

Exploring the repeat protein universe through computational protein design

TJ Brunette^{1,2*}, Fabio Parmeggiani^{1,2*}, Po-Ssu Huang^{1,2*}, Gira Bhabha³, Damian C. Ekiert⁴, Susan E. Tsutakawa⁵, Greg L. Hura^{5,6}, John A. Tainer^{5,7} & David Baker^{1,2,8}

A central question in protein evolution is the extent to which naturally occurring proteins sample the space of folded structures accessible to the polypeptide chain. Repeat proteins composed of multiple tandem copies of a modular structure unit¹ are widespread in nature and have critical roles in molecular recognition, signalling, and other essential biological processes². Naturally occurring repeat proteins have been re-engineered for molecular recognition and modular scaffolding applications^{3–5}. Here we use computational protein design to investigate the space of folded structures that can be generated by tandem repeating a simple helix–loop–helix–loop structural motif. Eighty-three designs with sequences unrelated to known repeat proteins were experimentally characterized. Of these, 53 are monomeric and stable at 95 °C, and 43 have solution X-ray scattering spectra consistent with the design models. Crystal structures of 15 designs spanning a broad range of curvatures are in close agreement with the design models with root mean square deviations ranging from 0.7 to 2.5 Å. Our results show that existing repeat proteins occupy only a small fraction of the possible repeat protein sequence and structure space and that it is possible to design novel repeat proteins with precisely specified geometries, opening up a wide array of new possibilities for biomolecular engineering.

In repeat proteins, the interactions between adjacent units define the shape and curvature of the overall structure⁶. While in nature the sequences of these units generally differ, stable repeat proteins with identical units have been designed for several families^{7–21} and, for leucine-rich repeats, customized units have allowed for the control of curvature²² and design of new architectures¹⁷. To our knowledge, all designed repeat protein structures to date have been based on naturally occurring families. These families may cover all stable repeat protein structures that can be built from the 20 amino acids or, alternatively, natural evolution may only have sampled a subset of what is possible.

To explore the range of possible repeat protein structures, we generated new repeat protein backbone arrangements and designed sequences predicted to fold into these structures (Fig. 1 and Extended Data Figs 1 and 2). Our designs are entirely *de novo*; they are not based on naturally occurring repeat proteins. The well-packed repeating structures that can be obtained from a simple helix–loop repeat unit are limited to straight rods, and hence we focused on the helix–loop–helix–loop unit from which repeat proteins with a wide diversity of curvatures can be generated. The lengths of the two helices were varied between 10 and 28 residues, and the lengths of the two turns from 1 to 4 residues. Starting conformations for four tandem repeats of each of the 5,776 ($19 \times 19 \times 4 \times 4$) independent combinations of helix and loop lengths were generated by setting the backbone torsion angles

to ideal helix values for helices and extended chain values for loops. Rosetta Monte Carlo fragment assembly²³ was carried out to generate compact structures; each Monte Carlo move was made at the equivalent position in each repeat to preserve symmetry²⁰. Rosetta design calculations²⁴ were then used to identify low-energy amino acid sequences with good core packing²⁵. At each step in the Monte Carlo-simulated annealing design process, a position is picked at random, and the current residue is replaced by a randomly selected amino acid and side-chain conformation (rotamer); a detailed all-atom energy function is then evaluated. Identical substitutions were carried out in each unit to maintain sequence identity between the four repeats; exposed hydrophobic residues in the N- and C-terminal repeats were switched to polar residues in a second round of sequence design to increase solubility. All steps in the design process were completely automated, and the calculations were carried out without manual intervention. Designs with low energies and complementary core side-chain packing were identified, and for the amino acid sequence of each of these designs, multiple independent Rosetta *de novo* folding trajectories²⁶ were carried out starting from an extended chain. The structures and energies of the sampled conformations map out an energy landscape for each protein (Extended Data Fig. 3).

Designed helical repeat proteins (DHRs) for which the design model had much lower energy than any other conformation sampled in the

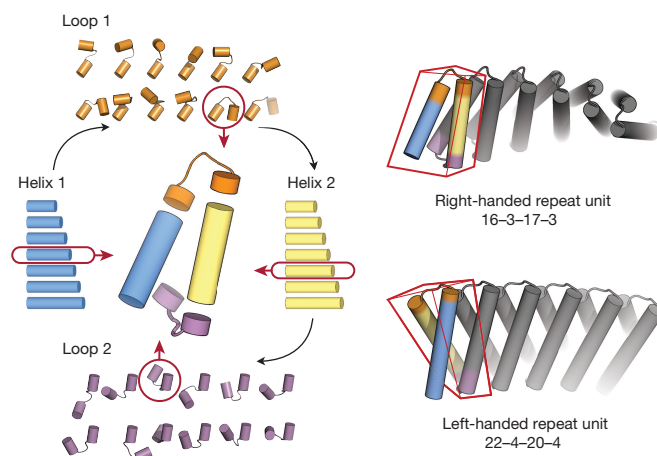


Figure 1 | Schematic overview of the computational design method. Helix–loop–helix–loop combinations are systematically sampled (left) and extended into repeating structures (right) using Rosetta Monte-Carlo fragment assembly. The red boxes on right indicate the individual repeat units; the numbers below, the lengths of helix 1, loop 1, helix 2, and loop 2 for these two examples.

¹Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. ²Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. ³Department of Cellular and Molecular Pharmacology, UCSF, San Francisco, California 94158, USA. ⁴Department of Microbiology and Immunology, UCSF, San Francisco, California 94158, USA. ⁵Molecular Biophysics & Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ⁶Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064, USA. ⁷Department of Molecular and Cellular Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. ⁸Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA.

*These authors contributed equally to this work.

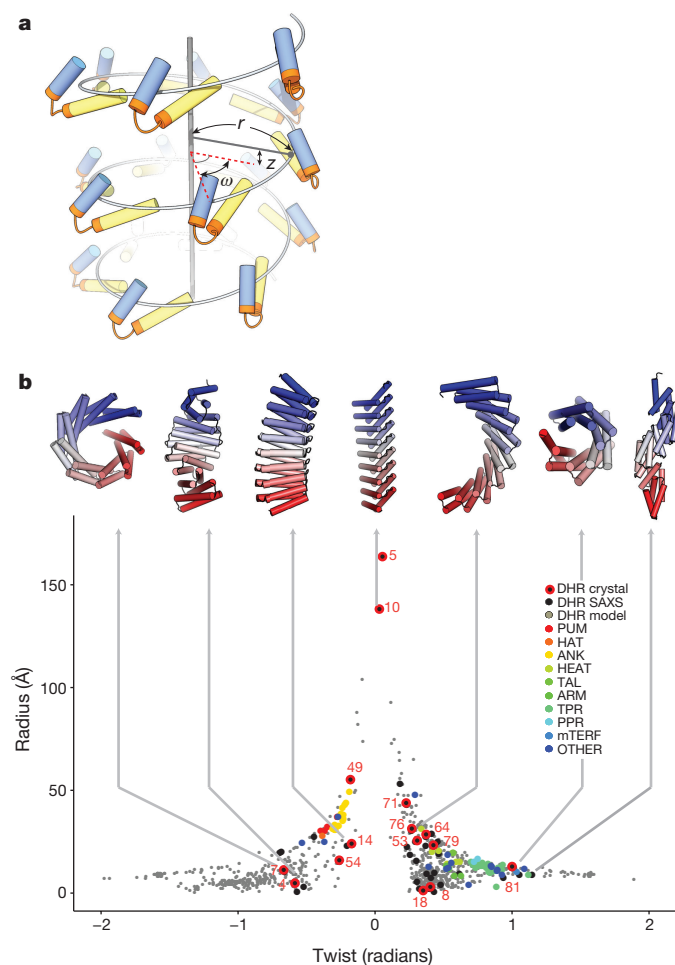


Figure 2 | The helical repeat protein universe. **a**, The geometry of a repeat protein can be described by the radius of the super-helix (r), the axial displacement (z) and the angular displacement or twist (ω) between repeat units. **b**, The 761 DHR models passing all the *in silico* filters are indicated by the small grey circles, the experimentally characterized DHR proteins confirmed by SAXS by large black circles, and those confirmed by X-ray crystallography, by black circles with red borders. The DHR proteins cover radius and twist ranges not found in native repeat protein families (colours). Designs forming right-handed super-helices have positive ω values; left-handed, negative ω values. ANK, ankyrin; ARM, armadillo; TPR, tetratricopeptide repeat; HAT, half TPR; PPR, pentatricopeptide repeat; HEAT, heat repeat; PUM, pumilio homology domain; mTERF, mitochondrial termination factor; TAL, transcription-activator-like effector; OTHER, alpha helical repeat proteins not in the other families. On top, representative experimentally validated designs with a variety of shapes.

de novo folding trajectories were selected and found to span a wide array of architectures. As the rigid body transform relating adjacent repeat units is identical throughout each design by construction, and since the repeated application to an object of an identical rigid body transformation produces a helical array, the designs all have an overall helical structure⁶. It is thus convenient to classify these architectures based on three parameters defining a helix²² (Fig. 2a): the radius (r), the twist between adjacent repeats around the helical axis (ω) and the translation between adjacent repeats along the helical axis (z). Because the repeat units are connected and form well packed structures, the three parameters are coupled. The arc length in the x - y plane spanned by a repeat unit is $\sim r\omega$, and the total length of a unit is $\sim \sqrt{(r\omega)^2 + z^2}$, hence the radius-twist distribution has a hyperbolic shape (Fig. 2b) with highly twisted structures having a smaller radius. Models with high r and high ω do not form a continuous protein core and are discarded during the backbone generation. Similarly, low-energy structures do not have high (>16 Å) z values as helices in adjacent repeats cannot then closely pack (Extended Data Fig. 4). Even with these geometric constraints, the design models span a wide range of helical parameters (Fig. 2b, grey), demonstrating that quite a diversity

