENIX[51]. Local rebuilding of all constructs was performed using the program COOT[52] , followed by refinement using the program PHENIX[51]. The final values for Rwork / Rfree are notated in Supplemental Table 1.2.

Chapter 2:

Design of T=4 de novo protein cages using pseudosymmetric hetero-oligomers

Abstract

Protein cages with a high triangulation (T) number are prevalent in Nature but they form only under icosahedral symmetry and are challenging to design because the components must present either multiple conformations or highly specific interactions. Here, we describe a stepwise design approach that starts from designing T = 1 cages with homo-oligomers under tetrahedral (T-sym), octahedral (O-sym) and icosahedral (I-sym) symmetries and expands them to T = 4 cages through pseudo-symmetrization of the oligomers. The formation of four components T = 4 cages with 48 subunits (T-sym), 96 subunits (O-sym) and 240 subunits (I-sym) are experimentally verified by cryo and negative-stain electron microscopy and match well with model structures. Diameters of these cages are 33nm, 43nm and 80nm, respectively. Our approach provides a general design route to expansion of protein cages using pseudosymmetric components.

Introduction

Protein cages composed of hundreds of subunits often emerge in nature as the form of virus capsids[53–59] and can also be computationally designed[60–65]. These cages are made by one or multiple unique components that are arranged under a point group symmetry in three dimensions, such as tetrahedral (T-sym), octahedral (O-sym) or icosahedral (I-sym) symmetry[55–57,59–63]. The size and number of subunits of the cages can be expanded following a geometrical principle[53], triangulation number (T), which is often defined (but not limited) under

icosahedral symmetry by the arrangement of hexameric and pentameric motifs. To design these cages with T > 1, there are several options. One is to design quasi-symmetric proteins that can adopt multiple conformations, as several viral capsid proteins do[53–55,58]. This requires solving a multi-state design problem which is generally known to be very challenging and is not yet well established[66–68]. Another route is to use multiple components with precisely defined interface combinations for a specific target T number cage. This requires designing hetero-oligomers with nearly symmetric structure (i.e. pseudo-symmetry) to minimize geometric discrepancy when they are assembled under the point group symmetry, but designed pseudo-symmetric oligomers are rare[69]. Recently, several high T number cages have been successfully made by AAB-type pseudosymmetric hetero-oligomers[70], but the use of AAB-type oligomers limits the interface specificity between oligomers, resulting in mixtures of cages with different T numbers. Furthermore, designed and naturally found high T number protein cages are limited to icosahedral symmetry; thus, it is desirable to develop a more general design approach for high T number cages with other point-group symmetries to expand the structural diversity of these cages and their applications.

We set out to develop a general design approach for T > 1 protein cages for tetrahedral, octahedral, and icosahedral symmetries. To this end, we developed a three-step hierarchical design strategy which leverages ABC-type pseudosymmetric hetero-oligomers to provide the precise interface specificity between components of T = 4 cages (Fig. 2.1). We first design T = 1 cages by docking trimeric homo-oligomers with a cyclic symmetry (C3) along the 3-fold symmetry axis of each symmetry (Fig. 2.1 left two columns)[60,64], which gives T3, O3 and I3 symmetric cages. In the second step, a part of each T = 1 cage is extracted by pseudosymmetrizing the homotrimers into ABC-type heterotrimers and removing the

trimer-trimer interface on one of the chains, which gives a higher order C3, C4 and C5 symmetric cyclic oligomers (crowns) (Fig. 2.1 third column). Lastly, the C3, C4 and C5 crowns are aligned along the 3-fold, 4-fold, or 5-fold symmetry axis of T-sym, O-sym and I-sym, respectively, and docked to an additional homotrimer aligned to the remaining 3-fold axis (Fig. 2.1 right column). This gives T = 4 cages for each cage symmetry, where a hexameric motif is located between edges of the crowns (*i.e.* $h = 2$ and $k = 0$ following the convention proposed by Caspar and Klug[53]).

BGL extension

Protein building blocks for this design approach should be able to dock to cages with various symmetries and be transformable to pseudosymmetric building blocks. To meet these requirements, we designed a structurally diverse set of C3 symmetric homo-oligomers that can be combined into pseudo-C3-symmetric hetero-oligomers. As base scaffolds, we chose the BGL family of trimers which have the same structure but with different amino acid sequences at the interface[69,71], and we fused diverse "arms" to these scaffolds to obtain structural diversity (Fig. 2.2A). We used designed helical repeat proteins (DHRs) as arms because they have various morphologies[72] and can be fused to the helical termini of the BGLs through HelixFuse[73]. This gives armed homo-oligomers about 10 nm in diameter depending on the size and the length of the DHRs. From the combinations of 13 BGLs and 37 DHRs, we computationally generated more than 2000 different armed BGLs and re-designed residues around the junction using Rosetta[74]. The junction designs were evaluated *in silico* based on a series of Rosetta metrics, including the number of helices interacting across the interface, buried surface across the junction and shape complementarity. Three armed versions of each of the 13 BGLs (*i.e.* total 39 designs, numbered from LK001 to LK039, see Supplemental Table 2.1) were chosen to test

experimentally, and we found that 37/39 are soluble under *E. coli.* expression and 33/39 showed a peak in size-exclusion chromatography (SEC) at a retention volume consistent with the size of the trimeric oligomers [of which 22 were monodisperse] (Supplemental Fig. 2.1). The oligomeric state was further confirmed by native mass spectrometry (nMS)[33], and 27/39 showed a signal at the correct mass of homotrimers (Supplemental Fig. 2.2). Structures of the oligomers were examined by small-angle X-ray scattering (SAXS)[34,35], and 17/18 had profiles matched well with those computed from the design models (Supplemental Fig. 2.3). Structural diversity of the armed BGLs was visually verified by negative-stain electron microscopy (nsEM), where different structural characters of DHR-arms (e.g. curved vs. straight) were identified from 2D class averages and 3D reconstructions (Fig. 2.2B,D, Supplemental Fig. 2.5). After the release of Alphafold2 (AF2) [75], we post-evaluated our oligomer designs, and from AF2 predictions, 7/17 (18/39 of the original set) showed good agreement to the design models (RMSD < 2.0 Å) and all but one were within 4.0 Å RMSD (Supplemental Figure 2.4). The high accuracy of the AF2 predictions was confirmed from the comparison to a crystal structure of LK031 (Fig. 2.2C). The model structure showed a slight mismatch at the arm junction from the crystal structure, resulting in a difference in the orientation of the arms (RMSD = 4.3 Å), but AF2 predicted the crystal structure more accurately (RMSD = 2.5 Å).

To design pseudosymmetric homo-oligomers, we used an interface transplantation approach to make an ABC-type hetero-oligomers with the same DHR arms, conserving C3 symmetry of the oligomers. This approach validated armed BGLs as a base scaffold (host) and transplants protomer-protomer interfaces from other BGLs (guest) (Fig. 2.2A), but conserving residues involving the junction where arms are fused; thus, additional sequence design is not needed. Before transplanting multiple different interfaces to a host scaffold at once, we first

check the compatibility between a host and a guest as a homo-oligomer by symmetrically transplanting the guest protomer interfaces into the host backbone. We selected seven armed BGL homotrimers (LK010, LK011, LK031,LK032, LK035, LK038, and LK039) as host scaffolds and four BGLs (BGL0, BGL17, BGL18, and BGL19) as guest interfaces because they showed high expression level and homogeneous oligomeric state. Among 28 possible combinations, we tested 18 samples in experiment and 14/18 showed a SEC peak at the correct oligomeric size and 11/18 showed a strong nMS signal at the correct mass of expected oligomers (Supplemental Fig. 2.2). After validating compatible combinations between a host scaffold and a guest interface, we chose LK032 as a host scaffold for testing hetero-oligomer transplantation (*i.e.* transplanting two different guest interfaces in a host scaffold) since it shows good compatibility with other guest interfaces. Five different combinations of hetero-oligomers (Supplemental Table 2.2) were experimentally tested. We assigned the different repeat number of the DHR-arm to each chain in order to give a significantly mass difference between chains and to visually confirm its hetero-oligomeric state under EM. We confirmed that two heterotrimers (LK059 and LK060) formed ABC-type heterotrimers as shown by SEC, SDS-PAGE, nMS (Fig. 2.2F, Supplemental Figure 2.6, Supplemental Figure 2.7). Additionally, both hetero-oligomers were visually verified to form homogenous species according to the presence of three different lengths of arm by 2D and 3D reconstruction of nsEM (Fig. 2.2F, Supplemental Figure 2.8).

Cage design

We first designed T = 1 cages with tetrahedral, octahedral and icosahedral symmetries using LK032 homotrimer, which is a base scaffold (host) of the pseudosymmetric LK059 and LK060. The homotrimer was computationally docked using RPXdock[76,77] under T3, O3 and I3 symmetry by sampling rotational and translational displacements along a 3-fold axis. We found

that its DHR arm provides good docking surfaces, where four helices interact across the docking interface, but each symmetry required different numbers of arm repeat units (1.5, 2.5 and 3.5 repeat units for T3, O3 and I3) due to the twist of the arm (Fig. 2.3B, 3F, 3G). Residues around the docking interface were designed using the recently developed ProteinMPNN sequence design neural network[78], and assembly to the target docking structure was validated with AF2. Experimentally, we tested 19 ProteinMPNN-designed LKs (LK063 - LK081) that show a backbone RMSD less than 2.0 Å between the design model and the AF2 prediction (Supplemental Table 2.3). For T3 systems (LK063 - LK069), all six showed a SEC peak at the expected retention volume (Supplemental Fig. 2.9), and we inspected them under nsEM and observed structurally homogeneous T3 cages (12 subunits, diameter: 13nm) (Fig. 2.3C). The 2D class averages and 3D reconstructed electron density map match well with the structure model (Fig. 2.3B,3C). For O3 systems (LK070 - LK077), 7/8 showed a sharp peak in SEC (Supplemental Fig. 2.9), LK071 and LK073 showed homogenous O3 cages (24 subunits, diameter: 20nm) under nsEM (Fig. 2.3F,G). Similarly, for I3 systems (LK078 - LK081), the formation of dodecahedron I3 cages (60 subunits, diameter: 40nm) was confirmed by nsEM (Fig. 2.3J,K) for LK078, but sometimes imperfect cages were also observed (Supplemental Fig. 2.10). We speculate that the staining and drying process for nsEM might affect the cage structure due to its large empty space inside the cage.

As a second step, we designed C3, C4 and C5 symmetric crowns. Instead of designing new docking interfaces for crowns, our design strategy is to reuse the docking interfaces verified from T = 1 cages by pseudo-symmetrizing the homo-oligomers forming the cages (Fig. 2.1 third column). The T = 1 cages were all designed based on LK032, thus the docking structures will be conserved even if LK032 is replaced to a LK032-based pseudo-symmetric heterotrimer.

Therefore, we replaced LK032 to LK060 and designed chain A (chA) and chB of LK060 to interact at the docking interface. This makes the arm of chC, which is designed to be soluble, point outward of the crown. The residues of chA and chB around the hetero-dimeric interface were designed using ProteinMPNN, and the designs were filtered by AF2 (RMSD < 2 Å). To avoid off-targets formed by self-interacting chains like chA-chA or chB-chB interface formation, we carried out additional AF2 tests with chA-only or chB-only systems and filtered out if at least one of the chains forms a self-interacting interface (Supplemental Fig. 2.M.3). We ordered genes for 19 sets of crowns that passed the AF2 filters (Supplemental Table 2.4), and the three chains of each set were expressed separately in independent *E. coli.* cultures. After growing, the amount of each protein was estimated by SDS-PAGE densitometry, and the cultures were mixed with a proper stoichiometry (chA:chB:chC = 1:1:1) for co-lysing and co-purification. In SEC, 4/5 (C3), 5/7 (C4) and 2/7 (C5) showed a peak at the expected crown size (Supplemental Fig. 2.11), and we identified the formation of crowns using nsEM (Supplemental Fig. 2.12). The 2D class average and 3D reconstructed electron density map matched well with the model structures of crowns (Fig. 2.3D,H,L).

Lastly, we designed T = 4 cages of tetrahedral, octahedral and icosahedral symmetries by docking the validated crowns to LK032 homo-oligomer (Fig. 2.1 last column). For docking, the C3, C4 and C5 crowns were aligned to 3-fold, 4-fold and 5-fold symmetry axes each symmetry, and LK032 was aligned to another 3-fold axis. Then, we generated docked structures by sampling rotational and translational displacements of the crowns and LK032 along the symmetry axis. This gives cage structures where the arms pointing outward from the crown (*i.e.* chain C of crowns) are docked to the arm of the new homotrimer. To find optimal docking interfaces, we conducted sampling by changing the length of arm and selected docking with the

highest RPXscore. Sequence of the new docking interface was designed using ProteinMPNN, and the assembly into the correct docking interface was validated using AF2 (RMSD < 2.0 Å). To minimize off-target formation, designs forming self-interacting interfaces in AF2 predictions were filtered out. We experimentally tested 14 sets of T = 4 cage designs that passed the AF2 filters (Supplemental Table 2.5). There are four unique chains in each cage (3 chains from a heterotrimer and one from a homotrimer), and each chain was independently expressed, mixed into a proper stoichiometry (1:1:1:1), co-lyzed and purified. From SEC, we collected fractions that contain large size assemblies (including the cages) and verified the formation of cages using nsEM. For T = 4 tetrahedral and octahedral cages, we observed 33 nm and 43 nm cages in diameter, respectively, and we confirmed that the 2D class averages and 3D reconstruction match well with the model structures (Fig. 2.4A-D). The T = 4 octahedral cage was further identified under cryoEM (Fig. 2.4G), and the reconstruction shows that the docking interface between crown and homotrimer is slightly twisted, resulting in C4 crowns slightly rotated counterclockwise around the 4-fold symmetry axis. Note that the tetrahedral and octahedral cages have concave shapes, which depends on the docking structure between oligomers, but they are topologically identical with T = 4 polyhedra (Supplemental Fig. 2.13). For T = 4 icosahedral cages, we observed cages homogenous in diameter (80 nm) in nsEM, which matches well with the model (Fig. 2.4E,F), but the structural quality of the cages was not good enough to clearly identify the arrangement of hexagonal and pentagonal motifs. Due to the large size of the cage (and large empty space), the staining and drying process for nsEM are likely to affect the quality of the structure, but we also speculate that the longer arm (4 repeats) of the icosahedral cages than that of tetrahedral and octahedral cages (3 repeats) might cause larger strain when a structural mismatch occurs at the docking interface.

Discussion

Our stepwise design approach using pseudo-symmetric hetero-oligomers provides a route to design four components T = 4 protein cages with tetrahedral, octahedral and icosahedral symmetries that have never been observed in nature nor designed. Structurally diverse homo-oligomers can be designed through helical fusion and pseudo-symmetrized into ABC hetero-oligomers that provide interface specificity for assembling a specific target T number cage. The design approach described here is not limited to T = 4 cages but can be extended to designing other T number cages. If structurally diverse and orthogonal pseudo-symmetric hetero-oligomers are given, T = 3, 7 or even higher T number cages can be designed using the same design principle. Therefore, we believe that more diverse and higher T number protein cages can be designed in the near future as various design strategies for pseudo-symmetric hetero-oligomers are increasingly reported recently[69]. We expect that the design approach and successfully designed T = 4 protein cages presented here will be a milestone toward designing unprecedented large and structurally complex protein cages with hundreds or even thousands of subunits.

Figures

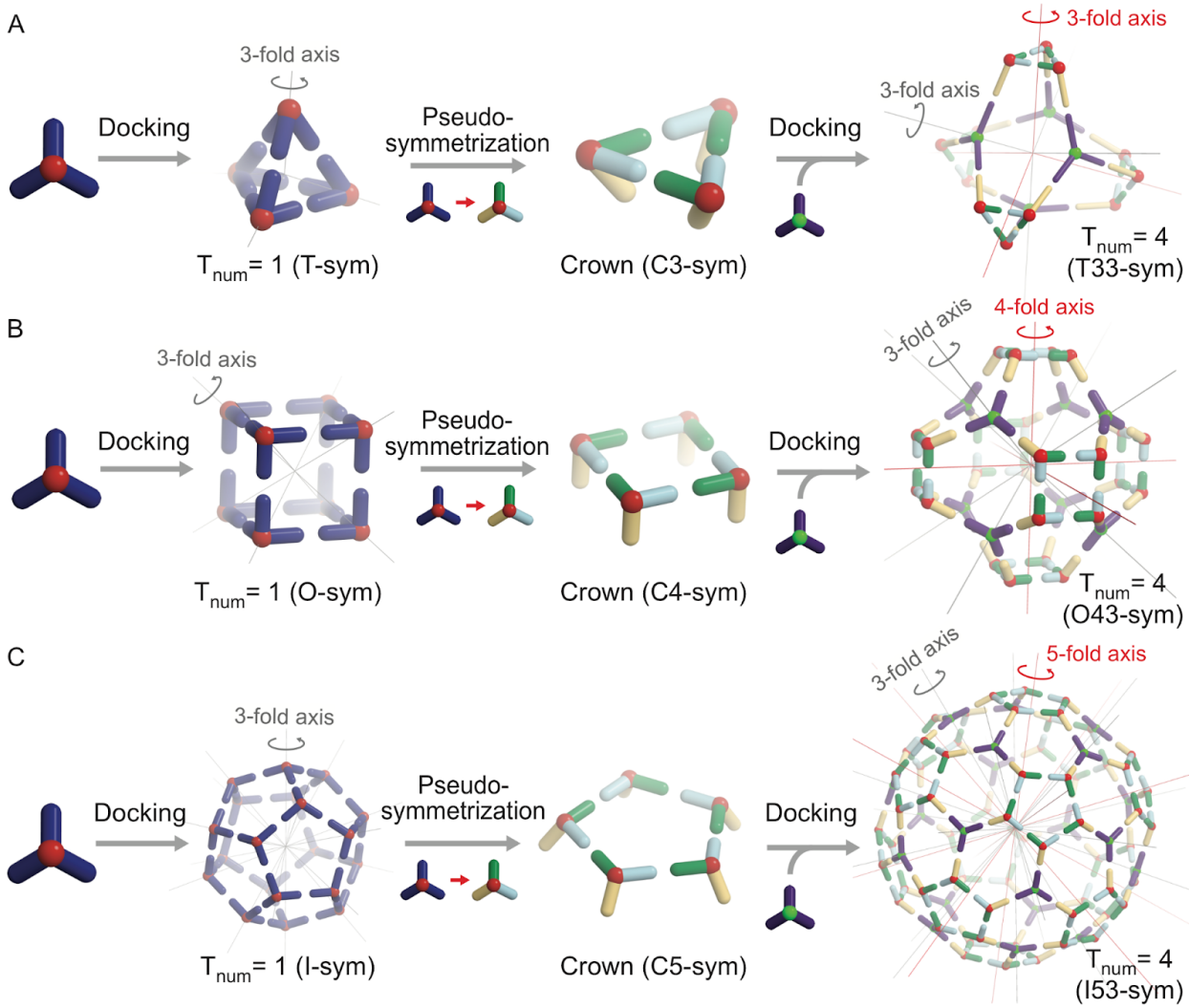


Figure 2.1: Overview of the stepwise design of T_{num} = 4 protein cages.

(**A**) C3 cyclic homo-oligomers are docked and designed into a T=1 tetrahedral cage. Replacing the homotrimers to ABC-type pseudosymmetric heterotrimers produces a C3 crown that has the same structure with a part of the T_{num} = 1 cage. The crown is docked to an additional homotrimer under T33 symmetry, which produces a T_{num} = 4 tetrahedral symmetric cage. Same design principle can be applied to (**B**) octahedral and (**C**) icosahedral symmetries.

