Title

Exploring the repeat protein universe through computational protein design

Journal

Nature, 528(7583)

ISSN

0028-0836

Authors

Brunette, TJ
Parmeggiani, Fabio
Huang, Po-Ssu
et al.

Publication Date

2015-12-01

DOI

10.1038/nature16162

Peer reviewed

Howard Hughes Medical Institute

Published as: *Nature*. 2015 December 24; 528(7583): 580–584.

Exploring the repeat protein universe through computational protein design

TJ Brunette[1,2,*], **Fabio Parmeggiani**[*,1,2], **Po-Ssu Huang**[*,1,2], **Gira Bhabha**[3], **Damian C. Ekiert**[4], **Susan E. Tsutakawa**[5], **Greg L. Hura**[5,6], **John A. Tainer**[5,7], and **David Baker**[1,2,8]

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

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

[3]Department of Cellular and Molecular Pharmacology, UCSF, San Francisco, California 94158, USA

[4]Department of Microbiology and Immunology, UCSF, San Francisco, California 94158, USA

[5]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

[6]Department of Chemistry and Biochemistry, University of California, Santa Cruz, California, USA

[7]Department of Molecular and Cellular Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA

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

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 unit[1] are widespread in nature and play critical roles in molecular recognition, signaling, and other essential biological processes[2]. Naturally occurring repeat proteins have been reengineered 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. 83 designs with sequences unrelated to known repeat proteins were experimentally characterized. 53 were monomeric and stable at 95 °C, and 43 have solution x-ray scattering spectra closely consistent with the design models. Crystal structures of 15 designs spanning a broad range of curvatures are in

Correspondence and requests for materials should be addressed to ; Email: dabaker@uw.edu
*these authors contributed equally

The authors declare no conflict of interest.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

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.E. setup crystallization trials and solved the crystal structures. F.P., S.T., G.H., J.T. collected and analyzed the SAXS data. F.P., T.B., P.-S.H. and D.B. wrote the manuscript with help from all the authors.

Crystal structures have been deposited in the RCSB protein databank with the accession numbers 5CWB (DHR4), 5CWC (DHR5), 5CWD (DHR7), 5CWF (DHR8), 5CWG (DHR10), 5CWH (DHR14), 5CWI (DHR18), 5CWJ (DHR49), 5CWK (DHR53), 5CWL(DHR54), 5CWM (DHR64), 5CWN (DHR71), 5CWO (DHR76), 5CWP (DHR79) and 5CWQ (DHR81).

close agreement with the design models with RMSDs 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 structure[6]. While in nature the sequences of these units generally differ, stable repeat proteins with identical units[7,8] have been designed for several families[9–21] and, for leucine rich repeats, customized designed units allow control of curvature[22] and new architectures[17]. All designed repeat structures to date have been based on naturally occurring repeat protein 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 Fig. 1–2). Our designs are entirely *de novo*; they are not based on naturally occurring repeat proteins. We focused on helix-loop-helix-loop as the basic repeating unit, as this is the simplest unit from which a wide diversity of curvatures can be generated (the simpler single helix-loop unit generates only straight rod-like models). 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 5776 (19×19×4×4) 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 assembly[23] was carried out to generate compact structures; each Monte Carlo move was made at the equivalent position in each repeat to preserve symmetry[20]. Rosetta design calculations[24] were then used to identify low energy amino acid sequences with good core packing[25]. 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 copy at each move 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, generating specialized capping repeats. 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 trajectories[26] 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 conformations sampled in the *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 structure[6]. It is thus convenient to classify these architectures based on three parameters defining a helix[22] (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 ~$r\cdot\omega$, and the total length of a unit is ~ sqrt($(r\omega)^2 + z^2$), hence the radius(r)-twist(ω) 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 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, colors 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 disulphide 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 N-terminal capping repeat, two internal repeats, and a C-terminal capping repeat including a 6-histidine tag. The proteins were expressed in *Escherichia coli* and purified by affinity chromatography. 74 of the 83 designs were expressed solubly and had the expected alpha helical CD spectrum at 25 °C, and 72 were stably folded at 95 °C (Supplementary Information, Experimental Data). 55 of these (66% of the original experimental set) were predominantly monomeric by analytical size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS); DHR49 and DHR76 were dimeric in solution. Structure stabilization with disulphide bonds did not systematically improve expression, solubility, or folding (Fig. 3a), probably because the designs are already very stable without disulphide bonds. Representative data on six of the designs are shown in Fig. 3b; the data on all 55 proteins is provided as Supplementary Information, 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 with Cα RMSDs from 0.7 Å to 2.5 Å and recapitulate the side chain orientations within the hydrophobic core (Fig. 4 and Extended Data Fig. 5). The designed disulfide bonds are all 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. 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 analyze all 55 proteins in solution, we used small angle X-ray scattering (SAXS)[27,28]. 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 data 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. To further assess the fit between models and experimental data, we employed the volatility ratio (Vr), which is more robust to experimental noise than the traditional χ^2 comparison used in SAXS[29]. We used the Vr values of the design models confirmed by crystallography for calibration; designs for which the Vr 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 Vr 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 (GuHCl) unfolding profiles. In contrast to almost all native proteins, four of the six designs do not denature at GuHCl 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 show here 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 remarkably 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, play 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.

Methods

Code availability

The Rosetta macromolecular modeling suite is available from 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, Rosetta_examples.

Similarity search

BLAST[30,31] and HHSEARCH[32] sequence similarity searches were performed with default settings. HHSEARCH was run on Pfam[33]. Sequence alignments were depicted using Jalview[34]. The structural similarity between designs and known helical repeat proteins was assessed by TM-align[35] on RepeatsDB[36] representative structures.