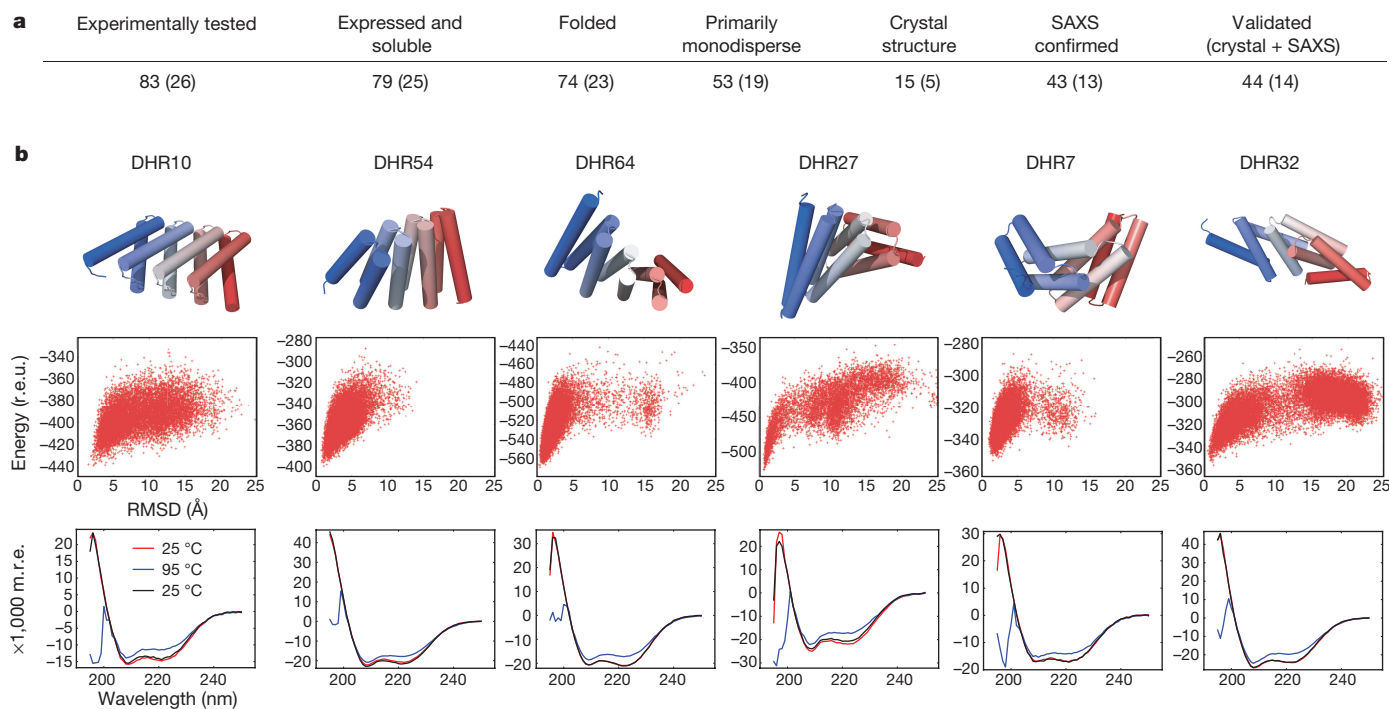


Figure 3 | Characterization of designed repeat proteins. **a**, Design success rate. Values for the subset with disulfide bonds are in parentheses. **b**, Results on six representative designs. Top row, design models. Middle row, computed energy landscapes. Energy is on y axis (r.e.u., Rosetta energy unit) and r.m.s.d. from design model on x axis. All six landscapes

are strongly funnelled into the designed energy minimum. Bottom row, circular dichroism spectra collected at 25 °C (red), 95 °C (blue; high dynode voltages reduce measurement accuracy below 200 nm) and back to 25 °C (black). The proteins do not denature within this temperature range (m.r.e., mean residue ellipticity; deg cm² dmol⁻¹ residue⁻¹).

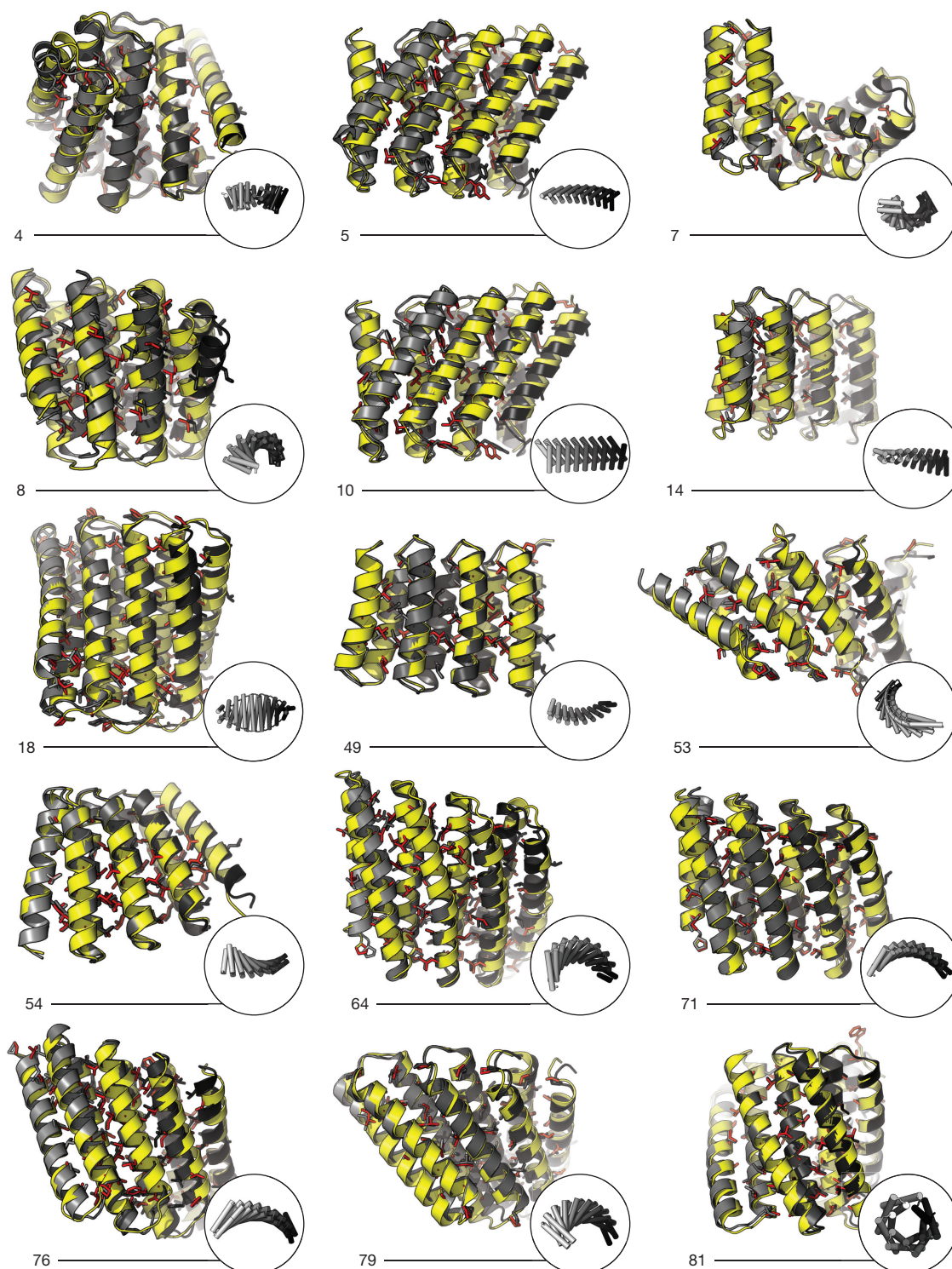


Figure 4 | Crystal structures of 15 designs are in close agreement with the design models. Crystal structures are in yellow, and the design models in grey. Insets in circles show the overall shape of the repeat protein. The r.m.s.d. values across all backbone heavy atoms are: 1.50 Å (DHR4), 1.73 Å (DHR5), 1.30 Å (DHR7), 2.28 Å (DHR8), 1.79 Å (DHR10), 2.38 Å (DHR14), 1.21 Å (DHR18), 0.87 Å (DHR49), 1.33 Å (DHR53), 0.93 Å (DHR54),

1.54 Å (DHR64), 0.67 Å (DHR71), 1.73 Å (DHR76), 1.04 Å (DHR79), 0.65 Å (DHR81). Hydrophobic side chains in the crystal structures (red) are close to those in the designs (grey) (Extended Data Fig. 5). The designed disulfide bonds are formed in the structures of DHR4 and DHR7 but not in the structures of DHR5 and DHR18 due to slight structural shifts relative to the design models.

of structures can be generated by tandem repeating a simple helix–loop–helix–loop unit. In contrast, native helical repeat proteins span a much narrower range of helical parameters (Fig. 2b, colours indicate different families) with very few straight (high r , low ω) or highly twisted (low r , high ω) geometries.

