Rational design of alpha-helical tandem repeat proteins with closed architectures

Lindsey Doyle[1], **Jazmine Hallinan**[1], **Jill Bolduc**[1], **Fabio Parmeggiani**[2,3], **David Baker**[2,3,4], **Barry L. Stoddard**[1], and **Philip Bradley**[1,3,5,*]

[1]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. Seattle WA 98109, USA

[2]Department of Biochemistry, University of Washington, Seattle WA 98195, USA

[3]Institute for Protein Design, University of Washington, Seattle WA 98195, USA

[4]Howard Hughes Medical Institute, University of Washington, Seattle WA 98195, USA

[5]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. Seattle WA 98019, USA

Abstract

Tandem repeat proteins, which are formed by repetition of modular units of protein sequence and structure, play important biological roles as macromolecular binding and scaffolding domains, enzymes, and building blocks for the assembly of fibrous materials[1,2]. The modular nature of repeat proteins enables the rapid construction and diversification of extended binding surfaces by duplication and recombination of simple building blocks[3,4]. The overall architecture of tandem repeat protein structures – which is dictated by the internal geometry and local packing of the repeat building blocks – is highly diverse, ranging from extended, super-helical folds that bind peptide, DNA, and RNA partners[5–9], to closed and compact conformations with internal cavities suitable for small molecule binding and catalysis[10]. Here we report the development and validation of computational methods for *de novo* design of tandem repeat protein architectures driven purely by geometric criteria defining the inter-repeat geometry, without reference to the sequences and structures of existing repeat protein families. We have applied these methods to design a series of closed alpha-solenoid[11] repeat structures (alpha-toroids) in which the inter-repeat packing geometry is constrained so as to juxtapose the N- and C-termini; several of these

designed structures have been validated by X-ray crystallography. Unlike previous approaches to tandem repeat protein engineering[12_20], our design procedure does not rely on template sequence or structural information taken from natural repeat proteins and hence can produce structures unlike those seen in nature. As an example, we have successfully designed and validated closed alpha-solenoid repeats with a left-handed helical architecture that – to our knowledge – is not yet present in the protein structure database[21].

Engineered proteins that contain closed repeat architectures represent a natural target for rational, geometry-guided design of repeat modules (Fig. 1) for several reasons. Closure results from simple constraints on the inter-repeat geometry: if we consider the transformation between successive repeats as being composed of a rotation (curvature) about an axis together with a translation (rise) parallel to that axis, then the rise must equal zero and the curvature multiplied by the number of repeats must equal 360°. Closed structures are stabilized by interactions between the first and last repeats, which obviates the need for capping repeats to maintain solubility and may make them more tolerant to imperfections in the designed geometry than open repeat architectures. Closed repeat arrays offer the advantages of rotational symmetry (for example, in generating higher-order assemblies) with the added control provided by a covalent linkage between subunits. Conversely, it may be possible to convert a monomeric closed repeat protein array into a symmetrical protein assembly by truncation (for example, converting a toroidal protein containing 'n' repeats into an equivalent homodimeric assembly containing 'n/2' repeats per subunit) if economy of protein length is required.

We developed an approach to geometry-guided repeat protein design (Fig. 2) that is implemented in the Rosetta molecular modeling package[22] and builds on published *de novo* design methodologies[23]. Key features include symmetry of backbone and side chain conformations extended across all repeats (allowing computational complexity to scale with repeat length rather than protein length); a pseudo-energy term that favors the desired inter-repeat geometry; clustering and resampling stages that allow intensified exploration of promising topologies; and an *in silico* validation step that assesses sequence-structure compatibility by attempting to re-predict the designed structure given only the designed sequence. Applying this design procedure produced a diverse array of toroidal structures (Fig. 2). We focused primarily on designs with left-handed bundles (Extended Data Fig. 1) since this architecture (closed, left-handed alpha-solenoid) appears to be absent from the structural database (SI Discussion). We selected five monomeric repeat architectures for experimental characterization: a left-handed 3-repeat family (**dTor_3x33L: d**esigned **Tor**oid with three **33**-residue repeats, **L**eft-handed), left- and right-handed 6-repeat families (dTor_6x35L and dTor_6x33R), a left-handed 9-repeat family (dTor_9x31L), and a left-handed 12-repeat design built by extending one of the 9-repeat designs by 3 repeats (dTor_12x31L). To enhance the likelihood of successful expression, purification, and crystallization, we pursued multiple designed sequences for some families, including a round of surface mutants for three designs that were refractory to crystallization (Extended Data Table 1).

We were able to determine five crystal structures for representatives from four monomeric designed toroid families (Fig. 3, Extended Data Fig. 2, and Extended Data Table 2). Close examination of the electron density for the structures, during and after refinement, indicated that most of these highly symmetrical designed proteins display significant rotational averaging within the crystal lattice (Extended Data Fig. 3), such that the positions corresponding to the loops that connect each repeated module are occupied by a mixture of continuous peptide and protein-termini. This lattice behavior was observed for most of the structures, but only appeared to significantly affect the refinement R-factors for a final multimeric construct (described below) consisting of multiple copies of the first 3 repeats of dTor_9x31L. In all cases, however, the positions and conformations of secondary structure and individual side chains, which are largely invariant from one repeat to the next, were clear and unambiguous in the respective density maps. Kajander *et al.* have reported[24] similar crystal averaging with associated disorder at protein termini in a set of structures for designed consensus TPR repeat proteins, albeit with translational averaging along a fiber axis rather than the rotational averaging observed here.

Comparison of the design models to the experimental crystal structures shows that all four designs form left-handed alpha-helical toroids with the intended geometries. The structural deviation between design model and experimental structure increases with increasing repeat number: from 0.6 Å for the 3-repeat design, to 0.9 Å for the 6-repeat design, to 1.1 Å for the 9-repeat and 12-repeat designs. Inspection of the superpositions in Figure 3 suggests that the design models are slightly more compact than the experimental structures, a discrepancy which becomes more noticeable as the number of repeats increases. This trend may reflect a tendency of the current design procedure to over-pack side chains during the sequence optimization step (perhaps due to under-weighting of repulsive electrostatic or van der Waals interactions). Nevertheless, the success of the 12x31L design implies that, at least for certain repeat modules, it is possible to control the geometry of the central pore by simply varying the number of repeats, without the need to re-optimize the sequence of individual repeats. Further characterization by size exclusion chromatography indicated that the 3- and 6-repeat designs form stable dimers in solution while the 9- and 12-repeat designs form monomers; all are thermostable (Extended Data Table 1 and Extended Data Figs. 4–6). Their behavior did not vary significantly as a function of protein or salt concentration, nor did they display a dynamic equilibrium between monomeric and dimeric states.

Our ability to successfully design several left-handed alpha-toroids demonstrates that the apparent absence of this fold from the current database of solved structures is not due to constraints imposed by the helical solenoid architecture or the toroidal geometry. It is possible that there exist in nature left-handed alpha-toroids whose folds have not been observed; it is also possible that this region of fold space has not been sampled during natural protein evolution. Indeed, left-handed alpha-helical tandem repeat bundles of any kind – open or closed – are rare relative to their right-handed counterparts (which are found in TPR, Armadillo, HEAT, PUF, and PPR structures, among others). Our search for left-handed helical solenoid repeats with multiple turns in the structural database yielded only the TAL effector[6,7] and mTERF[25] DNA binding domains (SI Discussion). The handedness of our designed toroids is due in part to the use of inter-helical turns whose geometry naturally imparts a handedness to the resulting helical bundle. The 3-residue 'GBB' (α_L-β-β)

turn type used in these designs prefers a left-handed dihedral twist between the connected helices, while the 'GB' turn found in dTor_6x33R correlates with right-handed geometry (Extended Data Fig. 1). Both these turn types are also compatible with canonical helix capping interactions[26,27], which may explain their selection by the design procedure (helix capping guarantees satisfaction of backbone polar groups and also strengthens sequence-encoding of local structure).

We explored the feasibility of splitting one of the larger monomeric designs into fragments that can assemble symmetrically to reform complete toroids comprising multiple copies of identical subunits. We selected the structurally characterized 9x31L design to split into a small 3-repeat subfragment, which was expected to then form a trimeric assembly. This 3-repeat fragment was expressed, purified, and formed diffraction-quality crystals. Upon determination of the experimental structure, we discovered that the design fragment formed an unexpected crystal packing arrangement composed of linked tetrameric rings (i.e., containing a total of 12 repeats per ring, Fig. 4a). Indeed, it was this unanticipated finding that led us to synthesize the monomeric 12x31L design whose characterization demonstrated that the designed 31 residue repeat sequence is indeed compatible with both 9- and 12-repeat monomeric toroidal geometries (and presumably 10- and 11- repeat geometries as well). The crystal structure of the 3-repeat fragment suggests that the 12x geometry may be preferred, and indeed this would be consistent with the apparent tendency of our design procedure to over-pack the design models.

