Hierarchical design of multi-scale protein complexes by combinatorial assembly of oligomeric helical bundle and repeat protein building blocks

Yang Hsia^{1,2,3,#}, Rubul Mout^{1,2,#}, William Sheffler^{1,2}, Natasha I. Edman^{1,2,5,6}, Ivan Vulovic^{1,2,4}, Young-Jun Park¹, Rachel L. Redler⁸, Matthew J. Bick^{1,2}, Asim K. Bera^{1,2}, Alexis Courbet^{1,2}, Alex Kang^{1,2}, TJ Brunette^{1,2}, Una Nattermann^{1,2,3}, Evelyn Tsai^{1,2}, Ayesha Saleem^{1,2}, Cameron M. Chow^{1,2}, Damian Ekiert^{8,9}, Gira Bhabha⁸, David Veesler¹, David Baker^{1,2,7*}

¹Department of Biochemistry, University of Washington, Seattle, WA 98195

²Institute for Protein Design, University of Washington, Seattle, WA 98195

³Biological Physics, Structure and Design Graduate Program, University of Washington, Seattle, WA 98195

⁴Molecular Engineering Graduate Program, University of Washington, Seattle, WA 98195

⁵Molecular and Cellular Biology Graduate Program, University of Washington, Seattle, WA 98195

⁶Medical Scientist Training Program, University of Washington, Seattle, WA 98195

⁷Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

⁸Department of Cell Biology and Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, NY 10016

⁹Department of Microbiology, New York University School of Medicine, New York, NY 10016

[#]These authors have contributed equally

*Corresponding author. E-mail: dabaker@uw.edu

29 **Abstract:**

30 A goal of *de novo* protein design is to develop a systematic and robust approach
31 to generating complex nanomaterials from stable building blocks. Due to their structural
32 regularity and simplicity, a wide range of monomeric repeat proteins and oligomeric
33 helical bundle structures have been designed and characterized. Here we describe a
34 stepwise hierarchical approach to building up multi-component symmetric protein
35 assemblies using these structures. We first connect designed helical repeat proteins
36 (DHRs) to designed helical bundle proteins (HBs) to generate a large library of
37 heterodimeric and homooligomeric building blocks; the latter have cyclic symmetries
38 ranging from C2 to C6. All of the building blocks have repeat proteins with accessible
39 termini, which we take advantage of in a second round of architecture guided rigid
40 helical fusion (WORMS) to generate larger symmetric assemblies including C3 and C5
41 cyclic and D2 dihedral rings, a tetrahedral cage, and a 120 subunit icosahedral cage.
42 Characterization of the structures by small angle x-ray scattering, x-ray crystallography,
43 and cryo-electron microscopy demonstrates that the hierarchical design approach can
44 accurately and robustly generate a wide range of macromolecular assemblies; with a
45 diameter of 43nm, the icosahedral nanocage is the largest structurally validated
46 designed cage to date. The computational methods and building block sets described
47 here provide a very general route to new *de novo* designed symmetric protein
48 nanomaterials.

49

50 Keywords: ***De novo* protein design, self-assembly, helical fusion, hierarchical**
51 **assembly, bottom up assembly**

52 **Introduction:**

53 There has been considerable recent interest in designing self assembling protein
54 nano structures and materials^{1,2}. Computational protein design has been used to create
55 proteins that self-assemble into a wide variety of higher order structures, from cyclic³
56 and dihedral symmetries⁴ to point group nanocages⁵⁻⁷, 1-dimensional fibers⁸, and
57 2-dimensional arrays⁹. The nanocages have been utilized in vaccine development^{10,11},
58 drug delivery¹², and as microscopy standards⁷. Most of these structures have been
59 created by symmetrically docking protein building blocks followed by sequence
60 optimization at the new interfaces^{3,5-7,9,13} using RosettaDesign¹⁴. However, interface
61 design remains challenging, and designable interface quality is heavily dependent on
62 how well the building blocks complement each other during design. An alternative
63 approach which avoids the need for designing new interfaces is to fuse oligomeric
64 protein building blocks with helical linkers; while this has led to a number of new
65 materials¹⁵, lack of rigidity has made the structures of these assemblies difficult to
66 precisely specify. More rigid junctions created by overlapping ideal helices and
67 designing around the junction region has resulted in more predictable structures^{16,17},
68 including closed ring dihedral structures which require even more precise structure
69 predictions¹⁸. This rigid fusion method, however, has its own set of challenges in
70 comparison to designing a new non-covalent protein-protein interface: first, for any pair
71 of protein building blocks, there are far fewer positions for rigid fusion than are for
72 unconstrained protein-protein docking limiting the space of possible solutions, and
73 second, while in the non-covalent protein interface case the space searched can be

74 limited by restricting building blocks to the symmetry axes of the desired nanomaterial,
75 this is not possible in the case of rigid fusions, making the search more difficult as the
76 number of building blocks increases.

77 A potential solution to the issue of having smaller numbers of possible fusion
78 positions for a given pair of building blocks in the rigid helix fusion method is to
79 systematically generate large numbers of building blocks having properties ideal for
80 helix fusion. Attractive candidates for such an approach are *de novo* helical repeat
81 proteins (DHRs)²² consisting of a tandemly repeated structural unit, which provide a
82 wide range of struts of different shape and curvature for building nanomaterials, and
83 parametric helical bundles (HBs)^{19–22} which provide a wide range of preformed
84 protein-protein interfaces for locking together different protein subunits in a designed
85 nanomaterial. Many examples of both classes of designed proteins have been solved
86 by x-ray crystallography, and they are typically very stable. We reasoned that by
87 systematically fusing DHR “arms” to central HB “hubs” we could generate building
88 blocks with a wide range of geometries and valencies that, because of the modular
89 nature of repeat proteins, enable a very large number of rigid helix fusions: given two
90 such building blocks with N- and C-terminally extending repeat protein arms, the
91 potentially rigid fusion sites are any pair of internal helical residues in the DHR arms.

92 With a large library of building blocks, the challenge is then to develop a method
93 to very quickly traverse all possible combinations of fusion locations. We present here
94 WORMS, a software package that uses geometric hashing of transforms to very quickly
95 and systematically identify the fusion positions in large sets of building blocks that

- 96 generate any specified symmetric architecture, and describe the use of the software to
- 97 design a broad range of symmetric assemblies.

