Scalable geometrically designed protein cages assembled via genetically encoded split inteins.

James N. Wright[1,3], Wan Ling Wong[1,3], Joseph A. Harvey[1], James A. Garnett[1], Laura S. Itzhaki[2] & Ewan R.G. Main[1,4]*

[1]School of Biological and Chemical Sciences
Queen Mary, University of London,
Mile End Road
London E1 4NS, U.K

[2]Department of Pharmacology
University of Cambridge
Tennis Court Road
Cambridge CB2 1PD, U.K.

[3]These authors contributed equally
[4]Lead Contact
*Correspondence: e.main@qmul.ac.uk

SUMMARY

Engineering proteins to assemble into user-defined structures is key in their development for biotechnological applications. However, designing generic rather than bespoke solutions is challenging. Here, we describe an expandable recombinant assembly system that produces scalable protein cages via split intein-mediated native chemical ligation. Three types of component are used: two complementary oligomeric “half-cage” protein fusions and an extendable monomeric “linker" fusion. All are composed of modular protein domains chosen to fulfil the required geometries, with two orthogonal pairs of split-intein halves to drive assembly when mixed. This combination enables both one-pot construction of two-component cages and stepwise assembly of larger three-component scalable cages. To illustrate the system’s versatility, trimeric half-cages and linker constructs comprising consensus-designed repeat proteins were ligated in one-pot and stepwise reactions. Under mild conditions rapid high yielding ligations were obtained, from which discrete proteins cages were easily purified and shown to form the desired trigonal bipyramidal structures.

Introduction

Nature has created ordered nanostructures to solve many of the challenges of life at the molecular scale. Examples include bacterial micro-compartments, actin filaments and virus particles. One route by which this order is achieved is the directed self-assembly of discrete protein building blocks into diverse polymeric structures ranging from fibres to networks and encapsulations. Two key features of such systems are: (i) the use of symmetry as a mechanism to both order specific assembly through the exquisite positioning of interacting interfaces and reduce the number of different building blocks required and (ii) a specific driving force to dock the protein building blocks at the proposed interfaces. By imitating and re-engineering these fundamental design elements synthetic biologists have been able to develop several methodologies to manipulate existing systems or design novel ones (examples include: (Bale et al., 2016; Banerjee and Howarth, 2017; Brodin et al., 2012; Brune and Howarth, 2018; Capito et al., 2008; Cortajarena et al., 2010; Giessen and Silver, 2016; Glover et al., 2016; Grove et al., 2010; Inostroza-Brito et al., 2015; Lee et al., 2018; Modica et al., 2018; Patterson et al., 2014; Phillips et al., 2012; Sawyer et al., 2013; Veggiani et al., 2016)). In particular, there has been much interest in the design of protein cages/encapsulations due to potential applications in areas as diverse as drug delivery and the compartmentalisation of enzymes to form novel micro-reactors. To this end, a number of system specific and/or computationally intensive solutions have been engineered. For example, protein cages have been assembled from a fusion of homodimeric and heterodimeric protein domains (Lai et al., 2012; Padilla et al., 2001), the rational and computational design of coiled-coil building blocks (Fletcher et al., 2013; Gradisar et al., 2013) and computationally designed from protein domains selected from the Protein Database (PDB) (Bale et al., 2016; King et al., 2012).

To create a more generic genetically encoded system of self-assembly that does not rely on the complicated modelling of protein-protein interfaces, we have previously designed Mxe GyrA intein-based fusion systems (Harvey et al., 2018; Phillips et al., 2012). In Nature, inteins are usually found in the middle of genes and post-translationally self-catalyse their excision and ligation of flanking polypeptides regions with a traceless irreversible peptide bond (via native chemical ligation). In contrast,

our system used inteins in a similar manner to a protecting group in solid state synthesis. The intein was placed at the C-terminus of the protein to be ligated (POI) and, when required, removed via induced self-excision to reveal a reaction-ready C-terminal thioester. To obtain directional self-assembly we designed two fusion proteins (component one and two). Component one contained the POI with a C-terminal intein, whereas component two contained the POI sandwiched between an N-terminal cysteine and a C-terminal intein. All fusion proteins are initially inert. Upon activation and subsequent mixing, the C-terminal thioester of component one spontaneously reacts with the N-terminal cysteine of component two resulting in their ligation (Harvey et al., 2018; Phillips et al., 2012).

Recent studies have described faster and higher yielding natural and engineered "split" intein variants (Aranko et al., 2014; Carvajal-Vallejos et al., 2012; Debelouchina and Muir, 2017; Stevens et al., 2016; Stevens et al., 2017; Thiel et al., 2014). These are intein domains that are divided into two separate polypeptide chains. The split-intein sequences are inert until complementary halves are combined, where upon they spontaneously fold together to produce an active intein. The newly-reactive intein self-catalyses its excision and ligates the two separate peptide chains together. The high affinity of each half intein for its partner results in high reaction rates and yields. Although the reaction does leave a short insertion of approximately 10 amino acids. Split inteins have been used with success for protein labelling, site-specific protein modification, protein cyclisation and linear protein semi-synthesis linking up to three modules (Aranko et al., 2014; Busche et al., 2009; Debelouchina and Muir, 2017; Demonte et al., 2015; Shah et al., 2011; Shi and Muir, 2005; Thiel et al., 2014; Vila-Perello et al., 2013). However, the use of split inteins to produce higher order linear assemblies has been hampered by the difficulties in identifying suitable split intein pairs that are both highly reactive and have high solubility, but are not cross reactive. For example, even the studies that linked three protein modules together with two ligations showed relatively slow reaction rates and relatively low final ligation yields (Busche et al., 2009; Demonte et al., 2015; Shah et al., 2011; Shi and Muir, 2005). Interestingly, a recent study has identified a number of new naturally occurring split-

intein pairs (for example gp41-1 and IMPDH) that do not cross react and that ligate rapidly and in high yields (Carvajal-Vallejos et al., 2012).

Here we present a modular recombinant protein system for the production of scalable encapsulations/cages. It uses simple symmetry design of protein fusions coupled with a driving force of two of the recently identified and highly reactive orthogonal split-intein pairs (IMPDH/gp41-1). This strategy creates a more general rather than bespoke system, which saves time on system-specific design and computational input. To validate our method and investigate the limits of the system, we constructed fusion proteins consisting of a monomeric linker and trimeric half-cage caps. These were composed of modular oligomeric/monomeric consensus-designed repeat proteins fused to differing split intein halves. All the fusions are initially inert until mixed with their complementary reaction partners, whereupon rapid high yielding ligations were obtained in mild conditions. By mixing differing fusions in a one-pot or stepwise manner both two-component cages and larger three-component scalable cages were produced, respectively. In addition to fully ligated discrete cages, cage formation reactions also generated partially ligated structures and "cross-ligated" extended protein networks. Optimisation of the ligation reactions coupled with a two-step purification enabled high proportions of discrete cages to be easily purified to homogeneity. Significantly, small angle x-ray scattering (SAXS) of the two-component cages showed that they adopt the expected 113 kDa trigonal bipyramidal cage with a central hollow cavity. Moreover, using the three-component system with two pairs of orthogonal split inteins enabled larger extended cages to be successfully assembled and purified using a stepwise process with no cross reactivity.

Results

System design

Our system is based on recombinantly expressing a minimum of three fusion proteins that use two pairs of orthogonal split inteins to drive irreversible assembly (Figure 1a-c & Supplementary Information [S.I.] Table 1). The designs consisted of pairs of complementary oligomeric half-cage "caps" and a monomeric linker. The half-cages comprised of an oligomerisation domain and one half of a split-intein pair (with an affinity tag for purification) sandwiching a rigid domain (Figure 1a & b). The oligomerisation domain specifies the geometry of the half cage by acting as the primary vertex with the sides composed of a rigid functionisable domain. In contrast, the linker fusion contained two orthogonal split intein halves (one on each termini) with the protein to be assembled sandwiched between them and an affinity tag for purification (Figure 1c). All fusions are initially inert, and they only react when mixed with a compatible construct.

