Engineering novel nanomaterials through d*e novo* design of hydrophobic scaffold proteins

Dane Zambrano

A dissertation

submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading committee:

Neil King, Chair

Alex Merz

Kelly K. Lee

Program Authorized to Offer Degree:

Biochemistry

Dane Zambrano

University of Washington

Abstract

Engineering novel nanomaterials through d*e novo* design of hydrophobic scaffold proteins

Dane Zambrano

Chair of the Supervisory Committee:

Neil King

Department of Biochemistry

Computationally designed protein nanoparticles leverage *de novo* designed protein subunits to build large oligomers which can be used for multivalent antigen display and cargo loading in the delivery of vaccines. Advancements in computational methods and design of membrane proteins increases the breadth of materials which can be used as subunit building blocks in these large oligomeric complexes. In this dissertation I describe engineering a protein scaffold with a hydrophobic pore and the rational design of a novel two-component nanoparticle utilizing a *de novo* designed transmembrane protein. Using machine learning and AI-guided approaches with two-component RPX Docking, AlphaFold 2 and RosettaFold Diffusion, I structurally characterized a novel two component transmembrane nanoparticle and to 4.16 Å by Cryo-EM. This technology marks the first generation of novel hydrophobic nanoparticles that is a step towards generating *in vitro* hybrid lipid-protein nanomaterials for the display of unique membrane proteins and lipid-conjugated moieties previously inaccessible through current designed nanoparticles.

TABLE OF CONTENTS

Acknowledgements

First and foremost, I have to thank my PhD advisor, Neil King. Your positive attitude and endless encouragement during times when my experiments did not return the results I was looking for has been a driving force during this PhD journey. Thank you for your patience when data came slowly and for believing in my projects even when I did not.

I want to extend my deepest gratitude to my committee members: Alex Merz, Libin Xu, David Veester and Kelly Lee for all your feedback and guidance. Between everyone, I had a wealth of knowledge and experience and I am very honored and grateful for every second, minute and hour you took out of your schedules to guide me.

Next, I would like to thank my collaborators at the Institute for Protein Design's Electron Microscopy Research Core (IPD EMRC). Specifically, I would like to thank Andrew Borst, Connor Weidle and Kenneth Carr for their support in my project. I would also like to take the time to acknowledge my other collaborators: Xinting Li (IPD) and Desiree Reid (Marty Lab, University of Arizona) who have provided me with mass spectrometry data and help with data interpretation. Thank you to: Kandise VanWormer, Hernan Nunez-Ortega, Rafael Ticzon, Andre Dubief, and Garrett Ruth for providing laboratory support as well as Luki Goldschmidt, Patrick Vecchiato, and Bulat Faezov for technical support.

I want to thank all past and present King lab members, who have made the lab a fun place to come to work. I would specifically like to thank: Thad, Quinton, Brooke, Dan Ellis, John, Chloe, Dan Humphrys, Adam, Cara, Grace, Marti, Bianka, Yuhan, Feli, Zijian and Sanela. You have all given me guidance and words of encouragement during my rotation and throughout my PhD. Thank you for being such great lab mates and friends. Having supportive role models and office mates really made the tough science days a little better, and the words of encouragement really made a difference.

Thank you for all of the friends I’ve made in Seattle that have made Washington State feel like home – you all break up the long science days with gym time, comic conventions, concerts, dancing, karaoke and volleyball. I’ve discovered new hobbies, food and adventure thanks to you all.

I want to thank my mother for absolutely everything she has done for me. You are an inspiration, mom. Every day I am reminded of your strength and resilience. It’s not easy to be a first-generation immigrant, and it’s also not easy to be the first in the family to pursue my PhD, but I’m reminded that you also were a “First” for many things in your family. I’m proud and honored to be your daughter.

The one constant during the turbulent PhD was my husband. Aside from my mother, no one has given me as much motivation and support as you. I’m very thankful to have your strength during the days when I was weak. Thank you for your endless support and patience. I love you.

♡

Chapter 1: Introduction

Naturally occurring protein scaffolds are structures with inherent self-assembling properties in which protein subunits interact with each other through weak, non-covalent interactions to form larger complexes. One example of a naturally occurring protein is ferritin, a ubiquitous iron storage protein found in bacteria and eukaryotes that self-assembles into a 24-subunit octahedral nanoparticle[1,2]. In highly acidic or basic environments ferritin nanoparticles disassemble and can be reassembled when the solution becomes more neutral, making it an ideal candidate to trap small molecules within its cavity[1]. Scientists have used this property to load ferritin particles with doxorubicin, an anticancer antibiotic, and target the drug to specific tissues by modifying key surface residues[1]. As well as carrying cargo on its interior, ferritin has also been modified to display antigenic proteins on its surface[2,3]. Through the genetic fusion of the flu antigen hemagglutinin (HA) to ferritin monomers, HA-ferritin nanoparticles enhanced the potency of neutralizing antibody responses in ferrets[2]. Although these examples show that natural nanoparticles are designable, targetable, and immunogenic, particle heterogeneity, improper folding of fused antigens, and antigen interference with nanoparticle self-assembly still need to be overcome[4].

Computational protein design is a field that has been inspired by naturally occurring self-assembling protein complexes and overcomes some challenges of using natural scaffolds[2,5–7]. In addition to similar benefits of natural scaffolds such as multivalent display and cargo loading, designed nanoparticles boast more control and modularity over the protein scaffolds that comprise the nanoparticle[8–10]. A commonly employed strategy for generating computationally designed nanoparticle assemblies involves the initial creation of smaller protein building blocks which self-assemble into higher-order structures[8,10–14]. These subunits form nanoparticles with precisely defined architectures and functional properties and are highly tunable[11–17]. As in naturally occurring protein complexes, designed nanoparticles take advantage of shape complementarity and weak, non-covalent interactions to self-assemble into larger oligomers[5,7,11]. By precisely designing the amino acid sequence of *de novo* proteins, it is possible to control the self-assembly process at the interface between the two subunits[14,16]. *De novo* nanoparticles face challenges in preclinical and early clinical development compared with nucleic acid and vector-based vaccine platforms, however, multiple intrinsic advantages of designed nanoparticle vaccines motivate their prioritization. For example, designed nanoparticles have been used for a number of biomedical applications such as multivalent antigen presentation for vaccine design leading to potent antibody responses[9,14,15,18–21]. The rational design of antigen presenting nanoparticles comes with the possibility of delivering *de novo* proteins as drugs and vaccines, however challenges in targeting and delivery of protein nanoparticles still remain[17,21].

Lipid-based nanoparticles (LNPs) are a class of designed nanoparticles that are highly adaptable and can be targeted to carry cargoes to different cellular compartments[22–25]. LNPs can be designed through the modification of lipid headgroups, lipid tails, or the addition of synthetic groups like polyethylene glycol (PEG) to stabilize, target, and functionalize LNPs for specific tasks[25,26]. LNPs offer several benefits, such as boosting a drug's solubility, increasing biodistribution and safeguarding drugs from in vivo degradation[22–27]. This technology has increased in popularity due to the rapid development and immunological success of mRNA vaccines against SARS-CoV-2 variants delivered by LNPs[22,28–30]. LNPs have also been used to deliver small proteins for therapeutic applications, suggesting that the LNP platform can carry a variety of cargoes with properties different from nucleic acids[31]. In one example, the small Designed Ankyrin Repeat Protein (DARPin), K27, was modified and encapsulated within LNPs where the K27-LNP was effective at reducing tumor load in mouse models[31]. The rapid development and biochemical properties make LNPs appealing for drug delivery applications, however the in vivo efficacy of liposomes carrying drugs can be highly impacted by size, surface charge, lipid composition, lamellarity, and stability[25,32]. The combination of LNP technology with the tuneability and modularity of protein-based nanoparticle design would be highly desirable.

In this dissertation I describe generating new unique scaffolds which take advantage of hydrophobic interactions between proteins and lipids. Here I characterize the design of two protein nanomaterials with functional groups building towards hybrid nanomaterials. These designed nanomaterials are the first generation of hybrid lipid-protein nanoparticles derived from *de novo* protein design.

Chapter 2: Toroid protein membrane mimetics for the display of membrane proteins

Background

Membrane mimetics have been used to characterize lipid-protein and protein-ligand interactions of membrane proteins[33–36]. The use of membrane mimetics in structural studies has provided insights to the effects of the membrane environment on protein structure[37,38]. HDL nanodiscs are composed of modified apolipoprotein which are aliphatic naturally occurring proteins that transport cholesterol and have been engineered as HDL nanodiscs to contain discoidal patches of lipid membrane[35,39]. Nanodiscs have been used to solubilize membrane-anchored antigens and peptides and deliver them as therapeutics in mouse models[40–42]. Mice vaccinated with HDL nanodiscs formulated with epitope protein generated antigen-specific cytotoxic T-lymphocyte responses, compared with the soluble epitope[41]. The measured responses of HDL nanodiscs revealed that epitope vaccination generated broad-spectrum T-cell responses when combined with immune checkpoint inhibitors[40,41]. However, due to the limited size of the lipid patch, membrane protein activity can alter in different types of membrane mimetics under the same lipid environment[43]. There is significant motivation to design a membrane mimetic that is highly defined and can be tuned to different sizes. Unlike HDL nanodiscs, toroids are highly designable and their circular structure has made them ideal scaffolds to build larger oligomeric assemblies[44–46]. Here we modeled toroid discs after HDL nanodiscs to contain discoidal patches of membranes. Compared with the flexibility and heterogeneity of HDL nanodiscs, toroids are more rigid in structure and provide a stable foundation for protein design[47].

Results

Expression of helical designed toroids modified with hydrophobic residues

The *de novo* designed toroid protein PDB ID 5BYO was the starting building block for toroid discs as it was structurally validated by X-ray crystallography and a similar height to HDL nanodiscs of 2 nm[35,47]. The original model of the toroid was a single-chain design with 12 repeating helix-turn-helix motifs[47]. We introduced hydrophobic point mutations on the interior-facing residues and tetramerized the design to have smaller oligomers and reduce solubility issues during protein expression. Five internal residues on the toroid were modified from polar residues to hydrophobic (**Figure 1A-B, Supplemental Figure 1A**). Of the designs chosen for protein expression, mutations that were fully hydrophobic did not express well (**Supplemental Figure 1A**). On the other hand, introducing a polar or aliphatic residue in the design produced three well-expressing proteins (**Figure 1A-B, Supplemental Figure 1A**). The three expressing candidates were purified by size exclusion chromatography (SEC) where two of the candidates had noticeable features on the SEC trace showing signs of aggregation (**Figure 1B**). Candidate RLLLR had a monodisperse SEC trace and eluted between 15 - 18 mL, a size corresponding to a size between 44,000 - 13,700 Da for the SEC purification column which was used for purification (**Supplemental Figure 1B**).

The lipid 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) was mixed with toroid to test the ability of the candidate RLLLR to capture lipids. We adapted an HDL nanodisc protocol and added a ratio of 20:1, 60:1 and 70:1 POPC lipids to toroid protein by molar concentration of protein[39]. Following SEC purification of the toroid-lipid mixture, samples were analyzed by native mass spectrometry to validate the size of the complex. The mass of a monomer for candidate RLLLR is 39,557 Daltons and an estimated 70 POPC molecules are expected to fit into the internal cavity of the protein for a total complex mass of 210,728 Daltons. However, native mass spectrometry of the toroid-lipid complex revealed a smaller size than expected, at 144,370 Daltons, well below our expected calculated mass (**Supplemental Figure 1B**). Repeated efforts to mix the protein with lipids did not yield reproducible toroid-lipid data by native mass spectrometry, thus, the Rosetta molecular modeling package was used to extend the helices to build taller toroid structures from 2 nm to 4 nm.

The height of eukaryotic plasma membranes ranges from 2.5 - 5 nm[48,49], therefore, increasing the height of the toroids provides 16 - 18 carbon lipids such as POPC a greater surface area for lipid-protein interactions (**Figure 1C**). Interestingly, one design for taller toroid proteins produced a doublet during SEC purification (**Figure 1D**). As only the interior residues of the toroids were modified as hydrophobic residues, the rest of the protein remained hydrophilic. We hypothesized that the doublet peak for the tall toroids was caused by stacking interactions of the proteins (**Figure 1E**)[47]. ns-EM for the main SEC peak and the shoulder peak revealed that both SEC peaks contain toroid proteins (**Supplemental Figure 1C**). We also characterized the SEC peak eluting at 11.3 mL using small-angle X-ray scattering (SAXS) which provided low resolution data of the toroid where the observed curves were consistent with the protein shape mass of the computationally predicted structure (**Supplemental Figure 1C**). To characterize the toroid sample to a higher resolution we used negative stain electron microscopy (ns-EM) where processed toroid particles show well formed rings for the SEC purification peak eluting at 11.3 mL which corresponds to the expected size of the designed toroid protein (**Figure 1E**). The ease of protein expression and purification of the hydrophobic toroid rings in both the short and tall design models was an encouraging result which suggests the adaptability of these proteins as design candidates to build modular nanodiscs.

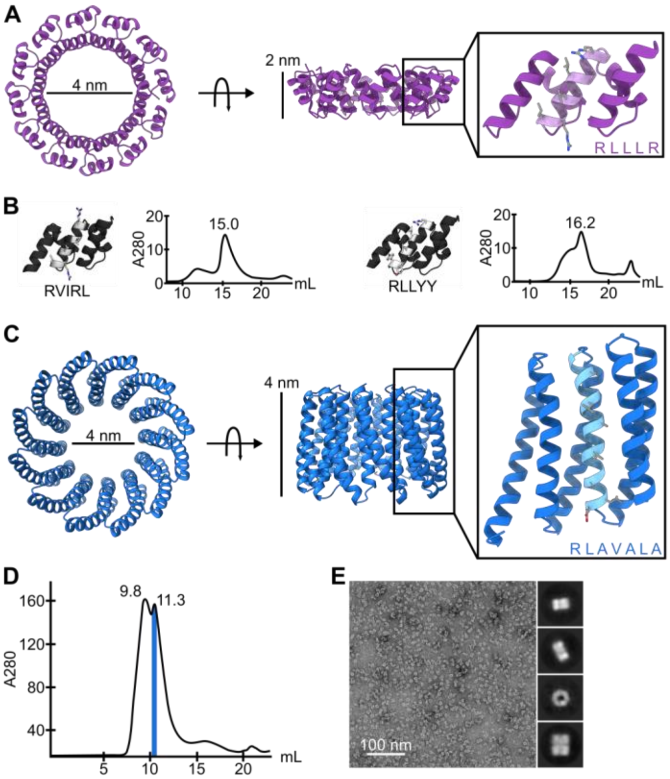


Figure 1: Short and long helical designed toroids modified with hydrophobic residues

(A) The toroid protein PDB ID 5BYO designed by Doyle et al, was used as the base model for the design of toroid nanodiscs[47]. The protein has an internal diameter of 4 nm and height of 2 nm. In each 9-helical bundle unit, the center internal-facing helix would be mutated with hydrophobic residues. (B) Two expressing toroid disc candidate designs of different mutations were characterized by SEC. SEC traces show multiple broad peaks suggesting protein aggregation. (C) The toroid helices were extended from 2 nm to 4 nm in height to provide a larger surface area for lipid-protein interactions. Seven residues on the interior helix were mutated. (D) SEC purification of tall toroid candidate RLAVALA. (E) ns-EM raw micrograph taken from a magnification of 57,000x and 2D class averages of the highlighted SEC peak in (D)

Lipid handles through maleimide conjugation of tall toroids

Although the simple addition of lipids to short toroid proteins did not produce toroid-lipid nanodiscs by native mass spectrometry, we investigated different strategies to contain lipids within the toroid discoidal cavity. To encapsulate DNA nanoparticles lipid "handles" were designed into the DNA nanoparticle scaffold by inserting lipid-conjugated oligonucleotides into the design[50]. The addition of lipid handles to these DNA nanoparticles was a proven strategy to instigate bilayer encapsulation around a large DNA nanoparticle[50]. We explored covalently conjugating n-dodecyl maleimide to the interior of our tall toroid proteins through a cysteine residue to employ a similar nucleation strategy as lipid handles (**Figure 2A**). 1 or 2 cysteine residues were introduced to the toroid candidate RLAVALA which was characterized by SEC and ns-EM. For the single mutation variant, a cysteine was introduced to the center helix and in the middle of the protein's height (**Figure 2B**). For the double mutation, cysteines were introduced on different helices of the same subunit (**Figure 2B**). The full toroid ring assembly of tetramers has a range of 4 or 8 cysteine residues for maleimide conjugation. Cysteine mutants were expressed and purified by SEC. The purified toroid proteins were mixed with n-dodecyl maleimide and the conjugation efficiency was assessed through liquid chromatography-mass spectrometry (LCMS). The masses for the conjugated protein samples reported by LCMS confirmed the successful conjugation of n-dodecyl maleimide to both of the cysteine constructs (**Figure 2C**).

Encouraged by the successful conjugation of n-dodecyl maleimide to the interior of the toroid constructs, we began titrating POPC lipids with our protein-maleimide conjugates. Lipid titrations for POPC were performed for the single cysteine toroid mutant ranging from 0:1 to 120:1 lipid-to-protein molar ratio values (**Supplemental Figure 2A**). After mixing and incubating POPC with toroid proteins, we purified the mixture using SEC. As the lipid ratio increased, a new SEC peak eluting earlier than the toroid protein emerged at approximately 9.5 mL (**Supplemental Figure 2A**). Initially this peak was believed to be a large complex of lipids within the toroid protein, however upon screening with ns-EM this larger peak contained only liposomes or lipid aggregates (**Supplemental Figure 2B-C**). While the titration of lipids did not produce intended results, this design method indicates preliminary data for the successful conjugation of n-dodecyl maleimide to our designed protein.

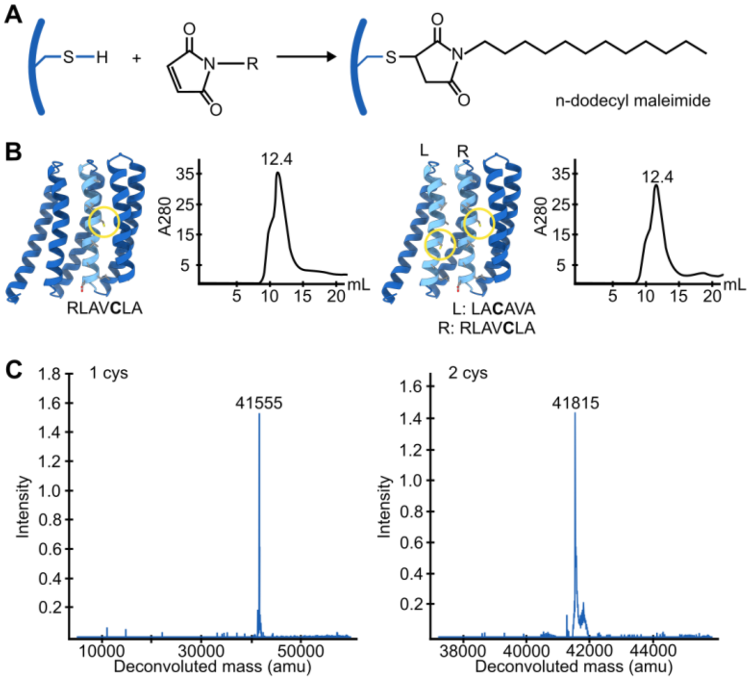


Figure 2: n-dodecyl maleimide conjugation to long toroid proteins

(A) Schematic of the maleimide reaction covalently linking n-dodecyl maleimide to the toroid protein. (B) 1 or 2 cysteine mutations were introduced to the toroid candidate protein RLAVALA. Cysteine mutations are circled in yellow. Cysteine mutants were purified by SEC. (C) LCMS data for the cysteine mutation constructs show successful conjugation of n-dodecyl maleimide to the toroid protein. The mass of the original hydrophobic construct is 41258 atomic mass units, while the mass of a single unit of n-dodecyl maleimide is 265 atomic mass units.

Discussion

Here we describe methods to generate designable membrane mimetics using *de novo* toroid proteins. These preliminary results indicate that we are able to tune and modify the base toroid design without losing protein expression. Furthermore, the successful conjugation of n-dodecyl maleimide shows that a hydrophobic motif can be inserted into the toroid ring and is an initial step towards nucleating a discoidal membrane bilayer within the toroid.

Initially, we were hopeful in our goal to produce membrane mimetics due to the simplicity and swiftness of structural characterization for the helical toroid proteins. However, with no high-throughput assay to validate successful lipid-protein interaction, it was difficult to assess whether we were ever successful at making a membrane mimetic. Furthermore, a limiting factor for lipid bilayer formation was likely the restrictive diameter of the toroid rings. HDL nanodisc diameters range from 4 - 16 nm and while we designed the toroid rings within the range of HDL nanodisc diameters, the small size and inflexibility of the toroid proteins compared with the HDL nanodisc protein could discourage lipid packing[51]. To overcome this size limitation, protein design of toroids with larger diameters should be prioritized.

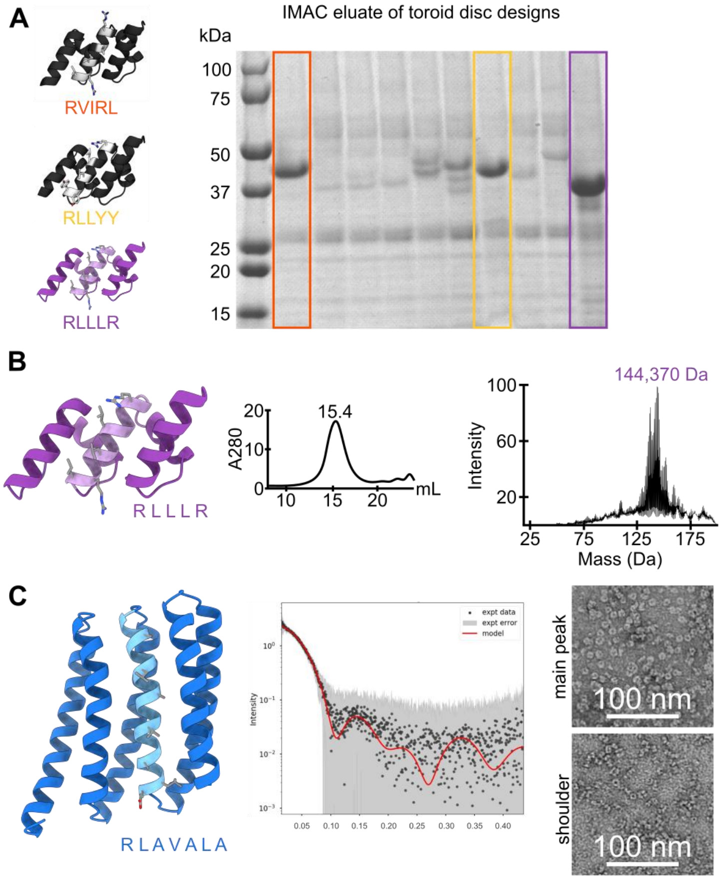
Another rationale why the tall churros were not able to compartmentalize lipids other than its size, may be due to the helical structure of the protein interior, thus we were inspired by the TIM barrel beta-turn-helix motif in our next round of toroid design (**Supplemental Figure 3A**). The beta sheet character of this design would provide the outer helices a large surface area for residue packing, while the interior hydrophobic residues of these new beta-toroids would also provide lipids with a planar surface for interaction (**Supplemental Figure 3A**). We modified RoseTTAFold diffusion (RFDiffusion) designed proteins previously characterized by cryogenic electron microscopy (cryo-EM)[52]. As in the helical toroid designs, we tetramerized the proteins and modified the interior residues as hydrophobic (**Supplemental Figure 3A**). Unlike helical toroids, all interior residues of beta-toroids could be mutated to hydrophobic residues and cysteine mutations were introduced without penalty of aggregation observed by SEC (**Supplemental Figure 3B**). However, the conjugation of n-dodecyl maleimide to the beta-toroids was unsuccessful. LCMS data of the protein conjugate revealed no mass of the protein oligomer and no proteins in complex with lipids (**Supplemental Figure 3B**).

There are still a lot of upstream studies that need to be done to further amplify the applications of toroid nanodiscs. Although the lipid handle strategy was unsuccessful in the context of toroid design, this technique may be applicable in other contexts to make hybrid nanomaterials[50]. These findings underscore the need for further optimization for the toroid protein height, internal diameter, and hydrophobicity before the next iteration of designs. Furthermore, high throughput assay development is necessary to be able to screen multiple designs and mutations in order to iterate on a faster design pipeline. On the other hand, these preliminary results show that toroid proteins are tolerable to hydrophobic mutations and are still good candidates for exploring designed membrane mimetics.

Acknowledgements

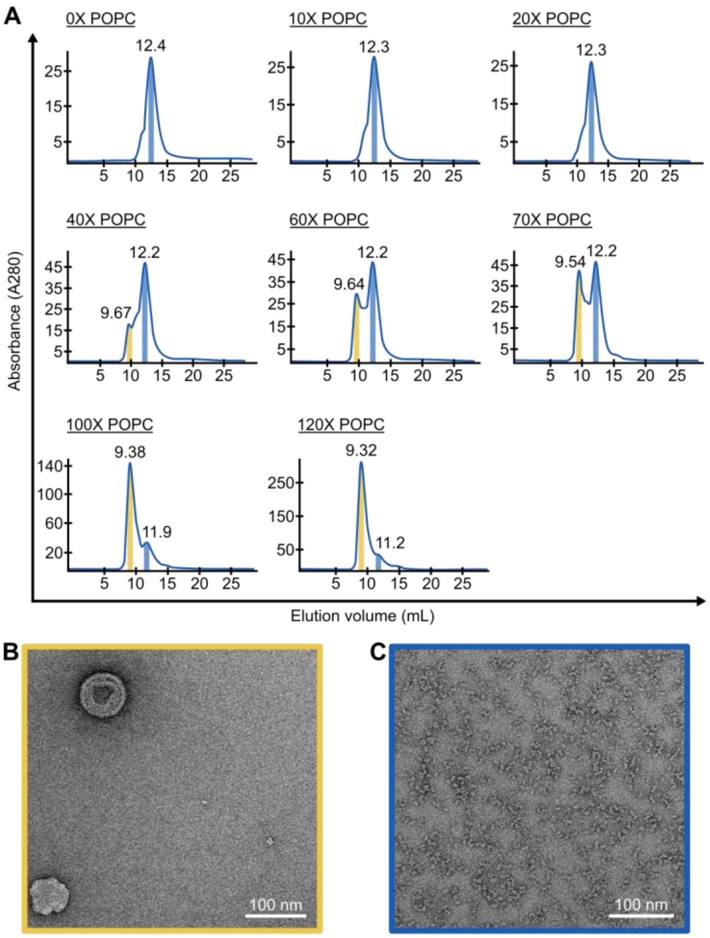
I would like to thank Thad Huber for designing the first iterations of the shorter toroid proteins. Before his departure, he set the foundation for the computational design pipeline which gave me a strong starting point to build the taller toroid models. I would also like to thank Helen Eisenach for lending me her design model for the beta-turn-sheet toroid rings and Ryan Kibler for sharing Python scripts to run RFDiffusion. Lastly, I would like to extend my gratitude to Masaharu Somiya for helpful discussions about lipids and liposome dynamics. All the LCMS and SAXS work in this chapter was possible because of the help of the Institute for Protein Design Mass Spectrometry Core and Protein Core.