98 **Results:**

99 We describe the development of methods for creating large and modular libraries
 100 of building blocks by fusing DHRs to HBs, and then using them to generate symmetric
 101 assemblies by rapidly scanning through the combinatorially large number of possible
 102 rigid helix fusions for those generating the desired architecture. We present the new
 103 methodology and results in two sections. In section one, we describe the systematic
 104 generation of homo- and hetero-oligomeric building blocks from *de novo* designed
 105 helical bundles, helical oligomers, and repeat proteins (Figure 1a). In the second
 106 section, we describe the use of these building blocks to assemble a wide variety of
 107 higher order symmetric architectures (Figure 1c).

108

109

Figure 1. Overview of the rigid hierarchical fusion approach. (a) Hetero- (yellow/green) and homo- (red) oligomeric helical bundles are fused to *de novo* helical repeat proteins (shades of blue) (left) to create a wide range of building blocks using HelixDock and HelixFuse (center). Symmetric units shown in grey. (b) Twenty representative HelixFuse outputs overlaid in groups of five to display the wide range of diversity that can be generated by using a single helical bundle core. (c) These are then further assembled into higher ordered structures through helical fusion (WORMS, right). The examples shown are cyclic crowns (top), dihedral rings (middle), and icosahedral nanocages (bottom).

120 **Section 1: Systematic generation of oligomeric building blocks**

121 To generate a wide variety of building blocks, we explored two different
122 methodologies for fusing DHRs to HBs (Figure 1a). The first is to dock the DHR units to
123 the HBs, redesign the residues at the newly created interface, and then build loops
124 between nearby termini (HelixDock, HD). The second protocol simplifies the process by
125 overlapping the helical termini of the DHRs and HBs and designing only the immediate
126 residues around the junction (HelixFuse, HF). As an example of the combinatorial
127 diversity that can be generated due to the large number of possible internal helical
128 fusion sites in a DHR (nearly all helical residues), a single terminus from a single helical
129 bundle (2L6HC3-12²⁰, N-terminus) combined with the library of 44 verified DHRs
130 resulted in 259 different structures (Figure 1b).

131 **HelixDock (HD) approach:** 44 DHRs with validated structures²³ and 11 HBs^{20,24}
132 (including some without pre-verified structures) were selected as input scaffolds for
133 symmetrical docking using a modified version of the sicdock software³. In each case, N
134 copies of the DHR, one for each monomer in the helical bundle, were symmetrically
135 docked onto the HB, sampling all six degrees of freedom, to generate star shaped
136 structures with repeat protein arms emanating symmetrically from the helical bundle in
137 the center. Docked configurations with linkable N- and C-termini within a distance cutoff
138 of 9Å with interfaces predicted to yield low energy designs²⁵ were then subjected to
139 Rosetta sequence design to optimize the residue identity and packing at the newly
140 formed interface. Designs with high predicted domain-domain binding energy and shape
141 complementarity²⁶ were identified, and loops connecting chain the termini were built

142 using the ConnectChainsMover¹⁷. Structures with good loop geometry (passing
143 worse9merFilter and FoldabilityFilter) were forward folded with RosettaRemodel²⁷
144 symmetrically, and those with sequences which fold into the designed structure *in silico*
145 were identified.

146 Synthetic genes encoding a subset of the selected designs with a wide range of
147 shapes were synthesized and the proteins expressed in *E. coli*. Of the 115 sequences
148 ordered successfully synthesized, 65 resulted in soluble protein. Those with poor
149 expression and/or solution behavior were discarded. Of the remaining, 39 had relatively
150 monodisperse Size Exclusion Chromatography (SEC) profiles that matched what was
151 expected from the design. Of the ones selected for small angle X-ray scattering (SAXS),
152 17 had profiles close to those computed from the design models (Figure S1-3). Design
153 C3_HD-1069, was crystallized and solved to 2.4 Å (Figure 2a). Although the two loops
154 connecting to the HB are unresolved in the structure, the resulting placement of the
155 DHR remains correct (unresolved loops were also present in the original HB structure
156 (2L6HC3_6)²⁰. The resolved rotamers at the newly designed interface between the HB
157 and DHR are also as designed.

158 **HelixFuse (HF) approach:** The same set of DHRs and HBs were
159 combinatorially fused together by overlapping the terminal helix residues in both
160 directions (“AB”: c-terminus of HB to n-terminus of DHR, “BA”: n-terminus of HB to
161 c-terminus of DHR)¹⁷. On the HB end, up to 4 residues were allowed to be deleted to
162 maximize the sampling space of the fusion while maintaining the structural integrity of
163 the oligomeric interface. On the DHR end, deletions up to a single repeat were allowed.

164 After the C-beta atoms are superimposed, a RMSD check across 9 residues was
 165 performed to ensure that the fusion results in a continuous helix. If no residues in the
 166 fused structure clash (Rosetta centroid energy < 10), sequence design was carried out
 167 at all positions within 8Å of the junction. This first step of the fusion sampling is wrapped
 168 into the Rosetta MergePDBMover¹⁷. After sequence design around the junction
 169 region^{14,28}, fusions were then evaluated based on the number of helices interacting
 170 across the interface (at least 3), buried surface (sasa > 800) across the junction, and
 171 shape complementarity (sc > 0.6) to identify designs likely to be rigid across the junction
 172 point. In total, the building block library generated *in silico* by HelixFuse using HB hubs
 173 and DHR arms in this set consists of 490 C2s, 1255 C3s, 107 C5s, and 87 C6s.

174 As a proof of concept, select fusions to C5 (5H2LD-10⁷) and C6 (6H2LD-8) (in
 175 press) helical bundles were tested experimentally, as structures of higher cyclic
 176 symmetries were historically more difficult to design thus resulting in a lack of available
 177 scaffolds. Contrarily, larger structures are easier to experimentally characterize via
 178 electron microscopy due to their size. A total of 65 designs whose genes encoding the
 179 designs were synthesized and subsequently expressed in *E.coli*, 45 were soluble, and
 180 23 were monodisperse by SEC. Of the ones that were selected for SAXS analysis, 7
 181 had matching SAXS profiles (Figure S4-5). Cryo-electron microscopy of C5_HF-3921
 182 followed by 3D reconstruction showed that the positions of the helical arms are close to
 183 the design model (Figure 2e, Figure S8 & 9). By negative-stain electron microscopy
 184 (EM), C5_HF-2101, C5_HF-0019, C6_HF-0075, and C6_HFuse-0080 (Figure 2f-i
 185 respectively) were class averaged and the top-down view clearly resembles that of the

186 designed model and its predicted projection map (Figure S10, S12, S13, and S14
187 respectively). From negative-stain EM class averaging, off-target states can sometimes
188 be observed; most obvious in C5_HF-0007 (Figure S11) and C6_HF-0075 (Figure S13),
189 and less in C5_HF-0019 (Figure S12), where in some cases an incorrect number of
190 DHR arms can be observed in the 2D class averages.

