A New Generation of Idealized Protein Building Blocks Enables Rational Nanomaterial Design

Timothy Huddy

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

Jihong Bai

Ning Zheng

Program authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

A New Generation of Idealized Protein Building Blocks Enables Rational Nanomaterial Design

Timothy Huddy

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

Humans build complex structures by utilizing intuitive building blocks. Bricks, wooden planks, plumbing fittings, and most other building materials all have features that help humans to use them and share ideas of how to work with them. In contrast, the design of novel protein nanomaterials has been difficult due to the irregular shapes and behaviors of naturally-evolved proteins. Even when new protein complexes can be modeled, it remains a challenge to assign sequences to the models for them to form correctly when expressed. In this work, we have made a new generation of protein building blocks with features that are amenable for building in the protein world. The blocks are designed helical repeat (DHR) proteins with explicitly-defined brick-like geometry. They are built from alpha helices that are straight (not supercoiled at all) and have their helical axes all parallel to each other. This strict specification allows us to model protein assemblies with simple blueprint representations. With this new building block set, we have been able to design and characterize new protein assemblies with unprecedented modularity.

TABLE OF CONTENTS

1. BACKGROUND AND MOTIVATION

Since before there was even a completely solved crystal structure of a protein, the protein alpha helix (**Figure 1**) has been an inspiration for rational thought in structural biology. Biochemists in as early as 1951 were able to inspect data from amino acid chemical structures and predict reasonable “alpha helix” polypeptide conformations that would satisfy backbone hydrogen bonding patterns in a repeating structural motif [1]. As the protein structural biology field gained more data and grew, it became attractive to see if new proteins could be made based on rational interpretation of the existing data. The regular structure and sequence of helices in coiled-coils (particularly leucine zippers) inspired scientists to make “back of the napkin”-style designs where they portrayed alpha helices as rotatable bodies that could display side chains with predictable geometry, which encouraged intuitive design choices based on if a position would be facing solvent or how closely it would be interacting with other residue positions and backbones of adjacent helices [2].

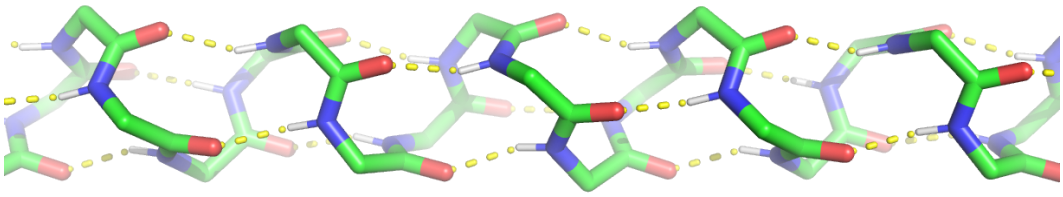


FIGURE 1. Model of a straight protein alpha-helix.
Here is a protein mainchain visualized with carbon atoms in green, nitrogen in blue, and oxygen in red. Yellow dotted line shows how this shape allows backbone hydrogen bonding between nitrogen and oxygen atoms.

It is not biochemically unfaithful to represent an alpha helix with a constant, linear central axis– in fact this has already been done both by design [3] and by nature [4], since this shape is in a reasonable Ramachandran space of torsion angles that helices have been seen to occupy.

Design of helices like this is often done with Crick parameters to describe the helical twist and coiling in an ideal way, so it follows that the knowledge to model these has existed since their discovery in the 1950s. Similarly, designing loops between helical bundles can be trivialized by using simple glycine and serine flexible linkers between adjacent helices with compatible terminal region directions [3]. So, why are there not many designs using very simple modeling of adjacent, parallel helices arrayed in 2D? This could enable protein design to accomplish 2D line-art style arrangements and even some 3D designs intuitively. Unfortunately, even when backbone modeling is reduced to a very simple form (**Figure 2**), the sequence design task remains formidable. The proteins must actually fold to the shapes that are described, and how is one to know which shapes are even accessible for a potential protein sequence to fold into? A primary goal of this work was to develop a design pipeline and set of building blocks that would enable us to reliably achieve target structures in this realm of shapes.

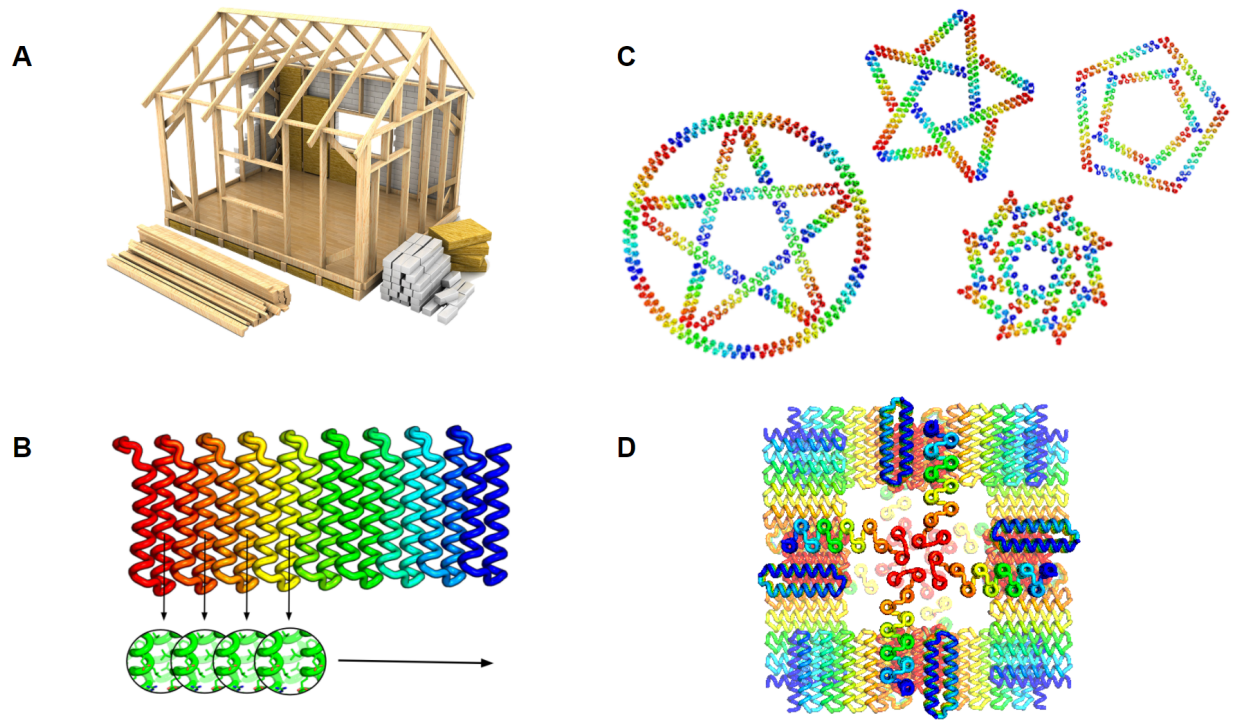


FIGURE 2. Schematic of simplified building.
A – *A cartoon of a house frame illustrates that simple wooden 2 by 4 planks can be used to build a complex structure*
B – *A rendering of a designed repeat protein shows how repeated interactions between identical helices can provide us with a protein shape that functions similarly to the 2 by 4 planks in the sense that it is straight and can be used at a variety of lengths*
C – *Illustration of the compatibility of blocks such as in (B) to be combined together in a line-art fashion*
D – *Ideality of building blocks can be applied to more complex structures if geometry is kept in mind, such as this octahedral cage that features straight DHRs coming from tetramers to make interactions at 90 degrees. In this shape, extension of the DHRs allows the whole cage to grow without disrupting geometry*

To add some perspective to the complexity of nanomaterial design, if you likened a house frame made with wooden planks (2-by-4's) to a protein structure, you could consider most house frame designs to be asymmetric as a whole (with some regular, repeated interactions), very high order (many unique wood pieces of varying length), and highly convergent/favorable in the sense that a blueprint can be made, understood, and used to produce the correct structure most of the time. Consider the ability of humans to design new protein complexes with these features, and there will be woeful underperformance when compared to the house frame. One huge

benefit that the house frame has is that it utilizes very ideal building blocks. When designing the structure, those wood pieces can be thought of mostly as linear objects of customizable length. These blocks can also reliably be joined together both at their ends and at any position along their length (such as a “T” shape junction). In order to simplify the design of new protein nanomaterials, we began our journey by making a new generation of protein building blocks with similar geometric features to the wooden planks.

The primary motivation for this project is one of basic science; that we should be able to design very idealized, sensible systems if we truly understand protein design. Having a simplified backbone generation and placement regime means that we can make geometrically-idealized design models that can, for example, close into a cyclic ring with zero closure RMSD. Then the goal becomes to determine which backbones can actually possibly work for this, and to assign correct sequences. Upon gaining enough insight in this area, we should be able to rationally design nanomaterials by explicitly defining relationships between blocks so that more complex interactions can be modeled besides just using straight sections.

A secondary motivation for this project is that protein nanomaterials design is lagging behind where it could be. DNA nanotechnology and origami has had the benefit of decades of research for bringing it up to the scale of industrial applications. Proteins can hold onto genetic fusions easily or even be engineered to directly interact with inorganic and organic substrates [5], whereas DNA nanotechnology has difficulty in doing these things rigidly. Improvements in the “buildability” of protein nanomaterials would grant them access to interesting industrial spaces, many of which have likely not even been fully considered due to old perceptions of what protein design can accomplish.

Protein nanomaterials also hold promise in the biomedical realm. Subunit vaccines from designed protein scaffolds offer an interesting ability to increase the potency of displaying immunogens to immune systems [6], and it will be beneficial moving forward to have a larger, more modular library of protein nanomaterials with which to do this. Designed protein complexes can also be useful scaffolds for a wider variety of ligands for a unique interaction with cell biology [7, 8], and this space will require improvements in protein complexity and size to reach new heights of multiplexed signal display.

2. TECHNOLOGIES AND METHODS

Computational: Generating images of protein models

All images of new protein models were made with: The PyMOL Molecular Graphics System, Version 2.0+ Schrödinger, LLC.

Computational: Backbone design

This project relies on using realistic models of straight alpha helices. We use Rosetta protein design suite, particularly the RosettaScripts implementation [9] to generate “blank slate” helical backbones that are generated from sets of Crick parameters [10, 11]. We tried two different parameter sets for straight helices. These are “alpha_helix” and “alpha_helix_100” parameters files from RosettaScripts; which both feature near ideal helical geometry with respect to hydrogen bonding and strain, but the _100 option has an exactly 100 degree turn per residue

(when looking “down the helix”), such that it can end up in perfect phase repetitions after 18 residues. 18 residues is the first multiple of 100 that ends up equal to an integer multiple of 360 degrees, meaning that a residue orientation ends up in the same phase again.This happens after the helix has gone through 5 turns.

18 residues * 100 degrees = 5 * 360 degrees

No difference was found in the computational design metrics on these 2 backbone sets or in experimental data, so either can be used. Both can be used to increase sampling.

Next, these helices were typically used as secondary structure elements in new straight DHR designs. The MakeBundle mover in RosettaScripts [12] was used to generate multiple helices (this can also be done in PyRosetta). Through the RepeatProteinPropagation mover [13], we can make repeat proteins from any structure that features the same structural elements represented more than once. So, the general pipeline for making a repeat protein that has explicitly defined repeat geometry is (**Figure 3**):

1) Decide on one of the secondary structural elements you want to repeat. In this case, it is newly generated straight alpha helices.
2) Make a clone of the structural element that represents the first time that element is repeated. The geometry between the 2 elements defines the trajectory of your repeat protein.
3) Add in additional secondary structural elements between the first “i” and “i+1” structural elements so that there is enough protein density to have a protein core between structural elements.
4) Loop them together and now you have a complete repeat unit + 1 copy of 1 of the structural elements (in this case an extra straight alpha helix). This is the minimal unit you need to create to make a repeat protein, as shown in **Figure 3 E**.

5) Use the RepeatPropagationMover and set the “start_pose_duplicate_residues” flag equal to the number of residues that you have extra, which equals the residue number of the structure clone added in (2). Set “numb_repeats” equal to how many repeat units you want outputted after. This generates a repeat protein backbone for you.

6) Perform sequence design on the backbone- detailed later.

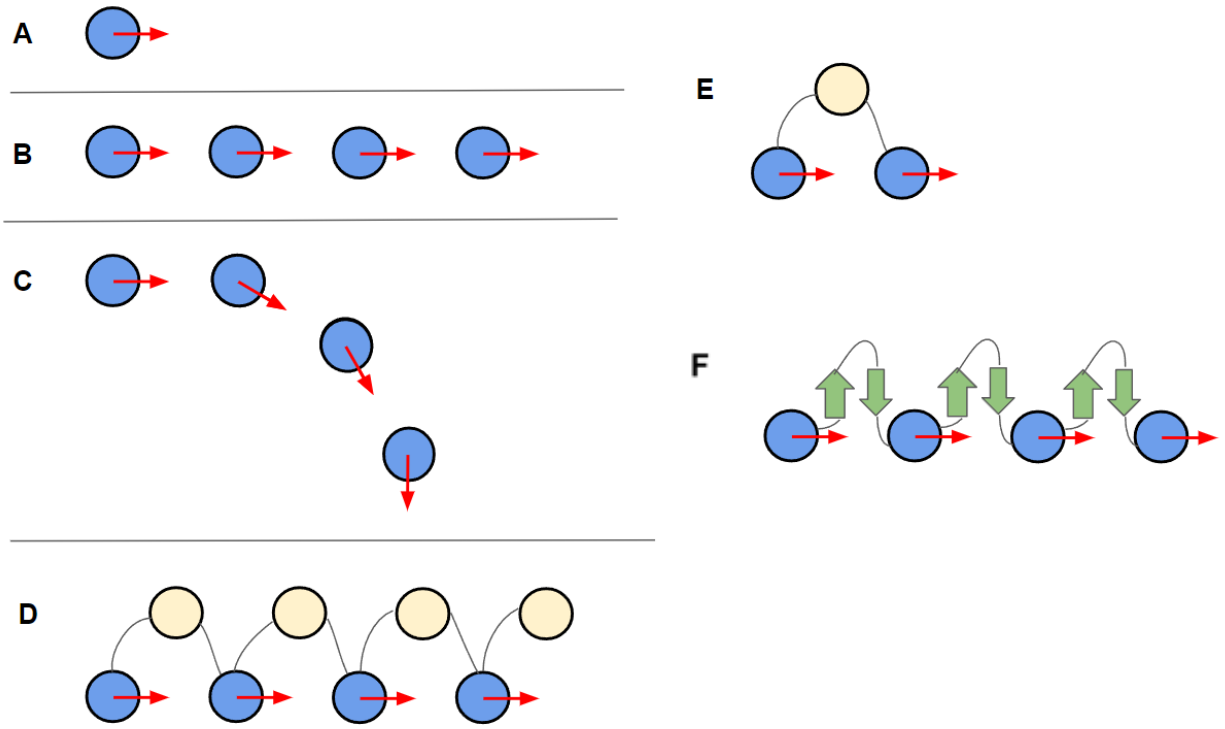


FIGURE 3. Method for making repeat structures of defined geometry.
A – *One can begin with a secondary structure element of choice, in this case we are looking down the axis of an alpha helix, which has an arbitrary rotational phase represented by an arrow*
B – *If a copy of the first structural element is made and displaced in some direction without changing the phase, then we can make repeat structures that extend linearly*
C – *If a copy of the first structural element is made and displaced in some direction with a defined change in the phase, then we can make repeat structures that curve with controlled dimensions*
D – *A complete repeat protein features the defined element for setting the repeating register plus additional structure to either make a sufficient protein core for rigidity, or to add functionality to a defined shape plan*
E – *The minimal backbone required to be laid down in order to use Repeat Propagation. This is 2 copies of the defining repeat structure element (blue color) + all of the structure between the 2 copies (cream color)*
F – *An example to show that this is not limited to all alpha helical proteins. Here, 2 beta strands are used between repeating helices. In fact, these same methods could be used with an all-beta protein without needing to do any adaptation.*

Based on examination of packing in previously designed helical bundles, it was decided to use 10 Å as the distance offset between helices to allow a good amount of room for hydrophobic packing interactions to occur in the core of the protein. To save on the amount of backbone sampling done, this distance was only coarsely sampled at increments of 9.5, 10.0, and 10.5 Å, while the bulk of the sampling for sake of finding good helix packing solutions was done by rotating helical phases relative to each other so that side chains would have their C-alpha to C-beta vector positions and directions sampled for how they project toward adjacent structural elements. This style was held for the bulk of the design done for this project. An aside is that this distance can be finely tuned for application. One generation of these was made with a repeating distance of 8.7 Å in order to match an inorganic material lattice parameter [14]. In all cases of these all alpha-helical topologies, loops were placed with default options of the ConnectChainsMover in RosettaScripts [15], which uses fragments of helix-loop-helix structure from natural protein solved structures to align with unlooped backbones for grafting in structured loops.

Computational: Sequence Design

At the start of the project, sequence design was attempted in many different ways within the capabilities of RosettaScripts. This worked for us with some designs, including several that yielded crystal structures. However it fell short with some of our more complex assemblies, so we recently adopted a new strategy- for this reason we will forgo detailing of the RosettaScripts methodologies here.

What was used recently instead is an in-development design script using a message-passing neural network (MPNN) that was trained on solved structures and sequences of protein complexes such that it could learn the language and spatial relationships of protein sequences

in order to be able to realistically assign sequence to new designed structures [16]. Default parameters for design were used, which are likely to be held very similar in the upcoming publication. In order to enforce concepts of symmetry or repeat design, residue identities were tied to each other with input pdb models that contained the respective symmetric copies of all chains or all the desired repeats of a repeat protein, respectively.

Computational: Design choice and filtering

Similar to the previous situation with sequence design, this project began with many Rosetta filters being used. Efforts were made to predict packing quality (hoping that energy of designed state was so low that it would be a preferred one) and secondary structure propensity (are helix-loop-helix sequence choices looking appropriate compared to natural proteins?), and to do some Rosetta forward folding for structure prediction [17]. All of these enrich for working designs to some extent, but are not hugely powerful. Even being very stringent on these, there are some topologies (some of our rings) where the structure prediction power with these methods is not strong enough to give us a lot of working designs because we cannot accurately assess with enough precision if separate interactions that are long distances apart will be likely to form with the right orientation or not.

Again, we have had fortune from machine learning advances. Now, we just take sequences from the MPNN sequence design and feed them into AlphaFold 2 [18]. Depending on the type of design being evaluated, we may input multiple chains to see if they are predicted to oligomerize correctly. There are several metrics from AlphaFold2 that can be used to assess the prediction quality, but we used its averaged position-specific confidence term (pLDDT) and the predicted structure models to compare RMSD to the design model and determine if the prediction was confident in that structure or not. In the future, we will investigate using more varied metrics. The

MPNN and AlphaFold2 are powerful enough that we can often achieve correctly folding structures experimentally with just these 2 methods for sequence design and evaluation.

In most cases of designs shown in this work, there were many, many similar looking samplings of helix phases and distances between each other and only a select few had the best packing and/or folding to be chosen. An aside- models 4 and 5 of Alpha Fold 2 have typically given us the best pLDDT scores for these straight helix topologies.

Computational: Modeling of protein complexes

Two methods were used for the majority of work done to design protein complexes from building blocks. In addition to the new straight DHR blocks, these methods also used some designed heterodimers and cyclic oligomer helical bundles.

The primary method used is WORMS, which takes in databases of building blocks that can be fused together by overlapping secondary structure. It allows for terminal regions to be trimmed so that many secondary structure elements can be scanned for potential overlaps, and then it checks if any fusions satisfy geometric criteria for a building goal [19]. DHRs are particularly powerful for this method because they can be input as a long chain with many repeats, and many potential sizes of the DHR can be sampled for satisfying geometric criteria, since they have an intact core whether they have only 2 repeats or as many as 16 repeats or higher. For example, with making interfaces in cyclic oligomers, DHRs can either be fused to heterodimers to use those interfaces repurposed as symmetric interfaces, or they can simply have one of their loops cut out so that it becomes a “split” DHR with an interface in the middle of it (“Crown” architecture for both of these, as described in the paper [19].

Besides the WORMS method of fusing chains together to make the appropriate shapes, there is also the ability to move and dock blocks together in appropriate spatial relationships. RPX Dock is a tool in development by Baker lab [20] to rapidly sample 3D space to find and score motifs from rigid body docking in many symmetries. For example, a high scoring dock for a Cyclic-3 symmetric design may feature 3 copies of a straight DHR docked head-to-tail in a triangle shape, which may have been selected for finding many motifs containing Valine, Leucine, or Isoleucine interactions on adjacent helices between copies of the DHR chains. The presence of residue pairs across interfaces at appropriate distance and angle relative to each other represents a good opportunity to find packing solutions because those orientations are the same as ones that held nice packing residues of good score in a training set of interaction motifs.

Computational: Codon optimization and construct design

For most of our designs, we have let the gene synthesis company (Genscript) do the codon optimization for expression. We have found similar results and greater versatility for big constructs by doing it ourselves with DNA Chisel [21]. This tool allows us to balance the incorporation of rare codons (typically bad for expression- some can cause stalling in translation) with synthesizability. Synthesis of large (> 2000 bp) and highly repetitive sequences will have a lot of failure unless we de-repeat the DNA sequence. Controlled incorporation of rare (but not the rarest) codons helps with this. A proxy that we used to see if our sequences were synthesizable enough was to use the complexity score from IDT's option to "Test Complexity" with custom gene synthesis orders [22].

Almost all DNA sequences we made were inserted into pET-29b(+) in E. coli between NdeI and XhoI restriction sites. This allows our genes to be under control of T7 promoter, and for us to have kanamycin resistance for colony selection after transformation. We used BL21 cells and

expressed with autoinduction [22]. This allows us to grow the culture with a pre-planned amount of dextrose and lactose in the media so that expression will be induced automatically after the culture has grown to peak density and metabolic rate.

Experimental: Protein production and purification

Transformed BL21 cells were usually inoculated into 50 mL of Terrific Broth II culture media with ~ 10 colonies and then shaken in an incubator overnight at 37 °C for 16-24 hours. Culture is collected into 50 mL conical tubes and spun at 4,000 x g to pellet the cells. Cells are resuspended in lysis buffer (300 mM NaCl, 20 mM Tris-HCl, 20 mM imidazole, pH 8) and then sonicated to lyse. Lysate is clarified by centrifuging at 14,000 x g. The soluble lysate supernatant is decanted into equilibrated IMAC columns (Ni-NTA resin and lysis buffer), washed with 10 column volumes of lysis buffer, and then eluted with elution buffer (50 mM EDTA, 300 mM NaCl, 20 mM Tris-HCl, 300 mM imidazole, pH 8).

For some samples such as fibers, the pelleted lysate was resuspended in 6M guanidine-HCl, which solubilizes some of the protein through chaotrope action/ denaturation so that it can be purified. Soluble protein in guanidine is dialyzed against sample buffer (50 mM NaCl, 25 mM Tris-HCl, pH 8) either individually or in presence of interaction partners so that interactions and folding could occur gradually during chemical annealing process.

Experimental: Protein characterization

Protein expression and purity was primarily assessed with native PAGE or SDS-PAGE, with follow-up mass spectrometry to confirm promising samples. Protein size and oligomeric state was measured with gel filtration on ÄKTA size exclusion chromatography instruments.Samples

that were promising at this stage were either concentrated for crystal tray setup, or looked at with negative-stain electron microscopy on a Talos L120C TEM. Negative stain datasets were averaged with cisTEM [23]. Collaborators performed cryo EM data collection on some designs that had convincing negative-stain EM data.

3. RESULTS

Characterization of new straight DHRs

For this building project, the first step was to design and characterize straight DHRs to determine if we could reliably design them to be as flat as needed so that they could act as the simplified linear materials in our building efforts downstream (**Figure 4**).