We selected for experimental characterization 83 designs spanning the range of α -helix and loop lengths and overall helical architectures; 26 of these contain disulfide bonds. BLAST searches against the NCBI databases yielded no hits with E values better than 0.0001 for 49 of the designs, and none of the hits found for the remaining designs were to

annotated repeat proteins. HHSEARCH comparisons of the designed repeat units to naturally occurring repeat families in Pfam yielded no hits with an *E* value better than 0.0005 (Supplementary Information Table 4). For each of the designs, we obtained a synthetic gene encoding an amino-terminal capping repeat, two internal repeats, and a carboxy-terminal capping repeat including a six-histidine tag. The proteins were expressed in *Escherichia coli* and purified by affinity chromatography. Of the 83 designs, 74 were expressed in a soluble form and had the expected α -helical circular dichroism spectrum at 25 °C, and 72 were stably folded at 95 °C (Supplementary Experimental Data). Fifty-three of these (64% of the original experimental set) were monomeric by analytical size-exclusion chromatography coupled to multi-angle light scattering; DHR49 and DHR76 were dimeric in solution. Structure stabilization with disulfide bonds did not systematically improve expression, solubility, or folding (Fig. 3a), probably because the designs are already very stable without disulfide bonds. Representative data on six of the designs are shown in Fig. 3b; the data on all 55 proteins is provided as Supplementary Experimental Data.

We solved the crystal structures of 15 of the designs (Fig. 4) with resolutions between 1.20 Å and 3.35 Å. The design models closely match the crystal structures over both the protein backbone (C α root mean square deviations (r.m.s.d.) from 0.7 Å to 2.5 Å) and the hydrophobic core side chains (Fig. 4 and Extended Data Fig. 5; crystal structure and design model side chains are in red and grey respectively). As is evident from the extended models in the Fig. 4 insets, the designs have different overall shapes: for example, DHR10 is linear and untwisted, DHR18 is linear and twisted, DHR8 forms a spiral and DHR81 is a flat toroid. The accuracy of the design models was sufficiently high that all of the crystal structures but DHR5 could be solved by molecular replacement. These repeat proteins are among the largest crystallographically validated protein structures designed completely *de novo*, ranging in size from 171 residues for DHR49 to 238 residues for DHR64. The crystal structures illustrate both the wide range of twist and curvature sampled by our repeat protein generation process and the accuracy with which these can be designed.

To characterize the structures for proteins that were reticent to crystallization and analyse all 55 proteins in solution, we used small-angle X-ray scattering (SAXS)^{27–30}. We collected SAXS profiles for each design, and compared them to scattering profiles calculated from the design models and from crystal structures. For 43 of the designs, the radius of gyration, molecular weight, and distance distributions computed from the SAXS data²⁹ corresponded to those computed from the models (Supplementary Information Table 6). For DHR49 and DHR76, we used the dimer orientation in the crystal for the fitting; the crystallographically confirmed DHR5 was unsuitable for SAXS as it formed higher-order species in solution. To further assess the fit between models and experimental data, we employed the volatility ratio (*V_r*), which is more robust to experimental noise than the traditional χ^2 comparison used in SAXS³⁰. We used the *V_r* values of the design models confirmed by crystallography for calibration; designs for which the *V_r* value between model and experimental data was less than 2.5 were considered successful. All 43 designs with radii, molecular weights, and distances consistent with the SAXS data are below the *V_r* threshold (Extended Data Fig. 6a). Furthermore, for almost all of the designs, the theoretical scattering profile computed from the design model more closely matches its own experimental scattering profile than the experimental scattering profiles of structurally dissimilar designs (Extended Data Fig. 6b, c).

The crystallographic and SAXS data together structurally validate 44 of the 55 designs that were folded and monodisperse—more than half of the 83 that were experimentally characterized. We randomly selected two designs confirmed by crystallography, two confirmed by SAXS, and two not confirmed by SAXS, and examined their guanidine hydrochloride unfolding profiles. In contrast to almost all native proteins, four of the six designs do not denature at guanidine hydrochloride concentrations up to 7.5 M; the other two, which were confirmed by SAXS but did

not yield crystals, have denaturation midpoints above 3 M (Extended Data Fig. 7). Hence, even the apparent failures are well-folded proteins; small amounts of association may be responsible for the discrepancies between computed and observed SAXS spectra rather than deviations from the design models.

We have shown that a wide range of novel repeat proteins can be generated by tandem repeating a simple helix–loop–helix–loop building block. As illustrated by the comparison of 15 design models to the corresponding crystal structures (Fig. 4), our approach allows precise control over structural details throughout a broad range of geometries and curvatures. The design models and sequences are very different from each other and from naturally occurring repeat proteins, without any significant sequence or structural homology to known proteins (Extended Data Fig. 8). This work achieves key milestones in computational protein design: the design protocol is completely automatic, the folds are unlike those in nature, more than half of the experimentally tested designs have the correct overall structure as assessed by SAXS, and the crystal structures demonstrate precise control over backbone conformation for proteins over 200 amino acids. The observed level of control over the repeating helix–loop–helix–loop architecture shows that computational protein design has matured to the point of providing alternatives to naturally occurring scaffolds, including graded and tunable variation difficult to achieve starting from existing proteins. We anticipate that the 44 successful designs described in this work (Extended Data Fig. 9), and sets generated using similar protocols for other repeat units, will be widely useful starting points for the design of new protein functions and assemblies.

Naturally occurring repeat protein families, such as ankyrins, leucine-rich repeats, TAL effectors and many others, have central roles in biological systems and in current molecular engineering efforts. Our results suggest that these families are only the tip of the iceberg of what is possible for polypeptide chains: there are clearly large regions of repeat protein space that are not sampled by currently known repeat protein structures. Repeat protein structures similar to our designs may not have been characterized yet, or perhaps may simply not exist in nature.

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 25 April; accepted 26 October 2015.

Published online 16 December 2015.

1. Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* **179**, 279–288 (2012).
2. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A census of protein repeats. *J. Mol. Biol.* **293**, 151–160 (1999).
3. Binz, H. K. et al. High-affinity binders selected from designed ankyrin repeat protein libraries. *Nature Biotechnol.* **22**, 575–582 (2004).
4. Varadachari, G., Tremmel, D., Hansen, S., Parmeggiani, F. & Plückthun, A. Designed Armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity. *J. Mol. Biol.* **424**, 68–87 (2012).
5. Cortajarena, A. L., Liu, T. Y., Hochstrasser, M. & Regan, L. Designed proteins to modulate cellular networks. *ACS Chem. Biol.* **5**, 545–552 (2010).
6. 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).
7. Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Höcker, B. & Baker, D. *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chem. Biol.* <http://dx.doi.org/10.1038/nchembio.1966> (2015).
8. Cortajarena, A. L. & Regan, L. Calorimetric study of a series of designed repeat proteins: modular structure and modular folding. *Protein Sci.* **20**, 336–340 (2011).
9. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503 (2003).
10. Mosavi, L. K., Minor, D. L. & Peng, Z. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA* **99**, 16029–16034 (2002).

11. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of stable α -helical arrays from an idealized TPR motif. *Structure* **11**, 497–508 (2003).
12. Urvoas, A. *et al.* Design, production and molecular structure of a new family of artificial alpha-helical repeat proteins (α Rep) based on thermostable HEAT-like repeats. *J. Mol. Biol.* **404**, 307–327 (2010).
13. Lee, S.-C. *et al.* Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proc. Natl Acad. Sci. USA* **109**, 3299–3304 (2012).
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.* **376**, 1282–1304 (2008).
15. Yadi, I. & Tawfik, D. S. Reconstruction of functional β -propeller lectins via homo-oligomeric assembly of shorter fragments. *J. Mol. Biol.* **365**, 10–17 (2007).
16. Coquille, S. *et al.* An artificial PPR scaffold for programmable RNA recognition. *Nature Commun.* **5**, 5729 (2014).
17. 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).
18. Lee, J. & Blaber, M. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc. Natl Acad. Sci. USA* **108**, 126–130 (2011).
19. Voet, A. R. D. *et al.* Computational design of a self-assembling symmetrical β -propeller protein. *Proc. Natl Acad. Sci. USA* **111**, 15102–15107 (2014).
20. Parmeggiani, F. *et al.* A general computational approach for repeat protein design. *J. Mol. Biol.* **427**, 563–575 (2015).
21. Tripp, K. W. & Barrick, D. Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *J. Mol. Biol.* **365**, 1187–1200 (2007).
22. Park, K. *et al.* Control of repeat-protein curvature by computational protein design. *Nature Struct. Mol. Biol.* **22**, 167–174 (2015).
23. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6**, e24109 (2011).
24. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
25. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
26. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution *de novo* structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
27. Rambo, R. P. & Tainer, J. A. Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annu. Rev. Biophys.* **42**, 415–441 (2013).
28. Hura, G. L. *et al.* Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nature Methods* **6**, 606–612 (2009).
29. Rambo, R. P. & Tainer, J. A. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481 (2013).
30. Hura, G. L. *et al.* Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. *Nature Methods* **10**, 453–454 (2013).

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

Acknowledgements We thank D. Kim and members of the protein production facility at the Institute for Protein Design. This work was facilitated through the use of advanced computational, storage and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. This work was supported in part by grants from the National Science Foundation (NSF) (MCB-1445201 and CHE-1332907), the Defense Threat Reduction Agency (DTRA), the Air Force Office of Scientific Research (AFOSR) (FA950-12-10112) and the Howard Hughes Medical Institute (HHMI-027779). F.P. was the recipient of a Swiss National Science Foundation Postdoc Fellowship (PBZHP3-125470) and a Human Frontier Science Program Long-Term Fellowship (LT000070/2009-L). SAXS work at the Advanced Light Source SIBLYS beamline was supported by the National Institutes of Health grant MINOS (Macromolecular Insights on Nucleic Acids Optimized by Scattering) GM105404 and by United States Department of Energy program Integrated Diffraction Analysis Technologies (IDAT). D.C.E. is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (Grant DRG-2140-12). G.B. is a recipient of the Merck fellowship of the Damon Runyon Cancer Research Foundation (DRG-2136-12) and is supported by NIH grant K99GM112982. J.A.T. is supported by a Robert A. Welch Distinguished Chair in Chemistry. We thank J. Holton for advice on S-SAD data collection, and the staff of ALS 8.2.1 and 8.3.1 for beamline support. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231. ALS beamline 8.3.1 is supported by the UC Office of the President, Multicampus Research Programs and Initiatives grant MR-15-338599 and the Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. ALS beamline 8.2.1 and the Berkeley Center for Structural Biology are supported in part by the National Institutes of Health, National Institute of General Medical Sciences, and the Howard Hughes Medical Institute.