We expect that designed alpha toroids may have potential applications as scaffolds for binding and catalysis and as building blocks for higher-order assemblies. Amino acids lining the central pores could be mutated to introduce binding or catalytic functionalities and/or sites of chemical modification. The modular symmetry of monomeric toroids could be exploited to array interaction surfaces with prescribed geometries: a designed interface on the external face of the 12x31L design, for example, could be replicated with 2, 3, 4, or 6-fold symmetry by repeating the interfacial mutations throughout the full sequence. Thus monomeric toroids could replace multimeric assemblies as symmetry centers in the assembly of protein cages; by breaking the symmetry of the interaction surfaces it may be possible to create more complex heterotypic assemblies with non-uniform placement of functional sites. Examination of the crystalline arrangements formed by our designed toroids suggests the potential for creating specific 1- and 2-dimensional assemblies: both the monomeric 9x31L and 12x31L crystals have channels extending continuously through the crystal formed from the pores in vertical stacks of toroids (Fig. 4b–c), with 2-dimensional layers of toroids running perpendicular to these stacks. Interface design could be applied to stabilize the crystal contacts seen in the existing structures thereby further stabilizing either the crystalline state or these 1- or 2-dimensional sub-assemblies[28,29]. Designed toroids with larger pores that crystallize in a similar manner might form crystal structures with channels capable of hosting guest molecules by covalent linkage or noncovalent binding. Stabilization of the concatemeric structure (Fig. 4a) formed by the 3-repeat fragment either by cross-linking or interface design could represent a path toward a variety of novel protein-based materials[30].

METHODS

Computational design

The repeat module design process applied here consists of an initial diversification round of large-scale sampling followed by filtering and clustering and then a second intensification round of sampling focused on successful topologies identified in the first round.

Fragment assembly—Starting backbone models for sequence design are built using a fragment assembly protocol which is based on the standard Rosetta *ab initio* protocol[31] with the following modifications: (1) fragment replacement moves are performed symmetrically across all repeats, guaranteeing that backbone torsion angles are identical at corresponding positions across repeats; (2) a pseudo-energy term (equal to the deviation between actual and desired curvature, in degrees, plus the deviation in rise multiplied by a factor of 5) is added to the potential to favor satisfaction of the geometric constraints; (3) the amino acid sequence used for low-resolution scoring is assigned randomly at the start of each simulation from secondary-structure specific distributions (helix: Ala+Ile+Leu+Asp+Ser, turn: Gly +Ser), which has the effect of increasing the diversity in helix packing distances and geometries compared with using a constant sequence such as poly-Val or poly-Leu. At the start of each independent design trajectory the lengths of the secondary structure elements and turns are chosen randomly, defining the target secondary structure of the repeat module and its length. Together with the number of repeats, this defines the total length of the protein and the complete secondary structure, which is used to select 3 and 9 residue backbone fragments for use in the low-resolution fragment assembly phase. The design calculations reported here sampled helix lengths from 7 to 20 residues, turn lengths from 1 to 5 residues, and total repeat lengths ranging from 20 to 40 residues.

Sequence design—The low-resolution fragment assembly simulation is followed by an all-atom sequence design stage consisting of two cycles alternating between fixed-backbone sequence design and fixed-sequence structure relaxation. Symmetry of backbone and side chain torsion angles and sequence identities is maintained across all repeats. Since the starting backbones for design are built by relatively coarse sampling in a low-resolution potential, sequences designed with the standard all-atom potential are dominated by small amino acids and the resulting structures tend to be under-packed. To correct for this tendency, a softened Lennard-Jones potential[32] is used for the sequence design steps, while the standard potential is used during the relaxation step. The Rosetta *score12prime* weights set was used as the standard potential for these design calculations.

Filtering and clustering—Final design models (typically 10–100,000 in this study) are first sorted by per-residue energy (total energy divided by the number of residues, to account for varying repeat length) and the top 20% are filtered for packing quality (sasapack_score<0.5), satisfaction of buried polar groups (buried unsatisfied donors per repeat<1.5, buried unsatisfied acceptors per repeat<0.5), and sequence-structure compatibility via a fast, low-resolution symmetric refolding test (40 trajectories, requiring at least 1 under an RMSD threshold of 2 Å for 3-repeat designs and 4 Å for larger designs). Designs that pass these filters are clustered by C-alpha RMSD (allowing for register shifts

when aligning helices with unequal lengths) in order to identify recurring architectures. The clusters are ranked by averaging residue energy, packing quality, and refolding success over all cluster members.

Resampling—During the intensification round of designs, representative topologies from successful design clusters are specifically resampled by enforcing their helix and turn lengths as well as their turn conformations (defined using a 5-state, coarse-grained backbone torsion alphabet[27]; Extended Data Fig. 1e) during fragment selection.

Large-scale refolding—Selected low-energy designs from the second round that pass the filters described above are evaluated by a large-scale refolding test in which 2,000–10,000 *ab initio* models are built by standard (asymmetric) fragment assembly followed by all-atom relaxation. Success is measured by assessing the fraction of low energy *ab initio* models with RMSDs to the design model under a length-dependent threshold.

Symmetry-breaking in the central pore—For designed toroids with an open, polar central pore, perfect symmetry may not allow optimal electrostatic interactions between nearby side chains corresponding to the same repeat position in successive repeats. We therefore explored symmetry-breaking mutations at a handful of inward-pointing positions via fixed-backbone sequence design simulations in which the length of the repeating sequence unit was doubled/tripled (for example, whereas perfect 6-fold repeat symmetry would require K-K-K-K-K-K or E-E-E-E-E-E, doubling the repeat length permits charge complementarity with K-E-K-E-K-E). Solutions from these designs were accepted if they significantly lowered the total energy.

Design model for dTor_12x31L—The **12x31L** design construct was generated by duplicating the final 3 repeats of the **9x31L** design. To build a "design model" for comparison with the experimentally determined structure, we followed the resampling protocol now forcing the **12x31L** amino acid sequence in addition to the number of repeats (12) and the helix and turn lengths (H14-L3-H11-L3) and turn conformations (GBB). Thus the sequence design steps were reduced to rotamer optimization (since the amino acid identities were fixed). This symmetric structure prediction process was repeated 10,000 times and the lowest-energy final model was taken as the computational model.

Surface mutations to enhance crystallization—For a single representative of the 3x31L and 6x31R families we performed lattice docking and design simulations to select mutations that might promote crystallization. Core positions were frozen at the design sequence. Candidate space groups were selected from those most commonly observed in the protein structural database. Theoretical models of crystal packing arrangements were built by randomly orienting the design model within the unit cell and reducing the lattice dimensions until clashes were encountered. Symmetric interface design was performed on these docked arrangements, and final designs were filtered by energy, packing, satisfaction of polar groups, and number of mutations from the original design model.

Handedness of tandem repeat helical bundles—To compute the handedness of helical bundles formed by tandem repeat proteins, we generate an approximate helical

bundle axis curve by joining the location of repeat-unit centers of mass in a sliding fashion along the protein chain. The handedness is then estimated by computing the directionality of the winding of the polypeptide chain about this axis curve.

Structural bioinformatics—To assess similarity between design models and proteins in the structural database, we performed searches using the structure-structure comparison program DALI[33] as well as consulting the protein structure classification databases CATH[34], SCOPe[35], and ECOD[36]. Further details are given in SI Discussion.

Code availability—Repeat protein design methods were implemented in the Rosetta software suite (www.rosettacommons.org) and will be made freely available to academic users; licenses for commercial use are available through the University of Washington Technology Transfer office.

Cloning and protein expression

The plasmids encoding individual constructs were cloned into previously described bacterial pET15HE expression vectors[37] containing a cleavable N-terminal His-tag and an ampicillin resistance cassette. Sequence-verified plasmids were transformed into BL21(DE3)RIL *E. coli* cells (Agilent Technologies) and plated on LB medium with ampicillin (100μg/mL). Colonies were individually picked and transferred to individual 10 mL aliquots of LB-Ampicillin media and shaken overnight at 37°C. Individual 10 mL aliquots of overnight cell cultures were added to individual 1 L volumes of LB-Ampicillin, which were then shaken at 37°C until the cells reached an OD_{600} value of 0.6 to 0.8. The cells were chilled for 20 minutes at 4°C, then IPTG was then added to each flask to a final concentration of 0.5 mM to induce protein expression. The flasks were shaken overnight at 16°C, and then pelleted by centrifugation and stored at −20°C until purification.