Protein expression and characterization

Genes were synthesized and cloned in vector pET21 by GenScript (Piscataway, NJ). Proteins were expressed in *E. coli* BL21(DE3), induced with 250 uM isopropyl-β-D-thiogalactopyransoide (IPTG) overnight at 22°C and purified by metal ion affinity chromatography (IMAC) and size exclusion chromatography (SEC) as described by Parmeggiani *et al.*[20] Cells were lysed by sonication and the clarified lysate was loaded on a NiNTA superflow column (Qiagen). Lysis and washing buffer was Tris 50mM, pH 8, NaCl 500mM, imidazole 30mM, glycerol 5% v/v. Lysozyme (2 mg/ml), DNAseI (0.2 mg/ml) and protease inhibitor cocktail (Roche) were added to the lysis buffer before sonication. Proteins were eluted in Tris 50 mM, pH 8, NaCl 500mM, imidazole 250mM, glycerol 5% v/v and dialyzed overnight either in tris 20mM, pH 8, NaCl 150mM. 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, Lakewood, NJ). Thermal denaturation was followed at 220nm in Tris 20mM, 50mM NaCl, pH 8. Proteins were considered folded if they had the expected alpha helical CD spectrum at 25°C and had either a sharp transition in thermal denaturation or a loss of less than 20% of 220 nm CD signal at 95°C. Chemical denaturation was monitored in a 1cm path-length

cuvette at 222nm with protein concentration of 0.05mg/ml in phosphate buffer 25mM NaCl 50mM 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 superdex200 increase for DHR59, 84, 93) (GE Healthcare) equilibrated in Tris 20mM, NaCl 150mM, 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). 74 of the 83 designs were expressed solubly and had the expected alpha helical CD spectrum at 25°C. 72 were stably folded at 95°C. DHR36 has Tm=75°C and DHR13 has a broad transition with Tm=62°C. Fifty-five of these were predominantly monodisperse. DHR49 and 76 were dimeric in solution. SDS-page gels, CD spectra, thermal denaturation and SEC 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 SEC 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 XDS[37] 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, due to the presence of 6 cysteine residues in the native protein, the DHR5 structure was solved by sulfur single wavelength anomalous dispersion (S-SAD) using a dataset collected at 7235 eV (Supplementary Information). Rigid body, restrained refinement with TLS and simulated annealing were carried out in Phenix[38]. Manual adjustment of the model was carried out in Coot[39]. The structures were validated using the Quality Control Check v2.8 developed by JCSG, which included Molprobity[40] (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

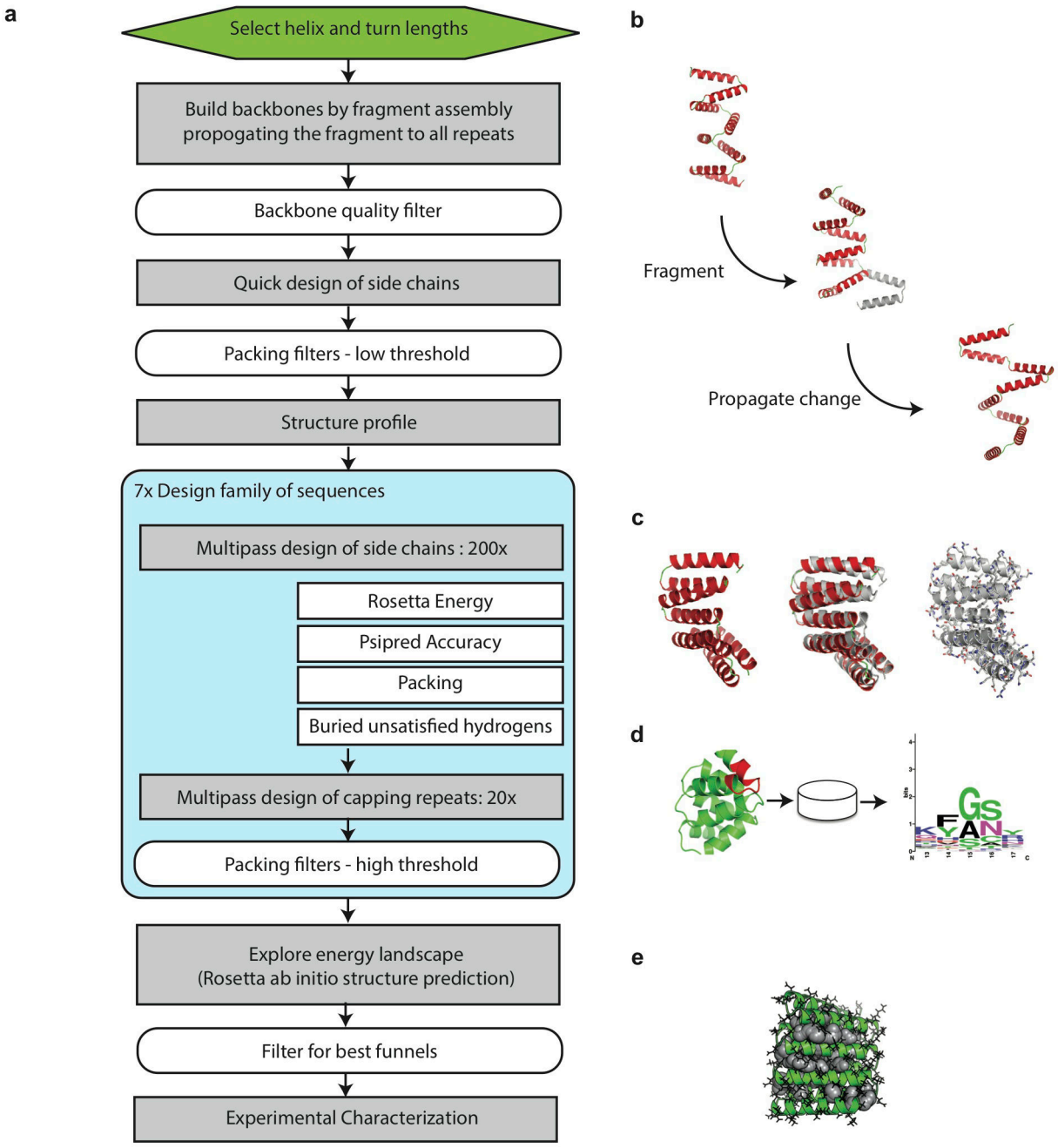
SAXS data on SEC-purified protein were collected at the SIBYLS 12.3.1 beamline at the Advanced Light Source, LBNL[28,41,42]. Scattering measurements were performed on 20 microliter samples and loaded into a helium-purged sample chamber, 1.5 m from the Mar165 detector. Data were collected on both the original gel filtration fractions and