Supplemental Data



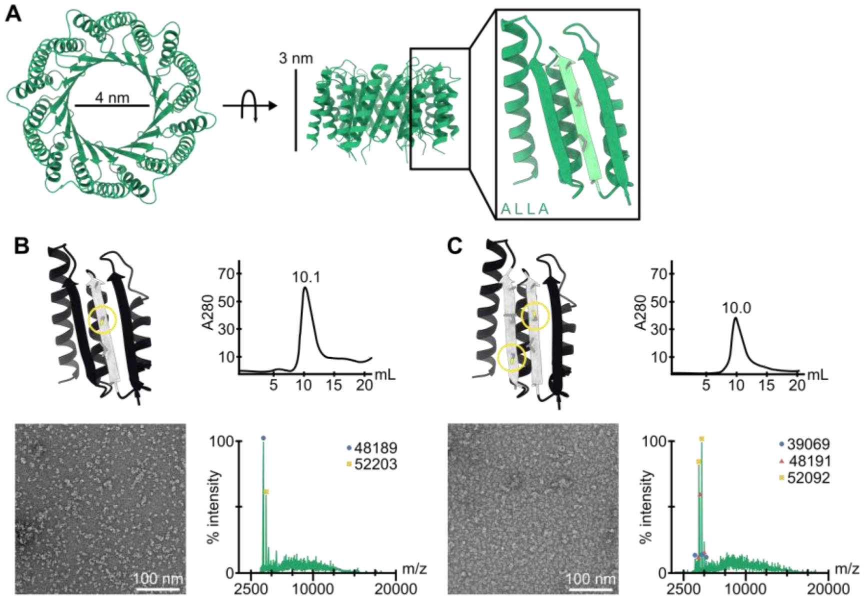
Supplemental Figure 1: Expression and characterization for short and tall helical toroid proteins

(A) SDS-PAGE of the IMAC eluate for 10 different toroid constructs. The constructs have different hydrophobic residue mutations on the interior helix of the protein. The three best expressing designs are highlighted and shown. (B) The best expressing design was construct RLLLR. The protein was purified by SEC and a ratio of 60:1 POPC lipids was added to the protein and characterized by native mass spectrometry (nMS). nMS data for the toroid-lipid sample was smaller than the expected size of the toroid-lipid complex. (C) Short toroid helices were extended and the tall toroid proteins were characterized by SAXS and ns-EM. ns-EM of the SEC main peak and the shoulder peak show toroid proteins present in both peaks.



Supplemental Figure 2: Maleimide-conjugated toroid lipid titration

(A) Lipid titration of POPC with single cysteine mutation tall toroid proteins conjugated with n-dodecyl maleimide. Lipid-to-protein molar ratios start at 0:1 and increase to 120:1 lipid-to-protein. After the 40:1 mixture, the emergence of a second peak becomes prevalent in SEC traces. (B) Raw ns-EM micrograph of the earlier SEC peak (yellow) reveals that the peak contains liposomes or lipid aggregates. (C) Raw ns-EM micrograph of the earlier SEC peak (blue) of tall toroid proteins.



Supplemental Figure 3: Beta-turn-helix motif scaffold for maleimide conjugation

(A) Beta-toroid protein design using RFDiffusion building blocks that was characterized by cryo-EM[52]. This design was tetramerized similarly to the helical toroids and the interior residues mutated to hydrophobic residues. (B) 1 or 2 cysteine residues were introduced to the interior beta-sheet to conjugate n-dodecyl maleimide. Mutants were purified by SEC and screened by ns-EM. Following the conjugation reaction, beta-toroids were assessed by LCMS, however masses analyzed did not correspond to the mass of the protein with the addition of n-dodecyl maleimide.

Materials and Methods

Internal residue mutation of short and tall toroid and beta-toroid proteins

Published design models for the short toroid protein (PDB ID 5BYO[47]) and the beta-toroid (HE0626[52]) were loaded into PyMol where the sequence and select feature were used to pick internal residues[53]. Alanine, valine and leucine were the most common mutations made.

Extension of the short toroid protein using Roseta

RoseTTAFold diffusion (RFDiffusion) was used to extend the helices of 5BYO by taking the input and desired thickness of the toroid into consideration. After obtaining RFDiffusion pdb file outputs, files were ran through Protein MPNN[54] then AlphaFold 2[55] where designed toroids were filtered based on backbone RMSD to the RFDiffusion outputs. RMSD values that were picked for *in vitro* testing fell between 0.987 and 1.57 Å. Selected pdb files were loaded into PyMol where the sequence and select feature were used to pick internal residues[53]. Alanine, valine and leucine were the most common mutations made.

Purification of short and tall helical toroid and beta-toroid proteins

Plasmids ordered from GenScript for the tetrameric toroid subunits were transformed into *E. coli* BL21(DE3) competent cells (NEB, Cat. C2527H) and plated on an agar plate treated with a final concentration of 50 µg/mL kanamycin. A single colony was picked from the transformed cells and grown in a 50 mL culture of LB broth (Lennox formulation) overnight at 37.0°C. 5 mL of the overnight culture was transferred into 500 mL of Terrific Broth (TB) media treated with a final concentration of 50 µg/mL kanamycin. Cells were grown in TB media at 37.0°C until an OD600 of 0.60 - 0.80 before induction with a final concentration of 1 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG) and grown for an additional 12 hours at 18.0°C. Cells were harvested at 5000 revolutions per minute (rpm) for 30 minutes and resuspended with a lysis buffer (25.0 mM Tris-HCl pH 8.00, 0.5 M NaCl, 1.00 mg/mL lysozyme and 1.00 mg/mL DNAse) then lysed using sonication. Cells were lysed at 4°C using a probe sonicator for a total of 5 minutes using 10.0 second sonication pulses and 10.0 second breaks. After cell lysis, the solution was clarified at 10,000*g* for 30 minutes. The supernatant was collected and ran through a nickel-nitrilotriacetic acid (Ni-NTA) IMAC resin with a bed volume of 5.00 mL (Qiagen, Cat. 30250). The IMAC resin was equilibrated with 5.00 column volumes of loading buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M NaCl, 10 mM imidazole) before the addition of cell lysate, then washed with 3 column volumes of the loading buffer. The bound protein was eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM NaCl, 0.50 M imidazole, 0.01 mM EDTA).

Addition of POPC to toroid proteins

1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) was ordered from Avanti Polar Lipids in chloroform at a concentration of 25 mg (850457). Different volume amounts for POPC depending on the molar concentration of toroid protein was calculated and was transferred to a 1-dram borosilicate glass vial. The chloroform was evaporated under a vacuum overnight. The thin film of lipid was rehydrated with toroid protein in the elution buffer and the mixture was left untouched for 1 hour before bath sonication at 20C for 10 minutes. Following bath sonication, the mixture was mixed slowly on a rotator for 1 hour before injecting the sample onto the Superdex 200 Increase 10/300 GL (Cytiva 28990944) for SEC purification.

ns-EM sample preparation for tall toroid protein (Figure 1E)

After SEC, the isolated toroid peak was diluted with the SEC buffer to 0.08 mg/mL. 3 µL of the diluted sample was applied to a freshly glow discharged 400 carbon square mesh grid (Electron Microscopy Sciences, EMS) and allowed to adsorb onto the grid for 30 seconds. Blotting paper (Wattman) was used to remove excess solution from the grid before the application of 3 µL 2% uranyl formate (2UF) to stain the grid for ns-EM. Excess stain solution was blotted immediately followed by another application of 3 µL 2UF. This was repeated once more, resulting in a total of 3 applications of 2UF. Following the final application of 2UF, all remaining solution was blotted and the sample was allowed to dry completely before imaging. There was no waiting period in between each application of 2UF, and after each blot new 2UF was applied to the grid immediately.

ns-EM data acquisition and processing for tall toroid protein (Figure 1E)

Negative stain electron microscopy (ns-EM) micrographs were collected on a Thermo Fisher Scientific Talos L120C 120 kV transmission electron microscope with a LaB_6 filament and CETA camera with a pixel size of 2.49 Å at 57,000x magnification. 188 micrographs were recorded at 57,000x magnification at a total dose rate 45.2 $e^-/Å^2$ for the octahedral nanoparticle. Micrographs were imported into CryoSPARC v4.7.0 [56] and Patch CTF Estimation was performed prior to blob picking, inspection, and extraction of 175,335 picks with a box size of 80 pixels. With CTF correction off, picks were classified into Reviewed 100 2D classes, 28 of which were selected to serve as templates for an additional round of particle picking. In this second round of particle picking, picks were inspected and extracted to 80 pixels before classification of 168,067 extracted particles into 100 2D classes with CTF correction off. 38 out of 150 of these 2D classes were selected as a second round of templates for another round of particle picking and extraction. In this last round, 172,098 particles were extracted to 80 pixels using the templates from the second round of extraction and classified into 50 2D classes with CTF correction off and maximum alignment resolution 6 Å to generate 2D class averages.

N-dodecyl maleimide conjugation to tall toroids and beta-toroids

n-dodecyl maleimide was ordered from ABCR (CAS 17616-03-4) as powder. The final ratio of maleimide-to-protein was 1:1 by molar concentration of the protein. n-dodecyl maleimide powder was measured using an analytical scale and mixed in DMSO before adding the mixture to the toroid protein 1:1 (c/c). The mixture incubated at room temperature for 1 hour before quenching the reaction with 0.5 mM TCEP and injecting the sample onto the Superdex 200 Increase 10/300 GL (Cytiva 28990944) for SEC purification.

LCMS and native mass spectrometry of toroids and toroid-lipid mixtures

The LCMS and native mass spectrometry data collected in this paper was in collaboration with the Mass Spectrometry Core Facilities at the Institute for Protein Design, University of Washington (LCMS) and the Marty Lab at the University of Austin (electrospray ionization native mass spectrometry).

Chapter 3: *De Novo* computational design and purification of two-component transmembrane nanoparticles

Adapted from

Zambrano D, Wargacki AJ, Humphrys D, Majumder S, Dowling QM, Carr KD, Weidle C, Borst AJ, King NP. *De novo* computational design and purification of two-component transmembrane nanoparticles. *Manuscript in preparation*.

Background

Enveloped protein nanoparticles (EPNs) are a class of hybrid nanomaterial composed from a single *de novo* designed protein subunit which assembles into a 60-mer icosahedral nanoparticle[57]. When transfected in mammalian cultures, the nanoparticles are enveloped by cell-derived membranes[57]. The hybrid lipid-protein nature of the EPNs is advantageous because it allows for the genetic modification of the protein scaffold *in silico*, thus allowing for the functionalization of EPN nanoparticles[57]. One limitation of the EPN platform is that extracellular vesicles made using EPNs are heterogeneous in size and cargo packaging in EPNs is still an active area of research[57]. The design of hybrid lipid-protein nanoparticles that can be formulated *in vitro* would be particularly useful due to the ability to modulate each building block with atomic-level accuracy, as well as control over the formulation of the lipids involved in nanoparticle preparation.

Two-component protein nanoparticles are built upon soluble protein building blocks and are constrained to using soluble symmetric subunits. As a result, membrane protein antigens used in nanoparticle designs first need to be stabilized and solubilized before they can be engineered for display on protein nanoparticles[15,18,19,58,59]. Recent advances in computational design have allowed us to design symmetric helical membrane proteins and beta barrels which can be adapted for specific functions[60–63]. These advancements in technology and design of membrane proteins using computational methods have opened the door to using more unique subunits as building blocks for designed nanoparticle complexes. Here we describe the design of a novel two-component octahedral nanoparticle using a *de novo* designed membrane protein as an underlying scaffold to the nanoparticle. This technology marks the first generation of transmembrane protein nanoparticles and is a step towards generating *in vitro* assembled EPN nanoparticles.

Results

In silico design of octahedral two-component membrane protein nanoparticles

One method to design protein nanoparticles is to dock cyclic oligomers into higher order structures and maximize positive contacts between the two subunits found at the interface. Here, the two components comprising the design of this nanoparticle were a *de novo* designed tetrameric membrane protein (PDB ID: 6B85) and a trimer of the aldolase 2-keto-3-deoxy-6-phosphogluconate (PDB ID: 1WA3), which has been used and characterized in previously designed nanoparticles with icosahedral and tetrahedral geometries[8,12,57]. Translational and rotational degrees of freedom were sampled for the 6B85 tetramer (referred to as O43-A) and the 1WA3 trimer (referred to as O43-B) before docking the components in octahedral symmetry (**Figure 3A-C**). Following docking of the two components, we initiated the design of specific residues at the interface between the two protein components. A feature of the *de novo* designed O43-A component is the ability to truncate the protein (O43-A-trunc) (**Figure 3D-E**)[60]. As a result, the truncated particle (O43-sol) and the transmembrane nanoparticle (O43-TM) have the same protein-protein interface between the two protein component building blocks (**Figure 3F**). Thousands of designs were evaluated based on score metrics such as shape complementarity and the change in Gibbs free energy (ddG) of the mutated interface residues. Designs were visually filtered to remove docks in which the transmembrane domain was not solvent-accessible. Ten of the top scoring designs from our filtering parameters were ordered for *in vitro* testing.

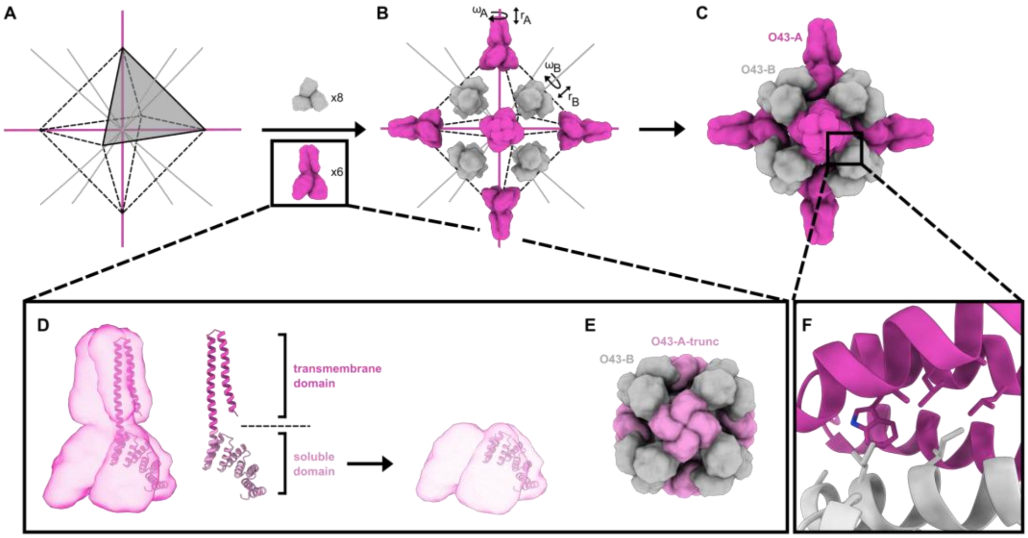


Figure 3: Two-component design for soluble and transmembrane octahedral protein nanoparticles.

(A) The O43 architecture comprises eight copies of the trimeric component and six copies of the tetrameric component (grey and pink) arranged in octahedral point group symmetry. The nanoparticle is composed of 48 total subunits. (B) Each oligomeric building block has two rigid body degrees of freedom, one translational (r) and one rotational (ω), that are systematically explored during docking. (C) Docking the components identifies interfaces with a high frequency of contacting residues. (D) The tetrameric component, O43-A, is a *de novo* designed transmembrane protein with two distinct domains: the soluble and transmembrane domains. O43-A can be truncated (O43-A-trunc) and the resulting protein-protein interface of the soluble nanoparticle, O43-sol, is identical to the interface of O43-TM. (E) The computational model of the O43-sol nanoparticle as a result of truncating O43-A. (F) Residues found in the interface of O43-A and O43-B are scored and manually modified to increase favorable interactions.

Characterization of designed O43-sol nanoparticles

The assembly interfaces between the two protein subunits are identical for the O43-sol and O43-TM nanoparticles, thus we proceeded with the expression and characterization of O43-sol to assess nanoparticle assembly independent of the transmembrane domain. Proteins for O43-A-trunc and O43-B were bicistronically expressed in *E. coli* and purified using histidine affinity tags on nickel immobilized metal affinity chromatography (IMAC) columns. Both components eluted from IMAC as shown by two bands represented in the Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) of the IMAC eluate fraction measuring approximately 15 kDa and 24 kDa (**Figure 4A**). These molecular weights were consistent with the masses of O43-A-trunc and O43-B respectively (**Figure 4A**). Further purification of the nanoparticle was accomplished by performing size exclusion chromatography (SEC) to isolate a single monodisperse peak at the predicted size. O43-sol nanoparticles eluted from SEC around 14.1 mL, earlier than a 660 kDa size standard(**Figure 4B**). Dynamic light scattering (DLS) and negative stain electron microscopy (ns-EM) were used to further validate the size and assembly of the nanoparticle isolated from the SEC peak fraction. The theoretical *in silico* diameter of O43-sol nanoparticles is approximately 19.7 nm, while the measured diameter found by DLS was 20.8 nm (**Figure 4C**). Monodisperse O43-sol nanoparticles were observed by ns-EM and processed into 2D and 3D reconstructions resolved at 21 Å (**Figure 4D-F**). Designs which did not express or assemble into nanoparticles during the solubility screen were considered viable candidates for addition of the transmembrane domain.

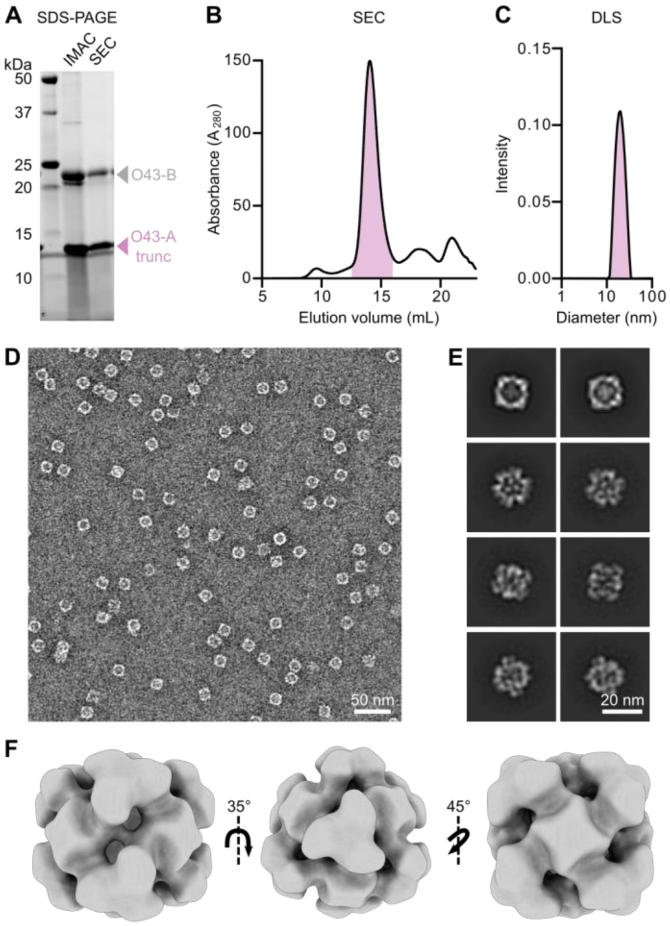


Figure 4: Purification and characterization of O43-sol nanoparticles by ns-EM.

(A) SDS-PAGE of IMAC and SEC eluate for O43-sol nanoparticles. Two bands labeled at 14 kDa and 24 kDa correspond to the O43-A-trunc (pink) and O43-B (grey) protein components respectively. The full annotated and unannotated SDS-PAGE can be found in **Supplemental Figure 4**. (B) SEC trace of protein absorbance at A_{280} as a function of elution volume with the nanoparticle highlighted in pink eluting at 14.1 mL. (C) DLS trace of O43-sol nanoparticles purified by SEC report a particle diameter of 20.8 nm. (D) ns-EM raw micrograph of the SEC peak fraction for O43-sol taken at 57,000X magnification. The scale bar represents 50 nm. (E) ns-EM 2D class averages and (F) 3D volume maps along the 2-, 3-, and 4-fold axes of symmetry for the O43-sol nanoparticle.

Purification and characterization of O43-TM nanoparticles

Following a revision of our purification protocols for the separate protein components for O43-TM, improved characterization of the nanoparticle could commence. Full length O43-A and O43-B were independently expressed and purified by IMAC and SDS-PAGE (**Figure 5A, Supplemental Figure 5A-C**). After mixing the two protein components together to allow for nanoparticle self-assembly, O43-TM was purified using SEC in a buffer containing 0.75% CHAPS detergent. A 2-fold improvement in protein purification of O43-TM was achieved when 100 mM glycine and 100 mM arginine were used as excipients in all buffers throughout protein purification[64]. The nanoparticle peak for O43-TM eluted from SEC around 13.2 mL, implying that the O43-TM nanoparticle with the addition of the transmembrane domain was larger than O43-sol nanoparticles (**Figure 5B**). The theoretical *in silico* diameter of O43-TM nanoparticles was 29.2 nm while the measured diameter found by DLS was 28.4 nm (**Figure 5C**). The SEC buffer containing CHAPS was run as a control, and we observed that the O43-TM sample contained detergent micelles measuring around 5.63 nm as well as a protein aggregation peak around 115 nm in diameter (**Figure 5C**). The SEC peak fraction was imaged by ns-EM where we identified nanoparticles with distinct spikes protruding from the underlying scaffold which resembled the design model (**Figure 5D-E, Supplemental Figure 5D-F**).

Encouraged by these results, O43-TM particles purified by SEC were imaged using Cryo-EM for higher resolution characterization. n-Dodecyl-β-D-Maltoside (DDM) detergent was used to prevent aggregation ofO43-TM nanoparticles for purification and Cryo-EM (**Figure 5F, Supplemental Figure A-H**). Transmembrane domain density was clearly observed in raw Cryo-EM movies as well as 2D class averages (**Figure 5G-H**). The unsharpened volume map revealed micelle density around the transmembrane domains of O43-TM while the sharpened density revealed clearly defined secondary structural elements of the transmembrane helices (**Figure 3H-J, Supplemental Figure 6I**). Subsequent 3D refinement yielded a final map with a global resolution of 4.16 Å, with local resolution ranging from 2.50 Å to 5.0 Å and an RMS deviation from the design model of 4.018 Å (**Supplemental Figure 6J**). Local resolution refinements for each component were also performed and further resolved by masking each nanoparticle component and applying symmetry expansion[56]. In the local refinement maps, the O43-A tetramer was resolved to a local resolution of 4.03 Å with Cα RMSD of 1.891 Å of the tetramer as viewed along the profile and a Cα RMSD of 1.495 Å for the monomer (**Figure 5K, Supplemental Figure 6K**). The O43-B trimer viewed along the C3 axis had a Cα RMSD of 1.036 Å (**Figure 5K, Supplemental Figure 6K**). These structural results, coupled with biophysical data indicate that the successful *in silico* design of O43-TM nanoparticles.

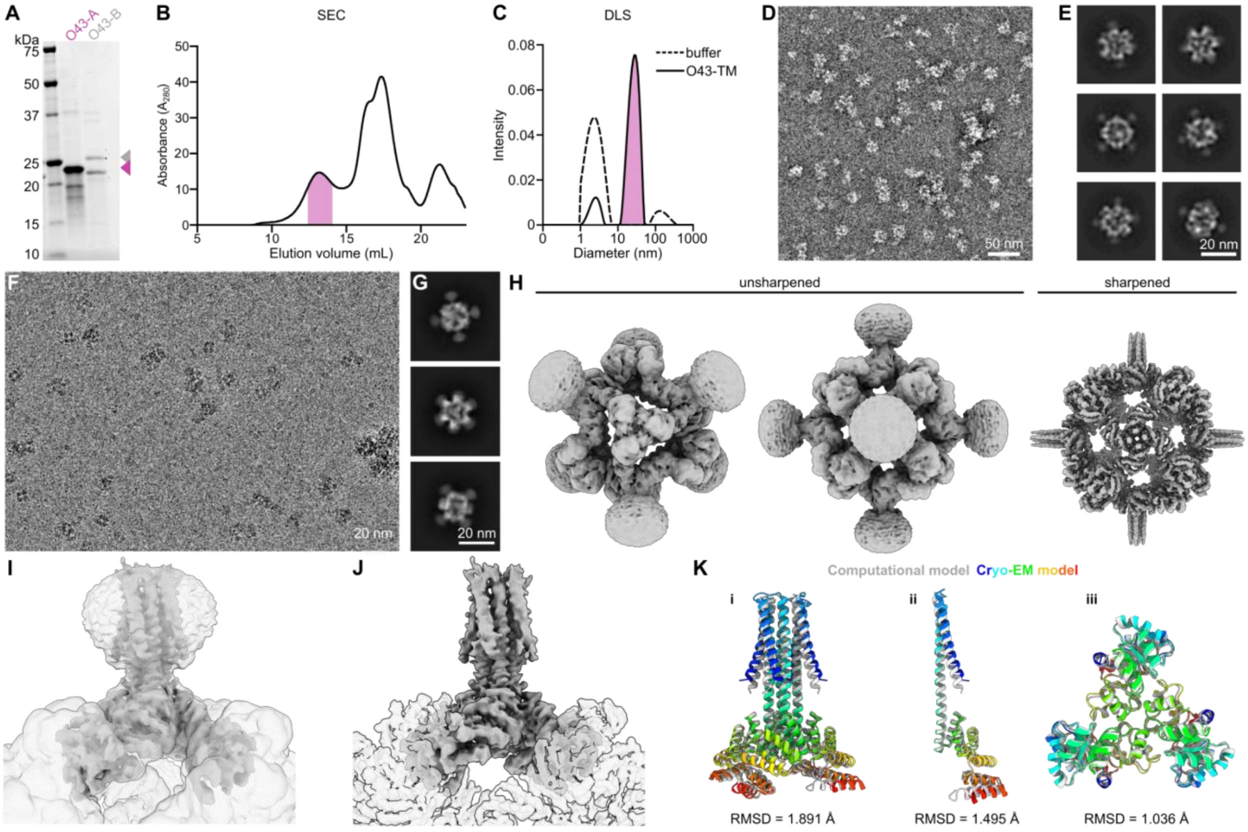


Figure 5: Purification and characterization of O43-TM nanoparticles by ns-EM and Cryo-EM.