Construct dTor_6x35L(SeMet), incorporating a single Methionine residue at position 168 in the original design construct, was generated using a 'QuikChange' site-directed mutagenesis kit (Agilent) and corresponding protocol from the vendor. The resulting plasmid construct was again transformed into BL21(DE3)RIL *E. coli* cells (Agilent Technologies) and plated on LB plates containing ampicillin (100μg/mL) and chloramphenicol (35 g/mL). Subsequent cell culture and protein expression in minimal media, along with incorporation of selenomethionine was incorporated during protein expression according to reference[38].

Purification

Cell pellets from 3L of cell culture were resuspended in 60 mL of PBS solution (140 mM NaCl, 2.5 mM KCl, 10 mM NaHPO_4, 2 mM KH_2PO_4) containing 10 mM imidazole (pH 8.0). Cells were lysed via sonication and centrifuged to remove cell debris. The supernatant was passed through a 0.2 micron filter, and then incubated on a rocker platform at 4° C for one hour after adding 3 mL of resuspended Nickel-NTA metal affinity resin (Invitrogen). After loading onto a gravity-fed column, the resin was washed with 45 mL of the same lysis buffer described above, and the protein was eluted from the column with three consecutive aliquots of PBS containing 150 mM imidazole (pH 8.0). Purified protein was concentrated to approximately 5 mg/mL to 25 mg/mL while buffer exchanging into 25mM Tris (pH 7.5)

and 200mM NaCl and then further purified via size exclusion chromatography using HiLoad 16/60 Superdex 200 column (GE).

Protein samples were then split in half; one sample was used directly for crystallization while the other had the His-tag removed by an overnight digest with biotinylated thrombin (Novagen), prior to additional crystallization trials. The digested sample was incubated for 30 minutes with streptavidin-conjugated agarose (Novagen) to remove the thrombin. All samples were tested for purity and removal of the His-tag via SDS PAGE. The final protein samples, both with and without the N-terminal poly-histidine affinity tag, were concentrated to values of 5 mg/mL to 25 mg/mL for crystallization trials.

Solution size and stability analysis

Proteins at a concentration of 4 to 10 mg/ml were run over a Superdex 75 10/300 GL column (GE Healthcare) in 25 mM Tris pH 8.0 plus 100 or 750 mM NaCl at a rate of 0.4 ml/min on an AKTAprime plus chromatography system (GE Healthcare). All fractions containing eluted toroid protein (visualized via electrophoretic gel analyses) were pooled, concentrated and run over the column a second time in order to assess their solution oligomeric behavior using protein with a minimal background of contaminants. Gel filtration standards (Bio-Rad) were run over the same column in in matching buffer, and the UV trace of the proteins was overlaid onto the standards using UNICORN 5 software (GE Healthcare).

For measurements of protein stability using circular dichroism (CD) spectroscopy, purified recombinant toroid constructs were diluted to between 10 to 20 μM concentration and dialyzed overnight into 10 mM potassium phosphate buffer at pH 8.0. Circular dichroism (CD) thermal denaturation experiments were performed on a JASCO J-815 CD spectrometer with a Peltier thermostat. Wavelength scans (190-250 nm) were carried out for each construct at 20°C and 95°C. Additional thermal denaturation experiments were conducted by monitoring CD signal strength at 206 nm over a temperature range of 4°C to 95°C (0.1cm path-length cell), with measurements taken every 2 degrees. Sample temperature was allowed to equilibrate for 30 seconds before each measurement.

Crystallization and data collection

Purified proteins were initially tested for crystallization via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech). Crystallization conditions were then optimized with constructs that proved capable of crystallizing in larger 24-well hanging drops. Out of 11 constructs that were purified to homogeneity, ten were crystallized, of which five yielded high quality x-ray diffraction that resulted in successful structure determination.

dTor_6x35L was crystallized in 160 mM Sodium Chloride, 100 mM Bis-Tris pH 8.5 and 24% (w/v) Polyethylene Glycol 3350 at a concentration of 26 mg/mL. The crystal was transferred to a solution containing 300 mM, then 500 mM Sodium Chloride and flash frozen in liquid nitrogen. Data was collected on a R-AXIS IV++ at wavelength 1.54 Å and processed on HKL2000[39].

dTor_6x35L(SeMet) was crystallized in 140 mM Sodium Chloride, 100 mM Tris pH 8.5 and 22% (w/v) Polyethylene Glycol 3350 at a concentration of 26 mg/mL. The crystal was transferred to a solution containing 300 mM, then 500 mM Sodium Chloride and flash frozen in liquid nitrogen. Data was collected at ALS Beamline 5.0.2 at wavelength 0.9794 Angstroms and processed on HKL2000[39].

dTor_3x33L_2-2 was crystallized in two different conditions, producing two different crystal lattices. The underlined first condition had 30% Polyethylene Glycol 3350, 100 mM Tris pH 6.5, 200mM NaCl with a protein concentration of 1.8mM. The protein was soaked in a 15% Ethylene Glycol cryoprotectant for one minute prior to being flash frozen in liquid nitrogen. Data was collect on a Saturn 944+ (Rigaku) at wavelength 1.54 Å for 180 degrees at phi=0 and another 180 degrees at phi=180. Data was then processed on HKL2000[39] out to 1.85Å in space group P212121.

The underlined second condition had 45% Polyethylene Glycol 400 and 100 mM Tris pH 7.7 with a protein concentration of 1.8mM. Protein crystal was flash frozen without being cryoprotected. Data was collect on a Saturn 944+ (Rigaku) at wavelength 1.54 Å for 180 degrees at phi = 0 and another 180 degrees at phi = 180. Data was then processed on HKL2000[39] out to 1.85Å in space group P43212.

dTor_9x31L_sub was crystallized in 100 mM Tris pH 8.5 and 15% (v/v) Ethanol at a concentration of 11.5 mg/mL. The crystal was transferred to a solution containing 75 mM Tris pH 8.5, 7.5% (v/v) Ethanol and 25% (v/v) Glycerol and flash frozen in liquid nitrogen. Data was collected at ALS Beamline 5.0.2 at wavelength 1.0 Å and processed on HKL2000[39] out to 2.9Å in space group P 41 21 2/P 43 21 2.

dTor_9x31L was crystallized in 0.1 M Sodium Citrate pH 5.4 and 1.0 M Ammonium Phosphate Monobasic at a concentration of 8.8 mg/mL in 3 ul drops containing 1 ul protein and 2 ul well solution. The crystal was transferred to a solution containing the well plus 25% (v/v) Glycerol and flash frozen in liquid nitrogen. Data was collected on a Saturn 944+ CCD at wavelength 1.54 Å and processed on HKL2000[39] out to 2.5Å in space group P 21 21 21.

dTor_12x31L was crystallized in 0.9 M Sodium malonate pH 7.0, 0.1 M HEPES pH 7.0 and 0.5% Jeffamine ED-2001 pH 7.0 at a concentration of 8.8 mg/mL in 2 ul drops containing 1 ul protein and 1 ul well solution. The crystal was transferred to a solution containing 0.675 M Sodium malonate pH 7.0, 0.075 M HEPES pH 7.0, 0.375% Jeffamine ED-2001 pH 7.0 and 25% Glycerol, and flash frozen in liquid nitrogen. Data was collected on a Saturn 944+ CCD at wavelength 1.54 Å and processed on HKL2000[39] out to 2.3Å in space group R 3:H.