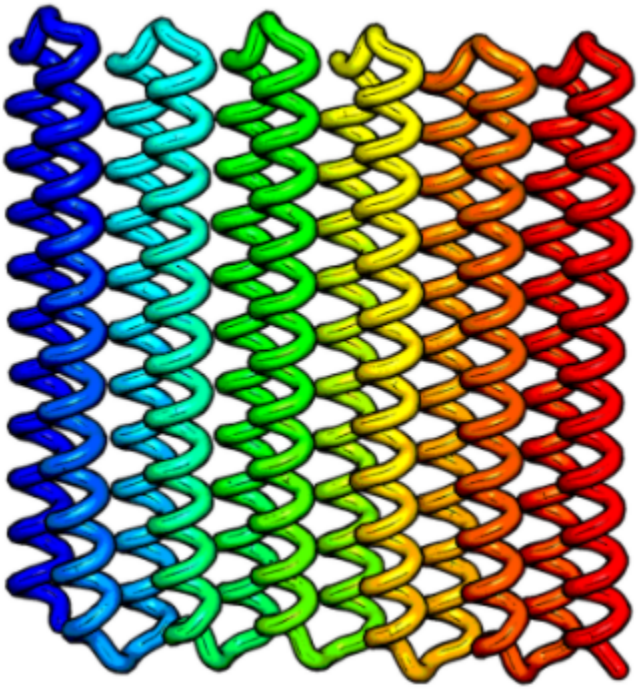


FIGURE 4. Model of one of the first tall repeat proteins ordered.
Colored in chain rainbow, the N terminus is at the blue end, and the C terminus is at the red end in this single-chain protein.

For the first gene order, a set of tall DHRs was made, with helices of ~ 40 residues long. All of them were expressed, and we checked to see if there was high yield of soluble product after IMAC by running SDS-PAGE (**Figure 5**)

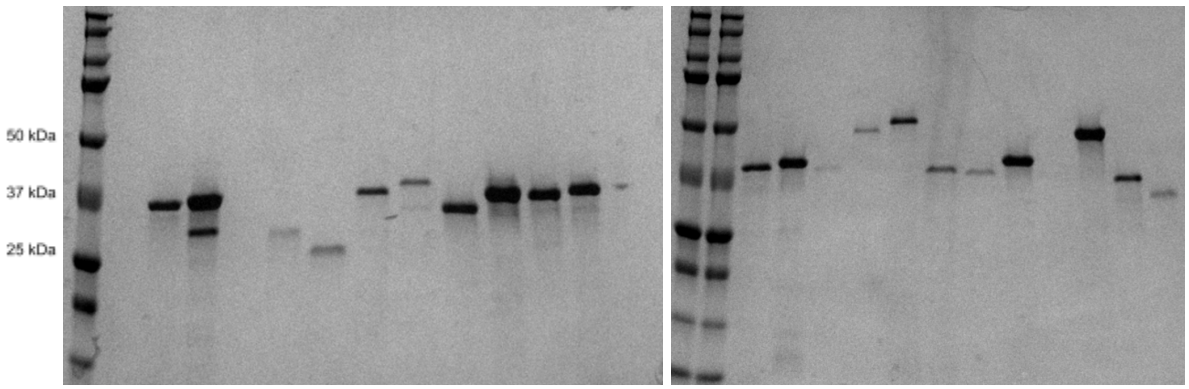


FIGURE 5. SDS-PAGE for IMAC-purified tall DHRs.
On left of each image is the Precision Plus protein ladder standard. Each other lane is a separate His-tagged DHR sample after eluting from Ni-NTA resin. Samples that yielded strong bands here ultimately expressed well enough to give enough protein for crystal tray setup from 50 mL cultures.

~16 of 25 expressed designs yielded an IMAC-elution band that was pure and intense enough to move forward with. 8 of 16 that were tested next showed good monomeric behavior on SEC S200 column at the expected elution volume, such as shown with sample THR10 (**Figure 6**)

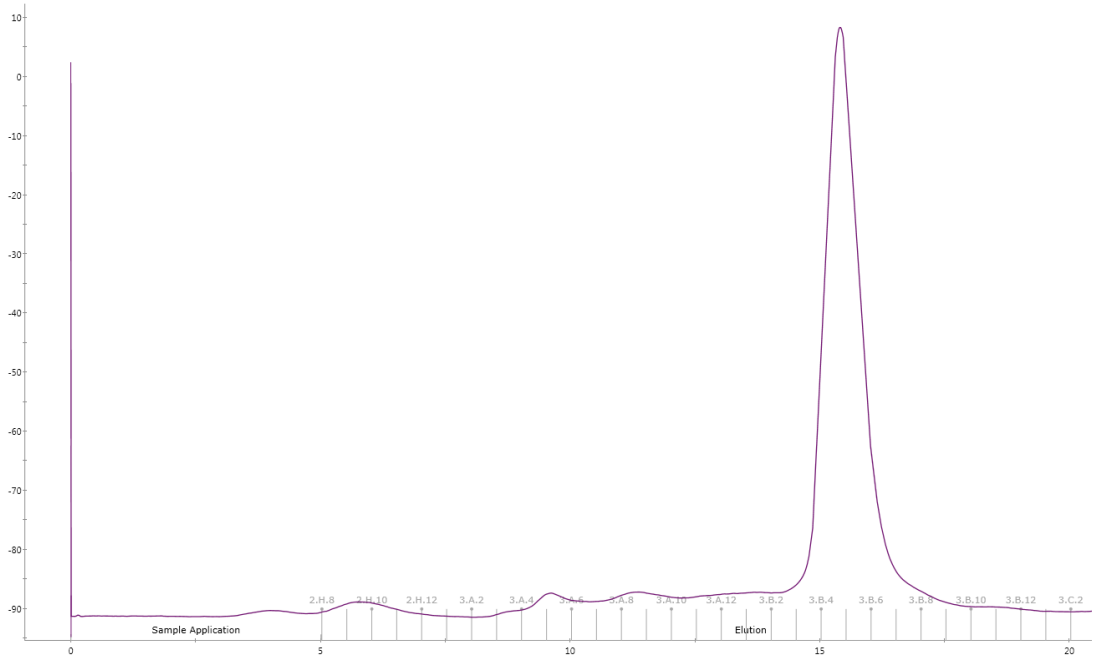
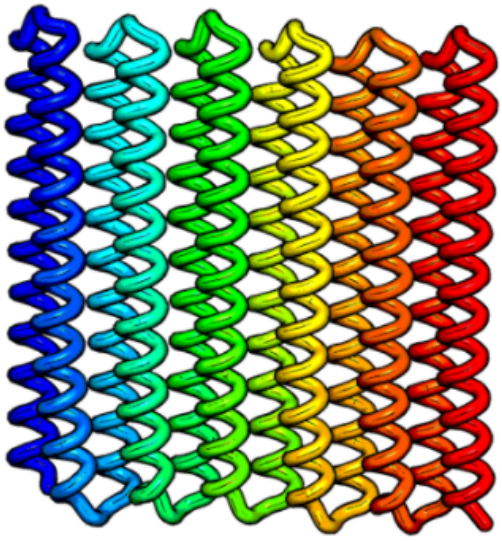


FIGURE 6. SEC elution profile for an example well-behaved monomeric DHR
Samples were run on S200 10/300 GL Increase column. Here the peak after 15 mL is sharp and represents a clean monomer according to comparison with size standards.

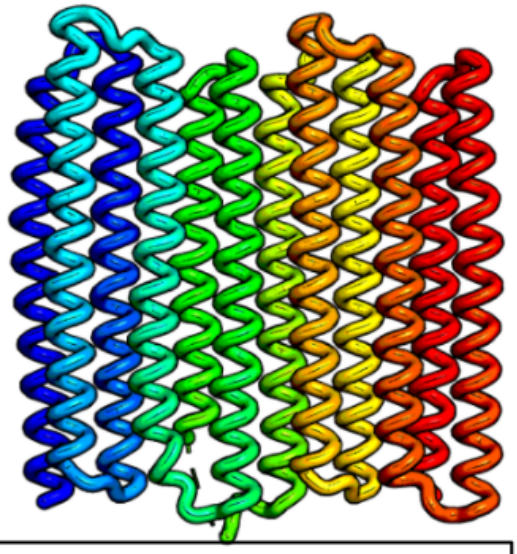
Successes at this stage were considered good enough to be used as building blocks for new designs that hopefully would validate them further. Crystal trays were set up for these designs and also for another set of repeat proteins with shorter helices, and ultimately we ended up with 4 crystal structures of straight DHRs (**Figure 7**).

35_DHR



2.6 Å RMSD xtal structure
0.6 Å RMSD to design model

37_DHR



2.8 Å RMSD xtal structure
1.3 Å RMSD to design model

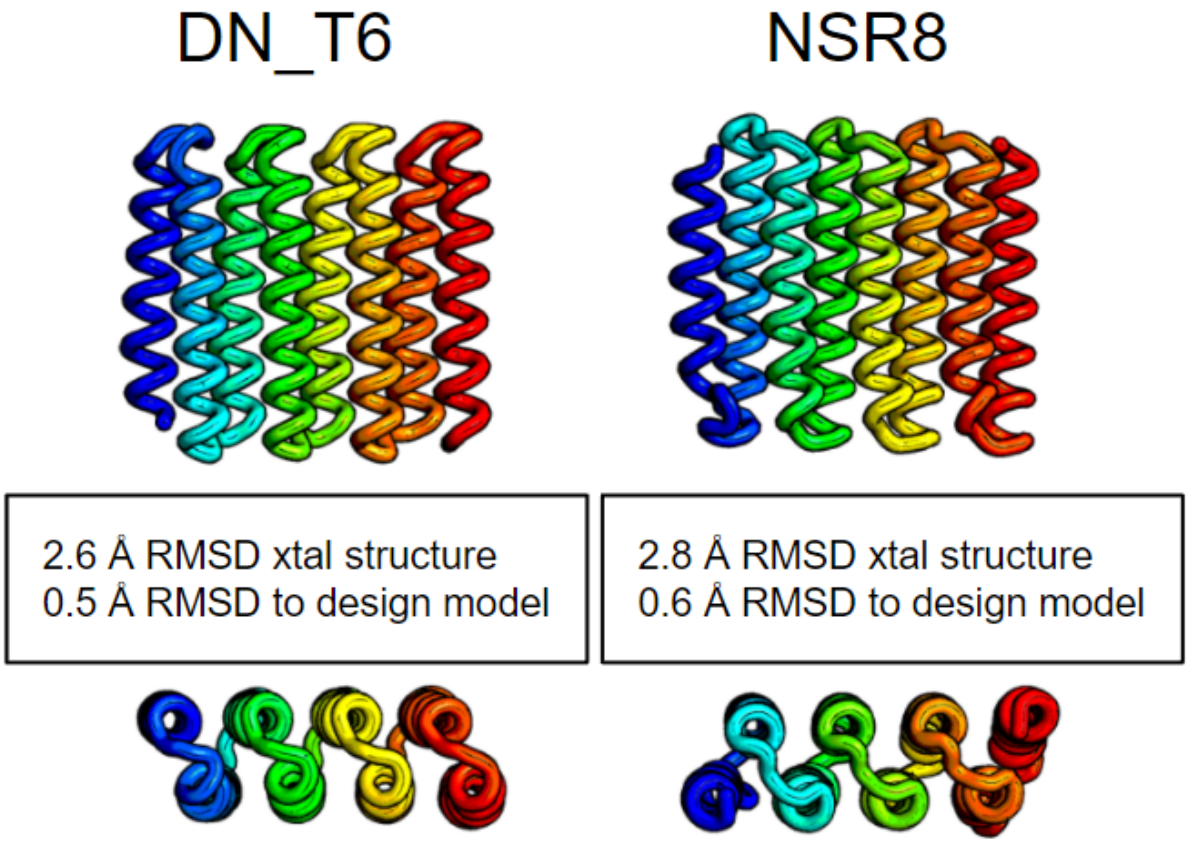
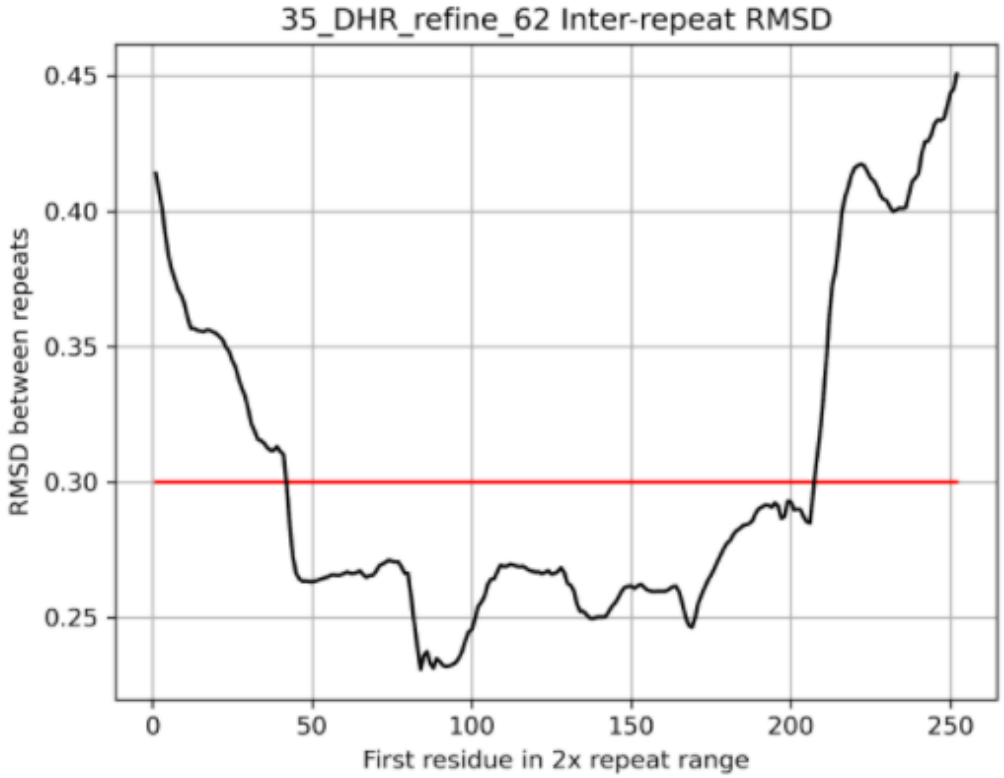


FIGURE 7. Crystal structures of 4 straight DHRs.
Here are 4 straight DHRs with the side and top views of their crystal structures shown along with their crystal resolution and full-atom RMSD to design models.

These structures looked overall how we hoped, but we wished to assess the straightness further. Harley Pyles in our lab took the experimental structures and determined what sliding window of 2 repeat units within the 4-repeat structure yielded the lowest RMSD between the 2 repeat units if overlaid. This is an attempt to make a perfectly repetitive structure informed by the crystal structure data as best as possible. The middle regions repeat better and the terminal regions deviate a bit (**Figure 8**).


35_DHR_refine_62 Inter-repeat RMSD
RMSD between repeats
0.45
0.40
0.35
0.30
0.25
0
50
100
150
200
250
First residue in 2x repeat range

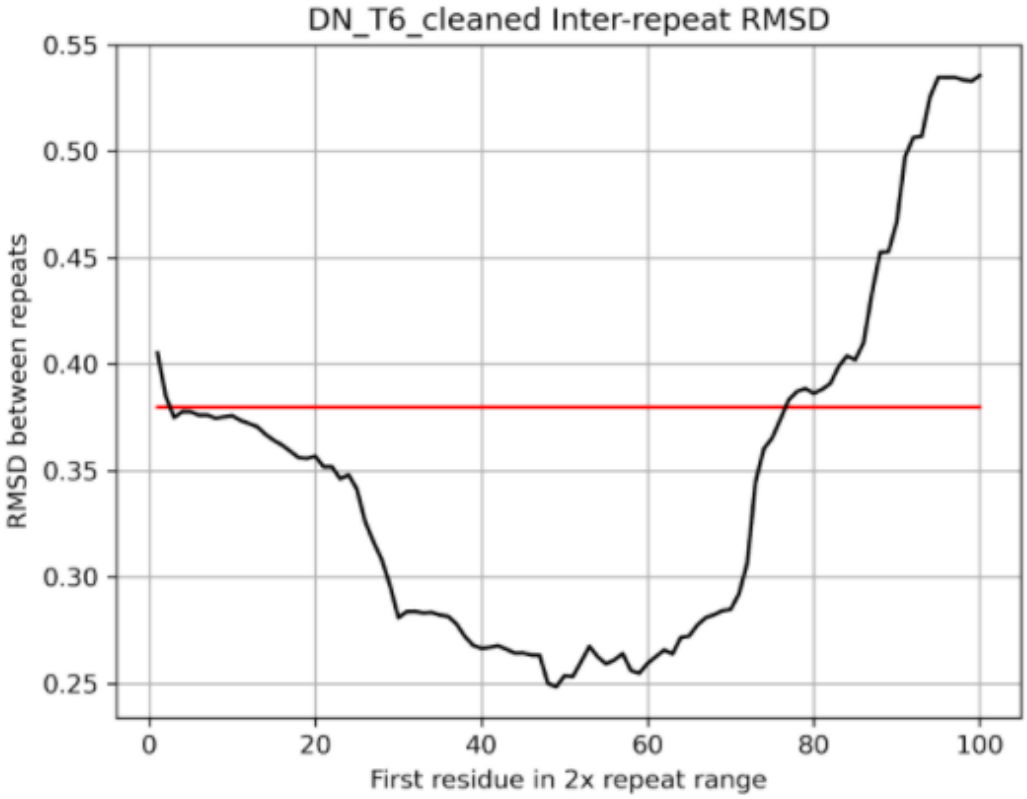

DN_T6_cleaned Inter-repeat RMSD
RMSD between repeats
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0
20
40
60
80
100
First residue in 2x repeat range

FIGURE 8. Plots to find the most repetitive section of DHR crystal structures
2 plots are shown here, each for a different protein. In both cases, the most consistent consecutive repeats were found near the middle of the structure. In both cases, the best 2 repeat units were very close to matching, with < 0.25 Angstrom RMSD in the experimental data.

With data from these, we generated models of the repeat proteins that fit the data and also were perfectly repetitive. In the case of sample DN_T6, we had a particularly straight protein model emerge. When the structure was propagated to 8 repeat units, deviations from perfect straightness were well satisfying our goals (**Figure 9**).

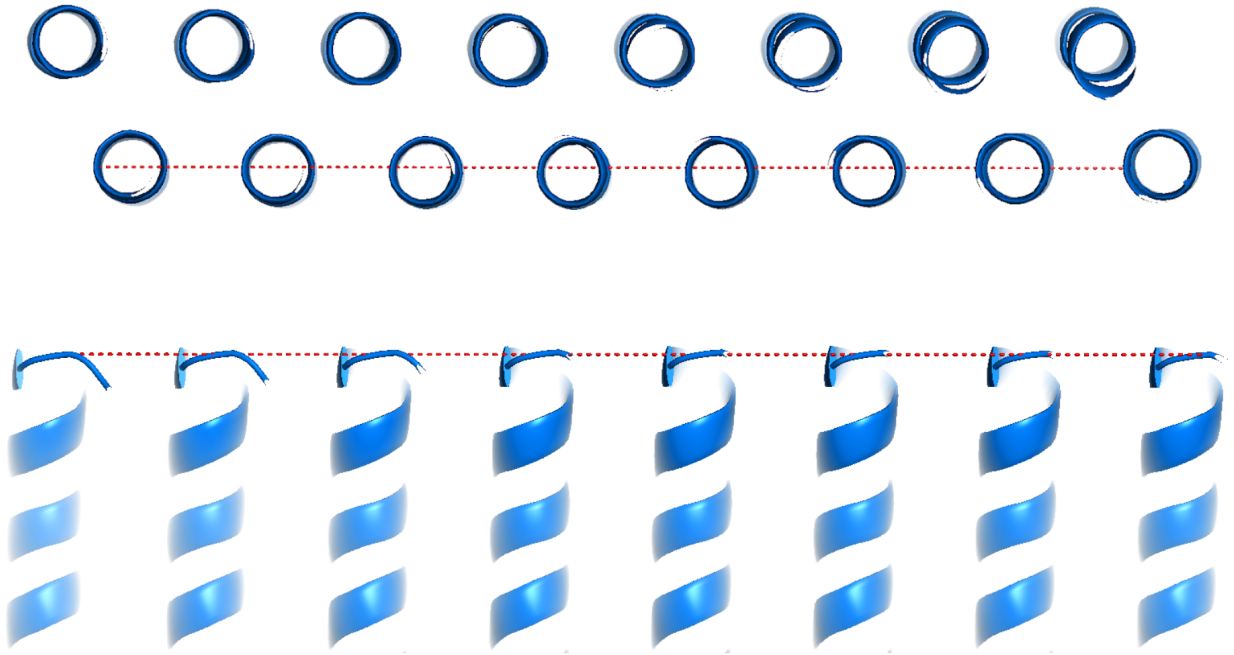


FIGURE 9. Straightness of repeat-idealized model of DN_T6 crystal structure.
Here are top and side views of 8 repeat units of DN_T6 as informed by the best repeating region of crystal structure (some structure is faded from view for clarity). Lines from beginning to end show that bowing/curving in either direction with 8 repeat units is well under 0.5 Angstroms

A caveat to these results is that crystal packing can influence the straightness of the experimental data. These bricks tend to pack particularly well against each other in crystalline form, usually with side-by-side docking that encourages straightness. However, more and more

DHRs were validated as parts of larger protein assemblies as this project pushed forward so this is unlikely to be suggesting any gap in our ability to achieve very straight, brick-like proteins.

Straight heterodimers added to the building toolbox

As we gained confidence that we could reliably make these straight helix structures, we moved on to make a set of heterodimers with similar helical features (**Figure 10**).

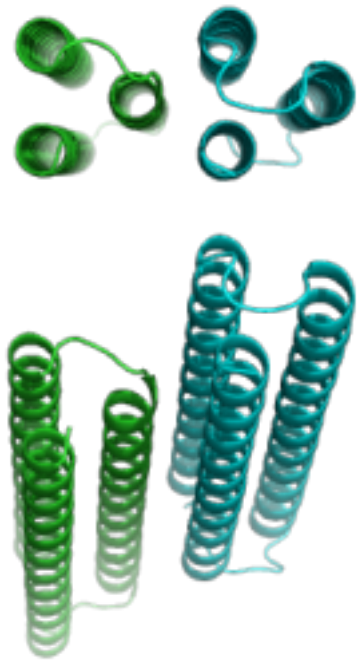


FIGURE 10. Model of a straight helix heterodimer.
These proteins were designed to be sets of paired 3-helix bundles that interact with each other via a straight/parallel interface where 1 helix from 1 bundle interfaces with 2 helices of the other bundle.

These were made with an interface of 1 helix interfacing with 2 so that they hopefully would have less propensity to homo-oligomerize than if both sides had the same number of helices interfacing with each other (because the interface on each side “expects” to see a different number of helices on the other side after designed). Also, the hope with this layout was that the

“1” side of the interface could represent the possibility for having interface residues grafted onto any straight helix in a structure that is particularly solvent exposed so that it could make this interaction, since only the outside residues of that 1 helix are needed for the green chain to form the interface (such as if a helix were at the outside corner of a 90 degree turning repeat protein junction, it could get the interface area grafted to act like the interaction helix of the green chain). Hydrogen bond networks were designed in the interface to give this more specificity of interaction than if they had just been all hydrophobic.

These were tested for association of the 2 chains by native gel. Chains were run both separately and combined to see if a new band appeared to represent a new species migrating differently through the gel (**Figure 11**).