191 We also applied the method to two non-helical bundle oligomers - 1wa3, a native
192 homo-trimer²⁹ and tpr1C4_pm3, a designed homo-tetramer²⁵. As described above, we
193 fused DHRs to the N-terminal helix of 1wa3 and the C-terminal helix of tpr1C4_pm3. For
194 1wa3, from the 13 designs were expressed for experimental validation, 10 displayed
195 soluble expression and showed clean monodispersed peaks by SEC. Through X-ray
196 crystallography, we were able to solve C3_nat_HF-0005 to 3.32Å resolution (Figure 2c).
197 A total of 16 tpr1C4_pm3 fusions were tested, 14 found to be soluble, and 10 displayed
198 monodispersed peaks by SEC. The best behaving designs were analyzed by electron
199 microscopy. C4_nat_HF-7900 was found to form monodisperse particles by cryo EM,
200 with the 3D reconstruction modeled to 3.7Å resolution (Figure 2d, Figure S5-S7). Both
201 the crystal structure of C3_nat_HF-0005 and the model of the cryo-EM reconstruction of
202 C4_nat_HF-7900 show very good matches near the oligomeric hub of the protein where
203 side chains are clearly resolved and as expected. However, it can be seen that they
204 deviate from the design model at the most distal portions of the structure. This is likely
205 due to the inherent flexibility of the unsupported terminal helices of the DHRs^{17,23,30} and
206 lever arm effects which increase with increasing distance from the fusion site (Figure
207 S15).

208 To extend the complexity of structures that can be generated, we built libraries of
 209 heteromeric two chain building block by fusing repeat proteins to two hetero-dimeric
 210 helical bundles (DHD-13, DHD-37)²¹ (Figure 1a). The fusion steps are identical, except
 211 for an additional step of merging the chain A and chain B fusions and checking for
 212 clashes and incompatible residues. In total, 2740 heterodimers were generated *in silico*
 213 to be part of the library. While the homo-oligomeric fusions are good building blocks for
 214 objects with higher order point group symmetries, hetero-oligomeric fusions are needed
 215 at segments without symmetry, such as building cyclic structures and/or connecting
 216 different axes of symmetry in higher order architectures (described below).

217 With a sufficiently high design success rate, the individual oligomers do not need
 218 to be experimentally verified before being used to build larger structures. Since all
 219 building blocks terminate in repeat proteins which can be fused anywhere along their
 220 length, the total number of possible three building block fusions which can be built from
 221 this set is extremely large, which could offset the degree of freedoms lost to symmetry
 222 constraints. The combined library for higher order oligomers consists of both HelixDock
 223 and HelixFuse generated building blocks; overall, the HelixFuse structures tended to
 224 have smaller interfaces across the junction, and thus less overall hydrophobicity than
 225 those generated by HelixDock. While the HelixFuse are less globular than their
 226 HelixDock counterparts, the smaller interface may contribute to the higher fraction of
 227 designs being soluble (~70% vs ~55%). The HelixDock method also requires an
 228 additional step of building a new loop between the HB and DHR, which is another
 229 potential source of modeling error, and takes significantly more computational hours.

230 Overall, the final fraction with single dominant species in SEC traces (examples shown
231 in Figure S1-S5) profiles are similar (~35%).

232

Figure 2. Homo-oligomer diversification by repeat protein fusion. Central oligomer units are shown in red and fused DHRs in blue. Design of (a) C3_HD-1069 (designed loop shown in yellow), (b) C3_HF_Wm-0024A (additional WORMS fusion shown in yellow), and (c) C3_nat_HF-0005. Overlay of the design model (purple/grey) and crystal structure (yellow/white) shows the overall match of the backbone. Inset shows the correct placement of the rotamers in the designed junction region. Design of higher order oligomer fusions (d) C4_nat_HF-7900 and (e) C5_HF-3921 as characterized by cryo-EM. C4_nat_HF-7900 design model (purple/grey) and Cryo-EM map (yellow/white), with inset highlighting the high resolution (~3.8Å) density. C5_HF-3921 inset showing density surrounding the designed junction. (f) C5_HF-2101, (g) C5_HF-0019, (h) C6_HF-0075, and (i) C6_HF-0080 showed good overall match to its negative-stain EM 2D class averages (top) from one direction; predicted projection map for comparison on the bottom.

247 **Section II: Assembly of higher order symmetric structures from repeat** 248 **protein-helical bundle fusion building blocks**

249 To generate a wide range of novel protein assemblies without interface design,
250 we took advantage of the protein interfaces in the library of building blocks described in
251 the previous section, which are oligomers with repeat protein arms. Assemblies are
252 formed by splicing together alpha helices of the repeat protein arms in different building
253 blocks. In our implementation, the user specifies a desired architecture and the
254 symmetries and connectivity of the constituent building blocks. The method then iterates
255 through splices of all pairs of building blocks at all pairs of (user specified, see methods)
256 helical positions; this very large set is filtered on the fly based on the rms of the spliced
257 helices, a clash check, off-architecture angle tolerance, residue contact counts around
258 splice, helix contact count, and redundancy; all of which can be user specified
259 parameters (see methods). The rigid body transform associated with each splice
260 passing the above criteria is computed; for typical pairs of building blocks allowing 100
261 residues, $100 \times 100 = 10,000$ unfiltered splices are possible.

262 Assemblies of these building blocks are modeled as chains of rigid bodies, using
263 the transform between coordinate frames of entry and exit splices, as well as transform
264 between entry splice and coordinate frames of the building blocks. Assemblies are built,
265 in enumerative fashion or with monte carlo, by simple matrix multiplication. For
266 efficiency, only prefiltered splices are used. This technique allows billions of potential
267 assemblies to be generated per cpu hour. Criteria for a given assembly design problem
268 can include any operation defined on the rigid body positions of the building blocks. In

269 this work, we use the transform from the start and end building blocks. To form Cx cyclic
 270 oligomers, the rotation angle of the transform must be $360/x$, and the translation along
 271 the rotation axis must be zero. To form tetrahedral, octahedral, icosahedral, and
 272 dihedral point group symmetries from cyclic building blocks, the symmetry axes of the
 273 start and end building blocks must intersect, and form the appropriate angle for the
 274 desired point group; for example, a 90° angle creates dihedral symmetry.