(A) SDS-PAGE of IMAC eluate for O43-TM nanoparticle components. Fractions correspond to the O43-A (pink) and O43-B (grey) proteins respectively. The full annotated and unannotated SDS-PAGE can be found in **Supplemental Figure 5A-C.** (B) SEC trace of protein absorbance at A_{280} as a function of elution volume with the nanoparticle highlighted in pink eluting at 13.2 mL for O43-TM nanoparticles. (C) DLS trace for

O43-TM nanoparticles purified by SEC with the nanoparticle reporting a diameter of 28.4 nm. O43-TM samples contain detergent micelles as demonstrated using a buffer control containing 0.75% CHAPS. (D) ns-EM raw micrograph of the SEC peak fraction for O43-TM. The scale bar represents 50 nm. (E) ns-EM micrographs for O43-TM nanoparticles were processed to generate 2D class averages and 3D reconstructions (**Supplemental Figure 5D-E**). (F) A raw representative Cryo-EM micrograph at 105,000x magnification with a scale bar representing 20 nm **(Supplemental Figure 6A-B**). (G) 2D class averages processed from Cryo-EM movies (**Supplemental Figure 6C-H**). (H) A comparison of the unsharpened and sharpened global maps for O43-TM along the 4-fold axis of symmetry. In the unsharpened maps along the 3- and 4-fold axes of symmetry we observe detergent micelle density around the transmembrane domains (**Supplemental Figure 6I**) (I) Sharpened local refinement map of the O43-A transmembrane protein docked into the unsharpened global map. In the local refinement map, the TM helices are able to be resolved through the detergent micelle. (J) Sharpened local refinement map of O43-A docked into the sharpened global map. (K) Global resolution for O43-TM nanoparticles was 4.07 Å with local resolutions ranging from 2.5 Å to 5.0 Å (**Supplemental Figure 6J**). Local refinement models of the nanoparticle subunits where the computational design (grey) was aligned to the Cryo-EM model (rainbow). (i) The O43-A tetramer as viewed along the profile has a Cα RMSD of 1.891 Å. (ii) The O43-A monomer Cα RMSD of 1.495 Å. (iii) The O43-B trimer viewed along the C3 axis has a Cα RMSD of 1.036 Å. More RMSD values for each protein component viewed from different angles can be found in **Supplemental Figure 6K**.

Discussion

Here we designed two-component nanoparticles using a transmembrane protein as part of the nanoparticle scaffold. This work establishes a foundation for the purification and characterization of transmembrane nanoparticles, and illustrates the feasibility of using transmembrane domains for nanoparticle design. This study presents, to our knowledge, one of the first examples of a transmembrane two-component protein nanoparticle utilizing *de novo* protein design.

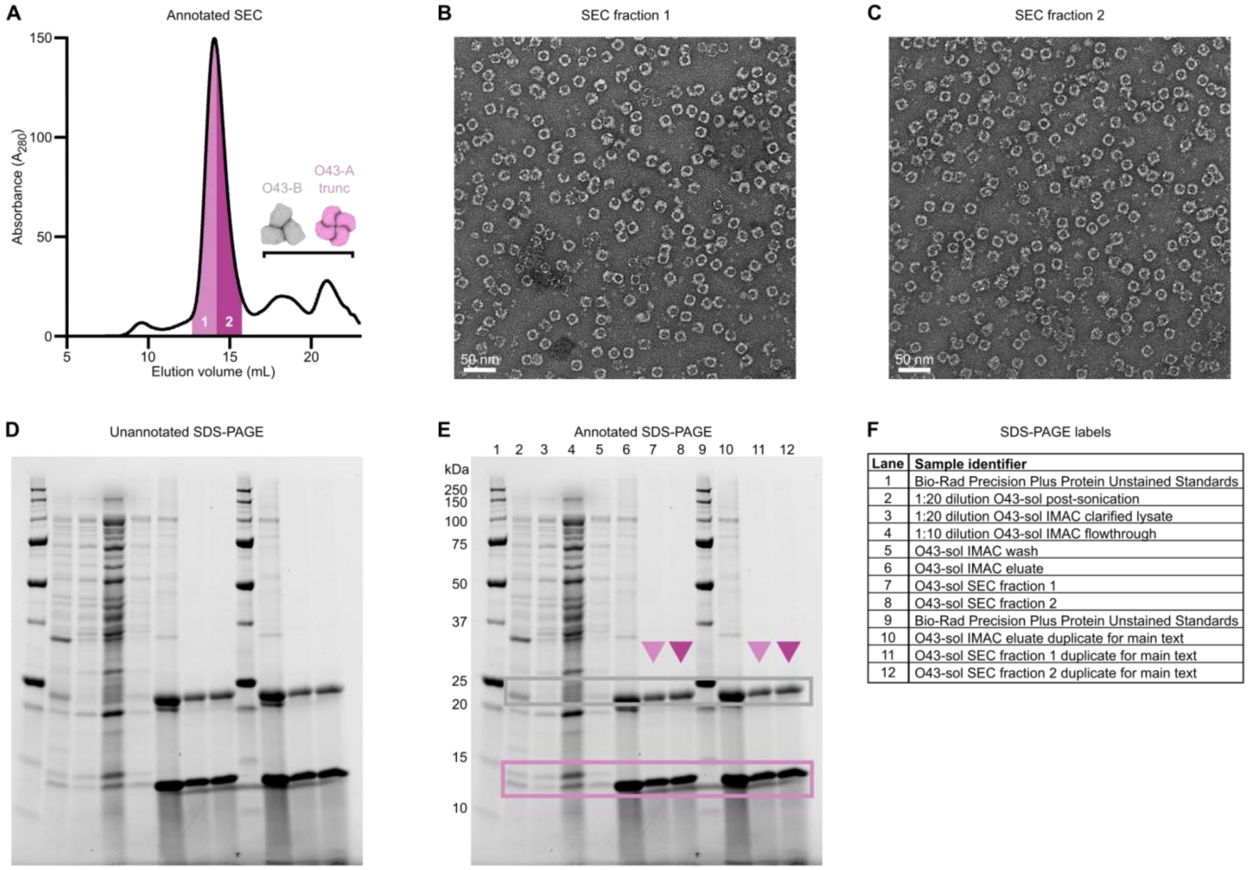
The goal of utilizing transmembrane scaffolds in nanoparticle design was to leverage the hydrophobic property of transmembrane domains in order to encapsulate the nanoparticle within a lipid bilayer (**Supplemental Figure 7**). Purifying a nanoparticle containing 24 copies of a multipass transmembrane protein was no trivial feat. After screening O43-sol nanoparticles we were very optimistic about the purification of O43-TM, however we noticed that our initial purification protocol contained an abundance of a contaminant protein[64]. Identifying the contaminant protein and finding the optimal mixture of excipients to use for improving our purification was the first hurdle to working with these transmembrane-displaying nanoparticles[64]. Following the identification of the contaminant, we were able to better isolate O43-TM and characterize the protein structure to 4.16 Å, a surprisingly high resolution for a nanoparticle containing 6 solvent-exposed transmembrane domains.

This work represents a step towards generating true hybrid protein-lipid nanoparticles. Our results establish the first *de novo* designed nanoparticle using a *de novo* transmembrane protein as a key scaffold forming the nanoparticle and demonstrates the potential of using computationally designed transmembrane proteins to design larger oligomers to be used for broader applications.

Acknowledgements

Thank you to Quinton Dowling who proposed this project to me during a coffee break and helped me to get started with the docking and design of the O43-TM nanoparticles. I would like to thank Adam Wargacki, Masaharu Somiya and Dan Humphreys for their excellent advice during difficult times of protein purification and methods of encapsulation. The Cryo-EM work here would not have been possible without my collaboration with the Institute for Protein Design Electron Microscopy Research Core, specifically Kenneth Carr, Connor Weidle and Andrew Borst.

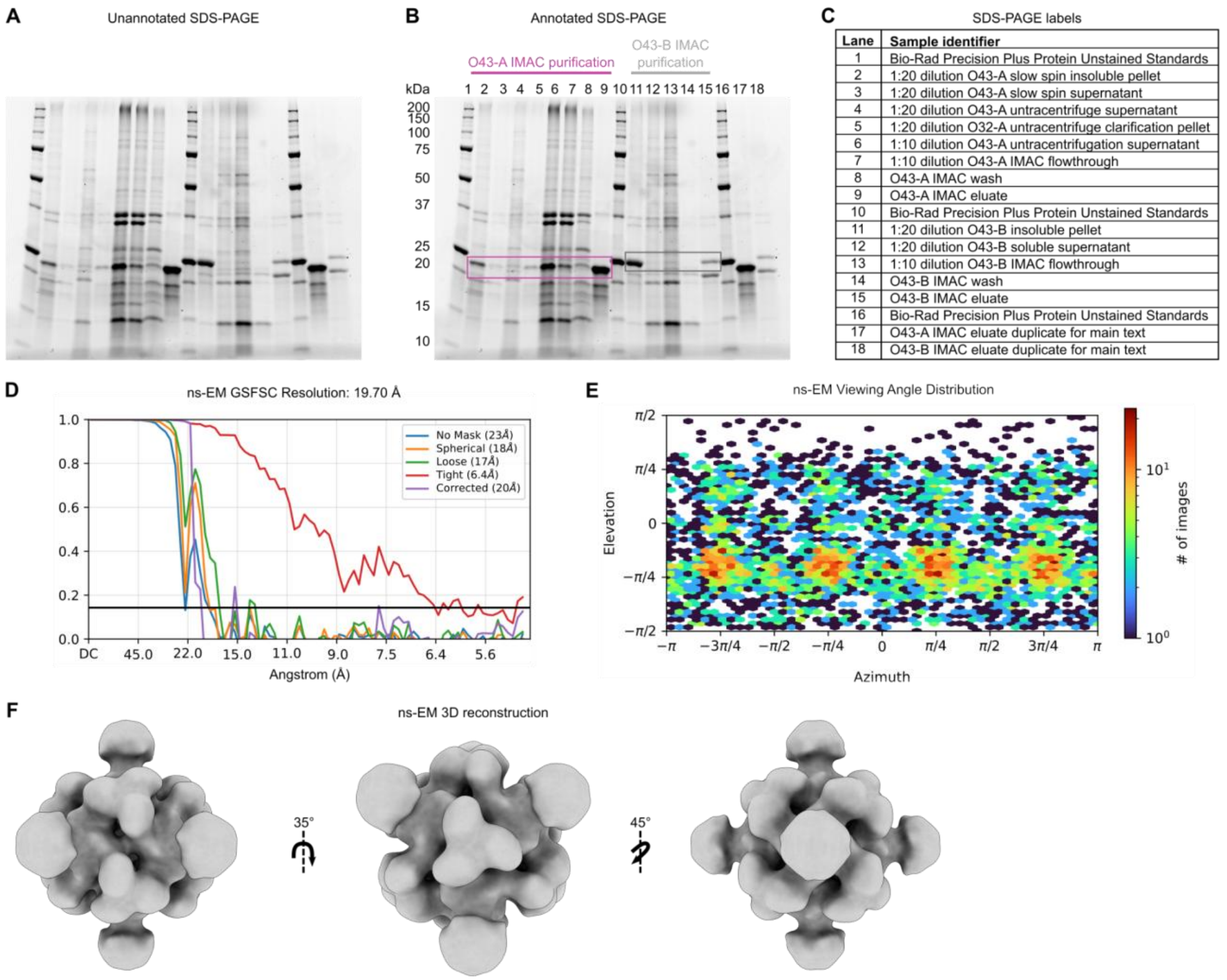
Supplemental Data



Lane	Sample identifier
1	Bio-Rad Precision Plus Protein Unstained Standards
2	1:20 dilution O43-sol post-sonication
3	1:20 dilution O43-sol IMAC clarified lysate
4	1:10 dilution O43-sol IMAC flowthrough
5	O43-sol IMAC wash
6	O43-sol IMAC eluate
7	O43-sol SEC fraction 1
8	O43-sol SEC fraction 2
9	Bio-Rad Precision Plus Protein Unstained Standards
10	O43-sol IMAC eluate duplicate for main text
11	O43-sol SEC fraction 1 duplicate for main text
12	O43-sol SEC fraction 2 duplicate for main text

Supplemental Figure 4: SDS-PAGE of bicistronic O43-sol IMAC purification

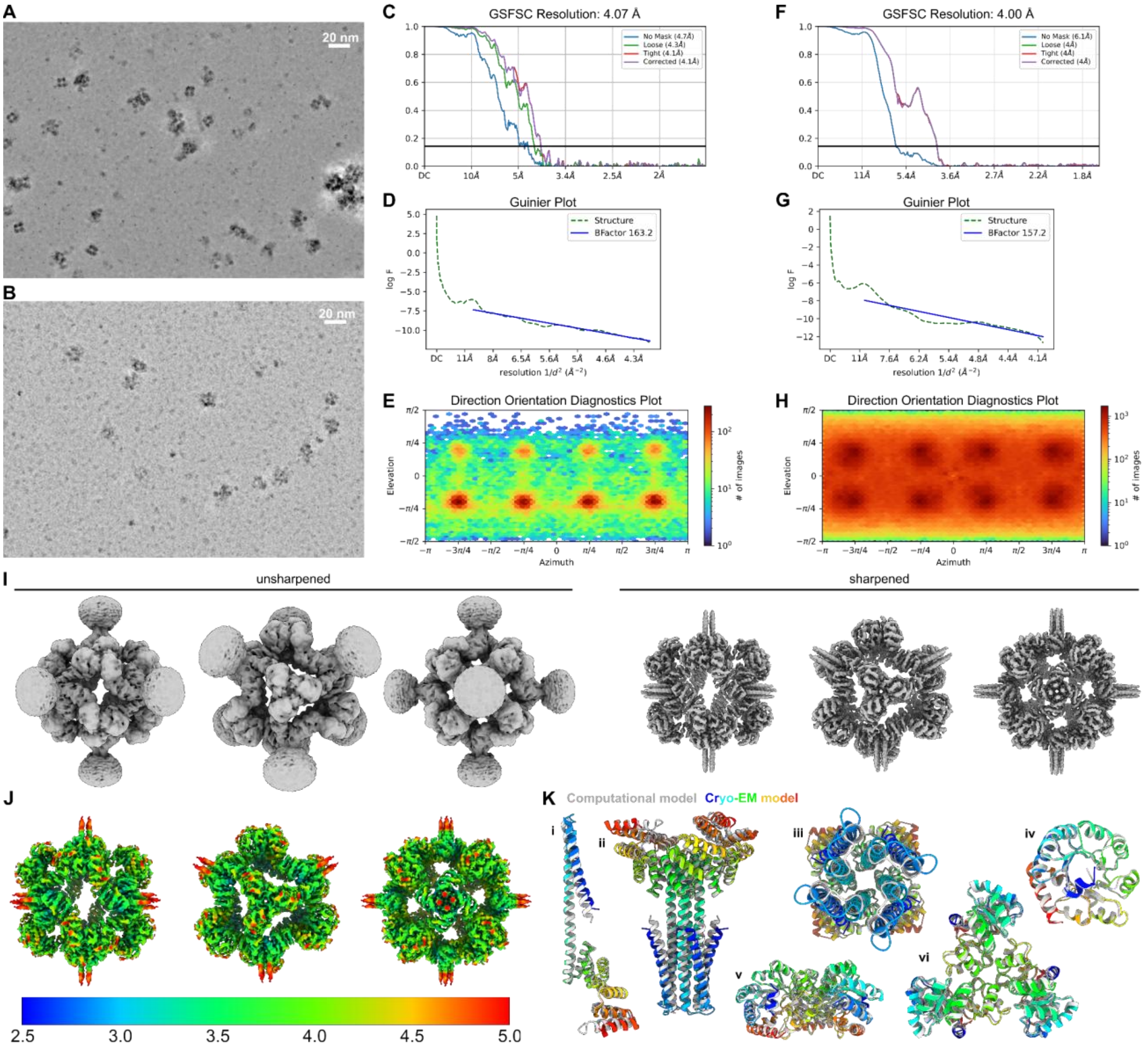
(A) Annotated SEC trace for the purification of O43-sol nanoparticles with the nanoparticle peak labeled as 2 different colors corresponding to separate SEC fractions. (B) ns-EM raw micrograph of SEC fraction 1 from the trace in (A) representing O43-sol nanoparticles. The white scale bar is 50 nm. (C) ns-EM raw micrograph of SEC fraction 2 from the trace in (A) representing O43-sol nanoparticles. The white scale bar is 50 nm. (D) Unannotated and (E) Annotated SDS-PAGE for O43-sol showing samples during different stages of IMAC and SEC purification. (F) Lane and Sample identifiers for the annotated SDS-PAGE in (E).



Lane	Sample identifier
1	Bio-Rad Precision Plus Protein Unstained Standards
2	1:20 dilution O43-A slow spin insoluble pellet
3	1:20 dilution O43-A slow spin supernatant
4	1:20 dilution O43-A untracentrifuge supernatant
5	1:20 dilution O32-A untracentrifuge clarification pellet
6	1:10 dilution O43-A untracentrifugation supernatant
7	1:10 dilution O43-A IMAC flowthrough
8	O43-A IMAC wash
9	O43-A IMAC eluate
10	Bio-Rad Precision Plus Protein Unstained Standards
11	1:20 dilution O43-B insoluble pellet
12	1:20 dilution O43-B soluble supernatant
13	1:10 dilution O43-B IMAC flowthrough
14	O43-B IMAC wash
15	O43-B IMAC eluate
16	Bio-Rad Precision Plus Protein Unstained Standards
17	O43-A IMAC eluate duplicate for main text
18	O43-B IMAC eluate duplicate for main text

Supplemental Figure 5: SDS-PAGE and ns-EM raw data for O43-TM purification and characterization.

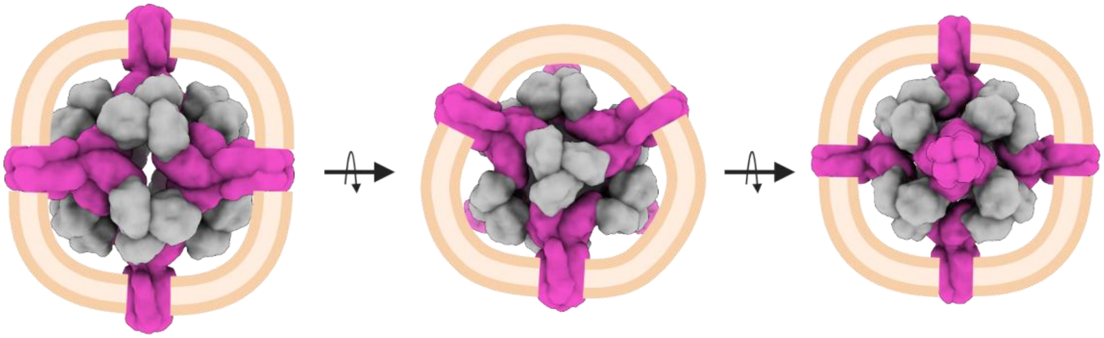
(A) Unannotated and (B) Annotated SDS-PAGE of the purification for O43-A and O43-B by IMAC. (C) Lane and Sample identifiers for the annotated SDS-PAGE in (B). (D) Fourier shell correlation (FSC) plot of the ns-EM 3D volume map. (E) View angle distribution plot of the ns-EM 3D volume map. (F) Low resolution ns-EM volume map of O43-TM nanoparticles purified by SEC in CHAPS detergent along the 2-, 3- and 4-fold axes of symmetry.



Supplemental Figure 6: Cryo-EM data processing for O43-TM

(A) Representative denoised Cryo-EM micrograph for O43-TM Dataset 1. (B) Representative denoised Cryo-EM micrograph for O43-TM Dataset 2 following improved purification (**Supplemental Table 1**) [64]. Final global refinement map statistics including (C) the Gold Standard Fourier Shell Correlation (GSFSC) plot showing an estimated resolution of 4.07 Å at a Fourier Shell Correlation (FSC) of 0.143 (D) Guinier Plot and (E) Direction Orientation Diagnostics. Final local refinement map statistics including (F) the GSFSC plot showing an estimated resolution of 4.00 Å at FSC 0.143, (G) Guinier Plot and (H) Direction Orientation

Diagnostics plot. (I) A comparison of the final unsharpened and sharpened global volume maps viewed along the 2-, 3- and 4-fold axes of symmetry. (J) Local resolution estimations of the sharpened global refinement map along the 2-, 3- and 4-fold axes of symmetry. The color bar shows regions with resolutions ranging from 2.5 Å to 5.0 Å. (K) Local refinement models of the nanoparticle subunits where the computational design (grey) is aligned to the Cryo-EM model (rainbow). Monomeric O43-A (i) has a Cα RMSD of 1.495 Å. Tetrameric O43-A as viewed along the profile (ii) and C4 axis (iii) has a Cα RMSD of 1.891 Å. Monomeric O43-B (iv) has a Cα RMSD of 0.764 Å. Trimeric O43-B as viewed along the profile (v) and C3 axis (vi) has a Cα RMSD of 1.036 Å. Local maps shown were produced in ChimeraX and have blobs smaller than size 10 hidden.



Supplemental Figure 7: Model of the encapsulated transmembrane nanoparticle

An illustrative model of the expected outcome for encapsulating a transmembrane nanoparticle. A lipid bilayer surrounds the protein nanoparticle around the transmembrane domains of O43-A. This representative model was made using BioRender[65].

Supplemental Table 1: O43-TM Cryo-EM Data Collection Statistics

Data Collection	**Dataset 1**	**Dataset 2**
Microscope	Titan Krios (FEI)	Titan Krios (FEI)
Voltage (kV)	300	300
Detector	K3 Gatan	K3 Gatan
Energy Filter	Gatan BioQuantum Gif	Gatan BioQuantum Gif
Magnification	105,000 X	105,000 X
Movie micrograph pixel size (Å)	0.843	0.843
Dose rate (e^-/Å^2/s)	8.4	11.8
No. of frames per Stack	99	79
Frame exposure time (s)	0.05	0.05
Movie micrograph exposure time (s)	5	5
Total dose (e^-/Å^2)	47.00	47.32
Defocus range (µm)	-0.8 to -1.8	-0.8 to -1.8
Number of movie micrographs	6,211	6,887
Global Cryo-EM Map		
Total Particles	42,340	
Symmetry	Octahedral	
Global Map Global Resolution Estimation (Å)	4.07	
Maximum Global Map Local Resolution (Å)	8.3	
Minimum Global Map Local Resolution (Å)	3.0	
Local Cryo-EM Map		
Total Particles	58,139	
Total Particle Poses	1,395,336	
Global Map Global Resolution Estimation (Å)	4.00	
Global Map Global Resolution Estimation (Å)	5.7	
Minimum Global Map Local Resolution (Å)	2.6	
Atomic Model Validation Metrics		
Bonds (RMSD)		
Bond Lengths (Å) (# > 4σ)	0.011	
Bond Angles (°) (# > 4σ)	1.766	
MolProbity Score	0.68	
Clash Score	0.21	
Ramachandran plot		
Outliers (%)	0.00	
Allowed (%)	2.46	
Favored (%)	97.54	
Rotamer outliers (%)	0.00	
Atomic Model		
Number of molecule A chains	24	
Residues per chain	210	
Number of molecule B chains	24	
Residues per chain	201	

Supplemental Sequence 1: O43-A-trunc sequence

ALAYVMLGLLLSLLNRLSLAAEAYKKAIELDPNDALAWLLLGSVLEKLKRLDEAAEAYKKAIELKPNDASAWKELGKVLEKLGRLDEAADAYAEALRLDPDDTEAWKELGKVLEKLGMLEAAAMAKKAAILLDPND*GS***EQKLISEEDL***GSGSGS*HHHHHH

MW: 17.4 kDa

Supplemental Sequence 2: O43-A sequence

MSKDTEDSRKIWRTIMLLLVFAILLSAIIWYQITTNPKTSQIATLLSMQLLLIALMLVVIALLLSRQTEQVAESIRRDVSALAYVMLGLLLSLLNRLSLAAEAYKKAIELDPNDALAWLLLGSVLEKLKRLDEAAEAYKKAIELKPNDASAWKELGKVLEKLGRLDEAADAYAEALRLDPDDTEAWKELGKVLEKLGMLEAAAMAKKAAILLDPND*GS***EQKLISEEDL***GSGSGS*HHHHHH

MW: 26.6 kDa

Supplemental Sequence 3: O43-B sequence

MEELFKKHKIVAVLRANSVEEAIEKAVAVFAGGVHLIEITFTVPDADTVIKALSVLKEKGAIIGAGTVTSVEQARKAVESGAEFIVSPHLDEEISQFAKEKGVFYMPGVMTPTELVKAMKLGHTILKLFPGEVVGPQFVKAMKGPFPNVKFVPTGGVDLDNVAEWFKAGVLAVGVGEALVDGEPRAVKIAAIAFRAVIKGATE*GS***EQKLISEEDL***GSGSGS*HHHHHH

MW: 24.1 kDa

The C-terminal of all sequences ends in a single GS linker (*italicised and underlined*) followed by a Myc-tag (**bold**). The Myc-tag and 6X histidine tag are separated by an addition of a 3xGS linker.

Materials and Methods

Computational design of O43 nanoparticles

Nanoparticle assemblies were generated by using 6B85 and 1WA3 cyclic oligomers as the starting subunits. The pair were aligned along the octahedral symmetry axes and allowed to rotate freely around the four-fold or three-fold symmetry axis before docking the subunits together. The nanoparticle is composed of a combination of six tetrameric building blocks and eight trimeric building blocks making up what we describe as the O43-TM nanoparticle. In total the O43 nanoparticles contained 48 subunits of proteins. Thousands of O43-TM docks were assessed and filtered by several parameters including: size of the interface between the subunits, the number of interacting residues, hydrophobicity of the interface, spacing of the backbone geometry at the interface, shape complementarity and the change in Gibbs free energy (ddG) of the mutated interface residues as previously described[8,11,12,15,18,66]. Designs where the transmembrane domain of O43-A was buried within the octahedral scaffold were disregarded. The top 10 nanoparticles after the application of these filtering parameters were chosen to be evaluated *in vitro*. For the top designs, C-terminal Myc-tag linked to 6x histidine-tagged plasmids were ordered from GenScript and transformed into *E. coli* competent cells.