samples concentrated ~2x–8x from individual fractions. Fractions prior to the void volume and concentrator eluates were used for buffer subtraction. Sequential exposures (0.5, 1, 2, and 5s) were taken at 12 keV to maximize signal to noise 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 insertion[23]. The results were clustered in Rosetta with a cluster radius of 2 Å and the cluster centers were used for comparison to the experimental data. We used FOXS[43,44] to calculate scattering profiles from cluster centers and fit them to the experimental data. The quality of fit between models and experimental SAXS data is usually assessed by the χ value[45], which, however, suffers from over-fitting in case of noisy datasets and domination of the low region of the scattering vector (q) on the value[27]. To avoid artificially low values that represent false positives, we instead used Volatility Ratio (Vr)[29] as primary metric for fit in the range of $0.015\ Å^{-1} < q < 0.25\ Å^{-1}$. Vr values of models with available crystal structures range from 0.7 to 2.3 (Supplementary Information Table 15). Vr=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 Vr similarity maps were obtained with a standardized fit procedure by averaging the scattering profile of the cluster centers from the five largest clusters and fitting the solvent hydration layer with parameters C1=1.015 and C2=2.0 for all the models. Vr was calculated in the range $0.04\ Å^{-1} < q < 0.3\ Å^{-1}$. The order of display was derived by shape similarity of original computational models using the program damsup[46] for superposition.

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

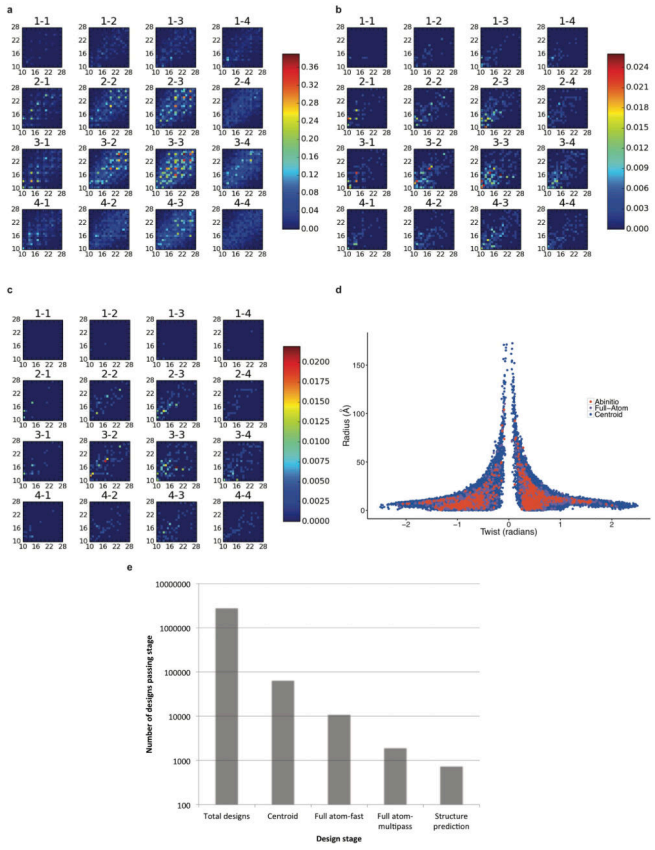
Extended Data



Extended Data Figure 1. Computational protocol for designing *de novo* repeat proteins