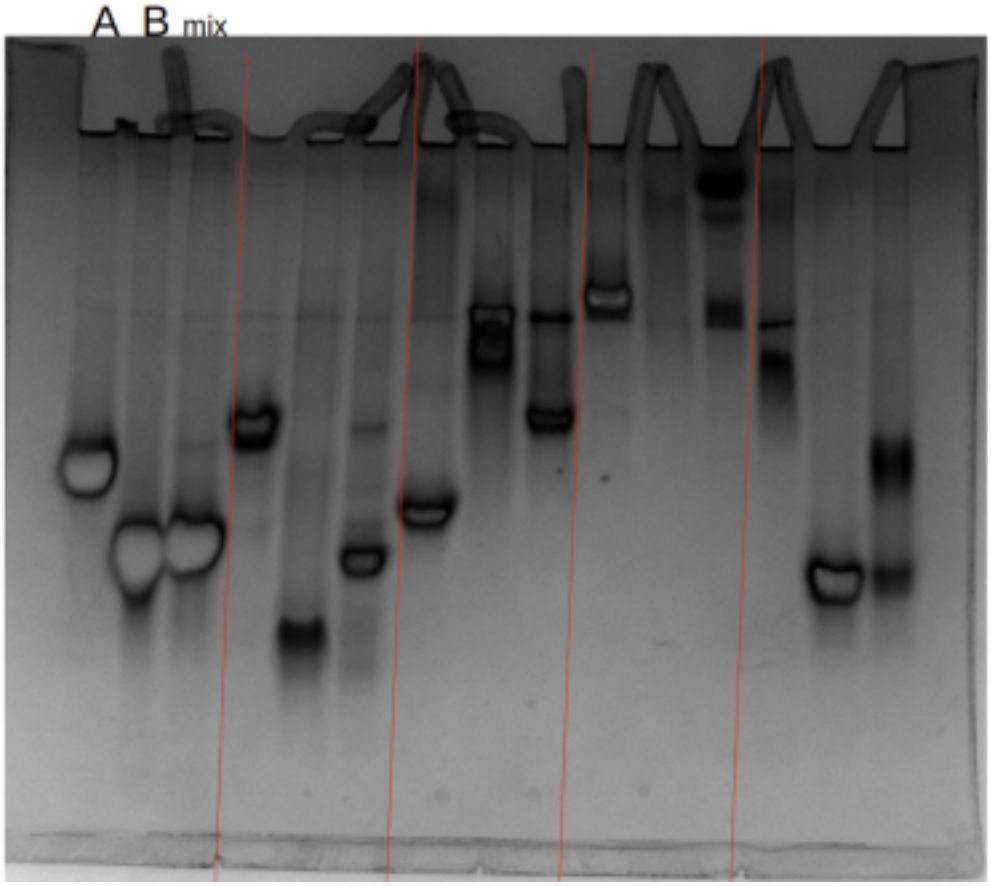


FIGURE 11. Native gel showing heterodimer interactions

Here we have data from 5 selected heterodimer pairs. Each sample has 3 lanes shown with the first lane being chain A alone, middle lane being chain B alone, and the third lane being the mixture of the two. Native gel shows differential migration of proteins based on both their folded size and charge, so a heterodimer will migrate as 1 thing differently from either component. Interaction can be shown by the appearance of a new band in the third lanes.

After we had obtained 10 designs that appeared to interact by native gel, I attempted an orthogonality matrix experiment with all-by-all combinations on many native gels (**Figure 12**)

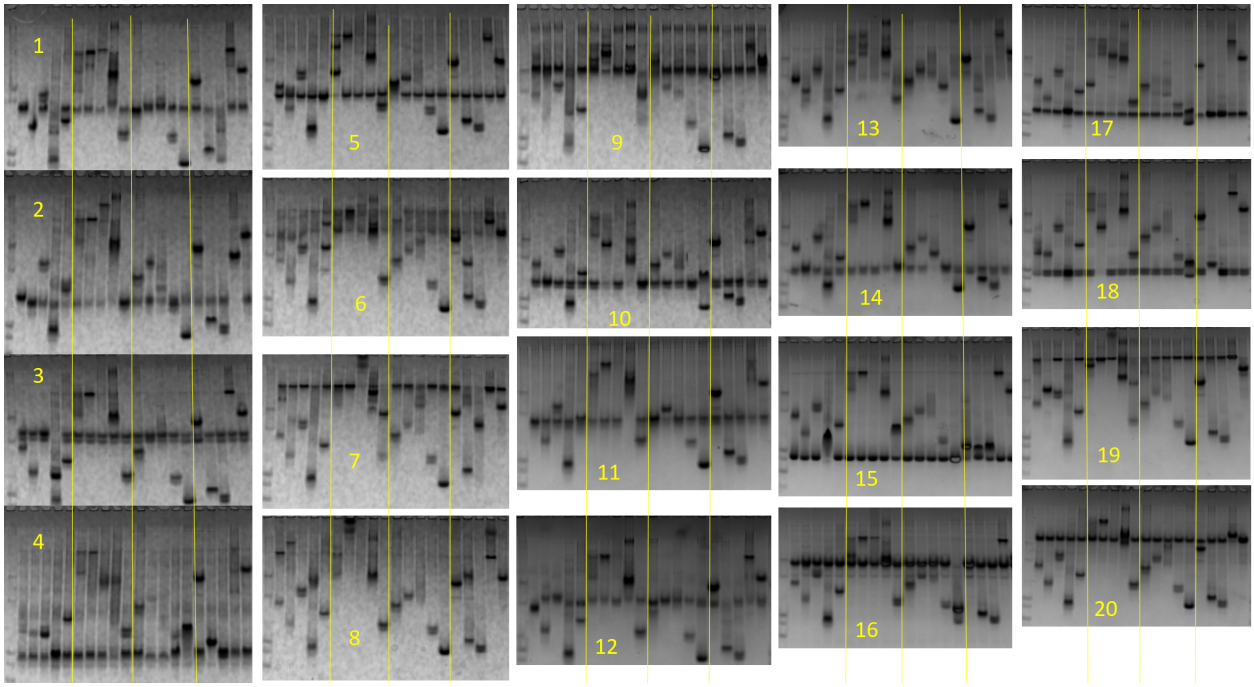


FIGURE 12. Native gels of each chain of 10 heterodimers, all by all.
Here 10 heterodimers that previously showed assembly evidence had each of their chains tested for assembly against all other chains. There are 20 gels and 20 lanes in each gel. Gel 1 has chain "1" in every lane and another sample added to each lane depending on the lane number. Lane 1 in each gel has chain "1", Lane 2 has chain "2" etc. Chain 1 and 2 are from one heterodimer design, chain 3 and 4 are from another one, and so on.

This data was analyzed qualitatively to generate an orthogonality matrix for these heterodimers (**Figure 13**). This yielded 13 potential combinations of 3 heterodimers that should not have unintended cross-talk if all were used at the same time in a protein complex (neither A nor B chain of any of the 3 heterodimers shows any non-cognate interaction).

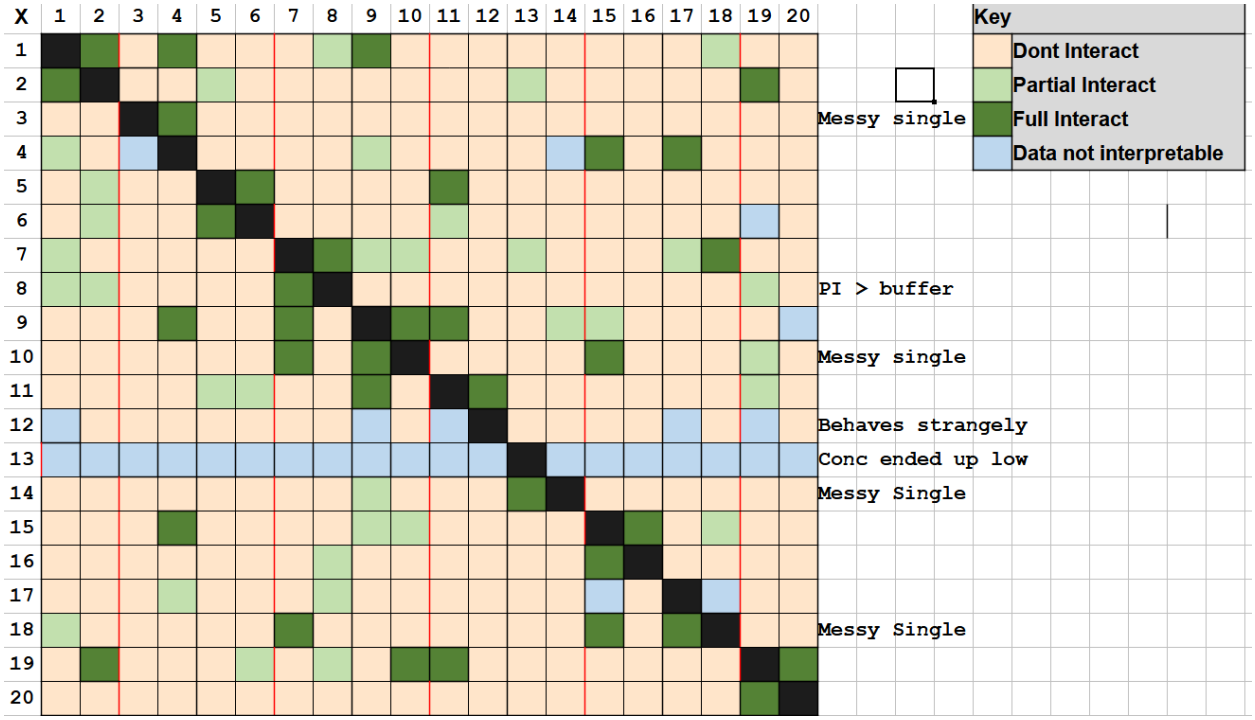


FIGURE 13. Qualitative orthogonality matrix from heterodimer data
The data from Figure 12 was interpreted and plotted here to represent interactions. Some gels had interesting features with their respective samples, for example the label "Messy Single" is used when an input chain displays behavior that shows evidence of more than one species ,likely exchanging between homodimer and monomer. Data is not mirrored across the diagonal- each interaction is represented twice in the native gel data. Rows are gel numbers and columns are lane numbers

With structurally verified repeat proteins and some new heterodimers made to be compatible with the repeat proteins, we expected to be able to make several robust designs, hopefully being able to reach higher orders of cyclic things than we could before (**Figure 14**).

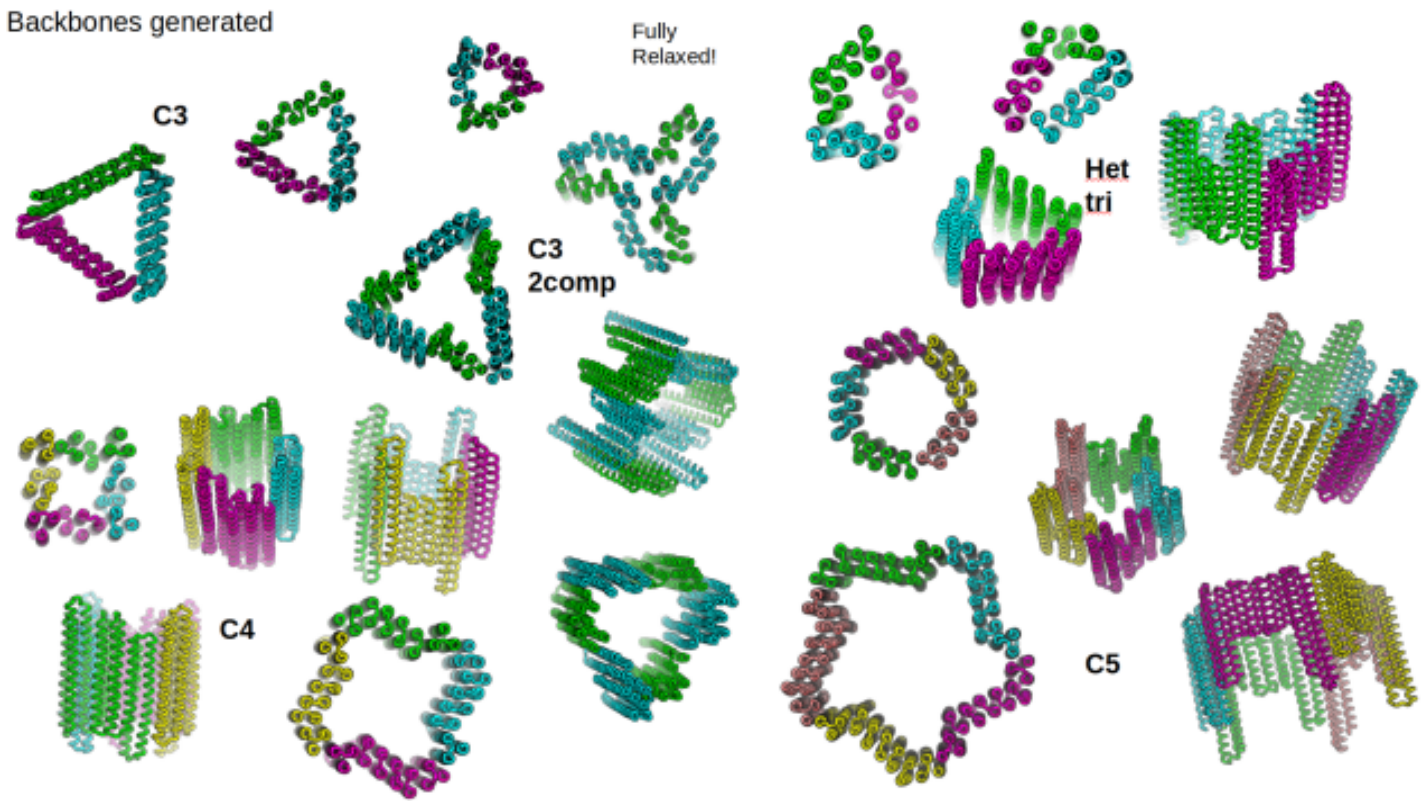


FIGURE 14. Varieties of structures that can readily be modeled with our tested building blocks.
Here there are more cyclic things including interesting 2 component C3. There are also asymmetric heterotrimers which use straight DHRs with combinations of heterodimers informed by the orthogonality matrix data

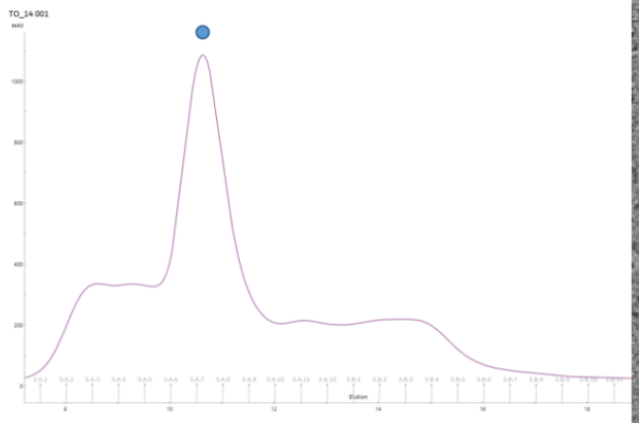
Unfortunately, not much time was spent on testing these kinds of designs. As the machine learning revolution advanced on us, we became wary of using this heterodimer set because their straightness was unlikely to be realistic according to AlphaFold 2 predictions. Still, this experimental pipeline will be useful to repeat with new structures for which we have high confidence in structural accuracy. Despite not being used to reach complex designs like the heterotrimer, some of these heterodimers were successful in giving us the structures shown in the next section.

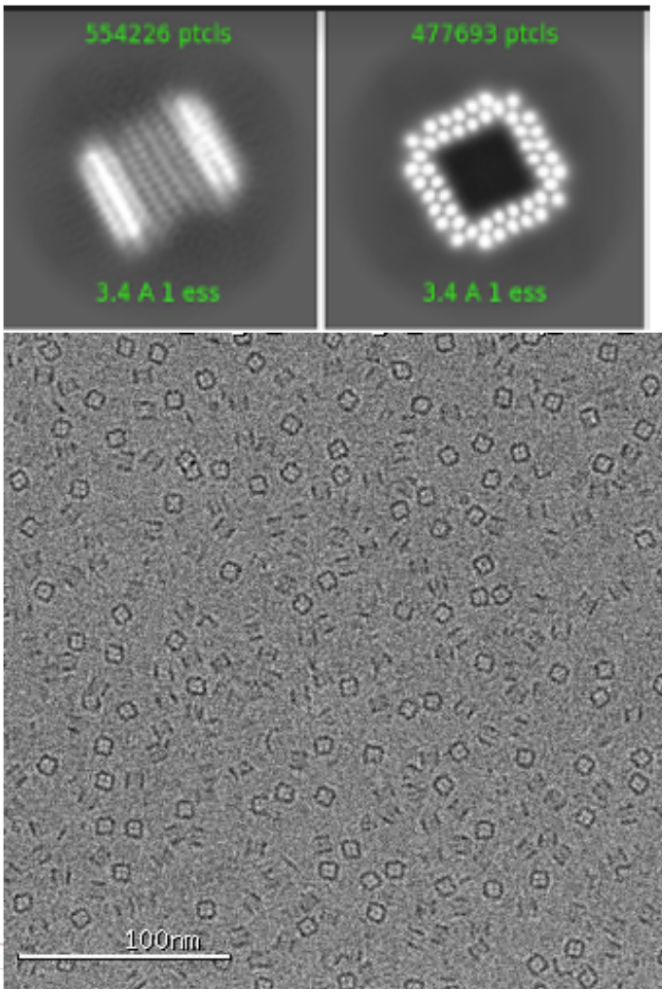
First generation of cyclic oligomers from idealized blocks

We first aimed to do some simple single-component cyclic symmetry homo-oligomers with combinations of new DHRs and heterodimers stitched together via WORMS protocol [19]. In addition to the heterodimers discussed above, some new ones were also added with “2 helix on 2 helix” interface. Because the heterodimers had not been fully structurally characterized, this was an important step before increasing in complexity. When these were first designed using the old Rosetta methods, 6 of 24 tested designs ended up yielding samples that were performing well enough to reach negative stain EM data collection.

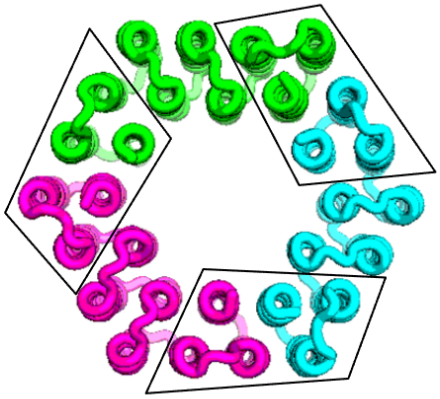
Below are several structures that reached this level:

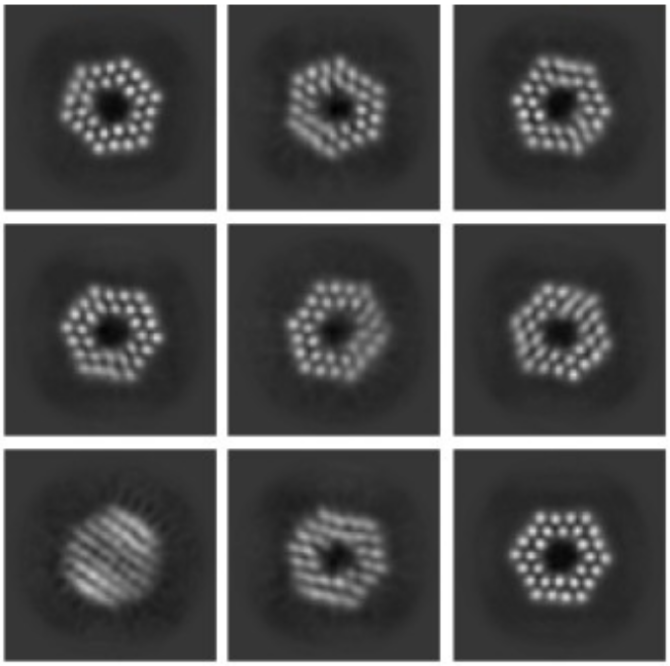
9.5 nm dia.
10.5 mL peak

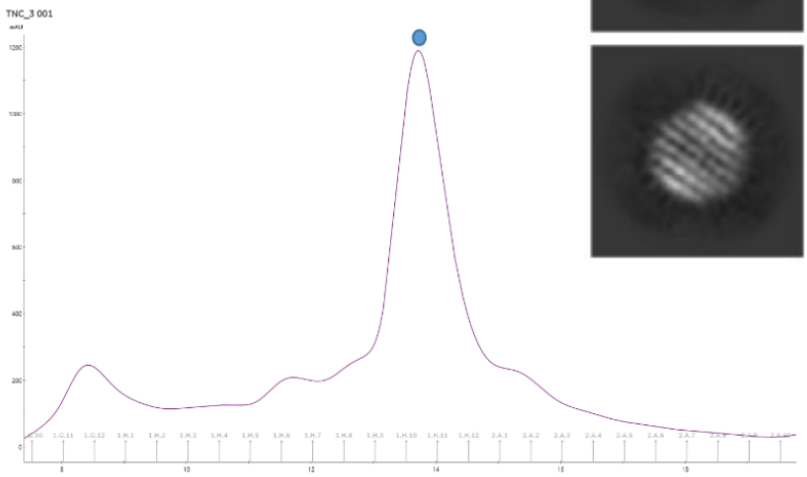




<u>THG4</u>, with S200 SEC profile, a cryo EM micrograph, and 2 class averages from cyo EM. Seeing helices as clearly as this in 2D averages is a very rare sight for the EM field.



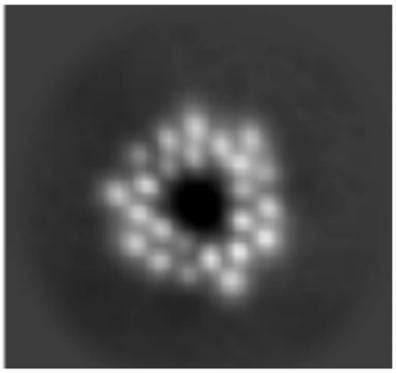


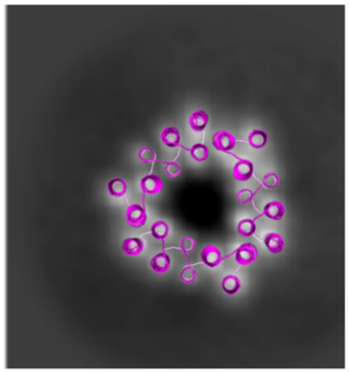


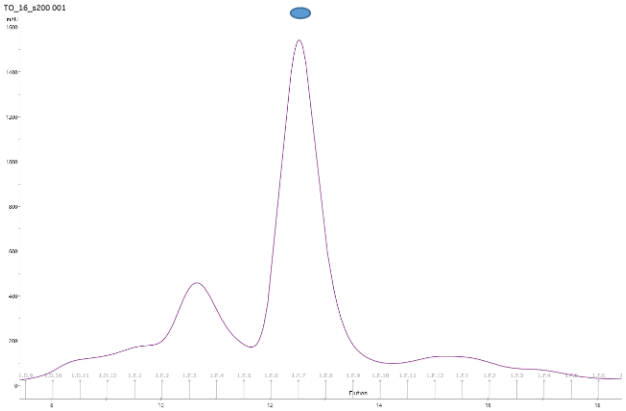
<u>THG6</u>, with S200 SEC profile and cryo EM class averages. Unfortunately only this view was obtained due to particle orientation bias. Protein model has boxes over the heterodimer parts to visualize what is heterodimer and what part is the connecting DHR.