Author Contributions P.-S.H., F.P. and D.B. conceived the *de novo* repeat protein design project. T.B., F.P., P.-S.H. and D.B. conceived the large scale conformational sampling approach. T.B. developed the algorithm with help from F.P. and P.-S.H. F.P. and T.B. expressed and characterized the design with help from P.-S.H. G.B. and D.C.E. setup crystallization trials and solved the crystal structures. F.P., S.E.T., G.L.H., J.T. collected and analysed the SAXS data. F.P., T.B., P.-S.H. and D.B. wrote the manuscript with help from all the authors.

Author Information Crystal structures have been deposited in the RCSB Protein Data Bank with the accession numbers 5CWB (DHR4), 5CWC (DHR5), 5CWD (DHR7), 5CWF (DHR8), 5CWG (DHR10), 5CWH (DHR14), 5CWI (DHR18), 5CWJ (DHR49), 5CWL (DHR53), 5CWL (DHR54), 5CWM (DHR64), 5CWN (DHR71), 5CWO (DHR76), 5CWP (DHR79) and 5CWQ (DHR81). 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 D.B. (dabaker@uw.edu).

METHODS

Code availability. The Rosetta macromolecular modelling suite is available from (<http://www.rosettacommons.org>). The design strategy is described in detail in the Supplementary Information. The Rosetta design code for each step is provided in Supplementary Information section 'Rosetta_examples'.

Similarity search. BLAST^{31,32} and HHSEARCH³³ sequence similarity searches were performed with default settings. HHSEARCH was run on Pfam³⁴. Sequence alignments were depicted using Jalview³⁵. The structural similarity between designs and known helical repeat proteins was assessed by TM-align³⁶ on RepeatsDB³⁷ representative structures.

Protein expression and characterization. Genes were synthesized and cloned in vector pET21 by GenScript (Piscataway). Proteins were expressed in *E. coli* BL21(DE3), induced with 250 μ M isopropyl- β -D-thiogalactopyranoside (IPTG) overnight at 22°C and purified by metal ion affinity chromatography (IMAC) and size-exclusion chromatography (SEC) as described in ref. 20. Cells were lysed by sonication and the clarified lysate was loaded on a NiNTA superflow column (Qiagen). Lysis and washing buffer was Tris 50 mM, pH 8, NaCl 500 mM, imidazole 30 mM, glycerol 5% v/v. Lysozyme (2 mg ml⁻¹), DNaseI (0.2 mg ml⁻¹) and protease inhibitor cocktail (Roche) were added to the lysis buffer before sonication. Proteins were eluted in Tris 50 mM, pH 8, NaCl 500 mM, imidazole 250 mM, glycerol 5% v/v and dialysed overnight either in Tris 20 mM, pH 8, NaCl 150 mM. Protein concentrations were determined using a NanoDrop spectrophotometer (Thermo Scientific). Except as indicated above, enzymes and chemical were purchased from Sigma-Aldrich. Secondary structure content, thermal stability and denaturation in presence of guanidine hydrochloride (GuHCl) were monitored by circular dichroism using an AVIV 420 spectrometer (Aviv Biomedical). Thermal denaturation was followed at 220 nm in Tris 20 mM, 50 mM NaCl, pH 8. Proteins were considered folded if they had the expected α -helical circular dichroism spectrum at 25°C and had either a sharp transition in thermal denaturation or a loss of less than 20% of 220 nm circular dichroism signal at 95°C. Chemical denaturation was monitored in a 1 cm path-length cuvette at 222 nm with protein concentration of 0.05 mg ml⁻¹ in phosphate buffer 25 mM NaCl 50 mM, pH 7. The GuHCl concentration was automatically controlled by a MicroLab titrator (Hamilton). Oligomeric state was assessed by analytical gel filtration coupled to multiple-angle light scattering (AFG-MALS). A Superdex 75 10/300 GL column (or Superdex 200 increase for DHR59, 84, 93) (GE Healthcare) equilibrated in Tris 20 mM, NaCl 150 mM, pH 8 was used on a HPLC LC 1200 Series (Agilent Technologies) connected to a miniDAWN TREOS (Wyatt Technologies). Protein molecular weights were confirmed by mass spectrometry on a LCQ Fleet Ion Trap Mass Spectrometer (Thermo Scientific). Of the 83 designs, 74 were expressed in a soluble form and had the expected α -helical circular dichroism spectrum at 25°C and 72 were stably folded at 95°C. DHR36 has $T_m = 75^\circ\text{C}$ and DHR13 has a broad transition with $T_m = 62^\circ\text{C}$. Fifty-five of these were predominantly monodisperse. DHR49 and 76 were dimeric in solution. SDS-page gels, circular dichroism spectra, thermal denaturation and size-exclusion chromatography profile *ab initio* folding funnel and SAXS data are shown as Supplementary Information for each of the 55 folded and monodisperse proteins.

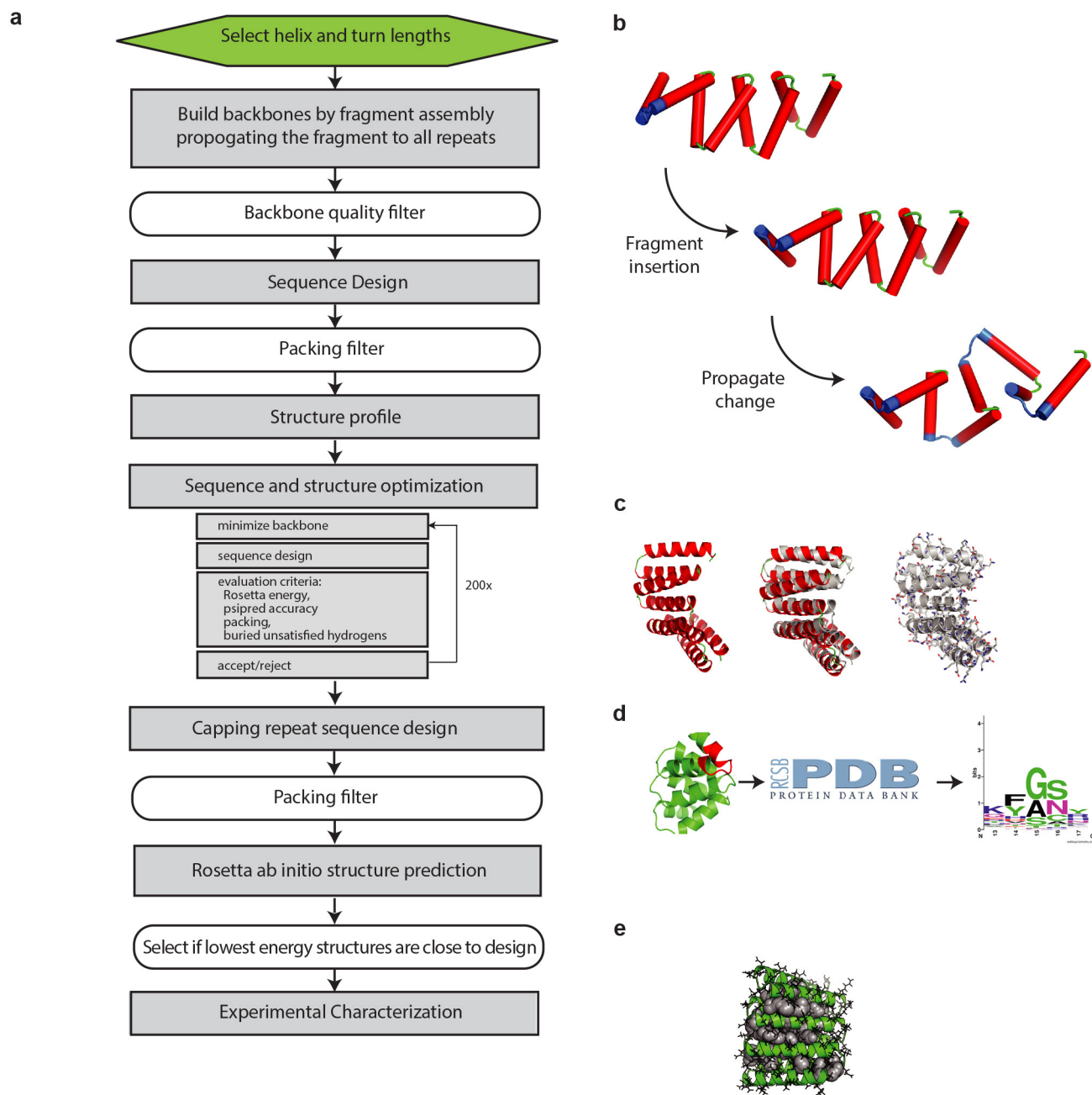
Crystallization. Proteins were purified using NiNTA resin and size-exclusion chromatography on a Superdex 75 column (GE healthcare). Pure fractions in the gel filtration buffer (20 mM Tris pH 8.0, 150 mM NaCl) were pooled and concentrated for crystallography. Final concentrations for each protein are shown in Supplementary Information Table 5. Initial crystallization trials were performed using the JCSG core I-IV screens at 22°C, and crystals were optimized if necessary. Drops were set up with the Mosquito HTS using 100 nl protein and 100 nl of the well solution. Crystals were cryoprotected in the reservoir solution supplemented with ethylene glycol, then flash cooled and stored in liquid nitrogen until data collection. All diffraction data were collected at the Advanced Light Source (ALS) at beamline 8.3.1 or beamline 8.2.1. Crystallization conditions, phasing method and space group information are shown in Supplementary Information Table 5. Data reduction was carried out using XDS³⁸ and HKL2000 (HKL Research). Most of the structures reported here were solved by molecular replacement using Phaser. Search models were generated by *ab initio* folding of the designed sequences in Rosetta and a set of the lowest energy 10–100 models was selected for molecular replacement trials. DHR5 was the only structure which could not be readily solved by molecular replacement. However, owing to the presence of six cysteine residues in the native protein, the DHR5 structure was solved by sulfur single wavelength anomalous dispersion (S-SAD) using a data set collected at 7,235 eV (Supplementary Information). Rigid body, restrained refinement with TLS and simulated annealing were carried out in Phenix³⁹. Manual adjustment of the model was carried out in Coot⁴⁰. The structures were validated using the Quality Control Check v2.8 developed by JCSG, which included Molprobity⁴¹ (publicly available at <http://smb.slac.stanford.edu/jcsg/QC/>). Data collection and final refinement statistics are shown in Supplementary Information Tables 6–14.

