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

Lindsey Doyle<sup>1</sup>, Jazmine Hallinan<sup>1</sup>, Jill Bolduc<sup>1</sup>, Fabio Parmeggiani<sup>2,3</sup>, David Baker<sup>2,3,4</sup>, Barry L. Stoddard<sup>1</sup> & Philip Bradley<sup>1,3,5</sup>

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<sup>1,2</sup>. The modular nature of repeat proteins enables the rapid construction and diversification of extended binding surfaces by duplication and recombination of simple building blocks<sup>3,4</sup>. 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<sup>5-9</sup>, to closed and compact conformations with internal cavities suitable for small molecule binding and catalysis<sup>10</sup>. 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<sup>11</sup> repeat structures ( $\alpha$ -toroids) in which the inter-repeat packing geometry is constrained so as to juxtapose the amino (N) and carboxy (C) termini; several of these designed structures have been validated by X-ray crystallography. Unlike previous approaches to tandem repeat protein engineering<sup>12–20</sup>, 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<sup>21</sup>.

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 a multiple of 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 modelling package<sup>22</sup> and builds on published *de novo* design methodologies<sup>23</sup>. 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 favours 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 (Supplementary Discussion). We selected five monomeric repeat architectures for experimental characterization: a left-handed 3-repeat family (dTor\_3x33L designed toroid with three 33-residue repeats, left-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 three 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



**Figure 1** | **Designed monomeric repeat architectures.** Side and top views of a representative design model from each family are shown in cartoon representation coloured from blue to red as the chain proceeds from the N to the C terminus. Design nomenclature is given in the main text.

<sup>1</sup>Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA. <sup>2</sup>Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. <sup>3</sup>Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. <sup>4</sup>Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. <sup>5</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98195, USA.

# **RESEARCH LETTER**



Figure 2 | Overview of the repeat module design process. Given a design target consisting of secondary structure types ( $\alpha/\alpha$  in this example), repeat number (6), and desired inter-repeat geometry (rise and curvature), 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 suboptimal per-residue energy (ENERGY/NRES), poor packing (PACKING\_SCORE), buried unsatisfied polar atoms (UNSAT\_POLARS), or low sequence-structure compatibility (SYM\_REFOLD\_RMSD, deviation between the final design model and the predicted structure of

connect each repeated module are occupied by a mixture of continuous peptide and protein termini. This lattice behaviour 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 three 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. Ref. 24 describes similar crystal averaging with associated disorder at protein termini in a set of structures for designed consensus tetratricopeptide repeat (TPR) proteins, albeit with translational averaging along a fibre axis rather than the rotational averaging observed here.

Comparison of the design models with 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- and 12-repeat designs. Inspection of the superpositions in Fig. 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 owing to under-weighting of repulsive electrostatic or van der Waals interactions). Nevertheless, the success of the 12x31L

the designed sequence; for details see Methods); (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; r.m.s.d., root mean squared deviation. Design cluster identifiers (for example, 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 five-state Ramachandran alphabet<sup>27</sup>; see Methods).

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 behaviour 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 α-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<sup>6,7</sup> and mTERF<sup>25</sup> DNA binding domains (Supplementary 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 three-residue 'GBB'

### LETTER RESEARCH



Figure 3 | Superposition of designed toroids (purple) and their refined crystallographic structures (green). Left: 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). Right: the same superpositions, enlarged to show the packing of side chains and helices between consecutive repeat modules.

 $(\alpha_L - \beta - \beta)$  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 of these turn types are also compatible with canonical helix capping interactions<sup>26,27</sup>, 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 (that is, containing a total of 12 repeats per ring; Fig. 4a). Indeed, it was this unanticipated finding that led us to synthesize the monomeric



**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).

12x31L design whose characterization demonstrated that the designed 31-residue repeat sequence is 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 two-, three-, four-, or six-fold symmetry by repeating the interfacial mutations throughout the full sequence. Thus monomeric toroids could replace multimeric assemblies as symmetry centres 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 one- and two-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 two-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 one- or two-dimensional sub-assemblies<sup>28,29</sup>. 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 towards a variety of novel protein-based materials<sup>30</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

#### Received 30 April; accepted 30 October 2015.

Published online 16 December 2015.

- Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A census of protein repeats. J. Mol. Biol. 293, 151–160 (1999).
- Kajava, A. V. Tandem repeats in proteins: from sequence to structure. J. Struct. Biol. 179, 279–288 (2012).
- Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. J. Struct. Biol. 134, 117–131 (2001).
- Grove, T.Z., Cortajarena, A. L. & Regan, L. Ligand binding by repeat proteins: natural and designed. *Curr. Opin. Struct. Biol.* 18, 507–515 (2008).
- Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501–512 (2002).
- Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* 335, 716–719 (2012).
- Deng, D. et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science 335, 720–723 (2012).
- Barkan, A. et al. A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet.* 8, e1002910 (2012).
- 9. Reichen, C., Hansen, S. & Plückthun, A. Modular peptide binding: from a comparison of natural binders to designed armadillo repeat proteins. *J. Struct. Biol.* **185**, 147–162 (2014).
- Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* 492, 193–198 (2001).
- Kobe, B. & Kajava, A. V. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* 25, 509–515 (2000).
- Main, É. R., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **11**, 497–508 (2003).
- Binz, H. K. et al. High-affinity binders selected from designed ankyrin repeat protein libraries. Nature Biotechnol. 22, 575–582 (2004).
- 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.* **376**, 1282–1304 (2008).
- Urvoas, A. *et al.* Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats. *J. Mol. Biol.* **404**, 307–327 (2010).
- Boersma, Y. L. & Plückthun, A. DARPins and other repeat protein scaffolds: advances in engineering and applications. *Curr. Opin. Biotechnol.* 22, 849–857 (2011).
- Rämisch, S., Weininger, U., Martinsson, J., Akke, M. & André, I. Computational design of a leucine-rich repeat protein with a predefined geometry. *Proc. Natl Acad. Sci. USA* **111**, 17875–17880 (2014).
- Voet, A. R. *et al.* Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl Acad. Sci. USA* **111**, 15102–15107 (2014).
- Park, K. et al. Control of repeat-protein curvature by computational protein design. Nature Struct. Mol. Biol. 22, 167–174 (2015).

- Parmeggiani, F. et al. A general computational approach for repeat protein design. J. Mol. Biol. 427, 563–575 (2015).
- Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
- Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487, 545–574 (2011).
- 23. Koga, N. et al. Principles for designing ideal protein structures. Nature **491**, 222–227 (2012).
- Kajander, T., Cortajarena, A. L., Mochrie, S. & Regan, L. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallogr. D* 63, 800–811 (2007).
- Jiménez-Menéndez, N. *et al.* Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. *Nature Struct. Mol. Biol.* 17, 891–893 (2010).
- 26. Aurora, R. & Rose, G. D. Helix capping. Protein Sci. 7, 21–38 (1998).
- Wintjens, R. T., Rooman, M. J. & Wodak, S. J. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.* 255, 235–253 (1996).
- Grove, T. Z., Regan, L. & Cortajarena, A. L. Nanostructured functional films from engineered repeat proteins. J. R. Soc. Interface 10, http://dx.doi.org/10.1098/ rsif.2013.0051 (2013).
- Lanci, C. J. *et al.* Computational design of a protein crystal. *Proc. Natl Acad. Sci.* USA **109**, 7304–7309 (2012).
- Abe, S. & Ueno, T. Design of protein crystals in the development of solid biomaterials. RSC Adv. 5, 21366–21375 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank Scientific Computing at the Fred Hutchinson Cancer Research Center for providing the computational infrastructure necessary for this project. This research was supported by the following research grants: National Institutes of Health R21GM106117 to P.B. and R01GM49857 to B.L.S., and Swiss National Science Foundation Postdoc Fellowship PBZHP3-125470 and Human Frontier Science Program Long-Term Fellowship LT000070/2009-L to F.P.

Author Contributions L.D., J.H., J.B. and F.P. expressed, purified, and characterized designed constructs. L.D., J.H. and J.B. performed crystal screening, collected diffraction data, and solved crystal structures. P.B. developed and implemented the repeat design algorithms. P.B. performed sequence design calculations with feedback from F.P. P.B., B.L.S. and D.B. supervised the research. P.B. conceived of the toroid design project with input from B.L.S. and D.B. P.B. wrote the manuscript with input from the other authors.