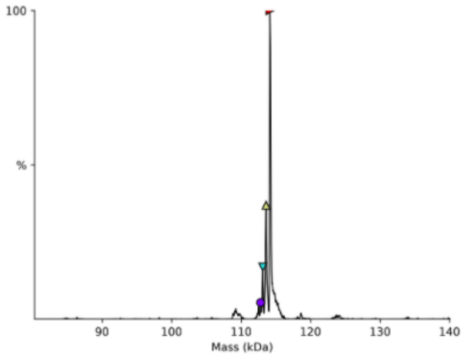
7.5 nm dia.
12.5 mL peak

Actually C3

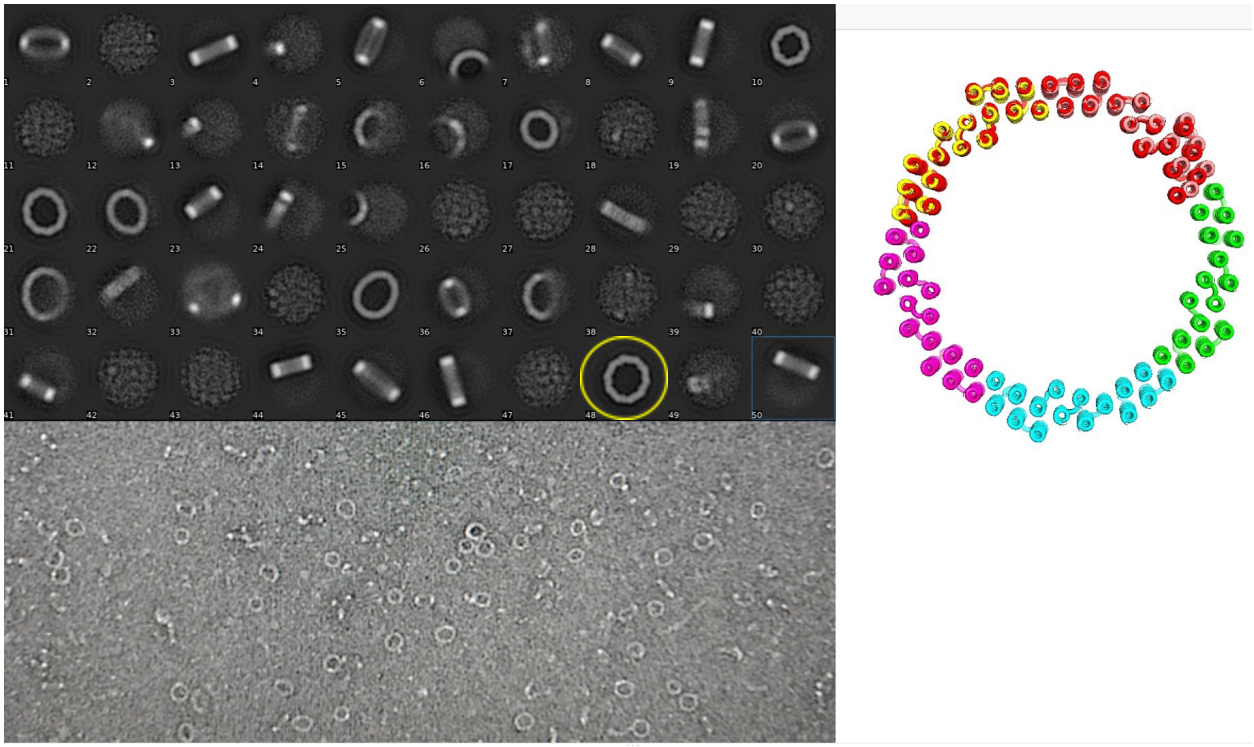








<u>TC4</u>, with S200 SEC profile, and both cryo EM class averages and native mass spec spectra indicating that it actually forms stably as C3. The cryo EM classes indicate the chain still folds to put the helices in the same place while maintaining straightness, but there is hinge-like rotation at interface which suggests this preferred to finish as a C3 during assembly.



TC5_2, shown with negative stain EM micrograph and class averages. In the protein model, 2 red chains are shown to reflect later structure prediction data that reflects how they are expected to interface with each other. Circled class corresponds to the correct C5 structure, there is also a robust tetramer shape at top right of averages.

It seems that the helices stay very well aligned and straight in almost all cases. However, helices are appearing to be able to hinge about each other to allow oligomeric states that are not the ones as designed. Besides trying to achieve the designed state as the most thermodynamically favorable, we are also fighting against kinetics. Often a smaller oligomeric state is accessible first while the chains are assembling and it is possible for the ring to close stably if the closure bonus of forming another interface is sufficient to overcome any strain that may be associated with making a smaller cyclic species (such as in TC5_2 where both C4 and C5 states were accessed).

Overall, this is encouraging that we can achieve a very “industrial” look of designed proteins- especially when looking at the micrograph of squares of THG4, it is striking how different this looks from most natural proteins. Several DHRs and heterodimers without structural data were tested in these cyclic configurations and were able to achieve reliable EM data. This is powerful because each time this happens we gain new validation of a cyclic oligomer, a DHR, and a heterodimer which can all be used for different purposes.

Investigating the modularity of new ideal designs

It is possible for both cyclic oligomers (as shown already) and nanocages to incorporate straight DHRs. Furthermore, if a straight DHR is used in the correct orientation, it will allow size change of the assembly by simple manipulations of how many times the DHR sequence is repeated. Here we have proof of this concept demonstrated by protein THG4 with cryo-EM of the base size as well as a version where 6 helices of repeat unit sequence has been added to the protein. We see that the shape has been maintained and the structure remains with high feature similarity (**Figure 15**).

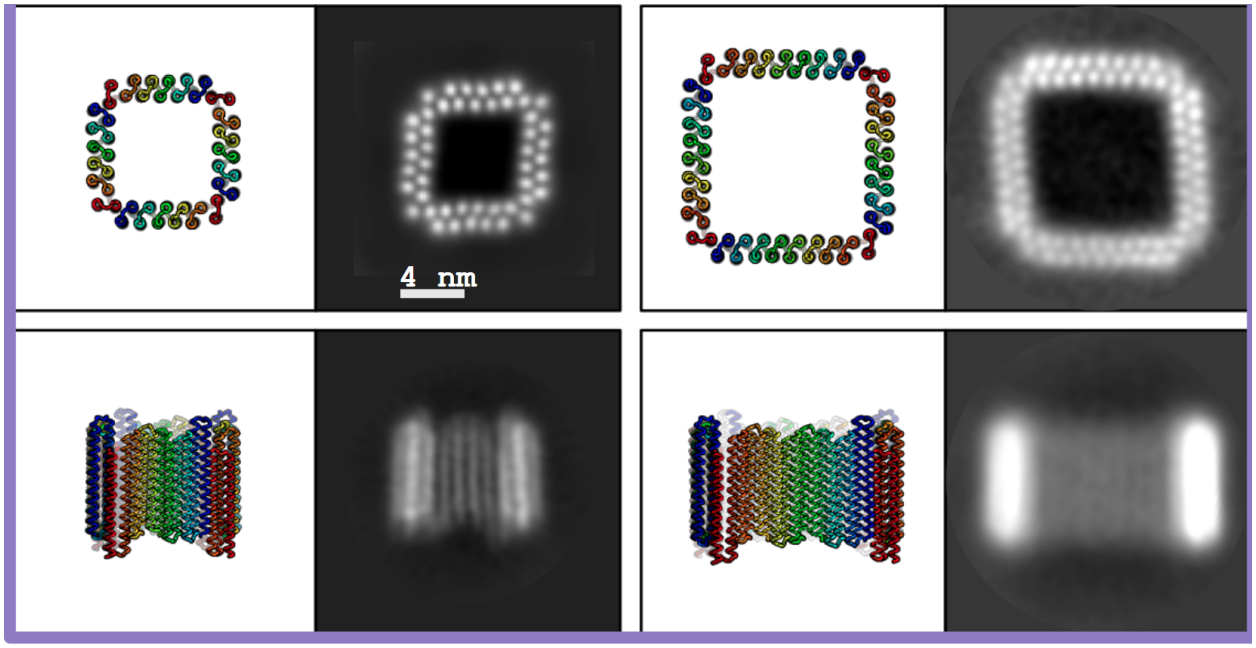


FIGURE 15. Extensible C4 "square" protein
Here on the left are top and side views of a base size of the C4 along with cryo EM class averages. On the right is the same C4 structure but with 2 repeat units added, which in this case is 6 helices.

Another interesting feature about this protein is that it uses long helices generated with the "alpha_helix_100" Crick parameters. This means the helices all have the ability to repeat their phase after 18 residues (as described in the Technologies and Methods section). So, it should be possible to cut out 18 residues from all the middle helices of the chain and maintain all the loops and packing interactions, just with less helix-helix packing. When we tried this with THG4, it elutes 0.5-1.0 mL later in SEC than the parent THG4 protein which indicates it is indeed slightly smaller but not by much, and it has cryo-EM data that supports that it can form the same square structure (**Figure 16**). However, it can be noted that there are much fewer particles in the EM class averages because this structure falls apart a bit in the EM sample environment. This could suggest either that using shorter helices has made the protein weaker overall, or that the shorter model is just more strained in the square conformation, possibly not having the same lowest-energy conformation as the parent.

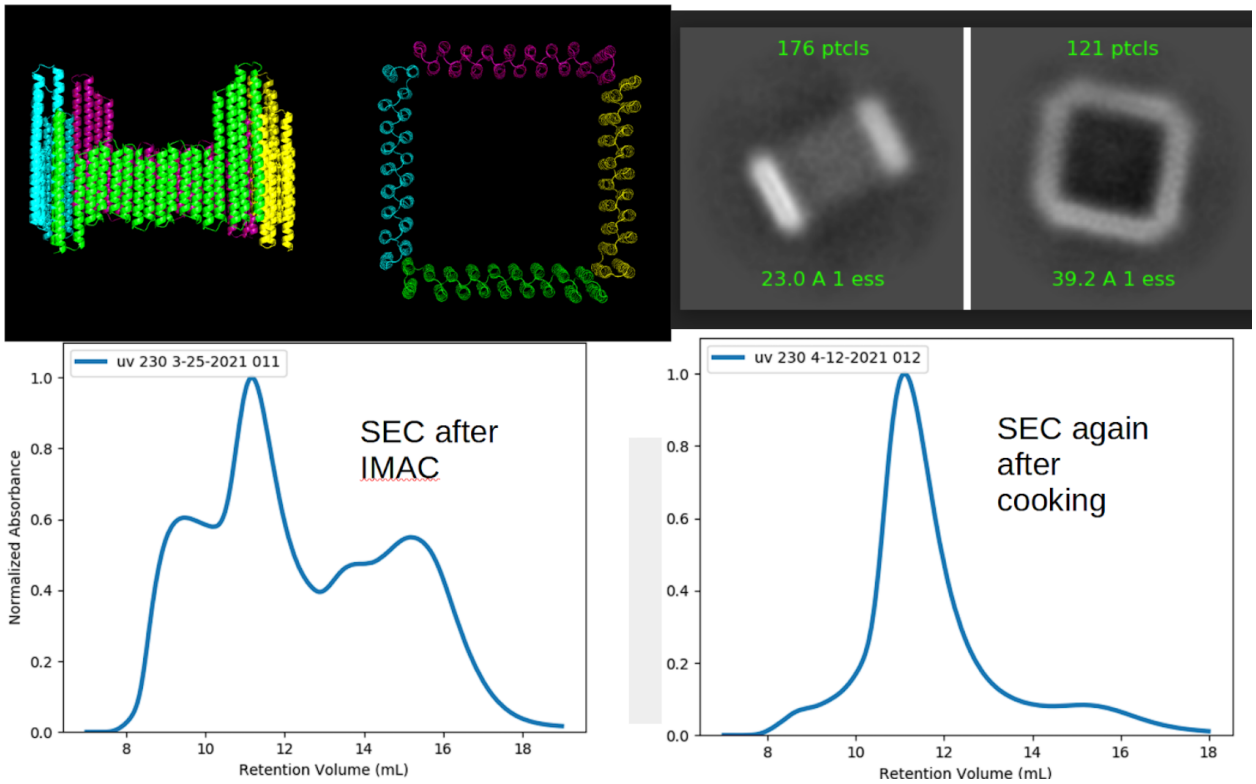


FIGURE 16. A variant of the extended C4 square, with shorter DHR helices
Here we have a model of the new C4 structure showing the truncated helices. The cryo EM class averages are from sparse particles because of there not being many intact squares on the grid. The left class probably includes both "upside-down" and "rightside-up" particles which would explain why the smaller "row" of helices in the middle is not visualized as a distinct feature. S200 SEC profiles show a clear major peak of appropriate size that remains and is cleaned up when the sample is thermally annealed.

There was another set of designs made that are cyclic oligomers, but are not rings. These are radially-extending oligomers featuring straight DHR arms. While the straight DHRs do not confer any new features to this type of design when it is looked at as a simple cyclic oligomer, these can give the extensibility feature to larger structures such as nanocages if they are incorporated into them correctly (**Figure 17**).

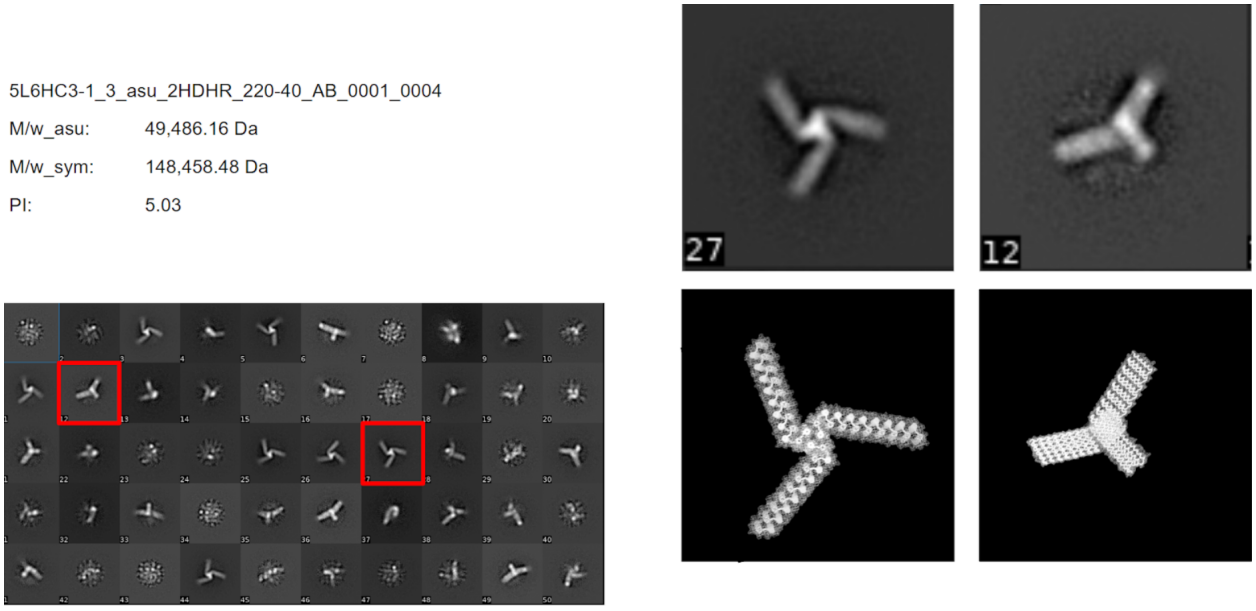


FIGURE 17. Straight helix C3 armed with straight DHRs
Here we have negative stain EM class averages alongside images of the design model. In this case, straight DHR was fused onto a straight-helix helical bundle designed by Scott Boyken [3].

We obtained several of these structures with convincing negative-stain EM averages. These were made with the idea in mind that nanocages could be built with them and they would have the extensibility property (grow cage size as a function of simple repeat number used in DHRs) if the extension vector of the DHR were aligned parallel to the plane between symmetry axes of nanocages (**Figure 18**).

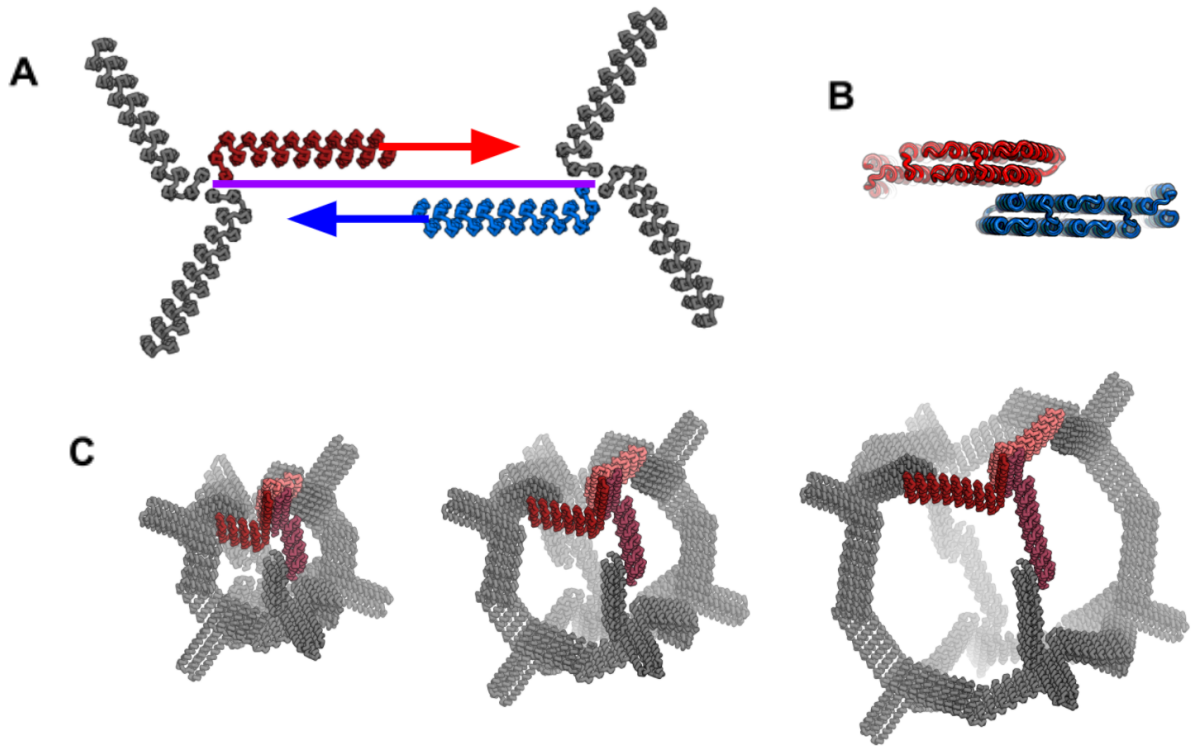
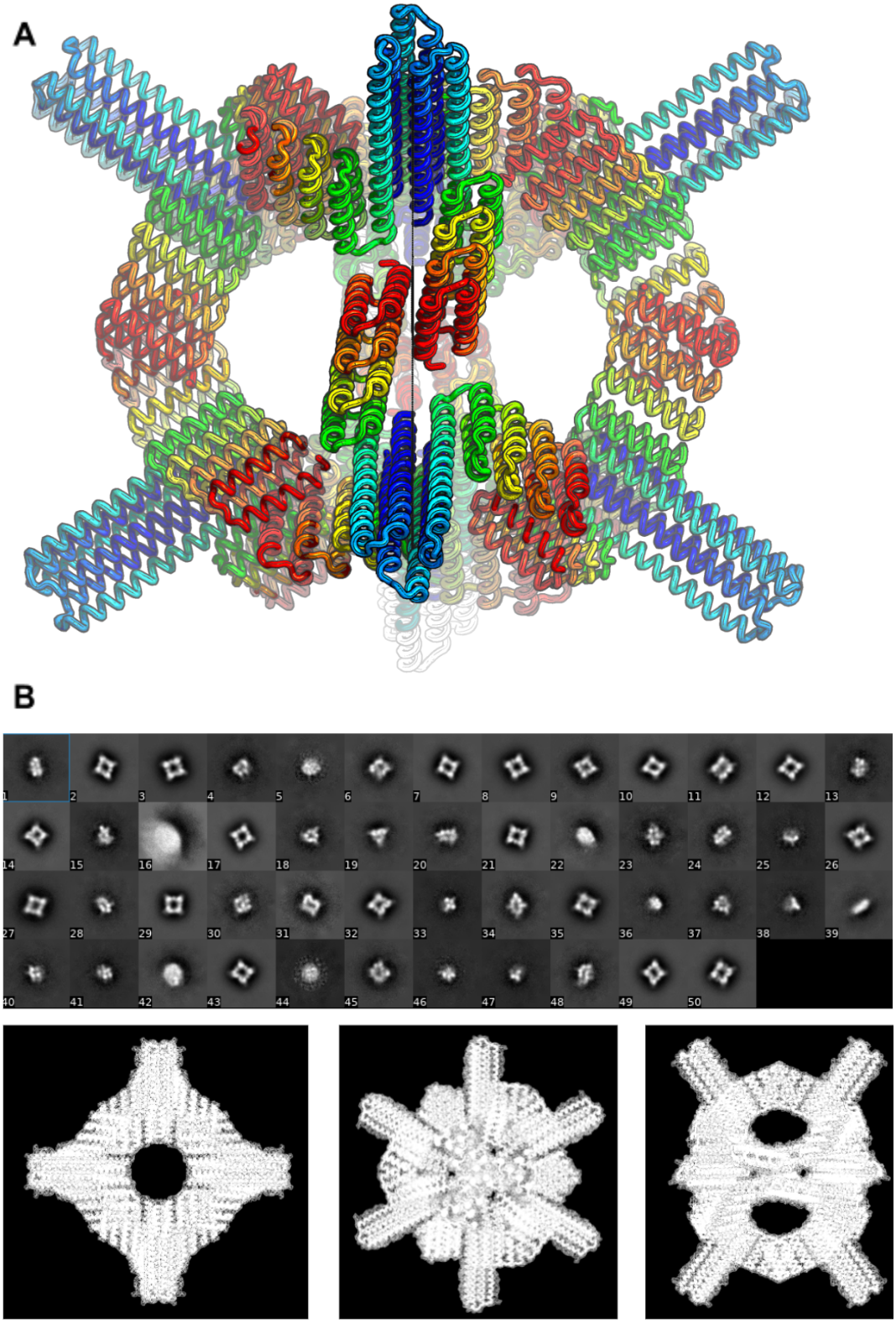


FIGURE 18. Cage extensibility paradigm
A – *In a nanocage with multiple cyclic components, a plane containing both of their symmetry axes (purple) represents the geometry that DHR extension vectors (Red or Blue) must be parallel to in order for DHR extension to grow the cage without breaking any geometry*
B – *It is OK for extending DHRs to meet in this configuration (here they are angled "up" toward the viewer) and any direction in this respect so long as they satisfy (A)*
C – *Example of extending cage, with an armed trimer in red. As the red trimer's arms extend, the overall cage shape and angles remain the same, and no interfaces are changed.*

For our first attempt at extensible nanocages, we did not constrain sampling to satisfy this extension vector alignment. Instead we used RPX dock to find interactions between these armed trimers that could support cage assembly. Even with only armed trimers as building blocks, you can design octahedral, tetrahedral, and icosahedral nanocages. From the top quality docks, we picked ones for which the extension vector was pointing less than 10 degrees off from ideal **(Figure 18 A)**.

The first cage to look at is an octahedral single-component cage made from one of our validated trimers with straight DHR arms. In this one, the alignment of DHR extension is not great with comparison to the direction between symmetry axes, about 8 degrees off (**Figure 19 A**). It was tested at base size “+0 repeats”, as well as with +2 and +4 repeat units. In this case a repeat unit is 2 helices of structure/sequence on the straight DHR arm. We tested this because we were curious about how much flexibility the arms have for allowing the cage to contort into place for assembly in the extended modes if they do not match up perfectly. The base size works quite nicely, and 2 larger sizes were tried. SEC data of the extensions seemed to indicate that a larger, full-cage species was present (**Figure 20**) and that it grew appropriately with the number of repeats growing. While negative stain and cryo-EM were good for the base size, we were unable to obtain good images of either of the 2 larger sizes despite multiple efforts. It is possible that our SEC data was misleading, or possibly the cages are too strained at larger size and the environment in EM makes them fall apart to the point that they cannot be reliably imaged as intact assemblies.