Phasing and refinement

The dTor_6x35L and both dTor_3x33L_2-2 structures were solved by Molecular Replacement with Phaser[40] via CCP4i[41] using the Rosetta-designed structure as a search model. The structures were then built and refined using Coot[42] and Refmac5[43], respectively.
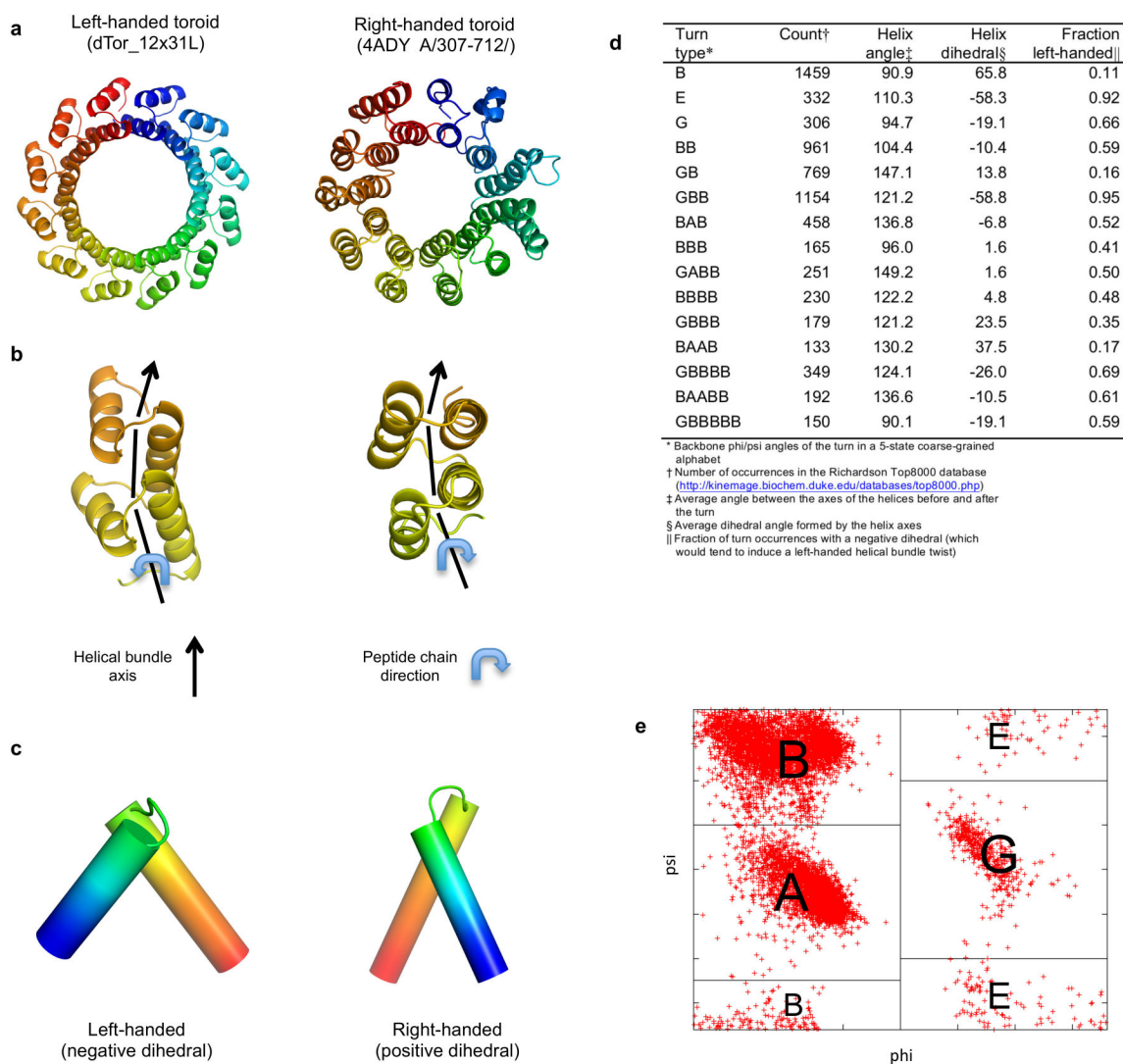
The structure of dTor_6x35L(SeMet) was solved by Molecular Replacement with Phaser[40] via PHENIX[44] using the best refined model of dTor_6x35L as a phasing model. The structure was then built and refined using Coot[42] and PHENIX[45], respectively.

The structures of dTor_9x31L_sub and dTor_9x31L were solved by Molecular Replacement with Phaser[40] via PHENIX[44] using the Rosetta-designed structure as a search model. The structure was then built and refined using Coot[42] and PHENIX[45], respectively.

The structure of dTor_12x31L was solved by Molecular Replacement with Phaser[40] via PHENIX[44] using a 4 repeat subunit the Rosetta-designed structure as a search model. The structure was then built and refined using Coot[42] and PHENIX[45], respectively.

Final Ramachandran statistics after refinement were as follows (given as %preferred, %allowed, %outliers): dTor_6x35L(SeMet): 98.06, 1.94, 0.0; dTor_3x33L_2-2a: 99.48, 0.0, 0.52; dTor_3x33L_2-2b: 98.96, 0.52, 0.52; dTor_9x31L_sub: 98.31, 1.69, 0.0; dTor_9x31L: 99.28, 0.36, 0.36; dTor_12x31L: 99.0, 1.0, 0.0.

Extended Data

Turn type*	Count†	Helix angle‡	Helix dihedral§	Fraction left-handed‖
B	1459	90.9	65.8	0.11
E	332	110.3	-58.3	0.92
G	306	94.7	-19.1	0.66
BB	961	104.4	-10.4	0.59
GB	769	147.1	13.8	0.16
GBB	1154	121.2	-58.8	0.95
BAB	458	136.8	-6.8	0.52
BBB	165	96.0	1.6	0.41
GABB	251	149.2	1.6	0.50
BBBB	230	122.2	4.8	0.48
GBBB	179	121.2	23.5	0.35
BAAB	133	130.2	37.5	0.17
GBBBB	349	124.1	-26.0	0.69
BAABB	192	136.6	-10.5	0.61
GBBBBB	150	90.1	-19.1	0.59

* Backbone phi/psi angles of the turn in a 5-state coarse-grained alphabet
† Number of occurrences in the Richardson Top8000 database (http://kinemage.biochem.duke.edu/databases/top8000.php)
‡ Average angle between the axes of the helices before and after the turn
§ Average dihedral angle formed by the helix axes
‖ Fraction of turn occurrences with a negative dihedral (which would tend to induce a left-handed helical bundle twist)
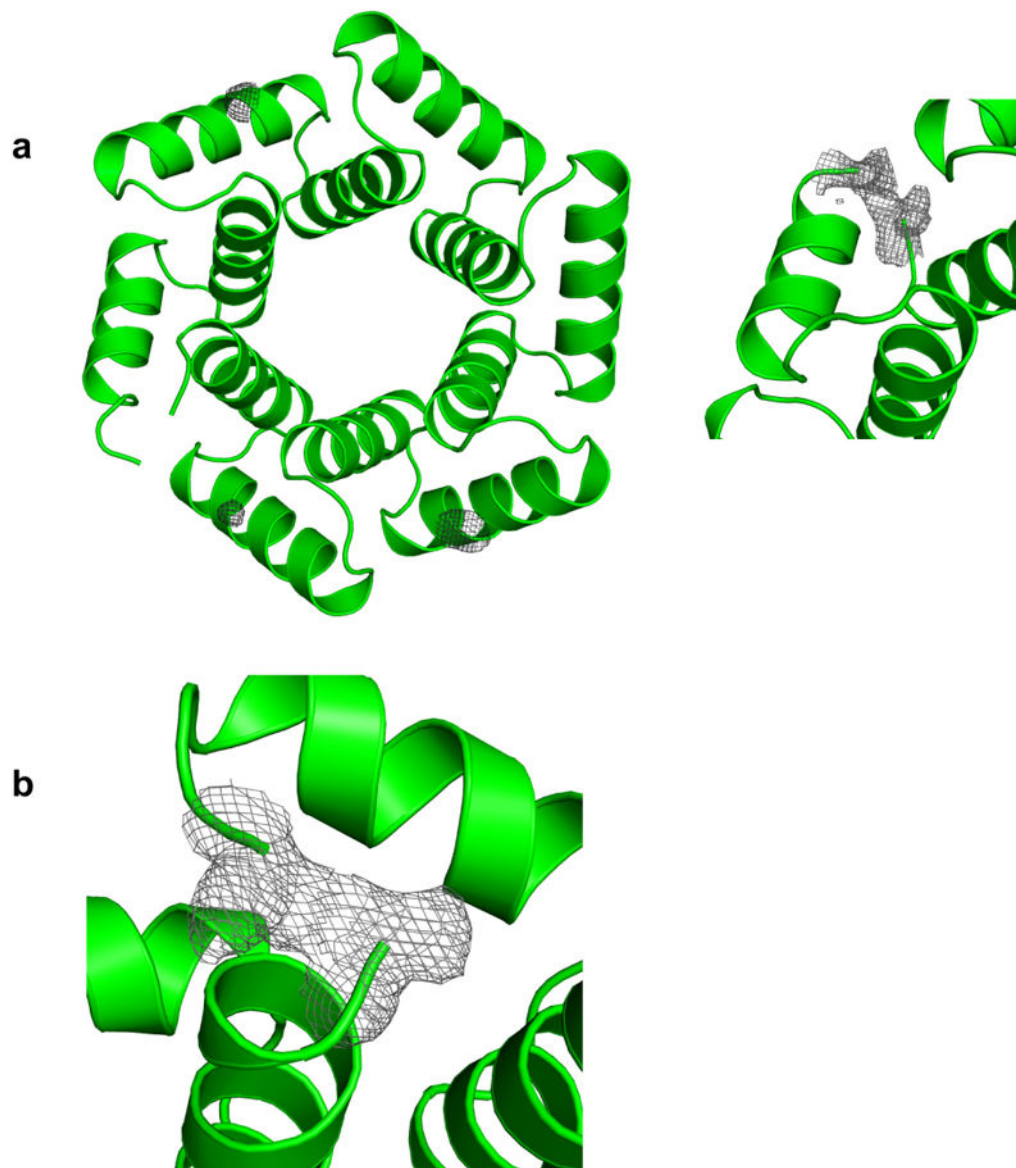
Extended Data Figure 1.