Fabrication of cages: Cages can be produced either in a one-pot two-component synthesis or via a sequential and iterative three-component reaction scheme (Figures 1e and 1f, respectively). In the two-component synthesis two compatible half-cages are mixed, whereupon their separate cognate split-inteins halves fold together to form an active intein. The active intein can then catalyse the ligation of the half-cages into the full cage product, whilst concurrently excising themselves (Figure 1e). The iterative sequential approach allows larger cages to be synthesised. Here, a half-cage is mixed with a compatible linker fusion (Figure 1f). The half-cage and linker ligate to produce a larger half-cage-linker with the removal of one split intein pair. Importantly, the half-cage-linker still contains a second split-intein half. The half-cage-linker can then either be iteratively extended through the addition of further linkers or formed into a cage through reaction with a suitable half-cage cap.

Trimeric half-cage caps & linker fusions

To validate our approach and explore its limits (structures formed, reaction speeds, efficiencies and yields), two sets of complementary trimeric half-cage caps and a linker fusion were designed (Figure 1a-c, S.I. Table 1). The half cage caps were composed of: (i) the homotrimer Monofoil-4-P (M4P) domain as the primary vertex (Figure 1d), (ii) consensus-designed tetratricopeptide repeat-containing protein CTPR3ΔS as the sides (Figure 1d), and one half of either (iii) **I**nosine-5'-**m**ono**p**hosphate **d**e**h**ydrogenase-1 [**IMPDH**] or (vi) **Gp 41** DNA Helicase [**gp41**-1] split-inteins fused to their termini. An alpha-helical linker, $(EAAAK)_2$, connects the M4P and CTPRΔS proteins, projecting the CTPRΔS units away from each other and thereby reducing the risk of misfolding. The half-cage caps were engineered such that they would react to form a trigonal bipyramidal caged product (Figure 1e). The linker fusion design was more simple. It consisted of the CTPR3ΔS side sandwiched between orthogonal gp41 and IMPDH split-intein halves (Figure 1c).

Choice of protein domains: M4P was selected as the primary vertex due to its high stability and orientation of its N- and C- termini (Lee and Blaber, 2011) (Figure 1d). These termini are solvent exposed and project out from the core structure in the same direction. Thus, extension from the vertex to form the sides of the cage can be achieved via the fusion of protein domains to either termini. The extensions do not disrupt homo-trimerisation and enable the creation of chimeras with complementary orientations (Figure 1e). CTPR3ΔS was chosen as the cage sides for its high stability, symmetric/rigid structure and the ability to dock onto itself in a linear N-to-C terminal manner (Kajander et al., 2005; Main et al., 2013). These attributes permit predictable scalable cage extension and, combined with a modify-able peptide binding pocket, a future route to functionalisation (Figure 1d). The IMPDH or Gp 41 split inteins were chosen because they have been shown to have fast rates of reaction over a range of pH conditions ($t_{1/2}$ = 8 secs and 4 secs, respectively), to be orthogonal at low concentrations (5 μM), and to have electrostatic distributions sufficiently different to suggest orthogonality should be maintained even at the higher protein concentrations (Carvajal-Vallejos et al., 2012; Dassa et al., 2009). A small, highly soluble and easy-to-

refold Chitin Binding Domain (CBD) was also placed next to each IMPDH split to increase solubility.

All fusion proteins expressed well and could be purified either under native conditions or under denaturing and then refolded with high purities and yields (S.I. Table 2). All were initially inert and showed no degradation or reactivity until their cognate split-intein partner was introduced in a reducing environment. Trimerisation of purified half-cage caps was confirmed by SEC analysis (S.I. Figure 1).

One-pot assembly of half-cage caps to form a two-component square bipyramidal cage

Ligation reactions were initiated by mixing purified cognate half-cages in equimolar concentrations from 1 μM to 200 μM under mild conditions (50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM DTT, 0 - 1 M Urea). Samples were taken at several timepoints over a 24-hour period and analysed by SDS-PAGE (Figure 2a-d, S.I. Figure 2). All reactions at all concentrations were rapid with high yields: gp41-mediated reactions were initially faster than IMPDH but gave lower overall yields. Both split intein-mediated ligations produced $\geq$ 65 % yield within 10 minutes, with the IMPDH reactions reaching $\geq$ 80 % and gp41 $\geq$ 70 % within 3 hours. After 3 hours, all reactions were close to completion with only small additional increases in yield when the reactions were left for 24 hours. Once the reactions reached completion, there was little difference in final yields across the protein concentration range of 1-100 μM for either IMPDH or gp41 ligations. At the higher concentration of 200 μM, the reactions produced some protein precipitation leading to a small reduction in yields. It is interesting to note that, although the yields were very high, we did not achieve the ~95 % values of the previous study (Carvajal-Vallejos et al., 2012). This difference may due to steric hindrance created by the trimeric structures.

Analysis of ligation reaction and purification of fully formed cages: Initially ligation reactions were analysed using size-exclusion chromatography (SEC) & SDS PAGE gels (Figure 2c-f, SI Figure 3a). This analysis showed that all ligations generated, to a greater

or lesser extent, heterogeneous mixture of differently sized proteins. That is, in addition to the expected fully ligated discrete cages and the excised split-inteins; the ligation reactions also produced networks of "cross-spliced" proteins and also partially ligated structures. It is interesting to speculate that the differing ligated products may stem from the irreversibility of the reaction. For example, when A-B-C subunits from one cap react with a-b-c from another, they can link in a correct manner giving a discrete cage (A-a, B-b and C-c). However, they can also react with other half-cages to form networks, produce partially ligated faulty "dead-end" structures (A-a, B-c) or a mixture of both.

To separate the ligation mixture, we employed a two-step process (Figure 2e-j, S.I. Figure 3). The first step separated the fully ligated assemblies from unreacted, partially ligated and excised split inteins via nickel affinity chromatography (Figure 2e-f). This was specifically facilitated by careful positioning of the affinity tags within the fusion proteins. The second step separated fully ligated cages from networks by size using SEC (Figure 2i-j). Both separation steps were highly effective: After the first affinity chromatography step, all ligation reactions contained > 95 % purity of fully ligated assemblies as assayed by SDS PAGE gel and anti-histidine affinity tag western blot (Figure 2e-f & S.I. Figure 3b, respectively). Moreover, when the resultant fully ligated products were subjected to SEC, those obtained using lower protein concentrations gave a better separated monodisperse peak. This peak had an elution volume that was consistent with the 113 kDa molecular weight expected for the cage (Figure 2i-j, S.I. Figure 3c-d). There was also good separation from the larger "cross-ligated" protein networks.

Interestingly, the two-step purification highlighted important differences between reactions mediated by IMPDH and gp41 split inteins. From analysis of the affinity purification SDS PAGE gels it is clear that the gp41-mediated ligations lead to more partial ligation and less cage closure than do the IMPDH-mediated ligations. For example, IMPDH-mediated ligations at lower protein concentrations (1 μM to 50 μM) all produced > 75 % fully ligated product. In contrast, gp41-mediated ligations at lower protein concentrations (1 μM and 10 μM) produced only ~ 50 % fully ligated product.

Thus, the faster ligation speed of the gp41 split inteins hinders discrete cage formation and leads to the production of higher proportions of partially ligated structures and networks.

Solution structures of ligated cages: The purified fully ligated cage structures were characterised using far-UV CD and SEC-SAXS (Size Exclusion Chromatography - Small Angled X-ray Scattering) (Figure 3). The far-UV CD spectra of the ligated cages show that: (i) they are highly alpha-helical, as one would expect from a protein containing 18 CTPR motifs, and, importantly, (ii) have exactly twice the ellipticity at 222 nm as that of the half cage caps that do not contain split-intein domains (Figure 3a). Thus, the ligation reaction has had no effect on the secondary structure of the CTPR proteins (had not caused any local unfolding). Guinier and Kratky plot analysis of the SAXS data confirmed that the purified cages were monodisperse and highly rigid. Moreover, the analysis shows that the cages are non-spherical and elongated with radius of gyration (R_g) of 3.85 nm and a maximum linear particle diameter (D_{max}) of 12.6 nm (S.I. Figure 4a-c, S.I. Table 3). This is in contrast with the SAXS data of the non-ligated half cage caps, whose Kratky plot analysis shows that their structures are highly dynamic (S.I. Figure 4d). Additionally, the molecular weight of the cages obtained from the SAXS data is in close agreement with that calculated from its amino acid sequence (110.5 KDa versus 113 KDa, respectively).