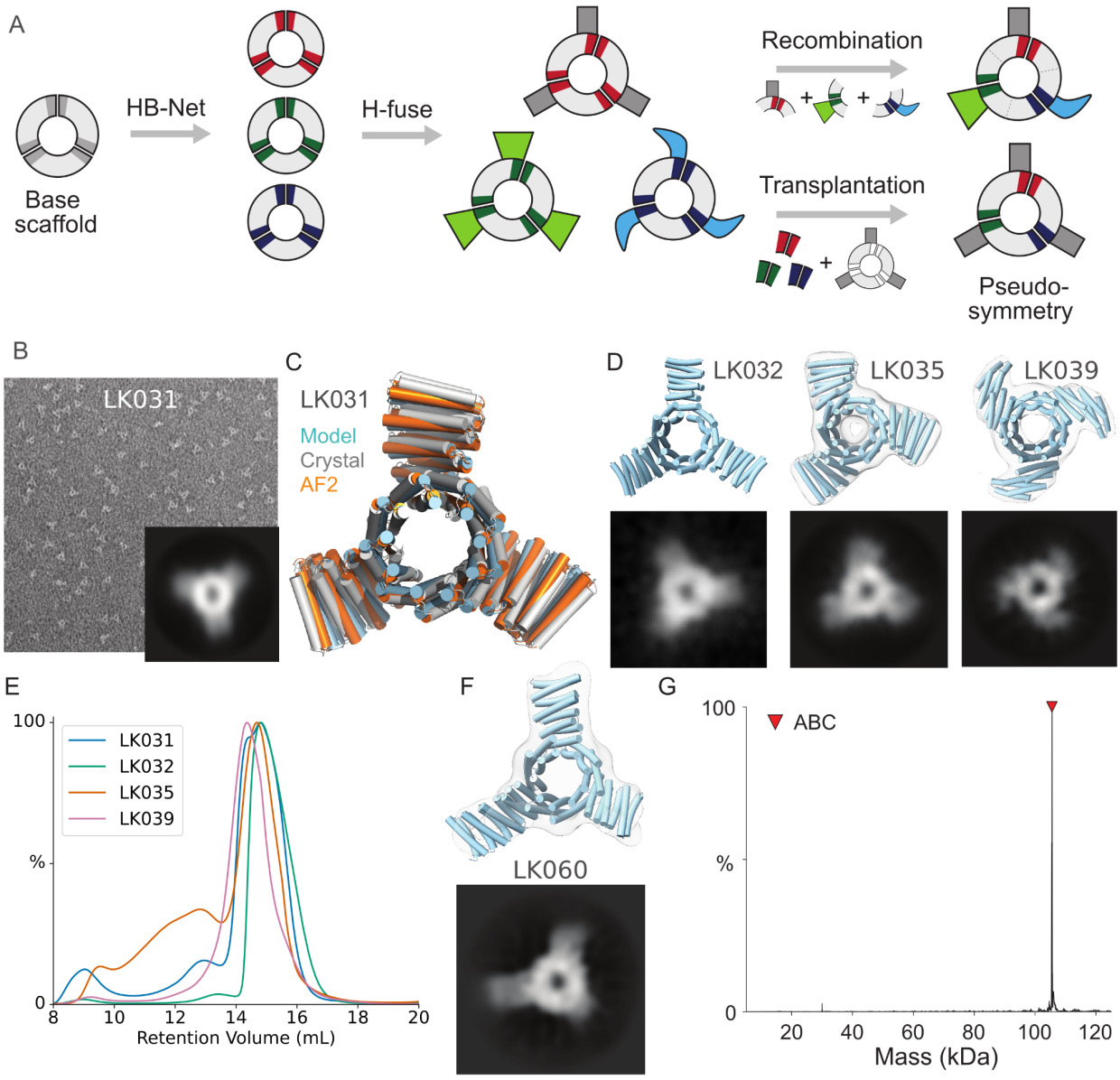


Figure 2.2: Structural diversification of homo-oligomers and their pseudo-symmetrization.

(**A**) Schematic illustration of design strategy for homo-oligomers with arms and pseudo-symmetric hetero-oligomers. (**B**) nsEM image of LK031 homotrimer and a 2D class average (inset) along 3-fold symmetry axis. Scale bar is 100nm. (**C**) Structure of LK031 solved by X-ray crystallography (white) compared to its design model (green) and AF2 prediction

(orange). (**D**) Top row: superpositions of the 3D reconstructed electron density map (gray cloud) to the structure model (blue) of LK032 (left), LK035 (middle) and LK039 (right). Bottom row: a 2D class average along a 3-fold symmetry axis. (**E**) SEC results of homotrimers. (**F**) ABC-type pseudo-symmetric hetero-oligomer (LK060) with same type of DHR arms but different lengths. Top: superposition of the 3D reconstructed electron density map (gray cloud) and model (blue). Bottom: a 2D class average along (pseudo-) 3-fold symmetry axis. (**G**) nMS result of LK060.

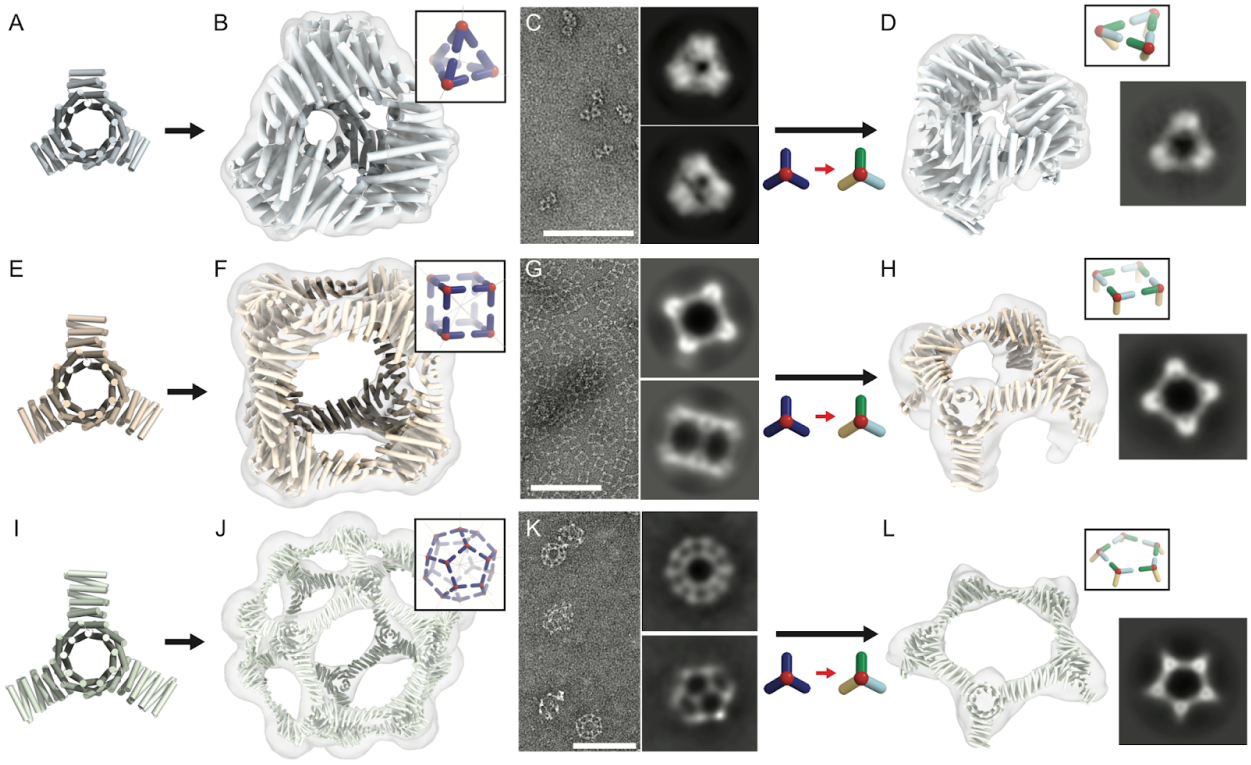


Figure 2.3: T=1 protein cages and cyclic crowns.

LK032 with (**A**) 1.5, (**E**) 2.5 and (**I**) 3.5 repeat units of arms are docked into (**B,C**) tetrahedral, (**F,G**) octahedral and (**J,K**) icosahedral T=1 cages. (**B,F,J**) Superpositions of 3D reconstructed electron density map (gray cloud) to the structure model of cages (colored). (**C**,**G**,**K**) (left) Negative stain electron micrographs and (right) characteristic 2D class averages of the cages. Scale bars are 100 nm. (**D**, **H**, **L**) C3, C4 and C5 crowns made by pseudosymmetric hetero-oligomers. (left) Superpositions of 3D reconstructed electron density map and the model crowns. (right) 2D class averages along 3-fold, 4-fold and 5-fold symmetry axes. The diameter of crowns is 11nm (C3), 20nm (C4) and 35nm (C5).

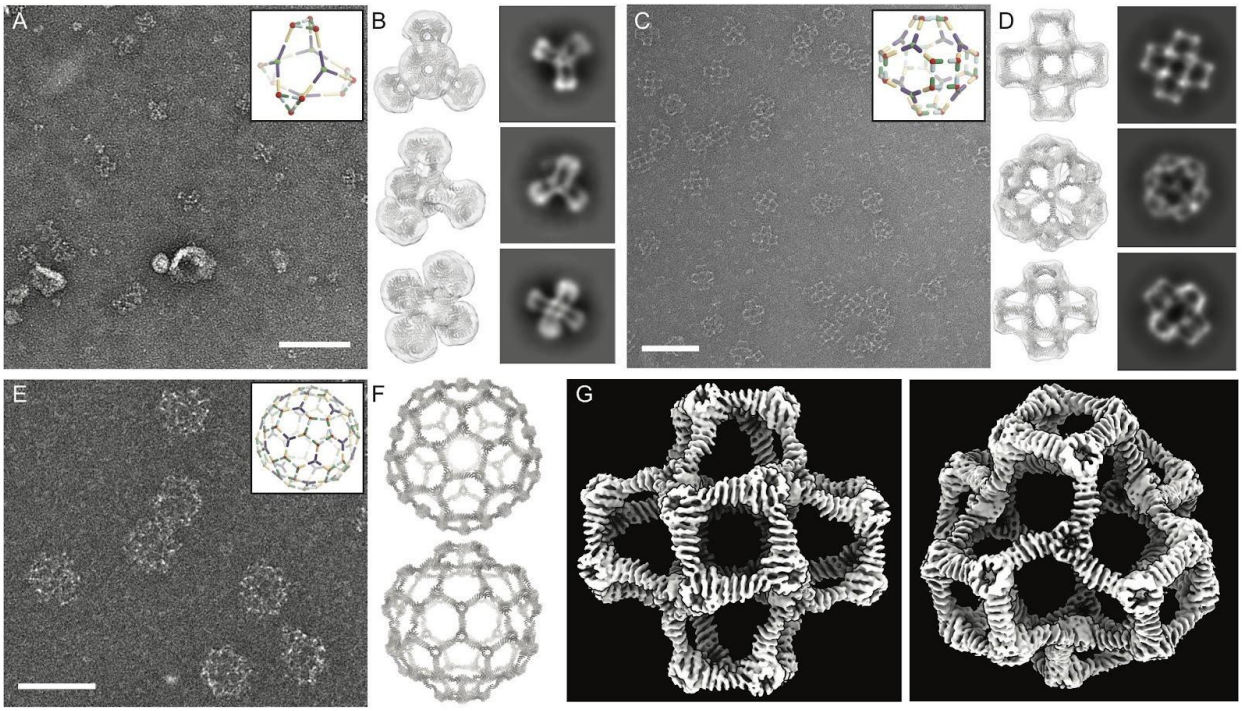
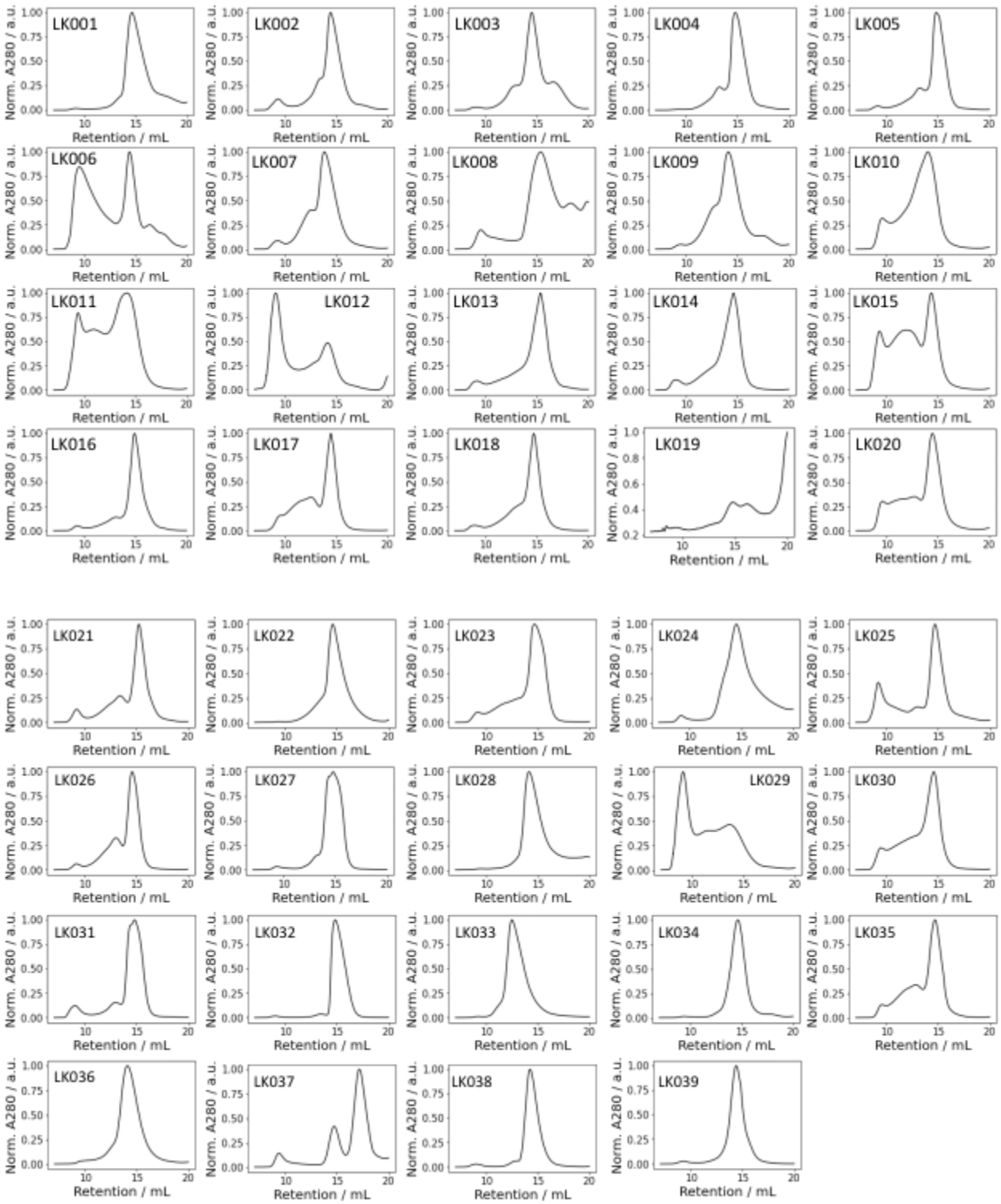
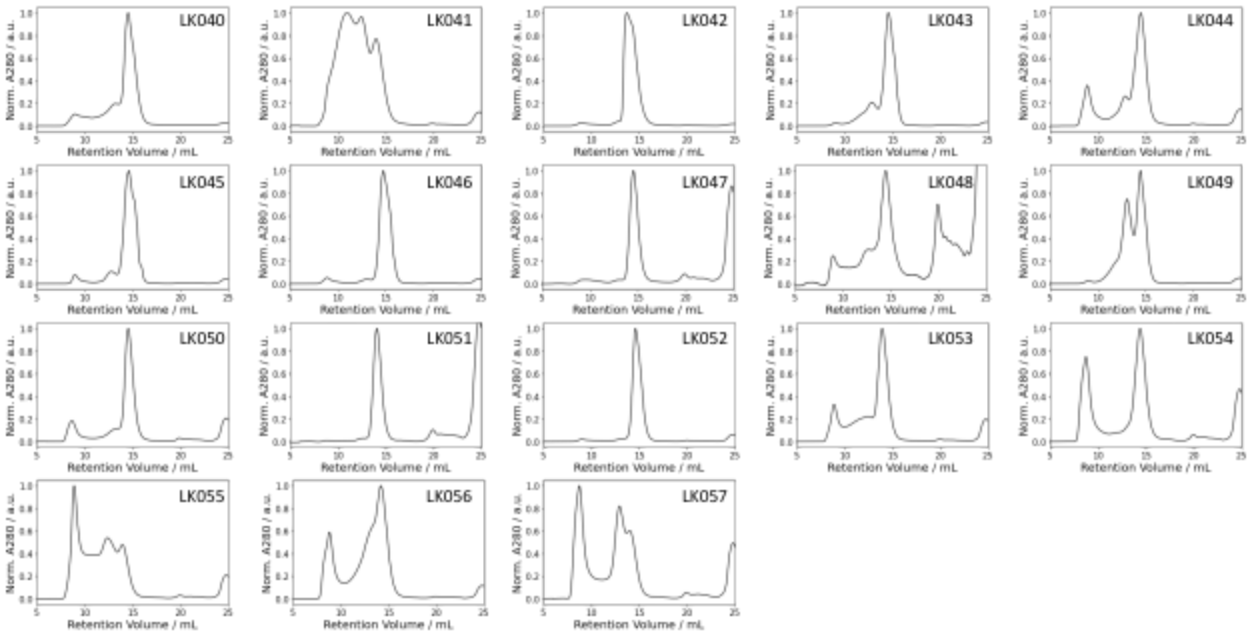


Figure 2.4: T=4 protein cages.

(**A, C, E**) T=4 protein cages with (**A**) tetrahedral, (**C**) octahedral and (**E**) icosahedral symmetry characterized by nsEM. Scale bars are 100 nm. (**B, D**) (left) 3D reconstructed electron density map (gray cloud) superpositioned to a structure mode (colored), and (right) characteristic 2D class averages. (**F**) Structure model of T=4 icosahedral cage. T=4 protein cage. (**G**) Cryo-EM characterization of T=4 octahedral cages viewed along (left) 4-fold and (right) 3-fold symmetry axes.