Handedness of alpha-helical bundles and helical linkers. (a) Design dTor_12x31L, shown on the left, has a left-handed helical bundle. The native toroid on the right, which has a right-handed bundle, is taken from the PDB structure 4ADY and corresponds to the PC repeat domain of the 26S proteasome subunit Rpn2[46]. (b) The handedness of a helical bundle is determined by the twist direction of the polypeptide chain as it wraps around the axis of the helical bundle. (c) Helical linkers characterized by a negative (positive) dihedral angle between the axes of the connected helices will, upon repetition, tend to impart a left-handed (right-handed) twist to the bundle. (d) Geometrical properties of the most common short alpha-helical linkers in the structural database indicate that certain turn types (e.g., 'E' and 'GBB') tend to form left-handed connections while others (e.g., 'GB' and 'BAAB') are associated with right-handed connections. Turn types are classified by mapping their backbone torsion angles to a coarse-grained alphabet[27] as shown in (e).
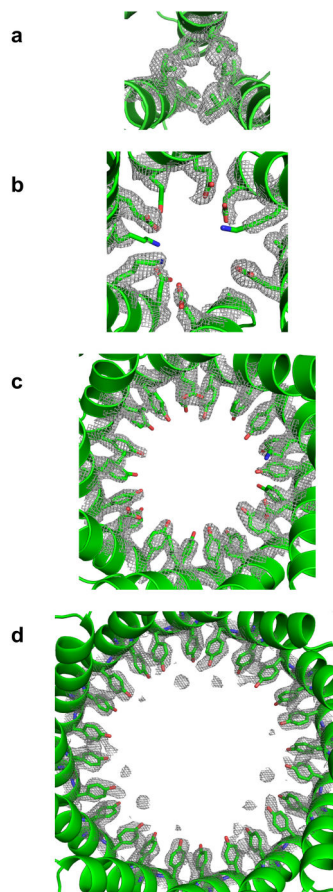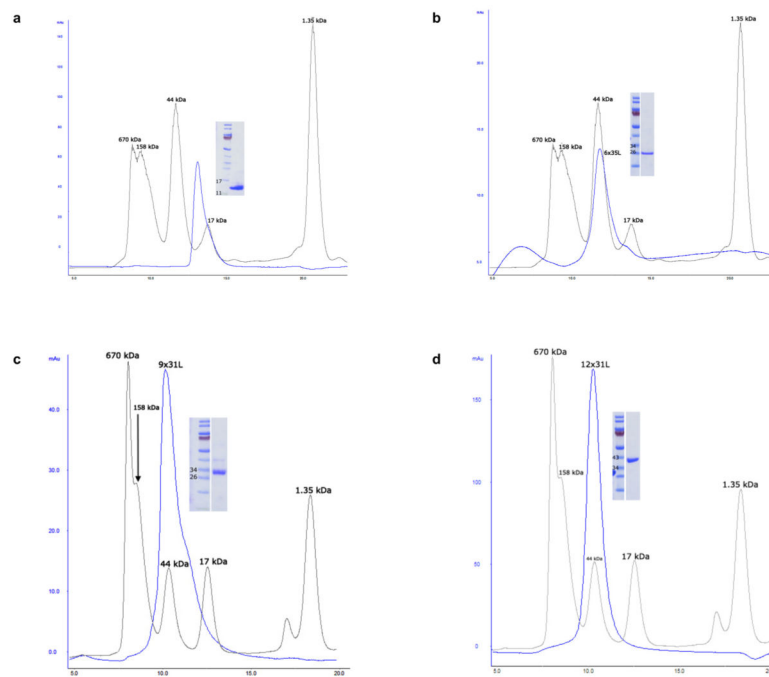

Extended Data Figure 2.

Unbiased 2Fo-Fc omit maps contoured around the side chains comprising the central pore regions for each crystallized toroid. The constructs shown are in the same order as in Figure 3.
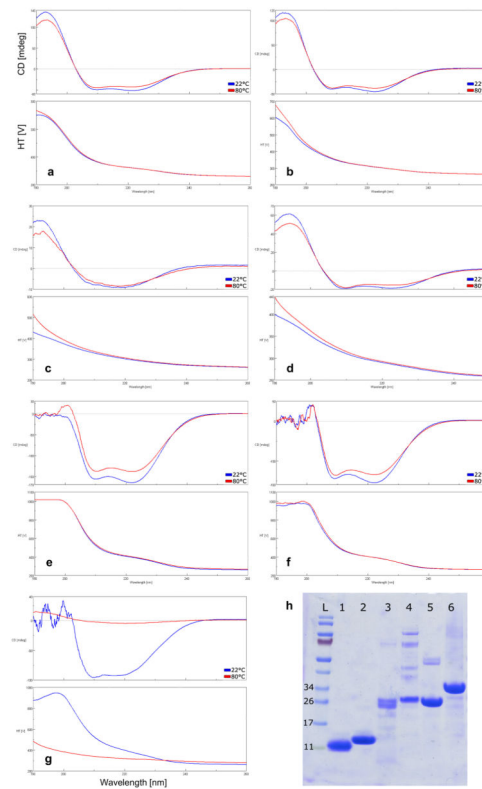
Extended Data Figure 3.
The crystallographic structures of highly symmetrical designed toroidal repeat proteins
display rotational averaging in the crystal lattice. (a) Electron difference density for
construct dTor_6x35L. The left panel shows anomalous difference Fourier peaks calculated
from data collected from a crystal of selenomethionine-derivatized protein. Although only
one methionine residue (at position 168) is present in the construct, strong anomalous
difference peaks (I/σI greater than 4.0) are observed at equivalent positions within at least 3
modular repeats. The right panel shows difference density extending across the modeled
position of the N- and C-termini in the refined model, indicating partial occupancy at that
position by a peptide bond. The other five equivalent positions around the toroidal protein
structure display equivalent features of density, indicating that each position is occupied by a
mixture of loops and protein termini. (b) Electron density for construct dTor_12x31L, again

calculated at a position corresponding to the refined N- and C-termini in the crystallographic model. As was observed for the hexameric toroid in (a), the electron density indicates a mixture of loops and protein termini.
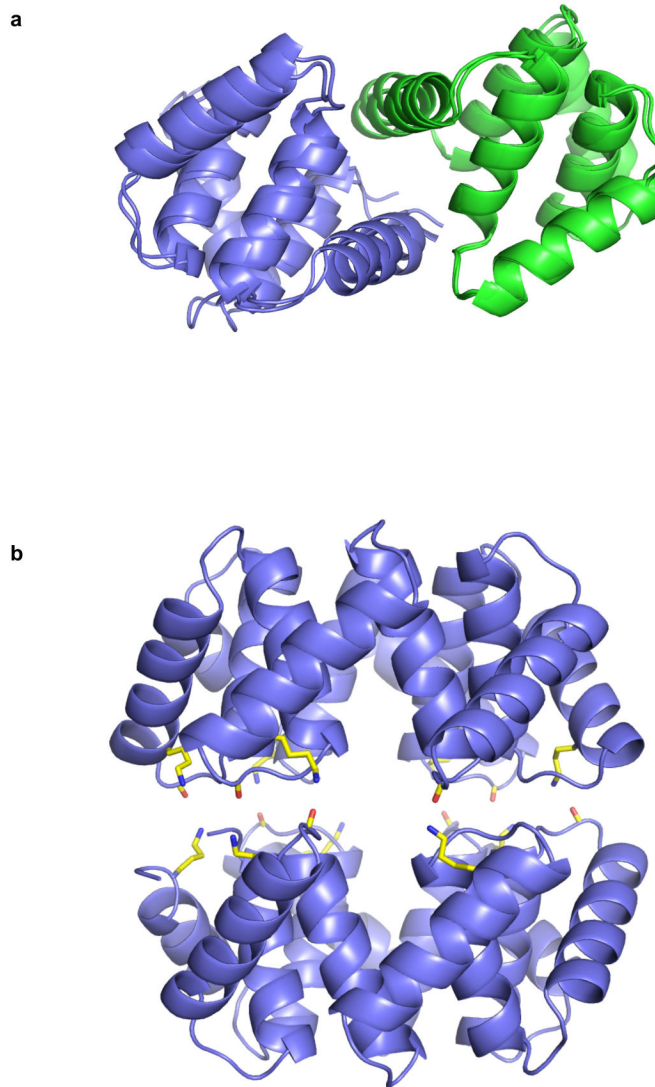

Extended Data Figure 4.
Size exclusion chromatography elution profiles for the four designed toroids whose crystal structures were determined. The elution profiles (blue traces) shown correspond to runs in high (750 mM) NaCl for dTor_3x33L_2-2 (a) and dTor_6x35L (b), while the elution profiles for dTor_9x31L (c) and dTor_12x31L (d) correspond to runs in lower (150 mM) NaCl. The superimposed elution profiles of standard protein size markers (brown traces) correspond to runs at those same salt concentrations, conducted on the same column and day. The inset in each panel displays the migration and relative purity of each construct used for the analysis.