275 This rapid symmetric architecture assembler through building block fusion has
 276 been implemented in a program called WORMS (Wm) which provides users with
 277 considerable control over building block sets, geometric tolerances, and other
 278 parameters and enables rapid generation of a wide range of macromolecular
 279 assemblies. The desired architecture is entered as a config file (or command line
 280 option) in the following format illustrated for a 3-part fusion with icosahedral symmetry:

```
281      ['C3_N',orient(None,'N')),('Het:CN',orient('C','N')),('C2_C',orient('C',None))]  
282      Icosahedral(c3=0,c2=-1)
```

283 The architecture is specified first, here an icosahedral structure constructed from
 284 a C3 and a C2 building block, and then how the selected building blocks types from the
 285 loaded databases are to be linked together (like a worm). In this example, a C3 building
 286 block with an available N-terminus 'C3_N' is to be fused to a hetero-dimeric building
 287 block 'Het:CN' via an available C-terminus, and the N-terminus of the same 'Het:CN' is
 288 in turn to be fused to a third C2 building block 'C2_C' through an available C-terminus.
 289 The 'None' designation marks that there are no additional unique connections to be
 290 made on that segment. Through the assignment of 'c3=0' and 'c2=-1', the first and last
 291 building blocks are declared as the C3 component and the C2 component, respectively.

The building blocks are cached the first time they are read in from the database files, which can range from a single entry per type to thousands, due to the combinatorial nature of the first fusion step. See supplementary information for more details regarding additional options, architecture definitions, and database syntax. With hundreds to thousands of building blocks each with ~100 residues available for fusion, the total number of three way fusions is on the order of greater than 10^{14} , so optimization of efficiency in both memory usage and CPU requirements was critical in WORMS software development.

Once building block combinations are identified that generate the designed architecture (within a user specifiable tolerance), explicit atomic coordinates are calculated and used for clash checking, redundancy filtering, and any other filtering that requires atomic coordinates. Models for each assembly passing user specified tolerances are constructed in Rosetta, scored and output for subsequent sequence design.

306

Generation of cyclic “crowns” (Crn): We generated C3, C4, and C5 assemblies with WORMS using two designed heterodimer fusions from HelixFuse, as described above. This resulted in head-to-tail cyclic ring structures (Figure 3a), generated by the following configuration (C3 as an example):

```
[('Het:CN',orient(None,'N')),('Het:CN',orient('C',None))]  
Cyclic(3)
```

Following fusion, the junction residues were redesigned to favor the fusion geometry and filtered as above. Seven C3s, seven C4s, and eight C5s were selected

315 and tested experimentally. All yielded soluble protein, and 6, 2, and 1 respectively
316 showed a single peak at the expected elution volume via SEC. We solved the structure
317 of the C3_Crn-05 to 3.19Å resolution (Figure 3b). The overall topology is as designed
318 and the backbone geometry at each of the three junctions is close to the design model.
319 A deviation at the tip of the undesigned heterodimeric HB is likely to be due to crystal
320 packing. C5_Crn-07 chromatographed as a single peak by SEC and was found to be
321 predominantly C5 by negative-stain EM (Figure 3d), but minor off-target species (C4,
322 C6, and C7) were also observed (Figure S16). Each of these structures experimentally
323 verifies three distinct helical fusions (two HelixFuse, one WORMS) from a previously
324 unverified building block library.

325 To further increase the diversity of the crown structures, we recursively ran
326 HelixFuse on both termini of C5_Crn-07 (Figure 3c). Six (6) N-terminal and 24
327 C-terminal fusions were selected and experimentally tested. All were soluble, but had
328 large soluble aggregate fractions when analyzed by SEC. When the peaks around the
329 expected elution volumes were analyzed by negative-stain EM, ring-like structures were
330 found in many of the samples. To facilitate EM structure determination, we combined a
331 c-terminal fusion (C5_Crn_HF-12) and an n-terminal (C5_Crn_HF-26) fusion to
332 generate C5_Crn_HF-12_26 (Figure 3c), which resulted in a much cleaner and
333 monodisperse SEC profile (Figure S17). Cryo-electron microscopy of 12_26 revealed
334 the major population of C5 (77%) structures in addition to C4 (1%), D5 (8%), and C6
335 (12%) subpopulations (Figure S17). We hypothesize that the D5 structure is due to
336 transient interactions of histidines placed on the loops for protein purification. The final

337 3D reconstruction to 5.6Å resolution shows that the major characteristics of the design
338 model are present, despite some splaying of the undesigned portion of the
339 heterodimeric HB relative to the design model (Figure 3d).

340

Figure 3. Design of cyclic "crown" (Crn) structures from heterodimeric building blocks. (a) Hetero-dimeric HB (green/yellow) fused with different DHRs (shades of blue) were fused together using WORMS by enforcing a specific overall cyclic symmetry (C3 and C5 shown). (b) The backbones of the crystal structure (yellow/white) of C3_Crn-05 overlaid with the design model (purple/grey). Insets show the backbone matching focused at each of the fusion locations.

349 (c) A C5 crown (C5_Crn-07, asymmetric unit in red) was fused to DHR units on either
 350 external ("C5_Crn_HF-12", blue arrow) or internal termini ("C5_Crn_HF-26", dark blue
 351 arrow). The two structures were then merged together to generate a double fusion
 352 ("C5_Crn_HF-12_26", black arrow). (d) Cryo-EM class average of the fused 12_26
 353 structure; the major C5 species shown. 3D reconstruction shows the main features of
 354 the designed structure are present, as is also evident in the class average (right).

355

Generation of two-component dihedral assemblies: Dihedral symmetry

protein complexes are attractive building blocks for making higher order 2D arrays and 3D crystal protein assemblies, and can be useful for receptor clustering in cellular engineering³¹. We first set out to design dihedral protein assemblies of D2 symmetry. A set of C2 homo-oligomers with DHR termini (described above) were fused with select *de novo* hetero-dimers (tj18_asym13, unpublished work) using WORMS (schematics shown in Figures 4a-b). The D2 rings harbored total 8 protein chains with 2 chains (two-component) as the asymmetric unit. To generate these rings, we used a database of building blocks containing 7 homo-dimers and 1 heterodimer using the following configuration:

```
[('C2_C',orient(None,'C')),('Het:CN',orient('N','C')),('C2_N',orient('N',None))
]
D2(c2=0, c2b=-1)
```

Of 208 outputs, we selected 6 designs to test, out of which three expressed as soluble two-component protein assemblies as indicated by Ni-NTA pulldown and subsequent SDS-PAGE experiments. Of these, two designs (designated as D2_Wm-01 and D2_Wm-02) eluted as expected by SEC and had SAXS profiles that matched with the designed models (Figure S18 & 19).