As the SAXS profile of the cage has a number of prominent features it allowed us to determine its shape to a higher resolution via two different approaches: (i) a comparison of the experimental SAXS profile to thirty manually generated atomic models of possible cage conformations using the program *Crysol* (Svergun et al., 1995) and (ii) a SAXS *ab initio* model re-constructed using the program *GASBOR* (Svergun et al., 2001). Excitingly, of our thirty manually generated atomic models, only those that closely resembled the intended designed trigonal bipyramidal structure were found to re-capitulate the SAXS experimental profile. i.e. M4P oligomer at the “primary” vertices with the CTPRs forming three symmetrical cage sides that enclose a central cavity (Data S1, Figure 3b-c). Models that were highly expanded cages or that did not contain a central cavity gave profiles that were very different from the experimental

data. The model that produces the closest fit between experimental and generated SAXS profiles used a continuous CTPR6ΔS as the cage sides (Figure 3c). Here, the CTPR3ΔS modules from the half cages dock upon ligation to form a single CTPR6 superhelix, rather than simply two linked CTPR3 domains like beads on a string. This produces an open cage with apertures of ≈ 35 Å between each side at the widest point and encloses a central cavity of ≈ 70 Å by 55-60 Å. Interestingly, the docked CTPR6 superhelix would also account for the increased rigidity of the cage in contrast to the dynamic half cage caps. The final χ^2 value between the model and experimental SAXS profiles was 1.66 (Figure 3b), with only a small discrepancy at the highest resolution SAXS data (suggesting an ambiguity between the modelled and exact rotation of the CTPR sides and their packing relative to their M4P vertices).

In comparison, the *ab initio GASBOR* modelling gave five solutions from thirty-two calculations that were supported by our biophysical data. Solutions were discarded when, for example, the CTPR/M4P domains would be required to fit protein density envelopes by either adopting non-native conformations or by ligating in a nonsensical formation (discarded examples are shown in S.I. Figure 4e-h). The five biophysically relevant solutions were averaged with *DAMAVER* (Volkov and Svergun, 2003) to produce a final model with excellent fit to the data (χ^2 = 1.06) (Figure 3d-e). Importantly, this final *ab initio* model also shows that the ligated cages form our intended structure with a protein density envelope that closely resembles the designed trigonal bipyramidal structure with a central hollow cavity. Moreover, the model envelope also fits extremely well with our manually generated atomic model (Figure 3f). Combined, the experimental data and the modelling confirm that the ligated cages form an open shell, with a central hollow cavity, that closely resembles the intended designed trigonal bipyramidal structure.

Scalable protein cages: Stepwise synthesis of larger caged structures

Given the high yielding success of the two-component split-intein mediated cage assembly, the next step was to investigate the stepwise enlargement of the cage structures. Using the scheme shown in Figure 1f & 4a, half-cage caps with orthogonal

gp41 and IMPDH split inteins were ligated in series with the linker fusion (a CTPR3ΔS module sandwiched between gp41 and IMPDH splits). Briefly, the first step reacted a half-cage with an excess of linker, followed by purification of fully ligated product via affinity chromatography. The larger cage was then formed by reacting the purified larger half cage in a 1:1 mixture with the second half-cage cap.

Both half-cages were trialled in the first-step ligation to the linker fusion. A 6:1 excess of linker produced the highest yields (Figure 4b & c), with the IMPDH-mediated ligation generating more product than did the gp41-mediated ligation. The 6:1 excess enables the IMPDH reaction to be driven to 90 % completion in 3 hours. High yields from each step are extremely important in any multi-step reaction and particularly so here, given that each trimeric half cage requires each of its three sides to react. Therefore, the IMPDH-mediated ligation was used as the first step. The resultant extended half-cage product was purified by affinity chromatography. Full ligation and purity was confirmed by anti-CBD affinity tag western blot and SDS PAGE gel (S.I. Figure 4i & Figure 4b, lane 9).

The extended half-cage was then further reacted with a gp41-tagged half-cage cap to close the cage (Figure 4d). The ligation yield of the second step is in line with the smaller gp41-mediated cage reactions (60 % after 3 hours). To favour the formation of discrete cages rather than networks, the cage closure was carried out with equimolar concentrations of reactants at 1 μM (as per the two-component gp41-mediated cage synthesis). The fully ligated product was purified, as previously, by affinity chromatography [confirmed by anti-histidine tag western blot & SDS PAGE gel (S.I. Figure 4i & Figure 4d, lane 8 &)] and characterised with SEC (Figure 4e). The results show that the cage closure was successful, with a substantial proportion of the fully ligated product forming discrete assemblies rather than extended networks. When the elution volume of the extended cage was compared to that obtained for the two-component cages a difference of 0.5 mL was observed (extended cages eluted at 11.4 mL, and the two-component cages eluted at 11.9 mL). Thus, as expected, the extended cages form a slightly larger structure than the two-component cages, but not so large as to indicate a dramatic change in conformation.

Discussion

Here we have demonstrated that genetically programmed NCL mediated via split-intein domains can be successfully utilised to assemble modular proteins designed with simple geometric symmetry into user-defined protein cages. To investigate the limits of the system, IMPDH and gp41 split inteins were used separately to drive two-component assembly of the half-cages modules (trefoil vertex and CTPR sides) and with an additional linker module for three-component assembly. Under mild conditions mixing compatible oligomeric protein fusions resulted in rapid and irreversible ligations with high yield of peptide bond-linked products. The fusion proteins were engineered to enable discrete fully ligated cages or extended half-cages to be easily and efficiently separated to homogeneity from each reaction in a two-step process. Significantly, this process generated the expected square bipyramidal structures ranging from 113 kDa (for the two-component, one reaction) to 150 kDa (three-component, two stepwise reactions). Furthermore, the cage resulting from the two-component ligation was shown to contain a central hollow cavity which could accommodate cargo up to 70 Å by 55-60 Å. Interestingly, although both split inteins gave high ligation yields, a greater yield of fully ligated discrete cages was obtained when lower reactant concentrations and the slower reacting IMPDH split intein were used. Likewise, for the stepwise assembly, IMPDH again gave higher yields.

The combined properties of the split inteins fusions and reaction/purifications yields show that: (i) the IMPDH mediated two-component system could be easily expanded to co-expression and assembly *in vivo,* and (ii) the three-component system permits a realistic scalable extension limit of three sequential ligations. The use of extendable CTPR sides additionally provides a method for loading cargo into the central hollow cavity of the nanostructures by using a binding module as the linker. For example, certain CTPR3 modules have had their pentapeptide binding pocket re-designed allow them to selectively recognise different pentapeptide tag sequences (Speltz et al., 2015). Thus, molecules of interest could be tagged with a pentapeptide sequence and then 'loaded' onto the nanostructure via the binding module, creating a generic loadable system. The use of proteins such as TPRs, which are natural binding proteins,

is an advantage of our assembly systems compared with others, for example those based on coiled-coils that do not possess such intrinsic binding capabilities.

In conclusion, our protein assembly system provides a more general route to producing protein cages that avoids many time-consuming and system-specific processes (for example, those requiring computational design). Moreover, by using a combination of the two highly reactive, orthogonal split inteins we expand the scope of assembly to stepwise scalable cage production. No bioconjugation, chemical modification or post-ligation refolding steps were required, and only a short sequence was inserted at the point of the NCL. Consequently, by using Nature's vast range of protein domains to engineer different geometric shapes, coupled with these high-yielding split inteins to drive the reaction, there are many opportunities for exploiting our self-assembling protein system.

Acknowledgements

We thank the beamline scientists at B21 of the Diamond Light Source, United Kingdom. LSI acknowledges the support of a Senior Fellowship from the UK Medical Research Foundation. ERGM and LSI labs acknowledge support from a Leverhulme Trust project grant. JW & JAH were supported by the Leverhulme Trust & QMUL Principal's Studentship, respectively. JAG was supported by the Medical Research Council (MR/M009920/1). We also wish to acknowledge Dr. Ruth Rose from the QMUL Protein Facility and Dr. Roberto Buccafusca from the QMUL Analytical Service Facility for their technical contributions.

Author Contributions

Conceptualization, ERGM, LSI; Methodology, JNW, WLW, ERGM and LSI; Investigation, JNW, WLW, JAH, JG, ERGM; Writing – Original Draft, ERGM; Writing – Review & Editing, JNW, WLW, JG, LSI and ERGM; Funding Acquisition, LSI and ERGM; Resources, LSI and ERGM; Supervision, ERGM.

Declaration of Interests

The authors declare no competing financial interests.

References

Aranko, A.S., Oeemig, J.S., Zhou, D., Kajander, T., Wlodawer, A., and Iwai, H. (2014). Structure-based engineering and comparison of novel split inteins for protein ligation. Molecular BioSystems *10*, 1023-1034.