A
B

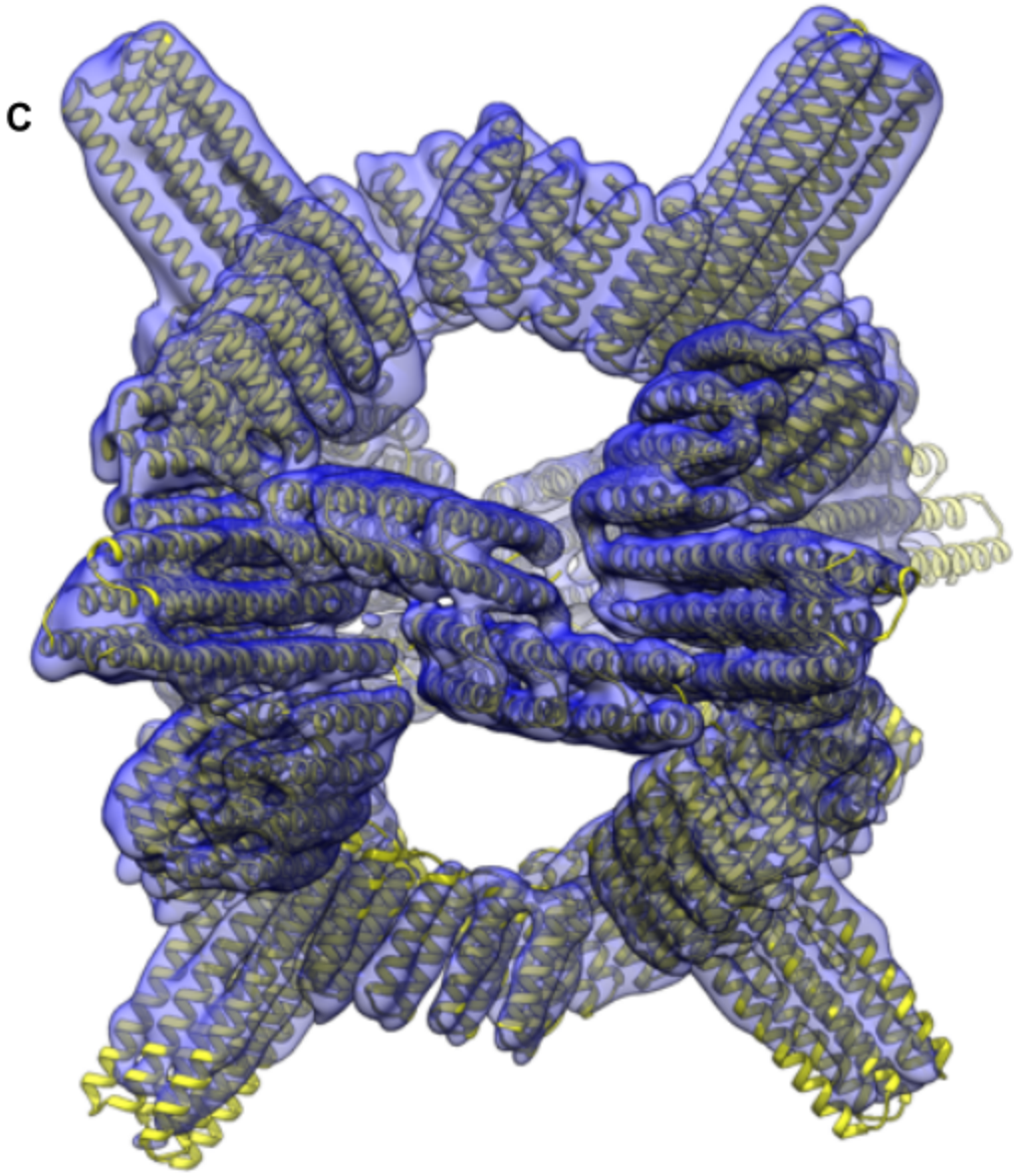


FIGURE 19. Base size "+0" octahedral nanocage
A – *Design model with line to show ideal extension direction*
B – *Negative stain EM classes compared to design model*
C – *A solved C1 model of the nanocage from cryo EM (density in blue) overlaid with design model in yellow*

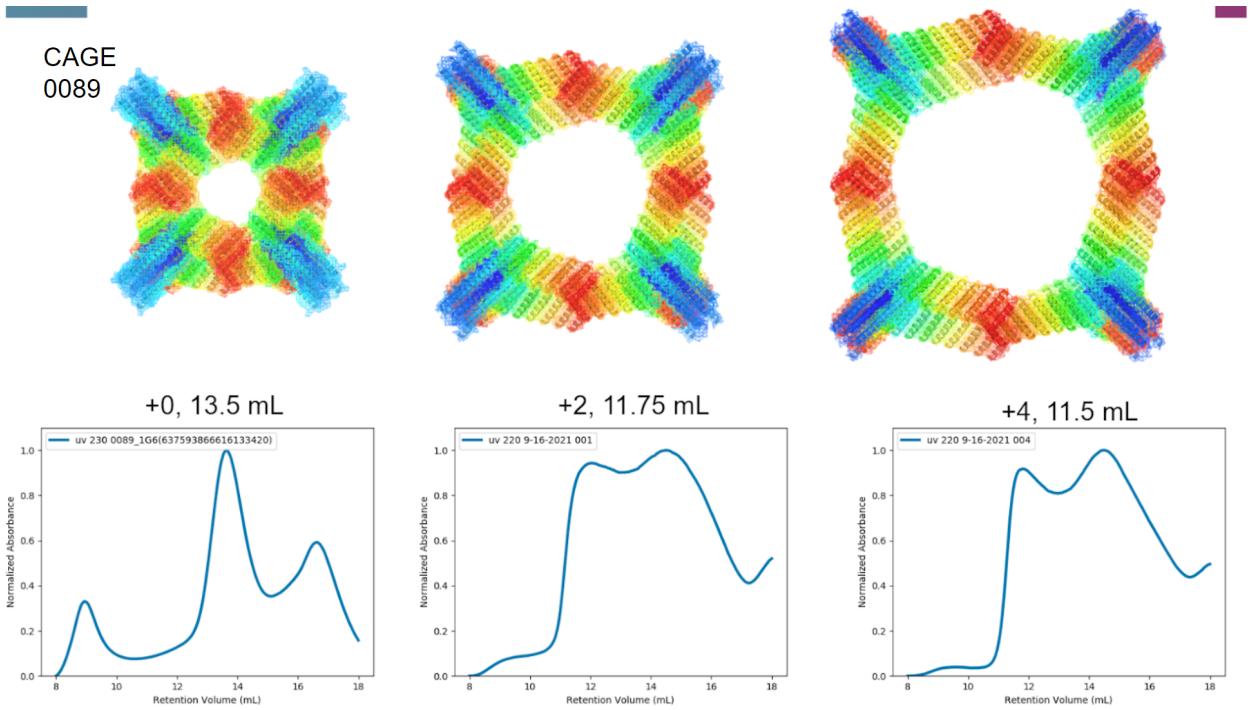


FIGURE 20. Octahedral nanocage extension SEC data
Here from left to right are the +0, +2, and +4 sizes of octahedral nanocage with design model on top and S6 SEC profile on bottom.

Another nanocage that we collected a lot of data on was a tetrahedral cage also made with just a single component straight arm trimer. This one had +0, +1, and +3 sizes tested, with an extension vector alignment that was slightly better than that of the octahedral cage, about 5.5 degrees now (**Figure 21**). The SEC of these samples was less convincing than with the previous cage at first glance because of the +0 size having a big early elution peak. However if the later peak from +0 sample SEC elution is looked at in EM, that one also contains nanocage and the +1 and +3 cages have sensible shifts of that peak in the SEC profiles (**Figure 22**)

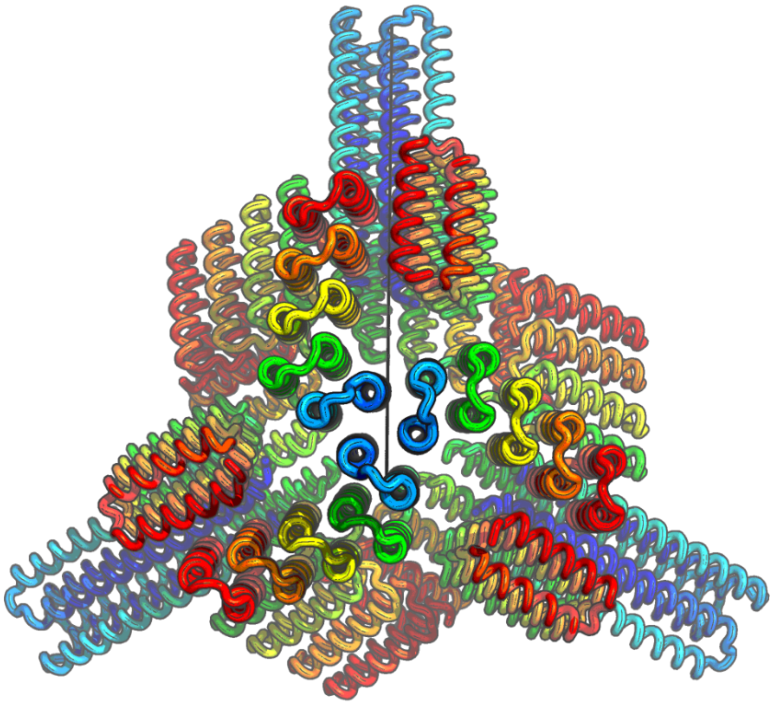


FIGURE 21. Tetrahedral nanocage model
This model features 4 straight-armed trimers and the ideal extension vector shown with faint black vertical line

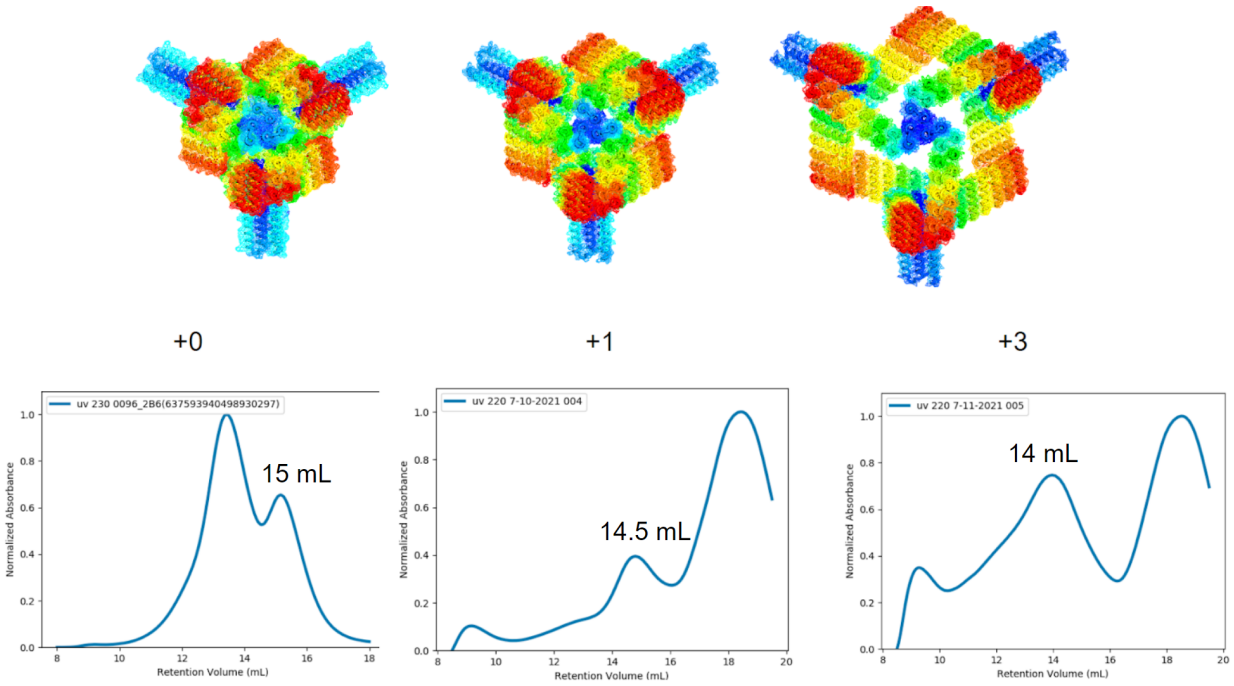


FIGURE 22. Tetrahedral nanocage extension SEC data
Here from left to right are the +0, +1, and +3 sizes of tetrahedral nanocage with design model on top and S6 SEC profile on bottom. Sample was taken from the peaks labeled with numbers for downstream work.

Our in-house EM was down when we analyzed these samples on SEC, so we sent protein from the presumed appropriate peaks directly to our collaborators for cryo-EM. From preliminary data in 2D classes, we observed views that indicated the appropriate size change was happening. Compared to design models, these looked very promising except for that the classes were missing density from the 4th trimer of the tetrahedral cage (**Figure 23**). This could be due to either the cages not forming fully, due to sample degradation in the EM environment (sometimes an issue of air-water interface in the sample), or due to flexibility/strain being concentrated in one area of the cage, making it harder to obtain density averaged there. We ended up not getting enough data on the +3 cage to make a 3D model of the full intact cage. We did, however, obtain full reconstructions of the +0 and +1 cage sizes from the cryo-EM. (**Figure 24**). In both of these cases, there is evidence of a trimer also being missing but only sometimes. It is hard to say if there is actually a trimer missing as often as indicated by class averages, or if 1

trimer of the structure tends to be more flexible and inconsistent due to possibly the strain for closure ending up in one area of the cage rather than evenly dispersed.

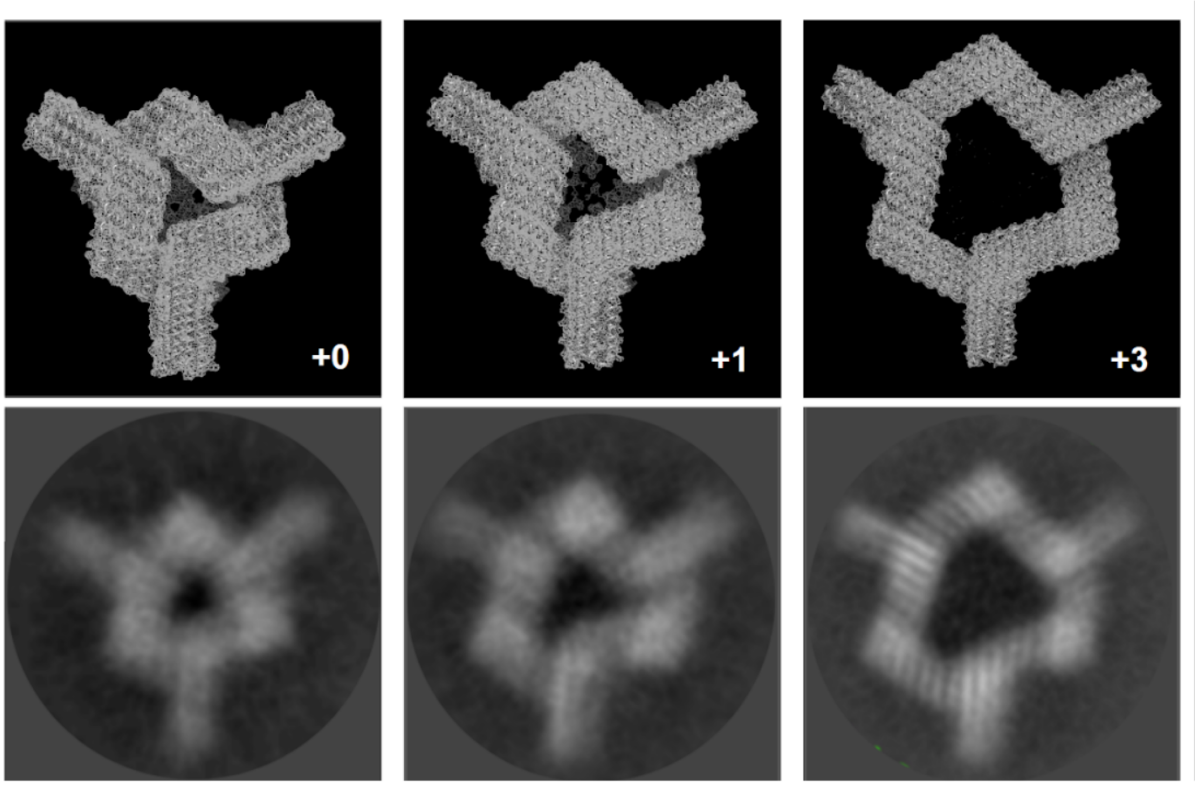
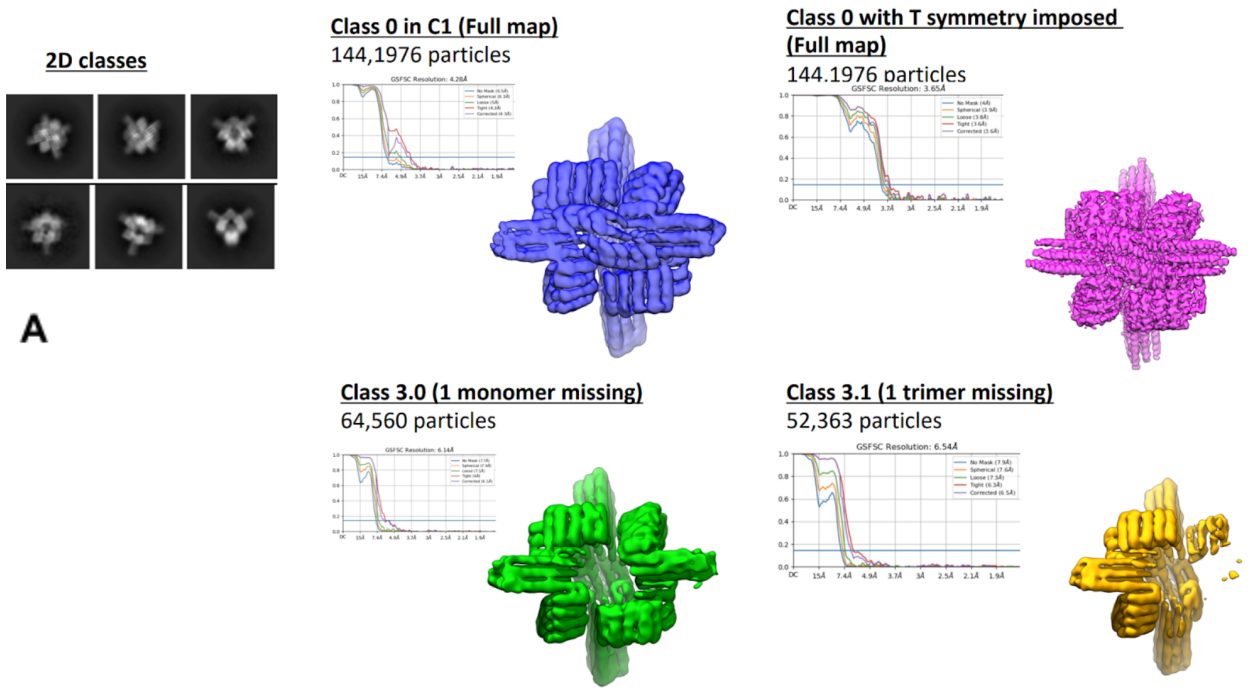


FIGURE 23. Selected cryo EM class averages from minimal data set collection for tetrahedral cage extension
Here from left to right are the +0, +1, and +3 sizes of tetrahedral nanocage with design model on top and matching cryo EM class average on bottom. Note that cryo EM classes show density for only 3 trimers (there should be more density in the middle). The design model views have had their own density of the 4th trimer faded to better visualize the match of the 3 trimer region to data.



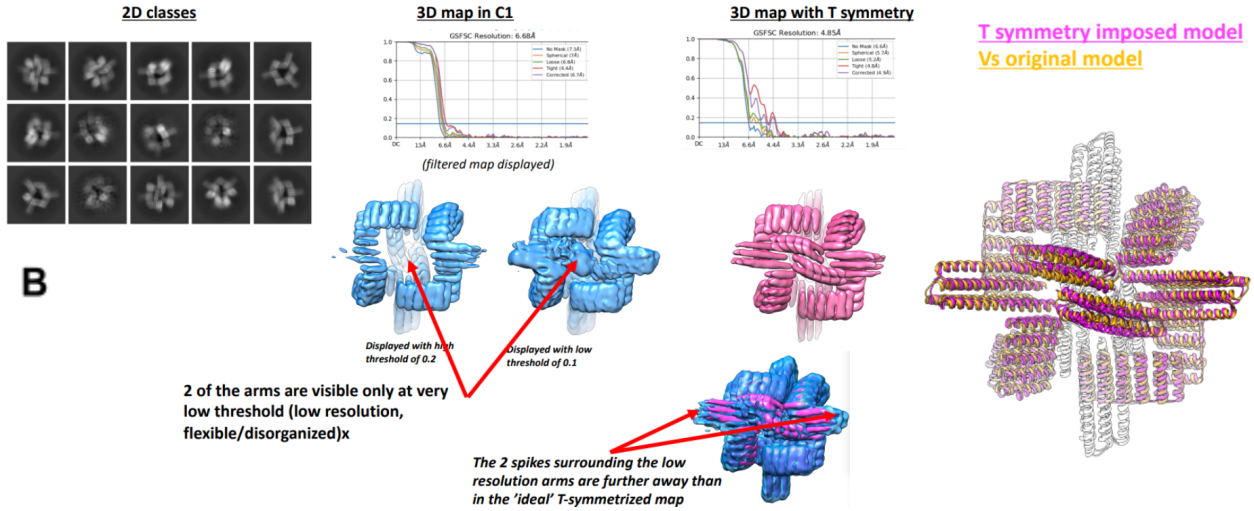


FIGURE 24. Cryo-EM structures of +0 and +1 tetrahedral nanocages
A – *Data for the +0 cage. Reconstructions of the cage are shown with (T symmetry imposed) and without (C1) symmetry. There are reconstructions for varieties that are missing density of either 1 monomer or 1 full trimer also*
B – *Data for the +1 cage. A T symmetric reconstruction is shown aligned to the design model. Evidence of the cage starting to splay apart is detected in C1 reconstruction.*

While the data on the +0 and +1 cage extension can be considered a moderate success for the extensibility concept applied to structures larger than cyclic rings, we also wanted to take advantage of the potential strain that we found in our cages from the EM data. The cage with missing trimer seemed appropriate to model as a C3 structure (it has 3 trimers making 3 interfaces), so we tried to solve a C3 structure of the +1 cage cryo data and found a nice density threshold that showed consistent C3 structure with arms pointing toward an unoccupied space (**Figure 25)**. The idea is that we could use this new position of the arms to build a new C3 structure to put a “hat” on the incomplete cage (**Figure 26**). The rationale behind doing this is that we have interesting C3 structures in the lab with C3 symmetric backbones that fit inside the desired area of the cage, and these backbones have also been tested with heteromerized sequences. So, it is possible to model an armed heterotrimer fitting as the last piece of the cage, which can be designed to not have strain as the original design presumably did. This gives us a large structure with a unique shape that is useful for EM structure determination and potential cell tomography applications. The complete structure represents a large object with asymmetric shape that is easy to classify into unique orientations when viewed in EM, so it is good for a structural chaperone for looking at other structures attached to it. Additionally there can be a high copy number of GFP or other reporter molecules on the cage that can give a big localization signal without having an avidity confoundance when a functional/target protein is fused to one of the heterotrimer hat chains (1 studied fusion protein of interest with 9 highly localized GFP molecules per fully assembled cage is possible, for example). This work is ongoing, as is more work on improved extensible cages.