To characterize the structures of D2_Wm-01 and D2_Wm-02 in more detail, we performed negative-stain EM and subsequent 2D averaging and 3D refinement. 2D averaging shows the resemblance of the designed model with the experiment-determined structures, whereas 3D refinement indicated accurate design of D2_Wm-01 and D2_Wm-02 at ~16 Å resolution (Figure 4c, Figure S19).

379 The homo dimeric building blocks used in D2_Wm-01 and D2_Wm-02 have large
380 interface areas (~35 residues long; 5 heptads). We sought to reduce the interface area
381 by truncating the helices to facilitate expression of the components and reduce off target
382 interactions. Deletion of one heptad from either of the homodimers of D2_Wm-01
383 (designated D2_Wm-01_trunc) resulted in a single and much narrower SEC peak of the
384 expected molecular weight (Figure S18). Negative-stain EM followed by 2D averaging
385 and 3D refinement indicated monodispersed particles with accurate structure as of the
386 designed model (Figure 4c).

Figure 4. Design of two-component dihedral rings using WORMS (Wm). (a) Two different homodimeric HBs (red) with DHR extensions (shades of blue) were aligned to their respective symmetrical axes with dihedral symmetry. An additional heterodimer (green/yellow) was placed between them and systematically scanned and fused together to design an 8-chain D2 ring. (b) The final asymmetric unit shown in green/yellow while the inset preserves the original colors. (c) Negative-stain EM followed by 2D average and 3D reconstruction of D2_Wm-01 and D2_Wm-01_trunc show that the major features of the designs were recapitulated (left) designed model, (middle) overlay of the designed models with the 3D reconstructions, (right) 2D averages.

Generation of one-component tetrahedral protein cages: Idealized ankyrin

homo-dimers²⁵ based on ANK1 and ANK3 and selected HBs²⁰ were combined to design one-component tetrahedral cages capable of hosting engineered DARPIN binding sites. For each combination, a monomeric ankyrin that perfectly matches the homo-dimer backbone was added as a spacer in between the homo-oligomers, thus extending the ankyrin homo-dimer by several repeats (Figure 5a). To set up this architecture, the following configuration can be used:

```
[('C2_N',orient(None,'N')),('Monomer',orient('C','N')),('C3_C',orient('C',None))
]
Tetrahedral(c2=0, c3=-1)
```

Due to the relatively small space of possibilities because of the limited building block set, only 27 valid fusion combinations were identified, of which 20 involved ankyrin homo-dimer extension at its N-terminus and the remaining 7 at its C-terminus. Eight (8) were selected by manual inspection for further sequence design at fusion regions and experimental characterization.

All 8 constructs were expressed and two were found to be soluble with mono-disperse elution profile peaks by SEC. The two promising structures were very similar, containing different helical bundles whose backbone geometry was identical, but with different internal hydrogen-bond networks. As the two were so similar, only one (T_Wm-1606) was selected for negative-stain EM and discrete particles were observed whose 2D class averages and 3D reconstruction to 20Å matched the computational model (Figure 5b). There was also good agreement between experimental SAXS profiles and profiles computed from the design model (Figure S20).

422

423 **Generation of two-component icosahedral protein cages:** Point group
 424 symmetry nanocages have been successfully designed using docking followed by
 425 interface design⁵⁻⁷. To build such structure using our building blocks with the smaller
 426 and weaker interfaces that give rise to cooperative assembly³²⁻³⁴, we systematically split
 427 each DHR at the loop in the center of four repeats, resulting in a hetero-dimeric
 428 structure with two repeats on each side. The resulting interfaces are considerably
 429 smaller than in for example our *de novo* designed helical bundles. The WORMS
 430 protocol was then applied using the C5, C3, and C2 HelixFuse libraries described above
 431 at their corresponding tetrahedral, octahedral, and icosahedral symmetry axes. The split
 432 DHRs were then sampled to be connected in the center to each of the two symmetrical
 433 oligomers (Figure 5c), using the configuration described above. Following fusion,
 434 sequence design was performed at each of the two new junctions.

435 57 total designs were selected for experimental characterization; 25 co-eluted by
 436 Ni-NTA chromatography, and of these 7 designs had large peaks in the void volume in
 437 SEC chromatography as expected for particles of this size. When the peaks were
 438 collected and re-analyzed with a Sephacryl 500 column, one design, I32_Wm-42
 439 (icosahedral architecture) was resolved into a void and a resolved peak (Figure S21).
 440 Cryo-EM analysis of the resolved peak reveals well formed particles that when
 441 reconstructed to 9Å resolution, accurately match the design model, including the distinct
 442 “S” shaped turn between the C3 and C2 axes (Figure 5d). This structure is considerably
 443 more open than previous icosahedral cages built by designing non-covalent interfaces

444 between homo-oligomers. For another design, T32_Wm-24, while cage was not formed,
 445 we were able to crystallize the polar-capped trimer component (C3_HF_Wm-0024A)
 446 and solve the structure by x-ray diffraction to 2.69Å (Figure 2B). The structure clearly
 447 shows that both of the newly designed junctions (from HelixFuse and WORMS) are as
 448 designed, matching the design model.

449 The 120 subunit I32_Wm-42 icosahedral nanocage has a molecular weight of 3.4
 450 MDa and a diameter of 42.7 nm and illustrates the power of our combined hierarchical
 451 approach. I32_Wm-42 is constructed from five building blocks (two helical bundles and
 452 three repeat proteins) combined via four unique rigid junctions; the EM structure
 453 demonstrates that all were modeled with reasonable accuracy. The combination of the
 454 HelixDock and HelixFuse helix fusion methods created a large set of over 1500
 455 oligomeric building blocks from which WORMS was able to identify combinations and
 456 fusion points that generated the icosahedral architecture; this example is notable
 457 because none of the oligomeric building blocks had been previously characterized
 458 experimentally. With fewer unknowns, either using less segments or a larger fraction of
 459 previously validated building blocks, we expect considerable improvement of the overall
 460 success rate.