Bale, J.B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T.O., Gonen, T., King, N.P.*, et al.* (2016). Accurate design of megadalton-scale two-component icosahedral protein complexes. Science *353*, 389-394.

Banerjee, A., and Howarth, M. (2017). Nanoteamwork: covalent protein assembly beyond duets towards protein ensembles and orchestras. Curr Opin Biotechnol *51*, 16-23.

Berrow, N.S., Alderton, D., and Owens, R.J. (2009). The precise engineering of expression vectors using high-throughput In-Fusion PCR cloning. Methods in molecular biology *498*, 75-90.

Brodin, J.D., Ambroggio, X.I., Tang, C., Parent, K.N., Baker, T.S., and Tezcan, F.A. (2012). Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat Chem *4*, 375-382.

Brune, K.D., and Howarth, M. (2018). New Routes and Opportunities for Modular Construction of Particulate Vaccines: Stick, Click, and Glue. Front Immunol *9*, 1432.

Busche, A.E., Aranko, A.S., Talebzadeh-Farooji, M., Bernhard, F., Dotsch, V., and Iwai, H. (2009). Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. Angew Chem Int Ed Engl *48*, 6128-6131.

Capito, R.M., Azevedo, H.S., Velichko, Y.S., Mata, A., and Stupp, S.I. (2008). Self-assembly of large and small molecules into hierarchically ordered sacs and membranes. Science *319*, 1812-1816.

Carvajal-Vallejos, P., Pallisse, R., Mootz, H.D., and Schmidt, S.R. (2012). Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. J Biol Chem *287*, 28686-28696.

Cortajarena, A.L., Wang, J., and Regan, L. (2010). Crystal structure of a designed tetratricopeptide repeat module in complex with its peptide ligand. The FEBS journal *277*, 1058-1066.

Dassa, B., London, N., Stoddard, B.L., Schueler-Furman, O., and Pietrokovski, S. (2009). Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. Nucleic Acids Res *37*, 2560-2573.

Debelouchina, G.T., and Muir, T.W. (2017). A molecular engineering toolbox for the structural biologist. Q Rev Biophys *50*, e7.

Demonte, D., Li, N., and Park, S. (2015). Postsynthetic Domain Assembly with NpuDnaE and SspDnaB Split Inteins. Appl Biochem Biotechnol *177*, 1137-1151.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and Development of Coot. Acta Crystallogr D Biol Crystallogr *66*, 486-501.

Fletcher, J.M., Harniman, R.L., Barnes, F.R., Boyle, A.L., Collins, A., Mantell, J., Sharp, T.H., Antognozzi, M., Booth, P.J., Linden, N.*, et al.* (2013). Self-assembling cages from coiled-coil peptide modules. Science *340*, 595-599.

Franke, D., Petoukhov, M.V., Konarev, P.V., Panjkovich, A., Tuukkanen, A., Mertens, H.D.T., Kikhney, A.G., Hajizadeh, N.R., Franklin, J.M., Jeffries, C.M.*, et al.* (2017). ATSAS

2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr *50*, 1212-1225.
Giessen, T.W., and Silver, P.A. (2016). A Catalytic Nanoreactor Based on in Vivo Encapsulation of Multiple Enzymes in an Engineered Protein Nanocompartment. Chembiochem *17*, 1931-1935.
Glover, D.J., Giger, L., Kim, S.S., Naik, R.R., and Clark, D.S. (2016). Geometrical assembly of ultrastable protein templates for nanomaterials. Nature communications *7*, 11771.
Gradisar, H., Bozic, S., Doles, T., Vengust, D., Hafner-Bratkovic, I., Mertelj, A., Webb, B., Sali, A., Klavzar, S., and Jerala, R. (2013). Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. Nature Chem Biol *9*, 362-366.
Grove, T.Z., Osuji, C.O., Forster, J.D., Dufresne, E.R., and Regan, L. (2010). Stimuli-Responsive Smart Gels Realized via Modular Protein Design. J Am Chem Soc *132*, 14024-14026.
Harvey, J.A., Itzhaki, L.S., and Main, E.R.G. (2018). Programmed Protein Self-Assembly Driven by Genetically Encoded Intein-Mediated Native Chemical Ligation. ACS Synth Biol *7*, 1067-1074.
Inostroza-Brito, K.E., Collin, E., Siton-Mendelson, O., Smith, K.H., Monge-Marcet, A., Ferreira, D.S., Rodriguez, R.P., Alonso, M., Rodriguez-Cabello, J.C., Reis, R.L*., et al.* (2015). Co-assembly, spatiotemporal control and morphogenesis of a hybrid protein-peptide system. Nat Chem *7*, 897-904.
Kajander, T., Cortajarena, A.L., Main, E.R., Mochrie, S.G., and Regan, L. (2005). A new folding paradigm for repeat proteins. J Am Chem Soc *127*, 10188-10190.
King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., Andre, I., Gonen, T., Yeates, T.O., and Baker, D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science *336*, 1171-1174.
Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr *36*, 1277-1282.
Lai, Y.T., Cascio, D., and Yeates, T.O. (2012). Structure of a 16-nm cage designed by using protein oligomers. Science *336*, 1129.
Lee, J., and Blaber, M. (2011). Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc Natl Acad Sci USA *108*, 126-130.
Lee, M.J., Mantell, J., Hodgson, L., Alibhai, D., Fletcher, J.M., Brown, I.R., Frank, S., Xue, W.F., Verkade, P., Woolfson, D.N*., et al.* (2018). Engineered synthetic scaffolds for organizing proteins within the bacterial cytoplasm. Nature Chem Biol *14*, 142-147.
Main, E.R., Phillips, J.J., and Millership, C. (2013). Repeat protein engineering: creating functional nanostructures/biomaterials from modular building blocks. Biochem Soc Trans *41*, 1152-1158.
Main, E.R.G., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. Structure *11*, 497-508.
Modica, J.A., Lin, Y., and Mrksich, M. (2018). Synthesis of Cyclic Megamolecules. J Am Chem Soc *140*, 6391-6399.
Padilla, J.E., Colovos, C., and Yeates, T.O. (2001). Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. Proc Natl Acad Sci USA *98*, 2217-2221.

Patterson, D.M., Nazarova, L.A., and Prescher, J.A. (2014). Finding the Right (Bioorthogonal) Chemistry. ACS Chem Biol *9*, 592-605.
Petoukhov, M.V., and Svergun, D.I. (2015). Ambiguity assessment of small-angle scattering curves from monodisperse systems. Acta Crystallogr D Biol Crystallogr *71*, 1051-1058.
Phillips, J.J., Millership, C., and Main, E.R.G. (2012). Fibrous Nanostructures from the Self-Assembly of Designed Repeat Protein Modules. Angew Chem Int Ed Engl *51*, 13132-13135.
Rambo, R.P., and Tainer, J.A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. Nature *496*, 477-481.
Sawyer, N., Chen, J., and Regan, L. (2013). All repeats are not equal: a module-based approach to guide repeat protein design. J Mol Biol *425*, 1826-1838.
Schrödinger, L. (2015). The PyMOL Molecular Graphics System, Version~1.8.
Shah, N.H., Vila-Perello, M., and Muir, T.W. (2011). Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. Angew Chem Int Ed Engl *50*, 6511-6515.
Shi, J., and Muir, T.W. (2005). Development of a tandem protein trans-splicing system based on native and engineered split inteins. J Am Chem Soc *127*, 6198-6206.
Speltz, E.B., Nathan, A., and Regan, L. (2015). Design of Protein-Peptide Interaction Modules for Assembling Supramolecular Structures in Vivo and in Vitro. ACS Chem Biol *10*, 2108-2115.
Stevens, A.J., Brown, Z.Z., Shah, N.H., Sekar, G., Cowburn, D., and Muir, T.W. (2016). Design of a Split Intein with Exceptional Protein Splicing Activity. J Am Chem Soc *138*, 2162-2165.
Stevens, A.J., Sekar, G., Shah, N.H., Mostafavi, A.Z., Cowburn, D., and Muir, T.W. (2017). A promiscuous split intein with expanded protein engineering applications. Proc Natl Acad Sci USA *114*, 8538-8543.
Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr *28*, 768-773.
Svergun, D.I. (1992). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. J Appl Crystallogr *25*, 495-503.
Svergun, D.I., Petoukhov, M.V., and Koch, M.H.J. (2001). Determination of domain structure of proteins from X-ray solution scattering. Biophys J *80*, 2946-2953.
Thiel, I.V., Volkmann, G., Pietrokovski, S., and Mootz, H.D. (2014). An atypical naturally split intein engineered for highly efficient protein labeling. Angew Chem Int Ed Engl *53*, 1306-1310.
Vagin, A.A., Steiner, R.S., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G.N. (2004). REFMAC5 dictionary: organisation of prior chemical knowledge and guidelines for its use. Acta Cryst. D *60*, 2284-2295.
Veggiani, G., Nakamura, T., Brenner, M.D., Gayet, R.V., Yan, J., Robinson, C.V., and Howarth, M. (2016). Programmable polyproteams built using twin peptide superglues. Proc Natl Acad Sci USA *113*, 1202-1207.
Vila-Perello, M., Liu, Z.H., Shah, N.H., Willis, J.A., Idoyaga, J., and Muir, T.W. (2013). Streamlined Expressed Protein Ligation Using Split Inteins. J Am Chem Soc *135*, 286-292.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.
Watanabe, T., Ito, Y., Yamada, T., Hashimoto, M., Sekine, S., and Tanaka, H. (1994). The roles of the c-terminal domain and type III domains of chitinase a1 from bacillus circulans wl-12 in chitin degradation. J Bacteriology, *176* (15):4465-4472.
Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A.*, et al.* (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr *67*, 235-242.