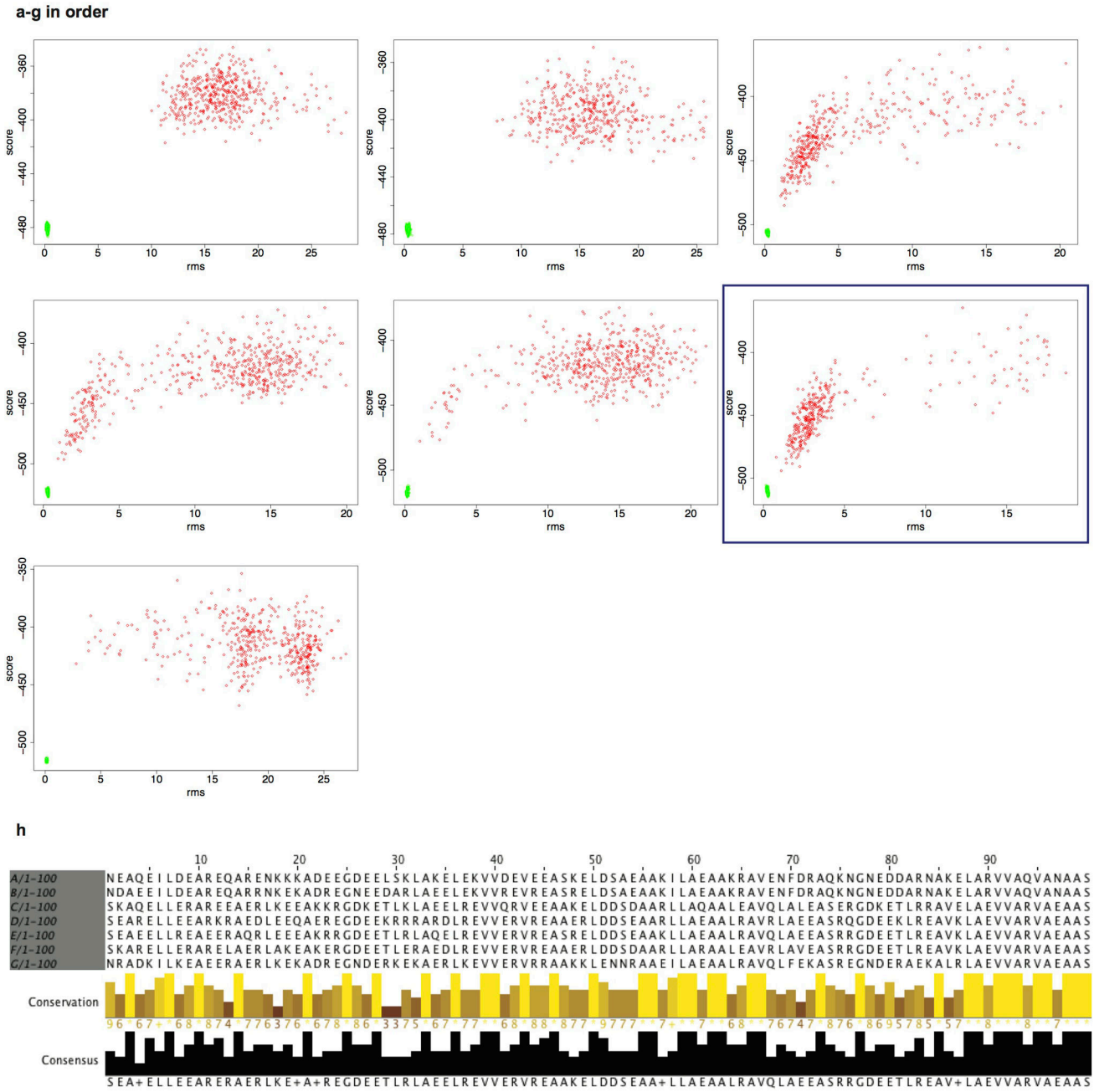
a, flowchart of the design protocol. The green box indicates user-controlled inputs, the grey boxes represent steps where protein structure is created or modified, and the white boxes indicate where structures are filtered. **b**, low resolution backbone build. **c**, quick full-atom design (grey) improves 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 PDB repository for structures within 0.5 Å RMSD. 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.



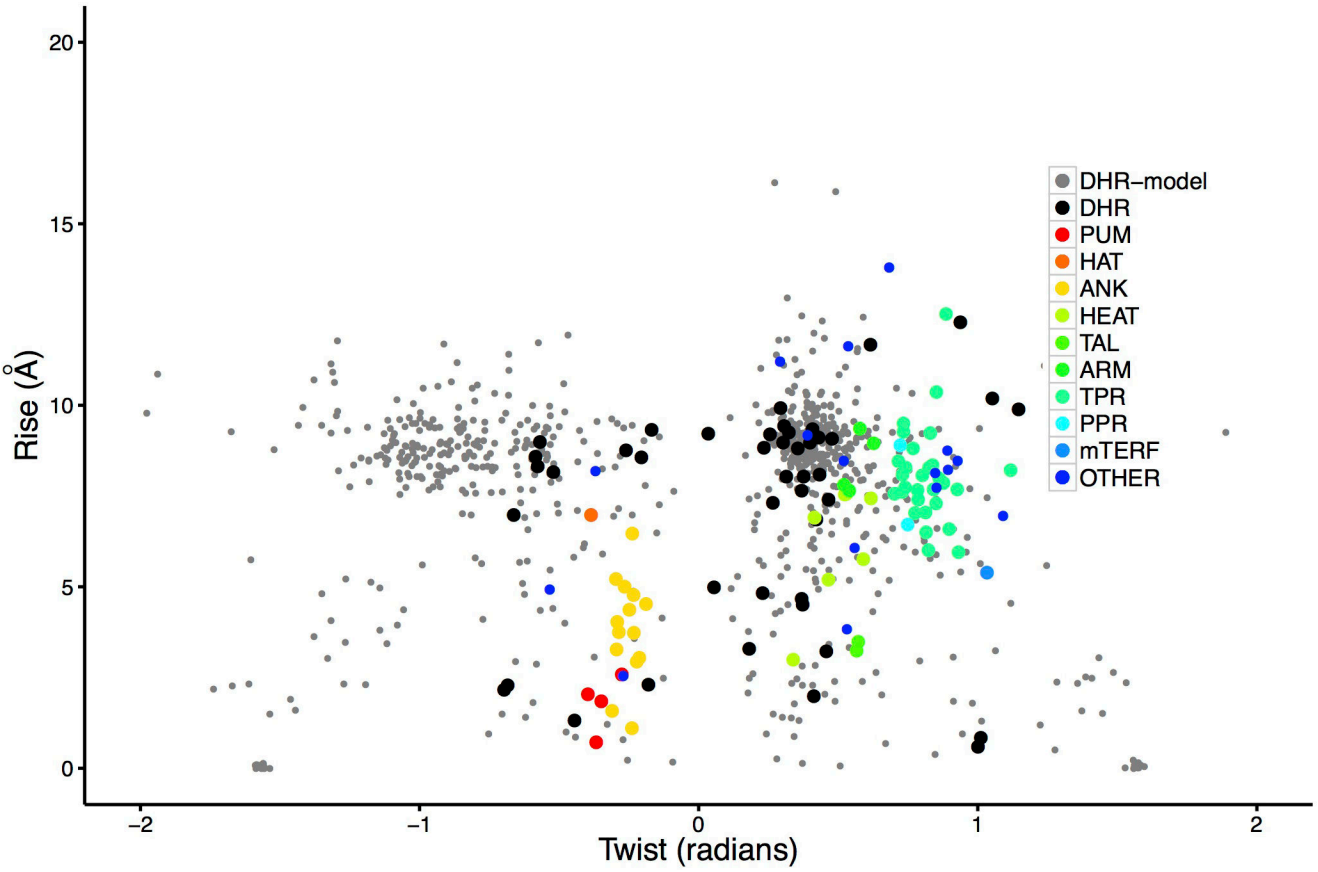
Extended Data Figure 2. Repeat space explored and model discrimination across design stages