SAXS. SAXS data on size-exclusion-chromatography-purified protein were collected at the SIBYLS 12.3.1 beamline at the Advanced Light Source, LBNL^{28,42,43}. Scattering measurements were performed on 20- μ l samples and loaded into a helium-purged sample chamber, 1.5 m from the Mar165 detector. Data were collected on both the original gel filtration fractions and samples concentrated ~ 2 – $8\times$ from individual fractions. Fractions before the void volume and concentrator eluates were used for buffer subtraction. Sequential exposures (0.5, 1, 2, and 5 s) were taken at 12 keV to maximize the signal-to-noise ratio with visual checks for radiation-induced damage to the protein. The data used for fitting were selected for having higher signal to noise ratio and lack of radiation-induced aggregation. In case of concentration dependency, the lowest concentration was used. Models for SAXS comparison were obtained by adding the flexible C-terminal tag present in the constructs to the original designs and the crystal structures, generating 100 trajectories for each starting model by Monte Carlo fragment insertion²³. The results were clustered in Rosetta with a cluster radius of 2 Å and the cluster centres were used for comparison to the experimental data. We used FOXS^{44,45} to calculate scattering profiles from cluster centres and fit them to the experimental data. The quality of fit between models and experimental SAXS data are usually assessed by the χ value⁴⁶, which, however, suffers from over-fitting in case of noisy data sets and domination of the low region of the scattering vector (q) on the value²⁷. To avoid artificially low values that represent false positives, we instead used volatility ratio (V_r)³⁰ as primary metric for fit in the range of $0.015 \text{ \AA}^{-1} < q < 0.25 \text{ \AA}^{-1}$. V_r values of models with available crystal structures range from 0.7 to 2.3 (Supplementary Information Table 15). $V_r = 2.5$ was selected as upper threshold to consider a design as validated by SAXS. An in-depth evaluation of SAXS curves including mass, radius of gyration, Porod number and probability distribution is described in detail in Supplementary Information.

Model profiles for V_r similarity maps were obtained with a standardized fit procedure by averaging the scattering profile of the cluster centres from the five largest clusters and fitting the solvent hydration layer with parameters $C1 = 1.015$ and $C2 = 2.0$ for all the models. V_r was calculated in the range $0.04 \text{ \AA}^{-1} < q < 0.3 \text{ \AA}^{-1}$. The order of display was derived by shape similarity of original computational models using the program damsip⁴⁷ for superposition.

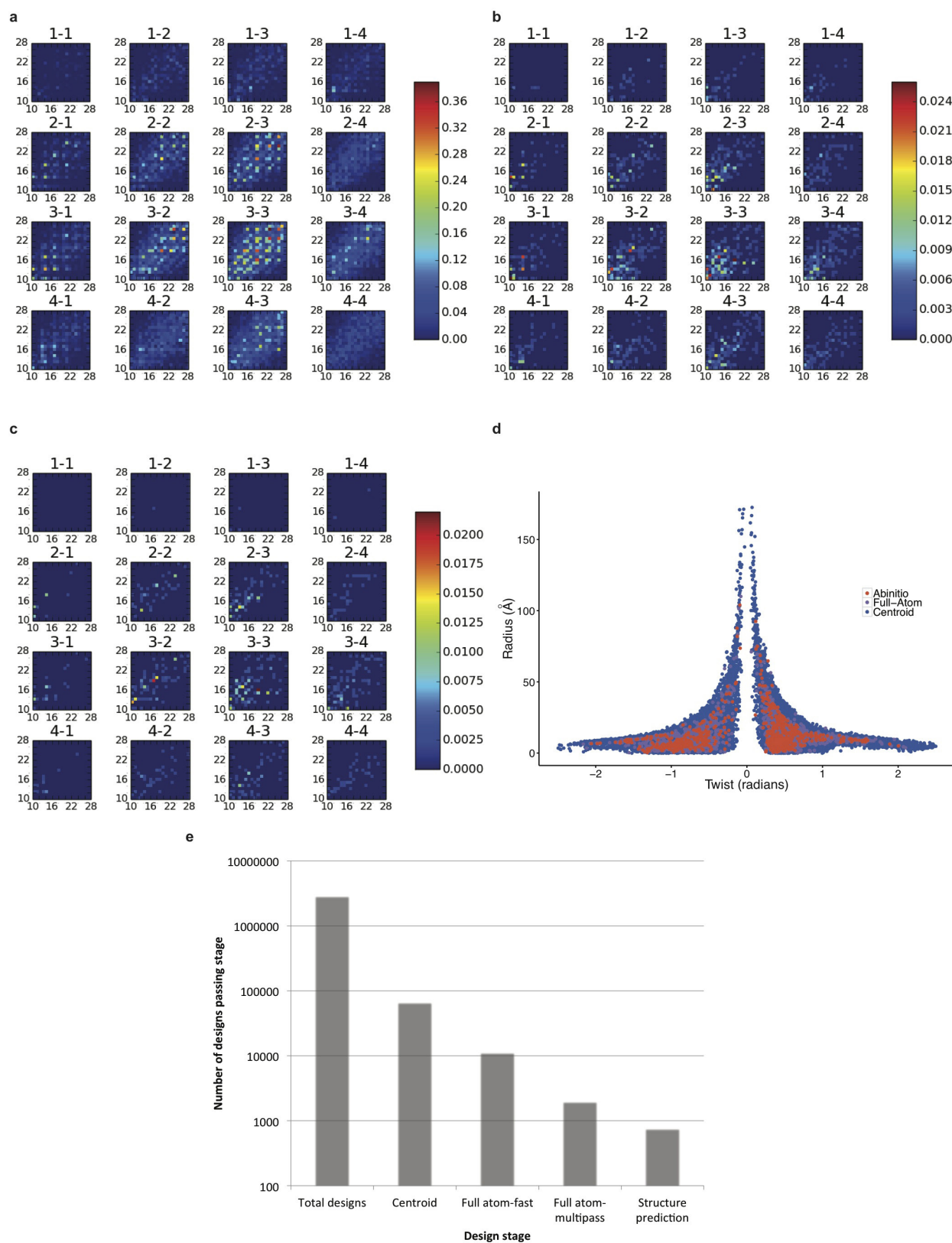
Additional details and discussions on computational design methods, DHR description, experimental characterization, crystallization and SAXS are provided as Supplementary Information.

- Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
- Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173–175 (2012).
- Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
- Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
- Di Domenico, T. *et al.* RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res.* **42**, D352–D357 (2014).
- Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
- Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
- Classen, S. *et al.* Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *J. Appl. Crystallogr.* **46**, 1–13 (2013).
- Classen, S. *et al.* Software for the high-throughput collection of SAXS data using an enhanced Blu-Ice/DCS control system. *J. Synchrotron Radiat.* **17**, 774–781 (2010).
- Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).
- Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–W544 (2010).
- Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOLE – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
- Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).



Extended Data Figure 1 | Computational protocol for designing *de novo* repeat proteins. **a**, Flowchart of the design protocol. The green box indicates user-controlled inputs, the grey boxes represent steps where protein structure is created or modified, and the white boxes indicate where structures are filtered. **b**, Low-resolution backbone build. **c**, Quick full-atom design (grey) improves the backbone model (red).