Figures

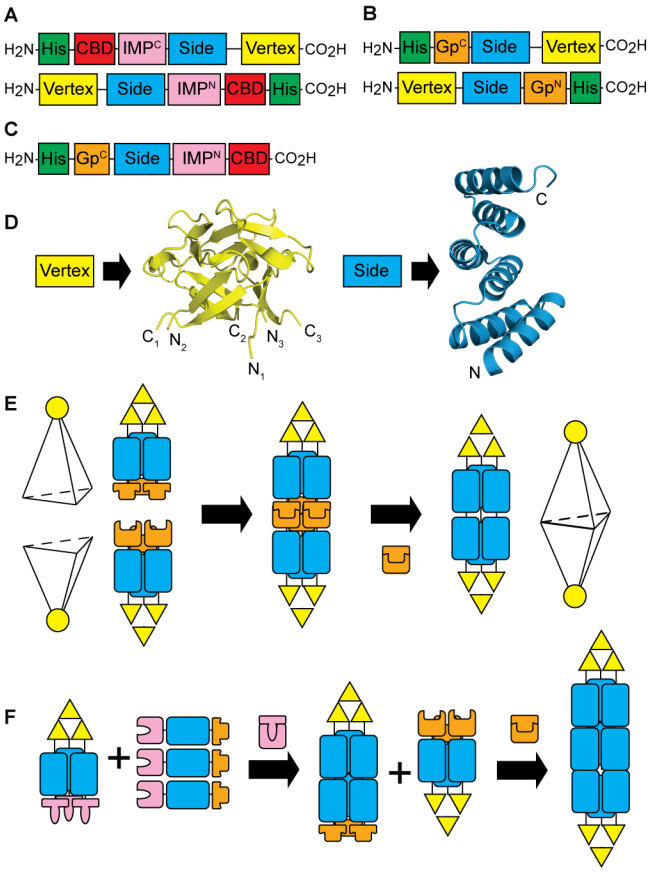

Figure 1: Schematic of designed recombinant fusion proteins and their formation into symmetric protein cages.

(a - b) Half-cage caps & Linker fusions: Two pairs of complementary half-cages and one linker fusion were constructed for this study. In (a & b) the compatible half-cage proteins are shown with (a) the IMPDH (IMP) and (b) the gp41 (Gp) split inteins. Each half cage consists of an oligomeric domain that acts as the primary vertex of the cage (yellow), a rigid domain that acts as the sides of the cage (blue), half of a split-intein pair for the chemistry required to join the cage halves (orange or pink) and affinity tags placed for initial fusion protein/final ligation product purification. The linker fusion (c) contains the protein to be assembled sandwiched between half an orthogonal pair of gp41 and IMPDH split inteins with affinity tags placed for initial fusion protein/final ligation product purification. In addition, a small, highly soluble and easy-to-refold Chitin Binding Domain (CBD) was placed next to each IMPDH split to aid solubility and retard aggregation.

(d) Structures of the primary vertex and side proteins used in this study: The initial vertex used was the designed homotrimer Monofoil-4-P (M4P) (3ol0.pdb (Lee and

Blaber, 2011)). The sides of each half cage were composed of the repeat protein CTPR3ΔS. The crystal structure of CTPR3 (1Na0.pdb (Main et al., 2003)) is shown without its final C-capping solvating helix to represent CTPR3ΔS.

(e) One-pot synthesis: From left to right - two complementary half-cage fusions with tripod-like structures. The M4P forms the vertex (yellow circle or three yellow triangles denoting each monomer), CTPR3ΔS units attached to each monomer of the M4P form the legs (blue rectangles) and the split intein pairs form the feet (orange). On mixing (1st arrow) the split inteins fold together and (2nd arrow) catalyse the ligation of the two cage halves whilst self-excising themselves from the complex. The product is a trigonal bipyramidal cage of two fused half-cages along one face via three common CTPR motif vertices formed on ligation.

(f) Stepwise extended cage synthesis: From left to right – compatible half cage & linker fusions are mixed (1st arrow). The compatible split intein pair fold together and catalyse the ligation of the linker to each of the cage "sides" whilst self-excising themselves from the complex. The product is an extended half cage which still contains reactive split intein halves on each extended "side". When mixed with a complementary half cage (2nd arrow) the second set of compatible split inteins fold and catalyse the 2nd ligation, forming the final extended cage product.

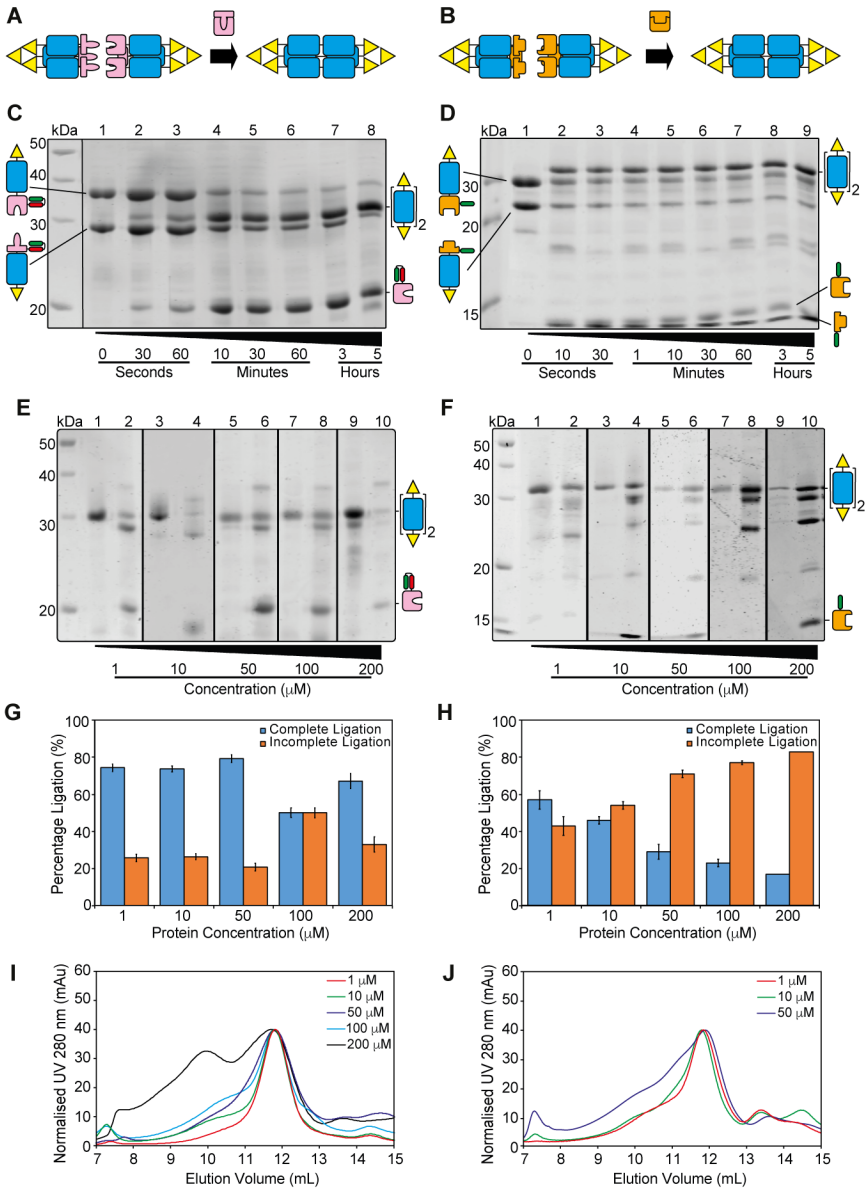

Figure 2: One-pot cage formation reactions, purification & isolation.