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 loop1 and loop2 lengths is indicated on top. X and Y axis indicate helix1 and helix2 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.



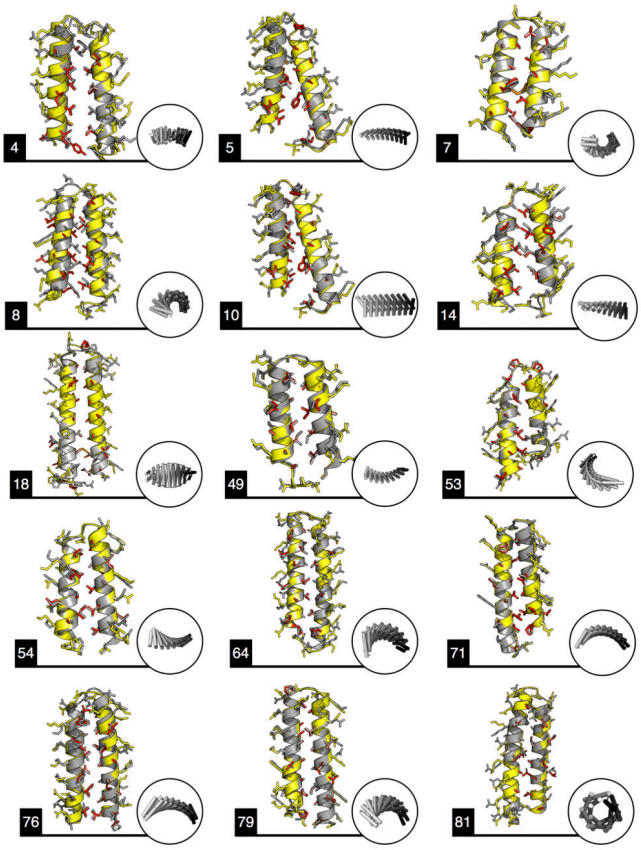
Extended Data Figure 3. Model validation by *in silico* folding

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 how often the consensus sequence occurs.



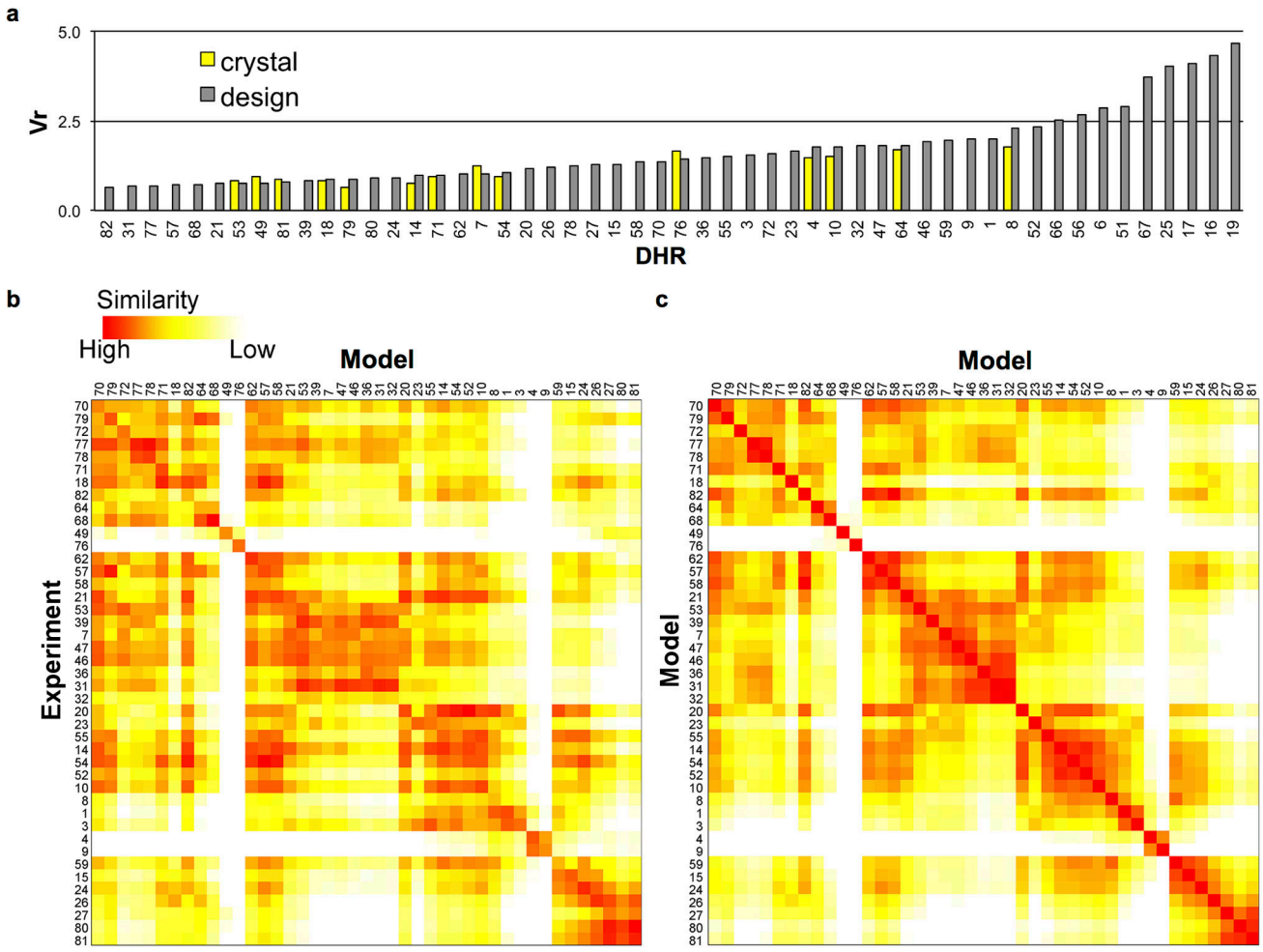
Extended Data Figure 4. Distribution of DHR axial displacement (z) and twist (ω)