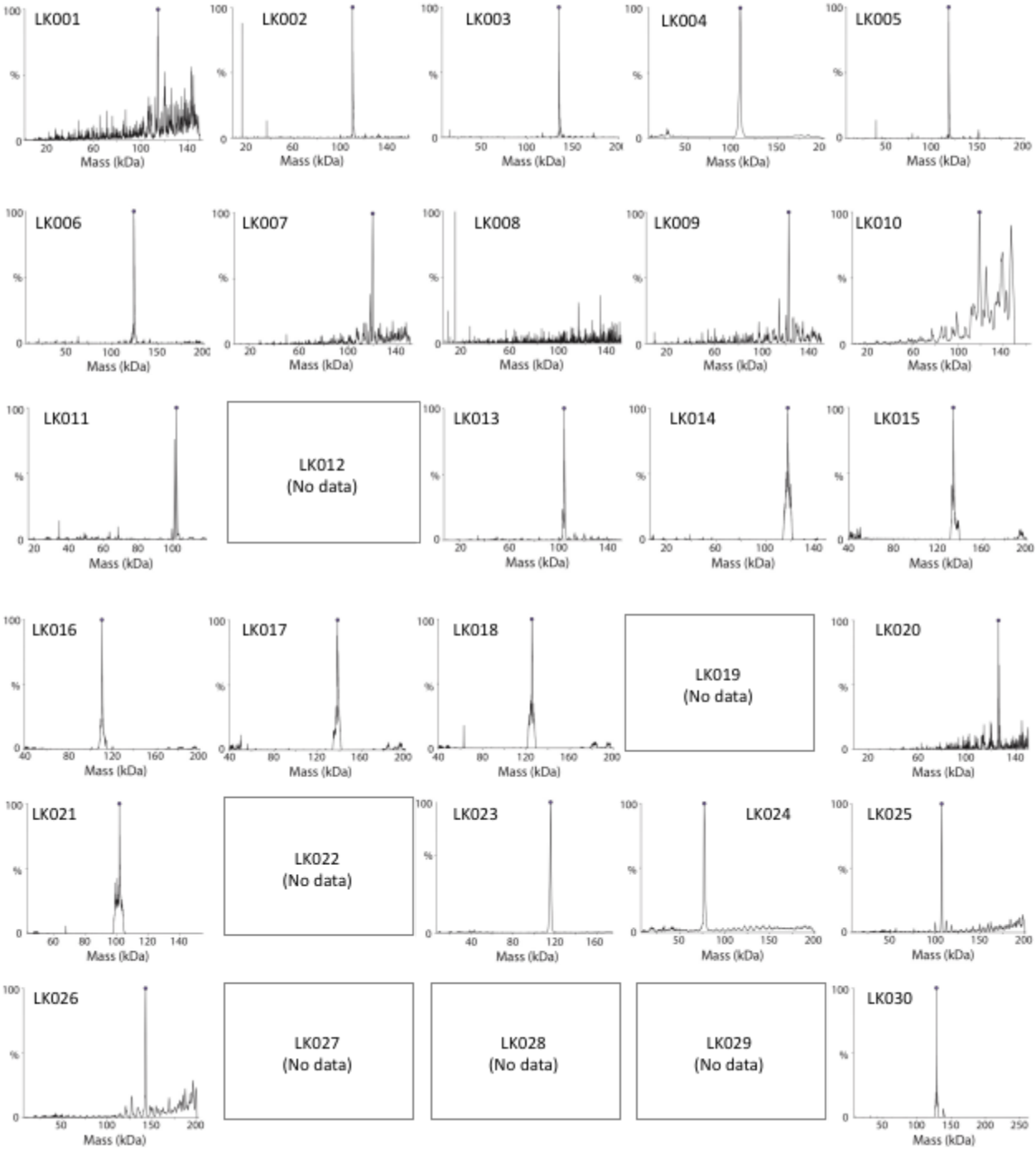
Supplemental Data

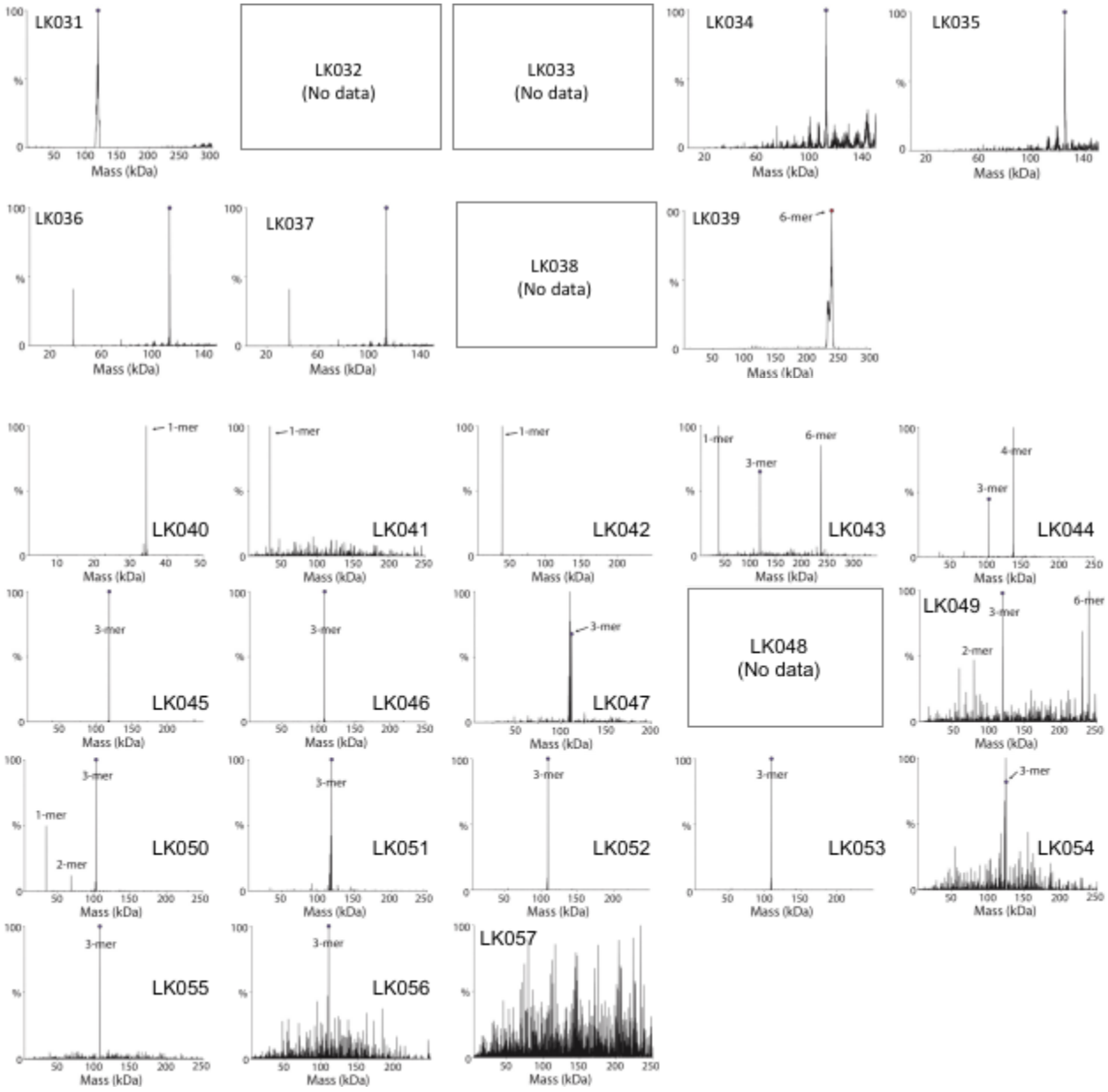




Supplemental Figure 2.1: SEC analysis of homotrimers

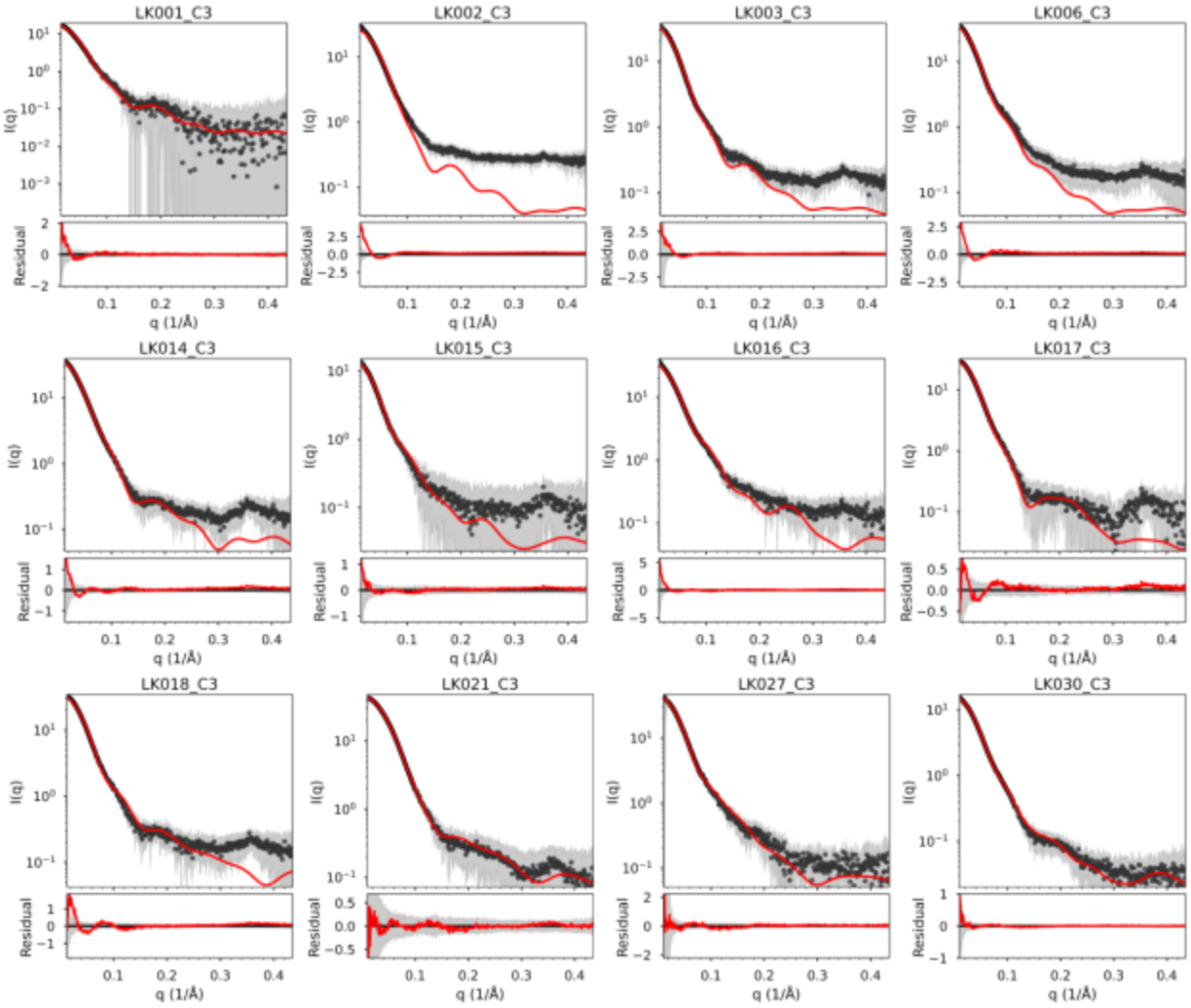
SEC traces of homotrimers (LK001 – LK039) obtained using S200 column.

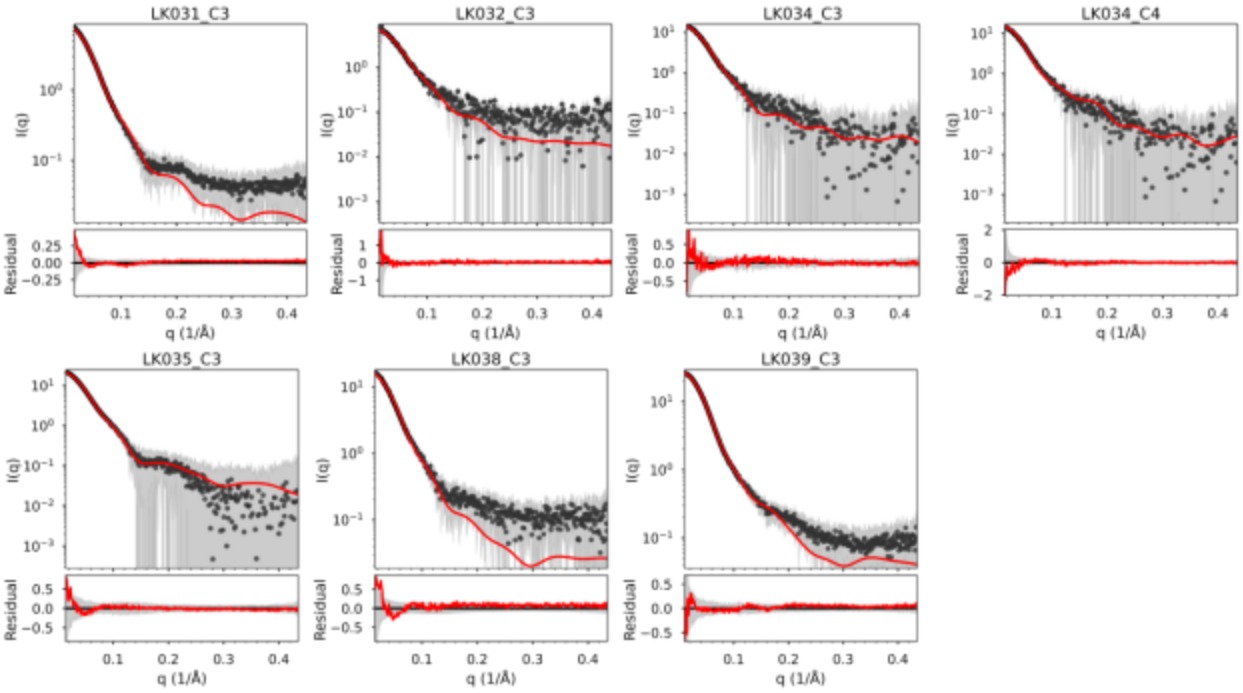

LK001
LK002
LK003
LK004
LK005
LK006
LK007
LK008
LK009
LK010
LK011
LK012
(No data)
LK013
LK014
LK015
LK016
LK017
LK018
LK019
(No data)
LK020
LK021
LK022
(No data)
LK023
LK024
LK025
LK026
LK027
(No data)
LK028
(No data)
LK029
(No data)
LK030
Mass (kDa)



Supplemental Figure 2.2: nMS of homotrimers

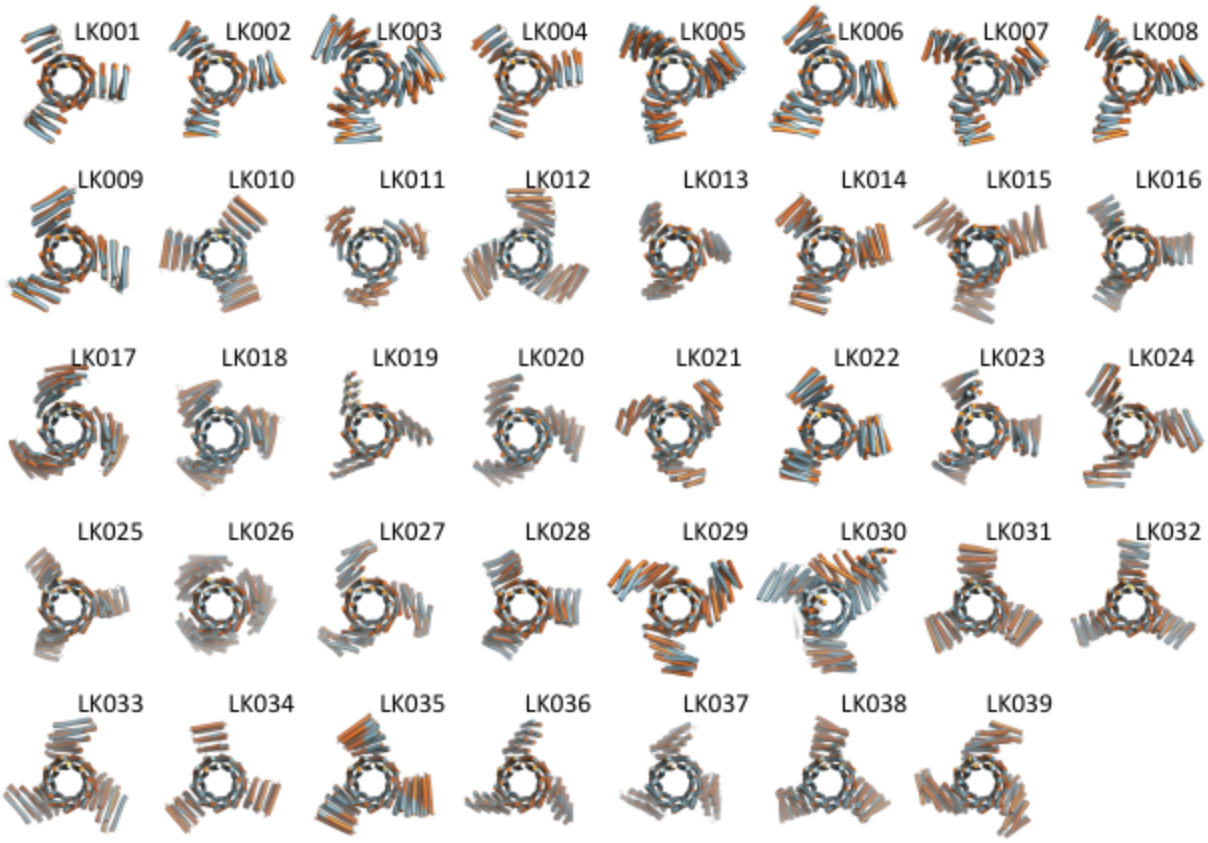
Plotting deconvoluted mass spectra. Signals for trimers are marked with a small dot.


LK001_C3
LK002_C3
LK003_C3
LK006_C3
LK014_C3
LK015_C3
LK016_C3
LK017_C3
LK018_C3
LK021_C3
LK027_C3
LK030_C3
I(q)
Residual
q (1/Å)



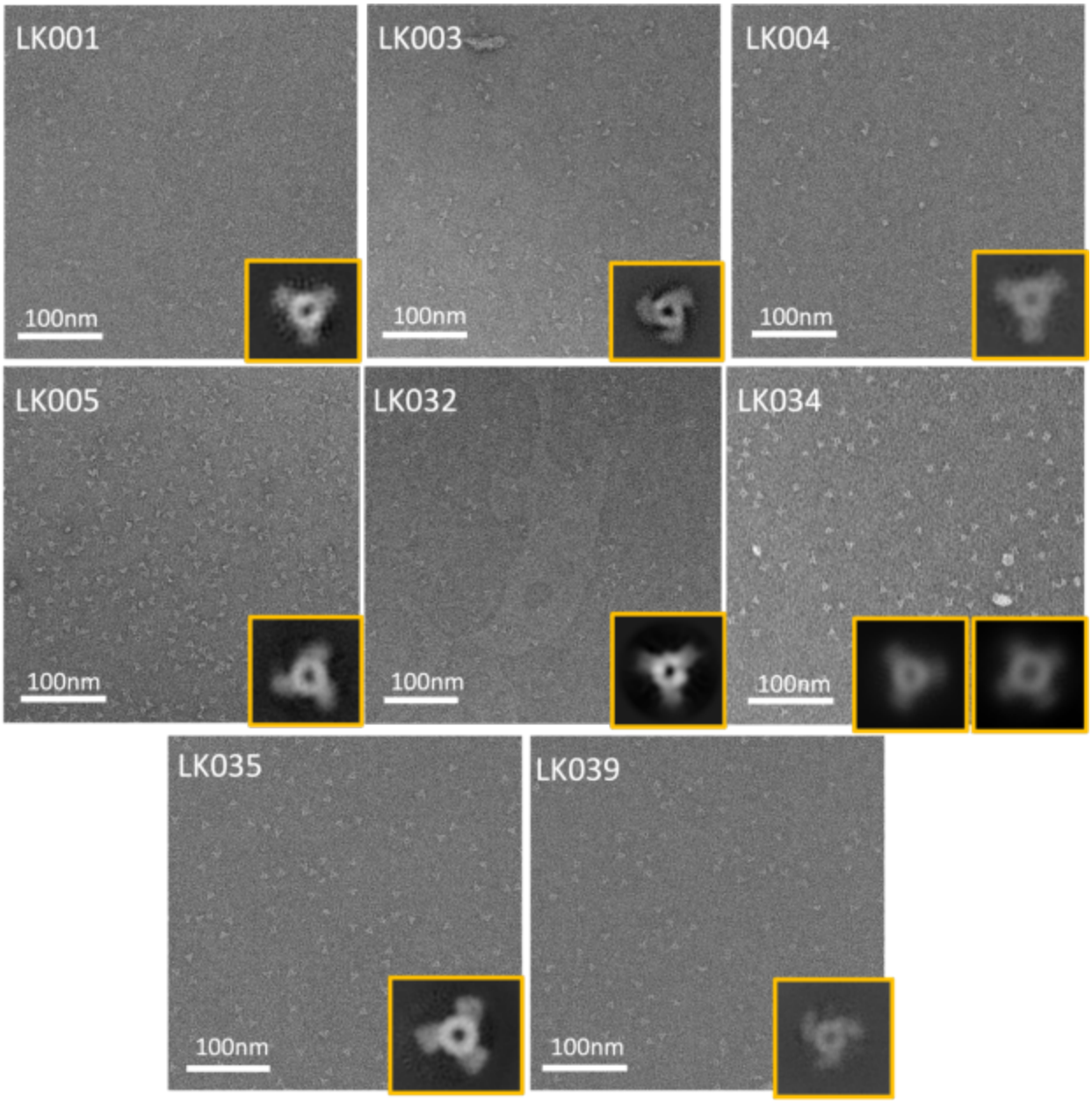
Supplemental Figure 2.3: SAXS of homotrimers

For each system, top panel is log scattering intensity (red: calculated from model, black dots: experimental data, grey: standard deviation of experimental data) and bottom panel is residual of the model fit to the data shown in linear scale.