Author Information Crystal structures determined in this study have been deposited in the RCSB Protein Data Bank under accession numbers 4YXX (dTor\_6x35L), 4YY2 (dTor\_3x33L\_2-2a), 4YY5 (dTor\_3x33L\_2-2b), 4YXY (dTor\_9x31L\_sub), 4YX2 (dTor\_9x31L), and 5BYO (dTor\_12x31L). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B. (pbradley@fredhutch.org).

#### **METHODS**

**Computational design.** The repeat module design process applied here consisted 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 were built using a fragment assembly protocol which is based on the standard Rosetta ab initio protocol<sup>31</sup> with the following modifications: (1) fragment replacement moves were performed symmetrically across all repeats, guaranteeing that backbone torsion angles were identical at corresponding positions across repeats; (2) a pseudoenergy term (equal to the deviation between actual and desired curvature, in degrees, plus the deviation in rise multiplied by a factor of 5) was added to the potential to favour satisfaction of the geometric constraints; (3) the aminoacid sequence used for low-resolution scoring was assigned randomly at the start of each simulation from secondary-structure-specific distributions (helix: Ala+Ile+Leu+Asp+Ser; turn: Gly+Ser), which had 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 were chosen randomly, defining the target secondary structure of the repeat module and its length. Together with the number of repeats, this defined the total length of the protein and the complete secondary structure, which was 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 was 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 was maintained across all repeats. Since the starting backbones for design were built by relatively coarse sampling in a low-resolution potential, sequences designed with the standard all-atom potential were dominated by small amino acids and the resulting structures tended to be under-packed. To correct for this tendency, a softened Lennard-Jones potential<sup>32</sup> was used for the sequence design steps, while the standard potential was 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 10,000-100,000 in this study) were first sorted by per-residue energy (total energy divided by the number of residues, to account for varying repeat length) and the top 20% 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 r.m.s.d. threshold of 2 Å for 3-repeat designs and 4 Å for larger designs). Designs that passed these filters were clustered by  $C-\alpha$  r.m.s.d. (allowing for register shifts when aligning helices with unequal lengths) to identify recurring architectures. The clusters were 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 were specifically resampled by enforcing their helix and turn lengths as well as their turn conformations (defined using a five-state, coarse-grained backbone torsion alphabet<sup>27</sup>; Extended Data Fig. 1e) during fragment selection.

Large-scale refolding. Selected low-energy designs from the second round that pass the filters described above were evaluated by a large-scale refolding test in which 2,000–10,000 *ab initio* models were built by standard (asymmetric) fragment assembly followed by all-atom relaxation. Success was measured by assessing the fraction of low-energy *ab initio* models with r.m.s.d. values 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 six-fold repeat symmetry would require K-K-K-K-K or E-E-E-E-E, doubling the repeat length allows 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 three 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 generated an approximate helical bundle axis curve by joining the location of repeat-unit centres of mass in a sliding fashion along the protein chain. The handedness was 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<sup>33</sup> as well as consulting the protein structure classification databases CATH<sup>34</sup>, SCOPe<sup>35</sup>, and ECOD<sup>36</sup>. Further details are given in Supplementary 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<sup>37</sup> containing a cleavable N-terminal His-tag and an ampicillin resistance cassette. Sequence-verified plasmids were transformed into BL21(DE3)RIL *Escherichia coli* cells (Agilent Technologies) and plated on lysogeny broth (LB) medium with ampicillin (100  $\mu$ g ml<sup>-1</sup>). 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 11 volumes of LB–ampicillin, which were then shaken at 37 °C until the cells reached an absorbance at 600 nm of 0.6–0.8. The cells were chilled for 20 min at 4 °C, then isopropyl- $\beta$ -D-thiogalactoside (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  $(100 \,\mu g \, ml^{-1})$  and chloramphenicol (35 g ml<sup>-1</sup>). Subsequent cell culture and protein expression in minimal media, along with incorporation of selenomethionine, was incorporated during protein expression according to ref. 38.