Figure 5. Design of assemblies with point group symmetry through helical fusion with WORMS. (a) Tetrahedron design schematic. A HB and a C2 homo-oligomeric made from ankyrin repeat proteins were aligned to their respective tetrahedral symmetry axis (red), and connected via fusion to Ankyrin repeat monomers (blue) to generate the target architecture. (b) 3D reconstruction reveals a well fitting map of T_Wm-1606. (c) Icosahedral design schematic. Libraries of unverified cyclic fusion homo-dimers and trimers were aligned to the corresponding icosahedral symmetry axes. Using WORMS, fusions to DHRs split in the center that hold the two homo-oligomers in the orientations which generate icosahedral structures were identified. (d) Cryo-EM 3D reconstruction of I32_Wm-42 closely matches the designed model.

473 **Discussion:**

474 Our general rigid helix-fusion based pipeline fulfills the promise of early
475 proposals^{16,35} in providing a robust and accurate procedure for generating large protein
476 assemblies by fusing symmetric building blocks and avoiding interface design, and
477 should streamline assembly design for applications in vaccine development, drug
478 delivery and biomaterials more generally. The set of structures generated here goes
479 considerably beyond our previous work with rigid helical fusions¹⁸, and the “WORMS”
480 software introduced here is quite general and readily configurable to different
481 nanomaterial design challenges. WORMS can be easily extended to other symmetric
482 assemblies including 2D arrays and 3D crystals, and should be broadly useful for
483 generating a wide range of protein assemblies.

484 DNA nanotechnology has had advantages in modularity and simplicity over
485 protein design because the basic interactions (Watson-Crick base pairing) and local
486 structures (the double helix) are always the same. Proteins in nature exhibit vast
487 diversity compared to duplex DNA, and correspondingly, re-engineering naturally
488 occurring proteins and designing new ones has been a more complex task than
489 designing new DNA structures. The large libraries of “clickable” building blocks-- helical
490 bundle - repeat protein fusions-- and the generalized WORMS software for assembling
491 these into a wide range of user specifiable architectures that we present in this paper
492 are a step towards achieving the modularity and simplicity of DNA nanotechnology with
493 protein building blocks. Although this modularity comes at some cost in that the building
494 blocks are less diverse than proteins in general, they can be readily functionalized by

495 fusion to protein domains with a wide range of functions. We show that it is possible to
 496 genetically fuse DHR “adapters” to natural proteins; these proteins can then be used in
 497 larger assemblies through WORMS with less likelihood of disrupting the original protein
 498 fold. Proteins of biological and medical relevance (binders like protein A, enzymes, etc.)
 499 can be used as components and combined with *de novo* designed HBs and DHRs to
 500 form nanocages and other architectures.

501 Moving forward, there are still a variety of challenges to address. The larger the
 502 set of building blocks for WORMS the more precisely the geometric constraints
 503 associated with the desired architecture can be achieved, and hence it is advantageous
 504 to use the very large *in silico* libraries of building blocks that can be created by helical
 505 bundle - repeat protein fusion rather than the very much smaller sets of fusions that can
 506 be experimentally characterized in advance (tens of thousands compared to tens). It will
 507 be important to understand how uncertainties in the structures of the *in silico* fusions
 508 translate into uncertainties in the structures of the resulting architectures, and more
 509 generally, how to further improve the fusion approach so that the *in silico* structures are
 510 nearly perfectly realized. As the assemblies become more complex with different
 511 building blocks and total number of subunits, more alternative structures become
 512 possible. Understanding how to achieve cooperative assembly and controlling for
 513 specificity of the desired assembly over alternatives will be an increasingly important
 514 challenge as the complexity of the target nanomaterials increases.

515

516

517 **Computational Methods Summary:**

518 **RosettaRemodel Forward Folding:** To test the extent to which the designed
 519 sequences encode the designed structure around the junction site, we used large scale
 520 *de novo* folding calculations. Due to computational limitations with standard full chain
 521 forward folding^{36,37}, we developed a similar but alternate approach for larger symmetric
 522 structures. Using RosettaRemodel²⁷ in symmetry mode (reversing the anchor residue
 523 for cases where the helical bundle was at the C-terminus), we locked all residues
 524 outside the junction region as rigid bodies, only allowing 40 residues starting from the
 525 end of the HB in the primary sequence direction of the DHR to be re-sampled. The
 526 blueprint file was set up to be agnostic of secondary structure in this segment of protein
 527 and we deleted all DHR residues past the first two helices after the rigid body region to
 528 reduce CPU cost. Each structure was set to at least 2000 trajectories to create a
 529 forward folding funnel.

530 **WORMS:** The WORMS software overall requires two inputs, a database of
 531 building block entries (format described in Supplementary Information in detail) and a
 532 configuration file (or command line options) as described in the main text to govern the
 533 overall architecture. While some segments can be of single building blocks of interest,
 534 to generate a wide variety of outputs, tens to thousands of entries per segment should
 535 be used. The number of designs generated also depends on the number of fusion
 536 points allowed, as the size of the space being sampled increases multiplicatively with
 537 the number of segments being fused. There are many options available to the user to
 538 control the fusions which are output as solutions; we have tuned the default options to

539 be relatively general-use (see Supplementary Information for description of options). A
 540 key parameter is the *tolerance*, the allowed deviation of the final segment in the final
 541 structure away from its target position given the architecture. For different geometries
 542 the optimal values vary; for example the same tolerance values involve more drastic
 543 error in icosahedral symmetry than cyclic symmetry. The WORMS code is specifically
 544 designed to generate fusions that have a protein core around the fusion joint; unless
 545 specified using the *ncontact_cut*, *ncontact_no_helix_cut*, and *nhelix_contacted_cut*
 546 option set, the code will not produce single extended helix fusions.

547

548 **Brief Experimental Methods:**

549 **Gene preparation:** All amino acid sequences derived from Rosetta were reverse
 550 translated to DNA sequences and placed in the pET29b+ vector. For two-component
 551 designs, all designs were initially constructed for bi-cistronic expression by appending
 552 an additional ribosome binding site (RBS) in front of the second sequence with only one
 553 of the components containing a 6xHis tag. Genes were synthesized by commercial
 554 companies: Integrated DNA Technologies (IDT), GenScript, Twist Bioscience, or Gen9.

555 **Protein expression and purification:** All genes were cloned into *E. coli* cells
 556 (BL21 Lemo21 (DE3)) for expression, using auto-induction³⁸ at 18° or 37°C for 16-24
 557 hours in 500mL scale. Post-induction, cultures were centrifuged at 8,000xG for 15
 558 minutes. Cell pellets were then resuspended in 25-30mL lysis buffer (TBS, 25mM Tris,
 559 300mM NaCl, pH8.0, 30mM imidazole, 0.25mg/mL DNase I) and sonicated for 2
 560 minutes total on time at 100% power (10 sec on/off) (QSonica). Lysate was then

centrifuged at 14,000xG for 30 minutes. Clarified lysates were filtered with a 0.7um syringe filter and put over 1-4mL of Ni-NTA resin (QIAGEN), washed with wash buffer (TBS, 25mM Tris, 300mM NaCl, pH8.0, 60mM imidazole), then eluted with elution buffer (TBS, 25mM Tris, 300mM NaCl, pH8.0, 300mM imidazole). Eluate was then concentrated with a 10,000 m/w cutoff spin concentrator (Millipore) to approximately 0.5mL based on yield for SEC.