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.



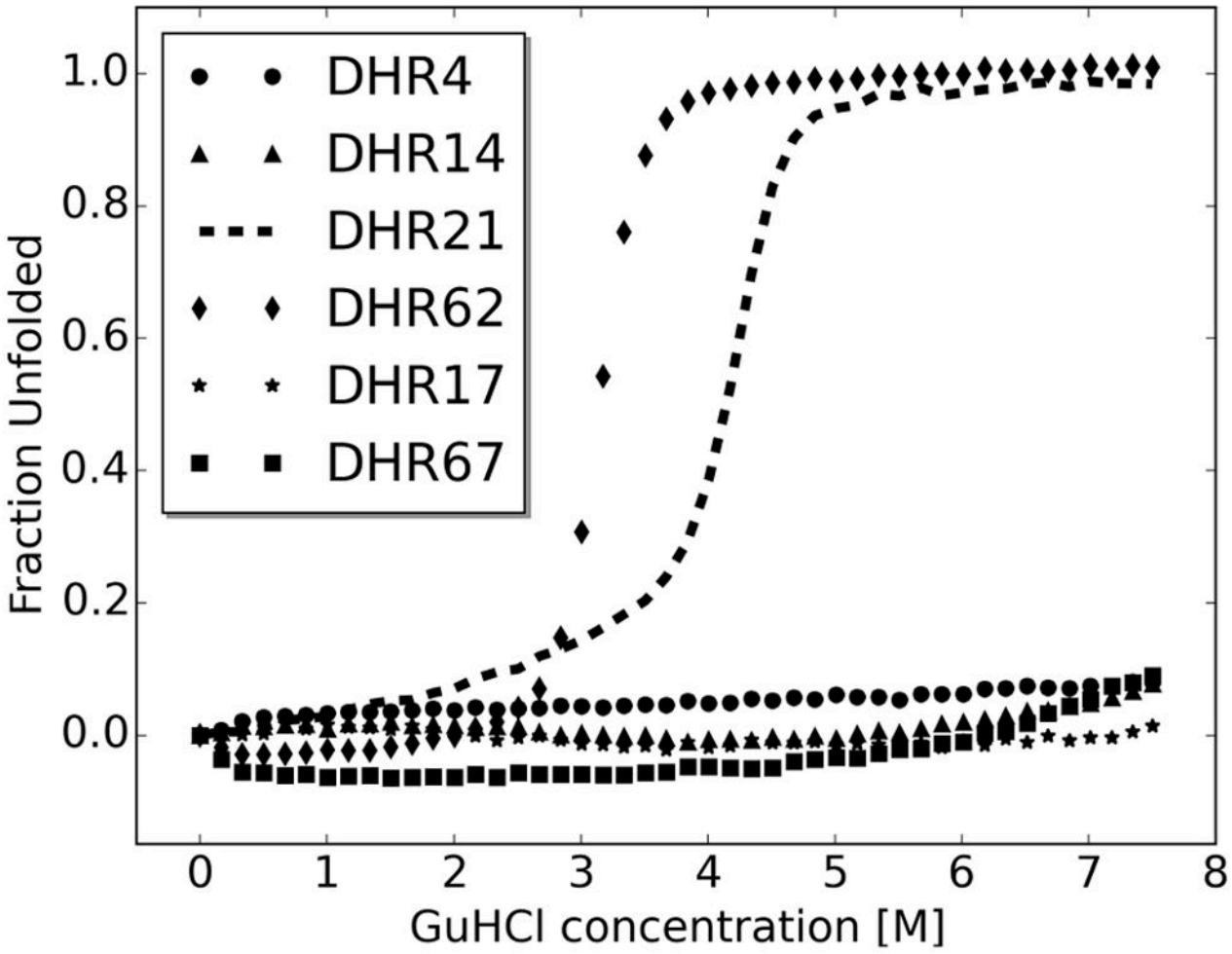
Extended Data Figure 5. Superposition between single internal repeats (second repeat) of designs (grey) and crystal structures (yellow)

Aliphatic and aromatic side chains are in red and cysteines are in orange. DHR7 and 18 show intra repeat disulphide 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 (e.g. DHR5, 8, 15), rotamers are by large correctly predicted.