(a+b) Schematic of half cage reactions to produce cage product with (a) IMPDH-mediated ligation or (b) gp-mediated ligation.

(c & d) Denaturing SDS PAGE gels following the time course of the ligation reaction between compatible half cage reactants driven by (c) IMPDH split intein pair or (d) gp-41-1 split intein pair. As the gels are denaturing, the bands of the half cages reactants and ligated products correspond to the size of "monomeric" half cages (one M4P monomer, CTPR3ΔS, split intein half and affinity tag) and a "dimeric" ligated product (two M4P monomers and two CTPR3ΔS units), respectively. These are shown schematically besides the bands using the same colour scheme as in Figure 1 & 2a-b.

[In (c) the black partition between molecular weight markers and lane 1 denotes the digital elimination of two superfluous lanes from the original gel.]

(e & f) Denaturing SDS PAGE gels showing the His-tag affinity purification of the fully ligated half cages from partially ligated products, unreacted half cages and excised split inteins. (e) refers to reactions driven by the IMPDH split intein pair at differing concentrations of half cage reactant (shown at the bottom of the gel) with (f) showing the same but using gp-41-1 split intein pair. As the gels are denaturing the bands correspond to "monomeric" half cages or "dimeric" ligated product (as per panels c & d). Gel lanes 1, 3, 5, 7 and 9 show the purified fully ligated product containing no affinity tag, whereas lanes 2, 4, 6, 8 and 10 show the removed partially ligated products, unreacted half cages and excised split inteins which all still possess affinity tags. [Ligation products that did not bind and ligation products/reactants that were eluted from the Ni^{2+} resin were concentrated to 1 mL prior to gel electrophoresis (expt. detail described in the Star Methods). Black partitions denote differing gels that are displayed next to each other of ease of viewing.].

(g + h) Quantification from SDS PAGE gels in (e & f) of the half cage ligation products that were either fully ligated or partially ligated (g = IMPDH mediated ligation and h = gp41-1 mediated ligation). Error bars equate to standard deviation of multiple repeat experiments (at least three in all cases).

(i & j) Analytical size exclusion chromatograms of the purified fully ligated half cage products (i = IMPDH-mediated ligation and j = gp41-mediated ligation). The unbound ligation fraction from the affinity purification step (g & h) were concentrated to 1 mL and 100 µL injected onto a Superdex G200 10/30 analytical column using standard ligation buffer (see SI).

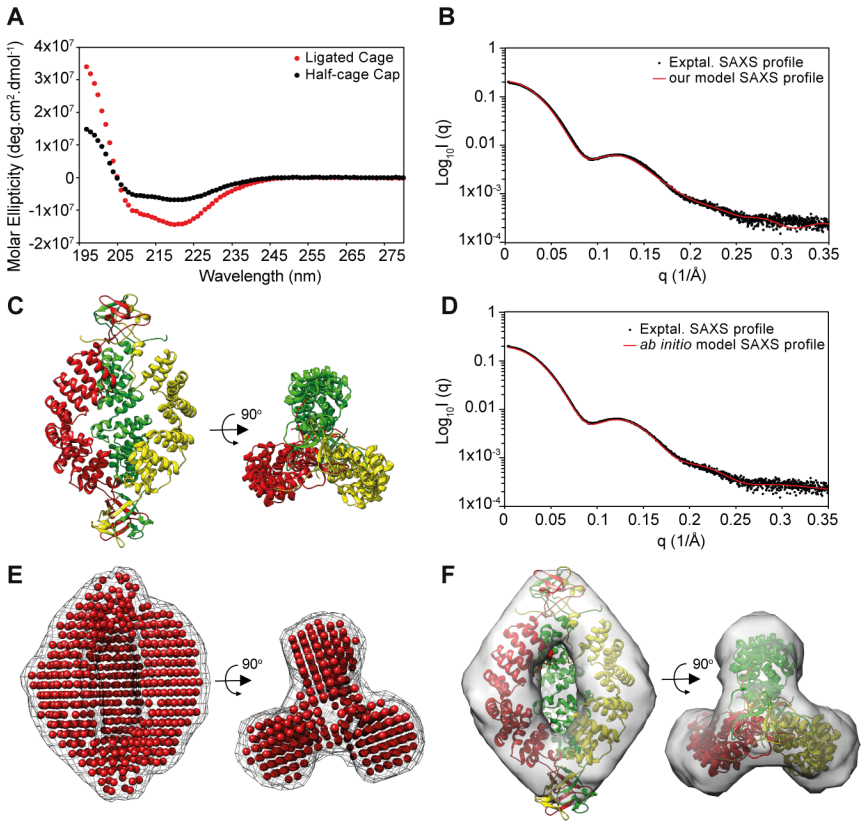

Figure 3: Structural characterisation of two-component ligated cages by SAXS and far UV CD.

(a) Far UV CD spectra of ligated cage in comparison to a half-cage cap (without split inteins). **(b)** Experimental SAXS profile (black circles) of ligated cages overlaid with our "best" ligated cage atomic model SAXS profile (red line) **(c)** Two orientations of our "best" ligated cage atomic model. **(d)** Experimental SAXS profile (black circles) of ligated cages overlaid the *ab initio* GASBOR generated model SAXS profile (red line). **(e)** Two orientations of the *ab initio* GASBOR generated model. **(f)** Two orientations of our "best" ligated cage atomic model docked into the *ab initio* GASBOR generated model.

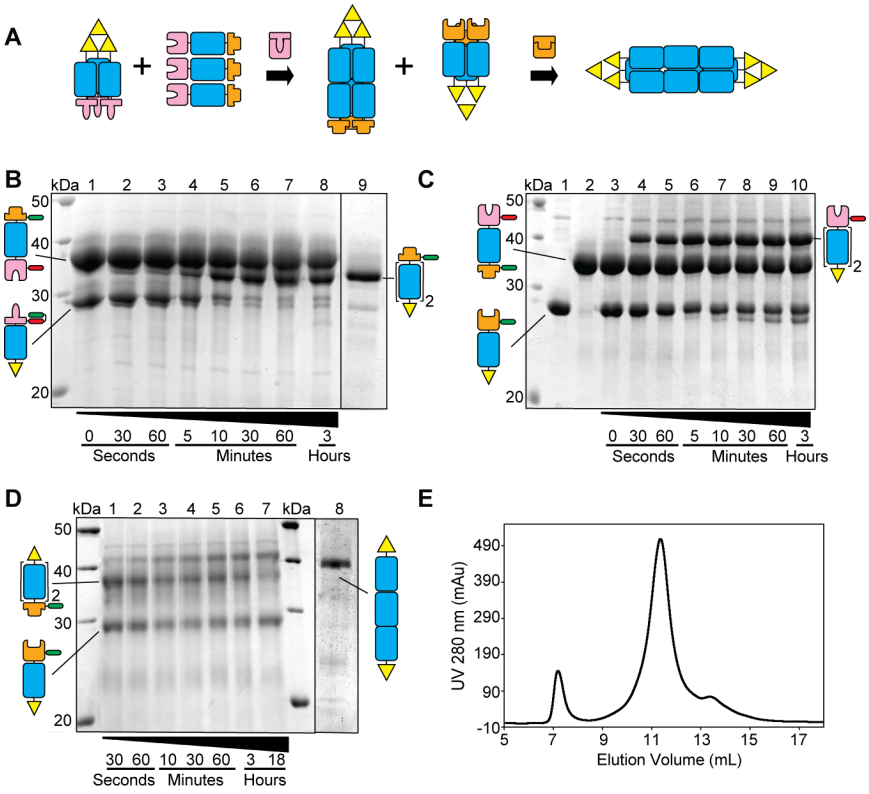

Figure 4: Stepwise extended cage formation reactions, purification and isolation.