**Purification.** Cell pellets from 3 l of cell culture were resuspended in 60 ml of PBS solution (140 mM NaCl, 2.5 mM KCl, 10 mM NaHPO<sub>4</sub>, 2 mM KH<sub>2</sub>PO<sub>4</sub>) 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 µm filter, and then incubated on a rocker platform at 4 °C for 1 h 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 5–25 mg ml<sup>-1</sup> while buffer exchanging into 25 mM Tris (pH 7.5) and 200 mM 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), before additional crystallization trials. The digested sample was incubated for 30 min with streptavidin-conjugated agarose (Novagen) to remove the thrombin. All samples were tested for purity and removal of the His tag via SDS–polyacrylamide gel electrophoresis. The final protein samples, both with and without the N-terminal poly-histidine affinity tag, were concentrated to values of  $5-25 \,\mathrm{mg}\,\mathrm{ml}^{-1}$  for crystallization trials. Solution size and stability analysis. Proteins at a concentration of  $4-10 \text{ mg ml}^{-1}$  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<sup>-1</sup> 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 to assess their solution oligomeric behaviour using protein with a minimal background of contaminants. Gel filtration standards (Bio-Rad) were run over the same column in matching buffer, and the ultraviolet trace of the proteins was overlaid onto the standards using UNICORN 5 software (GE Healthcare).

For measurements of protein stability using circular dichroism spectroscopy, purified recombinant toroid constructs were diluted to between 10 and  $20\,\mu$ M concentration and dialysed overnight into 10 mM potassium phosphate buffer at pH 8.0. Circular dichroism thermal denaturation experiments were performed on a JASCO J-815 circular dichroism spectrometer with a Peltier thermostat. Wavelength scans (190–250 nm) were performed for each construct at 20 °C and 95 °C. Additional thermal denaturation experiments were conducted by monitoring circular dichroism signal strength at 206 nm over a temperature range of 4–95 °C (0.1 cm path-length cell), with measurements taken every 2°. Sample temperature was allowed to equilibrate for 30 s 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, 10 were crystallized, of which 5 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<sup>-1</sup>. The crystal was transferred to a solution containing 300 mM, then 500 mM sodium chloride and flash frozen in liquid nitrogen. Data were collected on a R-AXIS IV++ at wavelength 1.54 Å and processed on an HKL2000 (ref. 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<sup>-1</sup>. The crystal was transferred to a solution containing 300 mM, then 500 mM sodium chloride and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.2 at wavelength 0.9794 Å and processed on an HKL2000 (ref. 39).

dTor\_3x33L\_2-2 was crystallized in two different conditions, producing two different crystal lattices. The first condition had 30% polyethylene glycol 3350, 100 mM Tris pH 6.5, 200 mM NaCl with a protein concentration of 1.8 mM. The protein was soaked in a 15% ethylene glycol cryoprotectant for 1 min before being flash frozen in liquid nitrogen. Data were collected on a Saturn 944+ (Rigaku) at wavelength 1.54 Å for 180° at  $\varphi = 0$  and another 180° at  $\varphi = 180$ . Data were then processed on an HKL2000 (ref. 39) out to 1.85 Å in space group  $P2_12_12_1$ .

The second condition had 45% polyethylene glycol 400 and 100 mM Tris pH 7.7 with a protein concentration of 1.8 mM. Protein crystal was flash frozen without being cryoprotected. Data were collected on a Saturn 944+ (Rigaku) at wavelength 1.54 Å for 180° at  $\varphi = 0$  and another 180° at phi = 180. Data were then processed on an HKL2000 (ref. 39) out to 1.85Å in space group  $P4_32_{12}$ .

dTor\_9x31L\_sub was crystallized in 100 mÅ Tris pH 8.5 and 15% (v/v) ethanol at a concentration of 11.5 mg ml<sup>-1</sup>. The crystal was transferred to a solution containing 75 mÅ Tris pH 8.5, 7.5% (v/v) ethanol and 25% (v/v) glycerol and flash frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.2 at wavelength 1.0 Å and processed on an HKL2000 (ref. 39) out to 2.9 Å in space group  $P4_{121}2/P4_{3}2_{12}$ .

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<sup>-1</sup> in 3µl drops containing 1µl protein and 2µl well solution. The crystal was transferred to a solution containing the well plus 25% (v/v) glycerol and flash frozen in liquid nitrogen. Data were collected on a Saturn 944+ charge-coupled device at wavelength 1.54 Å and processed on an HKL2000 (ref. 39) out to 2.5 Å in space group  $P2_12_12_1$ .