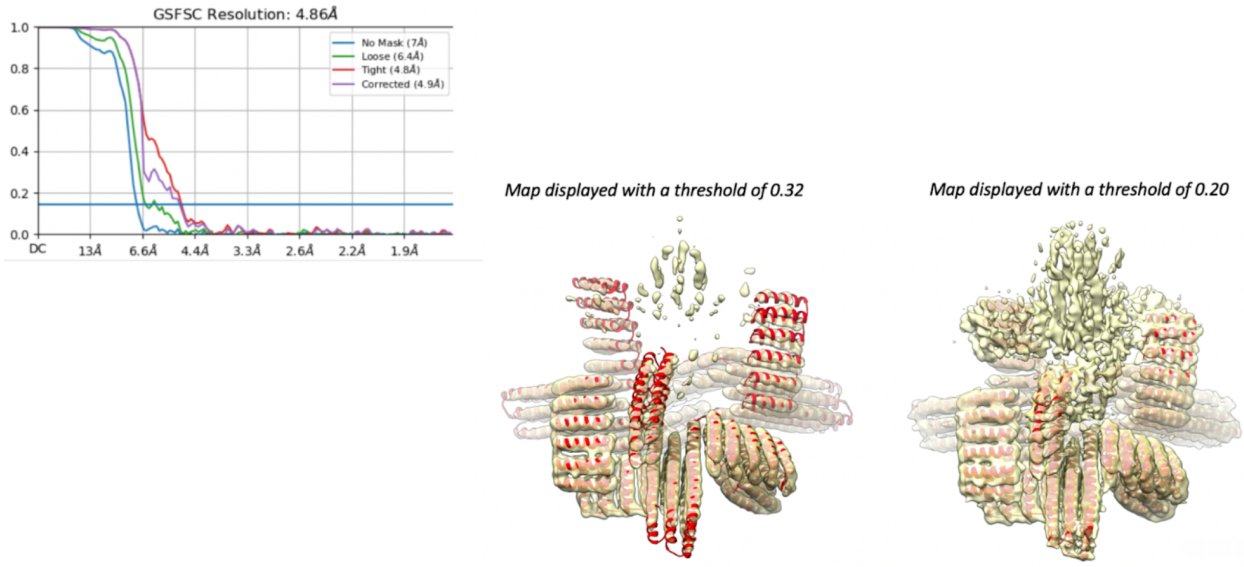


FIGURE 25. Reconstruction of the +1 tetrahedral nanocage with C3 symmetry imposed
The aim here was to get a good structure of the bottom ring that was faithful to the data so that the revised incomplete cage model could be represented with some symmetry but without having to strain to close.

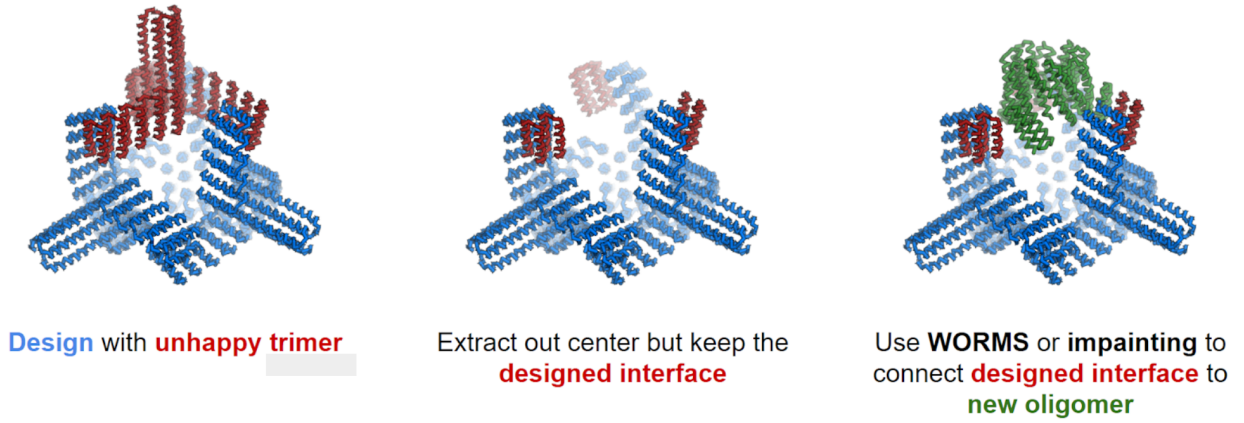


FIGURE 26. Plan to install "hat" on incomplete cage model
This utilizes newly informed locations of arms sticking up. The interface on the arms is fine as shown in the good part of the cage structure, but we want to hold them at the right orientation so as to not impose strain for closure.

Adding geometric relationships to building blocks to plan for larger structures

Up until this point, we have used the straight DHRs largely just for their feature of being able to be used as linear bricks that can help us change the size of larger protein assemblies. However, we can improve the degree to which they are tailor-made for particular building applications. For

instance, it is possible to use a larger repeat unit of 4 helices rather than the typical 2 helices so that there are more helical phases represented in the repeat unit. If we make a helical phase transform + rotation exactly as needed for an application, then the DHR can have the ability to make a turn of a controlled degree if it is fused with itself by aligning a helix on one copy with the controlled-transform partner to it on another copy (**Figure 27**).

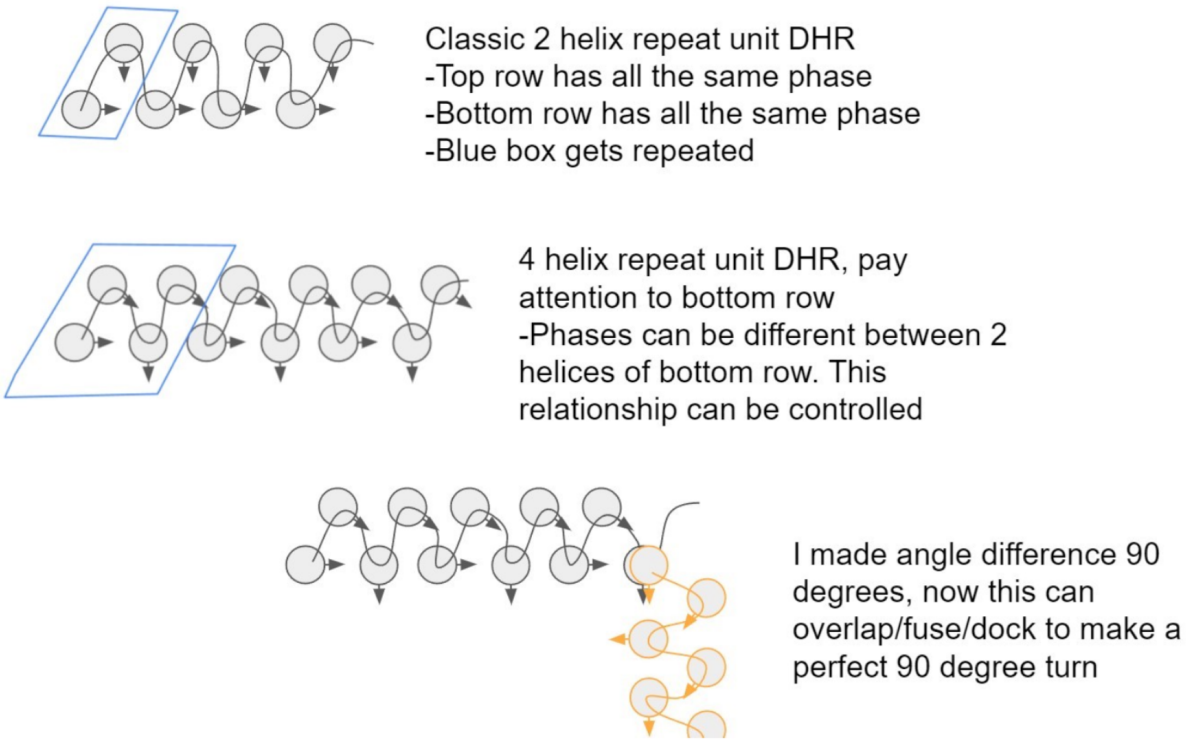


FIGURE 27. Plan for encoded angles in DHRs with 4 helix repeat unit
When there is an appropriate phase difference between 2 adjacent helices, that transform can be utilized in fusion protocols that rely on overlapping secondary structure

This kind of turn, if combined with splitting the DHRs to use their own repeat interface as an interface between chains, allows simplified creation of cyclic oligomers (**Figure 28**). In the past, we relied on combinations of many building blocks and luck to get rings that would satisfy geometric criteria in WORMS, but now we can achieve that with single building blocks that are designed for that purpose. This also suggests that we could methodically perform 2D line-art style protein designs if we follow this procedure.

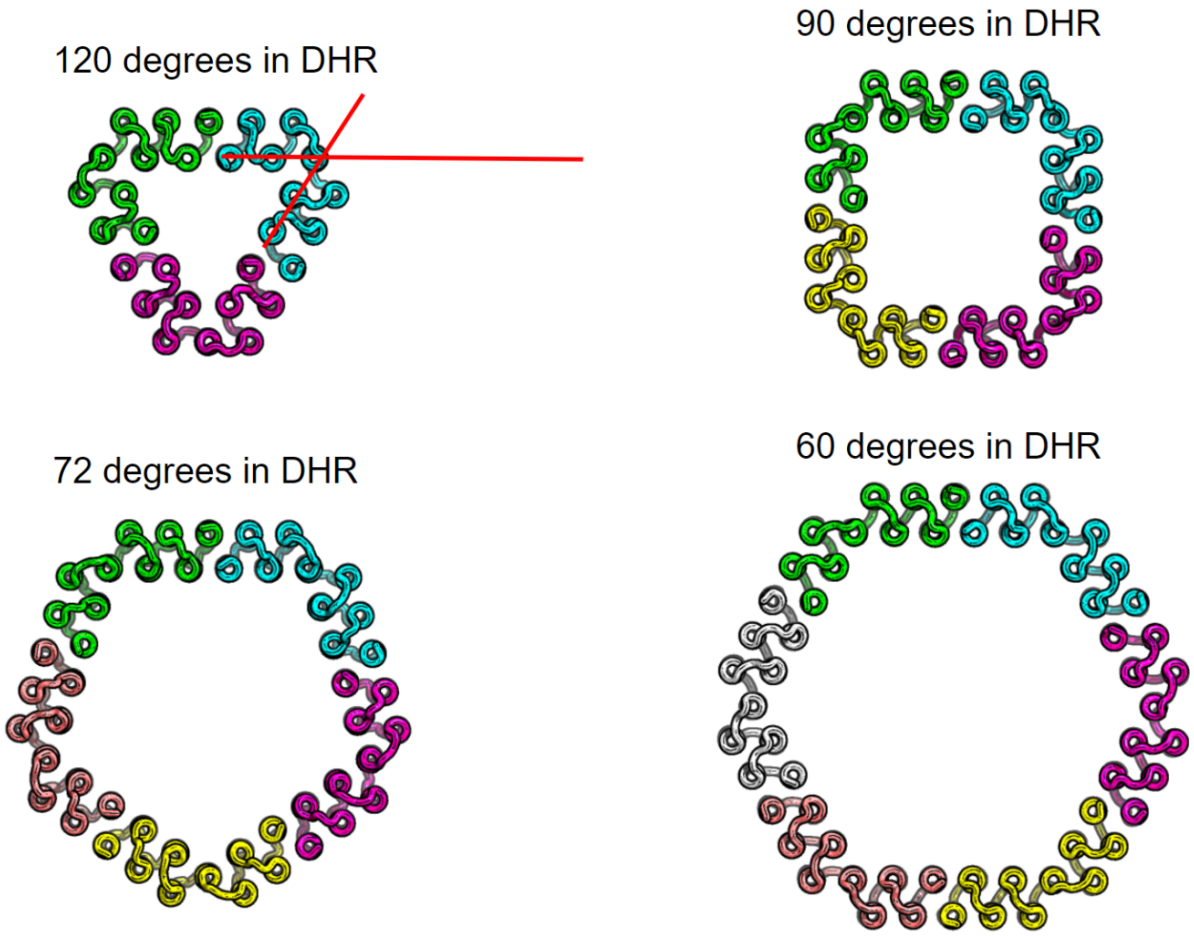


FIGURE 28. Explicitly encoded ring shapes from specialized straight DHRs
Here we see outputs from WORMS protocol where split DHRs with the appropriate encoded angles were fused with themselves, and solutions with "0 RMSD" interfaces were found, indicating it used the encoded angle correctly to find these as expected.

It is worth stating that from this point onward, we switched from Rosetta design and forward-folding to machine-learning based methods of MPNN for sequence design and Alpha Fold 2 structure prediction for design filtering. These design methods plus the design concept outlined here gave us robust expression and purification of this new ring set; even when a small scale expression was purified by cooking the sample we had enough ring sample remaining intact to do SEC and EM for most of them (**Figure 29**).

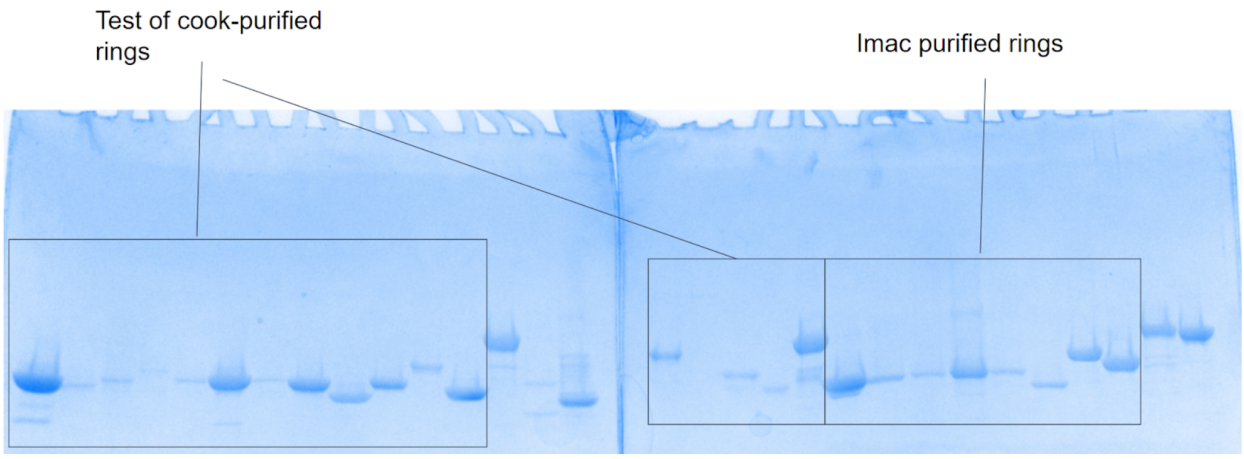


FIGURE 29. SDS-PAGE gel of purified angle-encoded rings
Here we have a variety of samples from the ring order set purified either by cooking the lysate and then collecting the soluble fraction after spinning down, or by doing traditional IMAC purification from the His-tag on the proteins. Both methods worked well, showing successful soluble purification of several ring samples. Even faint bands here represent enough sample to get SEC signal and negative stain EM data.

The C3 rings/triangles worked remarkably well, and 4 out of 5 samples that went to negative stain EM had very pronounced triangle-looking particles (**Figure 30**)

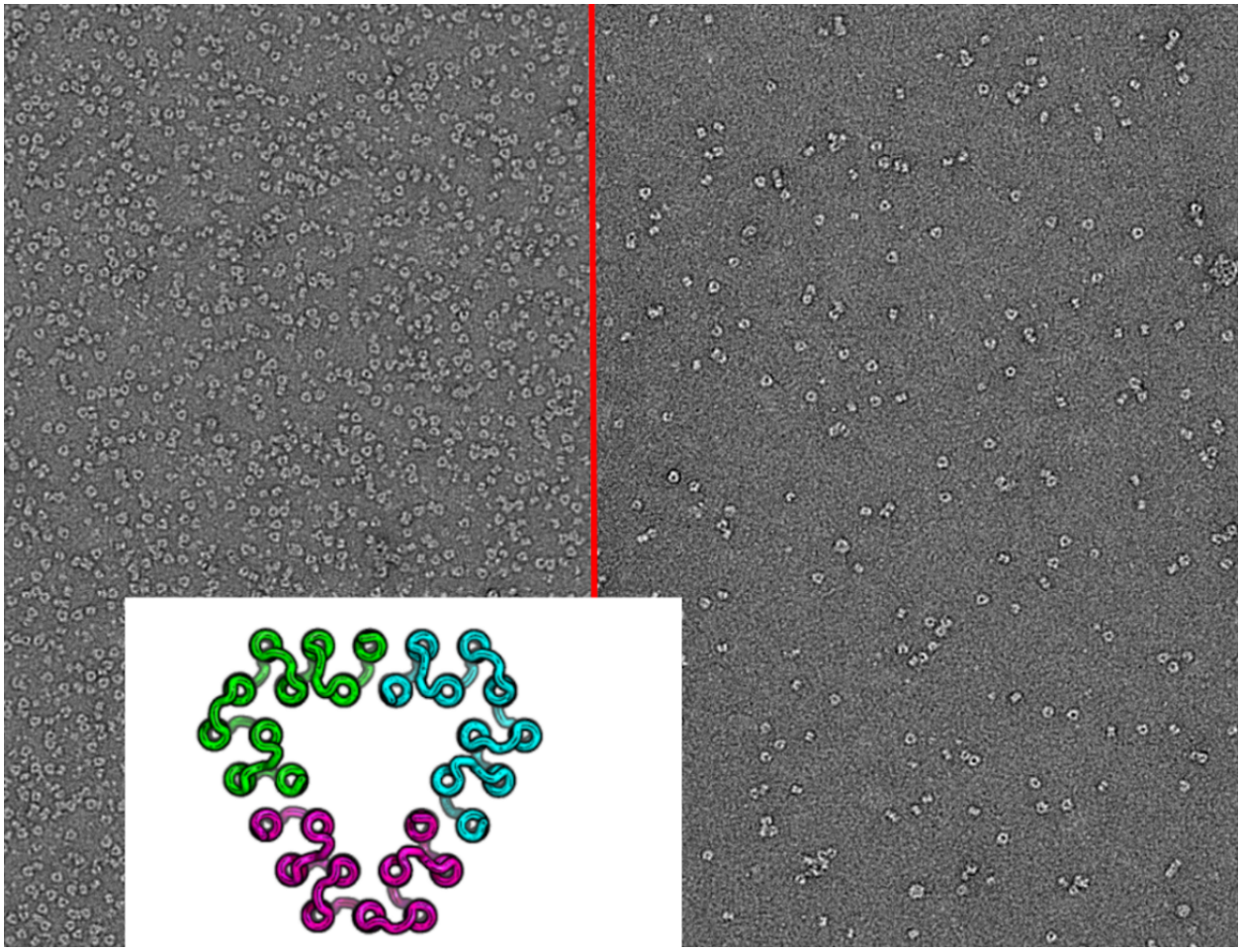


FIGURE 30. Micrographs of 120 degree angle trimers
Here are 2 representative micrographs from 2 different trimer designs, however both have the same number and configuration/position of helices.

So far one of these samples has been averaged from a data set collection with negative stain EM, and the triangle view with appropriate features at the corners as well as side views are visible (**Figure 31**). 3D model reconstruction with this data is in progress.

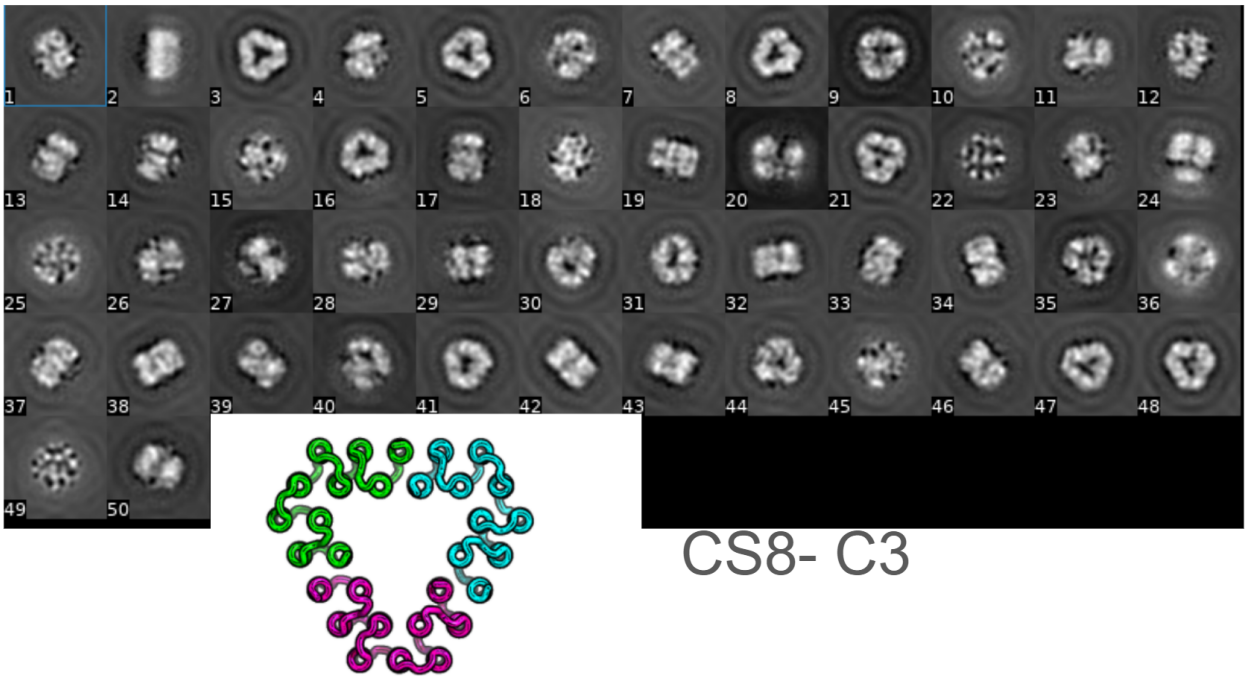
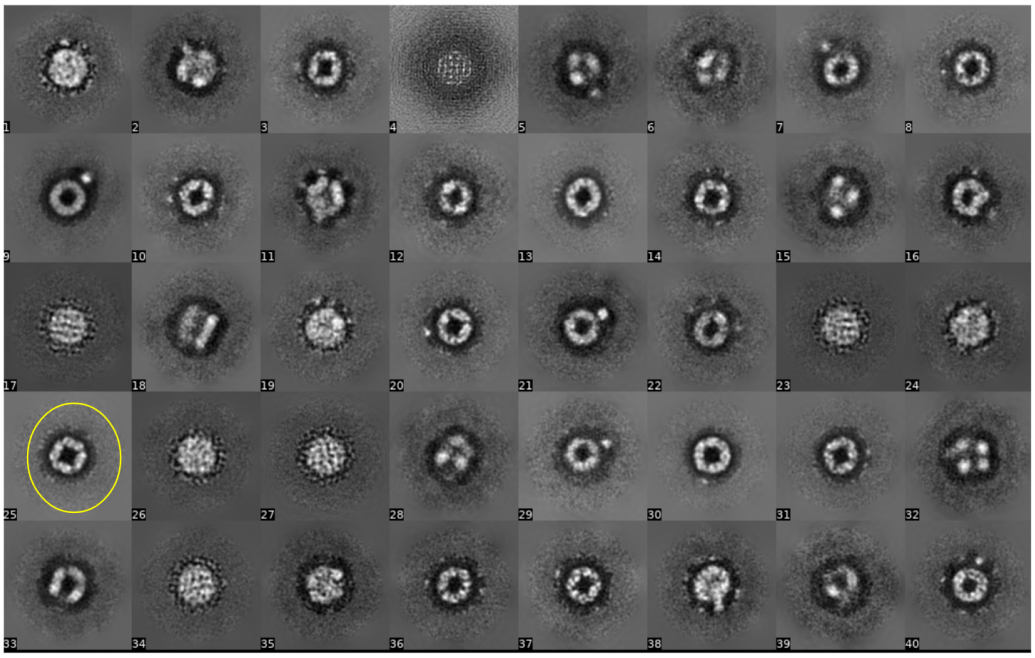


FIGURE 31. 2D class averages from negative stain EM of cook-purified angle-encoded C3 *Here we see several side and top views of the trimer that agree with the shape of the design model.*

The EM quality decreases for C4, C5, and C6 samples. This is likely due to them just requiring more interactions to achieve the desired state, so smaller oligomeric states are possible if the proteins flex enough as we saw before in previous rings. Also with larger rings, the difference in angle between adjacent ring sizes decreases, so there is less strain to access different shapes as the ring order increases. This is sometimes combated with chemical or thermal annealing so that subunits spend ample time finding each other and exchanging until they form un-strained closed structures. Still, we got data sets on a C4 ring and a C5 ring without trying anything experimentally out of the ordinary (these following structures were obtained from IMAC-purified samples without cooking). The C4 ring shows presumably the right species (**Figure 32**) while the C5 ring shows both pentamer and tetramer species and likely more flexibility overall, since we cannot see shape/corner features as cleanly in these class averages despite there being plenty of particles in the data set (**Figure 33**).