Purification of bicistronic O43-sol nanoparticles

Plasmids ordered from GenScript for the tetrameric soluble protein linked to the trimeric soluble protein (O43-A-trunc and O43-B) were transformed into *E. coli* BL21(DE3) competent cells (NEB, Cat. C2527H) and plated on an agar plate treated with a final concentration of 50 µg/mL kanamycin. A single colony was picked from the transformed cells and grown in a 50 mL culture of LB broth (Lennox formulation) overnight at 37.0°C. 5 mL of the overnight culture was transferred into 500 mL of Terrific Broth (TB) media treated with a final concentration of 50 µg/mL kanamycin. Cells were grown in TB media at 37.0°C until an OD600 of 0.60 - 0.80 before induction with a final concentration of 1 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG) and grown for an additional 12 hours at 18.0°C. Cells were harvested at 5000 revolutions per minute (rpm) for 30 minutes and resuspended with a lysis buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M KCl, 100 mM arginine, 100 mM glycine, 1.00 mg/mL lysozyme and 1.00 mg/mL DNAse) then lysed using sonication. Cells were lysed at 4°C using a probe sonicator for a total of 5 minutes using 10.0 second sonication pulses and 10.0 second breaks. After cell lysis, the solution was clarified at 10,000*g* for 30 minutes. The supernatant was collected and ran through a nickel-nitrilotriacetic acid (Ni-NTA) IMAC resin with a bed volume of 5.00 mL (Qiagen, Cat. 30250). The IMAC resin was equilibrated with 5.00 column volumes of loading buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M KCl, 100 mM arginine and 100 mM glycine) before the addition of cell lysate, then washed with 3 column volumes of the loading buffer. The bound protein was eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM KCl, 0.50 M imidazole, 0.15% n-Dodecyl-B-D-maltoside (DDM), 100 mM arginine and 100 mM glycine).

Purification of transmembrane O43-A nanoparticle component

Plasmids ordered from GenScript for the tetrameric transmembrane protein were transformed into BL21 Star (DE3)pLysS One Shot competent cells (Promega, Cat. L1195) and plated on an agar plate containing of 50 µg/mL kanamycin. A single colony was picked from the transformed cells and grown in a 50 mL culture of LB broth (Lennox formulation) overnight at 37.0°C. 5 mL of the overnight culture was transferred into 500 mL of Terrific Broth (TB) media treated with a final concentration of 50 µg/mL kanamycin. Cells were grown in TB media at 37.0°C until an OD600 of 0.60 - 0.80 before induction with a final concentration of 1 mM Isopropyl β-d-1-thiogalactopyranoside (IPTG) and grown for an additional 12 hours at 18.0°C. Cells were harvested at 5000 revolutions per minute (rpm) for 30 minutes and resuspended with a lysis buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M KCl, 100 mM arginine, 100 mM glycine, 1.00 mg/mL lysozyme and 1.00 mg/mL DNAse) then lysed using sonication. Cells were lysed at 4°C using a probe sonicator for a total of 5 minutes using 10.0 second sonication pulses and 10.0 second breaks. Following lysis by sonication, cells were sedimented at 10,000*g* for 15 minutes and the supernatant was collected for ultracentrifugation and clarified at 130,000*g* for 1.00 hour at 4.0°C. Resuspension buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M KCl, 100 mM arginine, 100 mM glycine and 2.00% n-Dodecyl-β-D-maltoside (DDM)) was used to resuspend the pellet membrane fraction after ultracentrifugation. This membrane suspension was placed on a rocker at 4.0°C overnight. The following day, the solution was spun for an additional 30 minutes at 14,000*g* as the last clarification step. The supernatant was collected and ran through an Ni-NTA IMAC resin (Qiagen, Cat. 30250) with a bed volume of 5.00 mL. The IMAC resin was equilibrated with 5.00 column volumes of resuspension buffer before the addition of clarified membrane supernatant, then washed with 3 column volumes of the resuspension buffer then eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM KCl, 0.50 M imidazole, 0.15% DDM, 100 mM arginine and 100 mM glycine).

Purification of soluble O43-B nanoparticle component

Plasmids ordered from GenScript for the trimeric soluble protein were transformed into *E. coli* BL21(DE3) competent cells (NEB, Cat. C2527H) and plated on an agar plate treated with a final concentration of 50 µg/mL kanamycin. A single colony was picked from the transformed cells and grown in a 50 mL culture of LB broth (Lennox formulation) overnight at 37.0°C. 5 mL of the overnight culture was transferred into 500 mL of Terrific Broth (TB) media treated with a final concentration of 50 µg/mL kanamycin. Cells were grown in TB media at 37.0°C until an OD600 of 0.60 - 0.80 before induction with a final concentration of 1 mM Isopropyl β-d-1-thiogalactopyranoside (IPTG) and grown for an additional 3 hours at 37.0°C. Cells were harvested at 5000 revolutions per minute (rpm) for 30 minutes and resuspended with a lysis buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M NaCl, 100 mM arginine, 100 mM glycine, 1.00 mg/mL lysozyme and 1.00 mg/mL DNAse) then lysed using sonication. Cells were lysed at 4°C using a probe sonicator for a total of 5 minutes using 10.0 second sonication pulses and 10.0 second breaks. After cell lysis, the solution was clarified at 10,000*g* for 30 minutes. The supernatant was collected and passed over a nickel-nitrilotriacetic acid (Ni-NTA) IMAC resin with a bed volume of 5.00 mL (Qiagen, Cat. 30250). The IMAC resin was equilibrated with 5.00 column volumes of loading buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M NaCl, 100 mM arginine and 100 mM glycine) before the addition of cell lysate, then washed with 3 column volumes of the loading buffer. The bound protein was eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 150 mM KCl, 0.50 M imidazole, 0.15% n-Dodecyl-B-D-maltoside (DDM), 100 mM arginine and 100 mM glycine).

Purification and characterization of O43 assembled nanoparticles

Purified proteins for O43-A were mixed together with O43-B at a 1:1.5 (c/c) by molar ratio for 1 hour at 4.0°C on a rocker to facilitate nanoparticle assembly. Purification of assembled nanoparticles was achieved using size exclusion chromatography (SEC) in an SEC buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM NaCl, 0.75% 3-((3-cholamidopropyl) dimethylammonio)-1-propanesulfonate (CHAPS), 100 mM arginine and 100 mM glycine) on a Cytiva Superose® 6 Increase 10/300 GL column. The SEC column was calibrated using the Cytiva Gel Filtration HMW Calibration Kit (Cat. 28403842) with a resolution limit of molecular masses ranging from 43 to 669 kilodaltons (kDa). The nanoparticle peak eluting from SEC was isolated and characterized by DLS, ns-EM and Cryo-EM.

ns-EM sample preparation for O43-sol

After SEC, the isolated nanoparticle peak was diluted with the SEC buffer from 0.31 mg/mL to 0.01 mg/mL. 3 µL of the diluted sample was applied to a freshly glow discharged 400 carbon square mesh grid (Electron Microscopy Sciences, EMS) and allowed to adsorb onto the grid for 30 seconds. Blotting paper (Wattman) was used to remove excess solution from the grid before the application of 3 µL 2% uranyl formate (2UF) to stain the grid for ns-EM. Excess stain solution was blotted immediately followed by another application of 3 µL 2UF. This was repeated once more, resulting in a total of 3 applications of 2UF. Following the final application of 2UF, all remaining solution was blotted and the sample was allowed to dry completely before imaging. There was no waiting period in between each application of 2UF, and after each blot new 2UF was applied to the grid immediately.

ns-EM data acquisition and processing for O43-sol

Negative stain electron microscopy (ns-EM) micrographs were collected on a Thermo Fisher Scientific Talos L120C 120 kV transmission electron microscope with a LaB_6 filament and CETA camera with a pixel size of 2.49 Å at 57,000x magnification. 173 micrographs were recorded at 57,000x magnification at a total dose rate 36 $e^-/Å^2$ for the octahedral nanoparticle. Micrographs were imported into CryoSPARC v4.7.0 [56] and Patch CTF Estimation was performed prior to blob picking, inspection, and extraction of 33,983 picks with a box size of 180 pixels (448.2 Å). With CTF correction off, picks were classified into Reviewed 150 2D classes, 19 of which were selected to serve as templates for an additional round of particle picking. In this second round of particle picking, picks were inspected and extracted to 180 pixels (448.2 Å) before classification of 42,337 extracted particles into 100 2D classes with CTF correction off. 5 out of 150 of these 2D classes were selected as a second round of templates for another round of particle picking and extraction. In this last round, 35,323 particles were extracted to 180 pixels (448.2 Å) using the templates from the second round of extraction and classified into 50 2D classes with CTF correction off and maximum alignment resolution 6 Å. 4 classes with 6,415 particles were used to generate three 3D octahedral ab-initio reconstruction classes. The best 3D class containing 5,275 particles was used for homogeneous refinement to generate a volume map using Octahedral symmetry with a maximum resolution threshold of 20 Å.

ns-EM sample preparation for O43-TM

After SEC, the isolated nanoparticle peak was diluted with the SEC buffer from 0.034 mg/mL to 0.01 mg/mL. 3 µL of the diluted sample was applied to a freshly glow discharged 400 carbon square mesh grid (Electron Microscopy Sciences, EMS) and allowed to adsorb onto the grid for 30 seconds. Blotting paper (Wattman) was used to remove excess solution from the grid before the application of 3 µL 2% uranyl formate (2UF) to stain the grid for ns-EM. Excess stain solution was blotted immediately followed by another application of 3 µL 2UF. This was repeated once more, resulting in a total of 3 applications of 2UF. Following the final application of 2UF, all remaining solution was blotted and the sample was allowed to dry completely before imaging. There was no waiting period in between each application of 2UF, and after each blot new 2UF was applied to the grid immediately.

ns-EM data acquisition and processing for O43-TM

ns-EM micrographs were collected on a Thermo Fisher Scientific Talos L120C 120 kV transmission electron microscope with a LaB_6 filament and CETA camera with a pixel size of 2.49 Å at 57,000x magnification. 143 micrographs were recorded at 57,000x magnification at a total dose rate 24.1 e-/Å^2 for the octahedral nanoparticle. Micrographs were imported into CryoSPARC v4.7.0 21 and Patch CTF Estimation was performed prior to blob picking, inspection, and extraction of 26,831 picks with a box size of 180 pixels (448.2 Å). Picks were classified into 50 2D classes, 7 of which were selected to serve as templates for an additional round of particle picking. In this second round of particle picking, picks were inspected and extracted to 180 pixels (448.2 Å) before classification of 12,953 extracted particles into 100 2D classes with CTF correction off. 12 out of 150 of these 2D classes were selected as a second round of templates for another round of particle picking and extraction. In this last round, 11,858 particles were extracted to 180 pixels (448.2 Å) using the templates from the second round of extraction and classified into 50 2D classes with CTF correction off. 5,118 particles were used to generate three 3D ab-initio reconstruction classes in C1. The best 3D class contained 5,104 particles which were homogeneously refined into a volume map using Octahedral symmetry with a maximum resolution threshold of 20 Å.

Cryo-EM sample preparation for O43-TM

O43-TM nanoparticles for the Cryo-EM datasets were purified in an SEC buffer containing 0.15% DDM instead of CHAPS detergent. For the first dataset, 3 µL of protein sample, at an estimated final concentration of 0.3 mg/mL, was applied to a glow-discharged R 2/2 Holey Carbon grid overlaid with a 2 nm layer of ultrathin continuous carbon (Quantifoil). Vitrification via plunge freezing into liquid ethane was performed using a Vitrobot Mark IV (Thermo Scientific) with a chamber temperature of 22°C, a chamber humidity of 100%, a wait time of 7.5 seconds, a blot time of 0.5 seconds, and a blot force of 0. The grid was clipped following standard protocols before being loaded onto a 300 kV Titan Krios (FEI) for imaging. For the second dataset, the same protocol was performed using the improved purity sample which had an estimated final concentration of 0.15 mg/mL.

Cryo-EM data acquisition and processing for O43-TM

For the first dataset, a total of 6,211 movies were collected using SerialEM (Mastronarde, 2003, Mastronarde, 2005) on a 300 kV Titan Krios (FEI) equipped with a K3 direct electron detector (Gatan) and a BioQuantum Gif energy filter (Gatan). The movies were recorded at 105,000x magnification, with a pixel (px) size of 0.843 Å/px. Each movie consisted of 99 frames recorded at a frame rate of 19.8 frames per second, with a dose rate of 9.4 e-/Å^2/s and an exposure time of 5 seconds, resulting in a total exposure dose of 47.0 e-/Å^2. For the second dataset, 6,887 additional movies were collected using SerialEM (Mastronarde, 2003, Mastronarde, 2005) on the same microscope with the same detector, energy filter, pixel size and magnification as before. Each movie consisted of 79 frames recorded at a frame rate of 19.8 frames per second, with a dose rate of 11.8 e-/Å^2/s and an exposure time of 4 seconds, resulting in a total exposure dose of 47.32 e-/Å^2.

Cryo-EM data processing for O43-TM

Two data sets were used to process data for O43-TM nanoparticles. For global refinement, the movies from the first dataset were imported into CryoSPARC v4.4 for data processing[56]. Initially, these movies were processed using the Patch Motion Correction job. Defocus and astigmatism values were estimated using the Patch CTF Estimation job with the default parameters. After two rounds of exposure curation, 4,264 exposures were selected for particle picking. Prospective particles were identified using blob picking with a diameter range of 200 to 400 Å. This process identified 808,774 prospective particles, which were then inspected and curated based on power histogram values and normalized cross-correlation scores. Following this curation, 456,319 particles were extracted at a box size of 600 pixels. These particles were classified into 150 classes with a batch size of 400 per class. 4 classes were selected to serve as templates for another round of particle picking with particle diameter set to 300 Å. This round yielded 1,468,115 prospective particles, which were curated down to 765,699 particles. These were extracted at 600 pixels and classified into 150 classes again. All high-quality particle classes (47,980 particles) were re-extracted with recentering enabled, resulting in 47,480 particles, which were then split into 150 classes, yielding 39,222 high-quality particles. Cryo-EM data processing of O43-TM particles follows protocols as previously described[64]. Of these high-quality particles, the 20,189 contained in 39 2D classes of the designed nanoparticle. These particles were combined with the particles isolated from the second dataset.

Movies from the second dataset were also imported into CryoSPARC v4.4 for data processing[56]. As with the first dataset, these movies were also processed using the Patch Motion Correction and Patch CTF Estimation jobs with default parameters. Exposures were curated twice, to yield 4,491 movies for particle particle picking. To ensure noise was not being selected, blob picking was once again used to identify candidate classes to use as templates during another round of particle picking. 22,151 particles composing 2D class averages of the designed protein were selected for combination with the first dataset to generate a total of 42,340 combined particles.

Particles in both datasets were used as inputs for a 3 class asymmetric ab-initio refinement with a maximum resolution of 20 Å. As the density could be observed to be octahedral and no symmetry breaking features were observed, future refinements were done with enforced octahedral symmetry. Firstly, a non-uniform refinement was run on the best output, using all 42,340 particles, which yielded a global Gold Standard Fourier Shell Correlation (GSFSC) resolution estimate of 4.15 Å. A subsequent non-uniform refinement with “Minimize over per-particle scale” and “Optimize per-group CTF params” set to True enabled us to derive a final volume with a global GSFSC resolution estimation of 4.07 Å.

Our best local refinement was derived following a re-processing of both datasets in parallel. First, each dataset was independently pre-processed using Import Movies, Patch Motion Correction, and Patch CTF estimation jobs identically to the way they were initially processed. Automated Curate Exposures jobs systemically rejected all movies with CTF fit resolutions below 6 Å. Each dataset was denoised using the Micrograph Denoiser job with default settings, before exposures were manually curated, reducing the first dataset to 4,630 movies and the second dataset to 5,382 movies. At this point, both datasets were combined for Blob Picking on all 10,012 movies with the input diameter narrowed to between 260 Å and 320 Å. 1,199,408 prospective particles were picked, which following manual inspection was reduced to 387,027. Prospective particles were then extracted to 640 pixels and down sampled to 160 pixels. Extracted regions were classified into 50 classes using default parameters, and 11 classes representing 64,362 particles were selected for further processing. These particles were sorted into 3 asymmetric ab-initio class averages, which were then re-sorted using Heterogeneous Refinement with “Force hard classification” set to True. These volumes were used as the input volumes for another round of asymmetric 3-volume Heterogeneous Refinement using a different particle set. That set was derived by using the best 4 classes of the 11 selected classes as templates for Template Picking with a diameter of 300 Å. Approved inspected picks were extracted to 640 pixels without down sampling.

The 58,139 particles that were sorted into the highest resolution volume were homogeneously refined using that volume and default parameters. This yielded an asymmetric volume with a global GSFSC of 7.37 Å, which was then non-uniformly refined using default parameters to a volume of 6.05 Å using the same particles. The handed-ness of the volume was observed to be inverted, so a Volume Tools job was used to z-flip the map, then another Non-uniform Refinement was run with “Minimize over per-particle scale” set to true, which yielded a global GSFSC of 5.97 Å. Once again, no symmetry breaking features were observed, so the volume and particles were once again refined using Non-uniform Refinement and Minimize over per-particle scale, however this time octahedral symmetry was enforced and a global GSFSC of 4.31 Å was derived. To prepare for symmetry expansion, a Volume Align Tools job was run with octahedral symmetry and symmetry alignment set to true to ensure the volume was perfectly centered. Then, the particles were symmetry expanded with octahedral symmetry selected yielding an expanded particle count of 1,395,336 particle-permutations. A mask surrounding one TM-domain containing tetramer was generated and imported as a mask and processed to have a 6-pixel dilation radius and a 24-pixel soft padding. A Volume Alignment Tools job was used to center the data to the mask’s center of mass, and then a Local Refinement was run which output an estimated GSFSC of 4.07 Å, but significantly improved map quality of the TM-domain containing component.

Cryo-EM model building for O43-TM

First, the local refinement map was fit to the density of the global refinement, and resampled to the grid of the global refinement. Then, the design model was relaxed into the best global volume map using ISOLDE in ChimeraX with the temperature of the simulation set to 1 K[67–69]. Secondary structure was repositioned as needed to eliminate most deviations between the secondary structure of the model and the secondary structure that could be observed in the global refinement map. Once the volume was observed to be adequately in density, manual refinement began. To refine the soluble component, one subunit was chosen and manually fit to the global density in Coot[70]. The O43-A was manually fit the same way, using the local refinement map in Coot and the global refinement map in regions where it was well resolved. Once the two components were satisfactorily refined, the model was brought back into ChimeraX where all unrefined subunits were deleted and then regenerated in updated positions by using symmetry operators to clone the manually refined subunits in an octahedral pattern[67,68]. The model was then brought into Phenix where Real Space Refinement was run relative to the global refinement map using zero macrocycles[71]. The output of that refinement was brought into ISOLDE where it was once again relaxed into the global density with the temperature set to 0 K and the map weight was reduced from 0.14 to 0.10. Iteratively, the model was processed using this Coot, then Phenix, then ISOLDE pattern with periodic Comprehensive Validation jobs run in Phenix until the model was acceptable[69–71]. Each time the model was inspected in Coot, residues that were determined to lack supporting density were trimmed to their Cβ carbons. Then the model and the global refinement were submitted to the wwPDB validation service to provide an additional metric of model quality.

Chapter 4: Identification of a protein contaminant using Cryo-EM and artificial intelligence to improve sample purification

Adapted from

Carr, K. D. et al. Protein identification using Cryo-EM and artificial intelligence guides improved sample purification. J. Struct. Biol. X 11, 100120 (2025).

Background

Understanding the details of a complex system is crucial when designing an experimental protocol, especially in the context of protein purification — a process fundamental to various scientific disciplines. The significance of protein purification lies in its ability to isolate recombinant proteins for downstream high-resolution structure determination, exploration of biochemical mechanisms, development of novel therapeutics, and the characterization of computationally designed proteins with specific functions. Despite their widespread use, protein purification protocols still face challenges, such as the co-purification of contaminant proteins from unknown origins[72]. This is particularly problematic in computational protein design, where it is often unclear whether these unknown co-purifying proteins represent off-target design states, cross-sample contamination, or naturally occurring proteins originating from the expression host.

During the characterization of the O43-TM nanoparticle we observed protein contamination throughout our attempts to characterize the nanoparticle. Upon initial purification attempts, we observed an unidentified protein co-eluting with our target nanoparticle, which was also found in several unrelated designed protein samples. To identify the protein, we employed a combination of Cryo-EM Single Particle Analysis (SPA), ModelAngelo, and the Protein Basic Local Alignment Search Tool (Protein BLAST). SPA is a technique that involves imaging thousands of individual biomolecules, then aligning and averaging their 2D projections to reconstruct a high-resolution 3D structure[73,74]. ModelAngelo, a machine-learning tool, automates atomic model building and aids in protein identification by analyzing sequence fragments from Cryo-EM 3D structure information[75]. ModelAngelo's ability to automatically build into Cryo-EM density maps in a sequence-agnostic manner makes it particularly well-suited for the identification of unknown proteins in complex samples where manual methods would be time-consuming or prone to errors[75]. Protein BLAST is a publicly available tool which allows users to reference protein sequences against a sequence database to find similar proteins[76]. By leveraging these tools, as well as orthogonal data from the Protein Data Bank (PDB), AlphaFold 3 (AF3), and a Western Blot, we successfully identified the contaminant protein in our sample[77,78]. This unambiguous identification allowed us to rationally refine our purification protocol to eliminate the contaminant from future preparations.

Results

Observation of an unknown co-eluting protein contaminant during electron microscopy

During the assembly and purification of the full transmembrane nanoparticle O43-TM, we discovered that a significant portion of the purified sample consisted of a smaller, unidentified protein complex, which also exhibited octahedral symmetry, as revealed by ns-EM (**Figure 6**). This smaller complex was also observed in various unrelated samples from the same laboratory, including samples of *de novo* designed fibrous proteins, cyclic oligomers, and two distinct icosahedral nanoparticles (**Supplemental Figure 8A-D**). In each of these other instances, the unknown protein was low in concentration relative to the designed protein, and thus could be largely ignored during downstream characterization and analysis. However, for the O43-TM nanoparticle the contaminant constituted the majority of the purified protein mass (**Figure 6A-D, Supplemental Table 2**). This high level of contamination greatly hindered our ability to accurately assemble and characterize the designed nanoparticle in the quantity and purity necessary for effective downstream biochemical, biophysical, structural, and functional applications. To improve the purification of this two-component nanoparticle, identifying the contaminating protein and its source became essential.

The intended designed two-component nanoparticle, O43-TM, had a predicted mass of 1.18 MDa, a diameter of 29.2 nm, and required the presence of detergent for solubilization and purification. Our nanoparticle preparation protocol necessitated that the two protein components first be purified independently via immobilized metal-affinity chromatography (IMAC). The purified components were initially characterized via SDS-PAGE before being combined to form the desired two-component nanoparticle assembly. This assembly was then purified using SEC, resulting in a final estimated total protein yield of 97 µg (**Supplemental Table 2**).

The SEC trace of the putatively purified O43-TM contained one relevant high molecular weight (MW) peak which eluted off the column between 10 mL and 15 mL, corresponding to a particle size greater than the 660 kDa MW standard for this column (**Figure 6B**). The SEC peak was analyzed using DLS to estimate the diameter of the purified nanoparticle fraction. To account for the presence of detergent micelles in the purification buffers, we performed DLS on the detergent-containing buffer used for SEC purification as a control, alongside the protein-containing peak isolated from SEC (**Figure 6C**). The DLS results showed three peaks with average diameters of 6.30 nm, 30.9 nm, and 223 nm. The smallest (6.30 nm) peak was identified as containing detergent micelles, consistent with the DLS profile of the control. The central peak (30.9 nm) was identified as likely being the designed O43-TM nanoparticle, while the largest peak (223 nm) corresponded to protein and micellar aggregation (**Figure 6C**). Nanoparticle structural characterization by liquid chromatography mass spectrometry (LCMS) was attempted, but results for the protein sample were difficult to interpret (see **Discussion**). ns-EM analysis of the SEC-purified sample revealed the presence of two distinct nanoparticles with diameters (measured diagonally across the 4-fold axis) of approximately 18 - 20 nm and 25 - 30 nm (**Figure 6D)**. The larger particle exhibited the predicted diameter and morphology for the on-target O43-TM nanoparticle, with spike-like proteins extending from an underlying octahedral nanoparticle scaffold (**Figure 6E**). However, the predominant species observed by ns-EM was a smaller, cube-like particle lacking the distinctive spiked features of our intended design (**Figure 6E-F**). As a result, the actual yield of O43-TM was estimated at approximately 3.79 µg from a total of 97 µg of protein purified via SEC (**Supplemental Table 2**). Furthermore, subsequent ns-EM 3D refinements produced a map of this smaller nanoparticle that revealed an octahedral assembly profile which deviated significantly from our intended design (**Figure 6F**).

Given the prevalence of the smaller octahedral complex compared to O43-TM, and its unexpected presence across multiple unrelated samples in the lab, we turned to Cryo-EM SPA to investigate whether this complex might be another designed octahedral nanoparticle assembly, potentially resulting from cross-contamination due to the shared use of purification equipment. Following Cryo-EM data collection, initial 2D class averages of the smaller nanoparticle revealed clearly defined secondary structural elements (**Figure 6G, Supplemental Figure 9A**). Subsequent 3D refinement yielded a final map with a global resolution of 2.51 Å, with the local resolution ranging from 2.88 Å at the periphery to 2.44 Å in the protein core (**Figure 6H, Supplemental Figure 9B-D, Supplemental Table 3**). However, despite the high resolution of the map, the identity of the co-eluting nanoparticle remained unclear in the absence of known sequence or structural information.