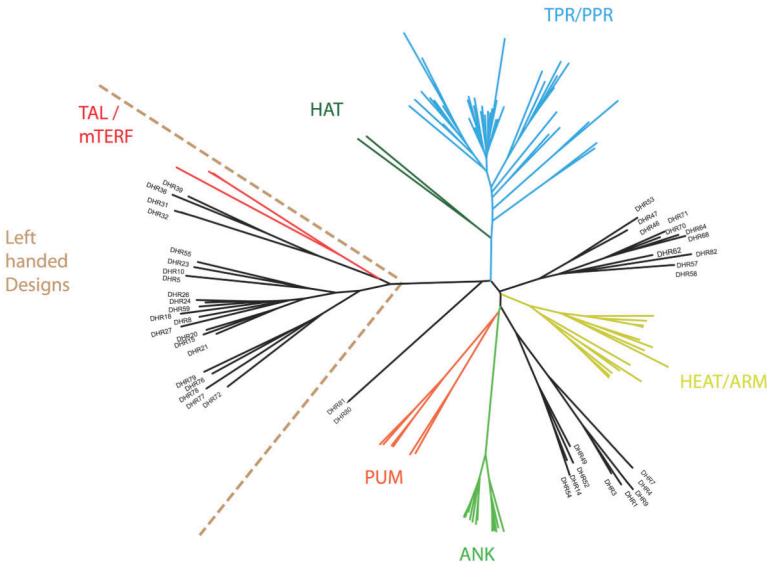
Extended Data Figure 6. Structural validation by SAXS

a, Vr 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** and **c**, pairwise Vr similarity maps[29] of 43 design models. **b**, experimental to model profile similarity, and **c**, model to model profile similarity. 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 color coded with red indicating best agreement and white lack of agreement.



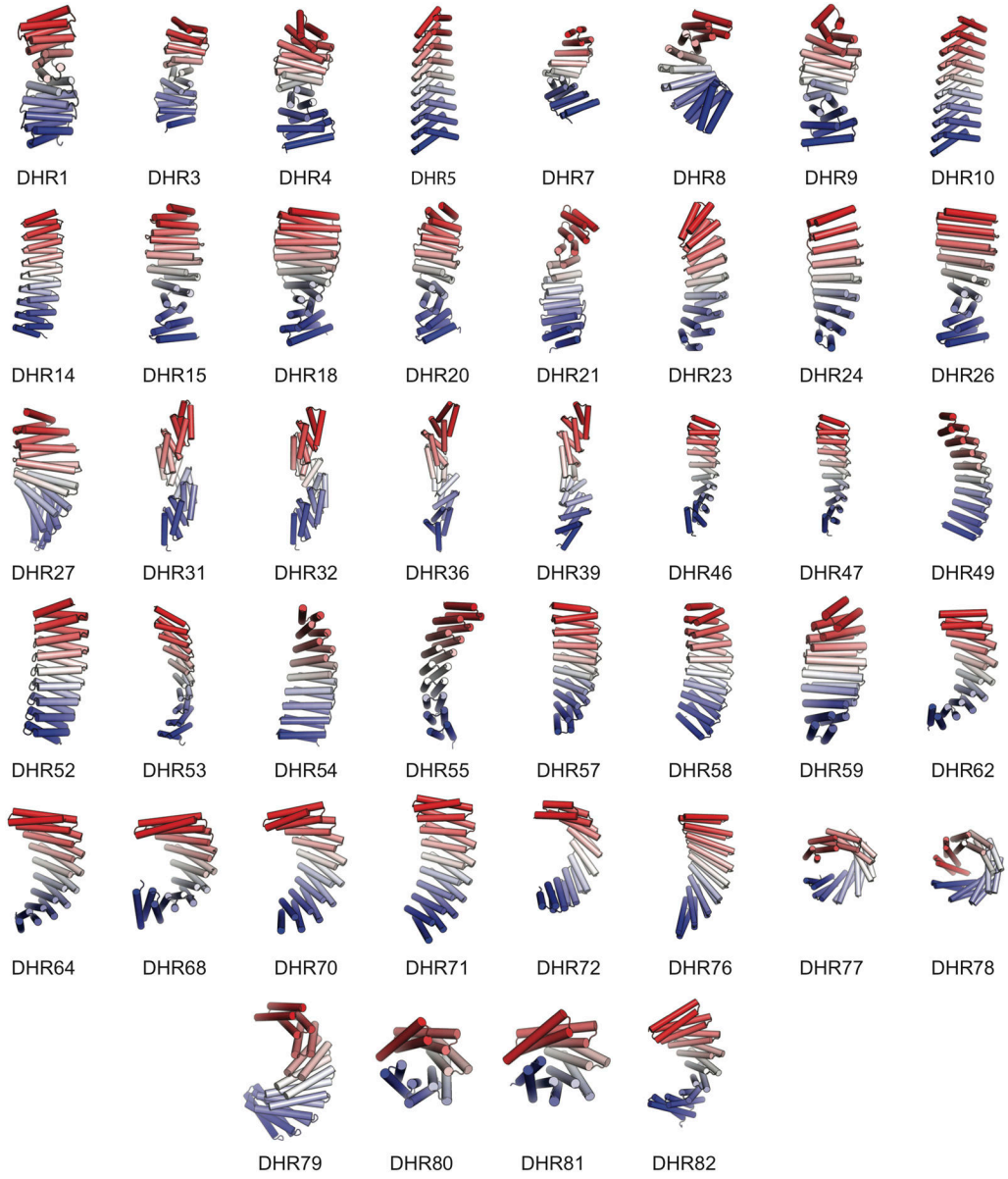
Extended Data Figure 7. Designs are stable to chemical denaturation by guanidine HCl (GuHCl)

Circular dichroism monitored GuHCl denaturant 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 alpha 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 Kajava[6], indicates the rotation of the main chain going from the N- to the C-terminal around the axis connecting the repeat centers 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank David Kim and members of the protein production facility at the Institute for Protein Design. This work was facilitated though 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 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 James 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 U.S. 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.