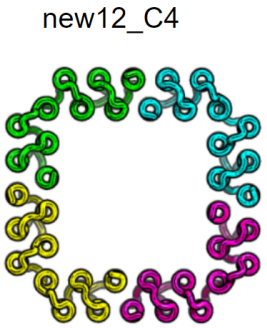


FIGURE 32. 2D class averages from negative stain EM of angle-encoded C4
Here we see several top views of the tetramer that agree with the shape of the design model. Some classes look more circular presumably because tilted views of the structure will reduce the definition of the corners

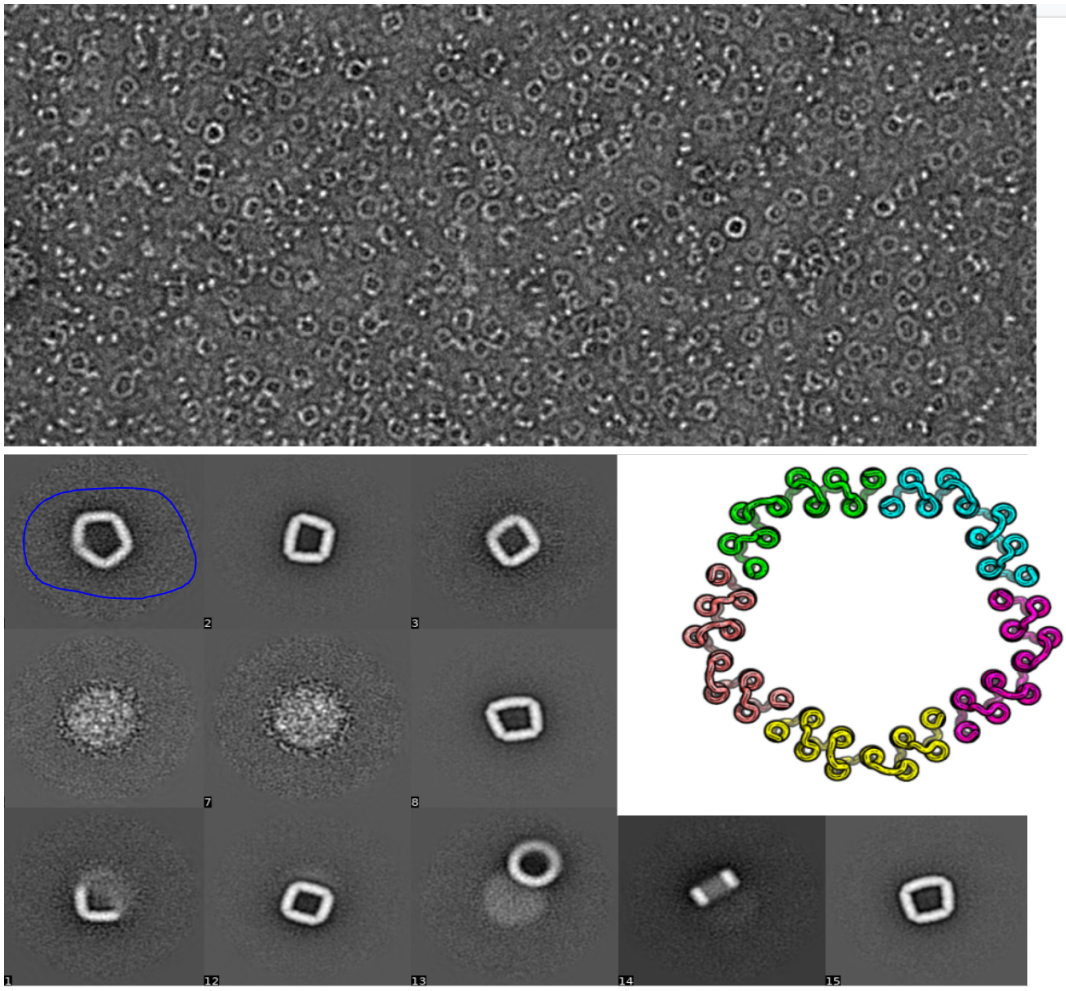


FIGURE 33. 2D class averages from negative stain EM of angle-encoded C5
Here we see side and top views of the structure that offer indication that the C5 ring can form as intended, but there are other species present. It seems possible that we see different ways that the tetramer rings have accommodated strain/buckling

These experiments show that it is reasonable to encode geometric features into building blocks for the purpose of making defined turns and angles. It is also possible to define repeat geometry for smoother curves and near-circular structures by using smaller angle changes that get used at every repeat unit. Here we made new 12-repeat closed “alpha-toroid” structures [24] which turn a defined 30 degrees at each repeat until they complete 360 degrees and finish a circular path to close in on themselves. We used repeat units of 4 helices and experimented with sampling how closely or loosely the helices pack with each other (**Figure 34**). These contrast to

previously designed “alpha-toroids” in that their structure is a planned consequence of building block design rather than a solution that is found from serendipitous structure fragment combinations. These represent the ability to make proteins with inner or outer diameters of whatever size is desired within this approximate 0-25 nm space. Additionally, because we can also define repeat units to do things like rise or fall in the same direction as the central axis, we could make tubes/fibrils of controlled diameter with the same idea just by additionally including a “rise” at each repeat. These potential tubes would appear helical/spiral in shape.

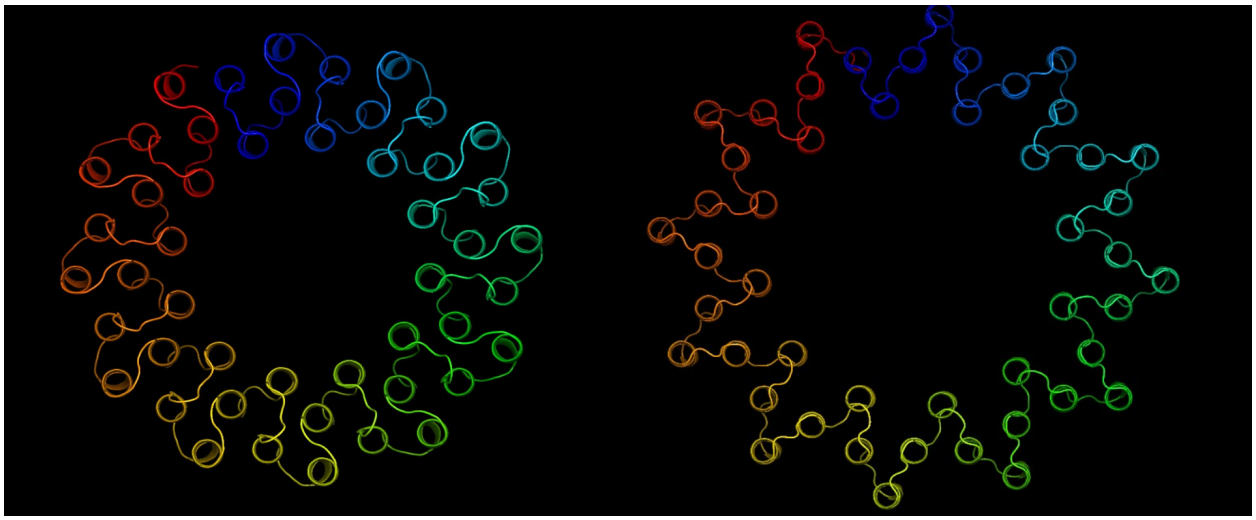


FIGURE 34. Defined alpha-toroids
Here are 2 toroid designs with the same number of helices but the left one is planned to be very dense and compact while the right one is keeping helices separated. These change the potential rigidity of the structure and how buried helices are, which affects how they get sequence designed.

One reason for picking a 12 repeat toroid to use is that it can theoretically be chopped up into a variety of cyclic oligomer states including C1, C2, C3, C4, C6, or C12 depending on if chains are expressed with either 12 repeat units linked together, or any factor of 12 instead (**Figure 35**). Experiments are ongoing to see how well this modularity actually works in practice.

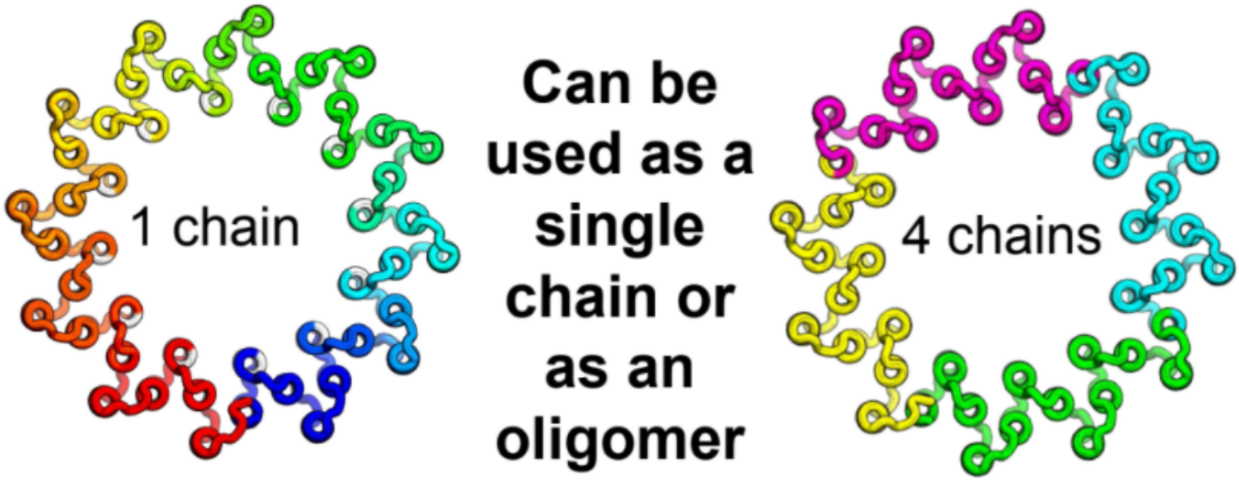


FIGURE 35. Alpha-toroids can be split up into factors of their repeat number
Here we show how a 12 repeat alpha toroid can be split into a C4 protein where each of the 4 chains contains 3 repeat units

We have 1 toroid structure with negative-stain EM averaging data so far on a C4 version (**Figure 36**). They were all ordered this C4 way at first just for having a nice, usual protein size of the subunit. Cryo-EM on this structure is pending, and we have others with different helical spacings that are in the validation pipeline currently. So far the 3D reconstruction from this data is too crude to conclude anything other than that it appears to be approximately the right donut-shape. Whether it is strained to be out of symmetry should be revealed by the cryo-EM data, and that should inform us whether it will be expected to perform well in other oligomeric states or not. For these designs we tested with AlphaFold 2 for 11, 12, and 13 repeat chains to assess if they had preference for closure as 12 repeats, so we hope they will not encounter difficulty in being tested in more complex states such as C6 or C12.

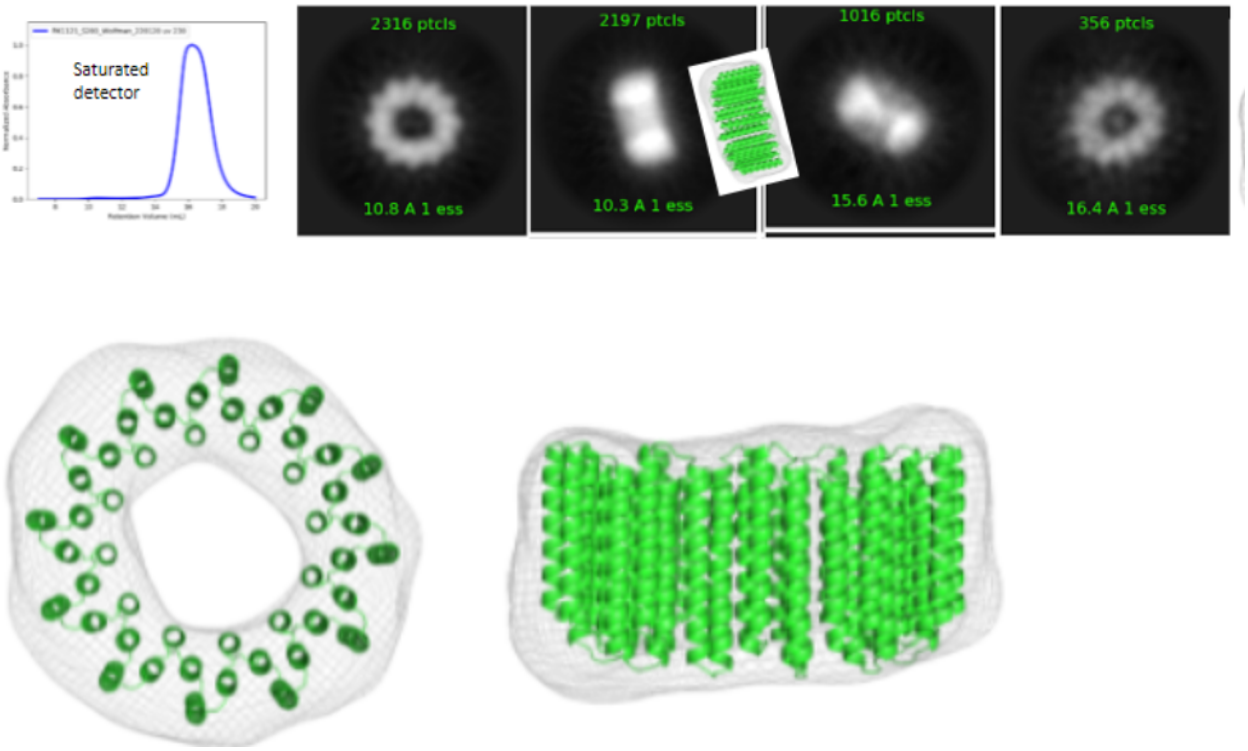


FIGURE 36. Data for a C4 alpha-toroid with 12 total repeats.
This toroid was monodisperse in S6 SEC and had negative stain 2D class averages that were clean enough to show the 12 humps in the "top-down" view. A 3D reconstruction was attempted but lack of alternative views hampered its progress.

The next goal we had for explicit, rational design of nanomaterials was to fix the problem we had with extensible nanocages. In our first designs, we relied on getting lucky solutions that were close to having the DHR extension vector close to where it needed to be for cages to have the simple extension feature. Now, we aim to eliminate the need for hunting for the needle in the haystack. Obviously, the needles we found in the first round were not sharp enough. Based on the premise that 2 DHR arms can make side-against-side interactions (as featured in our existing nanocage structures with straight DHR interactions), we can define a range of distances that we would expect for them to be offset relative to each other for making those interactions. Then, we can make a new set of armed cyclic proteins that project the DHR with an appropriate offset relative to radially projecting vectors from the cyclic protein (**Figure 37**). This means that if it is to interact with another DHR arm with both of their extension directions parallel to each

other, they will have the appropriate offset such that in the middle between their 2 extension vectors will be the axis/plane that connects from one cyclic center to another. These are made by positioning the DHR arms in the correct position in 3D space first, and then building in a new cyclic bundle that holds onto the arms the right way.

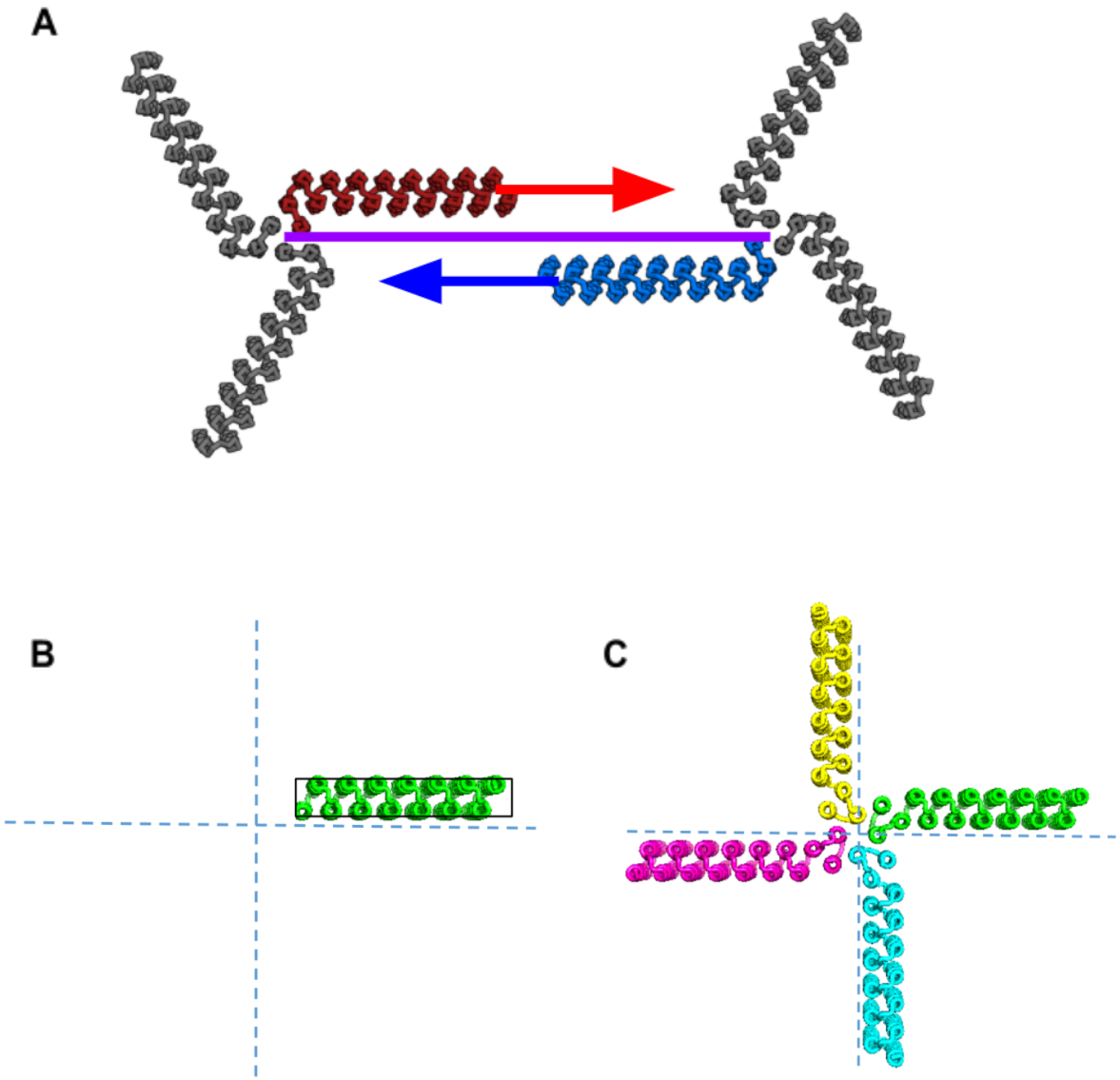


FIGURE 37. Setting up straight DHRs at good position in cyclic oligomer for future cage contacts
A – *Here is the cage extensibility requirement shown again. The DHR extension must be parallel to the purple plane. One nice solution, as shown, is to have DHRs interact side by side with the purple plane between them.*
B – *In planning a potential oligomer, we could pre-position the DHR with enough room at the center to build in new helices for cyclic bundle contacts, while holding the DHR at the appropriate offset from a radial axis so that it has room to make a parallel "handshake" interaction as in (A)*
C – *Sampling helices in the middle and looping them in to form a complete, new cyclic oligomer with straight DHR arms held at a position that is amenable to extensible cage design downstream*

New cyclic designs were computationally filtered by predicting the structure of a truncated oligomer with Alpha Fold 2 such that we would know what combination of new sampled helices in the middle would be predicted to actually begin the straight DHR arm trajectory in the expected direction (**Figure 38**).

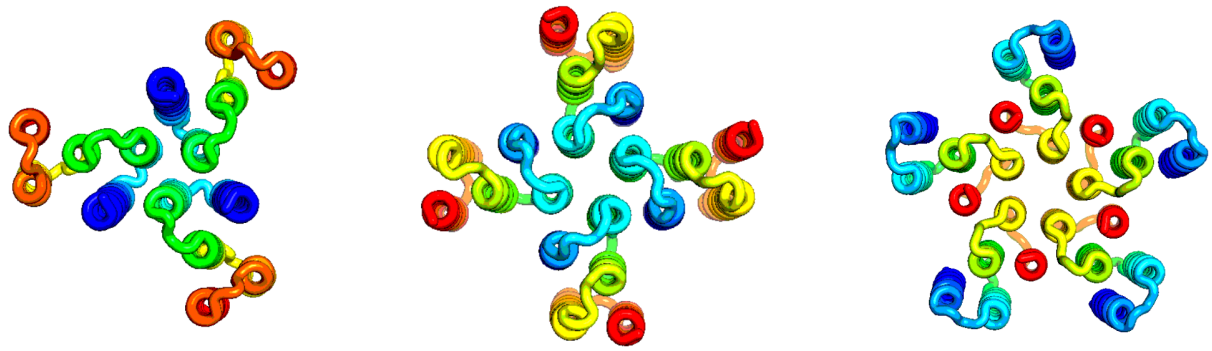


FIGURE 38. AlphaFold 2 predictions of new cyclic oligomer centers
To save compute time, we only folded the new cyclic hub + 1 repeat unit of DHR for all the newly sampled backbones. Here, we see that AlphaFold 2 without any symmetry enforced is able to predict very reasonable-looking cyclic oligomers with helices straight and parallel as desired.

When these oligomers with arms extended out on them are used to dock against each other into cage symmetries with RPX Dock, even in a naive search almost all of the top scoring outputs end up having the DHR arms aligned correctly (less than 3 degrees deviation threshold now, usually close to zero) for ideal cage extensibility **(Figure 39**). Cyclic designs of C3, C4, and C5 were made so that octahedral, tetrahedral, and icosahedral symmetries of T3, T33, O4, O3, O43, I5, I3, I53 were all possible. In all these cases, a new C2 axis interaction is formed in the docking procedure (as was intended by spacing the DHR arms at the right location to facilitate this interaction). All of these end up with distance between DHR arms as planned, but they find various angles out of plane with each other to fit the geometric needs of cage closure while still maintaining their extension vector alignment.