(a) Schematic of stepwise extended cage synthesis: From left to right – compatible half cage & linker fusion are mixed (1st arrow). The compatible pair of split inteins fold together and catalyse the ligation of the linker to each of the cage "sides" whilst self-excising themselves from the complex. The product is an extended half cage which still contains reactive split intein halves on each extended "side". This is isolated and mixed with a complementary half cage (2nd arrow). As previously, the split inteins fold and catalyse the 2nd ligation, forming the final extended cage product.

(b & c) Denaturing SDS PAGE gels following the time course of the ligation reaction between half cage and linker reactants driven by (b) IMPDH split intein pair (lane 9 shows a separate gel of the purified fully ligated product) or (c) gp-41-1 split intein pair. As per Figure 2, the denaturing gel means the bands of the half cages reactants and ligated products correspond to the size of "monomeric" half cages (a M4P monomer, one CTPR3ΔS, one split intein half and one affinity tag) and a "dimeric" ligated product (M4P monomers, two CTPR3ΔS units, one split intein half and one affinity tag), respectively. These are shown schematically besides the bands using the same colour scheme as in Figure 1.

(d) Denaturing SDS PAGE gel following the time course of the ligation reaction between purified extended half cage and linker reactants driven by gp-41-1 split intein pair. Lane 8 shows a separate gel of the purified fully ligated product. As in b & c, the bands correspond to the size of “monomeric” extended half cage and a “trimeric” ligated final cage product. These are shown schematically besides the bands using the same colour scheme as in Figure 1.

(e) Analytical size exclusion chromatogram of the purified fully ligated extended cage product shown in panel d, lane 8. The sample was concentrated to 1 mL and 100 µL injected onto a Superdex G200 10/30 analytical column using standard ligation buffer (see SI).

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Ewan Main (e.main@qmul.ac.uk)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacteria Strains

XL2-Blue and Bl21(DE3) C41 cells were cultured in 2YT medium in the presence of appropriate antibiotics.

METHOD DETAILS

Construction & Production of half-cage and linker fusions

Construction of fusion protein genes & vectors: Genes were synthesised by Life Technologies (UK) [now Thermofisher] and the final expression plasmids generated by either: (i) sequential insertion of the aforementioned genes via 5' NheI and 3' SpeI BglII restriction sites into a prepared pOPIN vector (Berrow et al., 2009) or (ii) sub-cloned into a pOPIN vector that had been customised to include a multiple cloning site sandwiched between split intein genes and His/CBD affinity tags [CBD tag sequence used was (Watanabe *et al.*, 1994)]. All vectors were verified by DNA sequencing (Genewiz). S.I. Table 1 shows an overview of the constructs (Please note that the split inteins were cloned with an additional 5 amino acid extein sequence required for high yields of spliced product).

Protein production: Plasmids were transformed into the *E. coli* C41 cells (Lucigen) and grown in YT media at 37 °C until the A_{600nm} reached ~1. Expression was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. After induction, for native purification the temperature was lowered to either 16

°C overnight or 30 °C for 5 hours (S.I. Table 2). For denatured purification the temperature was maintained at 37 °C for between 4 hrs and overnight (S.I. Table 2). Cells were harvested by centrifugation (10 mins at 10,000 g) and the cell pellet was resuspended in 50 mM Tris (pH 8), 300 mM NaCl (native) or 6M GuHCl, 50mM Tris (pH 8), 300mM NaCl (denaturing), snap frozen and stored at -80 °C.

Protein purification: The resuspended cell pellets were thawed and lysed via sonication on ice. Insoluble matter was removed by centrifugation (30 min at 39,000 x g). The expressed protein fusions were then then purified from the supernatant using Nickel-IDA resin (using either standard native or denaturing protocol as per Qiagen/ThermoFisher Scientific manual). For denaturing purification, proteins were refolded either via stepping down the denaturant concentration whilst bound to the Nickel-IDA resin or via dialysis. DTT was added to the native, refolded or denatured elutions (final concentration 5 mM). If further purification was required, size exclusion chromatography (SEC) was performed using a HiLoad 16/60 Superdex 200 prep grade on an AKTA Pure or Purifier systems (GE). Protein purity was assayed through SDS denaturing PAGE electrophoresis. Protein concentrations and thus yields were determined from the absorbance at 280 nm using the extinction co-efficient calculated from the amino acid sequence. The concentration of proteins refers to that of a monomer (even in situations where the fusion protein forms oligomeric species, unless otherwise stated). All proteins were stored at 4 °C for use up to 24 hours. Where it was necessary to store for more than 24 hours, 1 mL protein aliquots were flash-frozen and stored at -80 °C, in an appropriate buffer for downstream applications.

Ligation reactions

All ligations, unless otherwise stated, were carried out in mild conditions using a ligation buffer of 50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM DTT, 1 M urea and

incubated at room temperature (25 °C) – the small concentration of urea increased yields via reducing aggregation. The reactions were left to proceed under mild agitation. Where yields are quoted they are calculated as averages of multiple repeat experiments (at least three in all cases).

One-pot cage synthesis & 1st step purification: Ligation reactions of two complementary half-cage caps was initiated by mixing equimolar amounts of each. The effect of using different split inteins to mediate ligations, changing protein concentration and urea concentrations were then investigated whilst using the same ligation buffer (except where urea concentration was varied). Protein concentrations and reaction volumes were varied as follows:

(i) 200 µM reacted in 1.5 mL, (ii) 100 µM reacted in 1.5 mL, (iii) 50 µM reacted in 4.5 mL, (iv) 10 µM reacted in 20 mL and (v) 1 µM reacted in 100 mL.

Once the ligation was complete the resultant reaction mixture was purified using a nickel HisTrap column (GE) attached to an AKTA Pure or Purifier systems (GE). Both the fully ligated product, which did not bind to the HisTrap column, and partially ligated products/ reactants that were bound and then eluted from the HisTrap column were collected and concentrated to 1 mL prior to further characterisation.

Stepwise cage synthesis & 1st step purification: The first step in the larger cage formation reacts one half cage cap with the linker protein. This reaction was trialled with both gp41- and IMPDH-tagged half cage caps at 33 µM, where the linker protein was in an excess of 9:1, 6:1 and 3:1. In addition, the effect of changing urea concentrations was also investigated (0 to 2 M). The highest yield of ligated product was achieved with IMPDH-mediated ligations in a 1:6 half-cage to linker ratio and 1 M urea (Figure 4b). Once the first step ligation was complete the resultant reaction mixture was purified using a chitin column (NEB) under gravity. After the IMDPH mediated first step and purification the cage structure was closed by reacting 1 µM of fully ligated larger cage cap with 1 µM of the compatible gp41-tagged half cage cap. Once the ligation was complete the resultant reaction mixture was purified using a nickel HisTrap column (GE) attached to an AKTA Pure systems (GE).

Characterisation of reactants, products & reaction yields from ligation reactions

Ligation reactions were monitored and the products characterised by denaturing SDS-PAGE electrophoresis, western blot, size exclusion chromatography (SEC), small angled x-ray scattering (SAXS), far UV circular dichroism (C.D.) and where possible MALDI Mass spectrometry.

Denaturing SDS-PAGE gels: Aliquots at differing time points were removed from ligation reactions and halted via a combination of the addition of 2x SDS-PAGE non-reducing loading buffer (20% glycerol, 125 mM Tris HCl pH 6.8, 4% SDS, 0.2% bromophenol blue) and boiling. For time = 0 reading, samples were boiled prior to mixing to prevent artefactual ligation. SDS-PAGE was performed on 14 % - 18 % Tris-glycine gels and stained using Coomassie Brilliant blue G-250. Band intensity was measured using the LI-COR Odyssey Infrared Imaging System and Image Studio Light (ver 5.2.5) software. Integrated intensity values (I) corresponding to each protein band were thereby obtained.

Reaction yields from SDS-PAGE gels: Equation 1 was used to obtain the percentage of ligated product formed:

$$\% \text{ Yield} = \left[\frac{\left(\frac{I_P}{MWt_P}\right)}{\left[\left(\frac{I_P}{MWt_P}\right)+\left(\frac{I_R}{MWt_R}\right)\right]}\right] \times 100 \qquad (1)$$

where I_P is the integrated intensity of the ligated product, MWt_P is the molecular weight of the ligated product, I_R is the integrated intensity of the most consumed reactant and MWt_R is the molecular weight of most consumed reactant. Equation 1 assumes that the binding of Coomassie stain (and, therefore, the intensity) is linearly related to the molecular weight of each NCL protein.