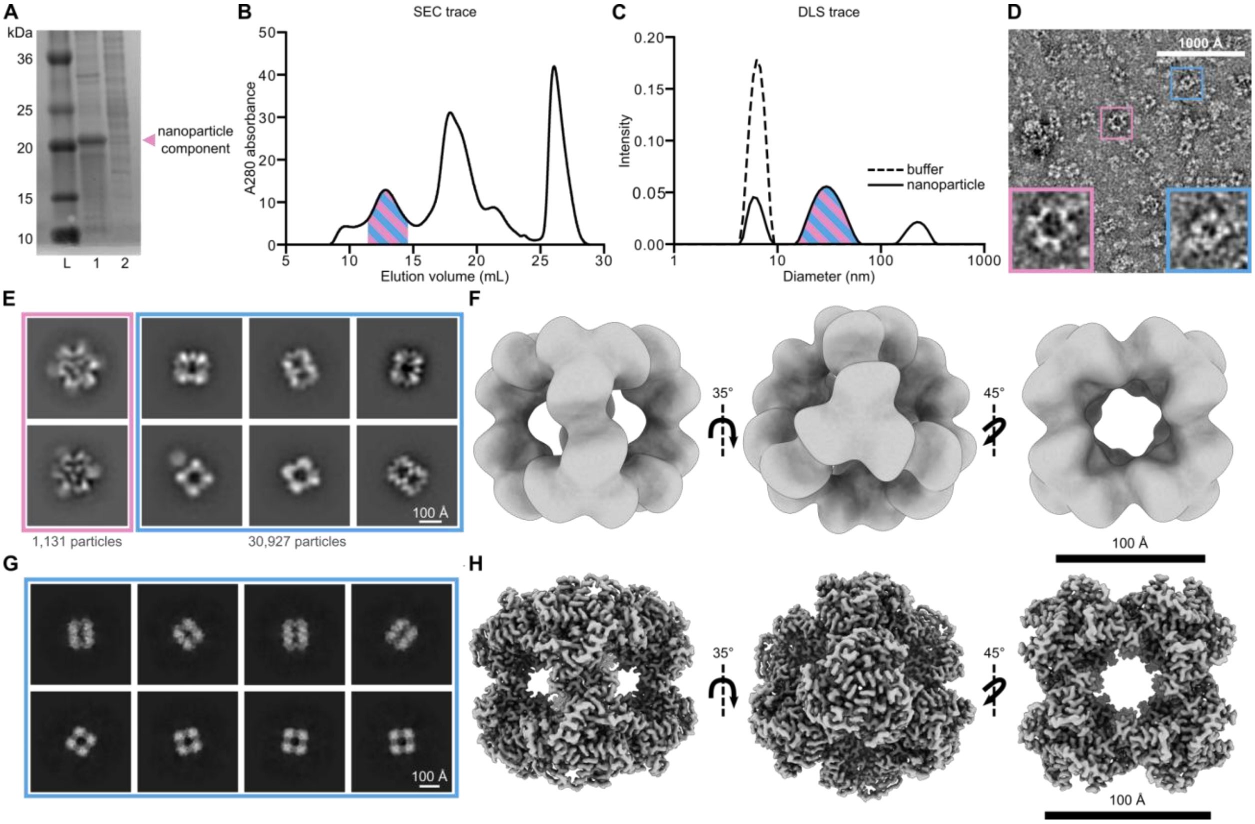


Figure 6: Characterization of an unknown co-eluting protein via electron microscopy.

(A) SDS-PAGE of the IMAC eluate for the two designed protein components (O43-A and O43-B) prior to mixing for assembly into the target designed octahedral nanoparticle (O43-TM). There is a clearly defined band in lane 1 for the one of the two components, O43-A, of the designed nanoparticle. The second component O43-B, shown in lane 2, is present in lower amounts and is less pure. (B) SEC trace of the assembled nanoparticle. The peak corresponding to the nanoparticle is highlighted in stripes to represent the co-elution of the contaminant with O43-TM. (C) DLS trace of the SEC-purified sample highlighted in stripes to represent the contaminant and O43-TM diameters are not separated efficiently by DLS. The buffer (dotted line) was run as a control and contains detergent micelles found at 6.30 nm in diameter. (D) A portion of a ns-EM micrograph of the heterogeneous sample. The pink box represents O43-TM and the blue box represents the contaminant nanoparticle. (E) 2D ns-EM class averages of O43-TM and contaminant nanoparticle species. A total of 1,131 O43-TM nanoparticles were processed and are represented here by two 2D class averages. 30,927 particles were processed for the contaminant species and are represented here by six 2D class averages. Corresponding particle numbers are reflective of all particles of each species in the dataset, not only those of the displayed classes. (F) 3D ns-EM map along the 2-, 3-, and 4-fold axes of symmetry of the contaminant nanoparticle generated using CryoSPARC v4.4. (G) 2D Cryo-EM class averages and (H) 3D Cryo-EM reconstruction viewed along the 2, 3, and 4-fold axes of symmetry of the contaminant nanoparticle using CryoSPARC v4.5. (Blue = contaminant protein; Pink = O43-TM nanoparticle).

ModelAngelo structure-to-sequence identification of the contaminant protein

To identify the unknown nanoparticle in the absence of a corresponding amino acid sequence or prior structural information, we turned to the automated model-building software ModelAngelo (**Figure 7A**)[75]. ModelAngelo's primary mode uses deep learning to integrate available structural data and supplied sequence information to automatically build atomic models into Cryo-EM density maps[75]. It can alternatively operate in a sequence-agnostic manner, automatically generating and building polypeptide sequence fragments into well-resolved regions of Cryo-EM density maps when sequence information is unavailable. Running ModelAngelo in this manner on the 2.51 Å Cryo-EM map of the contaminant protein generated 92 all-atom sequence fragments of varying lengths, along with secondary structure predictions that were in strong agreement with the density of the unknown nanoparticle (**Figure 7B**). We aligned these output chain-fragments using ClustalOmega to generate a multiple sequence alignment which was then input into WebLogo, which produced a frequency plot displaying the relative abundance of amino acids at each position (**Supplemental Figure 10A**)[79]. From this frequency plot we constructed a consensus sequence using the most prevalent amino acid at each position (**Supplemental Figure 10B**). This consensus sequence was input into Protein BLAST, which returned a list of potential matches for the unknown protein contaminant[76]. Remarkably, 98 of the top 100 Protein BLAST results identified dihydrolipoamide succinyltransferase (DLST) (**Supplemental Table 4**). DLST is an octahedral subunit of the E2 component of the α-ketoglutarate dehydrogenase complex (KGDHC), which plays a pivotal role in the tricarboxylic acid (TCA) cycle, converting α-ketoglutarate to succinyl-Coenzyme A and reducing nicotinamide adenine dinucleotide[80–82]. The complex has three structured domains linked together by a chain of unstructured residues, with the C-terminal domain being the catalytic domain where multiple DLST subunits interact to form the full 24-subunit assembly[82]. Notably, DLST has previously been observed as a contaminant protein in samples purified from *E. coli* expression systems[72]. Following the Protein BLAST result, a pairwise sequence alignment using EMBOSS Needle between the Uniprot sequence A0A140NDX4 of DLST from *E. coli* BL21 (DE3) and the ModelAngelo-derived consensus sequence was performed[76,83]. This pairwise sequence alignment excluded the first 170 residues of DLST that were unresolved in our Cryo-EM map. The alignment revealed 60.3% identity and 69.3% similarity, a strong indicator that the observed contaminant was DLST (**Supplemental Figure 10C**)[83].

To further validate the identity of the unknown protein as DLST, we next compared our Cryo-EM density map with previously published structures of the DLST catalytic domain, as well as AlphaFold 3 predicted structures of the UniProt DLST sequence (**Figure 7C-D**)[76,78,83]. The highest resolution structure of DLST in the PDB (1SCZ; X-ray Crystallography; 2.20 Å) was selected for comparison and revealed significant agreement in secondary structure positions when docked into the Cryo-EM map (**Figure 7C**)[84]. Similarly, AlphaFold 3 (AF3) predictions using the catalytic domain of the A0A140NDX4 UniProt sequence also showed high agreement with our Cryo-EM map (**Figure 7D**)[78,83]. A Western Blot was finally performed on all fractions collected during IMAC purification in an attempt to biochemically confirm that the contaminate was DLST (**Supplemental Figure 11, Supplemental Table 5**). Consistent with computational predictions, the anti-DLST Western blot revealed a prominent band near 50 kDa in the IMAC eluate, corresponding to the expected mass of a single DLST subunit (**Supplemental Figure 11**). Together, the computational and experimental results unambiguously confirmed the identity of the co-eluting octahedral protein as DLST.

To generate our final DLST catalytic domain model, the DLST sequence derived from *E. coli* BL21(DE3) was used as the input for another round of automated model building, this time utilizing the sequence-guided feature in ModelAngelo with the UniProt sequence for DLST (**Figure 7A**)[75,83]. The output from ModelAngelo was then further refined to produce the final model of the DLST catalytic domain

(**Supplemental Figure 14**). Aligning this final model to both the 1SCZ monomer and AF3 outputs revealed RMSD values of 0.51 Å and 0.52 Å, respectively (**Figure 7E-F**).

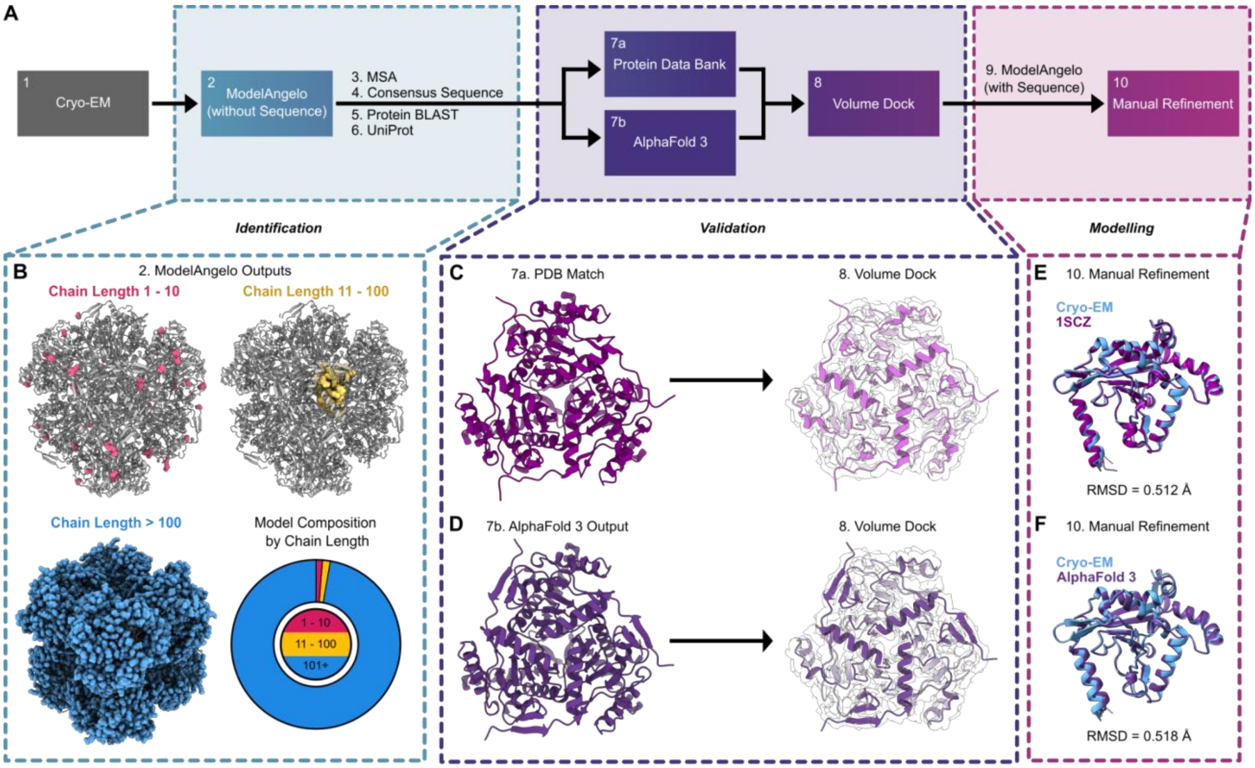


Figure 7: Structure-to-sequence workflow for the unambiguous identification of DLST.

(A) An overview of the 10-step sequence-to-structure workflow used to identify DLST and build our atomic model used for structural analysis. (B) The sequence-agnostic ModelAngelo output using the 2.51 Å Cryo-EM map for the contaminant protein. Chains between 1 to 10 residues (pink), 11 to 100 residues (yellow), and longer than 100 residues (blue) are displayed in sphere view along the octahedral 3-fold axis of symmetry. An accompanying pie chart displays the percentage of all residues belonging to chains of those length ranges. (C-D) Published DLST crystal structure PDB:1SCZ (C) and AlphaFold 3 prediction model of DLST UniProt sequence A0A140NDX4 (D) each docked into the Cryo-EM density map of the unknown co-eluting protein. (E-F) A single subunit of our Cryo-EM model aligned to 1SCZ (RMSD 0.512 Å) (E) and the AlphaFold 3 model (RMSD 0.518 Å) (F).

Revised purification methods increased purity of a *de novo* two component protein nanoparticle

Following the identification of the contaminant protein as DLST, we turned to the literature to explore potential strategies to exclude it from our future purifications. While previous studies noted DLST's co-purification during IMAC of many proteins, none offered a rationale for its co-elution, nor did any outline an effective protocol to eliminate it from the purification pipeline[72]. In an effort to determine the underlying reason for DLST's co-elution with multiple *de novo* protein samples, we first conducted a detailed analysis of its surface properties. Given the use of IMAC to purify our Histidine-tagged designed O43-TM nanoparticle, we initially hypothesized that the surface histidine content of DLST might contribute to its binding during the IMAC purification process. Clusters of surface-exposed histidine residues could potentially interact with the nickel chelating resin in a manner similar to histidine tags on recombinant proteins, leading to its unintended co-purification[85,86]. However, previous reports analyzing the surface histidine content of the DLST did not find any large histidine patches to support this as a potential mechanism for co-elution from IMAC columns[72]. Despite the known lack of histidine clusters, we considered that non-specific interactions between single histidines on the surface of DLST may have interacted with the IMAC column (**Supplemental Figure 12A**). To test this, we applied an elution gradient ranging from 0 M to 0.5 M imidazole during IMAC purification, aiming to isolate the individual protein components and reduce DLST's hypothesized weak binding to the resin. However, DLST was still detected in ns-EM samples after purification (**Supplemental Figure 12B-C**). Furthermore, in samples purified with an imidazole gradient, few to no on-target two-component nanoparticle assemblies were observed (**Supplemental Figure 12C, Supplemental Table 2**).

Given the numerous charged patches on the surface of DLST, we next tested the adjustment of sodium chloride (NaCl) concentrations in our buffers during cell lysis and IMAC purification to determine if increased ionic levels would selectively disrupt any non-specific ionic interactions occurring between DLST and the IMAC column (**Supplemental Figure 12D**)[86]. After separately purifying each protein component using

IMAC with high salt buffers, followed by SEC purification of the nanoparticle, no DLST particles were detected in the sample by ns-EM (**Supplemental Figure 12E-F**). However, while the higher NaCl concentrations eliminated DLST, it also significantly reduced the total yield of assembled O43-TM nanoparticles (**Supplemental Figure 12F, Supplemental Table 2**).

To exclude DLST while also preserving or improving the concentration of the O43-TM nanoparticle, we next tested whether non-specific interactions between DLST and the on-target designed two-component nanoparticle were potentially occurring throughout purification. To do so, we opted to include excipients to all of our purification buffers – a common technique for purification of recombinant proteins to enhance yield and reduce protein aggregation[87–90]. In particular, amino acids such as glycine, threonine, arginine, glutamate, and histidine, have been successfully used as excipients to reduce non-specific protein-protein interactions, flocculation, and protein aggregation[87,90]. They have also been shown to enhance protein stability in solution, particularly when working with computationally designed protein nanoparticles[90].

100 mM glycine and 100 mM arginine were added to all purification buffers to mitigate potential off-target interactions between our nanoparticle and DLST, which resulted in the near-complete removal of DLST in all ns-EM micrographs of these purified samples (**Figure 8A**). Strikingly, this approach significantly improved the final yield of our two-component nanoparticle assembly to an estimated 220 µg of total protein (**Supplemental Table 2**). ns-EM 2D class averages of over 70,000 particles revealed that the designed two-component nanoparticle now became the overwhelmingly dominant species, making up 99.5% of extracted particles—a nearly 58-fold increase in on-target purified protein yield with excipients used throughout purification (**Figure 8B-C, Supplemental Table 2**). In contrast, under standard purification conditions without excipients, the target nanoparticle accounted for only 3.53% of extracted particles (**Figure 8B-C, Supplemental Table 2**). Western Blot analyses comparing the purification of protein components with and without the use of excipients also indicated the near-complete removal from DLST in samples purified using 100 mM glycine and 100 mM arginine (**Figure 8D, Supplemental Figure 11**). To our knowledge, this is the first reported method for effectively removing DLST as a co-purifying protein

contaminant during recombinant protein production from *E. coli* - a challenge that, if left unaddressed, could potentially hinder accurate downstream characterization of many natural and designed protein systems. This result underscores the crucial role excipients played in eliminating DLST from our purification process while significantly improving the overall yield of the O43-TM nanoparticle.

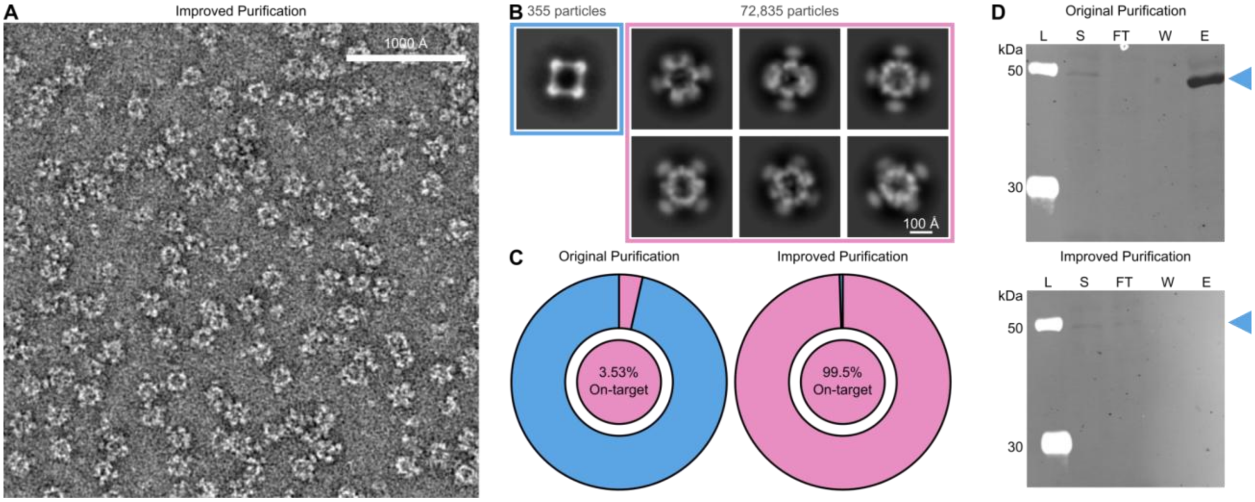


Figure 8: Modifications to the protein purification protocol result in an increased sample purity.

(A) Representative ns-EM micrograph utilizing the optimized purification protocol. (B) Corresponding 2D ns-EM class averages of the O43-TM nanoparticle and DLST, with corresponding particle numbers listed for each. (C) Pie charts comparing the relative abundance of O43-TM against DLST as processed using the original purification protocol and the improved purification protocol. (Blue = contaminant protein, DLST; Pink = O43-TM). (D) Cropped anti-DLST Western Blot. IMAC soluble (S), flowthrough (FT), wash (W) and elution (E) fractions were run on SDS-PAGE followed by Western Blot. The LiCor Chameleon® 700 Pre-stained Protein Ladder (L) was used. Blue labels arrows indicate the molecular weight of a single subunit of DLST. A full annotated Western Blot as well as the accompanying SDS-PAGE gel can be referenced in **Supplemental Figure 11** and **Supplemental Table 5**.

Effect of Cryo-EM resolution on accurate DLST identification and potential broader applicability

Following our successful exclusion of DLST, we aimed to explore the potential broader applicability of this ModelAngelo-to-BLAST identification approach, particularly under conditions where generating near-atomic resolution Cryo-EM data might not be feasible. Establishing the effectiveness of this workflow across a range of resolutions would increase its utility for various experimental setups, including those that often struggle to achieve high resolution. Specifically, we evaluated the ability of the pipeline to accurately identify DLST using Cryo-EM data spanning a broad resolution range by low-pass filtering the original 2.51 Å Cryo-EM map to resolutions between 3.00 Å and 9.00 Å, which were then used as inputs for the ModelAngelo-to-BLAST pipeline **(Figure 9A, Supplemental Figure 13)**. The average length of the sequence fragments generated by ModelAngelo was inversely proportional to the map resolution, with lower resolution input maps yielding shorter length chain fragments **(Figure 9B, Supplemental Table 6-7)**. Following analysis of Protein BLAST results generated using the consensus sequences, the proportion of the top 100 results corresponding to DLST was plotted against their associated 3D map resolution values **(Figure 9C, Supplemental Sequences 4, Supplemental Tables 8-13)**[76]. This analysis revealed that at resolution thresholds of 3.00 Å, 4.00 Å, 4.50 Å, and 4.75 Å, Protein BLAST accurately identified DLST as the most likely identity of the unknown protein in 99%, 99%, 98%, 95%, and 97% of cases, respectively **(Figure 9C)**. However, at 5.00 Å and at all lower resolutions, the consensus sequences did not yield any significant Protein BLAST results, indicating a lack of detectable sequence similarity among known proteins for sequence fragments generated by ModelAngelo at these resolutions **(Figure 9C)**.

In addition to conducting Protein BLAST searches on consensus sequences, we also evaluated the ability of three other structure-agnostic approaches to identify DLST. These included the hmmsearch function built into ModelAngelo, external hmmsearch queries based on HMM profiles of the ten longest chains generated by ModelAngelo, and Protein BLAST searches using the same ten chains of each model [75]. For resolutions better than 4.50 Å, three of the four methods demonstrated high identification rates: hmmsearch within ModelAngelo, hmmsearch using the 10 longest chains, and Protein BLAST queries on the 10 longest chains. The latter two methods performed exceptionally well, achieving nearly 100% success rates **(Figure 9C)**. The hmmsearch function within ModelAngelo, however, exhibited a lower identification rate (~70%)

within this range **(Figure 9C)**. At lower resolutions, the performance of all methods declined, with external hmmsearch and Protein BLAST correctly identifying DLST as the contaminant in fewer than 20% of cases by 6.00 Å **(Figure 9C)**. The hmmsearch function within ModelAngelo maintained a higher identification rate (~50%) at resolutions as low as 6.50 Å, outperforming the other methods in this range **(Figure 9C)**. Ultimately, none of the tested methods were successful at identifying DLST at resolutions worse than 6.50 Å **(Figure 9C)**. These findings suggest that combining methods can optimize the identification of unknown proteins in Cryo-EM samples, with the choice of method tailored to the resolution of the starting structure. The flexibility of these four methods across a wide resolution range underscores their broad applicability in addressing diverse biological and experimental challenges in protein identification, even when near-atomic resolution cryo-EM data is unavailable.

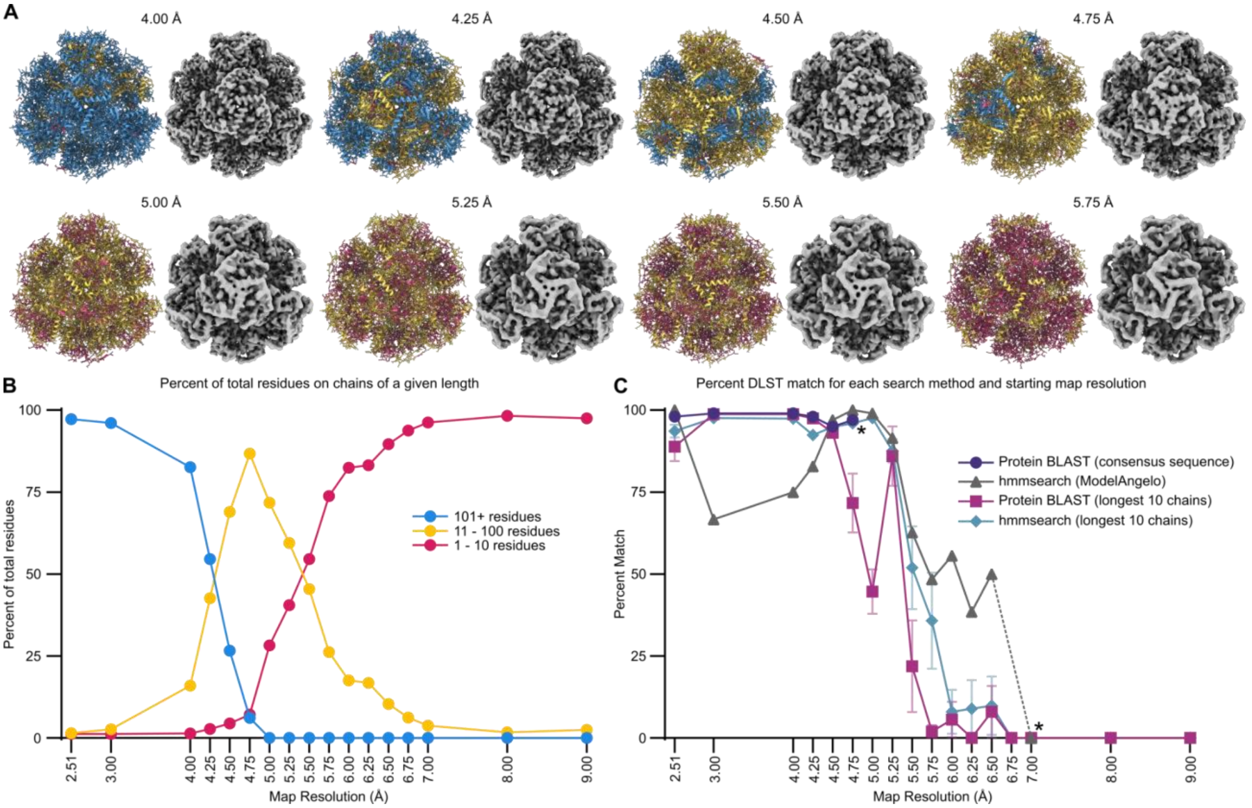


Figure 9: Impact of Cryo-EM resolution and search method on accurate DLST identification.

The final Cryo-EM map was low-pass filtered to 3.00 Å, 8.00 Å, 9.00 Å, and each 0.25 Å interval between 4.00 Å and 7.00 Å. Each map was used as the input for the computational steps of our workflow, comparing the efficacy of three identification approaches: Protein Blast using a consensus sequence, Protein BLAST using the longest 10 chains, and hmmsearch using HMM profiles generated by ModelAngelo for the longest 10 chains. Chains with 1–10 residues are shown in pink, 11–100 residues in yellow, and chains longer than 100 residues in blue. (A) Low-pass filtered Cryo-EM maps and models between 4.00 Å and 6.75 Å. Additional maps and models spanning a broader resolution range can be found in **Supplemental Figure 13**. (B) Line graph illustrating the percentage of residues organized by chain length for each generated model. Raw data used for the line graph can be found in **Supplemental Table 6**. (C) Line graph comparing the efficacy of the four identification methods across resolutions. Each data point represents the average percent identification accuracy for DLST using the specified method, with error bars representing the 95% confidence interval for methods which had more than one data point per resolution. Asterisks (*) represent the lowest resolution data that returned any results for that search method. A dashed line marks the range between 6.50 Å and 7.00 Å on the hmmsearch (ModelAngelo) data to communicate that no results were returned for that method at 6.75 Å and consequently 6.50 Å is the lowest resolution that was able to achieve a non-zero score for that method.