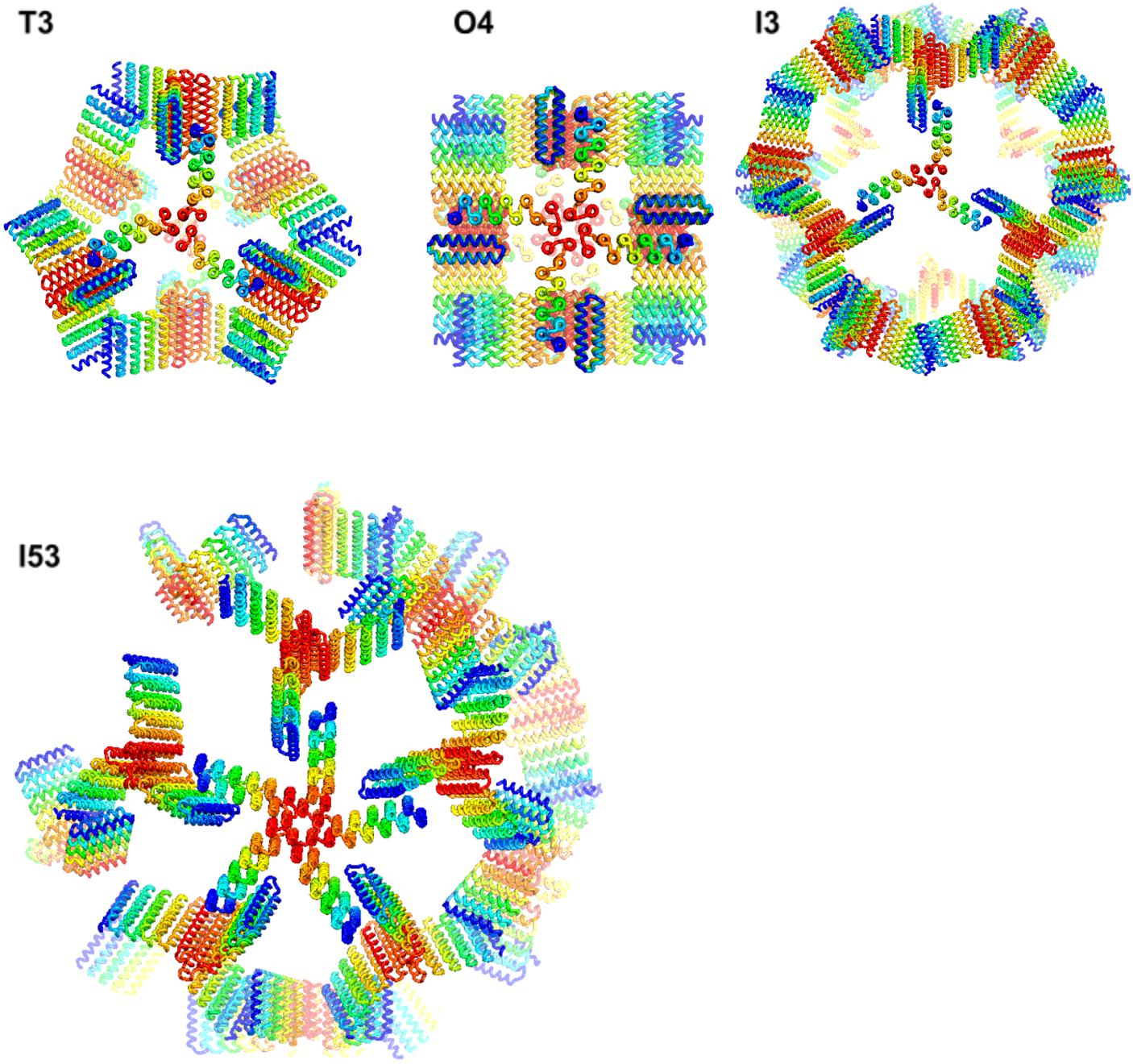


FIGURE 39. Idealized pre-planned nanocage docks
Here a variety of cage architectures are shown, all having DHR extension vectors from the cyclic oligomers pointing very close in alignment with the path between symmetry axes of the cages. Even 2 component cage symmetries such as I53 work with these building blocks.

These designs had the new docked interfaces designed with MPNN and are now being tested.

Adding interfaces to the sides of repeat proteins

The sides/surfaces of repeat proteins are an underutilized area for protein interface design. Successful designs on the surface of a DHR represent interfaces that can be grafted onto any repeat unit, perhaps multiple times throughout the length of the DHR. To investigate this feature, we made some “train track” DHR structures that feature un-capped split-DHR rails that are linked together by C2 straight DHR ties (**Figure 40**).

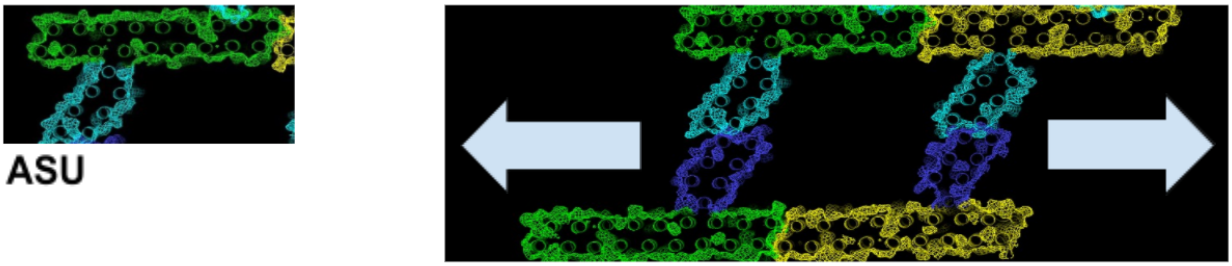


FIGURE 40. Train track schematic
Here we can use a 2 component system to make a train track fiber. Green/Yellow are the same chain- a split DHR with interfaces to interact with itself continuously on both ends- but color differentiated for clarity. Cyan/Blue are 2 identical chains of newly designed straight DHR-containing C2 designs. The only design that needs done is docking and designing the C2 to the split DHR and making sure they are straight/parallel. Then the train track shape just happens by virtue of the building block style.

Because the C2 is interfacing with the side of the repeat protein, we can change how many C2 ties are in a given length of train track by changing how many times the interface is repeated on the rails. Similarly, the C2 DHR lengths can change to spread apart or bring together the 2 rails more or less. This could be an interesting system for studying rates of fiber growth as a function of these features which affect the kinetics and rigidity of the fibers.

When first tested, these designs had poor solubility of the rail components, presumably because of lanky rails forming and flexibly tangling and such. In order to better assess the interactions of the 2 components of this system, both components were purified in denatured condition (or close to denatured) in 6 M guanidine-HCl, added together in molar ratio to match the fiber model, and then dialysed against a standard Tris buffer with low salt. This led to soluble yield of both components and assemblies that could be seen on negative-stain EM to be correct (**Figure 41**).

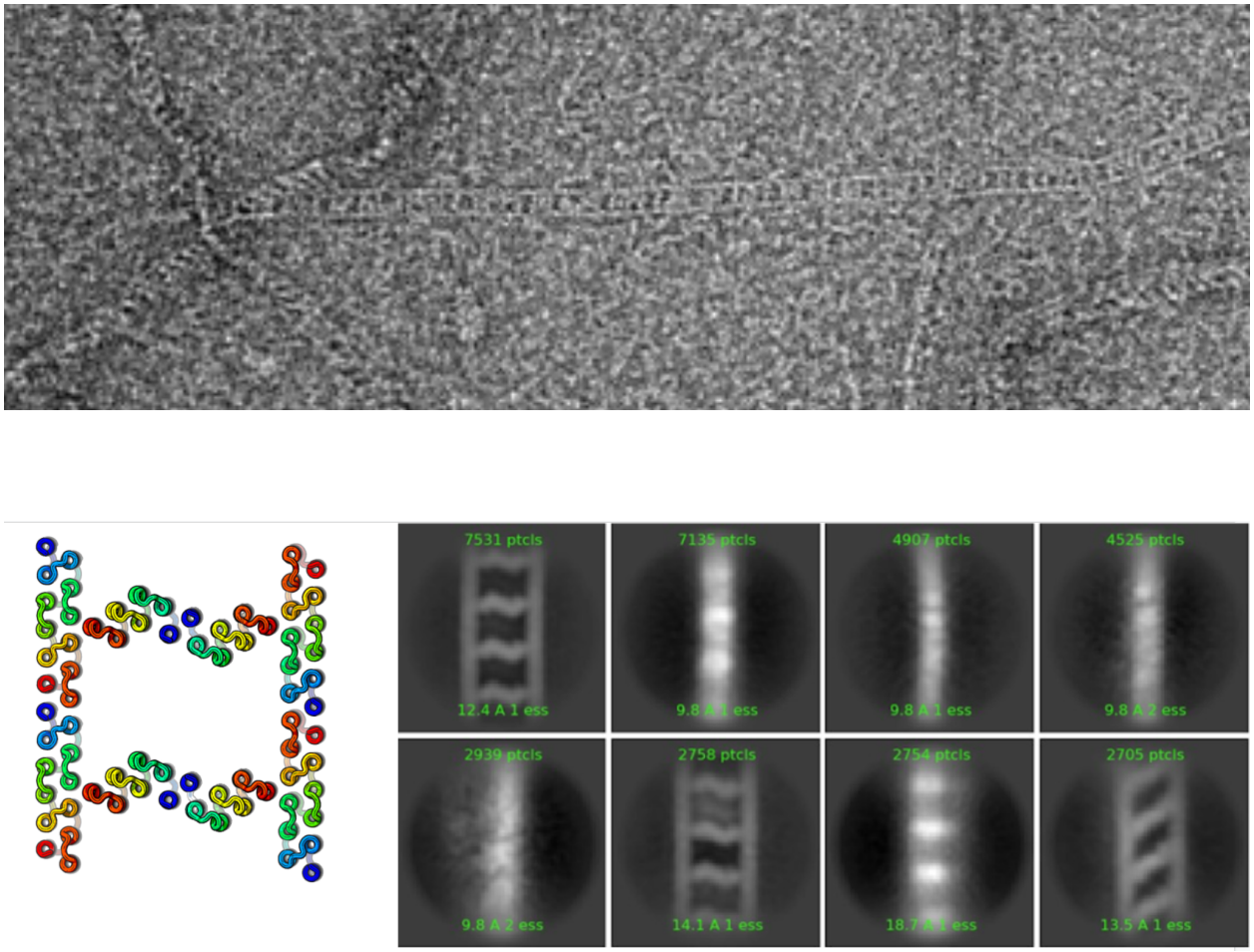


FIGURE 41. Train track structural data
On top we see a negative stain micrograph of a long train track fiber- each distance between rungs is about 4.5 nm. This averaged well and gave class averages that resemble the displayed design model.

We are following up on this success with variants of the tie size and spacing. Additionally we will try it with GFP attached to components so that fiber length can be assessed with other techniques, which can inform us which combination of tie spacings and sizes yield the longest fibers and the most robust assembly.

We are also trying a set of designs that feature DHRs docked cyclically against each other, and then all those cyclic designs docked “all-against-all” to each other (with trimming of extra DHR length to fit, as will be described in upcoming RPX Dock paper [20]) to see if we can achieve buttressed 2-component structures. This is another way to utilize interfaces on the sides of repeat proteins that could be powerful for designing unique structures. Some even had a 3rd component designed, a toroid generated by our “explicit” method to match size and encircle the new docked structure (**Figure 42**).

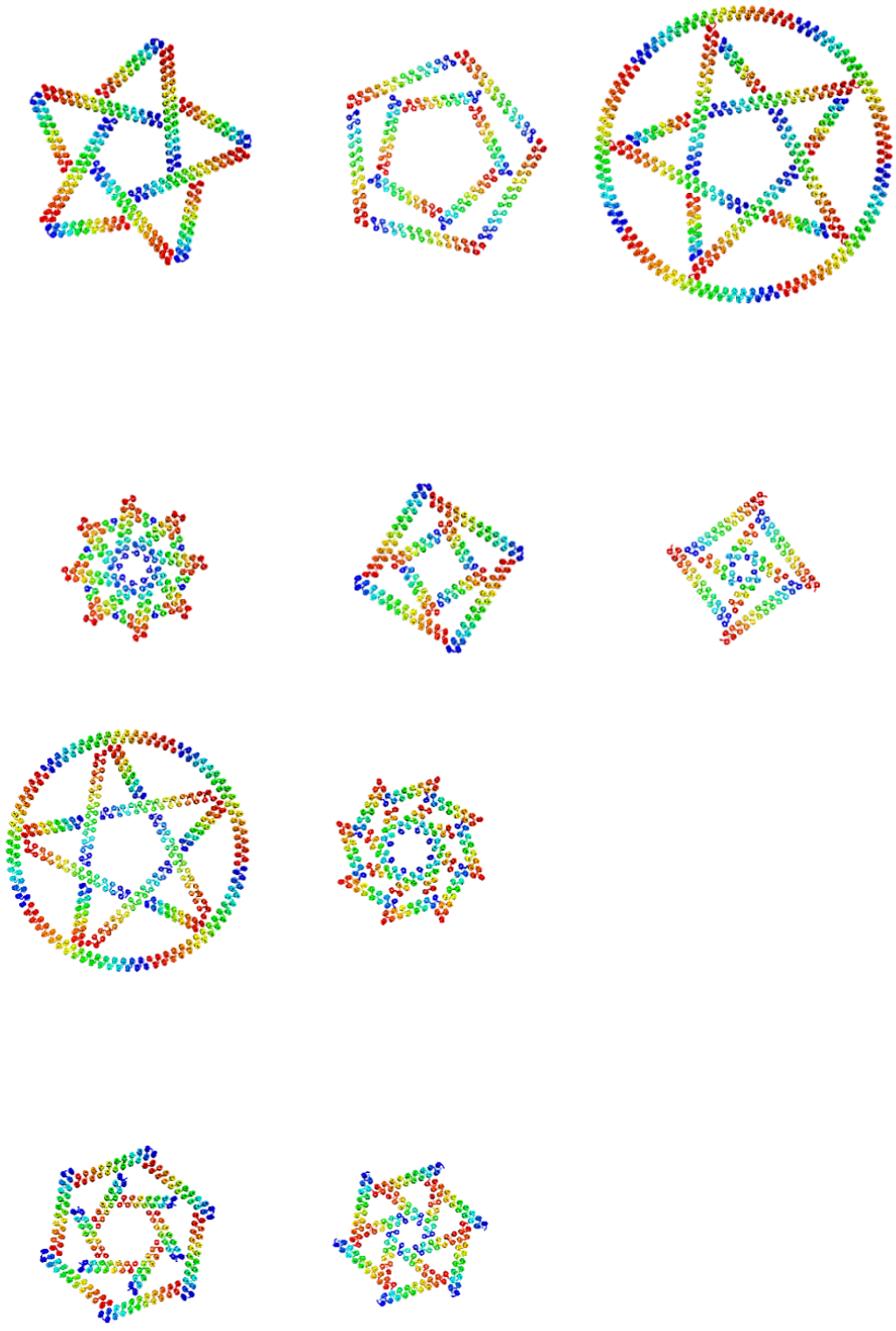


FIGURE 42. New rings with buttressing
Here are a variety of ring shapes from C4 thru C8 symmetry with different degrees of buttressing. The rings that have a circular protein around their edges are 3-component, and others have 2 components.

Based on how consistently the C2 ties were occupying their spots in the rails in the train track structural data, we also believe that making a track “ring” that consists of 2 toroid curved rails with ties between them will be a good way to make highly cooperative ring structures that have reinforcement for the correct shape. These are the next designs that we are excited to make (**Figure 43**).

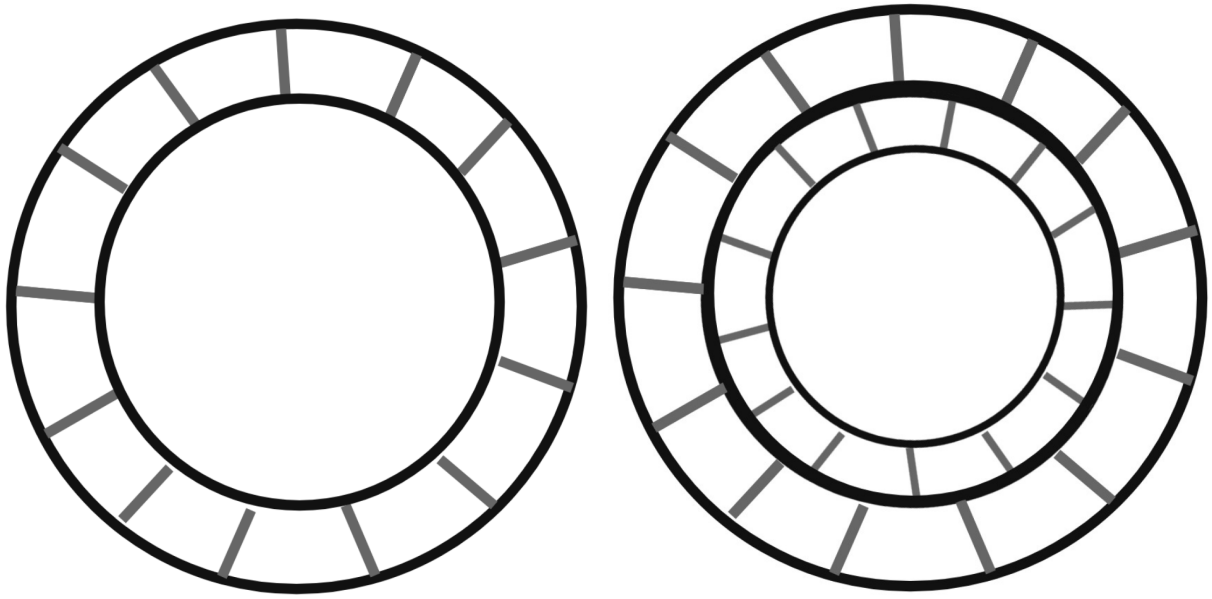


FIGURE 43. Schematic for new cooperative rings
On the left we have a ring concept showing if the train track design idea were applied to 2 concentric toroids of an appropriate difference in diameter so that repeating struts could be designed between them. Possibly to make this work we would need to have the exterior toroid repeat units be larger than the interior ones. On the right is a similar schematic but showing another layer added if we wanted to test for increased rigidity and cooperativity.

4. ONGOING WORK AND FUTURE DIRECTIONS

This work is ongoing with several members of the team still active. Data is still coming in for many of the types of structures detailed here. We think this is an exciting time because we are at the stage where it seems most of our designs of this simplified fold are likely folding correctly, and now we get to be creative about kinetics, assembly pathways, and experimental conditions to actually make large structures happen. It will also be particularly useful to stay up to date with publications regarding how best to use new machine learning methods, as the field is exploding with potential from these advancements.

We believe that we are carving out a small chunk of the protein design space that can actually work intuitively and enables creative structure design without the user having to master every arcane element there is to know about protein design. We hope that the modularity of these designs will facilitate experiments to help us understand the biophysics of various modes of protein complex assembly, particularly with designed proteins. If adequate mastery is gained over this space, it is hoped that the potential of protein nanomaterials can be more easily visualized and conceptualized so that they gain the ability to fulfill practical needs in industrial and biochemical nanomaterials markets.

5. ACKNOWLEDGEMENTS

Yang Hsia – extending nanocage work, armed trimers, and conceptual aid

Ryan Kibler– toroid designs and conceptual aid

Justas Dauparas– MPNN design program

TJ Brunette– early mentorship

Scott Boyken– early mentorship

Zibo Chen– early mentorship

Deepesh Nagarajan– DHR characterization

Asim Bera, Alex Kang– crystal structure solving

Rachel Redler– Cryo EM data collection and analysis

Nicolas Coudray– Cryo EM analysis

Phil Leung– DHR characterization

Neville Bethel

Lukas Milles

Basile Wicky

Andrew Borst

Kandise VanWormer

Austin Smith

David Baker

Lance Stewart

Zari Magness

Brian Coventry

Joe Watson

Harley Pyles

6. CITATIONS

1. PAULING L, COREY RB, BRANSON HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*. 1951;37(4):205-211. doi:10.1073/pnas.37.4.205

2. West MW, Hecht MH. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci*. 1995;4(10):2032-2039. doi:10.1002/pro.5560041008

3. Boyken SE, Chen Z, Groves B, et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity [published correction appears in Science. 2016 May 20;352(6288). pii: aag1318. doi: 10.1126/science.aag1318]. *Science*. 2016;352(6286):680-687. doi:10.1126/science.aad8865

4. Barnes CA, Shen Y, Ying J, et al. Remarkable Rigidity of the Single α-Helical Domain of Myosin-VI As Revealed by NMR Spectroscopy. *J Am Chem Soc*. 2019;141(22):9004-9017. doi:10.1021/jacs.9b03116

5.Pyles H, Zhang S, De Yoreo JJ, Baker D. Controlling protein assembly on inorganic crystals through designed protein interfaces. *Nature*. 2019;571(7764):251-256. doi:10.1038/s41586-019-1361-6

6. Walls AC, Fiala B, Schäfer A, et al. Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell*. 2020;183(5):1367-1382.e17. doi:10.1016/j.cell.2020.10.043

7. Correnti CE, Hallinan JP, Doyle LA, et al. Engineering and functionalization of large circular tandem repeat protein nanoparticles. *Nat Struct Mol Biol*. 2020;27(4):342-350. doi:10.1038/s41594-020-0397-5

8. Divine R, Dang HV, Ueda G, et al. Designed proteins assemble antibodies into modular nanocages. *Science*. 2021;372(6537):eabd9994. doi:10.1126/science.abd9994

9. Fleishman SJ, Leaver-Fay A, Corn JE, et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One*. 2011;6(6):e20161. doi:10.1371/journal.pone.0020161

10. Huang PS, Oberdorfer G, Xu C, et al. High thermodynamic stability of parametrically designed helical bundles. *Science*. 2014;346(6208):481-485. doi:10.1126/science.1257481

11. Grigoryan G, Degrado WF. Probing designability via a generalized model of helical bundle geometry. *J Mol Biol*. 2011;405(4):1079-1100. doi:10.1016/j.jmb.2010.08.058

12. MakeBundle Mover's author(s): Vikram K. Mulligan, Systems Biology, Center for Computational Biology, Flatiron Institute [vmulligan@flatironinstitute.org]. Reference: https://new.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/Movers/movers_pages/MakeBundleMover

13.. Brunette TJ, Parmeggiani F, Huang PS, et al. Exploring the repeat protein universe through computational protein design. *Nature*. 2015;528(7583):580-584. doi:10.1038/nature16162

14. Publication pending by Fatima Davila, Amy Stegman, Harley Pyles et al in David Baker's group

15. Reference: https://new.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/xsd/mover_ConnectChainsMover_type

16. Publication pending by Justas Dauparas et al in David Baker's group

17. Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev,Jimin Pei, David Kim, Elizabeth Kellogg, Frank DiMaio, Oliver Lange, Lisa Kinch, Will Sheffler, Bong-Hyun Kim, Rhiju Das, Nick V. Grishin, and David Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. (2009) Proteins 77 Suppl 9:89-99.

18. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

19. Hsia, Y., Mout, R., Sheffler, W. *et al.* Design of multi-scale protein complexes by hierarchical building block fusion. *Nat Commun* 12, 2294 (2021). https://doi.org/10.1038/s41467-021-22276-z

20. Rpxdock: a fast and versatile computational method for protein docking. Publication pending by Will Sheffler, Yang Hsia, Erin Yang et al in David Baker's group

21. Reference: https://edinburgh-genome-foundry.github.io/DnaChisel/notes.html

22. Reference: https://www.idtdna.com/site/order/geneentry

22. Studier FW. Protein production by auto-induction in high density shaking cultures. Protein Expression and Purification. 2005 May;41(1):207-234. DOI: 10.1016/j.pep.2005.01.016. PMID: 15915565.

23. Grant T, Rohou A, Grigorieff N. *cis*TEM, user-friendly software for single-particle image processing. *Elife*. 2018;7:e35383. Published 2018 Mar 7. doi:10.7554/eLife.35383

24. Doyle L, Hallinan J, Bolduc J, et al. Rational design of α-helical tandem repeat proteins with closed architectures. *Nature*. 2015;528(7583):585-588. doi:10.1038/nature16191

7. GLOSSARY OF COMMON ABBREVIATIONS USED

DHR - Designed Helical Repeat protein. Proteins with repetitive alpha helical secondary structure.

SEC - Size Exclusion Chromatography. Proteins travel through porous resin and are retained differentially according to their size. Larger proteins travel faster.

SDS - PAGE - Sodium Dodecyl Sulfate Polyacrylamide gel electrophoresis. A “denaturing” gel in which the SDS helps unfold and uniformly coat the protein in negative charge so it can migrate through a gel based primarily on its polymer length.

EM - Electron microscopy. Both negative-stain and cryo EM are used in this work. Negative stain EM uses a carbon film on which the protein sample is laid and then treated with a heavy-metal containing stain (such as Uranyl Formate). Then, the stain stain scatters the electron beam strongly, while the protein sample makes spaces that are devoid of stain and can be easily visualized. Cryo EM involves freezing the protein sample in amorphously frozen solvent, which allows the density of the proteins to be the relatively strong beam scattering feature.

C, T, O, I *(in the context of symmetries)*- These refer to cyclic, tetrahedral, octahedral, and icosahedral symmetries respectively. Any numbers after the letters indicate the oligomer size of any protein component(s) comprising the symmetric structure.

RMSD - Root-mean-square deviation. A similarity measure between molecules; lower is more similar.