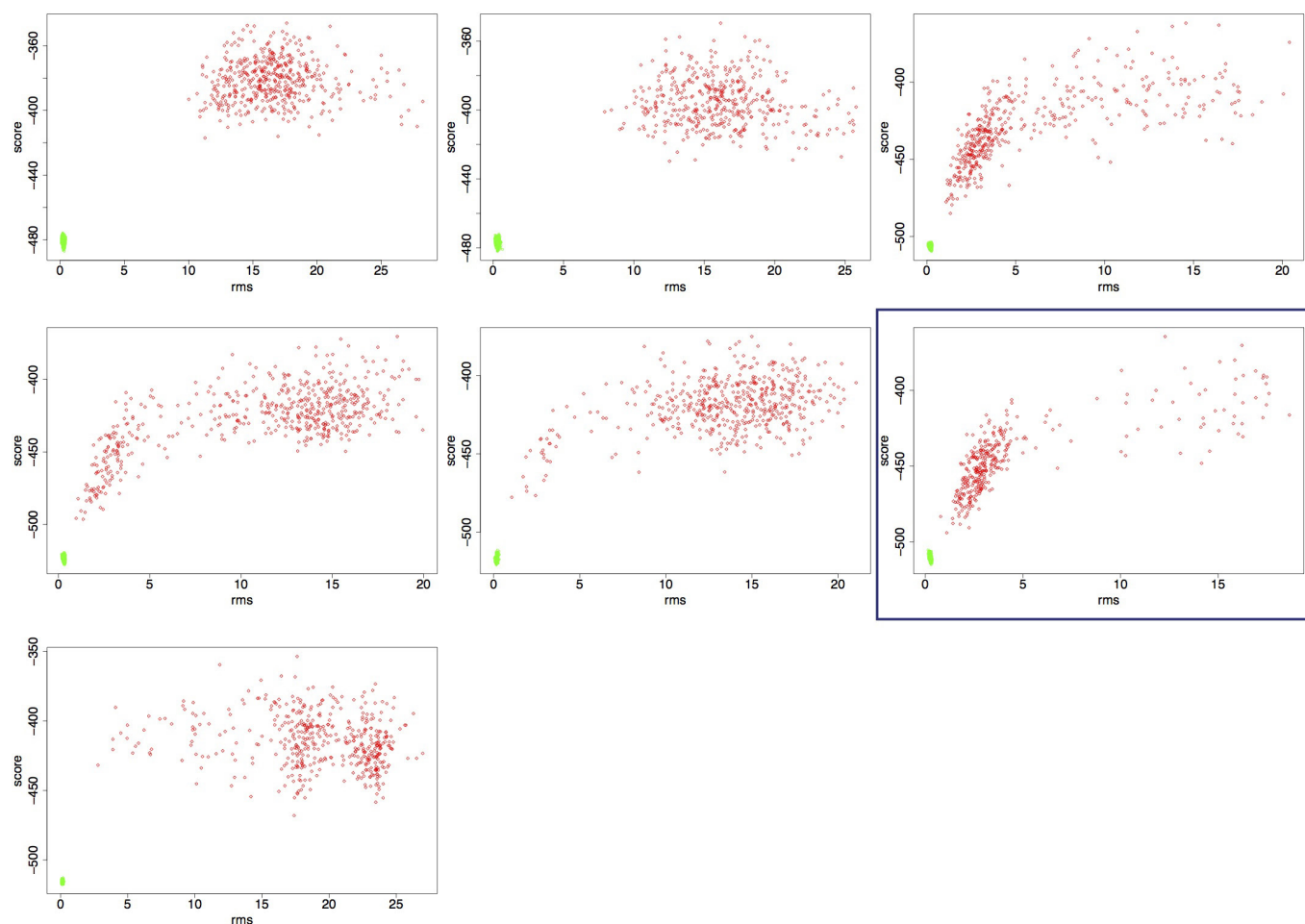
The superposition in the middle highlights the structural changes introduced. **d**, Structural profile: a 9-residue fragment is matched against the Protein Data Bank repository for structures within 0.5 Å r.m.s.d. The sequences from these structures are used to generate a sequence profile that influences design. **e**, Packing filters were used to discard designs with cavities in the core, illustrated as grey spheres.



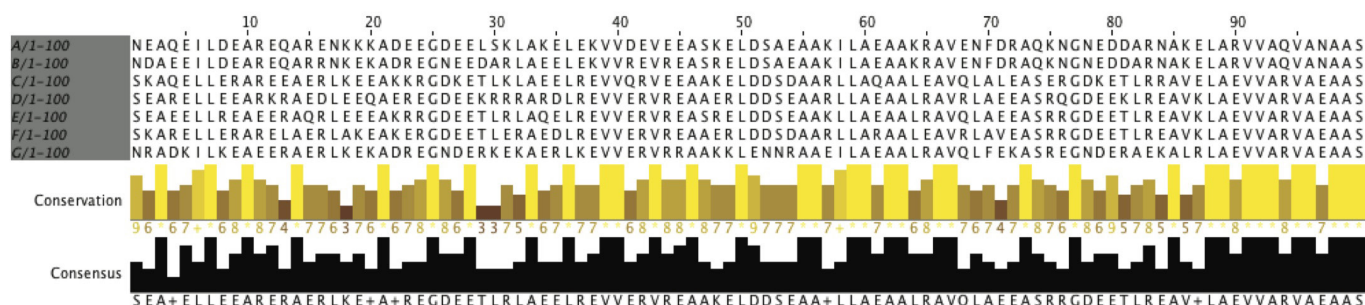
Extended Data Figure 2 | Repeat space explored and model discrimination across design stages. a–c, Percentage of models accepted at backbone building or centroid (a), design (b) and *ab initio* (c) stages. Models are divided according to secondary structure length. The combination of loop 1 and loop 2 lengths is indicated on top. x and y axis indicate helix 1 and helix 2 lengths, respectively. The fraction of models

in the bin that passed the selection stage is indicated in the side bar. Generally, one-residue loops and large differences between helix lengths reduce the number of selected models. d, Distribution of radius and twist of models in the three stages. e, Number of models passing design stages (log scale). From ~2.8 million structures, 761 are accepted.