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  $8.8 \text{ mg ml}^{-1}$  in

 $2 \mu$ l drops containing 1  $\mu$ l protein and 1  $\mu$ l 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 were collected on a Saturn 944+ charge-coupled device at wavelength 1.54 Å and processed on an HKL2000 (ref. 39) out to 2.3 Å 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<sup>40</sup> via CCP4i<sup>41</sup> using the Rosetta-designed structure as a search model. The structures were then built and refined using Coot<sup>42</sup> and Refmac5<sup>43</sup>, respectively.

The structure of dTor\_6x35L(SeMet) was solved by Molecular Replacement with Phaser<sup>40</sup> via PHENIX<sup>44</sup> using the best refined model of dTor\_6x35L as a phasing model. The structure was then built and refined using Coot<sup>42</sup> and PHENIX<sup>45</sup>, respectively.

The structures of dTor\_9x31L\_sub and dTor\_9x31L were solved by Molecular Replacement with Phaser<sup>40</sup> via PHENIX<sup>44</sup> using the Rosetta-designed structure as a search model. The structure was then built and refined using Coot<sup>42</sup> and PHENIX<sup>45</sup>, respectively.

The structure of dTor\_12x31L was solved by Molecular Replacement with Phaser<sup>40</sup> via PHENIX<sup>44</sup> using a 4-repeat subunit the Rosetta-designed structure as a search model. The structure was then built and refined using Coot<sup>42</sup> and PHENIX<sup>45</sup>, respectively.

Final Ramachandran statistics after refinement were as follows (given as % preferred, % allowed, % outliers, respectively): 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.

- Simons, K. T., 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. 268, 209–225 (1997).
- Dantas, G. et al. High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J. Mol. Biol. 366, 1209–1221 (2007).
- Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. Trends Biochem. Sci. 20, 478–480 (1995).
- Sillitoe, I. et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43, D376–D381 (2015).
- Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 42, D304–D309 (2014).
- Cheng, H. *et al.* ECOD: an evolutionary classification of protein domains. *PLOS Comput. Biol.* **10**, e1003926 (2014).
- Mak, A. N., Lambert, A. R. & Stoddard, B. L. Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kl. *Structure* 18, 1321–1331 (2010).
- Walden, H. Selenium incorporation using recombinant techniques. Acta Crystallogr. D 66, 352–357 (2010).
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* 276, 307–326 (1997).
- McCoy, A. J. et al. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674 (2007).
- Winn, M. D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D 67, 235–242 (2011).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D 66, 486–501 (2010).
- Skubák, P., Murshudov, G. N. & Pannu, N. S. Direct incorporation of experimental phase information in model refinement. Acta Crystallogr. D 60, 2196–2201 (2004).
- Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* 66, 213–221 (2010).
- 45. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
- He, J. et al. The structure of the 26S proteasome subunit Rpn2 reveals its PC repeat domain as a closed toroid of two concentric α-helical rings. Structure 20, 513–521 (2012).

# LETTER RESEARCH



**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 Protein Data Bank structure 4ADY and corresponds to the PC repeat domain of the 26S proteasome subunit Rpn2 (ref. 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 (for example, 'E' and 'GBB') tend to form left-handed connections whereas others (for example, 'GB' and 'BAAB') are associated with right-handed connections. Turn types are classified by mapping their backbone torsion angles to a coarse-grained alphabet<sup>27</sup> as shown in **e**.



Extended Data Figure 2 | Unbiased  $2F_o - F_c$  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 Fig. 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. Left: anomalous difference Fourier peaks calculated from data collected from a crystal of selenomethioninederivatized protein. Although only one methionine residue (at position 168) is present in the construct, strong anomalous difference peaks  $(I/\sigma I$  greater than 4.0) are observed at equivalent positions within at least three modular repeats. Right: difference density extending across the modelled 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.

**RESEARCH LETTER** 





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.

# LETTER RESEARCH



Extended Data Figure 5 | Purification and characterization of designed toroids. a–g, CD wavelength scan from 260 to 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 after metal affinity purification. Lane L, molecular mass protein standards (in kilodaltons); 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.