D2 proteins went through an extra round of bulk purification. Concentrated protein was heated at 90 °C for 30 minutes to further separate bacterial contaminants. Samples were then allowed to cool down to room temperature and any denatured contaminants were removed by centrifuging at 20,000xG.

Size exclusion chromatography (SEC): All small oligomers were passed through a Superdex200 Increase 10/300 GL column (Cytiva) while larger assemblies were passed through a Superose 6 Increase 10/300 GL column (Cytiva) on a AKTA PURE FPLC system. The mobile phase was TBS (TBS, 25mM Tris, 300mM NaCl). Additionally, for the icosahedral assembly, an additional custom packed 10/300 Sephacryl500 column (Cytiva) was used to separate out the void. Samples were run at a speed of 0.75mL/min and eluted with 0.5mL fractions.

Protein Characterization: See supplementary information for detailed methods regarding SAXS sample preparation, electron microscopy, and x-ray crystallography.

580

581 **ACKNOWLEDGMENTS**

582 This work was supported by the National Science Foundation (NSF) award 1629214
583 (DB), a generous gift from the Audacious Project, the Open Philanthropy Project
584 Improving Protein Design Fund, the National Institute of General Medical Sciences
585 (R01GM120553 to D.V.), the National Institute of Allergy and Infectious Diseases
586 (HHSN272201700059C to DV), a Pew Biomedical Scholars Award (DV), an
587 Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs
588 Wellcome Fund (DV) and the University of Washington Arnold and Mabel Beckman
589 cryo-EM center. YH was supported in part by a NIH Molecular Biology Training Grant
590 (T32GM008268). RM is a recipient of the Washington Research Foundation (WRF)
591 Innovation fellowship and his research is funded in part by the US DOE BES Energy
592 Frontier Research Center CSSAS (The Center for the Science of Synthesis Across
593 Scales) located at the University of Washington (award number DESC0019288). UN
594 was supported in part by PHS NRSA (T32GM007270) from NIGMS. AC is a recipient of
595 the Human Frontiers Science Program Long Term Fellowship and a Washington
596 Research Foundation Senior Fellow.

597 This work was conducted at the Advanced Light Source (ALS), a national user facility
598 operated by Lawrence Berkeley National Laboratory on behalf of the Department of
599 Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis
600 Technologies (IDAT) program, supported by DOE Office of Biological and Environmental
601 Research. Additional support comes from the National Institute of Health project
602 ALS-ENABLE (P30 GM124169) and a High-End Instrumentation Grant S10OD018483.

603 We thank staff at Advanced Photon Source beamline NE-CAT 24-ID-E for data
604 collection. Northeastern Collaborative Access Team beamline supported by NIH grants
605 P30GM124165 and S10OD021527, and DOE contract DE-AC02-06CH11357. We also
606 want to thank Banumathi Sankaran at the Advanced Light Source (ALS) beamline 8.2.2
607 at Lawrence Berkeley National Laboratory for data collection. The Berkeley Center for
608 Structural Biology is supported in part by the National Institutes of Health (NIH),
609 National Institute of General Medical Sciences, and the Howard Hughes Medical
610 Institute. The Advanced Light Source (ALS) is supported by the Director, Office of
611 Science, Office of Basic Energy Sciences and US Department of Energy under contract
612 number DE-AC02-05CH11231.

613 We thank Kristen Dancel-Manning and Alice Liang of the NYU Microscopy Laboratory,
614 William Rice and Bing Wang of the NYU Cryo-EM Laboratory, Kashyap Maruthi, Ed
615 Eng, Laura Yen, and Misha Kopylov of the New York Structural Biology Center, and
616 members of the Bhabha/Ekiert labs for assistance with grid screening and data
617 collection and helpful discussions. We especially thank Nicolas Coudray of the
618 Bhabha/Ekiert lab for helpful discussions and guidance regarding EM data processing.
619 Some of this work was performed at the Simons Electron Microscopy Center and
620 National Resource for Automated Molecular Microscopy located at the New York
621 Structural Biology Center, supported by grants from the Simons Foundation
622 (SF349247), NYSTAR, and the NIH National Institute of General Medical Sciences

623 (GM103310) with additional support from Agouron Institute (F00316), NIH (OD019994),
624 and NIH (RR029300).

625 We would like to thank the Rosetta@Home user base for donating their computational
626 hours to run our forward folding simulations. Thanks to George Ueda for the
627 unpublished tj18_asym13 heterodimer. An additional thanks to Robby Divine and Josh
628 Lubner for support in the documentation and development WORMS.

629

630 **AUTHOR CONTRIBUTIONS**

631 YH, RM, and DB wrote the manuscript. YH, WS, and TB developed the HelixDock
632 protocol; YH, RM, NIE, IV, and UN made designs and characterized experimentally with
633 assistance from ET, AS, and CMC in protein production. YH and TB developed the
634 HelixFuse protocol. IV developed the helical fusion method in .NET; WS and DB
635 implemented it into the WORMS protocol; YH and RM assisted developing in its
636 application. UN and ET crystallized and MJB solved the C3_HD-1069 structure. AK
637 crystallized C3_nat_HF-0005, C3_HF_Wm-0024A, and C3_Crn-05. AB solved the
638 crystal structure for C3_nat_HF-0005 and C3_HF_Wm-0024A. MJB solved the structure
639 for C3_Crn-05. RLR, assisted by DE and GB, performed negative-stain and cryo-EM for
640 all HelixFuse structures presented. YH designed and characterized crown structures
641 and icosahedral cage; RM the dihedral structures, IV the tetrahedral cage. RM and AC
642 performed initial EM screening of dihedral, cyclic and icosahedral WORMS structures.

643 YJP, assisted by DV, performed negative-stain and cryo-EM for all WORMS structures
644 presented. DB guided the project.