Discussion

Addressing the challenge of co-eluting protein contaminants during protein purification is essential for the production and accurate experimental characterization of recombinant proteins. In the case of the purification of O43-TM discussed here, the presence of a co-purifying protein artificially inflated the estimated concentration of the designed components required for nanoparticle assembly, leading to inconsistent assembly conditions and inaccurate estimations of the final O43-TM nanoparticle concentration **(Figure 6, Supplemental Table 2)**. Thus, it became necessary to identify and exclude this protein from our purifications so that we could more effectively produce and characterize our on-target design. By using the ModelAngelo-to-BLAST pipeline and cross-referencing our Cryo-EM data with experimentally determined PDB structures, AlphaFold 3 predictions, and traditional biochemical analysis, we unambiguously identified the co-purifying protein as DLST and optimized our purification protocols to eliminate it.

Before obtaining secondary structure information from Cryo-EM and implementing the ModelAngelo-to-BLAST pipeline, we hypothesized that the co-eluting protein might be the result of cross-sample contamination from another computationally designed protein purification. During our investigation, we explored several alternative methods for determining the identity of the unknown protein. One study utilizing Cryo-EM had successfully identified a number of unknown proteins found in native cell extracts utilizing the software package Omokage[91]. However, this approach requires existing structural information to confidently identify the unknown protein species[91,92]. Furthermore, a retrospective analysis of our data revealed that Omokage misidentified the contaminant protein in most cases, proposing potential identifications that included both computationally designed and natural proteins, though DLST also appeared among these candidates. These misidentifications underscore the lower success rate of this approach compared to the more accurate sequence-based methods benchmarked in this study.

We also considered traditional experimental methods for our initial identification of the contaminant, such as liquid chromatography-tandem mass spectrometry (LC-MS/MS) and electrospray ionization mass spectrometry (ESI-MS). While these techniques are highly valuable, they posed significant anticipated challenges for our system, primarily due to our reliance on detergents to stabilize our O43-TM nanoparticle. Indeed, detergents are well-known to complicate mass spectrometry workflows and the interpretation of final data, adding significant complexity to an already complicated purification process[93]. Unlike the Omokage and mass spectrometry methods, the ModelAngelo-to-BLAST pipeline did not rely on prior PDB structural information, and enabled us to accurately identify the contaminant using only our experimental Cryo-EM data and publicly available sequence databases. Further validation using AlphaFold 3 models was as robust as PDB models during docking into the Cryo-EM density map, demonstrating that ModelAngelo-to-BLAST identifications can be verified even without the availability of pre-existing PDB information or manual building[78,92]. Additionally, while Western Blotting was used to biochemically confirm the computational predictions, its success was heavily contingent upon the computational results correctly guiding the selection of the appropriate primary antibody. Thus, the Western blot served primarily as a validation tool, reinforcing the reliability of the computational pipeline used here.

An advantage of this computational approach to contaminant identification is its flexibility across a wide range of resolutions. While we initially relied on a near-atomic resolution Cryo-EM map for the ModelAngelo-to-BLAST pipeline, we found that significantly lower resolutions were still sufficient to accurately identify DLST in our sample, consistent with previous benchmarking results for ModelAngelo **(Figure 9C, Supplemental Table 4)**[75]. This robustness across resolutions highlights the adaptability of the pipeline, making it particularly valuable for experimental setups where achieving near-atomic resolution data is not feasible due to sample heterogeneity, limited imaging time, or other technical constraints.

Our analysis further revealed that a tailored combination of methods provided the highest level of confidence across varying resolutions. Notably, Protein BLAST using consensus sequences excelled at resolutions better than 5.00 Å for correctly identifying DLST. In contrast, the hmmsearch function within ModelAngelo was most reliable within the 5.00 Å to 6.50 Å resolution range, but somewhat unexpectedly, showed reduced reliability at higher resolutions. This diminished performance was due to the hmmsearch function frequently misidentifying the contaminant as dihydrolipoamide acetyltransferase (DLAT) at resolutions between 2.50 Å and 5.0 Å. This issue highlighted the importance of leveraging multiple complementary approaches, as each method demonstrated strengths and weaknesses for this sample depending on the resolution range. Notably, methods that analyzed the longest chains of each DLST model—either through Protein BLAST or hmmsearch—offered minimal to no additional practical benefit across any resolution range compared to the other two approaches, while also requiring significantly more time and effort to prepare.

The ability of these pipelines to function effectively across such a broad range of resolutions demonstrates their potential as practical tools for addressing contaminant identification challenges. This is particularly relevant for diagnosing sample purity issues in non-routine purifications or experimental scenarios where cryo-EM data collection is feasible or preferred for biochemical or experimental reasons, yet achieving near-atomic resolution may not be practical.

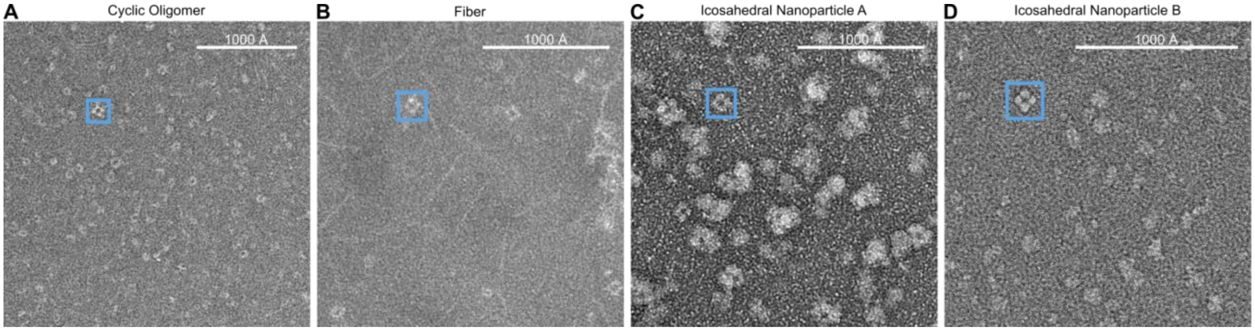
Regarding the high-resolution Cryo-EM structure of the DLST catalytic domain determined in this study, we note that minimal differences were observed between our structure and other previously reported crystallographic and Cryo-EM structures of the DLST catalytic domain **(Supplemental Figure 14)**. The resolution we achieved for DLST using octahedral refinement allowed for the clear construction of a detailed intramolecular water network, which was particularly well-resolved at the catalytic pocket **(Supplemental Figure 14A-D)**. Comparison of this water network with the highest resolution published structure revealed key differences in the assignment of H_2O atomic coordinates **(Supplemental Figure 14E-F)**[84]. These differences may reflect differences in the protein crystallization conditions versus the native-like environment of Cryo-EM sample preparation, potentially offering valuable insights for future studies aimed at understanding this enzyme's mechanistic activity. Additionally, asymmetric data processing of DLST reduced both local and global resolution of the catalytic domain but uncovered an intriguing, previously unobserved segment of density along one of the 4-fold axes of DLST **(Supplemental Figure 15)**. This unidentified density, which globally resolves to between 5.50 Å and 8.50 Å, measures approximately 20 Å along its longest axis, and is positioned proximally to DLST residues E224, I229, R230, K299, V303, R306, and D307 **(Supplemental Figure 15)**. Notably, the portion of this density that resolves to better than 6.50 Å—the lowest resolution threshold at which we identified DLST (a significantly larger protein)—is extremely small. This limited resolution poses challenges to its identification, with further attempts to characterize this density falling beyond the scope of this study. Despite this, we believe these findings are important and provide a foundation for future studies aimed at elucidating the structure, dynamics, and function of this enzyme in greater detail.

Ultimately, by leveraging advanced tools such as Cryo-EM, ModelAngelo, Protein BLAST, the PDB, and AlphaFold 3, we transformed a challenging purification problem into an opportunity to refine our techniques and enhance the quality of our results. This integrated approach underscores the potential of combining Cryo-EM with emerging AI and sequence-based tools as robust alternatives for tackling long-standing biochemical challenges—particularly in purifying proteins from complex biological systems, even when near-atomic resolution data cannot be achieved.

Acknowledgements

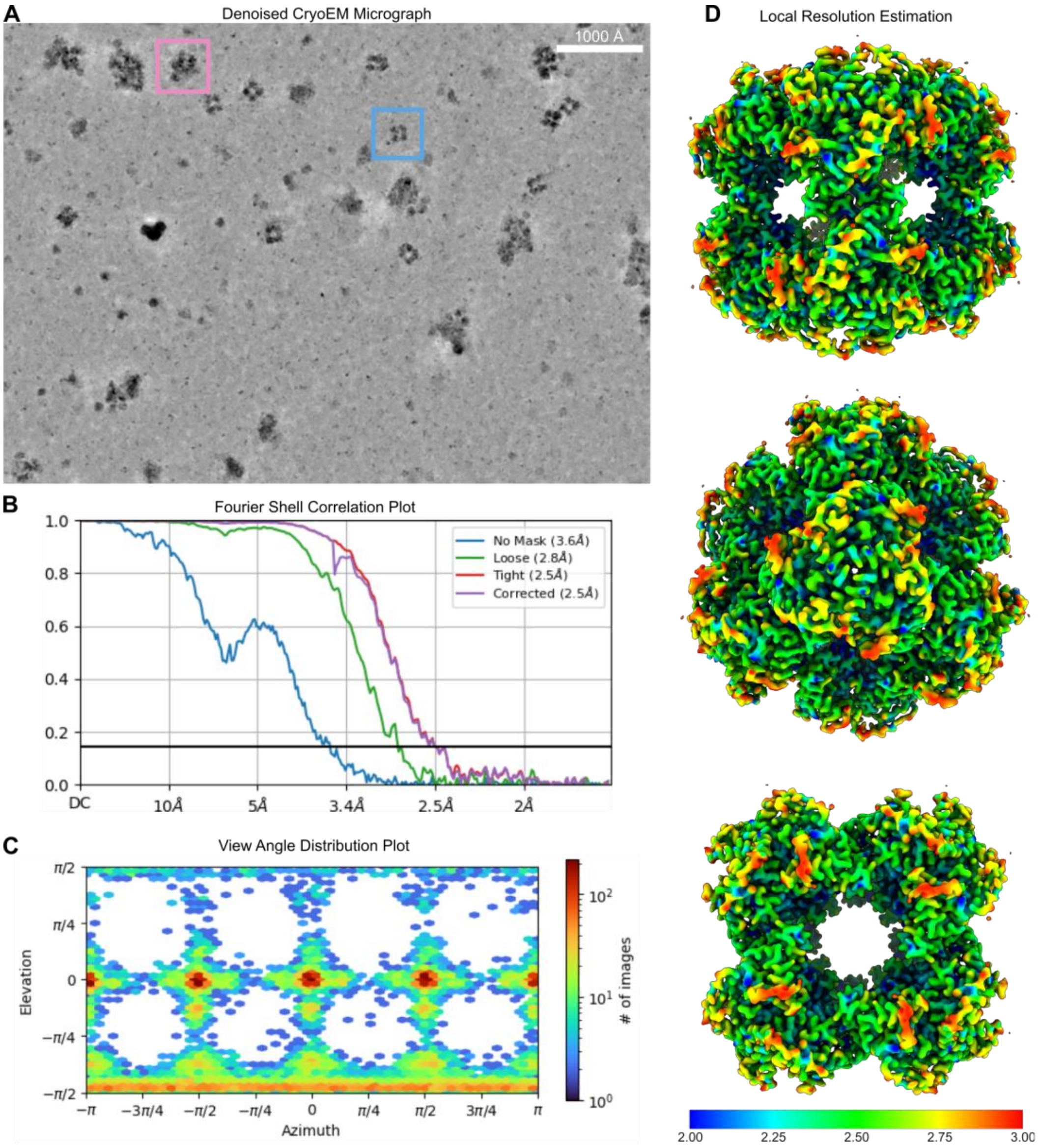
I would like to thank the developers of ModelAngelo, without whom I would probably still be banging our heads against the bench, desperately cleaning SEC columns in an attempt to remove the contaminant from our *de novo* designed protein samples. I want to thank Kenneth Carr, who was very stubborn and insistent that I could find a way to remove the contaminant, and Andrew Borst, for his endless encouragement on this project.

Supplemental Data



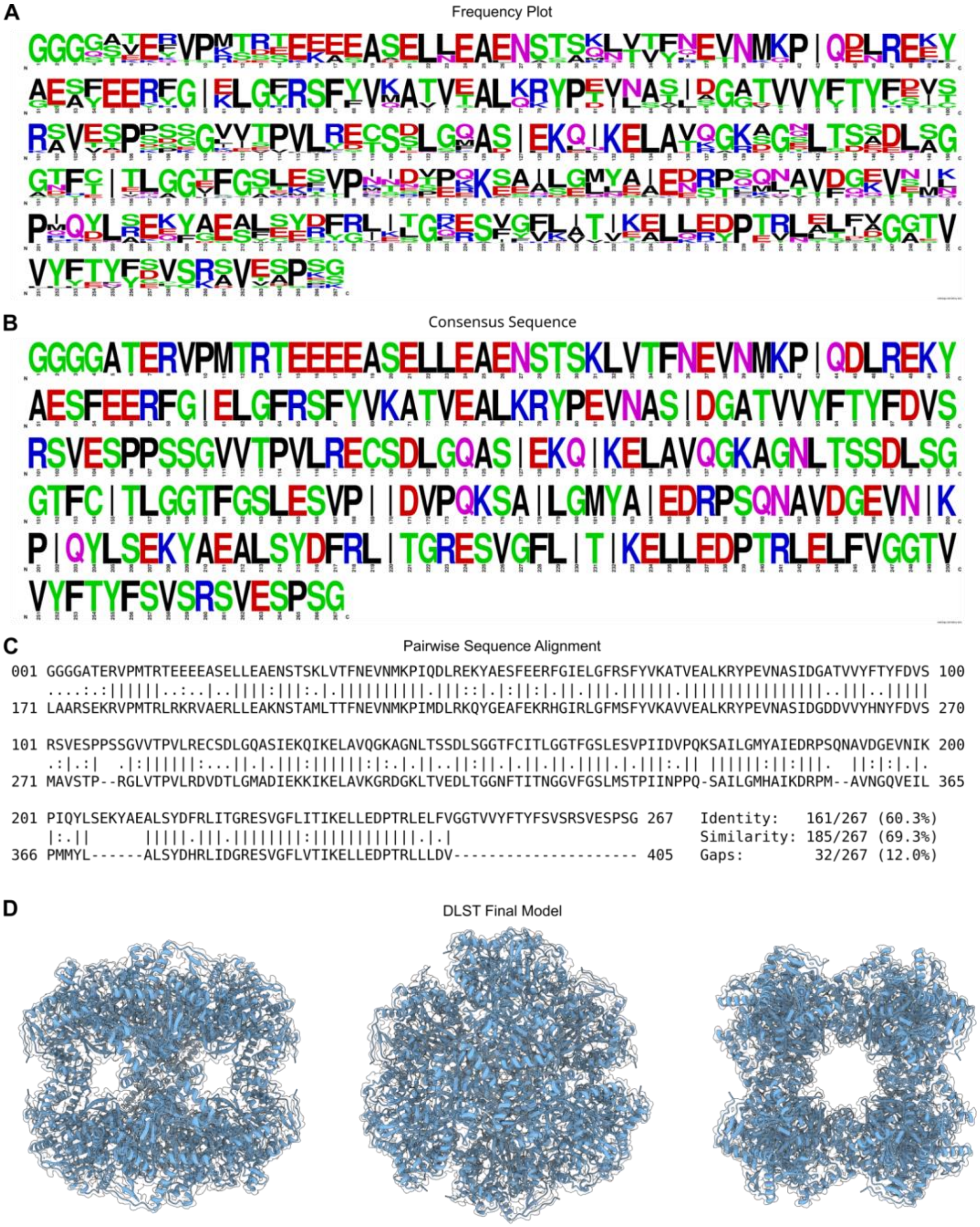
Supplemental Figure 8: Octahedral contaminant protein observed in additional *de novo* protein samples by negative stain electron microscopy.

Sections of representative ns-EM micrographs demonstrating observation of the octahedral contaminant protein in multiple samples of *de novo* designed proteins imaged at various magnifications, with 1000 Å scale bars sized proportionally. (A) A sample of a *de novo* cyclic-oligomeric protein imaged at 45,000x magnification. (B) A sample of a *de novo* protein fiber imaged at 57,000x magnification. (C) A sample of an icosahedral nanoparticle imaged at 57,000x magnification. (D) A sample of a different *de novo* icosahedral nanoparticle imaged at 73,000x magnification.



Supplemental Figure 9: Cryo-EM Data processing metrics.

(A) Denoised representative Cryo-EM micrograph at 105,000x magnification showing particles of both the designed two-component nanoparticle and DLST. (Blue = contaminant protein; Pink = on-target designed two-component nanoparticle). (B) Fourier shell correlation (FSC) plot of the Cryo-EM map. (C) View angle distribution plot of the Cryo-EM map. (D) Local resolution estimate of the Cryo-EM map with a 0.143 FSC cutoff. Colors range from blue at 2.00 Å to red at 3.00 Å.

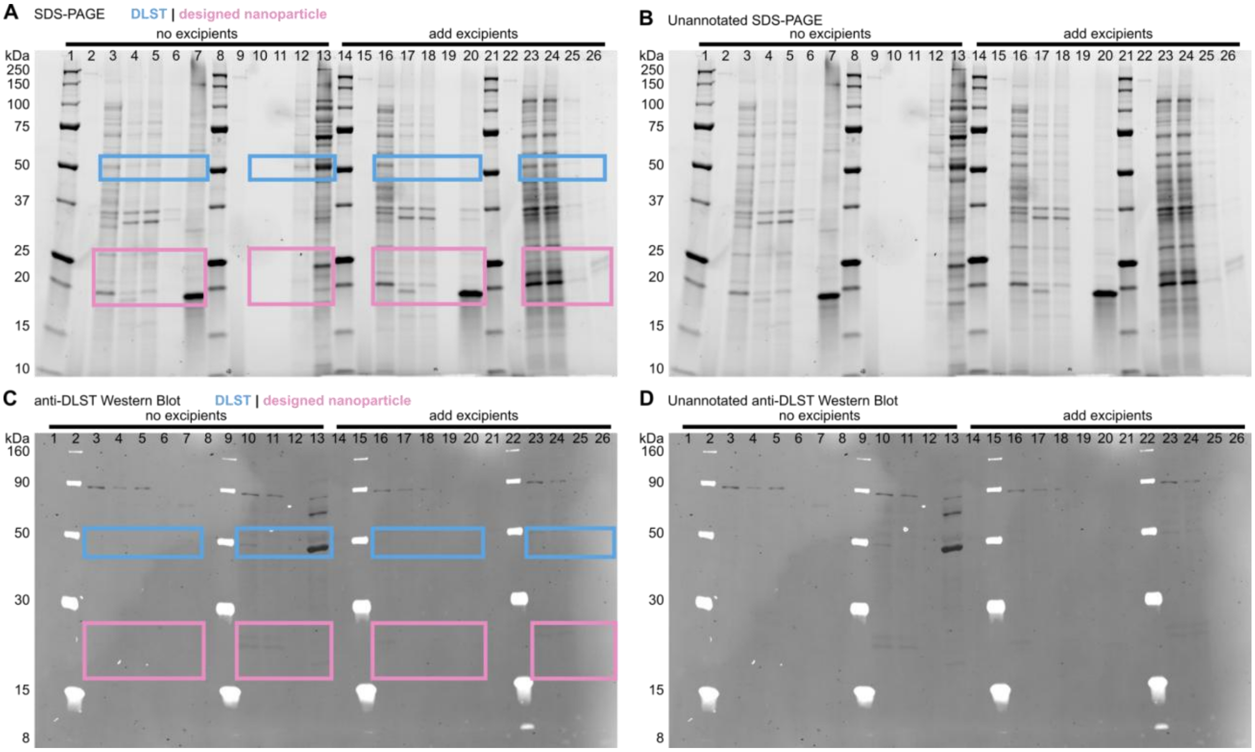


```
001 GGGGATERVPMTRTEEEEASELLEAENSTSKLVTFNEVNMKPIQDLREKYAESFEERFGIELGFRSFYVKATVEALKRYPEVNASIDGATVVYFTYFDVS 100
    ....:.:|||||..:..|..||||:|||:.|.||||||||||.|||::|.|:||:|.||.|||.||||||.|||||||||||||||..|||..|||||
171 LAARSEKRVPMTRLRKRVAERLLEAKNSTAMLTTFNEVNMKPIMDLRKQYGEAFEKRHGIRLGFMSFYVKAVVEALKRYPEVNASIDGDDVVYHNYFDVS 270

101 RSVESPPSSGVVTPVLRECSDLGQASIEKQIKELAVQGKAGNLTSSDLSGGTFCITLGGTFGSLESVPIIDVPQKSAILGMYAIEDRPSQNAVDGEVNIK 200
    .:|.:|  .|:||||||:...||.|.|||:||||||:|:.|.||..||:||.|.||.||.||||.|.|||:.|| ||||||:||:|||.  ||:|:|.|.
271 MAVSTP--RGLVTPVLRDVDTLGMADIEKKIKELAVKGRDGKLTVEDLTGGNFTITNGGVFGSLMSTPIINPPQ-SAILGMHAIKDRPM--AVNGQVEIL 365

201 PIQYLSEKYAEALSYDFRLITGRESVGFLITIKELLEDPTRLELFVGGTVVYFTYFSVSRSVESPSG 267   Identity:   161/267 (60.3%)
    |:.||      |||||.|||.||||||||:||||||||||||.|.|                         Similarity: 185/267 (69.3%)
366 PMMYL------ALSYDHRLIDGRESVGFLVTIKELLEDPTRLLLDV--------------------- 405   Gaps:        32/267 (12.0%)
```

Supplemental Figure 10: Sequences derived from ModelAngelo's sequence-agnostic method used to identify and model DLST from the Cryo-EM volume map.

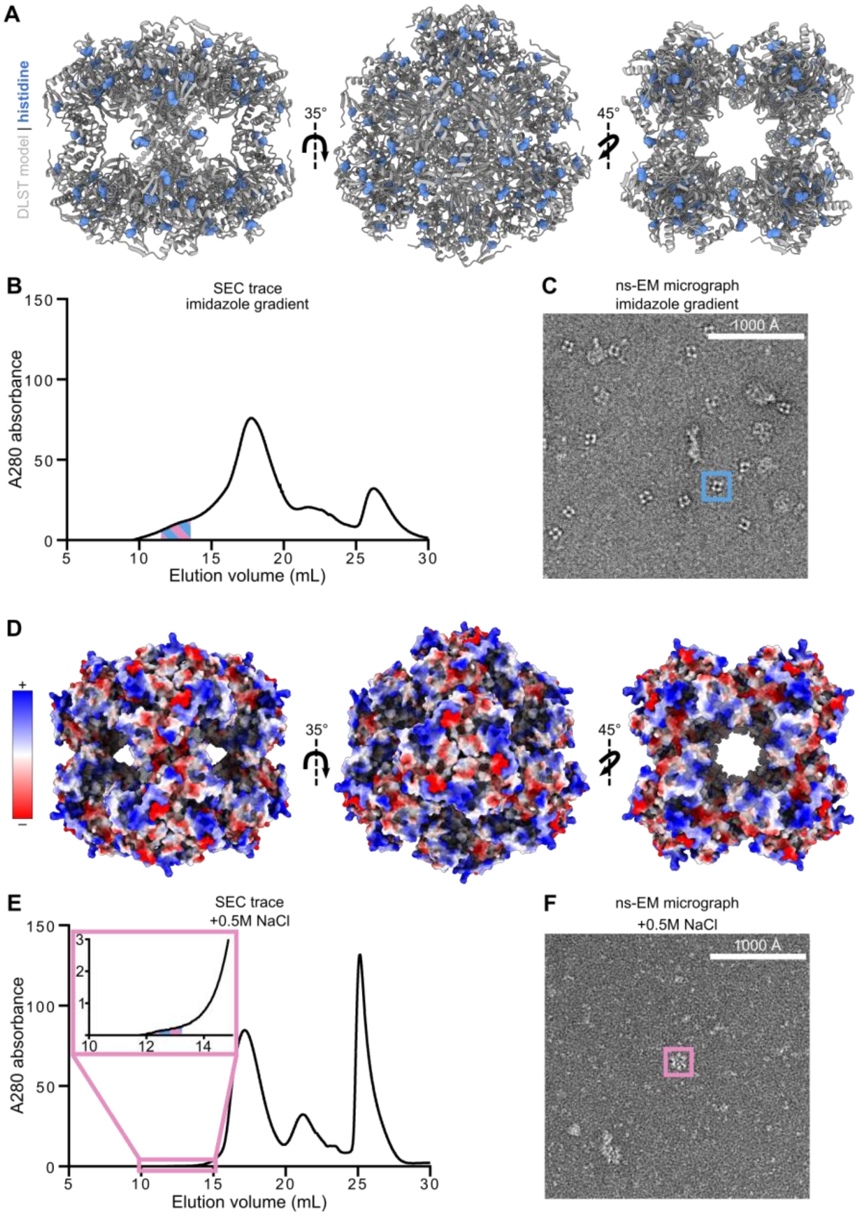
(A) ModelAngelo sequence fragments were aligned in an MSA to generate a frequency plot displaying relative abundance of amino acids at each position. (B) A consensus sequence for one chain was generated using amino acid frequency data from (A) at each position. This consensus sequence was the input for

Protein BLAST[76]. (C) After Protein BLAST identified DLST, UniProt was searched for the corresponding DLST sequence for the BL21 (DE3) expression vector. The UniProt sequence A0A140NDX4 was aligned with the consensus sequence[83]. The first alignment revealed a gap spanning the first 170 residues of the UniProt sequence so an additional alignment was performed using only the UniProt residues at position 171 or later. (D) The final built model of the DLST catalytic domain docked into the 2.51 Å density map along the 2-, 3-, and 4-fold axes of symmetry.