а

b





**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 coloured 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		Υ	Ν
dTor_3x33L_1-1	3	33	Left	Y	Y		Ν	
dTor_3x33L_2	3	33	Left	Y	Y		Υ	Ν
dTor_3x33L_2-1	3	33	Left	Y	Ν			
dTor_3x33L_2-2	3	33	Left	Y	Y	D	Y	Y
dTor_3x33L_2-3	3	33	Left	Y	Ν			
dTor_3x33L_2-4	3	33	Left	Y	Ν			
dTor_3x33L_3	3	33	Left	Y	N/A			
dTor_6x33R_1	6	33	Right	Y	Y		Υ	Ν
dTor_6x33R_1-1	6	33	Right	Y	Ν			
dTor_6x33R_1-2	6	33	Right	Y	Ν			
dTor_6x33R_1-3	6	33	Right	Y	Ν			
dTor_6x33R_2	6	33	Right	Y	Ν			
dTor_6x33R_3	6	33	Right	Y	Ν			
dTor_6x33R_4	6	33	Right	Ν				
dTor_6x35L	6	35	Left	Y	Y	D	Υ	Y
dTor_6x35L(SeMet)	6	35	Left	Y	Y		Y	Y
dTor_9x31L	9	31	Left	Y	Y	М	Υ	Y
dTor_12x31L	12	31	Left	Y	Y	М	Υ	Y

\*Construct was successfully overexpressed.

†Construct was successfully purified to homogeneity and concentrated to at least 1 mg ml<sup>-1</sup>.

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

Sconstruct 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 Cell dimensions	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
$ \begin{array}{l} a, b, c ( \AA ) \\ \alpha, \beta, \gamma ( ^{\circ} ) \\ \text{Resolution } ( \AA ) \dagger \\ \\ R_{\text{merge}} \\ I / \sigma I \\ \text{Completeness } ( \% ) \\ \text{Redundancy} \end{array} $	63.5, 85.3, 80.5 90.0, 90.0, 90.0 50.0-2.26 (2.30-2.26) 0.045 (0.159) 39.9 (13.8) 98.1 (97.9) 3.8 (3.6)	$\begin{array}{c} 63.5, 85.1, 80.5\\ 90.0, 90.0, 90.0\\ 50.0\text{-}2.18\\ (2.26\text{-}2.18)\\ 0.059 \ (0.323)\\ 29.7 \ (8.41)\\ 99.7 \ (99.2)\\ 13.7 \ (11.6) \end{array}$	$\begin{array}{c} 37.1, 68.6, 152.4\\ 90.0, 90.0, 90.0\\ 50.00-1.85\\ (1.90-1.85)\\ 0.056 \ (0.500)\\ 20.3 \ (4.34)\\ 90.6 \ (95.9)\\ 6.0 \ (7.0) \end{array}$	$\begin{array}{c} 40.2, 40.2, 217.7\\ 90.0, 90.0, 90.0\\ 50-2.78\\ (2.88-2.78)\\ 0.048 \ (0.136)\\ 27.0 \ (15.0)\\ 98.9 \ (98.2)\\ 12.3 \ (10.6) \end{array}$	$102.8, 102.8, 93.9\\90.0, 90.0, 90.0\\50.0-3.2\\(3.3-3.2)\\0.056 (0.461)\\31.3 (6.48)\\100.0 (100.0)\\14.8 (15.1)$	$\begin{array}{c} 41.7, 72.0, 86.2\\ 90.0, 90.0, 90.0\\ 50.0-2.50\\ (2.54-2.50)\\ 0.079 \ (0.292)\\ 30.4 \ (5.66)\\ 99.2 \ (91.2)\\ 10.0 \ (4.50) \end{array}$	95.4, 119.4, 76.3 90.0, 110.9, 90.0 50.0-2.50 (2.54-2.50) 0.048 (0.298) 27.2 (3.7) 98.9 (87.6) 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 R <sub>work/</sub> R <sub>free</sub> No. atoms		11137 23.8/29.6	29249 22.7/28.2	4760 19.3/26.7	8662 29.96/34.5	9355 22.5/32.8	27183 21.42/25.4
Protein Ligand/ion Water B-factors		1476 - -	3038 8 139	1480 - 50	2292 - -	2011	5608 - 166
Protein Ligand/ion Water R.m.s deviations		43.7 - -	36.6 61 52.4	26 - 56	108.2	35.9	42.1 43.8
Bond lengths (Å) Bond angles (°)		0.0142 1.6908	0.017 1.708	0.017 1.918	0.002 0.5	0.008 1.038	0.002 0.49

\*Each structure was determined from a single crystal. †Highest resolution shell is shown in parenthesis.