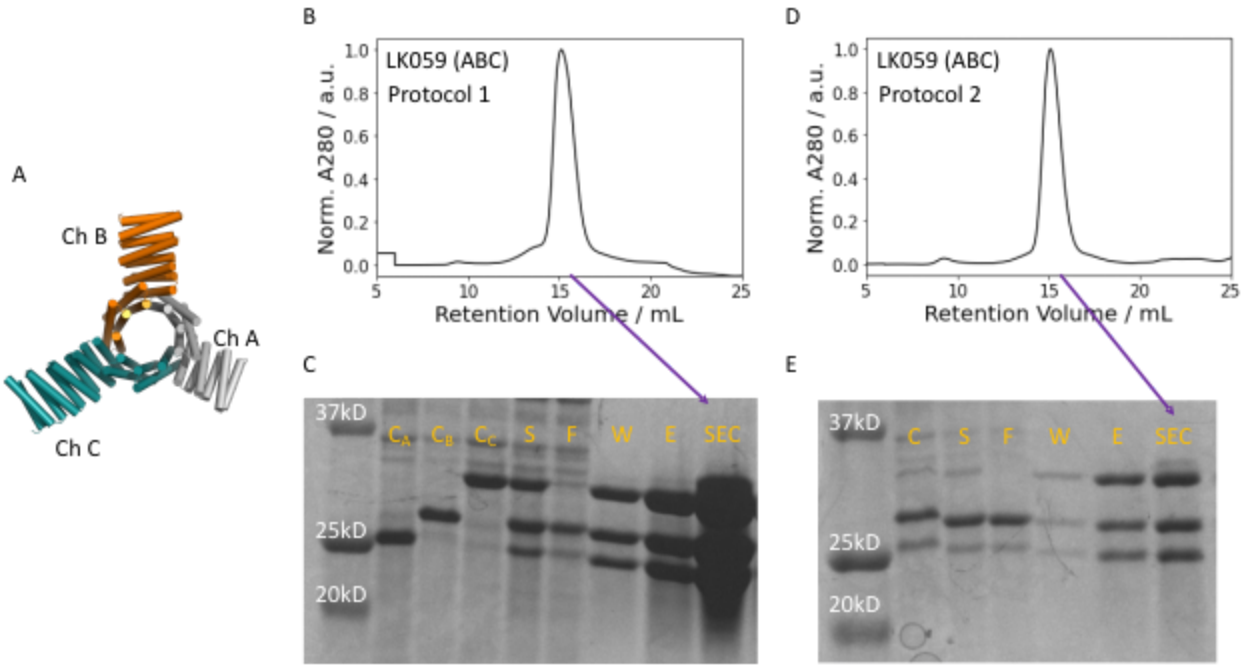
Supplemental Figure 2.4. AF2 analysis of homotrimers

Overlays of model structure (light blue) of homotrimers (LK001 – LK039), and their AF2 predictions (orange).



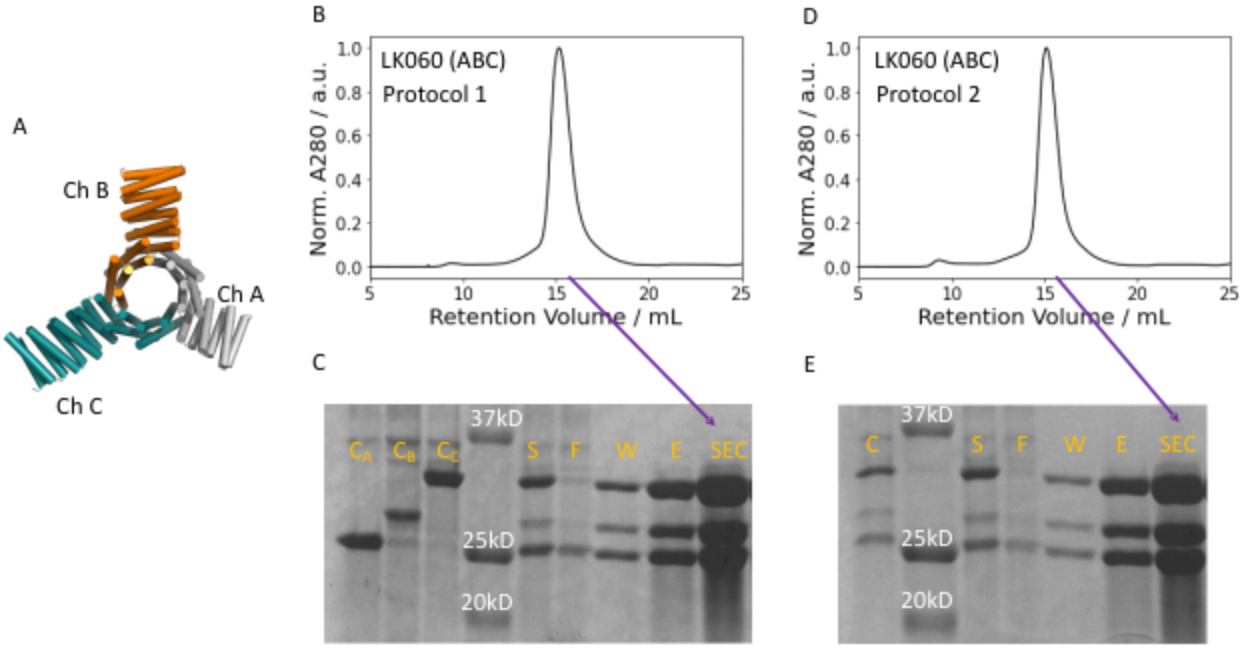
Supplemental Figure 2.5. nsEM analysis of homotrimers

Negative stained electron micrographs (nsEM) of homotrimers (LK001, LK003, LK004, LK005, LK032, LK034, LK035, LK039). Insets are 2D average classes along 3-fold symmetry axis of the oligomers. Note that LK034 is a mixture of C3 and C4 (off-target) oligomers.



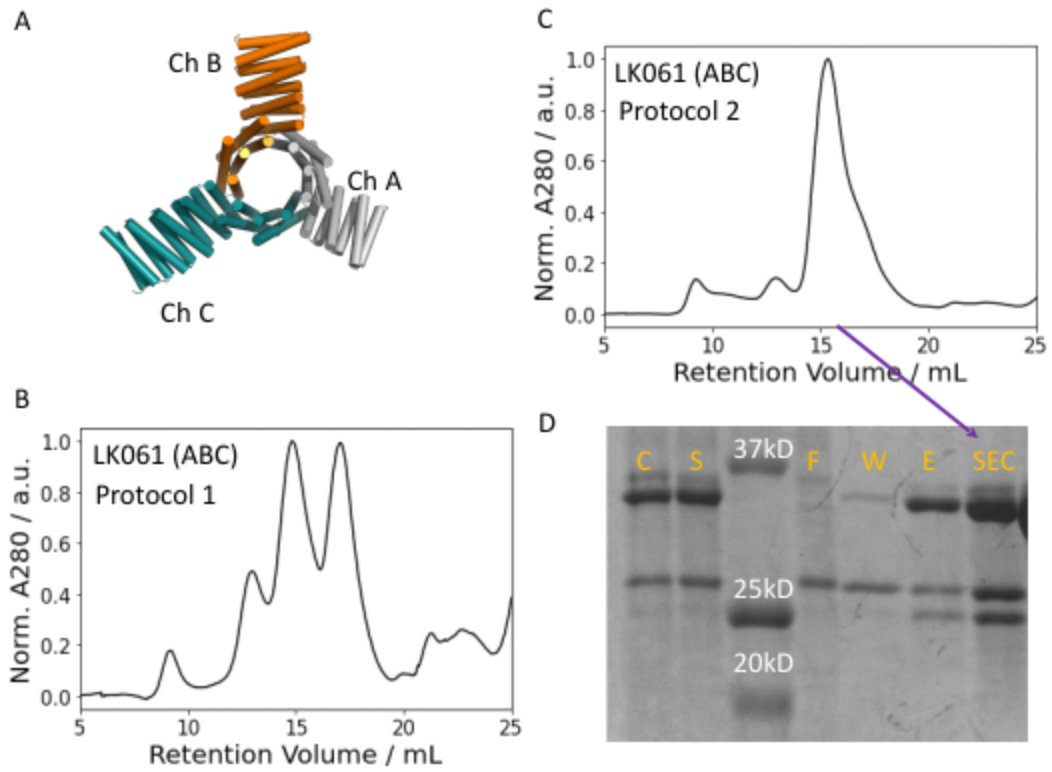
Supplemental Figure 2.6.1: SEC of heterotrimers (LK059)

(A) Model structure. (B) SEC trace and (C) SDS-PAGE results from protocol 1. (D) SEC trace and (E) SDS-PAGE results from protocol 2. (C, E) C: E. Coli culture. S: Soluble fraction. F: flow through from IMAC. W: Washed fraction. E: Elution. SEC: SEC peak. C_x indicates E. Coli culture of chain-X before mixing with other chains



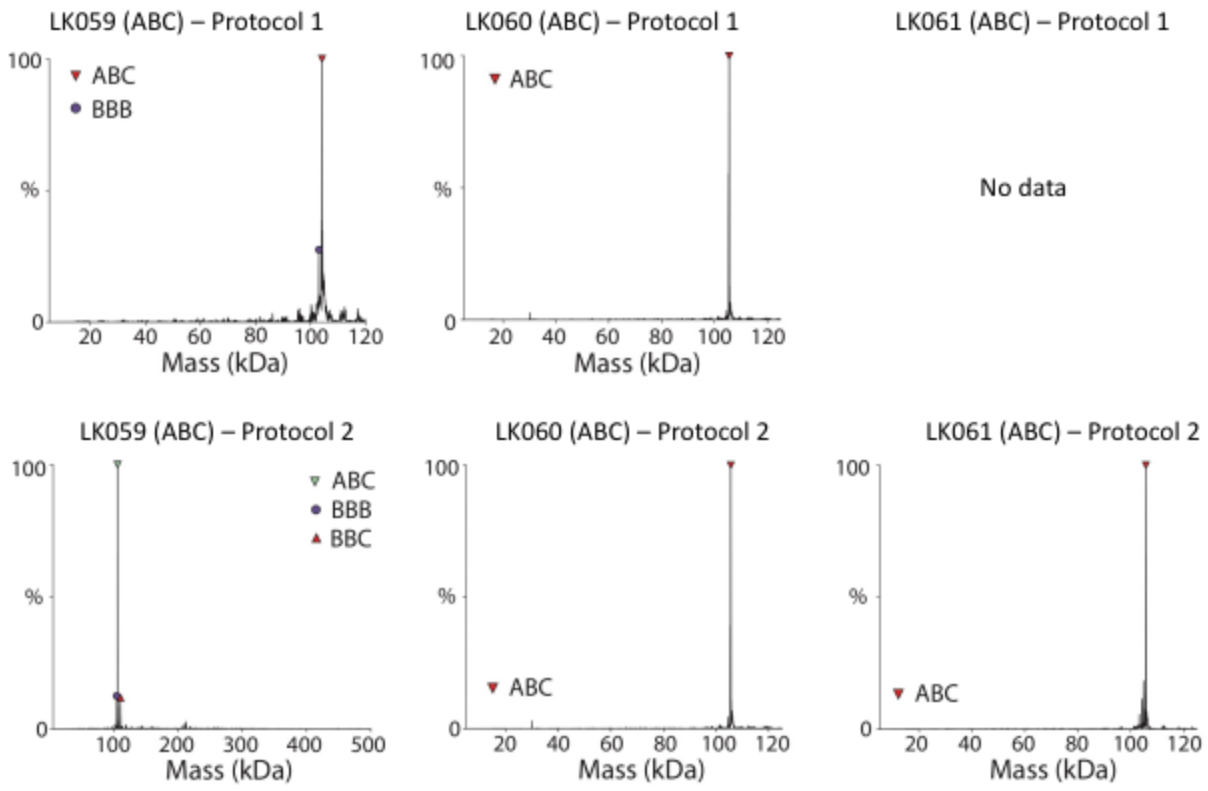
Supplemental Figure 2.6.2: SEC of heterotrimers (LK060)

(A) Model structure. (B) SEC trace and (C) SDS-PAGE results from protocol 1. (D) SEC trace and (E) SDS-PAGE results from protocol 2. (C, E) C: E. Coli culture. S: Soluble fraction. F: Flow through from IMAC. W: Washed fraction. E: Elution. SEC: SEC peak. C_X indicates E. Coli culture of chain-X before mixing with other chains.



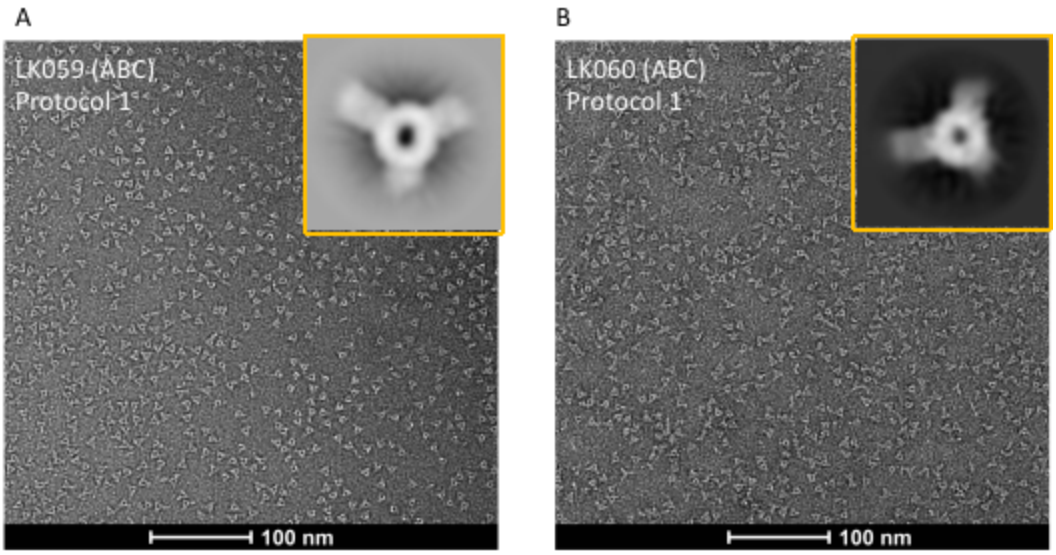
Supplemental Figure 2.6.3: SEC of heterotrimers (LK061)

(A) Model structure. (B) SEC trace from protocol 1. (C) SEC trace and (D) SDS-PAGE results from protocol 2. (D) C: E. Coli culture. S: Soluble fraction. F: Flow through from IMAC. W: Washed fraction. E: Elution. SEC: SEC peak.



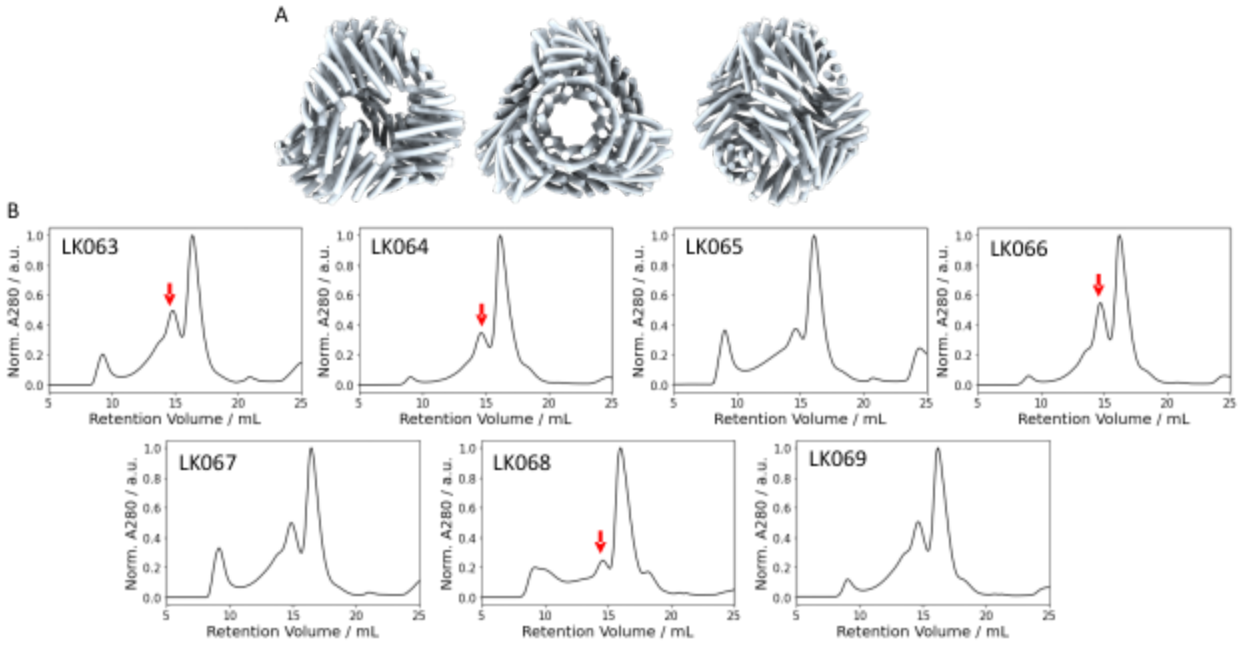
Supplemental Figure 2.7: nMS of heterotrimers

Native mass spec results of pseudosymmetric heterotrimers (LK059, LK060, LK061) made from (top row) protocol 1 and (bottom row) protocol 2.



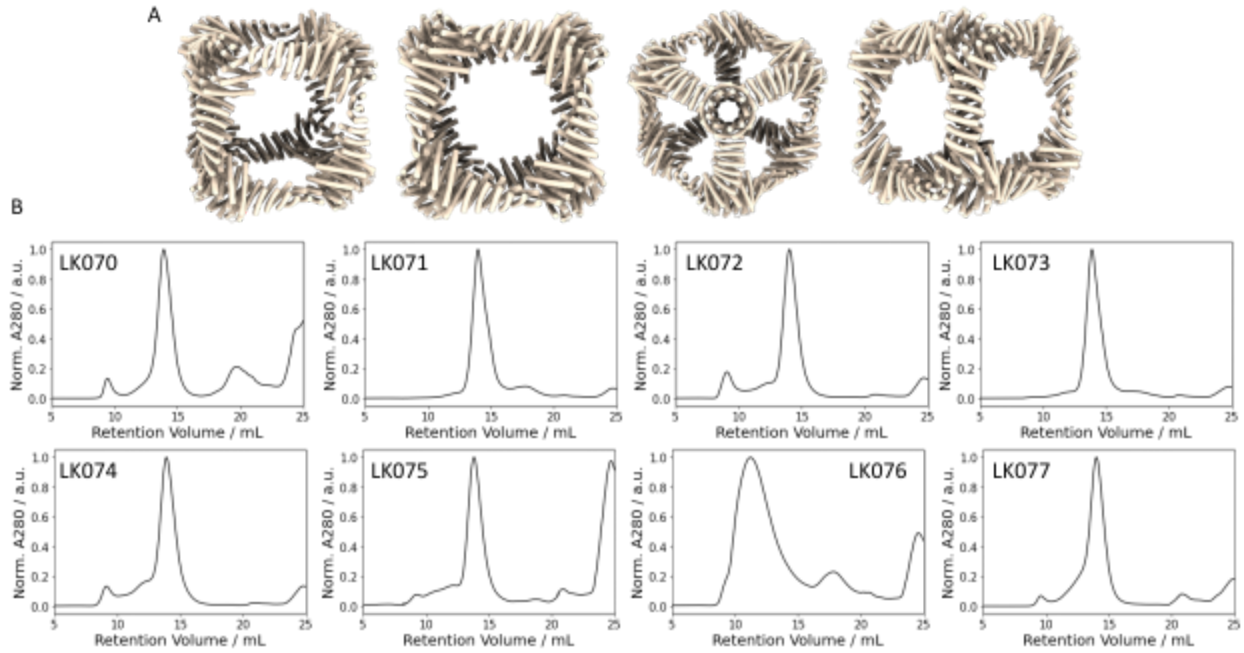
Supplemental Figure 2.8: nsEM of heterotrimers

nsEM of pseudosymmetric heterotrimers (LK059 and LK060) obtained from protocol 1. Insets are 2D average classes along 3-fold symmetry axis of the oligomers.



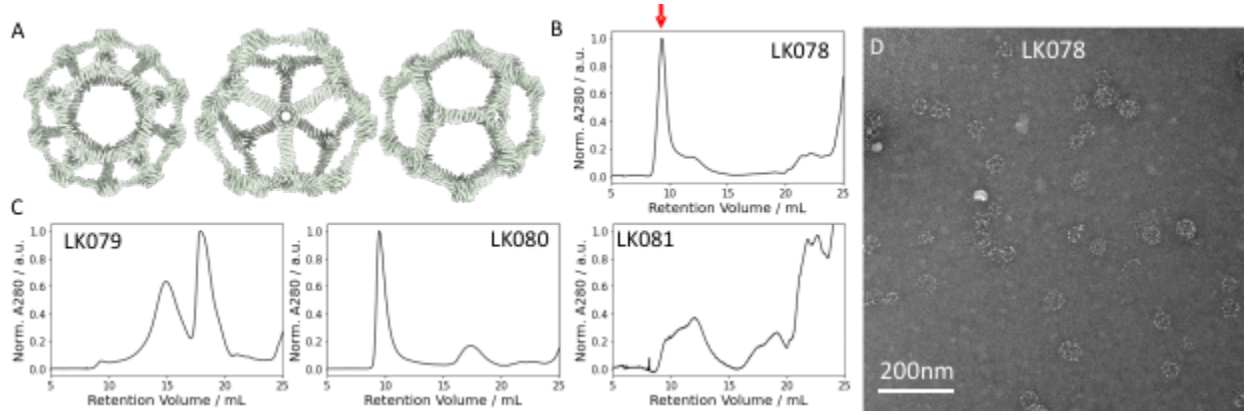
Supplemental Figure 2.9.1: SEC of T=1 tetrahedral cages

(**A**) Structure model of T=1 tetrahedral cages (symmetry: T3). (**B**) SEC traces of the cages (LK063 – LK069) obtained from S6 column. Transparent grey box (~ 15ml) of each trace indicates an expected retention volume of the cage. Red arrows indicate where nsEM in Fig. S2.10 were obtained.