Extended Data Figure 5.
Purification and characterization of designed toroids. (a–g) CD wavelength scan from 260-190 nm of several designed toroids and a positive control protein at 22°C (blue) and 80°C (red). (a) dTor_9x31L_sub; (b) dTor_3x33L_2-2; (c) dTor_6x33R_1; (d) dTor_6x35L; (e) dTor_9x31L; (f) dTor_12x31L; (g) positive control. (h) Bis-Tris Gel (4-12%) showing designed toroids immediately following metal affinity purification. Lane L, Mw protein standards (kDa); lane 1, dTor_9x31L_sub; lane 2, dTor_3x33L_2-2; lane 3, dTor_6x33R_1; lane 4, dTor_6x35L; lane 5, dTor_9x31L; lane 6, dTor_12x31L.

Extended Data Figure 6.
Potential dimerization interfaces observed in crystal packing interactions. (a) Superposition of monomer-monomer packing interactions for the dTor_3x33L_2-2 design observed in two entirely different crystal forms. (b) Stacking interactions between two dTor_6x35L subunits observed in the crystal structure; lysine residues interacting with backbone carbonyl groups in the partner monomer are shown in stick representation and colored yellow along with their interaction partners.

Extended Data Table 1

Characterization of designed constructs

ID	No. of repeats	Repeat length	Bundle handedness	Expressed[*]	Purified[†]	Oligomeric state[‡]	Crystals[§]	Structure[‖]
dTor_9x31L_sub[¶]	3	31	Left	Y	Y	M/D[#]	Y	Y
dTor_3x33L_1	3	33	Left	Y	Y		Y	N

ID	No. of repeats	Repeat length	Bundle handedness	Expressed*	Purified[†]	Oligomeric state[‡]	Crystals[§]	Structure[//]
dTor_3x33L_1-1	3	33	Left	Y	Y		N	
dTor_3x33L_2	3	33	Left	Y	Y		Y	N
dTor_3x33L_2-1	3	33	Left	Y	N			
dTor_3x33L_2-2	3	33	Left	Y	Y	D	Y	Y
dTor_3x33L_2-3	3	33	Left	Y	N			
dTor_3x33L_2-4	3	33	Left	Y	N			
dTor_3x33L_3	3	33	Left	Y	N/A			
dTor_6x33R_1	6	33	Right	Y	Y		Y	N
dTor_6x33R_1-1	6	33	Right	Y	N			
dTor_6x33R_1-2	6	33	Right	Y	N			
dTor_6x33R_1-3	6	33	Right	Y	N			
dTor_6x33R_2	6	33	Right	Y	N			
dTor_6x33R_3	6	33	Right	Y	N			
dTor_6x33R_4	6	33	Right	N				
dTor_6x35L	6	35	Left	Y	Y	D	Y	Y
dTor_6x35L(SeMet)	6	35	Left	Y	Y		Y	Y
dTor_9x31L	9	31	Left	Y	Y	M	Y	Y
dTor_12x31L	12	31	Left	Y	Y	M	Y	Y

*
Construct was successfully overexpressed

[†]Construct was successfully purified to homogeneity and concentrated to at least 1 mg/mL.

[‡]Dominant solution species, as assessed by size-exclusion chromatography (SEC; Extended Data Fig. 4); M: monomer, D: dimer.

[§]Construct crystallized

[//]Crystals diffracted and structure determination was successful

[¶]The 3-repeat subfragment of dTor_9x31L

[#]Concentration-dependent monomer/dimer equilibrium

Extended Data Table 2

Crystallographic Statistics

	dTor_6x35L	dTor_6x35L(SeMet)	dTor_3x33L_2-2a	dTor_3x33L_2-2b	dTor_9x31L_sub	dTor_9x31L	dTor_12x31L
Data collection *							
Space group	C 2 2 2₁	C 2 2 2₁	P 2₁ 2₁ 2₁	P 4₃ 2₁ 2	P 4₃ 2₁ 2	P 2₁ 2₁ 2₁	C 2
Cell dimensions							
a, b, c (Å)	63.5, 85.3, 80.5	63.5, 85.1, 80.5	37.1,68.6,152.4	40.2, 40.2, 217.7	102.8, 102.8, 93.9	41.7, 72.0, 86.2	95.4, 119.4, 76.3
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 110.9, 90.0
Resolution (Å) [†]	50.0–2.26 (2.30–2.26)	50.0–2.18 (2.26–2.18)	50.00–1.85 (1.90–1.85)	50–2.78 (2.88–2.78)	50.0–3.2 (3.3–3.2)	50.0–2.50 (2.54–2.50)	50.0–2.50 (2.54–2.50)
R_{merge}	0.045 (0.159)	0.059 (0.323)	0.056 (0.500)	0.048 (0.136)	0.056 (0.461)	0.079 (0.292)	0.048 (0.298)
$I/\sigma I$	39.9 (13.8)	29.7 (8.41)	20.3 (4.34)	27.0 (15.0)	31.3 (6.48)	30.4 (5.66)	27.2 (3.7)
Completeness (%)	98.1 (97.9)	99.7 (99.2)	90.6 (95.9)	98.9 (98.2)	100.0 (100.0)	99.2 (91.2)	98.9 (87.6)
Redundancy	3.8 (3.6)	13.7(11.6)	6.0 (7.0)	12.3 (10.6)	14.8 (15.1)	10.0 (4.50)	3.7 (3.0)
Refinement							
Resolution (Å)		43.0–2.18 (2.23–2.18)	76.2–1.85 (1.90–1.85)	54.42–2.78 (2.85–2.78)	29.95–3.2 (3.7–3.2)	29.98–2.5 (2.6–2.5)	30.6–2.5 (2.59–2.50)
No. reflections		11137	29249	4760	8662	9355	27183
R_{work}/R_{free}		23.8/29.6	22.7/28.2	19.3/26.7	29.96/34.5	22.5/32.8	21.42/25.4
No. atoms							
Protein		1476	3038	1480	2292	2011	5608
Ligand/ion		-	8	-	-	-	-

	dTor_6x35L	dTor_6x35L(SeMet)	dTor_3x33L_2-2a	dTor_3x33L_2-2b	dTor_9x31L_sub	dTor_9x31L	dTor_12x31L
Water	-	139	50	-	-	166
B-factors							
Protein	43.7	36.6	26	108.2	35.9	42.1
Ligand/ion	-	61	-	-	-	-
Water	-	52.4	56	-	-	43.8
R.m.s deviations							
Bond lengths (Å)	0.0142	0.017	0.017	0.002	0.008	0.002
Bond angles (°)	1.6908	1.708	1.918	0.5	1.038	0.49

*
Each structure was determined from a single crystal.

†
Highest resolution shell is shown in parenthesis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol. 1999; 293:151–160. DOI: 10.1006/jmbi.1999.3136 [PubMed: 10512723]

2. Kajava AV. Tandem repeats in proteins: from sequence to structure. J Struct Biol. 2012; 179:279–288. DOI: 10.1016/j.jsb.2011.08.009 [PubMed: 21884799]

3. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. J Struct Biol. 2001; 134:117–131. DOI: 10.1006/jsbi.2001.4392 [PubMed: 11551174]

4. Grove TZ, Cortajarena AL, Regan L. Ligand binding by repeat proteins: natural and designed. Curr Opin Struct Biol. 2008; 18:507–515. DOI: 10.1016/j.sbi.2008.05.008 [PubMed: 18602006]

5. Wang X, McLachlan J, Zamore PD, Hall TM. Modular recognition of RNA by a human pumilio-homology domain. Cell. 2002; 110:501–512. [PubMed: 12202039]

6. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. The crystal structure of TAL effector PthXo1 bound to its DNA target. Science. 2012; 335:716–719. DOI: 10.1126/science. 1216211 [PubMed: 22223736]

7. Deng D, et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science. 2012; 335:720–723. DOI: 10.1126/science.1215670 [PubMed: 22223738]

8. Barkan A, et al. A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. PLoS Genet. 2012; 8:e1002910. [PubMed: 22916040]

9. Reichen C, Hansen S, Pluckthun A. Modular peptide binding: from a comparison of natural binders to designed armadillo repeat proteins. J Struct Biol. 2014; 185:147–162. DOI: 10.1016/j.jsb. 2013.07.012 [PubMed: 23916513]

10. Wierenga RK. The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett. 2001; 492:193–198. [PubMed: 11257493]

11. Kobe B, Kajava AV. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem Sci. 2000; 25:509–515. [PubMed: 11050437]

12. Main ER, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of stable alpha-helical arrays from an idealized TPR motif. Structure. 2003; 11:497–508. [PubMed: 12737816]