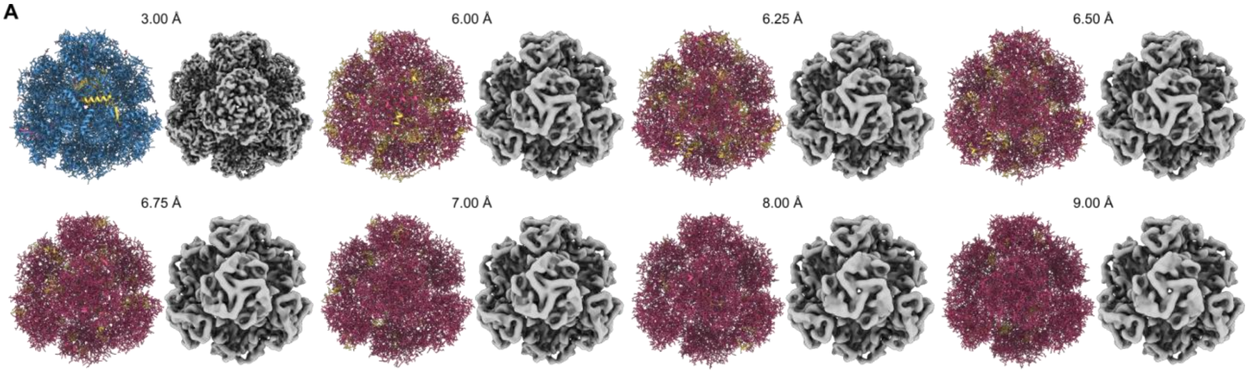
Supplemental Figure 11: Full SDS-PAGE and Western Blot of improved purification of the designed two-component nanoparticle purified with and without excipients.

Labels for individual lanes are located in **Supplemental Table 5**. (A-B) Annotated and Unannotated SDS-PAGE of IMAC purification of each component of the designed two-component nanoparticle. Both protein components of the designed nanoparticle fall between 25.7 kDa and 23.2 kDa and are generally represented by the pink boxes. The DLST catalytic subunit is labeled around 50 kDa by the blue boxes. (C-D) Annotated and unannotated anti-DLST Western Blot from the SDS-PAGE gel transfer.



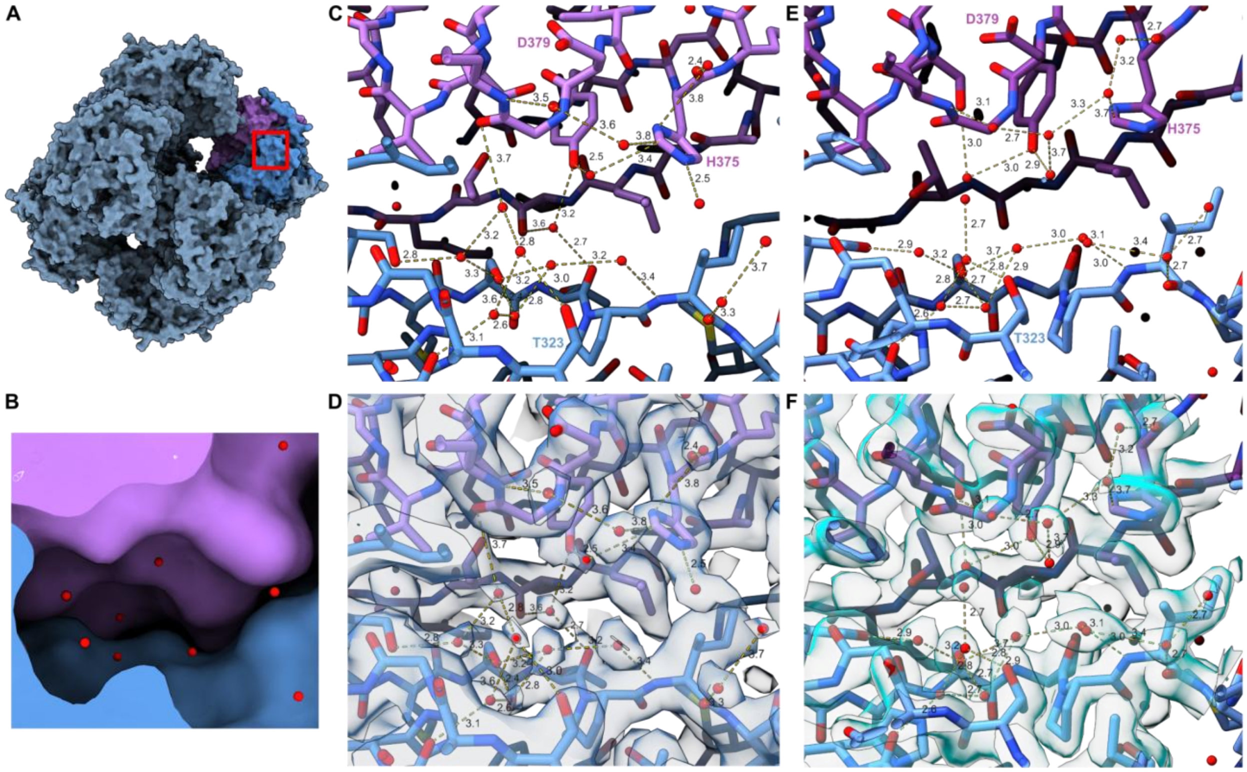
Supplemental Figure 12: Improving purification of the designed nanoparticle through modification of purification protocols to remove DLST.

Various purification attempts of the on-target nanoparticle were conducted, evaluating buffer conditions to eliminate the co-purification of DLST. (A) Location of histidines (blue) in DLST along the 2-, 3-, and 4-fold axis of symmetry. (B) SEC trace of the designed nanoparticle after protein components were purified with an elution gradient from 0.00 M - 0.50 M imidazole. (C) The highlighted fraction from SEC purification was taken for ns-EM. This is a cropped view of a ns-EM micrograph taken at 57,000x magnification. The majority of observed particles were of DLST. (D) Electrostatic potential of the catalytic domain of DLST with positively charged residues shown in blue and negatively charged residues in red along the 2-, 3-, and 4-fold axes of symmetry. (E) SEC trace of the octahedral nanoparticle after cells were lysed with 0.5 M NaCl. IMAC buffers to purify components also contained 0.5 M NaCl. Zoomed-in window highlighting the SEC fraction taken for ns-EM in the next panel. (F) The highlighted fraction from SEC where the expected elution time of the nanoparticle was taken for ns-EM. This is a cropped view of a ns-EM micrograph taken at 57,000x magnification. Particle total yield and total concentration of assembled material was significantly diminished **Supplemental Table 2)**. We did not observe the presence of DLST in these micrographs. (Blue = contaminant protein, DLST; Pink = on-target two component nanoparticle).



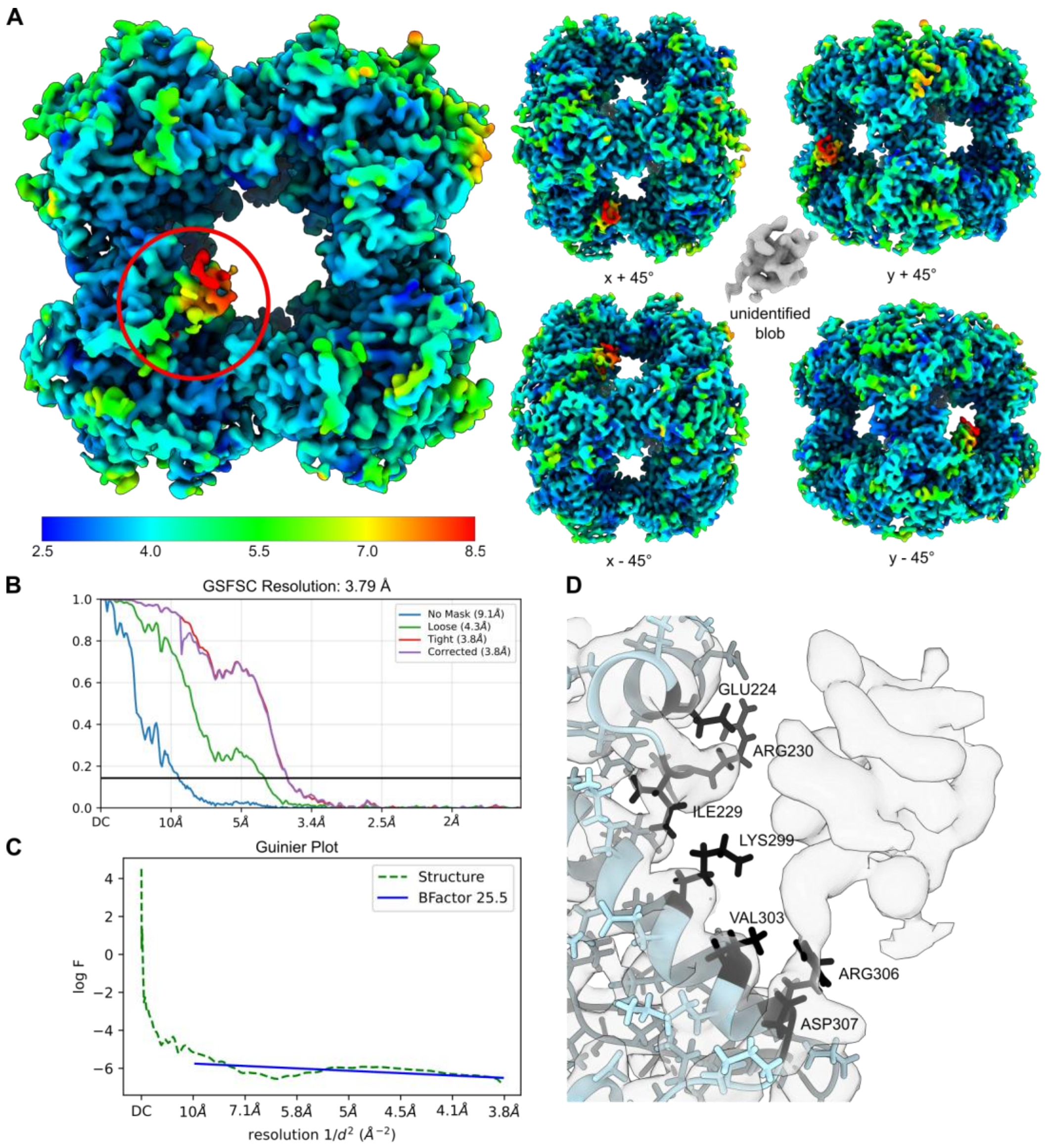
Supplemental Figure 13: Additional maps and models.

(A) Additional low-pass filtered maps and associated ModelAngelo output models between 3.00 Å and 9.00 Å not shown in **Figure 9**.



Supplemental Figure 14: A comparison of the active site water network coordinates for DLST.

Water network of the DLST active site comparing the 2.51 Å Cryo-EM structure discussed here with the 2.20 Å crystal structure (PDB: 1SCZ)[84]. The catalytic residues HIS375, ASP379, THR323 are labeled. (A) Our Cryo-EM structure surface view, with a red box showing the location of the active site. (B) A zoomed in view of the active site highlighted in (A) using a cutaway surface view. (C) The Cryo-EM structure of DLST built from the 2.51 Å Cryo-EM map. (D) The structure in (C) fit within the Cryo-EM map. (E) 1SCZ model taken from the deposited structure[84]. (F) 1SCZ fit within the 2Fo-Fc map[84].



Supplemental Figure 15: Asymmetric refinement of DLST catalytic domain.

(A) Local resolution estimation of asymmetrical refined DLST catalytic domain (sharpened using DeepEMhancer) and the unidentified blob along the 4-fold axis, and the 2-fold axes nearest the face with the blob. The scale bar shows local resolution ranging from 2.50 Å (blue) to 8.50 Å (red). The unidentified blob is circled in red in the 4-fold axis image and as stand-alone density (grey) isolated from the rest of the model. (B) Gold Standard Fourier Shell Correlation (GSFSC) of the asymmetric map with a global resolution estimation of 3.79 Å at FSC 0.143. (C) Guinier Plot of the asymmetric map showing a BFactor of 25.5. (D)

Cutaway-view showing the final DLST structure docked into the asymmetric density. The residues nearest to the blob were identified as GLU224, ILE 229, ARG230, LYS299, VAL303, ARG306, and ASP307.

Supplemental Table 2: Estimated protein yields from different purification methods based on particles picked and sorted by ns-EM 2D classification

Purification Method	**Total Yield (µg)**	**Designed Protein**			**Contaminant Protein**		
		Particles Extracted (%)	**Mass Extracted (%)**	**Estimated Protein Yield (µg)**	**Particles Extracted (%)**	**Mass Extracted (%)**	**Estimated Protein Yield (µg)**
Original Purification	97	3.53	3.91	3.79	96.5	96.1	93.2
+ 0 mM - 0.5 M Imidazole gradient	75	--	--	--	--	--	--
+ 0.5 mM NaCl	Below detection for UVis	1 particle observed	--	--	--	--	--
+ 100 mM Gly + 100 mM Arg	220	99.5	99.6	219	0.49	0.44	0.96
Improvement w/ excipients (fold)	***2.27***	***28.2***	***25.5***	***57.7***	***0.01***	***0.00***	***0.01***

Supplemental Table 3: Cryo-EM Data Collection, Map, and Model Statistics.

Sample Preparation	
Instrument	Vitrobot Mk 4
Temperature (C)	22°
Humidity (%)	100
Blot Force (s)	0
Blot Time (s)	0.5
Wait Time (s)	7.5
Grid Type	Quantifoil 2/2 Holey Carbon +2nm Carbon Film
Data Collection	
Instrument	FEI Titan Krios
Camera	Gatan K3
Pixel Size (Å/pixel)	0.843
Energy Filter	Bioquantum
Voltage (kV)	300
Dose Rate (e^{-1} pix^{-1} s^{-1})	8.4
Frame Rate (s)	0.05
Total Dose (e^{-}/Å^{2})	47
Frames per Stack	75
Defocus Range (µm)	0.8 - 1.8
Movies Recorded	6,211
Cryo-EM Map	
Particles in Final Map	18,809
Symmetry	Octahedral
Global Resolution (Å)	2.51
Maximum Local Resolution (Å)	2.88
Minimum Local Resolution (Å)	2.44
Atomic Model	
Number of Chains	24
Residues per Chain	233
Bond Lengths (Å)	0.004
Bond Angles	1.005
MolProbity Score	1.51
Clash Score	5.22
Ramachandran plot Outliers (%)	0.00
Ramachandran plot Allowed (%)	0.43
Ramachandran plot Favored (%)	99.57
Rotamer Outliers (%)	1.99
Asymmetric Cryo-EM Map	
Particles in Final Map	19,033
Symmetry	C1 / Asymmetric
Global Resolution (Å)	3.79
Maximum Local Resolution (Å)	2.50
Minimum Local Resolution (Å)	8.50

Supplemental Table 4: Percent DLST identification using Protein BLAST (consensus sequence), hmmsearch (ModelAngelo), Protein BLAST (longest 10 chains), and hmmsearch (longest 10 chains).

		Protein BLAST (consensus sequence)	**hmmsearch (ModelAngelo)**	**Protein BLAST (longest 10 Chains)**										**hmmsearch (longest 10 chains)**									
Chain Number		*consensus*	*all chains*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Map Resolution (Å)	2.51	98.0	100	98.0	97.0	98.0	98.0	99.0	98.0	94.0	73.0	70.0	64.0	96.0	98.0	96.0	98.0	98.0	99.0	96.0	86.0	82.0	87.0
	3.00	99.0	66.7	99.0	99.0	99.0	99.0	99.0	99.0	99.0	98.0	98.0	99.0	98.0	98.0	98.0	97.0	97.0	98.0	98.0	97.0	98.0	97.0
	4.00	99.0	75.0	98.0	99.0	99.0	99.0	99.0	99.0	98.0	99.0	99.0	99.0	97.0	97.0	98.0	97.0	98.0	97.0	98.0	97.0	98.0	97.0
	4.25	98.0	82.9	99.0	98.0	98.0	98.0	98.0	94.0	99.0	99.0	99.0	93.0	97.0	93.0	95.0	86.0	94.0	86.0	95.0	93.0	87.0	98.0
	4.50	95.0	97.00	97.0	98.0	98.0	98.0	92.0	90.0	88.0	87.0	91.0	92.0	94.0	95.0	95.0	94.0	96.0	94.0	96.0	95.0	94.0	96.0
	4.75	97.0	100	98.0	93.0	93.0	88.0	80.0	86.0	30.0	27.0	86.0	36.0	95.0	99.0	96.0	97.0	95.0	95.0	96.0	96.0	95.0	96.0
	5.00	--	99.0	50.0	48.0	33.0	27.0	79.0	26.0	26.0	46.0	28.0	84.0	98.0	98.0	98.0	96.0	98.0	98.0	98.0	97.0	98.0	97.0
	5.25	--	91.5	95.0	100	93.0	92.0	--	--	--	50.0	--	--	96.0	94.0	97.0	94.0	94.0	80.0	69.0	97.0	68.0	87.0
	5.50	--	62.7	78.0	93.3	--	--	2.00	0.00	0.00	2.00	0.00	0.00	93.0	95.0	89.0	79.0	86.0	20.4	26.0	31.0	0.00	0.00
	5.75	--	48.4	--	0.00	0.00	0.00	0.00	18.0	0.00	0.00	0.00	0.00	91.0	0.00	93.88	0.00	0.00	81.0	92.0	0.00	0.00	0.00
	6.00	--	55.6	0.00	0.00	0.00	0.00	1.00	0.00	55.0	0.00	0.00	0.00	1.37	0.00	0.00	0.00	11.6	0.00	67.0	0.00	0.00	0.00
	6.25	--	38.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	79.0	0.00	--	0.00	0.00	1.67	0.00	0.00
	6.50	--	50.0	1.00	0.00	0.00	79.0	0.00	0.00	0.00	0.00	0.00	0.00	8.16	0.00	0.00	81.0	--	0.00	0.00	0.00	0.00	0.00
	6.75	--	--	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	7.00	--	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	--	0.00	0.00	0.00	0.00
	8.00	--	--	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	--	--	--	--	--
	9.00	--	--	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	--	--	--	0.00

Supplemental Table 5: SDS-PAGE and anti-DLST Western Blot lane annotations.

Lane	Annotation
1	Bio-Rad Precision Plus Protein Unstained Standards
2	LiCor Chameleon® 700 Pre-stained Protein Ladder
3	1:20 dilution nanoparticle component 1 soluble lysate
4	1:10 dilution nanoparticle component 1 fraction loaded on IMAC
5	1:5 dilution nanoparticle component 1 IMAC flowthrough
6	Nanoparticle component 1 IMAC wash
7	Nanoparticle component 1 IMAC eluate
8	Bio-Rad Precision Plus Protein Unstained Standards
9	LiCor Chameleon® 700 Pre-stained Protein Ladder
10	1:20 dilution nanoparticle component 2 fraction loaded on IMAC
11	1:20 dilution nanoparticle component 2 IMAC flowthrough
12	Nanoparticle component 2 IMAC wash
13	Nanoparticle component 2 IMAC eluate
14	Bio-Rad Precision Plus Protein Unstained Standards
15	LiCor Chameleon® 700 Pre-stained Protein Ladder
16	1:20 dilution nanoparticle component 1 soluble lysate + excipients
17	1:10 dilution nanoparticle component 1 fraction loaded on IMAC + excipients
18	1:5 dilution nanoparticle component 1 IMAC flowthrough + excipients
19	Nanoparticle component 1 IMAC wash + excipients
20	Nanoparticle component 1 IMAC eluate + excipients
21	Bio-Rad Precision Plus Protein Unstained Standards
22	LiCor Chameleon® 700 Pre-stained Protein Ladder
23	1:20 dilution nanoparticle component 2 fraction loaded on IMAC + excipients
24	1:20 dilution nanoparticle component 2 IMAC flowthrough + excipients
25	Nanoparticle component 2 IMAC wash + excipients
26	Nanoparticle component 2 IMAC eluate + excipients

Additional Supplemental Sequences (4) and Tables (6-13) are included as additional supplemental material.

Materials and Methods

Purification protocol and protocol modifications for O43-TM

For the designed O43-B, *E. coli* BL21(DE3) cells (NEB, Cat. C2527H) were transformed with the trimer plasmid. A colony was picked and cultures were grown in Terrific Broth media at 37.0°C until an OD600 of 0.60 - 0.80 before induction with a final concentration of 1 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG) and grown for an additional 3 hours at 37.0°C. Cells were harvested at 5000 revolutions per minute (rpm) for 15 minutes and resuspended with the lysis buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M NaCl, 100 mM arginine, 100 mM glycine, 1.00 mg/mL lysozyme and 1.00 mg/mL DNAse) then lysed using sonication. Lysed cell resuspensions were sonicated for a total of 5 minutes using 10.0 second sonication pulses and 10.0 second breaks. After cell lysis, the solution was pelleted at 10,000*g* for 30 minutes. The supernatant was collected and ran through a nickel-nitrilotriacetic acid (Ni-NTA) IMAC resin with a bed volume of 5.00 mL (Qiagen, Cat. 30250). The IMAC resin was equilibrated with 5.00 column volumes of loading buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M NaCl, 100 mM arginine and 100 mM glycine) before the addition of cell lysate, then washed with 3 column volumes of the loading buffer. The bound protein was eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM KCl, 0.50 M imidazole, 0.15% n-Dodecyl-B-D-maltoside (DDM), 100 mM arginine and 100 mM glycine).

For O43-A, plasmids were transformed into BL21 Star (DE3) pLysS One Shot competent cells (Promega, Cat. L1195). Cells were grown for 3 hours at 37.0°C until an OD600 of 0.80 - 1.00 before induction with a final concentration of 1 mM Isopropyl β- d-1-thiogalactopyranoside (IPTG) and grown for an additional 18 hours at 18.0°C. The cells were then lysed in the same fashion as described above. Following lysis by sonication, cells expressing the tetramer were spun at 10,000*g* for 15 minutes and the supernatant was collected for ultracentrifugation at 130,000*g* for 1.00 hour at 4.0°C. A resuspension buffer (25.0 mM Tris-HCl pH 8.00, 0.15 M KCl, 100 mM arginine, 100 mM glycine and 2.00% DDM) was used to resuspend the pellet after ultracentrifugation. This resuspension was placed on a rocker at 4.0°C overnight. The following day, the resuspended solution was spun for an additional 30 minutes at 14,000*g*. The supernatant was collected and ran through an Ni-NTA IMAC resin with a bed volume of 5.00 mL. The IMAC resin was equilibrated with 5.00 column volumes of resuspension buffer before the addition of cell lysate, then washed with 3 column volumes of the resuspension buffer and eluted with 1.50 column volumes of elution buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM KCl, 0.50 M imidazole, 0.15% n-Dodecyl-B-D-maltoside (DDM), 100 mM arginine and 100 mM glycine).

Purified proteins were mixed together at a 1:1 (c/c) by molar ratio for 1 hour at 4.0°C to facilitate nanoparticle assembly. Further purification of the nanoparticles was achieved using size exclusion chromatography (SEC) in an SEC buffer (25.0 mM Tris-HCl pH 8.00, 0.15 mM NaCl, 0.10% DDM, 100 mM arginine and 100 mM glycine) on a Cytiva Superose® 6 Increase 10/300 GL column. The SEC column was calibrated using the Cytiva Gel Filtration HMW Calibration Kit (Cat. 28403842) with molecular masses ranging from 43,000 to 669,000 Daltons. The nanoparticle peak purified on the SEC was isolated and characterized by DLS, ns-EM and Cryo-EM.

All lysis, resuspension and purification buffers resulting in the co-purification of DLST with the on-target nanoparticle did not contain 100 mM glycine and 100 mM arginine. Otherwise, the purification protocols described here are the same for samples where co-purification of DLST with our two-component nanoparticle were observed.

Western Blot of IMAC-purified components containing DLST

After each step of Ni-NTA purification of the protein nanoparticle components with and without the use of excipients, 500 µL of sample was saved for SDS-PAGE gel on a Criterion™ TGX Stain-Free™ Precast Any kD gel (Cat. 5678125). Soluble protein lysate and NTA flowthrough fractions were diluted to prevent overloading the sample signal on the gel. The NTA Wash and Elution fractions of the protein components were run on the gel without dilution. 10 µL of each sample was mixed with 10 µL of 2x Laemmli Sample Buffer (Brio-Rad, Cat. 1610737) and boiled for 10 minutes at 95°C. The gel was loaded with 10 µL Bio-Rad Precision Plus Protein Unstained Standards (Bio-Rad, Cat. 1610363) ladder for SDS-PAGE applications, 4 µL of LiCor Chameleon® 700 Pre-stained Protein Ladder (LiCor, Cat. 928-70000) for Western Blot applications, and 10 µL of each boiled sample.

A Western Blot was performed following stain-free imaging of the SDS-PAGE gel. The gel was transferred onto a 0.2 µm Nitrocellulose membrane (Bio-Rad, Cat. 1620112) and blocked for 1 hour using a 1% solution of Blocking Buffer (1% w/v Blocking-Grade Blocker (Bio-Rad, Cat. 1706404) dissolved in 25 mM Tris-HCl pH 8.0, 150 mM NaCl and 0.05% Tween-20 (TBS-T)). A polyclonal C-terminal DLST antibody (Antibodies Online, Cat. ABIN2788818) was used as the primary antibody to detect DLST in the samples. The primary antibody was incubated with the membrane at a 1:100 dilution in the blocking buffer for 1 hour. The membrane was washed three times with TBS-T and 25 mL of new blocking buffer was added with a 1:1000 dilution of the secondary antibody Goat anti-Rabbit IgG FITC (Antibodies Online, Cat. ABIN101988) to visualize DLST. Western Blots were imaged on the LiCor Odyssey® M using the 700 nm wavelength channel to visualize the ladder and 488 nm channel to detect the FITC fluorescence signal.

ns-EM data acquisition and processing for samples containing DLST

Negative stain electron microscopy (ns-EM) micrographs were collected on a Thermo Fisher Scientific Talos L120C 120 kV transmission electron microscope with a LaB_6 filament and CETA camera with a pixel size of 2.49 Å at 57,000x magnification. 164 micrographs were recorded at 57,000x magnification at a total dose rate 40.8 $e^-/Å^2$ for the octahedral nanoparticle. Micrographs were imported into CryoSPARC v4.4 and Patch CTF Estimation was performed prior to blob picking, inspection, and extraction of 22,158 picks with a box size of 180 pixels (448.2 Å)[56]. With CTF correction off, picks were classified into 50 2D classes, four of which were selected to serve as templates for an additional round of particle picking. In this second round of particle picking, picks were inspected and extracted to 180 pixels (448.2 Å) before classification of 55,802 extracted particles into 150 2D classes with CTF correction off. 5 out of 150 of these 2D classes were selected as a second round of templates for another round of particle picking and extraction. In this last round, 49,262 particles were extracted to 180 pixels (448.2 Å) using the templates from the second round of extraction and classified into 50 2D classes with a batch size of 400 particles and CTF correction off. 30,927 particles were used to homogeneously refine a map into Octahedral symmetry with a maximum resolution threshold of 20 Å. Particles were aligned into three 3D octahedral ab-initio reconstruction classes. For subsequent ns-EM data processing for DLST, we followed a similar pipeline outlined here.