a-g in order

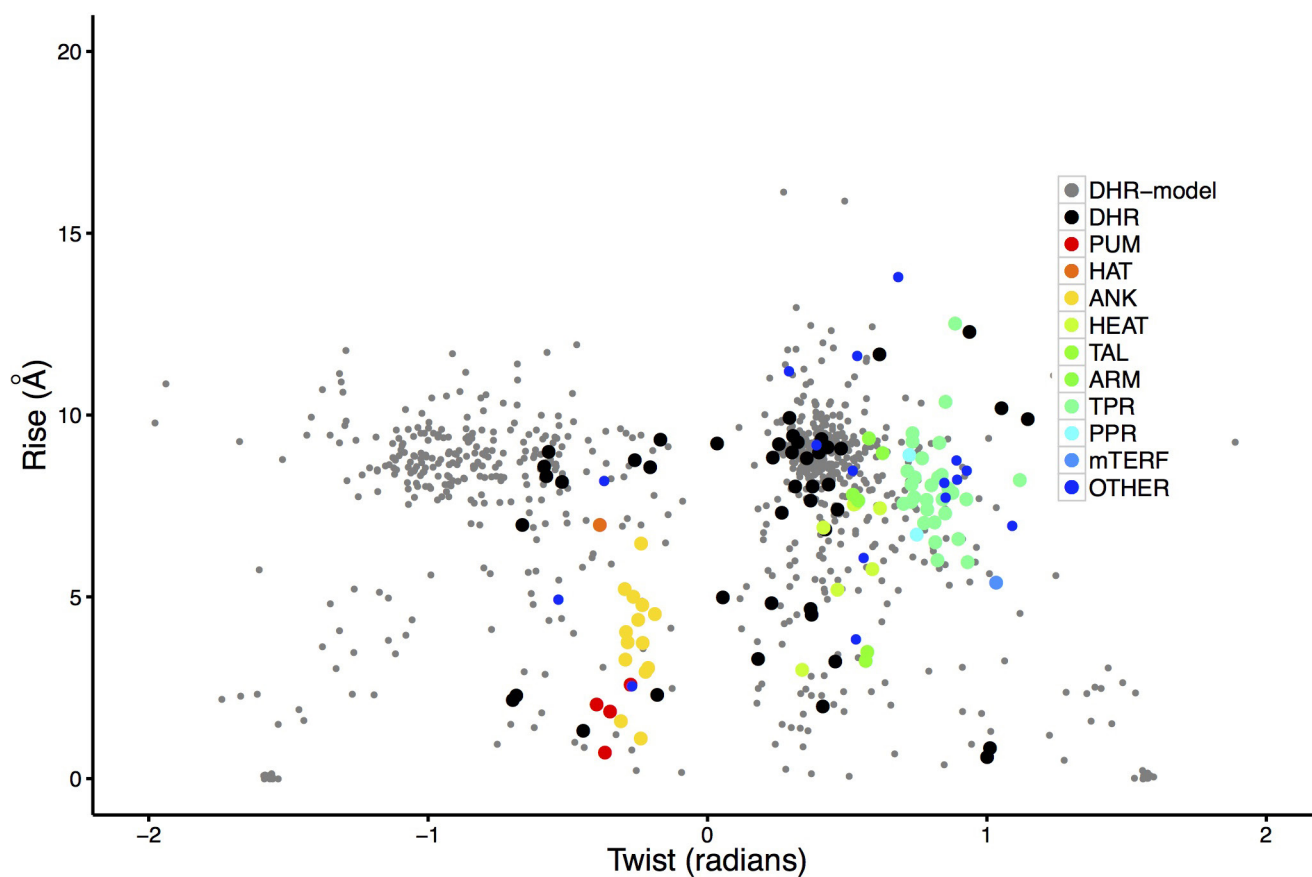


h

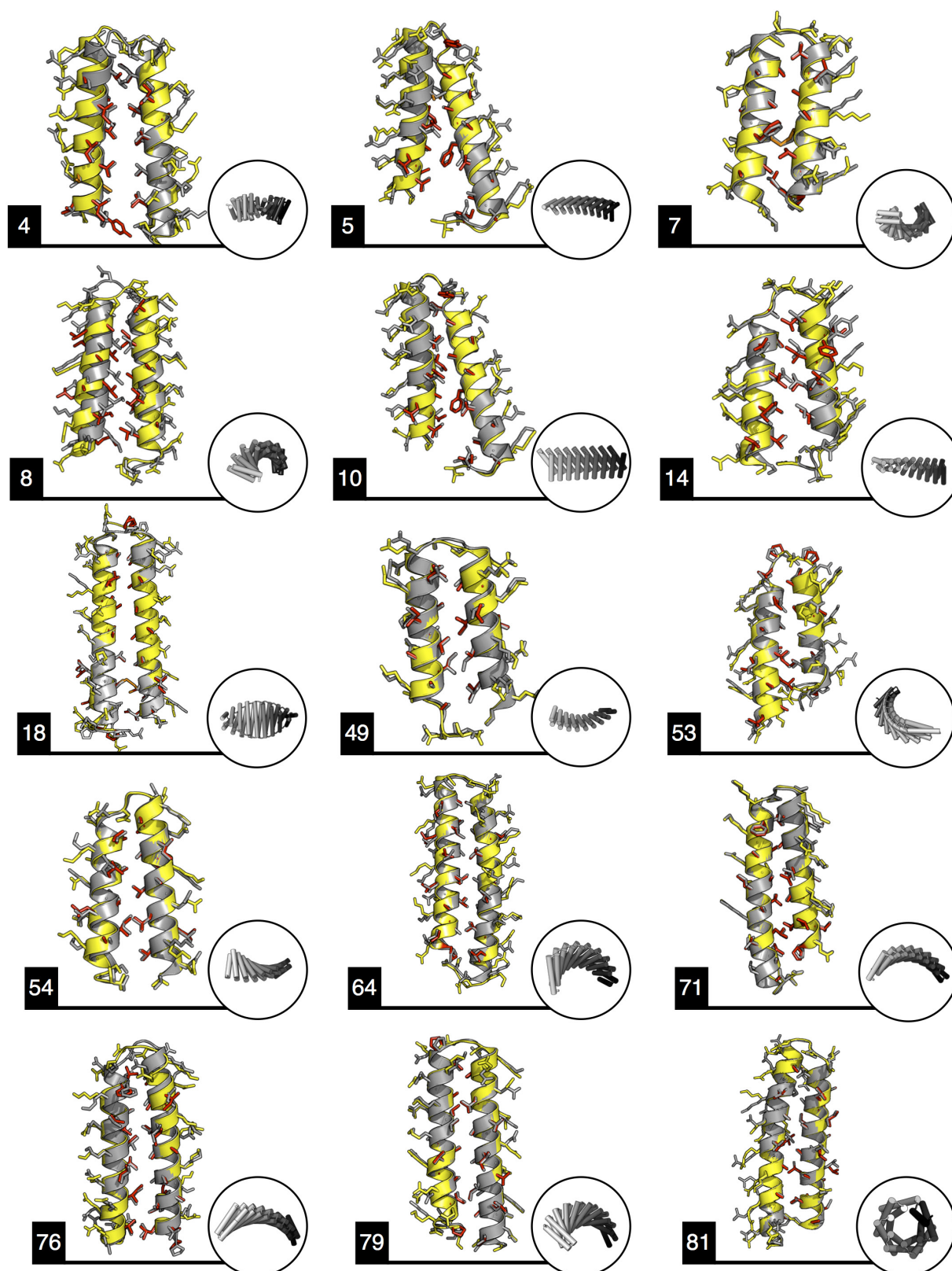


Extended Data Figure 3 | Model validation by *in silico* folding. To assess folding robustness seven sequence variants were made for each design. **a–g**, Illustrate the energy landscape explored by Rosetta *ab initio*. In red are the protein models produced by *ab initio* search, in green by side-chain repacking and minimization (relax). Models in deep global energy minima near the relaxed structures are considered folded.

The variant with highest density of *ab initio* models near the relax region was chosen for experimental characterization (blue box). **h**, Jalview sequence alignment of the first 100 residues of the variants. The yellow bar height indicates sequence conservation, while the black bar represents how often the consensus sequence occurs.

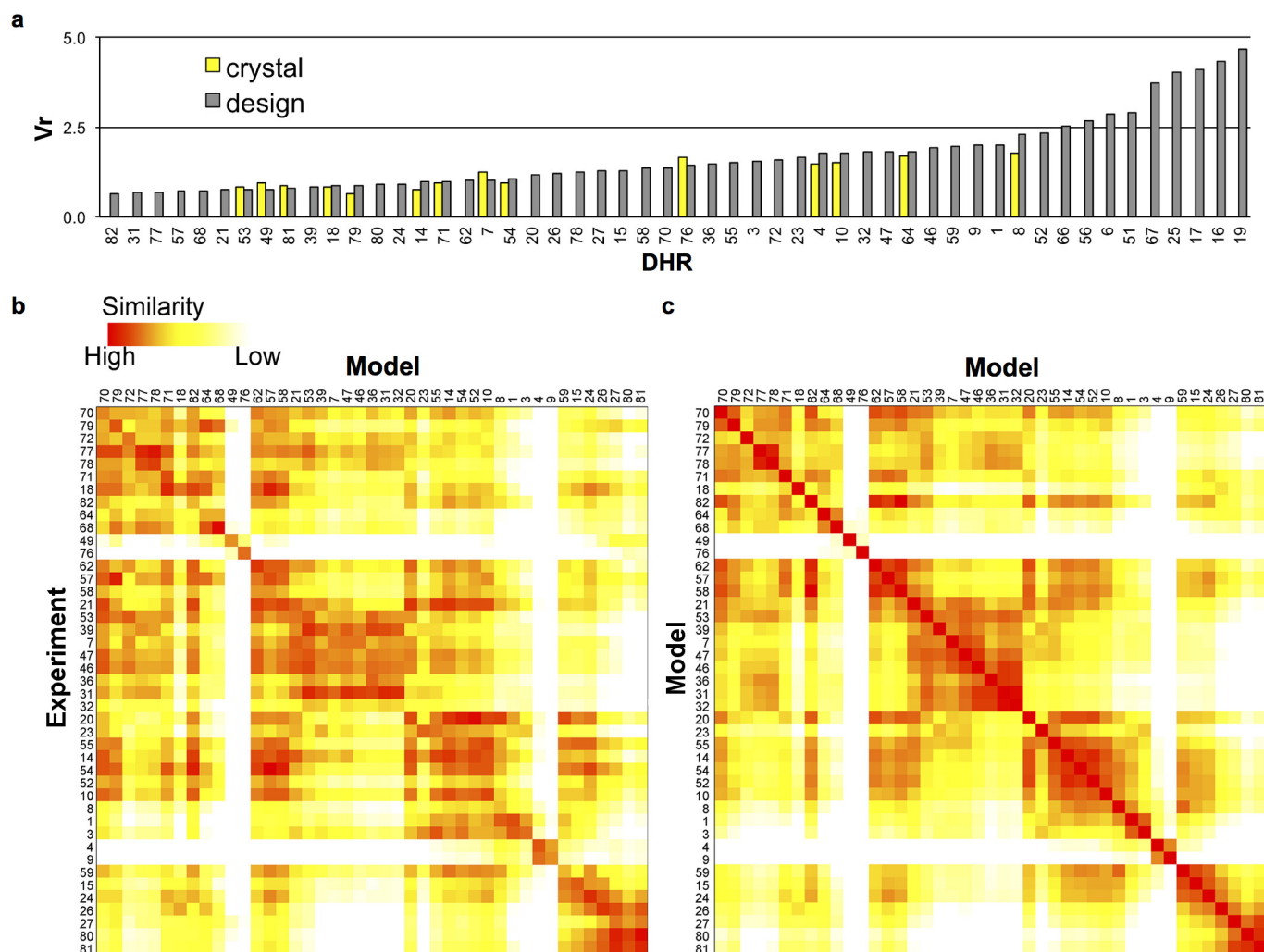


Extended Data Figure 4 | Distribution of DHR axial displacement (z) and twist (ω). Parameters for repeat protein family representatives were extracted as described in the Supplementary Information. The DHR models are the 761 proteins validated by *in silico* folding.



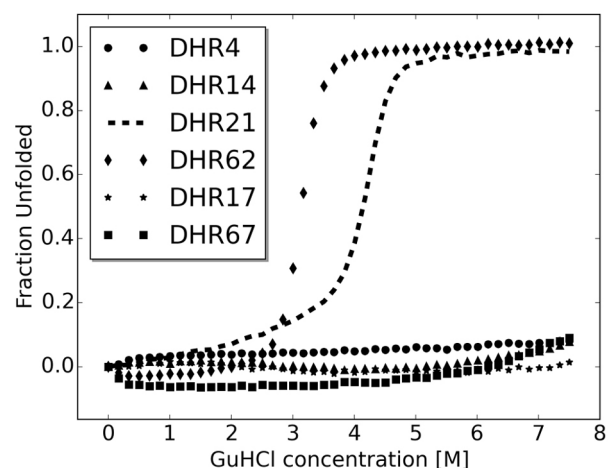
Extended Data Figure 5 | Superposition between single internal repeats (second repeat) of designs (grey) and crystal structures (yellow). Aliphatic and aromatic side chains are in red and cysteines are in orange. DHR7 and 18 show intra repeat disulfide bonds while DHR4 and 81 form

inter-repeat cystines. DHR5 does not form the expected S-S bond. Core side chains in design recapitulate the conformation observed in the crystal structures. Even when the backbone is shifted (for example, DHR5, 8, 15), rotamers are by large correctly predicted.