Supplemental Figure 2.9.2: SEC of T=1 octahedral cages

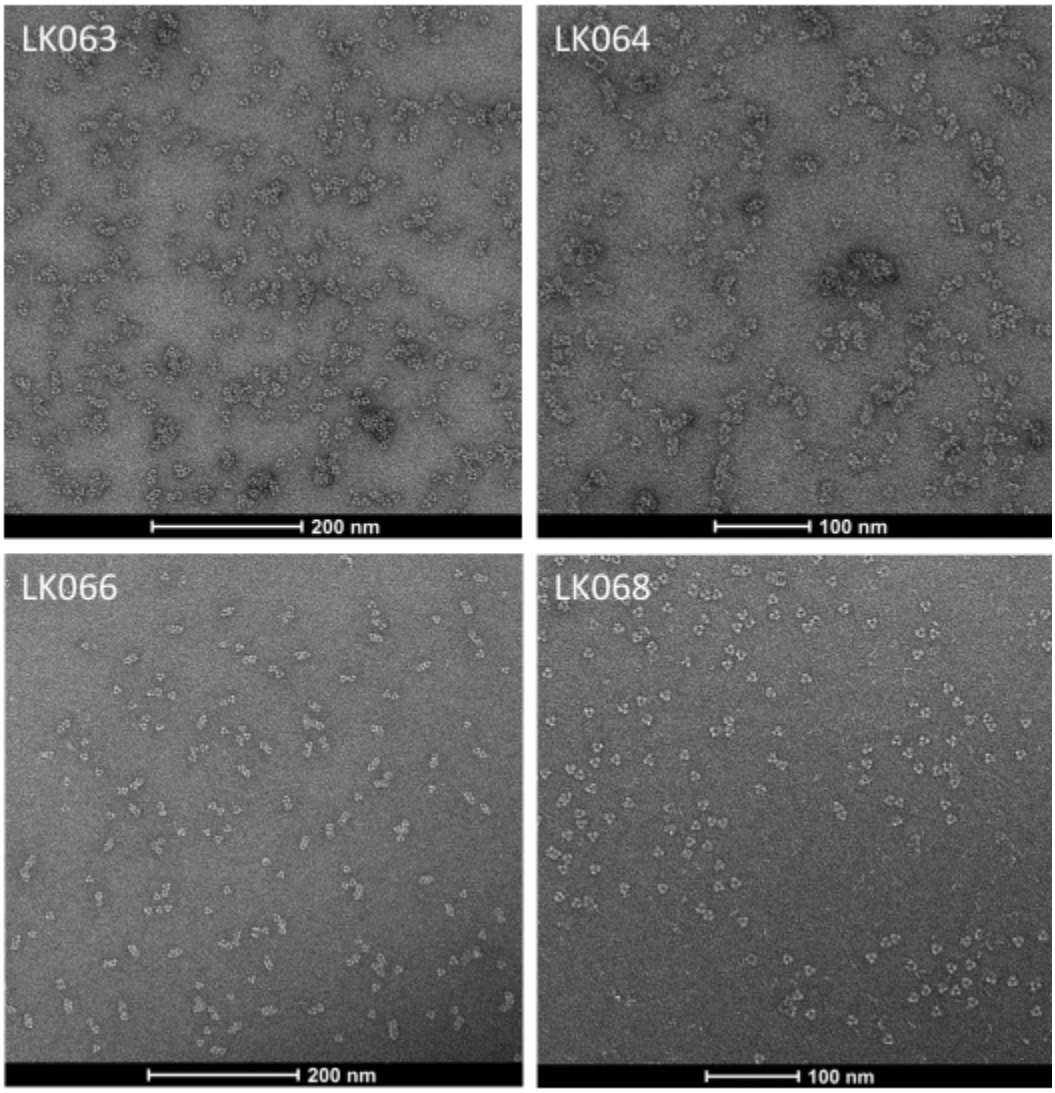
(**A**) Structure model of T=1 octahedral cages (symmetry: O3). (**B**) SEC traces of the cages (LK070 – LK077) obtained from S6 column. Transparent grey box (~ 13ml) of each trace indicates an expected retention volume of the cage.



Supplemental Figure 2.9.3: SEC and EM of T=1 tetrahedral cages

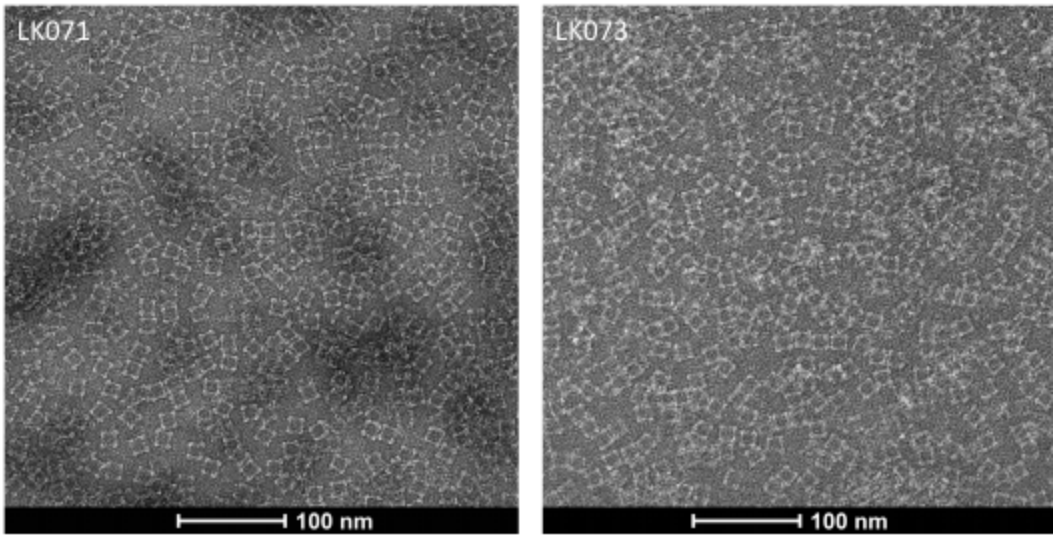
(**A**) Structure model of T=1 icosahedral cage (symmetry: I3). (**B, C**) SEC traces of the cages (LK078 – LK081) obtained from S6 column. Transparent grey box (~ 9ml) of each trace

indicates an expected retention volume of the cage. (**D**) An nsEM image obtained from SEC peaks (red arrow) shown in (B).



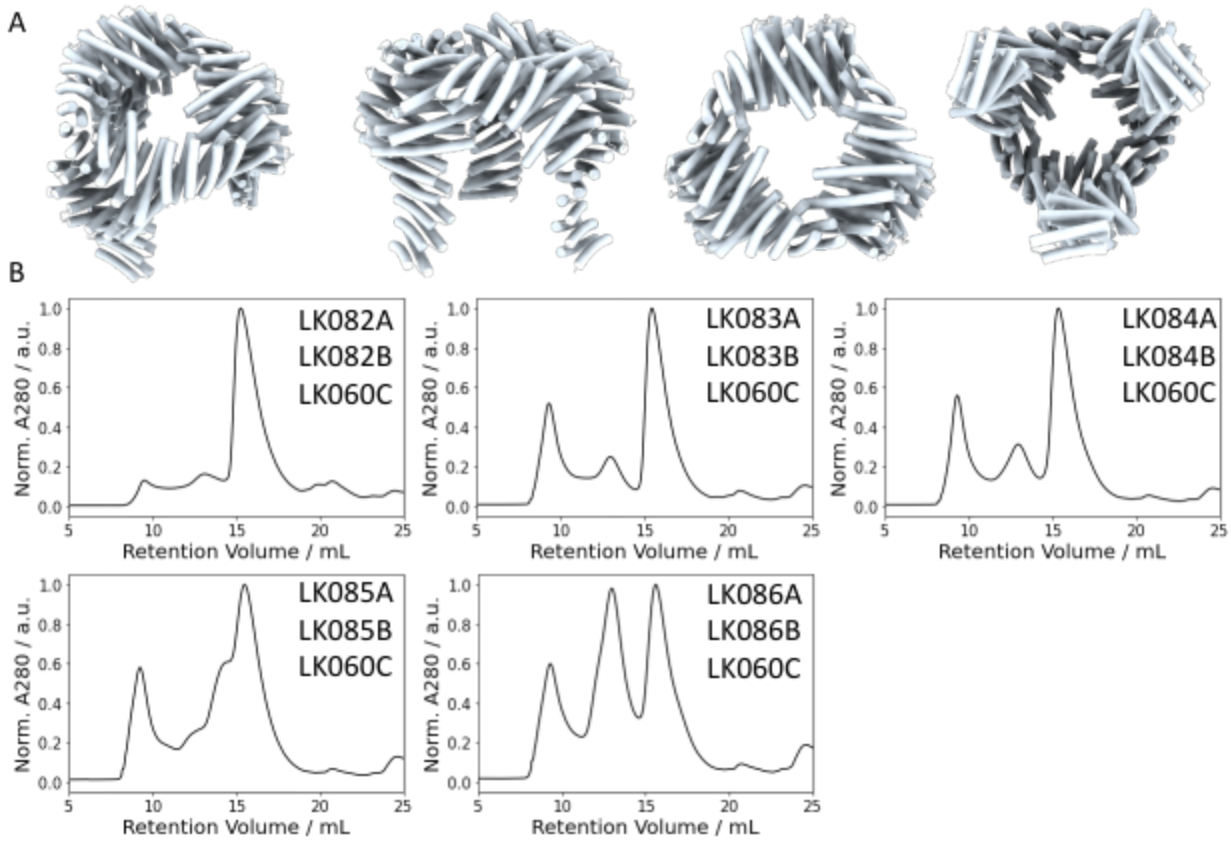
Supplemental Figure 2.10.1: nsEM of T=1 tetrahedral cages

nsEM images of tetrahedral cages (T=1, Symmetry: T3) (LK063, LK064, LK066, LK068) obtained from SEC peaks shown in Fig. S2.9.1.



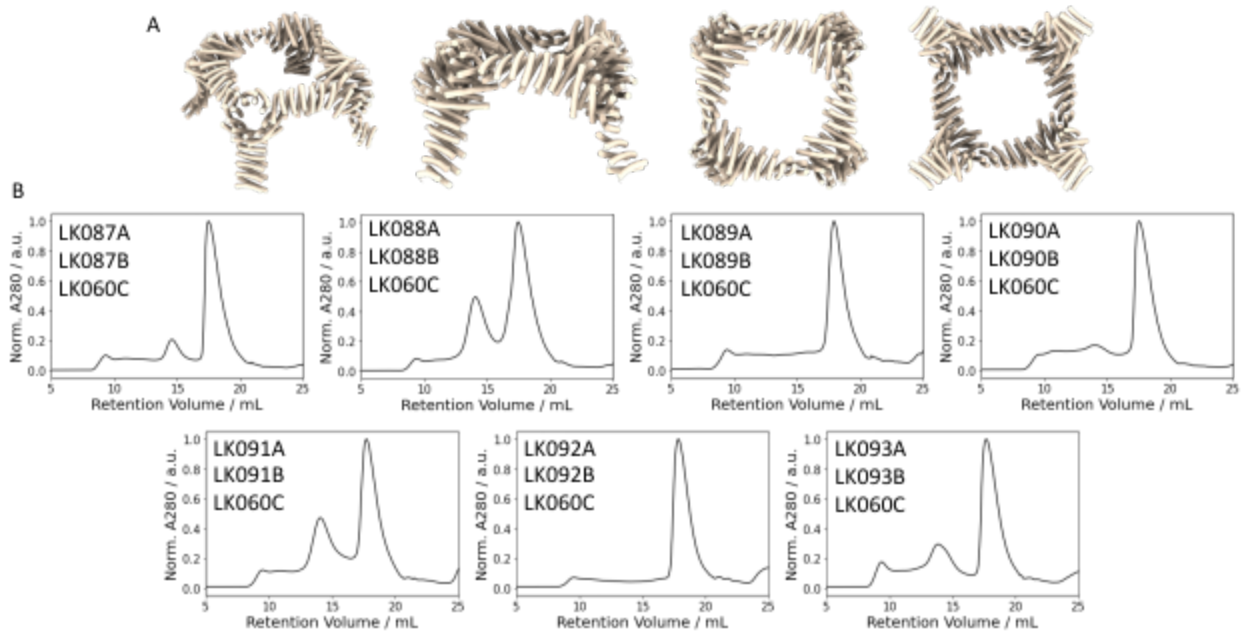
Supplemental Figure 2.10.2: nsEM of T=1 octahedral cages

nsEM images of octahedral cages (T=1, Symmetry: O3) (LK071 and LK073) obtained from SEC peaks shown in Fig. S2.9.2.



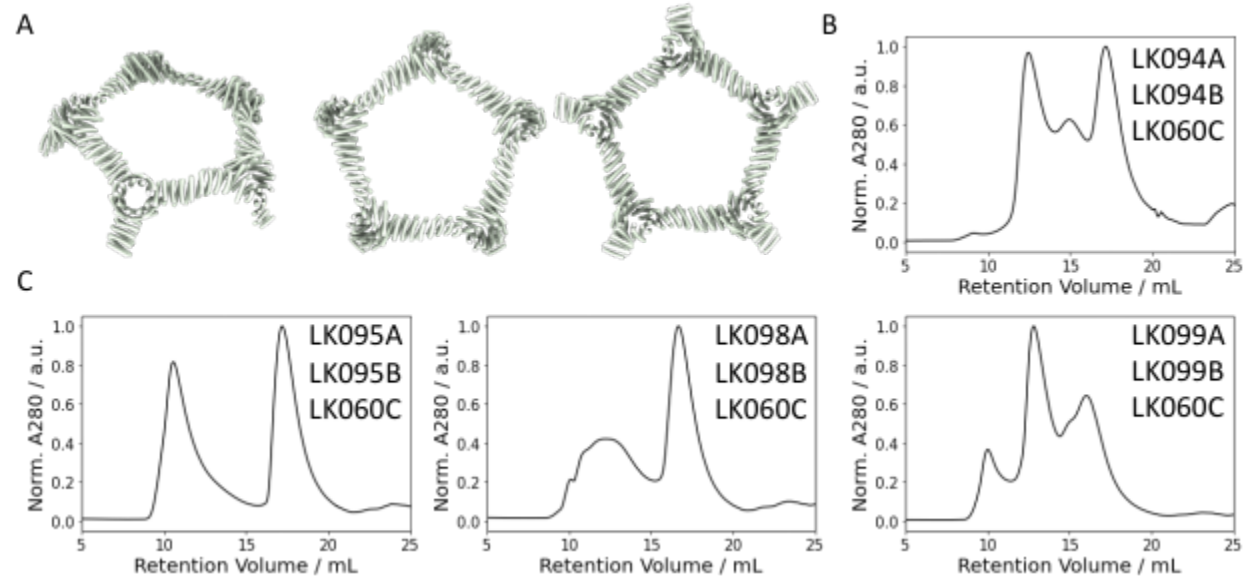
Supplemental Figure 2.11.1: SEC of C3 crowns

(**A**) Structure model of C3 crowns. (**B**) SEC traces of the crowns (LK082 – LK086) obtained from S200 column. Transparent grey box (~ 13ml) of each trace indicates an expected retention volume of the crown.



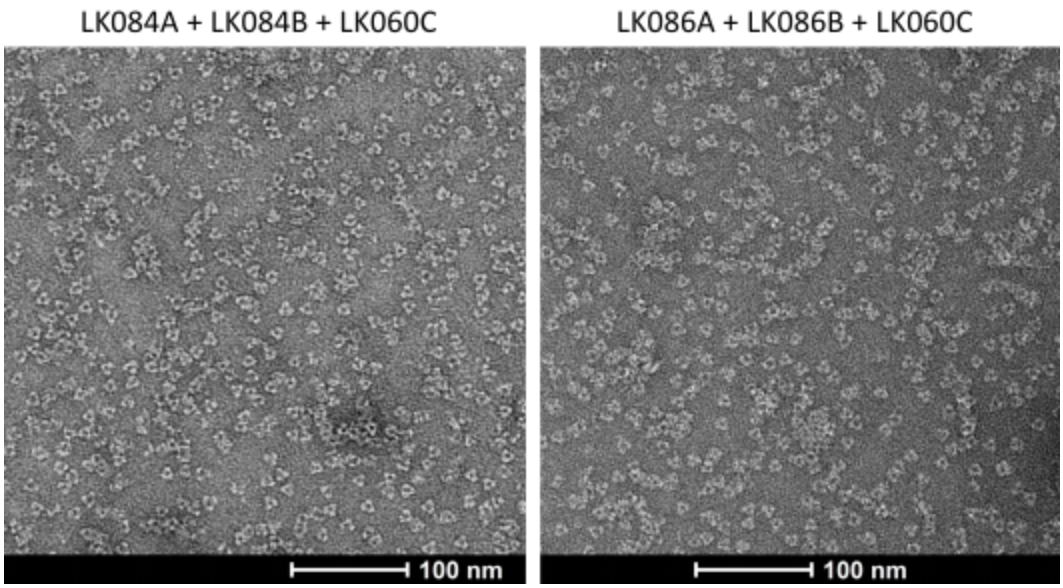
Supplemental Figure 2.11.2: SEC of C4 crowns

(**A**) Structure model of C4 crowns. (**B**) SEC traces of the crowns (LK087 – LK093) obtained from S6 column. Transparent grey box (~ 14ml) of each trace indicates an expected retention volume of the crown.



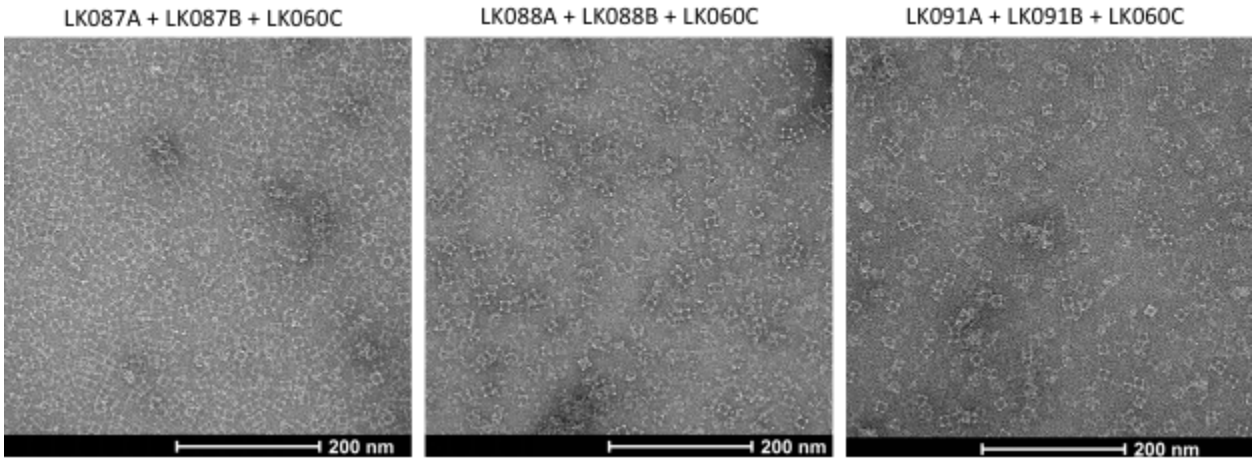
Supplemental Figure 2.11.3: SEC of C5 crowns

(**A**) Structure model of C5 crowns. (**B, C**) SEC traces of the crowns (LK094 – LK099) obtained from (**B**) S6 column and (**C**) S200 column. Transparent grey box of each trace indicates an expected retention volume of the crown.