645

646 **ONLINE CONTENT**

647 Crystallography data:

648 C3_HD-1069 (**6XT4**)

649 C3_HF_Wm-0024A (**6XI6**)

650 C3_nat_HF-0005 (**6XH5**)

651 C3_Crn-05 (**6XNS**)

652 Electron microscopy data:

653 C4_nat_HF-7900 (**6XSS, EMD-22305**)

654 C5_HF_3921 (**EMD-22306**)

655 C5_Crn_HF-12_26 (**EMD-XXXX**)

656 D2_Wm-01 (**EMD-XXXX**)

657 D2_Wm-01_trunc (**EMD-XXXX**)

658 D2_Wm-02 (**EMD-XXXX**)

659 T_Wm-1606 (**EMD-XXXX**)

660 I32_Wm-42 (**EMD-XXXX**)

661 **COMPETING INTERESTS**

662 The authors declare no competing interests.

663

664 **ADDITIONAL INFORMATION**

665 Supplementary files:

666	Supplementary information	*.pdf
667	HelixDock sequence design	*.xml, *.symdef
668	HelixDock loop closure	*.xml
669	HelixFuse	*.xml
670	WORMS sequence design	*.xml, *.xml
671	Design models	*.pdb (as zip files)

672

REFERENCES:

1. Baker, D. What has de novo protein design taught us about protein folding and biophysics?
Protein Sci. **28**, 678–683 (2019).
2. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
3. Fallas, J. A. *et al.* Computational design of self-assembling cyclic protein homo-oligomers.
Nat. Chem. **9**, 353–360 (2017).
4. Sahasrabudhe, A. *et al.* Confirmation of intersubunit connectivity and topology of designed
protein complexes by native MS. *Proc. Natl. Acad. Sci.* **115**, 1268–1273 (2018).
5. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials.
Nature **510**, 103–108 (2014).
6. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein
complexes. *Science* **353**, 389–394 (2016).
7. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535**, 136–139
(2016).
8. Shen, H. *et al.* De novo design of self-assembling helical protein filaments. *Science* **362**,
705–709 (2018).
9. Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays
mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365–1368 (2015).
10. Ueda, G. *et al.* Tailored Design of Protein Nanoparticle Scaffolds for Multivalent
Presentation of Viral Glycoprotein Antigens.
<http://biorxiv.org/lookup/doi/10.1101/2020.01.29.923862> (2020)
doi:10.1101/2020.01.29.923862.
11. Marcandalli, J. *et al.* Induction of Potent Neutralizing Antibody Responses by a Designed

698 Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17
699 (2019).

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

702 13. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with
703 atomic level accuracy. *Science* **336**, 1171–1174 (2012).

704 14. Leaver-Fay, A. *et al.* Rosetta3. in *Methods in Enzymology* vol. 487 545–574 (Elsevier,
705 2011).

706 15. McConnell, S. A. *et al.* Designed Protein Cages as Scaffolds for Building Multienzyme
707 Materials. *ACS Synth. Biol.* **9**, 381–391 (2020).

708 16. Youn, S.-J. *et al.* Construction of novel repeat proteins with rigid and predictable
709 structures using a shared helix method. *Sci. Rep.* **7**, 2595 (2017).

710 17. Brunette, T. *et al.* Modular repeat protein sculpting using rigid helical junctions. *Proc.*
711 *Natl. Acad. Sci.* **117**, 8870–8875 (2020).

712 18. Vulovic, I. *et al.* Generation of ordered protein assemblies using rigid three-body fusion.
713 <http://biorxiv.org/lookup/doi/10.1101/2020.07.18.210294> (2020)
714 doi:10.1101/2020.07.18.210294.

715 19. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical
716 bundles. *Science* **346**, 481–485 (2014).

717 20. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular
718 hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).

719 21. Chen, Z. *et al.* Programmable design of orthogonal protein heterodimers. *Nature* **565**,
720 106–111 (2019).

721 22. Thomson, A. R. *et al.* Computational design of water-soluble α -helical barrels. *Science*

- 722 **346**, 485–488 (2014).
- 723 23. Brunette, T. *et al.* Exploring the repeat protein universe through computational protein
724 design. *Nature* **528**, 580–584 (2015).
- 725 24. Boyken, S. E. *et al.* De novo design of tunable, pH-driven conformational changes.
726 *Science* **364**, 658–664 (2019).
- 727 25. Fallas, J. A. *et al.* Computational design of self-assembling cyclic protein
728 homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).
- 729 26. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J.*
730 *Mol. Biol.* **234**, 946–950 (1993).
- 731 27. Huang, P.-S. *et al.* RosettaRemodel: A Generalized Framework for Flexible Backbone
732 Protein Design. *PLoS ONE* **6**, e24109 (2011).
- 733 28. Coventry, B. & Baker, D. *Protein sequence optimization with a pairwise decomposable*
734 *penalty for buried unsatisfied hydrogen bonds.*
735 <http://biorxiv.org/lookup/doi/10.1101/2020.06.17.156646> (2020)
736 doi:10.1101/2020.06.17.156646.
- 737 29. Fullerton, S. W. B. *et al.* Mechanism of the Class I KDPG aldolase. *Bioorg. Med. Chem.*
738 **14**, 3002–3010 (2006).
- 739 30. Geiger-Schuller, K. *et al.* Extreme stability in de novo-designed repeat arrays is
740 determined by unusually stable short-range interactions. *Proc. Natl. Acad. Sci.* **115**,
741 7539–7544 (2018).
- 742 31. Correnti, C. E. *et al.* Engineering and functionalization of large circular tandem repeat
743 protein nanoparticles. *Nat. Struct. Mol. Biol.* **27**, 342–350 (2020).
- 744 32. Zlotnick, A. To Build a Virus Capsid. *J. Mol. Biol.* **241**, 59–67 (1994).
- 745 33. Zlotnick, A., Johnson, J. M., Wingfield, P. W., Stahl, S. J. & Endres, D. A Theoretical

746 Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly [†]. *Biochemistry*
747 **38**, 14644–14652 (1999).

748 34. Ceres, P. & Zlotnick, A. Weak Protein–Protein Interactions Are Sufficient To Drive
749 Assembly of Hepatitis B Virus Capsids [†]. *Biochemistry* **41**, 11525–11531 (2002).

750 35. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self
751 assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci.* **98**,
752 2217–2221 (2001).

753 36. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β
754 sheets. *Science* **355**, 201–206 (2017).

755 37. Marcos, E. & Silva, D.-A. Essentials of *de novo* protein design: Methods and
756 applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1374 (2018).

757 38. Studier, F. W. Protein production by auto-induction in high-density shaking cultures.
758 *Protein Expr. Purif.* **41**, 207–234 (2005).