Cryo-EM sample preparation for samples containing DLST

Prior to freezing, 3 µL of protein sample, at an estimated final concentration of 0.3 mg/mL, was applied to glow-discharged Quantifoil 2/2 holey carbon grids overlaid with a thin layer of additional carbon. Vitrification was performed using a Vitrobot MkIV at 22°C and 100% humidity, with a wait time of 7.5 seconds, a blot time of 0.5 seconds, and a blot force of 0, before being immediately plunge-frozen into liquid ethane. The sample grids were clipped following standard protocols before being loaded into a 300 kV Titan Krios for imaging.

Cryo-EM data acquisition and processing for samples containing DLST

A total of 6,211 movies were collected using SerialEM on a 300 kV FEI Titan Krios equipped with a Gatan K3 direct electron detector and a BioQuantum Gif energy filter[94]. The movies were recorded at 105,000x magnification, with a pixel size of 0.843 Å/pixel. Each movie consisted of 99 frames recorded at a frame rate of 19.8 frames per second, with a dose rate of 9.4 e-/Å^2/s and an exposure time of 5 seconds, resulting in a total exposure dose of 47.0 e-/Å^2. All movies were imported into CryoSPARC v4.4 for data processing[56]. Initially, the movies were corrected for Patch Motion Correction. Defocus and astigmatism values were estimated using the Patch CTF with the default parameters. After two rounds of exposure curation, 4,264 exposures were selected for particle picking. Prospective particles were identified using blob picking with a diameter range of 200 to 400 Å, corresponding to the expected sizes of both our intended and co-purified proteins. This process identified 808,774 prospective particles, which were then inspected and curated based on power histogram values and normalized cross-correlation scores. Following this curation, 456,319 particles were extracted at a box size of 600 pixels. These particles were classified into 150 classes with a batch size of 400 per class. 4 classes were selected to serve as templates for another round of particle picking with particle diameter set to 300 Å. This round yielded 1,468,115 prospective particles, which were curated down to 765,699 particles. These were extracted at 600 pixels and classified into 150 classes again. All high-quality particle classes (47,980 particles) were re-extracted with recentering enabled, resulting in 47,480 particles, which were then split into 150 classes, yielding 39,222 high-quality particles. Of these, 19,033 of the unknown protein contaminant particles were reconstructed into 3 ab-initio 3D C1 reconstruction classes. All particles were homogeneously refined into the best Ab-initio volume with octahedral symmetry enabled, achieving a resolution of 2.62 Å. Local CTF refinement was performed, followed by another round of homogeneous refinement, improving the resolution to 2.59 Å. Reference Based Motion Correction was then performed which reduced the number of particles to 18,809. Subsequently, Global CTF refinement was run for two iterations with Fit Anisotropic Mag set to true. One final homogenous refinement was run in Octahedral symmetry with Minimize over per-particle scale and Optimize per-group CTF parameters set to true, yielding a map with a final estimated resolution of 2.51 Å. This map was then sharpened using deepEMhancer on the “highRes” model and filtered to 2.51 Å. All other parameters were left as default.

Asymmetric refinement of the catalytic domain for DLST

To produce an asymmetric model of the catalytic domain, best ab-initio volume was also homogeneously refined using the same 19,033 particles used for octahedral refinement, but without a symmetry operator enforced. The 3.79 Å homogenous refinement was sharpened using DeepEMhancer to produce our final asymmetric map.

Identification of DLST from octahedral nanoparticle samples

ModelAngelo was run on the final Cryo-EM map, sharpened using deepEMhancer, in a sequence-agnostic manner to generate sequence segments fit to the density as a cif file[75,95]. PyMol was used to generate a FASTA sequence from the cif file which was aligned by the multiple sequence alignment program Clustal

Omega[79]. The aligned sequences were then exported and used by sequenced logo program WebLogo to generate a frequency plot of the aligned sequences[96]. The most frequent amino acid present at each residue was used to create a consensus sequence. This sequence was input into the National Institutes of Health (NIH)'s Protein BLAST and the result strongly indicated that the most likely candidate protein was dihydrolipoamide succinyltransferase (**Supplemental Table 4**)[76]. PDB assembly 1SCZ of the catalytic domain of dihydrolipoamide succinyltransferase was aligned to our Cryo-EM volume and used to visually confirm the identity of the contaminant[82,84].

Generation of low-pass filtered maps for resolution-based analysis

To generate maps of the contaminant with resolutions ranging from 3.00 to 9.00 Å, the 2.51 Å Cryo-EM map (prior to sharpening with deepEMhancer) was low-pass filtered using the Volume Tools job in CryoSPARC to 3.00 Å, 8.00 Å, 9.00 Å, and each 0.25 Å interval between 4.00 Å and 7.00 Å. The maps output by these jobs were then used as inputs for ModelAngelo without an input sequence to produce models containing chain fragments and corresponding HMM profiles, and from which consensus sequences could be generated using the protocol designed for the 2.51 Å map. Protein BLAST analysis of the consensus sequences for each model.

Each consensus sequence was individually used as the input for a Protein BLAST search on the non-redundant protein sequences (nr) database with default parameters and with no taxonomic restriction applied[76]. The descriptions table for each result was downloaded and trimmed to the 100 results with the lowest E value. The protein names (from the descriptions column) were then searched for one of the key terms identified for unambiguous DLST identification. All results were formatted in lowercase, then screened to identify the presence of any of the following search terms: "dihydrolipoamide succinyltransferase", "dihydrolipoyllysine-residue, succinyltransferase", "dihydrolipoyltranssuccinylase", "dihydrolipoyltranssuccinate transferase", or "dihydrolipoyltranssuccinase". To prevent potential ambiguity or mis-identification in results, terms such as "2-oxo acid dehydrogenase subunit E2" were not considered matches unless they also contained one of the above terms. A percent DLST match score was assigned to each chain from each method representing the number of the top 100 results that could be unambiguously identified as DLST. For results which contained fewer than 100 entries, the percent DLST match was determined based on all of the entries available. For example, the Protein BLAST result for chain 2 of the 5.50 Å model was assigned a score of 93.3 as 14 of the 15 entries returned matched DLST (**Supplemental Table 4**). Because results with no entries would have an indivisible score (as the non-existent matching entries would have to be divided by zero total entries), those results are scored with a double dash (--), and were excluded from **Figure 9C** (**Supplemental Table 4**).

Protein BLAST analysis for the longest 10 chains of each model

The FASTA sequence of each model output by ModelAngelo was sorted by chain length, and the sequences of the longest 10 chains were recorded for subsequent analysis. Each sequence was then used as the input for a novel Protein BLAST search and assigned a subsequent DLST match score using the same methods and parameters that were used for the consensus sequences (**Supplemental Table 4**).

hmmsearch analysis (in ModelAngelo) for each model

The hmmsearch function within ModelAngelo was executed on each model referencing *E. coli (BL21 DE3)* reference proteome UP000002032[75]. The all_hits.csv results were - like with searches using Protein BLAST - trimmed to the 100 entries with the lowest E value, or all entries for results with less than 100 entries and DLST match scores were assigned using the same criteria as for Protein BLAST.

hmmsearch analysis (in HMMER 3.4) for the longest 10 chains of each model

Using the chain names of the longest 10 chains for each model identified during Protein BLAST analysis, the corresponding HMM profiles generated by ModelAngelo for the longest 10 chains of each model were identified for use with hmmsearch directly through HMMER 3.4 against sequence database UniProtKB/TrEMBL[95,97]. The output files were then processed in the same manner as the Protein BLAST and ModelAngelo hmmsearch outputs.

Cryo-EM Structural Refinement

Following identification of the contaminant protein as DLST, the *dlst* gene sequence from BL21 DE3 *E. coli* was pulled from UniProt (entry A0A140NDX4) and used as the input sequence for automated model building using ModelAngelo along with the 2.51 Å deepEMhancer sharpened volume map[67,69,97]. The resulting model was imported into PyMOL to extract its FASTA sequence, the chains of which were then aligned with the UniProt gene sequence once more to verify that deviations to the sequence were not introduced[53,98]. Then the model was relaxed using ISOLDE in ChimeraX to ensure the model fit moderately well to the density[67–69,99]. Subsequently, one chain of the model was isolated and retained while all other copies were deleted. Symmetry operators were applied to the remaining chain in ChimeraX to repopulate the density with the 23 additional subunits[67]. The model was then iteratively processed using Coot, Phenix, and ISOLDE to accurately orient the backbone and sidechains[69–71]. After each iteration of processing, the A chain was preserved and other chains were first deleted, and then repopulated using ChimeraX symmetry operators[67,68]. Periodically, between iterations, the model was submitted to the wwPDB validation service to provide an additional metric of model quality[77]. Once the iteratively produced model was deemed by the authors to be of sufficient quality, a mask was generated in ChimeraX of the 2.51 Å map sharpened (without using deepEMhancer) preserving only regions within 6 Å of chain A. A network of waters with density present in this mask was built in Coot and then symmetrized over the whole model in ChimeraX[67,68,70]. The symmetrized model was examined in ChimeraX and Coot to identify regions where waters were duplicated (as waters were built along all interface between chain A and the adjacent chains), all but one copy of each of these waters were deleted such that following an additional round of symmetrization the final model possessed no duplicates. The full model was refined one last time in Phenix using the 2.51 Å map sharpened (without using deepEMhancer) map[71]. The distance between each water and nearby non-Hydrogen atoms was measured to verify that the waters were not placed too near or far from the model. The model was manually inspected in Chimera as a final quality check and then the final model was once again submitted to the wwPDB validation service to provide a quantitative assessment of model quality prior to deposition[77,99]. The final structure was deposited in the Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB) under accession numbers 9DZ8 and EMD-47326 respectively.

Chapter 5: Conclusions

In this work, I present the development and application of advanced computational strategies to design transmembrane nanoparticles and membrane mimetic nanoparticles, encompassing both the preliminary design of circular toroid protein membrane mimetics and the first generation of transmembrane two-component nanoparticles.

A central achievement of this dissertation was the successful purification and structural characterization of the O43-TM particle, a novel protein material using *de novo* designed transmembrane proteins as one of the base components of the nanoparticle. Building upon the methodologies established for two-component protein design, similar approaches were applied to generate designs for the O43-TM nanoparticles[8,11,12]. Despite the challenges associated with the purification of a *de novo* designed transmembrane protein, I successfully isolated O43-A and proceeded to elucidate a Cryo-EM structure of the transmembrane nanoparticle solubilized in detergent with a global resolution of 4.16 Å. Our structure of O43-TM clearly shows detergent micelles surrounding the transmembrane helices of the O43-A component, an exciting glimpse into what the hybrid particle can look like. However, during the first rounds of purification I observed a distinct protein contaminant present in our samples. The integration of machine learning and AI tools such as AlphaFold 3 and ModelAngelo facilitated the improved characterization of O43-TM by allowing us to identify the contaminant and improve the purification and yield[64]. To our knowledge, this transmembrane nanoparticle is the first of its kind and an exciting first iteration towards achieving *in vitro* nanoparticle membrane encapsulation which will open the door to display native membrane proteins on nanoparticle surfaces.

I also used computational design to build novel materials from unique building blocks. These toroid building blocks were encoded with hydrophobic regions that would be localized areas for lipid-protein interactions. Beyond designing O43-TM, this work also focused on the rational design of toroid proteins that can incorporate lipids into the nanoparticle scaffold by coding specific motifs into the protein’s design. Besides being able to modify the height of toroids, I observed that introducing hydrophobic mutations into the protein was not a detriment to protein expression. The ease of computational design and purification of toroid proteins makes them an ideal candidate for making a *de novo* designed membrane mimetic. Furthermore, because toroids are rigid repeat proteins it is possible to multimerize toroids into spherical nanoparticles and leverage multivalent display of lipid patches[45–47]. Similarly to encapsulated nanoparticles, the presentation of lipid patches through toroids can provide a wider breadth for the types of antigens or drug conjugates that can be displayed on designed nanoparticles.

Taken together, this dissertation illustrates a step towards utilizing unique building blocks in oligomeric protein design. This opens the door for designing nanoparticles which can be encapsulated by lipid bilayers and would be an ideal platform for displaying native antigens, creating a membrane surface for targeting, and perhaps increase bioavailability much like the advantages of LNPs[22,25,57]. Although many steps remain to validate a toroid protein membrane mimetic, this work lays a foundation for generating new designed scaffolds suitable for holding lipid bilayers. As computational protein design continues to evolve, the design of hybrid nanoparticles will become increasingly more attainable. Experimental evaluation emerges as a key paradigm for the next generation of hybrid nanoparticle design. These findings underscore the need for further structural characterization, membrane formulation investigation, and assay development to refine the design and characterization of true hybrid lipid-protein nanoparticles.

References

1. Zhen, Z. *et al.* RGD-modified apoferritin nanoparticles for efficient drug delivery to tumors. *ACS Nano* **7**, 4830–4837 (2013).
2. Kanekiyo, M. *et al.* Self-assembling influenza nanoparticle vaccines elicit broadly neutralizing H1N1 antibodies. *Nature* **499**, 102–106 (2013).
3. Powell, A. E. *et al.* A single immunization with spike-functionalized ferritin vaccines elicits neutralizing antibody responses against SARS-CoV-2 in mice. *ACS Cent. Sci.* **7**, 183–199 (2021).
4. Rodrigues, M. Q., Alves, P. M. & Roldão, A. Functionalizing ferritin nanoparticles for vaccine development. *Pharmaceutics* **13**, 1621 (2021).
5. Lai, Y.-T., King, N. P. & Yeates, T. O. Principles for designing ordered protein assemblies. *Trends Cell Biol.* **22**, 653–661 (2012).
6. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2217–2221 (2001).
7. Zhang, S. Fabrication of novel biomaterials through molecular self-assembly. *Nat. Biotechnol.* **21**, 1171–1178 (2003).
8. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
9. Chao, C. W. *et al.* Protein nanoparticle vaccines induce potent neutralizing antibody responses against MERS-CoV. *Immunology* (2024).
10. Wargacki, A. J. *et al.* Complete and cooperative in vitro assembly of computationally designed self-assembling protein nanomaterials. *Nat. Commun.* **12**, 883 (2021).
11. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012).
12. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein dodecahedron. [corrected]. *Nature* **535**, 136–139 (2016).
13. de Haas, R. J. *et al.* Rapid and automated design of two-component protein nanomaterials using ProteinMPNN. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2314646121 (2024).
14. Yang, E. C. *et al.* Computational design of non-porous pH-responsive antibody nanoparticles. *Nat.*

Struct. Mol. Biol. 1–9 (2024) doi:10.1038/s41594-024-01288-5.

15. Ueda, G. *et al.* Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens. *Elife* **9**, e57659 (2020).
16. Khmelinskaia, A., Wargacki, A. & King, N. P. Structure-based design of novel polyhedral protein nanomaterials. *Curr. Opin. Microbiol.* **61**, 51–57 (2021).
17. Wang, J. Y. J. *et al.* Improving the secretion of designed protein assemblies through negative design of cryptic transmembrane domains. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2214556120 (2023).
18. Ellis, D. *et al.* Stabilization of the SARS-CoV-2 Spike receptor-binding domain using deep mutational scanning and structure-based design. *Front. Immunol.* **12**, 710263 (2021).
19. Walls, A. C. *et al.* Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382.e17 (2020).
20. Butterfield, G. L. *et al.* Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
21. Hendricks, G. G. *et al.* Computationally designed mRNA-launched protein nanoparticle immunogens elicit protective antibody and T cell responses in mice. *Sci. Transl. Med.* **17**, eadu2085 (2025).
22. Hou, X., Zaks, T., Langer, R. & Dong, Y. Lipid nanoparticles for mRNA delivery. *Nat. Rev. Mater.* **6**, 1078–1094 (2021).
23. Semple, S. C. *et al.* Rational design of cationic lipids for siRNA delivery. *Nat. Biotechnol.* **28**, 172–176 (2010).
24. Pardi, N. *et al.* Zika virus protection by a single low-dose nucleoside-modified mRNA vaccination. *Nature* **543**, 248–251 (2017).
25. Mehta, M. *et al.* Lipid-based nanoparticles for drug/gene delivery: An overview of the production techniques and difficulties encountered in their industrial development. *ACS Mater. Au* **3**, 600–619 (2023).
26. Shah, S., Dhawan, V., Holm, R., Nagarsenker, M. S. & Perrie, Y. Liposomes: Advancements and innovation in the manufacturing process. *Adv. Drug Deliv. Rev.* **154-155**, 102–122 (2020).
27. Kowalski, P. S., Rudra, A., Miao, L. & Anderson, D. G. Delivering the messenger: Advances in technologies for therapeutic mRNA delivery. *Mol. Ther.* **27**, 710–728 (2019).

28. Chen, K. *et al.* MRNA vaccines against SARS-CoV-2 variants delivered by lipid nanoparticles based on novel ionizable lipids. *Adv. Funct. Mater.* **32**, 2204692 (2022).

29. Tenchov, R., Bird, R., Curtze, A. E. & Zhou, Q. Lipid Nanoparticles─From liposomes to mRNA vaccine delivery, a landscape of research diversity and advancement. *ACS Nano* **15**, 16982–17015 (2021).

30. Swetha, K. *et al.* Recent advances in the lipid nanoparticle-mediated delivery of mRNA vaccines. *Vaccines (Basel)* **11**, 658 (2023).

31. Haley, R. M. *et al.* Lipid nanoparticle delivery of small proteins for potent in vivo RAS inhibition. *ACS Appl. Mater. Interfaces* **15**, 21877–21892 (2023).

32. Chelliah, R. *et al.* Liposomes for drug delivery: Classification, therapeutic applications, and limitations. *Next Nanotechnology* **8**, 100209 (2025).

33. Keener, J. E. *et al.* Chemical additives enable native mass spectrometry measurement of membrane protein oligomeric state within intact nanodiscs. *J. Am. Chem. Soc.* **141**, 1054–1061 (2019).

34. Gault, J. *et al.* High-resolution mass spectrometry of small molecules bound to membrane proteins. *Nat. Methods* **13**, 333–336 (2016).

35. Denisov, I. G. & Sligar, S. G. Nanodiscs for structural and functional studies of membrane proteins. *Nat. Struct. Mol. Biol.* **23**, 481–486 (2016).

36. Postis, V. *et al.* The use of SMALPs as a novel membrane protein scaffold for structure study by negative stain electron microscopy. *Biochim. Biophys. Acta* **1848**, 496–501 (2015).

37. Hiotis, G., Notti, R. Q., Bao, H. & Walz, T. Nanodiscs remain indispensable for Cryo-EM studies of membrane proteins. *Curr. Opin. Struct. Biol.* **92**, 103042 (2025).

38. Notti, R. Q. & Walz, T. Native-like environments afford novel mechanistic insights into membrane proteins. *Trends Biochem. Sci.* **47**, 561–569 (2022).

39. Bayburt, T. H., Grinkova, Y. V. & Sligar, S. G. Self-assembly of discoidal phospholipid bilayer nanoparticles with membrane scaffold proteins. *Nano Lett.* **2**, 853–856 (2002).

40. Guo, Z. *et al.* Cancer cell membrane nanodiscs for antitumor vaccination. *Nano Lett.* **23**, 7941–7949 (2023).

41. Kuai, R., Ochyl, L. J., Bahjat, K. S., Schwendeman, A. & Moon, J. J. Designer vaccine nanodiscs for

personalized cancer immunotherapy. *Nat. Mater.* **16**, 489–496 (2017).

42. Noh, I. *et al.* Cellular nanodiscs made from bacterial outer membrane as a platform for antibacterial vaccination. *ACS Nano* **17**, 1120–1127 (2022).
43. Dalal, V. *et al.* Lipid nanodisc scaffold and size alter the structure of a pentameric ligand-gated ion channel. *Nat. Commun.* **15**, 25 (2024).
44. Rankovic, S. *et al.* Computational design of bifaceted protein nanomaterials. *Nat. Mater.* **24**, 1635–1643 (2025).
45. Lee, S. *et al.* Four-component protein nanocages designed by programmed symmetry breaking. *Nature* **638**, 546–552 (2025).
46. Huddy, T. F. *et al.* Blueprinting extendable nanomaterials with standardized protein blocks. *Nature* **627**, 898–904 (2024).
47. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528**, 585–588 (2015).
48. Mitra, K., Ubarretxena-Belandia, I., Taguchi, T., Warren, G. & Engelman, D. M. Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4083–4088 (2004).
49. Yesylevskyy, S. O., Rivel, T. & Ramseyer, C. The influence of curvature on the properties of the plasma membrane. Insights from atomistic molecular dynamics simulations. *Sci. Rep.* **7**, 16078 (2017).
50. Perrault, S. D. & Shih, W. M. Virus-inspired membrane encapsulation of DNA nanostructures to achieve in vivo stability. *ACS Nano* **8**, 5132–5140 (2014).
51. Hagn, F., Nasr, M. L. & Wagner, G. Assembly of phospholipid nanodiscs of controlled size for structural studies of membrane proteins by NMR. *Nat. Protoc.* **13**, 79–98 (2018).
52. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
53. PyMOL. http://www.pymol.org/pymol.
54. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

55. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
56. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
57. Votteler, J. *et al.* Designed proteins induce the formation of nanocage-containing extracellular vesicles. *Nature* **540**, 292–295 (2016).
58. Brouwer, P. J. M. *et al.* Enhancing and shaping the immunogenicity of native-like HIV-1 envelope trimers with a two-component protein nanoparticle. *Nat. Commun.* **10**, 4272 (2019).
59. 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).
60. Lu, P. *et al.* Accurate computational design of multipass transmembrane proteins. *Science* **359**, 1042–1046 (2018).
61. Xu, C. *et al.* Computational design of transmembrane pores. *Nature* **585**, 129–134 (2020).
62. Liu, Y. *et al.* Bottom-up design of calcium channels from defined selectivity filter geometry. *bioRxivorg* (2024) doi:10.1101/2024.12.19.629320.
63. Li, Y. *et al.* Water, solute, and ion transport in DE Novo-designed membrane protein channels. *ACS Nano* **19**, 2185–2195 (2025).
64. Carr, K. D. *et al.* Protein identification using Cryo-EM and artificial intelligence guides improved sample purification. *J. Struct. Biol. X* **11**, 100120 (2025).
65. Scientific Image and Illustration Software. https://www.biorender.com/.
66. Dosey, A. *et al.* Combinatorial immune refocusing within the influenza hemagglutinin RBD improves cross-neutralizing antibody responses. *Cell Rep.* **42**, 113553 (2023).
67. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
68. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
69. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018).

70. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
71. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).
72. Bolanos-Garcia, V. M. & Davies, O. R. Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochim. Biophys. Acta* **1760**, 1304–1313 (2006).
73. Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21**, 129–228 (1988).
74. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294**, 5181–5197 (2019).
75. Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).
76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
77. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
78. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
79. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
80. Sheu, K. F. & Blass, J. P. The alpha-ketoglutarate dehydrogenase complex. *Ann. N. Y. Acad. Sci.* **893**, 61–78 (1999).
81. Knapp, J. E. *et al.* Expression, purification, and structural analysis of the trimeric form of the catalytic domain of the Escherichia coli dihydrolipoamide succinyltransferase. *Protein Sci.* **9**, 37–48 (2000).
82. Andi, B. *et al.* Structure of the dihydrolipoamide succinyltransferase catalytic domain from Escherichia coli in a novel crystal form: a tale of a common protein crystallization contaminant. *Acta Crystallogr. F Struct. Biol. Commun.* **75**, 616–624 (2019).

83. UniProt. https://www.uniprot.org/uniprotkb/A0A140NDX4/entry.

84. Schormann, N. *et al.* Improved structural model for the catalytic domain of E.coli dihydrolipoamide succinyltransferase. Preprint at https://doi.org/10.2210/pdb1scz/pdb (2004).

85. Coskun, O. Separation techniques: Chromatography. *North. Clin. Istanb.* **3**, 156–160 (2016).

86. Bornhorst, J. A. & Falke, J. J. [16] Purification of proteins using polyhistidine affinity tags. in *Methods in Enzymology* 245–254 (Elsevier, 2000). doi:10.1016/s0076-6879(00)26058-8.

87. Jorgensen, L., Hostrup, S., Moeller, E. H. & Grohganz, H. Recent trends in stabilising peptides and proteins in pharmaceutical formulation - considerations in the choice of excipients. *Expert Opin. Drug Deliv.* **6**, 1219–1230 (2009).

88. Ishihara, T. & Hosono, M. Improving impurities clearance by amino acids addition to buffer solutions for chromatographic purifications of monoclonal antibodies. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **995-996**, 107–114 (2015).

89. Falconer, R. J., Chan, C., Hughes, K. & Munro, T. P. Stabilization of a monoclonal antibody during purification and formulation by addition of basic amino acid excipients. *J. Chem. Technol. Biotechnol.* **86**, 942–948 (2011).

90. Arakawa, T. *et al.* Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. *Biophys. Chem.* **127**, 1–8 (2007).

91. Suzuki, H., Kawabata, T. & Nakamura, H. Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDB. *Bioinformatics* **32**, 619–620 (2016).

92. Skalidis, I. *et al.* Cryo-EM and artificial intelligence visualize endogenous protein community members. *Structure* **30**, 575–589.e6 (2022).

93. Loo, R. R. O., Dales, N. & Andrews, P. C. The effect of detergents on proteins analyzed by electrospray ionization. in *Protein and Peptide Analysis by Mass Spectrometry* 141–160 (Humana Press, New Jersey, 2003). doi:10.1385/0-89603-345-7:141.

94. Mastronarde, D. N. SerialEM: A program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position. *Microsc. Microanal.* **9**, 1182–1183 (2003).

95. Sanchez-Garcia, R. *et al.* DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).

96. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

97. Eddy, S. R. hmmer.org. *HMMER: biosequence analysis using profile hidden Markov models* (2007).

98. Goldstein, T., Studer, C. & Baraniuk, R. FASTA: A generalized implementation of forward-backward splitting. *arXiv [cs.MS]* (2015).

99. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).