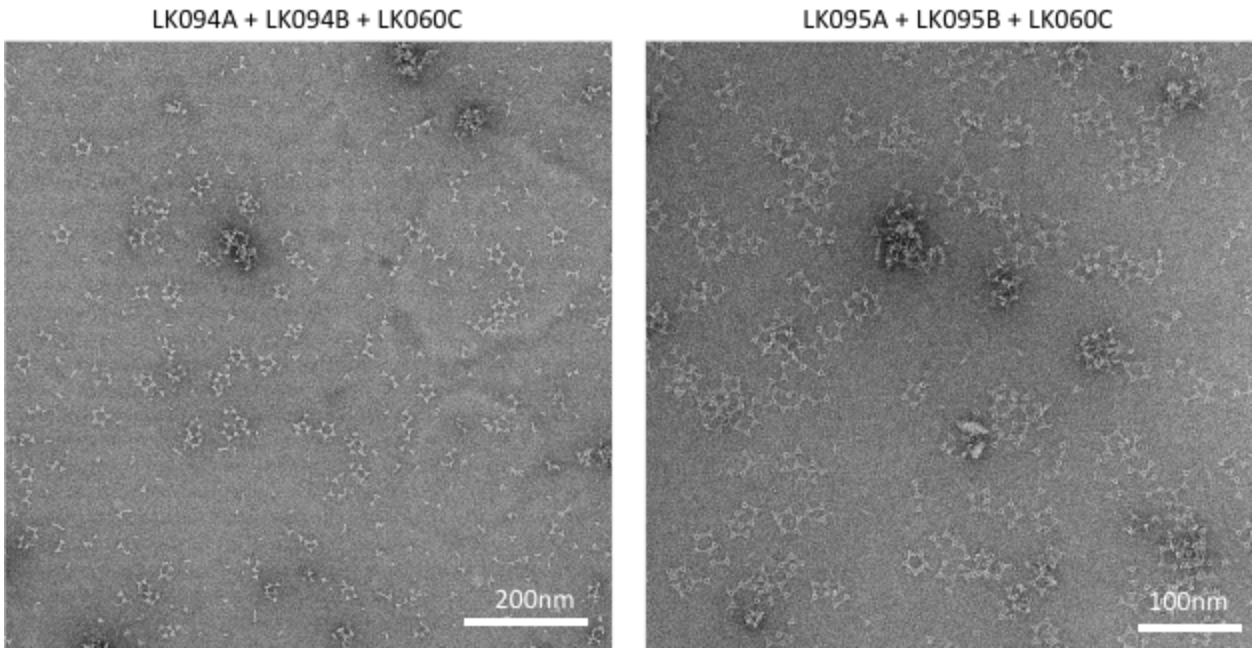
Supplemental Figure 2.12.1: nMS of C3 crowns

nsEM images of C3 crowns (LK084 and LK086) obtained from SEC peaks shown in Fig. S2.11.1.



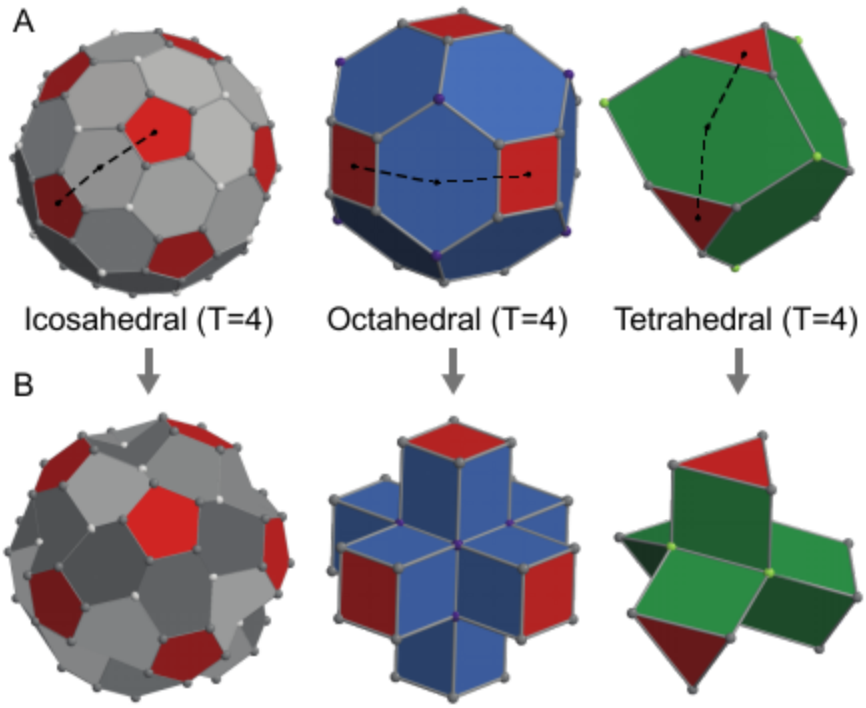
Supplemental Figure 2.12.2: nMS of C4 crowns

nsEM images of C4 crowns (LK087, LK088, LK091) obtained from SEC peaks shown in Fig. S2.11.2.



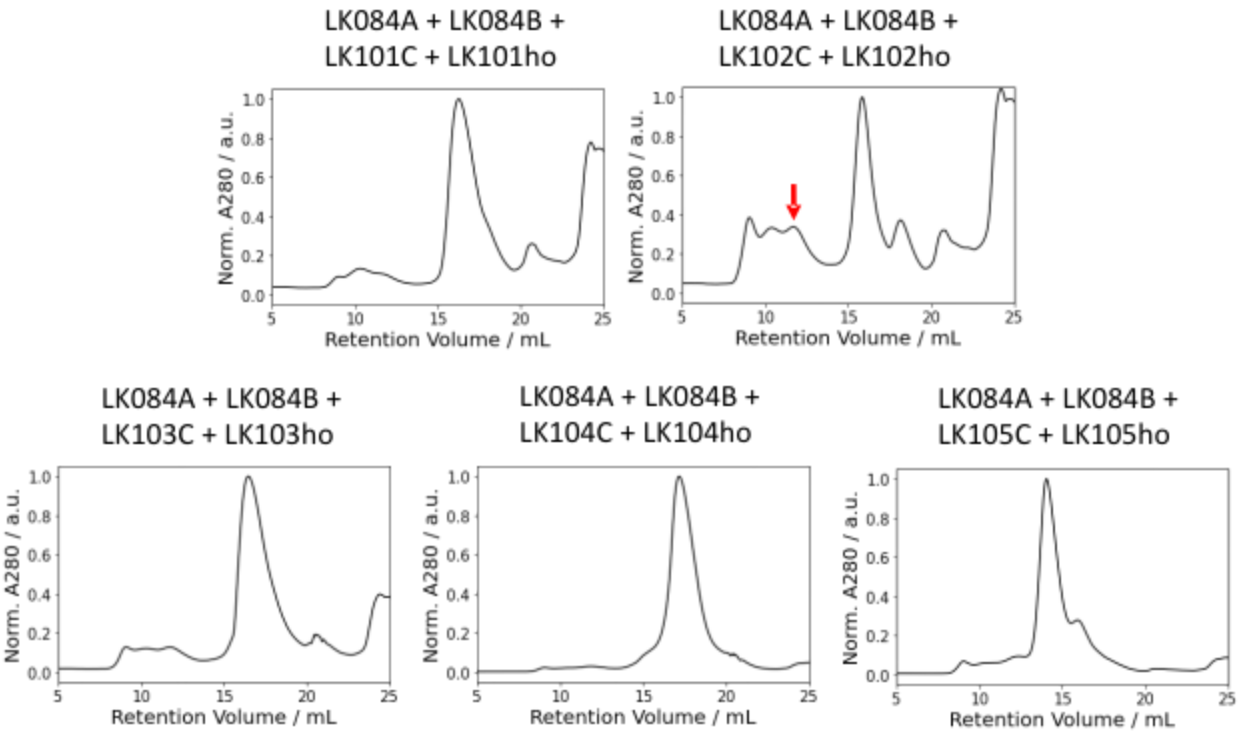
Supplemental Figure 2.12.3: nMS of C5 crowns

nsEM images of C5 crowns (LK094, LK095) obtained from SEC peaks shown in Fig. S2.11.3.



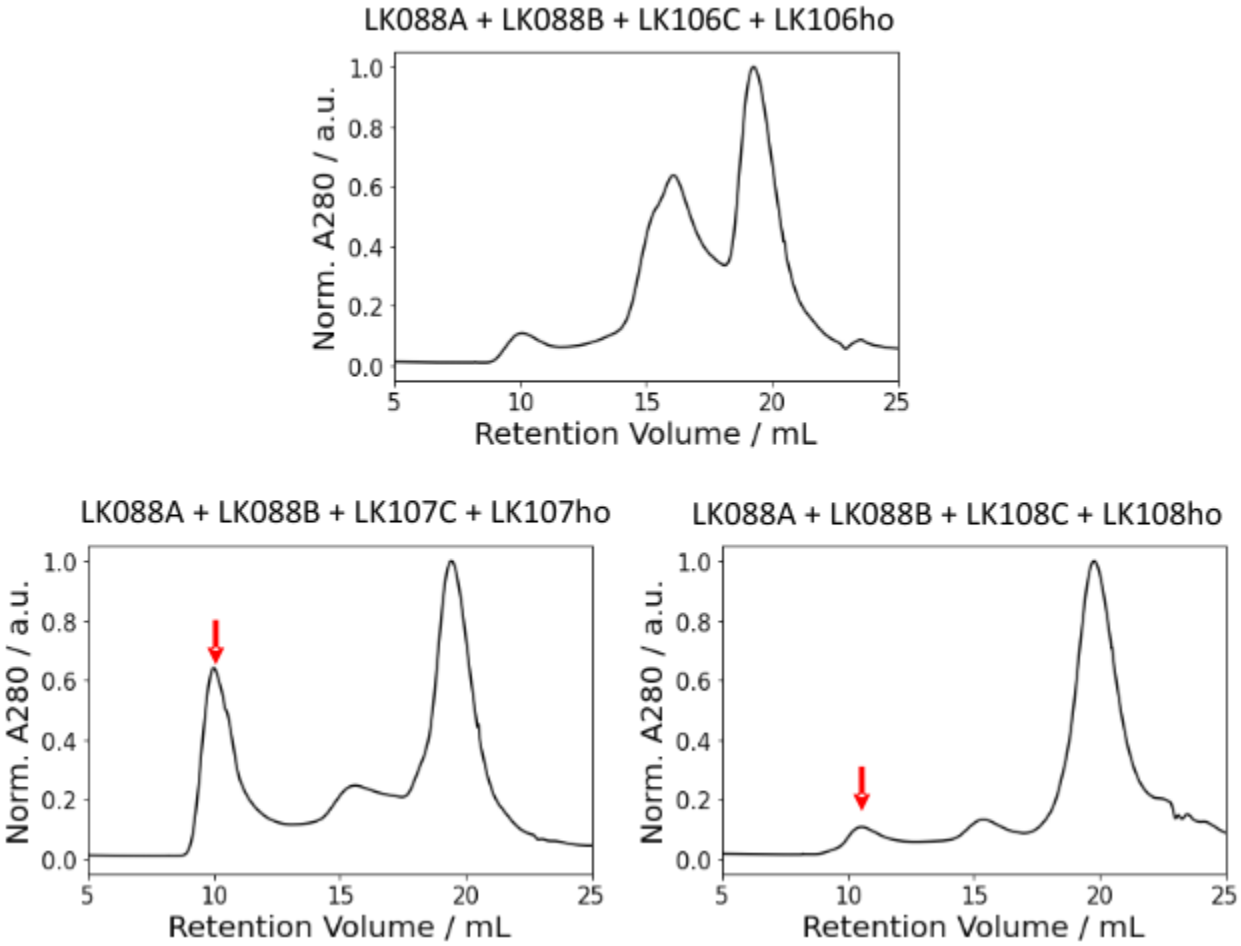
Supplemental Figure 2.13: Topologies of T=4 polyhedra

T=4 polyhedra with (**A**) convex and (**B**) concave shapes.



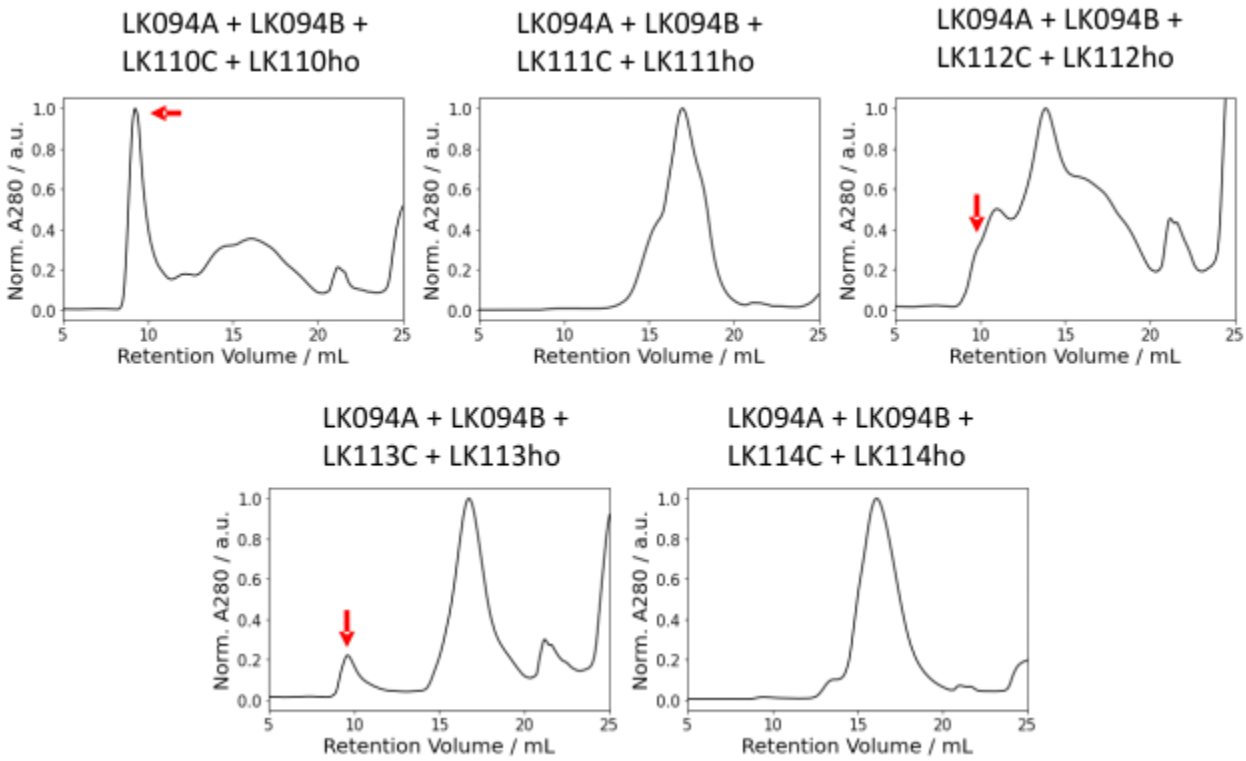
Supplemental Figure 2.14.1: SEC of T=4 tetrahedral cages

SEC traces of the T=4 tetrahedral cages (LK101 – LK105) obtained from S6 column. Transparent grey box of each trace indicates an expected retention volume of the cage. A red arrow indicates the sample where the nsEM image of Fig. 2.4A was obtained.



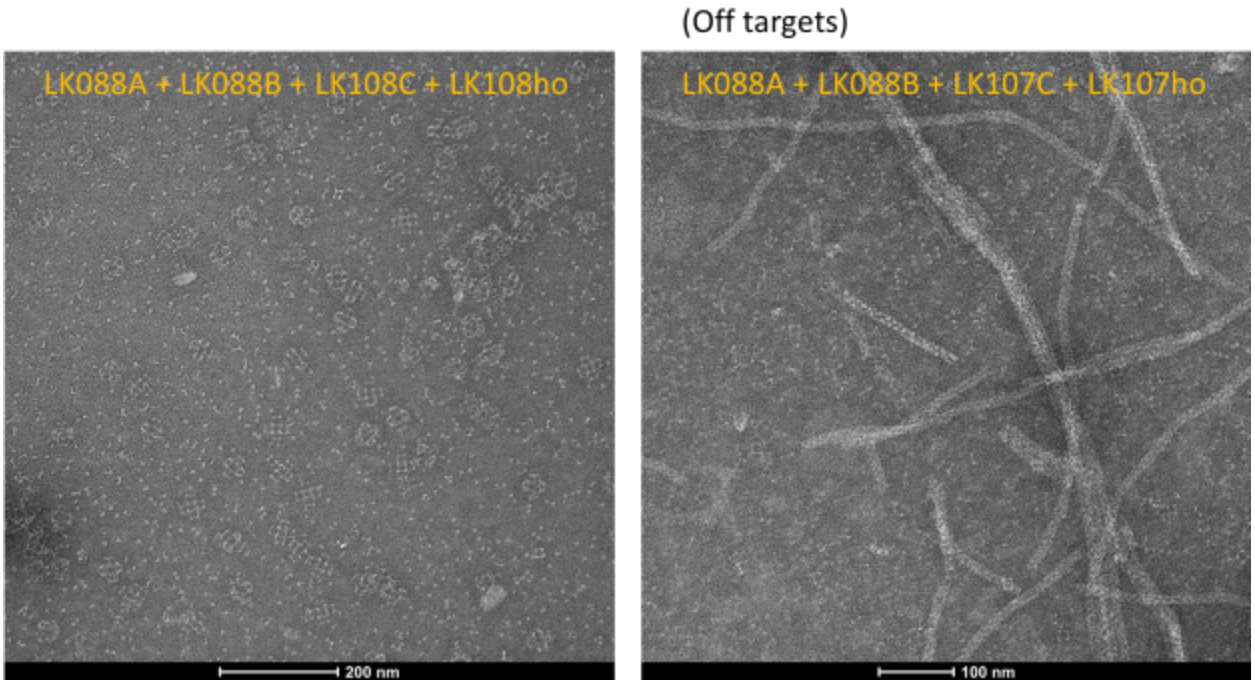
Supplemental Figure 2.14.2: SEC of T=4 octahedral cages

SEC traces of the T=4 octahedral cages (LK106 – LK108) obtained from S6 column. Transparent grey box of each trace indicates an expected retention volume of the cage. Red arrows indicate the peak where the nsEM images of Fig. 2.4C and Fig. S2.15.1 was obtained.



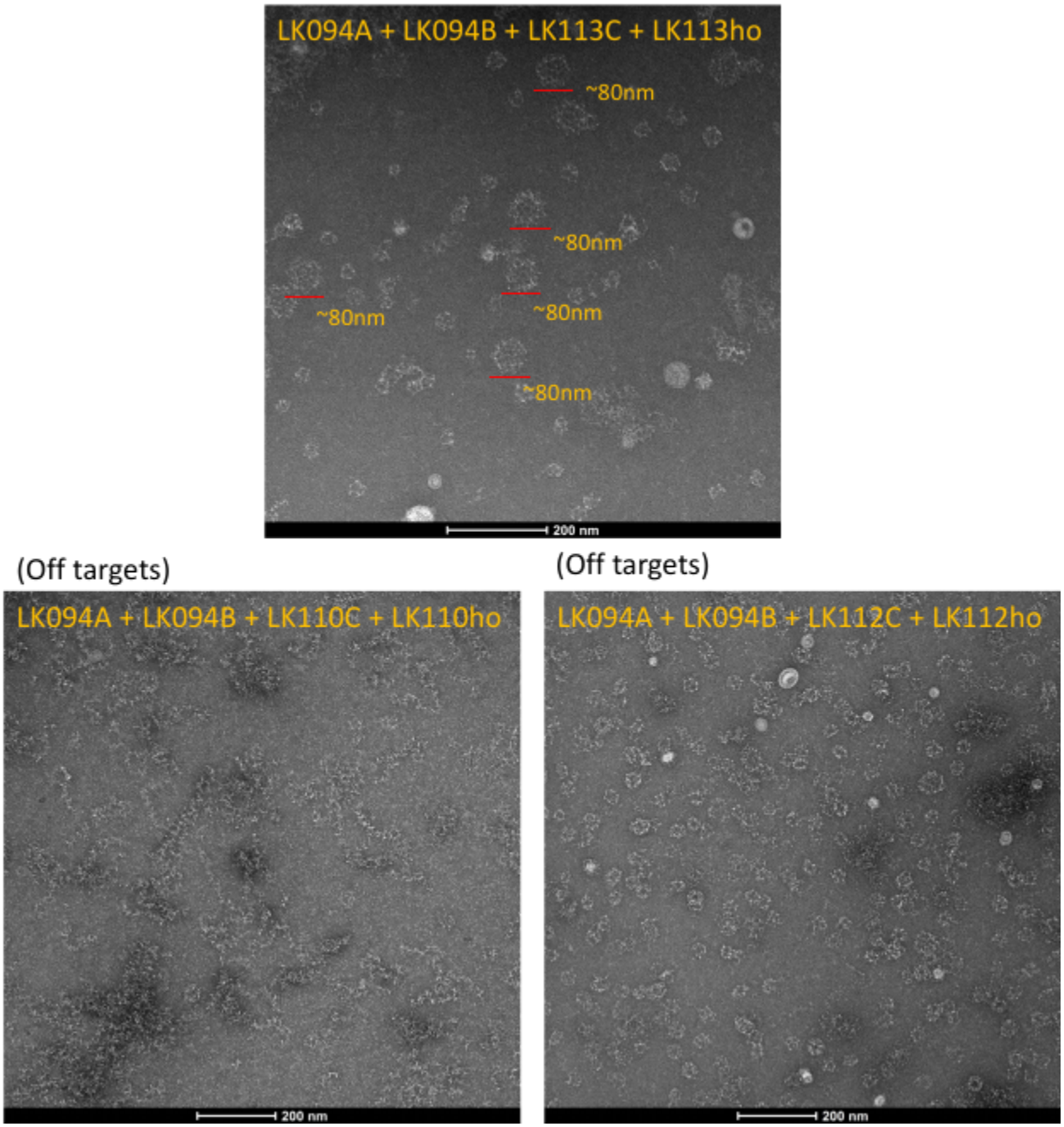
Supplemental Figure 2.14.3: SEC of T=4 icosahedral cages

SEC traces of the T=4 icosahedral cages (LK110 – LK114) obtained from S6 column. Transparent grey box of each trace indicates an expected retention volume of the cage. Red arrows indicate the peak where the nsEM images of Fig. 4E and Fig. S2.15.2 was obtained.