Western Blots: SDS Page gels were run as above and transferred to the membrane using the Bio-Rad Trans-Blot® Turbo™ Transfer System using its preset mixed molecular weight program (1.5 A, 25 V for 7 minutes). After blocking with 5 % milk in 1 X PBS, 0.1 % Tween (PBST) for an hour, the membrane was washed twice with PBST

and incubated with 1:1000 PBST of either Monoclonal Anti-CBD Tag antibody or Monoclonal Anti-His antibody produced in mouse (NEB & Sigma, respectively) at room temperature for 1 hour. Membranes were then washed twice with PBST for 5 minutes and incubated with 1:20000 PBST of IRDye® 680LT Goat anti-Mouse IgG for 1 hour. Blots were then washed with PBST and PBS before being imaged with a LI-COR Odyssey Infrared Imaging System.

Analytical Size Exclusion Chromatography (SEC): Analytical SEC was carried out using the Superdex^{TM} 200 10/30 column attached to an AKTA Pure or Purifier system (GE). 100 µl of protein sample (50 - 100 µM) was loaded onto the column and run in 50 mM Tris pH8, 300 mM NaCl, 5 mM DTT, 0 - 2 M Urea. UV absorption peaks were processed using Unicorn software (v5.0) and analysed, before being exported and plotted using Microsoft Excel. When compared on the same plot, UV 280 nm signals were normalised to a maximum of 50 mAU for clarity. Size was estimated using the Amersham low molecular weight gel filtration calibration kit containing the following standards: Albumin 67 kDa, Ovalbumin 43 kDa, Chymotrypsinogen A 25 kDa. To quantify the relative position of each peak the V_e/V_o was calculated and plotted against the $\log_{10}$ of the molecular weight (kDa). A linear trend line was drawn and the equation used to calculate the molecular weights of unknown samples. (V_e = elution volume (ml) where the maxima of the UV absorption peak appears and V_o = void volume (ml) of the mobile phase.)

Mass Spectrometry: Matrix assisted laser desorption/ionisation - time of flight mass spectrometry (MALDI-TOF/MS) was carried out to confirm the masses of recombinant proteins and post-splice reaction products where possible using a Bruker 2000 MALDI-TOF mass spectrometer. To remove/reduce buffer components, samples were prepared either by (i) dilution using 50% aceto-nitrile, 1% or 0.1% trifluoroacetic acid in water (Sigma) or (ii) using EMD Millipore Zip-Tip® pipette tips according to the manufacturer's instructions. 1 µl of sample from either dilution or Zip-Tip® was mixed 1:1 with saturated sinapinic acid matrix and spotted onto a Bruker MALDI-TOF/MS steel plate and allowed to air dry. The target plate was loaded into the Bruker 2000 MALDI-TOF mass spectrometer. Using positive ion mode the gain and laser power was adjusted until optimal signal/noise ratio was achieved. The TOF was operated in the

reflectron or linear mode and each spectrum was an average of 300 laser shots. The spectra were calibrated using the Protein standard 1 & 2 (Bruker). The data was extracted and visualised using R, Bruker Flex Analysis or Microsoft Excel software.

(iii) Circular Dichroism (CD): Protein concentrations of 1 – 10 µM in 10 mM Tris-HCl pH 8, 50 mM NaCl and 5 mM DTT were analysed in a 0.5 mm path length cuvette using a Chirascan™ CD Spectrometer (Applied Photophysics Ltd, UK). For each sample a spectrum from 195 – 280 nm was recorded with points taken at 1.0 nm intervals and 0.5 sec per point scanning time. The averaged spectrum of 3 repeats was taken for each sample. Data was converted to molar ellipticity using Equation 2:

$$\theta_{molar} = 100 \times \theta_{obs} \ / \ M \times l \qquad (2)$$

where θ_{molar} is the molar ellipticity in deg cm^2 $dmol^{-1}$, θ_{obs} is the observed CD signal in millidegrees, l is the path length in cm and M is the molar concentration of protein.

Size exclusion chromatography small angle X-ray scattering (SEC-SAXS)

SAXS cage samples were prepared by ligating 100 mL of 10 µM IMPDH tagged half cage caps and purified using the same affinity and size exclusion chromatography steps outlined above. Purified cages were concentrated to 10 mg/ml and dialysed 10 mM Tris pH 8.0, 50 mM NaCl, 5mM DTT. SAXS experiments were recorded on beamline B21 at the Diamond Light Source (DLS), UK, coupled to a Shodex KW403-4F size exclusion column. Data were measured at 20 °C with a wavelength of 0.99 Å and a 3 s exposure time per frame on a Pilatus 2 M two-dimensional detector at 4.014 m distance from the sample, corresponding to a momentum transfer range of $0.004 < q < 0.4$ $Å^{-1}$ ($q = 4\pi \sin \theta \lambda^{-1}$, 2θ is the scattering angle). Elution peak and buffer selection, and subsequent buffer subtraction, intensity normalization, and data merging were performed in ScÅtter (Rambo and Tainer, 2013). Further analysis was carried out with a q range of $0.018 < q < 0.35$ $Å^{-1}$. The radius of gyration (R_g) and scattering at zero angle (I(0)) were calculated from the analysis of the Guinier region by AUTORG (Konarev et al., 2003; Svergun, 1992). The distance distribution function (P(r)) was subsequently obtained using GNOM (Konarev et al., 2003; Svergun, 1992),

yielding the maximum particle dimension (D_{max}). To demonstrate the absence of concentration-dependent aggregation and interparticle interference we inspected Rg over the elution peaks and performed our analysis on frames where Rg was most stable. The Porod exponent and molecular weight were calculated within the SCATTER (DLS, UK) and ATSAS package (Franke et al., 2017), respectively. All data collection and processing statistics are listed in S.I. Table 3.

Comparison of experimental SAXS profile to generated atomic cage models*:* We compared the experimental cage SAXS profile to 30 manually generated differing atomic models of possible designed cage conformations using the program *Crysol* (Svergun et al., 1995) (Data S1). The 30 designed cages models were constructed by first manually positioning the crystal structure of M4P (PDB 3OL0) with either CTPR3 minus the C-terminal solvating helix for non "docked" cage interfaces (PDB 1NA0) or a CTPR6 unit from CTPR8 for "docked" cage interfaces (PDB: 2FOT) using PyMol (Schrödinger, 2015). COOT was then used to add connecting sequences between the domains using its Rigid Body Fit Zone functionality (Emsley et al., 2010) and three-fold symmetry enforced using PyMol. The chains were renumbered (PDBSet, CCP4 suit (Winn et al., 2011)) and rigid body refined with REFMAC5 (Vagin et al., 2004; Winn et al., 2011) to obtain final the lowest energy confirmation. These final lowest energy structures were then converted to a SAXS profile by *Crysol* and compared to the experimentally determined profile (Figure 3 and Data S1). Modelling statistics of the "best" model are listed in S.I. Table 3.

Ab initio shape determination from SAXS*:* The structure of cages formed from half cage ligations were expected to display 32-symmetry at low resolution, which increases the likelihood of ambiguity in low-resolution particle shapes reconstructions and is reflected by having an ambiguity score of 1.6 (Petoukhov and Svergun, 2015). Therefore, *ab initio* modelling using Gasbor (Svergun et al., 2001) was repeated 32 times applying 32-symmetry restraints and models were selected that were supported by our biophysical data. Solutions were discarded when, for example, the CTPR/M4P domains would be required to either adopt non-native conformations or be ligated in a nonsensical formation to fit the calculated protein density envelopes (discarded

examples are shown in S.I. Figure 4e-h). This resulted in five solutions that were then aligned and averaged using DAMAVER (Volkov and Svergun, 2003). *Ab initio* modelling statistics are listed in S.I. Table 3.

QUANTIFICATION AND STATISTICAL ANALYSIS

Accuracy and reproducibility of the ligation reactions: Quantification of accuracy and reproducibility of the yields and rates of the various ligation reactions were obtained by conducting each reaction in at least triplicate and calculating the standard deviation. These are displayed as error bars in Figure 2g-h / S.I Figure 2.

SAXS: Modelling statistics and final χ^2 value between the model and experimental SAXS profiles (shown in S.I. Table 3 and Data S1) were calculated as described in Method Details.