13. Binz HK, et al. High-affinity binders selected from designed ankyrin repeat protein libraries. Nat Biotechnol. 2004; 22:575–582. DOI: 10.1038/nbt962 [PubMed: 15097997]

14. Parmeggiani F, et al. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J Mol Biol. 2008; 376:1282–1304. DOI: 10.1016/j.jmb.2007.12.014 [PubMed: 18222472]

15. Urvoas A, et al. Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (alphaRep) based on thermostable HEAT-like repeats. J Mol Biol. 2010; 404:307–327. DOI: 10.1016/j.jmb.2010.09.048 [PubMed: 20887736]

16. Boersma YL, Pluckthun A. DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr Opin Biotechnol. 2011; 22:849–857. DOI: 10.1016/j.copbio.2011.06.004 [PubMed: 21715155]

17. Ramisch S, Weininger U, Martinsson J, Akke M, Andre I. Computational design of a leucine-rich repeat protein with a predefined geometry. Proc Natl Acad Sci U S A. 2014; 111:17875–17880. DOI: 10.1073/pnas.1413638111 [PubMed: 25427795]

18. Voet AR, et al. Computational design of a self-assembling symmetrical beta-propeller protein. Proc Natl Acad Sci U S A. 2014; 111:15102–15107. DOI: 10.1073/pnas.1412768111 [PubMed: 25288768]

19. Park K, et al. Control of repeat-protein curvature by computational protein design. Nat Struct Mol Biol. 2015; 22:167–174. DOI: 10.1038/nsmb.2938 [PubMed: 25580576]

20. Parmeggiani F, et al. A general computational approach for repeat protein design. J Mol Biol. 2015; 427:563–575. DOI: 10.1016/j.jmb.2014.11.005 [PubMed: 25451037]

21. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]

22. Leaver-Fay A, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–574. DOI: 10.1016/B978-0-12-381270-4.00019-6 [PubMed: 21187238]

23. Koga N, et al. Principles for designing ideal protein structures. Nature. 2012; 491:222–227. DOI: 10.1038/nature11600 [PubMed: 23135467]

24. Kajander T, Cortajarena AL, Mochrie S, Regan L. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. Acta Crystallogr D Biol Crystallogr. 2007; 63:800–811. DOI: 10.1107/S0907444907024353 [PubMed: 17582171]

25. Jimenez-Menendez N, et al. Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. Nat Struct Mol Biol. 2010; 17:891–893. DOI: 10.1038/nsmb. 1859 [PubMed: 20543826]

26. Aurora R, Rose GD. Helix capping. Protein Sci. 1998; 7:21–38. DOI: 10.1002/pro.5560070103 [PubMed: 9514257]

27. Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in proteins. J Mol Biol. 1996; 255:235–253. DOI: 10.1006/jmbi.1996.0020 [PubMed: 8568871]

28. Grove TZ, Regan L, Cortajarena AL. Nanostructured functional films from engineered repeat proteins. J R Soc Interface. 2013; 10 ARTN 20130051.

29. Lanci CJ, et al. Computational design of a protein crystal. Proc Natl Acad Sci U S A. 2012; 109:7304–7309. DOI: 10.1073/pnas.1112595109 [PubMed: 22538812]

30. Abe S, Ueno T. Design of protein crystals in the development of solid biomaterials. Rsc Adv. 2015; 5:21366–21375. DOI: 10.1039/c4ra16748a

31. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997; 268:209–225. DOI: 10.1006/jmbi.1997.0959 [PubMed: 9149153]

32. Dantas G, et al. High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Biol. 2007; 366:1209–1221. DOI: 10.1016/j.jmb.2006.11.080 [PubMed: 17196978]

33. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci. 1995; 20:478–480. [PubMed: 8578593]

34. Sillitoe I, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015; 43:D376–381. DOI: 10.1093/nar/gku947 [PubMed: 25348408]

35. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014; 42:D304–309. DOI: 10.1093/nar/gkt1240 [PubMed: 24304899]

36. Cheng H, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014; 10:e1003926. [PubMed: 25474468]

37. Mak AN, Lambert AR, Stoddard BL. Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kI. Structure. 2010; 18:1321–1331. DOI: 10.1016/j.str.2010.07.006 [PubMed: 20800503]

38. Walden H. Selenium incorporation using recombinant techniques. Acta Crystallogr D. 2010; 66:352–357. DOI: 10.1107/S0907444909038207 [PubMed: 20382987]

39. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Method Enzymol. 1997; 276:307–326. DOI: 10.1016/S0076-6879(97)76066-X

40. Mccoy AJ, et al. Phaser crystallographic software. J Appl Crystallogr. 2007; 40:658–674. DOI: 10.1107/S0021889807021206 [PubMed: 19461840]

41. Winn MD, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr D. 2011; 67:235–242. DOI: 10.1107/S0907444910045749 [PubMed: 21460441]

42. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D. 2010; 66:486–501. DOI: 10.1107/S0907444910007493 [PubMed: 20383002]

43. Skubak P, Murshudov GN, Pannu NS. Direct incorporation of experimental phase information in model refinement. Acta Crystallogr D. 2004; 60:2196–2201. DOI: 10.1107/S0907444904019079 [PubMed: 15572772]

44. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D. 2010; 66:213–221. DOI: 10.1107/S0907444909052925 [PubMed: 20124702]

45. Afonine PV, et al. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D. 2012; 68:352–367. DOI: 10.1107/S0907444912001308 [PubMed: 22505256]

46. He J, et al. The structure of the 26S proteasome subunit Rpn2 reveals its PC repeat domain as a closed toroid of two concentric alpha-helical rings. Structure. 2012; 20:513–521. DOI: 10.1016/j.str.2011.12.015 [PubMed: 22405010]
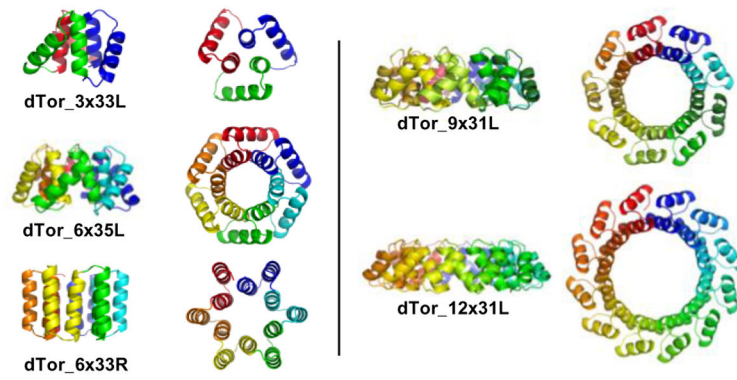
Figure 1. Designed monomeric repeat architectures

Side and top views of a representative design model from each family are shown in cartoon representation colored from blue to red as the chain proceeds from N- to C-terminus. Design nomenclature is given in the main text.
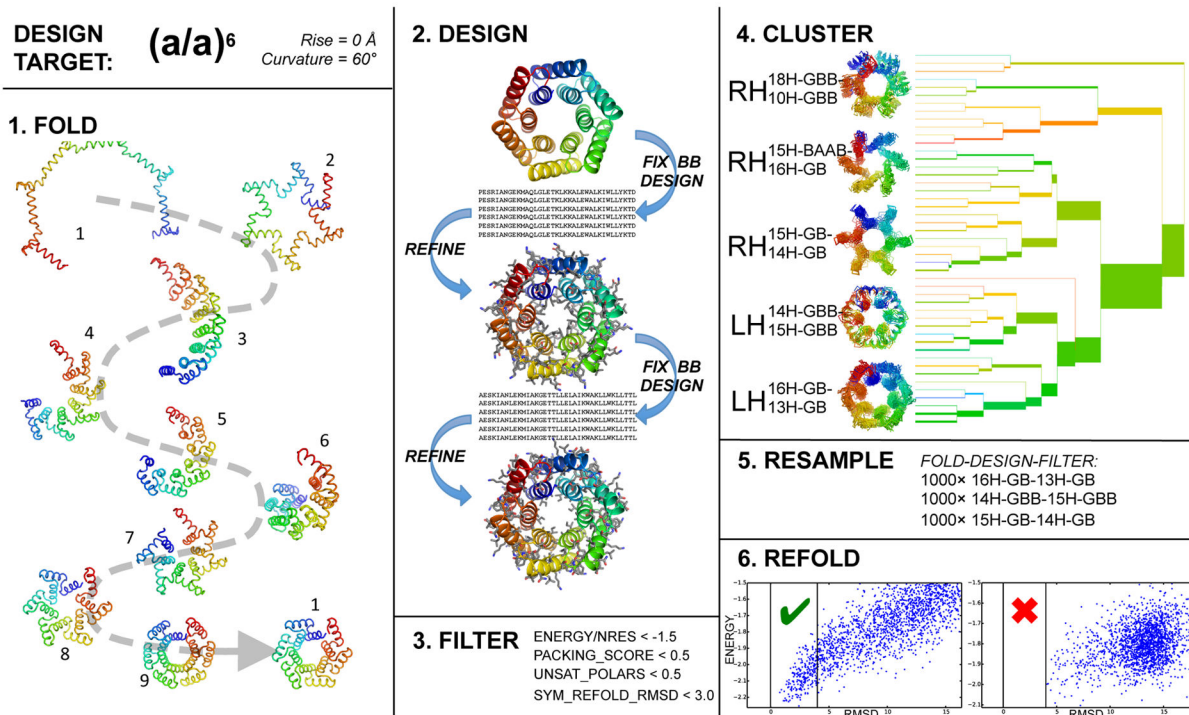
Figure 2. Overview of the repeat module design process