Supplemental Figure 2.15.1: nsEM of T=4 octahedral cages

nsEM images of T=4 octahedral cages (LK107 - right, LK108 - left) obtained from SEC peaks shown in Fig. S2.14.2. Right figure shows off-target formation.



Supplemental Figure 2.15.2: nsEM of T=4 icosahedral cages

nsEM images of T=4 icosahedral cages (LK110 – bottom left, LK112 – bottom right, LK113 – top) obtained from SEC peaks shown in Fig. S2.14.3. Bottom figures show off-target formations.

Supplemental tables

included in Supplemental Files:

Supplemental Table 2.1: Homotrimer metadata

Supplemental Table 2.2: Heterotrimer metadata

Supplemental Table 2.3: T=1 metadata

Supplemental Table 2.4: Crown metadata

Chain C = LK060C

Supplemental Table 2.5: T=4 metadata

Chain ‘ho’ indicates the protomer that forms homotrimer.

Tetrahedral: Chain A = LK084A; Chain B = LK084B

Octahedral: Chain A = LK088A; Chain B = LK088B

Icosahedral: Chain A = LK094A; Chain B = LK094B

Methods

In silico design of homo-oligomers

1.1. Helical fuse

The HelixFuse protocol from Hsia *et al.*[73] was used to combinatorially fuse together a set of 13 BGLs and 38 DHRs by structurally superimposing terminal helix residues[41] in either direction ("AB": N-terminus of a DHR to the C-terminus of a BGL; "BA": C-terminus of a BGL to the N-terminus of a DHR), only requiring the DHR to extend outward from the center of the BGL. Up to 11 residues of the BGL and an entire repeat of the DHR were allowed to be deleted during the procedure. After the C-beta atoms are superimposed, a RMSD check across 6 residues was performed to ensure that the fusion results in a continuous helix. Outputs were filtered to remove significant clashes (Rosetta centroid energy < 5.0), ensure at least 3 helices in contact with the overlapped region, and remove fusions with poor shape complementarity (<0.5).
List of DHRs used[72]: DHR01, DHR03, DHR07, DHR08, DHR09, DHR10, DHR14, DHR15, DHR18, DHR20, DHR21, DHR23, DHR24, DHR26, DHR27, DHR31, DHR32, DHR36, DHR39, DHR46, DHR49, DHR52, DHR53, DHR54, DHR55, DHR57, DHR59, DHR62, DHR64, DHR68, DHR70, DHR71, DHR72, DHR76, DHR77, DHR79, DHR80, DHR82
List of BGLs used: BGL0, BGL01, BGL02, BGL03, BGL05, BGL06, BGL08, BGL09, BGL14, BGL15, BGL17, BGL18, BGL19

1.2. Sequence design (Rosetta)

Sequence design was carried out with RosettaScripts[74] (hfuse_sym_hdf5.xml) with sequence choices informed by fragments with similar sequence (via StructProfileMover). Because of the presence of hydrogen bond networks in the BGL interfaces, a scoretype designed to avoid buried unsatisfied hydrogen bond networks was used[46] and the locations of HBNets

were not allowed to be changed. The interface between the BGL and DHR defined by InterfaceByVector and any residues superimposed over were redesigned and sidechain optimized with a basic fixed backbone relax (sidechain minimize -> repack (no design) -> sidechain minimize). The redesign was done again with a fixed core to optimize surface and boundary. After sequence design, designs were selected with more stringent shape complementarity (sc > 0.55), no more than 50 alalines, and at most 10 continuous apolar residues to identify designs likely to be well behaved and rigid across the junction point. In total, the building block library generated *in silico* by HF consists of 1556 AB and 589 BA fusions.

In silico design of pseudosymmetric hetero-oligomers

The basic idea of designing pseudosymmetric hetero-oligomers is to transplant three different protomer-protomer interfaces of BGLs (*1*) (guest) into a LK homotrimer scaffold (host), which gives three different chains that form a heterotrimer but the same structure with the homotrimer. We first tested the compatibility between a guest interface and a host scaffold as a homo-oligomer, following steps below:

(i) Identifying residues (of the LK host) involved in protomer-protomer interfaces.

(ii) Identifying residues (of the LK host) involved in DHR-arm binding.

(iii) The residues (of the LK host) identified in (i) but not involving in (ii) are defined as 'mutable residues'.

(iv) Identifying residues (of the BGL guest) involved in protomer-protomer interfaces.

(v) Mutating the 'mutable residues' to the residues in (iv).

A schematic diagram of the interface transplantation process is shown in Fig. S2.M.1. After experimentally validating interface transplanted homotrimers in experiment, we combined them into heterotrimers (Fig. S2.M.2). For a heterotrimer, a host and three guests are selected (Table

S2.M.1), and the protomer-protomer interface of each guest is transplanted into one of the protomer-protomer interfaces of the host.

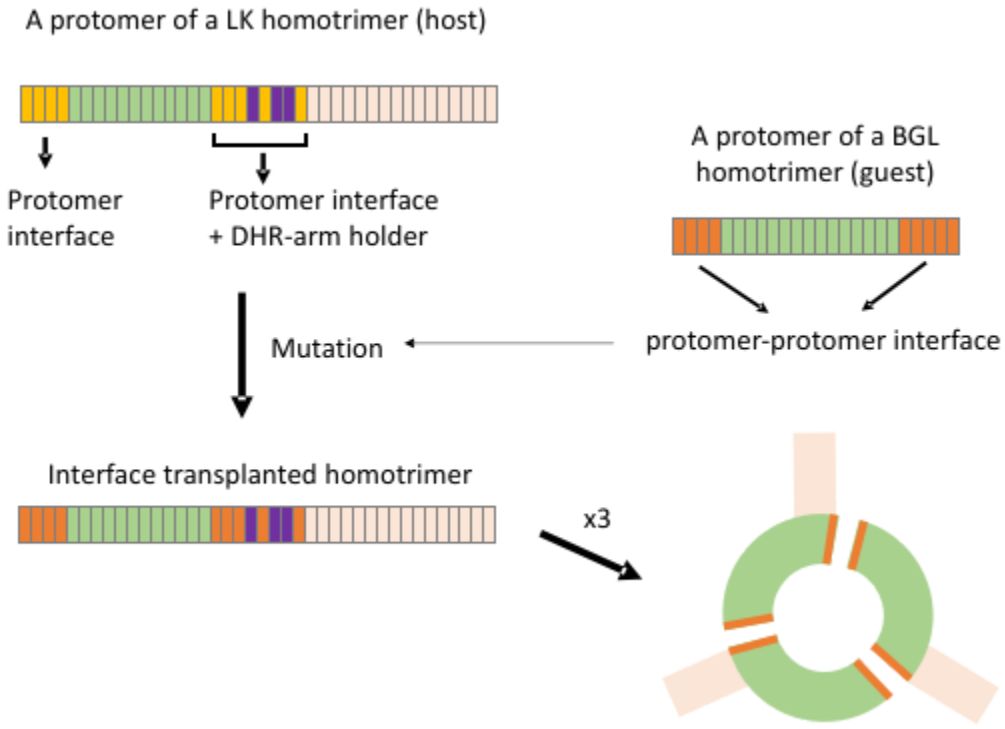


Fig. S2.M.1. Schematic diagram of the interface transplantation of homotrimers.

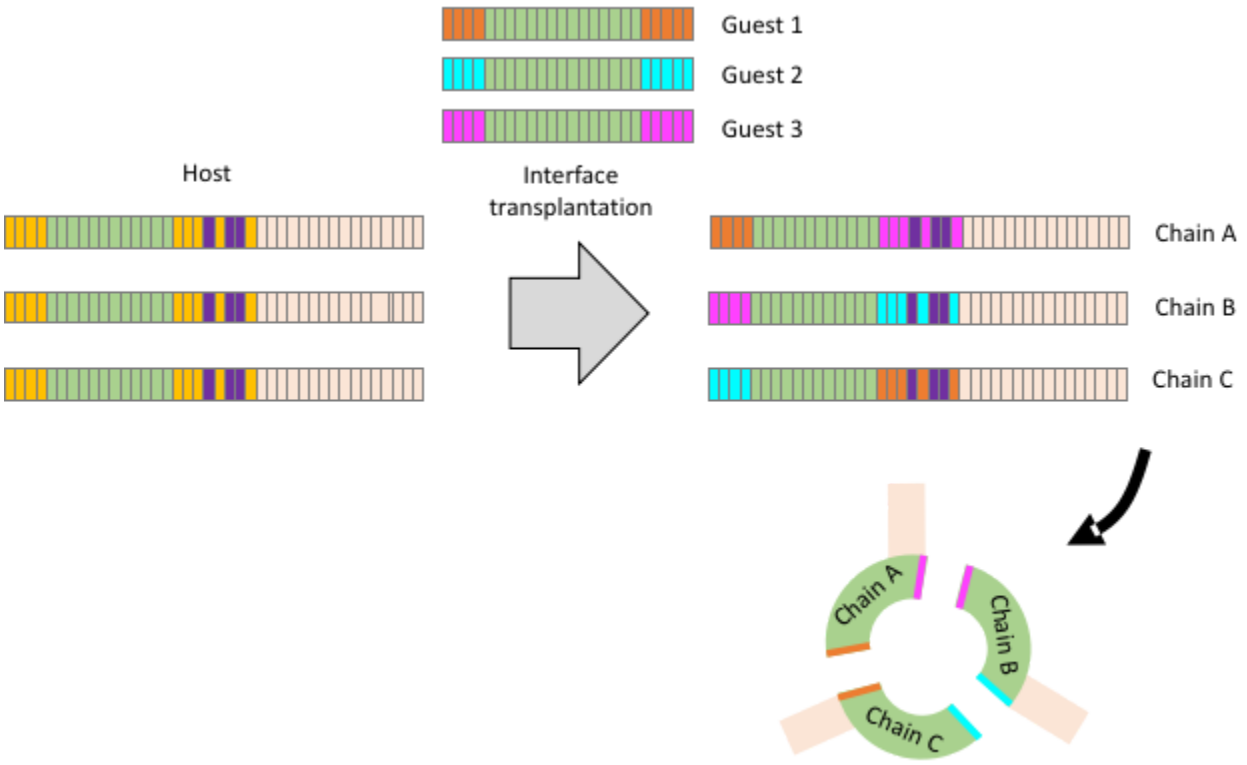


Fig. S2.M.2. Schematic diagram of the interface transplantation of heterotrimers.

ID	Host	Guest 1	Guest 2	Guest 3
LK058	LK032	BGL00	RK387	RK388
LK059	LK032	BGL00	RK387	RK389
LK060	LK032	BGL00	RK388	RK387
LK061	LK032	BGL00	RK389	RK387
LK062	LK032	RK387	RK388	RK389

Table S2.M.1. Host and guest combinations for the interface transplantation of heterotrimers.

In silico design of cages and crowns

3.1. Docking

Cages docking was carried out using the residue-pair transform docking (RPXdock) which is available at https://github.com/willsheffler/rpxdock, and details of the RPXdock can be found in a recently published paper[77]. Briefly, the RPXdock samples docking configurations using input building blocks for a given symmetry by translating and rotating the building blocks along given symmetry axes. For T=1 cages, T3, O3 and I3 symmetries were used for tetrahedral, octahedral, and icosahedral symmetric cages, respectively. Input scaffolds are LK032-based backbones, which are the same backbones with LK066 (for T3), LK071 (for O3), and LK078 (for I3). For crowns, we did not run additional dockings. Instead, we reused the docking interfaces verified from T = 1 cages. For T=4 cages, T33, O43 and I53 symmetry were used. For T33, input scaffolds are: (1) C3 crown (made by backbones of LK084A, LK084B, LK102C) along a 3-fold symmetry axis, (2) a homotrimer (made by backbones of LK102ho) along another 3-fold symmetry axis. For O43, input building blocks are: (1) C4 crown (made by backbones of LK088A, LK088B, LK108C) along a 4-fold symmetry axis, (2) a homotrimer (made by

backbones of LK108ho) along a 3-fold symmetry axis. For I53, input building blocks are: (1) C5 crown (made by backbones of LK094A, LK094B, LK113C) along a 5-fold symmetry axis, (2) a homotrimer (made by backbones of LK113ho) along a 3-fold symmetry axis. Each docking configuration is scored by RPX-score, and dockings with top 3 scores for each symmetry were selected to carry out sequence design.

3.2. Sequence design

Sequence design for the cages was carried out using a deep learning based software ProteinMPNN which is available at https://github.com/dauparas/ProteinMPNN, and details (algorithms etc.) are well described in a recently published paper[79]. ProteinMPNN was used for designing only for residues forming the protein-protein interfaces for the cages or crowns. For T=1 cage design, protomer sequences were tied to be identical to design homo-dimeric interfaces, and cysteines were disallowed. For crown design, we reused the docking interfaces verified in T=1 cages, but now protomer sequences were not tied to design hetero-dimeric interfaces, and cysteines were disallowed. For T=4 cage, the docking interface between chain C of crowns and homo-trimer were designed, and designed sequences were not tied to design hetero-dimeric interfaces. Cysteines were disallowed. For each design, 32 sequences for each temperature (0.1, 0.2, 0.4, 0.6, 0.8 and 1.0) were generated.

The quality of the newly designed interfaces was assessed by AlphaFold2 (AF2)[75] using model 4 and 5 (https://github.com/deepmind/alphafold). The new interfaces for cages and crowns are located at DHR-arm part (not the BGL ring part), thus we conducted AF2 only with the DHR-arm parts instead of running entire protomers to reduce computation time. For homo-dimeric interface design, we selected designs that show backbone RMSD < 2.0 (computed with pymol) to the original protomer in AF2 prediction and N_{Met} < 2 at the interface for

experimental test. For hetero-dimeric interface designs, we additionally checked the self-interacting of the designed interface of each chain, using AF2 (Fig. S2.M.3). We only used the designs that show a weak self-interaction.

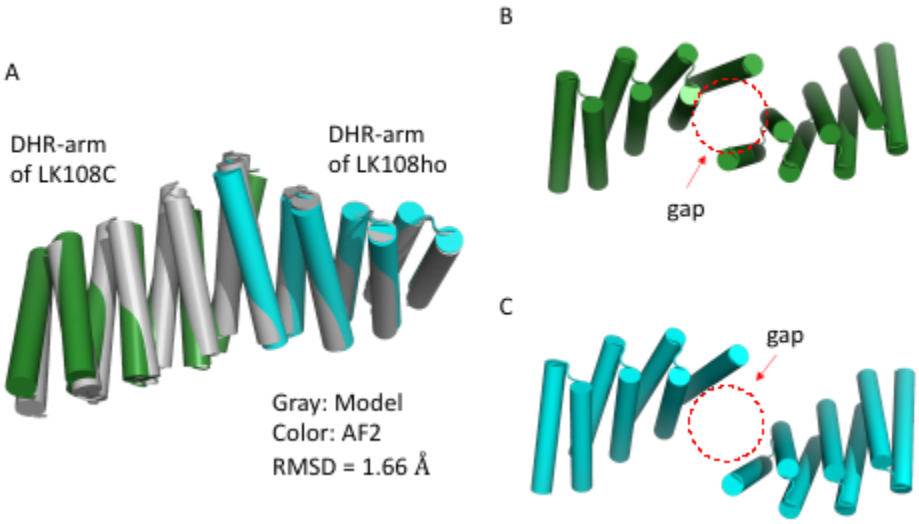


Fig. S2.M.3. An example of computational validation of designed interface using AF2. Only DHR-arm part was used for AF2 prediction. (**A**) Interface between LK108C and LK108ho. Model (Gray) and AF2 prediction (green and cyan). (**B-C**) In self-association check, AF2 predict that the homo-dimeric interface of (**B**) LK108C and (**C**) LK108ho show a gap at the interface. We assumed that this gap indicates a weak (or no) self-interaction. For all other hetero-dimeric interface designs of this work, we only used the interface designs that show the weak self-interaction.

Experimental methods and materials

The experimental methods and materials used in this study are mostly same methods and materials described in the previous chapter.

4.1. Buffer and media recipes

All buffers and media were made using Milli-Q filtered water.

Autoinduction media (TBM-5052):

1.2% [wt/vol] tryptone, 2.4% [wt/vol] yeast extract, 0.5% [wt/vol] glycerol, 0.05% [wt/vol] D-glucose, 0.2% [wt/vol] D-lactose, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, 2 mM $MgSO_4$, 10 μM $FeCl_3$, 4 μM $CaCl_2$, 2 μM $MnCl_2$, 2 μM $ZnSO_4$, 400 nM $CoCl_2$, 400 nM $NiCl_2$, 400 nM $CuCl_2$, 400 nM Na_2MoO_4, 400 nM Na_2SeO_3, 400 nM H_3BO_3

Lysis buffer:

25 mM Tris, 300 mM NaCl, 20 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Wash buffer:

25 mM Tris, 300 mM NaCl, 40 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Elution buffer:

25 mM Tris, 300 mM NaCl, 300 mM imidazole, 100 mM EDTA, 10% glycerol, pH 8.0 at room temperature

Het Lysis buffer:

25 mM Tris, 300 mM NaCl, 10 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Het Wash buffer:

25 mM Tris, 300 mM NaCl, 20 mM imidazole, 10% glycerol, pH 8.0 at room temperature

Het Elution buffer:

25 mM Tris, 300 mM NaCl, 50 mM imidazole, 10% glycerol, pH 8.0 at room temperature

SEC buffer:

25 mM Tris, 300 mM NaCl, pH 8.0 at room temperature

SAXS buffer:

25 mM Tris, 300 mM NaCl, 2% glycerols, pH 8.0 at room temperature

TEV buffer:

25 mM Tris, 100 mM NaCl, 0.5 mM EDTA, 1mM DTT, pH 8.0 at room temperature

4.2. Construction of synthetic genes

All synthetic genes were ordered from either Integrated DNA Technologies Inc. (Coralville, IA, USA) (IDT) using pET29b+ vectors with kanamycin resistance, and genes were reverse translated and codon optimized using Domesticator (https://github.com/rdkibler/domesticator_3). See Supplementary Tables 2.1 - 2.5 for full protein sequences. Where indicated, hexa-histidine tag (histag) was added for purification, and a tobacco etch virus protease (TEVp) cleavage site (ENLYFQG) was added for the cases that the histag needs to be removed, such as crystallization, SAXS.

4.3. Protein expression

Plasmids (100ng) were transformed into chemically competent *E. coli* expression strain, BL21(DE3) or BL21(DE3)Star, for protein expression following manufacturer's protocol, with the exception of using 10ul competent cells per reaction. Following transformation and recovery, the entire transformation products were used to inoculate 1 mL Luria-Bertani (LB) medium containing 100 ug/mL kanamycin and grown at 37°C with shaking at 225 rpm overnight. 500ul of overnight cultures were diluted into 50 mL TBM-5052 supplemented with 100 ug/mL kanamycin in 250 mL baffled flasks and incubated at 37°C with shaking at 225 rpm for 18-24 hours.

For systems using hetero-oligomers (pseudosymmetric hetero-oligomers, crowns, and T=4 cages), we used two different protocols: *protocol 1* and *protocol 2*.

Protocol 1: Each component was transformed and incubated independently. After the incubation, we first checked relative expression level between components using SDS-PAGE (section 4.5) based on the intensity of bands. Then, the incubated E. coli cultures of each component were mixed with a correct stoichiometry for target assemblies (1:1:1 for hetero-trimers and crowns. 1:1:1:1 for T=4 cages). After mixing the E. coli cultures, we harvested proteins following *Section 4.4*.