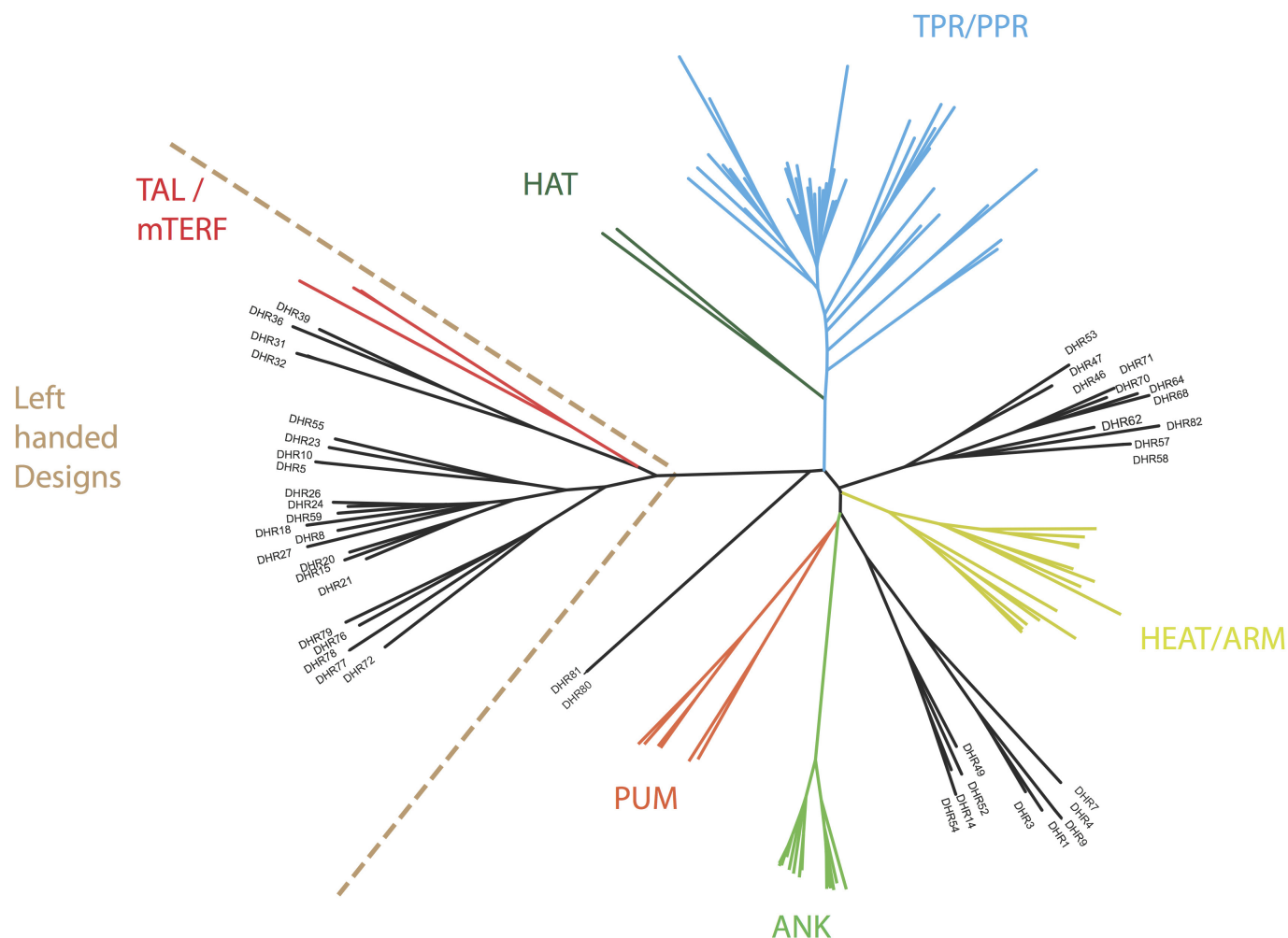


Extended Data Figure 6 | Structural validation by SAXS. **a**, V_r values for the fit of SAXS profiles to design models, in dark grey, and crystal structures, in yellow. For 43 designs, models are within the range defined by crystal structures. DHR49 and DHR76 form dimers in solution and the models employed the configuration observed in the crystal structures. Designs showing aggregation on the scattering profiles, including DHR5 for which the structure was solved, were not included in this figure. **b**, **c**, Pairwise V_r similarity maps³⁰ of 43 design models. **b**, Experimental-to-model profile similarity (**b**) and model-to-model profile similarity (**c**).

Models that are similar to each other show correlation off-diagonal in **c**, and the same pattern is observed when compared to experimental data in **b**. The order of display was obtained by clustering the original designed models by structural similarity. The ability to reproduce characteristic patterns within a large set of designs indicates that the models are capturing the relative structural similarities between proteins in solution. The scores are colour coded with red indicating best agreement and white lack of agreement.

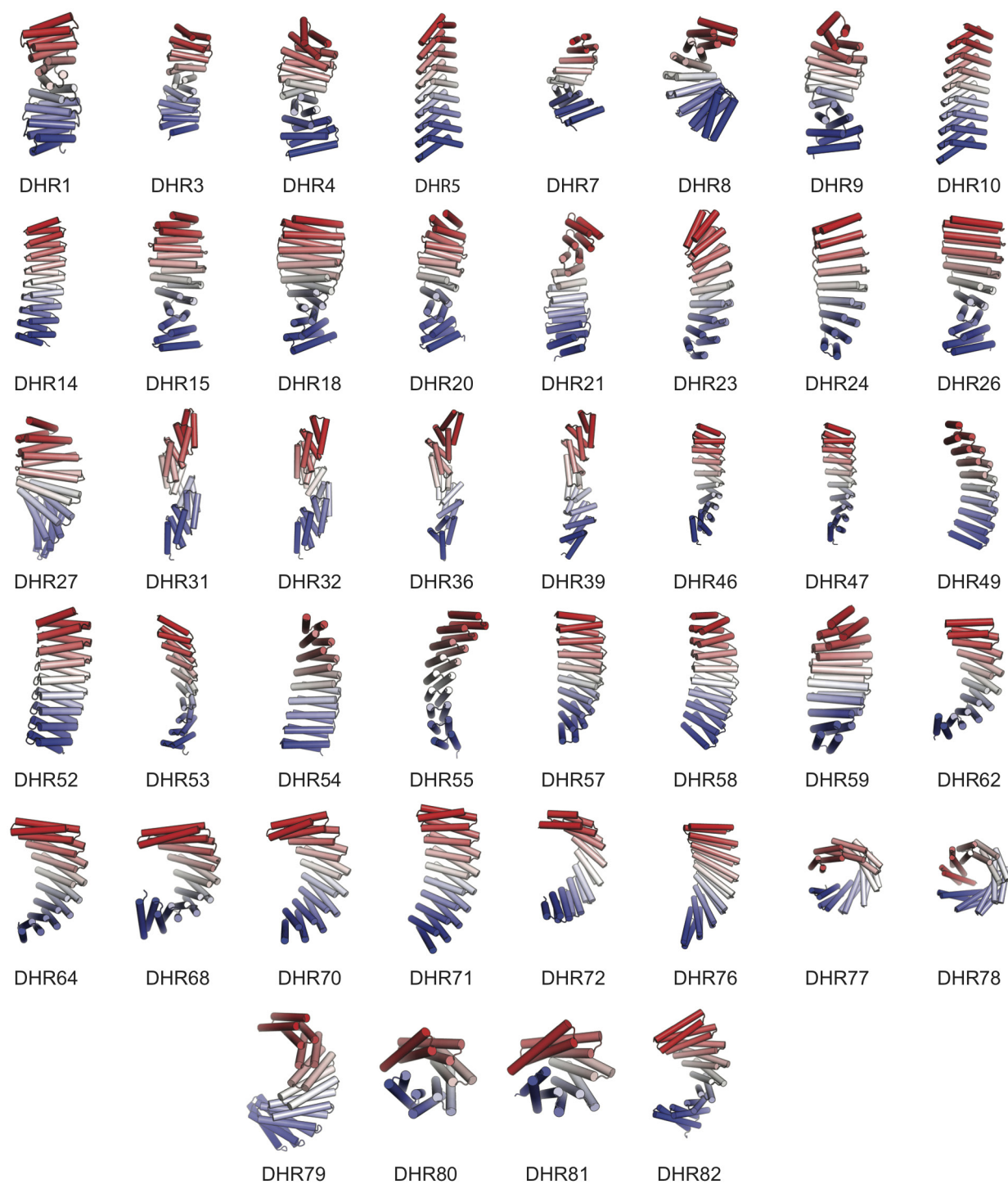


Extended Data Figure 7 | Designs are stable to chemical denaturation by guanidine hydrochloride (GuHCl). Circular-dichroism-monitored GuHCl denaturation experiments were carried for two designs for which crystal structures were solved (DHR4 and DHR14), two with overall shapes confirmed by SAXS (DHR21 and DHR62), and two with overall shapes inconsistent with SAXS (DHR17 and DHR67). In contrast to almost all native proteins, four of the six proteins do not denature at GuHCl concentrations up to 7.5 M. Both designs not confirmed by SAXS were extremely stable to GuHCl denaturation and hence are very well-folded proteins; the discrepancies between the computed and experimental SAXS profiles may be due to small amounts of oligomeric species or variation in overall twist.



Extended Data Figure 8 | Structural similarity between DHRs and repeat protein families. DHRs cluster separately from existing repeat proteins. DHRs are equally distributed between right-handed and left-handed repeats, as referred to the repeat handedness, in contrast to known α -helical repeat proteins, which are mostly right-handed. This result indicates that the handedness observed in known families is not

an intrinsic limitation of repeat proteins structures. Repeat handedness, as defined by Kobe and Kajava⁶, indicates the rotation of the main chain going from the N- to the C-terminal around the axis connecting the repeat centres of mass. The structural similarity tree was built using pairwise comparison as measured by TM-score.



Extended Data Figure 9 | Extended versions of models validated by SAXS and crystallography. DHRs were characterized as containing four repeats but the number of internal repeats can be increased without additional design steps. Extended models highlight the differences in twist and radius between the validated designs.