Given a design target consisting of secondary structure types, repeat number, and desired inter-repeat geometry, the main steps of the design methodology are (1) symmetric fragment assembly to generate starting backbone conformations; (2) all-atom sequence design and structure relaxation; (3) filtering to eliminate designs with poor packing, buried unsatisfied polar atoms, or low sequence-structure compatibility; (4) clustering to identify recurring packing arrangements; (5) intensified sampling of architectures identified in the clustering step; (6) final design assessment by large-scale re-prediction of the designed structure starting from the designed sequence. Design cluster identifiers (e.g., 14H-GBB-15H-GBB) record the length of the alpha-helices (14H and 15H) and the backbone conformations of the connecting loops (using a coarse-grained 5-state Ramachandran alphabet[27]; see Methods).
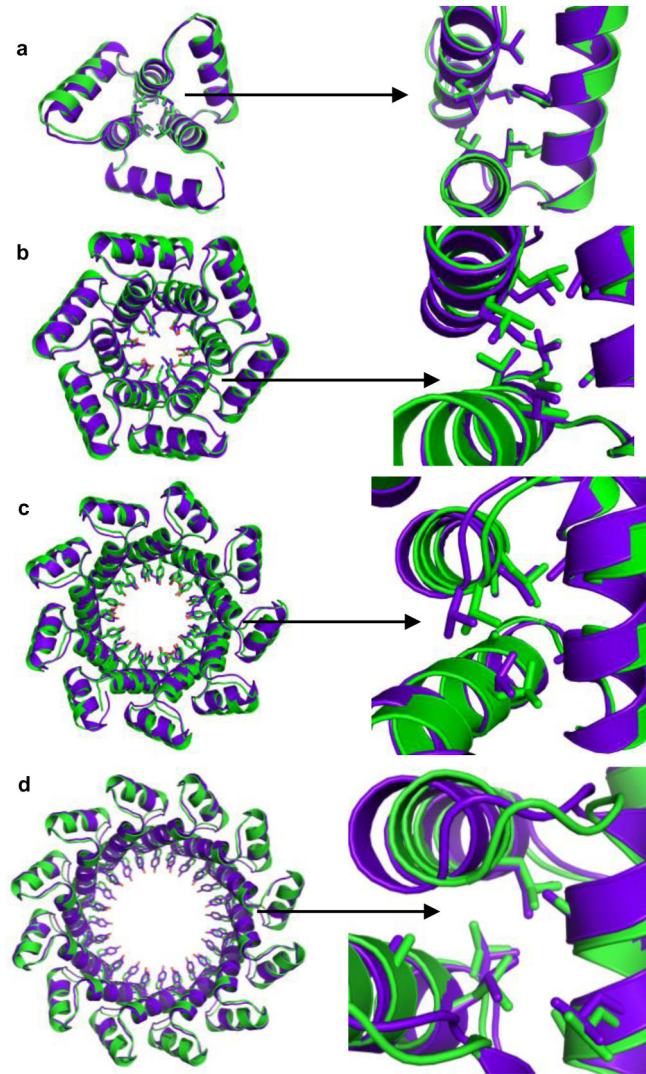
Figure 3. Superposition of designed toroids (purple) and their refined crystallographic structures (green)

The left panels show the overall superposition of the entire protein backbone, with the side chains that line the innermost pore shown for both models (a, dTor_3x33L; b, dTor_6x35L; c, dTor_9x31L; d, dTor_12x31L). The right panels show the same superpositions, enlarged to show the packing of side chains and helices between consecutive repeat modules.
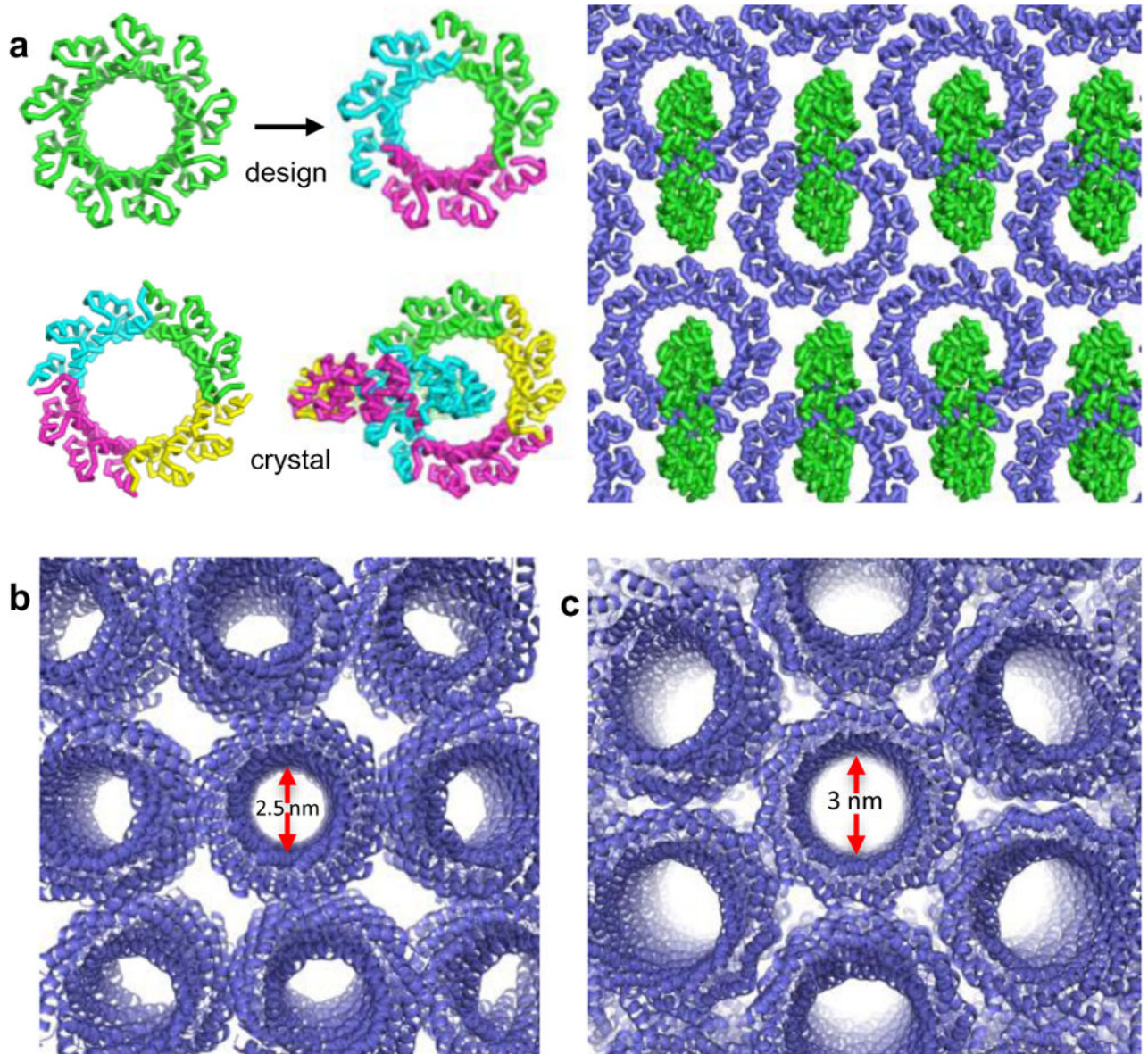
Figure 4. Crystal packing geometries of designed toroids

(a) Rather than forming the expected trimeric toroid ("design"), the 3-repeat sub-fragment of dTor_9x31L associated in the crystal as two linked tetrameric rings ("crystal") which pack into the layers visualized on the right (the full crystal is then formed from stacks of these layers). Continuous channels are assembled from stacked toroids in the crystals of the monomeric 9x31L and 12x31L designs ((b) and (c) respectively).