Protocol 2: Plasmids of all components were transformed into E. coli together. For example, in hetero-trimer expression (3 components – A, B, C), 300ng of plasmids (100ng of A + 100ng of B + 100ng of C) were transformed into 10ul E. coli cells and incubated as described in this section. Then, proteins were harvested following *Section 4.4*.

4.4. Protein purification

Immobilized metal affinity chromatography (IMAC)

Cultures were harvested by centrifugation at 4000 rcf for 10 minutes, culture supernatant decanted, and pellets resuspended to 30 mL in Lysis buffer. 300ul PMSF (100mM in 100% EtOH) is added immediately prior to sonication at 70% power for 5 minutes. “Lysate” fractions are saved, and then lysates were clarified by ultracentrifugation at 18,000 rcf at 12°C for at least 30 minutes and applied to 1.5 mL Ni-NTA resin (Qiagen) pre-equilibrated with Lysis buffer and packed into Econo-Pac columns (Bio-Rad) for gravity chromatography. The columns were washed twice with 15 mL Wash buffer and eluted with 10 mL Elution buffer. Hetero-oligomers (LK058-LK062) were purified according to a similar procedure, except they used the “Het” variants of the the Lysis, Wash, and Elution buffers as this was found to improve the yield of single-his-tagged complexes over non-specific multiple-his-tagged complexes which can arise at

high concentrations and likely dominate binding to the Ni-NTA resin. Samples prepared for crystallization were treated similarly, except 500 mL cultures were used and lysate was divided among six gravity columns.

Size-exclusion chromatography (SEC)

Samples were concentrated using 10k MWCO spin concentrators and were purified using a Superdex 200 10/300 increase (for oligomers) or Superdex 6 10/300 increase (for crowns and cages) columns (Cytiva) in SEC buffer using an ÄKTA pure system (Cytiva). SEC traces were also used to qualitatively determine homogeneity and quantitatively measure total yield by A280 absorbance integrated over the collected fractions using Unicorn (Cytiva).

TEVp cleavage

Purification and mass tags were buffer exchanged into TEV buffer and cleaved with TEVp at a ratio of 1 mg TEV per 100 mg substrate for 24-72h at room temperature. After TEV cleavage, samples were exchanged into Lysis buffer and passed over a Ni-NTA gravity column and washed with 10ml lysis buffer. Flowthrough was collected, concentrated using 10k MWCO spin concentrators, and purified once again by SEC.

4.5. Sample analysis

SDS-PAGE

Samples were diluted 1:1 with 2x Laemmli Sample Buffer (Bio-Rad) without Beta-mercaptoethanol and 15ul were loaded onto AnykD™ Criterion™ TGX™ Precast Midi Protein Gels (Bio-Rad). Ladder was 10ul of Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (Bio-Rad). Gels were run at 300V for 18 minutes, then stained using an

eStain™ L1 Protein Staining System (Genscript). Stained gels were imaged using a Chemidoc XRS+ (Bio-Rad)

Liquid chromatography mass spectrometry (LC-MS)

To identify the molecular mass of each protein and thus verify sample identity and integrity, intact mass spectra was obtained via reverse-phase LC/MS on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column, and subsequently deconvoluted by way of Bioconfirm using a total entropy algorithm.

Native mass spectrometry (nMS)

Oligomeric state of SEC-purified and LC-MS verified samples was analyzed by on-line buffer exchange MS in 200 mM ammonium acetate using a Vanquish ultra-high performance LC system coupled to a Q Exactive ultra-high mass range Orbitrap mass spectrometer (Thermo Fisher Scientific). A self-packed buffer exchange column was used (P6 polyacrylamide gel, BioRad). The recorded mass spectra were deconvolved with UniDec version 4.2+.

Negative stain electron microscopy (nsEM)

SEC purified samples were diluted to (0.005 mg/ml for oligomers, 0.05 – 0.1 mg/ml for crowns and cages) using SEC buffer immediately before application for 45s to glow discharged thick carbon film-coated 400 mesh copper grids (CF400-CU TH) (Electron Microscopy Sciences). Grids were then stained and dried immediately twice using 2% uranyl formate. Dried grids were screened on a 120 kV Talos L120C transmission electron microscope. The E. Pluribus Unum (EPU) (FEI Thermo Scientific) software was used for automated data collection. Data processing was carried out in CryoSPARC™ (Structura Biotechnology Inc).

Cryo-EM

Small angle X-ray scattering (SAXS)

TEVp-cleaved (and optionally AEC purified) samples were re-purified by SEC in SAXS buffer and concentrated using thoroughly washed 10k MWCO small spin concentrators; the flowthrough of concentration was used as blanks for buffer subtraction. Scattering measurements were performed at the SIBYLS 12.3.1 beamline at the Advanced Light Source as part of the HT-SAXS program. The Xray wavelength (λ) was 1.27 Å, and the sample-to-detector distance was 1.5 m, corresponding to a scattering vector q ($q = 4\pi \sin\theta/\lambda$, where 2θ is the scattering angle) range of 0.01 to 0.3 Å−1. A series of exposures, in equal subsecond time slices, were taken of each well: 0.3 second exposures for 10 seconds resulting in 32 frames per sample. For each sample, data was collected for two different concentrations to test for concentration dependent effects; "low" concentration samples were ~2.5 mg/mL and "high" concentration samples were ~ 5 mg/mL. Data was processed using the SAXS FrameSlice online server. FoXS[48] was used to compare design models to experimental scattering profiles and calculate quality of fit (χ) values. The SAXS Similarity online server was used to compute the similarities of scattering profiles to each other and calculate quality of fit (χ) values.

X-ray crystallographic

Crystals of LK031 were grown using protein purified as described above and TEVp cleaved. Protein samples dispensed in 1 uL drops at purification concentrations were mixed with equal volume of a crystallization solution and set in hanging drops. Vapor phase equilibration of the resulting drops against a 1 mL reservoir of the same crystallization solution resulting in growth of crystals. The crystals were flash cooled in liquid nitrogen after transfer into a cryoprotective solution. Diffraction data were collected on a Pilatus areas detector at the Advanced Light Source (ALS) synchrotron facility at beamline 5.0.2. The resulting data sets extend to 4.5 Å resolution. Data was processed using the program HKL2000[49] or Aimless[50]. The

placement of subunits was determined using the molecular replacement algorithm in program PHENIX[51]. Local rebuilding of all constructs was performed using the program COOT[52], followed by refinement using the program PHENIX[51]. Final values for Rwork/Rfree are 0.279/0.319. Please inquire for crystal growth and protection conditions.

Conclusion

The method described in Chapter 1 of creating hetero-oligomers by the efficient and easy validation of related homo-oligomers is currently unlocking new paths forward for the design of large nanostructures. I have shown one application in the use of modularly replaceable symmetric and pseudosymmetric building blocks in the stepwise design of T=4 nanocage (Chapter 2), but the structural implications are not limited to nanocages. Work is also under way by Yang Hsia and Paul Kwon to create 2D arrays from symmetric oligomers with pseudosymmetric variants, which will enable them to switch between an unbounded 2D array and a bounded cyclic oligomer with the same geometry as the 2D array's unit cell[80]. Not only is this an interesting design challenge and another good example of the relationship between higher and lower order symmetries, but it will provide an easier platform for prototyping modifications to 2D arrays, which are incredibly useful for displaying patterns of molecules but which themselves can be difficult to work with because array formation and nonspecific aggregation cannot be distinguished with many common biochemical techniques.

Pseudosymmetric structures will also provide a bridge between the world of symmetry, so often used by Nature and protein designers alike (for good reason), and asymmetry. It can often be useful to have a single "handle" on a symmetric structure, such as an antibody or a symmetric receptor, for the purposes of attaching them to other proteins or structures or ensuring single point modification. Designing a pseudosymmetric protein to interact with the symmetric target (or more likely designing a symmetric binder to interact with a symmetric target, then pseudosymmetrizing the binder) would be an effective way to accomplish this which requires no complicated in vitro modifications or purifications, which are currently needed for this task.

Core to the reason why this pseudosymmetrization method has been so effective is because it modularized protein structure and enabled interface design to be abstracted. When protein designers can abstract elements of their design problem, they can focus on the harder and more rewarding aspects and advance the field more efficiently. Modular protein assembly extends far beyond what I presented in this dissertation, and Tim Huddy, Yang Hsia, myself, and others at the Baker Lab are now attempting to push to its farthest practical limits.

Acknowledgements

Forgive me if I forget to mention your name. The list of people I've interacted with and received help from throughout the years is deep and continuously growing. Basically, if we've ever spoken or exchanged messages, I thank you.

I've been blessed to have several fantastic teams of mentors and colleagues supporting me. From the beginning, Zibo Chen, Scott Boyken, and Marc Lajoie provided endless help and support in more ways than I can count. I want to thank Zibo in particular for being my most involved mentor and friend. Nick Woodall, Basile Wicky, and Ajasja Ljubetič supported me when I was struggling to find a project that I actually enjoyed working on, and which shortly blossomed into the work that I presented in Chapter 1. Finally, I am thankful for my latest mentor and collaborator, Sangmin, who worked hand-in-hand with me on the content of Chapter 2 and is helping me through the final stretch of graduate school and paper writing. I've benefited tremendously from our relationship and I couldn't have been luckier to have been paired up with him.

Throughout my entire PhD, the great connector, David Baker, has been a constant force, asking me if I've met so-and-so, and driving me to innovate and leave my comfort zone. It's really special to have someone guiding you who is so attuned to the science and the process as he is. He is clearly doing something right at the Institute for Protein Design, and I get chills when I think about the ever accelerating rate of innovations coming out of his lab.

I've had the pleasure of being involved with several fruitful collaborations for regular data collection with the Wysocki Lab at Ohio State University for nMS and SIBYLS at LBNL for SAXS. It's been wonderful working with Greg Hura and Kathryn Burnett at SIBYLS and Florian Busch, Andrew Norris, Nicholas Horvath, Stella Lai, Sean Cleary, Marius Kostelic, and

Vicki in the Wysocki Lab. I'm also thankful to my personal collaborators Madison Kennedy, Lindsey Doyle, Phil Bradley, and Barry Stoddard at the Fred Hutch, Andrew Hunt at Northwestern, Agnese Curatolo and Ryan Krueger (the other RK) for our shared projects.

I've had many great collaborators and friends within the Baker Lab. My closest pals, Yang Hsia and Tim Huddy, have been amazing to work with over the past year or two and I hope the fun continues. I've also had the pleasure of working with/mentoring/hanging out with/learning from Phil Leung, Adam Broerman, Ryan McHugh, Sam Tipps, Elias Kinfu, Helen Eisenach, and Jinwei Xu, to name a few. They are surpassing me in so many ways. Sam Pellock keeps me grounded.

An acknowledgements section would not be complete without thanking the BPSD program and the venerable grad student wrangler, Erin Kirchner. I must also thank the outstandingly helpful leadership, administrative, lab support, and IT staff at the IPD and the IPD Core Labs scientists. The IPD doesn't work without any of them, but in particular Kandise VanWormer, Luki Goldschmidt, Lance Stweart, and Lauren Carter, who will stop at nothing to make sure everyone has the support they need. Thank you to Mengyu Wu and Andrew Borst, the former and current EM Core leads, who taught me negative stain electron microscopy, which has been transformative for me and my projects, and now especially to Andrew for teaching me CryoEM. This is only the start of an exciting journey.

My cohort mates, Robby Divine, Gabi Reggiano, Inez Pranoto, Halli Benasutti, Carson Adams, and James Griffin deserve a special thanks. From weekend hangouts to winter cabin retreats, this group has made grad school more fun than it should have been. I look forward to keeping this talented group of people as my friends for the rest of my life.

I must express my gratitude to my mom and dad, Laurie and Bruce Kibler, my brother Eric, my grandma Athalie, and my partner Olivia Thibeault for everything they've done for me throughout my life. Being so far from my family has been so difficult, and I know I've not been the best communicator during these last five years, but they've stood by me and I've felt their love the entire time. I love them and cherish them.

References

1. Langan, R. A. *et al.* De novo design of bioactive protein switches. *Nature* **572**, 205–210 (2019).
2. Kirkpatrick, R. L. *et al.* Conditional Recruitment to a DNA-Bound CRISPR–Cas Complex Using a Colocalization-Dependent Protein Switch. *ACS Synth. Biol.* **9**, 2316–2323 (2020).
3. Lajoie, M. J. *et al.* Designed protein logic to target cells with precise combinations of surface antigens. *Science* **369**, 1637–1643 (2020).
4. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).
5. Chen, Z. *et al.* De novo design of protein logic gates. *Science* **368**, 78–84 (2020).
6. Chen, Z. *et al.* Programmable design of orthogonal protein heterodimers. *Nature* **565**, 106–111 (2019).
7. Quijano-Rubio, A. *et al.* De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
8. Hsia, Y. *et al.* Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **12**, 2294 (2021).
9. Bermeo, S., Favor, A. & Baker, D. De novo heterotrimeric helical bundles. *Revis.*
10. Lee, S., Teich, E. G., Engel, M. & Glotzer, S. C. Entropic colloidal crystallization pathways via fluid–fluid transitions and multidimensional prenucleation motifs. *Proc. Natl. Acad. Sci.* **116**, 14843–14851 (2019).
11. Cockburn, D. W. *et al.* Structure and substrate recognition by the Ruminococcus bromii amylosome pullulanases. *J. Struct. Biol.* **213**, 107765 (2021).
12. Levy, E. D. & Teichmann, S. A. Chapter Two - Structural, Evolutionary, and Assembly

Principles of Protein Oligomerization. in *Oligomerization in Health and Disease* (eds. Giraldo, J. & Ciruela, F.) vol. 117 25–51 (Academic Press, 2013).

13. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion. *Science* **289**, 1546–1550 (2000).
14. Wilusz, C. J. & Wilusz, J. Lsm proteins and Hfq: Life at the 3' end. *RNA Biol.* **10**, 592–601 (2013).
15. Fromont-Racine, M. *et al.* Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95–110 (2000).
16. Khusial, P., Plaag, R. & Zieve, G. W. LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends Biochem. Sci.* **30**, 522–528 (2005).
17. Williams, G. J. *et al.* Structure of the heterotrimeric PCNA from Sulfolobus solfataricus. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **62**, 944–948 (2006).
18. Dowling, Q. & King, N. Computational design and assembly of massive protein nanocages. *Prep.*
19. Lee, S. *et al.* Design of T=4 de novo protein cages using pseudosymmetric hetero-oligomers. *Prep.*
20. Fallas, J. A. *et al.* Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).
21. Mohan, K. *et al.* Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. *Science* **364**, eaav7532 (2019).
22. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network–mediated specificity. *Science* **352**, 680–687 (2016).

23. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

24. Reinke, A. W., Grant, R. A. & Keating, A. E. A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *J. Am. Chem. Soc.* **132**, 6025–6031 (2010).

25. Dawson, W. M. *et al.* Coiled coils 9-to-5: rational *de novo* design of α-helical barrels with tunable oligomeric states. *Chem. Sci.* **12**, 6923–6928 (2021).

26. Lombardi, A., Bryson, J. W. & DeGrado, W. F. De novo design of heterotrimeric coiled coils. *Pept. Sci.* **40**, 495–504 (1996).

27. Sahtoe, D. D. *et al.* Reconfigurable asymmetric protein assemblies through implicit negative design. *Science* **375**, eabj7662.

28. Bolon, D. N., Grant, R. A., Baker, T. A. & Sauer, R. T. Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci.* **102**, 12724–12729 (2005).

29. Hallinan, J. P. *et al.* Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces. *Commun. Biol.* **4**, 1240 (2021).

30. Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J. Chem. Theory Comput.* **14**, 2751–2760 (2018).

31. Park, H., DiMaio, F. & Baker, D. CASP11 refinement experiments with ROSETTA. *Proteins Struct. Funct. Bioinforma.* **84**, 314–322 (2016).

32. Grigoryan, G. & DeGrado, W. F. Probing Designability via a Generalized Model of Helical Bundle Geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).

33. VanAernum, Z. L. *et al.* Rapid online buffer exchange for screening of proteins, protein

complexes and cell lysates by native mass spectrometry. *Nat. Protoc.* **15**, 1132–1157 (2020).

34. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007).
35. 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.
36. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis and testing. *Science* **357**, 168–175 (2017).
37. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
38. Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376**, 383–390 (2022).
39. Moyer, A. stapler.
40. Correnti, C. E. *et al.* Engineering and functionalization of large circular tandem repeat protein nanoparticles. *Nat. Struct. Mol. Biol.* **27**, 342–350 (2020).
41. Brunette, T. *et al.* Modular repeat protein sculpting using rigid helical junctions. *Proc. Natl. Acad. Sci.* **117**, 8870–8875 (2020).
42. Vanderstraeten, J. & Briers, Y. Synthetic protein scaffolds for the colocalisation of co-acting enzymes. *Biotechnol. Adv.* **44**, 107627 (2020).
43. Smulski, C. R. *et al.* Hetero-oligomerization between the TNF receptor superfamily members CD40, Fas and TRAILR2 modulate CD40 signalling. *Cell Death Dis.* **8**, e2601 (2017).

44. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLOS ONE* **6**, e20450 (2011).
45. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6**, 685–689 (1953).
46. Coventry, B. & Baker, D. Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLOS Comput. Biol.* **17**, e1008061 (2021).
47. Hiranuma, N. *et al.* Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1340 (2021).
48. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).
49. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
50. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013).
51. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
52. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
53. Caspar, D. L. D. & Klug, A. Physical Principles in the Construction of Regular Viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).
54. Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. Tomato bushy stunt virus at 2.9 Å resolution. *Nature* **276**, 368–373 (1978).
55. Rayment, I., Baker, T. S., Caspar, D. L. D. & Murakami, W. T. Polyoma virus capsid

structure at 22.5 Å resolution. *Nature* **295**, 110–115 (1982).

56. Liddington, R. C. *et al.* Structure of simian virus 40 at 3.8-Å resolution. *Nature* **354**, 278–284 (1991).
57. Johnson, J. E. Functional implications of protein-protein interactions in icosahedral viruses. *Proc. Natl. Acad. Sci.* **93**, 27–33 (1996).
58. Prasad, B. V. V. *et al.* X-ray Crystallographic Structure of the Norwalk Virus Capsid. *Science* **286**, 287–290 (1999).
59. Baker, T. S., Olson, N. H. & Fuller, S. D. Adding the Third Dimension to Virus Life Cycles: Three-Dimensional Reconstruction of Icosahedral Viruses from Cryo-Electron Micrographs. *Microbiol. Mol. Biol. Rev.* **63**, 862–922 (1999).
60. King, N. P. *et al.* Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336**, 1171–1174 (2012).
61. Lai, Y.-T., Cascio, D. & Yeates, T. O. Structure of a 16-nm Cage Designed by Using Protein Oligomers. *Science* **336**, 1129–1129 (2012).
62. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).
63. Lai, Y.-T. *et al.* Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat. Chem.* **6**, 1065–1071 (2014).
64. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535**, 136–139 (2016).
65. 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).

66. Ambroggio, X. I. & Kuhlman, B. Computational Design of a Single Amino Acid Sequence that Can Switch between Two Distinct Protein Folds. *J. Am. Chem. Soc.* **128**, 1154–1161 (2006).

67. Allen, B. D. & Mayo, S. L. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21375.

68. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A Generic Program for Multistate Protein Design. *PLoS ONE* **6**, e20937 (2011).

69. Kibler, Ryan *et al.* Stepwise design of pseudosymmetric protein hetero-oligomers. *Prep.* (2022).

70. Dowling, Quinton *et al.* Generating multicomponent, high T-number protein nanocages by hierarchical design of pseudosymmetric components. *Prep.* (2022).

71. Hallinan, J. P. *et al.* Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces. *Commun. Biol.* **4**, 1240 (2021).

72. Brunette, T. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).

73. Hsia, Y. *et al.* Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **12**, 2294 (2021).

74. Macromolecular modeling and design in Rosetta: recent methods and frameworks | Nature Methods. https://www.nature.com/articles/s41592-020-0848-2.

75. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

76. Fallas, J. A. *et al.* Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).

77. Sheffler, Will *et al.* Rpxdock: a fast and versatile computational method for protein docking. *Prep.* (2022).

78. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* eadd2187 (2022) doi:10.1126/science.add2187.

79. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

80. Chen, Z. *et al.* Self-Assembling 2D Arrays with de Novo Protein Building Blocks. *J. Am. Chem. Soc.* **141**, 8891–8895 (2019).