References

1. Kajava AV. Tandem repeats in proteins: From sequence to structure. J Struct Biol. 2012; 179:279–288. [PubMed: 21884799]
2. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats1. J Mol Biol. 1999; 293:151–160. [PubMed: 10512723]
3. Binz HK, et al. High-affinity binders selected from designed ankyrin repeat protein libraries. Nat Biotechnol. 2004; 22:575–582. [PubMed: 15097997]
4. Varadamsetty 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. 2012; 424:68–87. [PubMed: 22985964]
5. Cortajarena AL, Liu TY, Hochstrasser M, Regan L. Designed Proteins To Modulate Cellular Networks. ACS Chem Biol. 2010; 5:545–552. [PubMed: 20020775]
6. Kobe B, Kajava AV. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem Sci. 2000; 25:509–515. [PubMed: 11050437]
7. Wetzel SK, Settanni G, Kenig M, Binz HK, Plückthun A. Folding and Unfolding Mechanism of Highly Stable Full-Consensus Ankyrin Repeat Proteins. J Mol Biol. 2008; 376:241–257. [PubMed: 18164721]
8. Cortajarena AL, Regan L. Calorimetric study of a series of designed repeat proteins: Modular structure and modular folding. Protein Sci. 2011; 20:336–340. [PubMed: 21280125]
9. Binz HK, Stumpp MT, 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. 2003; 332:489–503. [PubMed: 12948497]
10. Mosavi LK, Minor DL, Peng Z. Consensus-derived structural determinants of the ankyrin repeat motif. Proc Natl Acad Sci. 2002; 99:16029–16034. [PubMed: 12461176]
11. Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure. 2003; 11:497–508. [PubMed: 12737816]
12. 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. 2010; 404:307–327. [PubMed: 20887736]
13. Lee SC, et al. Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. Proc Natl Acad Sci. 2012; 109:3299–3304. [PubMed: 22328160]
14. Parmeggiani F, et al. Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. J Mol Biol. 2008; 376:1282–1304. [PubMed: 18222472]
15. Yadid I, Tawfik DS. Reconstruction of Functional β-Propeller Lectins via Homo-oligomeric Assembly of Shorter Fragments. J Mol Biol. 2007; 365:10–17. [PubMed: 17054983]

16. Coquille S, et al. An artificial PPR scaffold for programmable RNA recognition. Nat Commun. 2014; 5
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. 2014; 111:17875–17880. [PubMed: 25427795]
18. Lee J, Blaber M. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc Natl Acad Sci. 2011; 108:126–130. [PubMed: 21173271]
19. Voet ARD, et al. Computational design of a self-assembling symmetrical β-propeller protein. Proc Natl Acad Sci. 2014; 111:15102–15107. [PubMed: 25288768]
20. Parmeggiani F, et al. A General Computational Approach for Repeat Protein Design. J Mol Biol. 2015; 427:563–575. [PubMed: 25451037]
21. Tripp KW, Barrick D. Enhancing the Stability and Folding Rate of a Repeat Protein through the Addition of Consensus Repeats. J Mol Biol. 2007; 365:1187–1200. [PubMed: 17067634]
22. Park K, et al. Control of repeat-protein curvature by computational protein design. Nat Struct Mol Biol. 2015; 22:167–174. [PubMed: 25580576]
23. Huang PS, et al. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. PLoS ONE. 2011; 6:e24109. [PubMed: 21909381]
24. Leaver-Fay A, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–574. [PubMed: 21187238]
25. Huang PS, et al. High thermodynamic stability of parametrically designed helical bundles. Science. 2014; 346:481–485. [PubMed: 25342806]
26. Bradley P, Misura KMS, Baker D. Toward High-Resolution *de Novo* Structure Prediction for Small Proteins. Science. 2005; 309:1868–1871. [PubMed: 16166519]
27. Rambo RP, Tainer JA. Super-Resolution in Solution X-Ray Scattering and Its Applications to Structural Systems Biology. Annu Rev Biophys. 2013; 42:415–441. [PubMed: 23495971]
28. Hura GL, et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods. 2009; 6:606–612. [PubMed: 19620974]
29. Hura GL, et al. Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. Nat Methods. 2013; 10:453–454. [PubMed: 23624664]
30. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
31. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]
32. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012; 9:173–175. [PubMed: 22198341]
33. Punta M, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–D301. [PubMed: 22127870]
34. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009; 25:1189–1191. [PubMed: 19151095]
35. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33:2302–2309. [PubMed: 15849316]
36. Di Domenico T, et al. RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res. 2014; 42:D352–D357. [PubMed: 24311564]
37. Kabsch W. XDS. Acta Crystallogr Sect D. 2010; 66:125–132. [PubMed: 20124692]
38. Adams PD, et al. PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr Sect D. 2002; 58:1948–1954. [PubMed: 12393927]
39. Emsley P, Cowtan K. *Coot*: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004; 60:2126–2132. [PubMed: 15572765]
40. Chen VB, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D. 2010; 66:12–21. [PubMed: 20057044]

41. 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. 2013; 46:1–13. [PubMed: 23396808]

42. Classen S, et al. Software for the high-throughput collection of SAXS data using an enhanced Blu-Ice/DCS control system. J Synchrotron Radiat. 2010; 17:774–781. [PubMed: 20975223]

43. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. Biophys J. 2013; 105:962–974. [PubMed: 23972848]

44. Schneidman-Duhovny D, Hammel M, Sali A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. 2010; 38:W540–W544. [PubMed: 20507903]

45. Svergun D, Barberato C, Koch MHJ. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr. 1995; 28:768–773.

46. Petoukhov MV, et al. New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr. 2012; 45:342–350. [PubMed: 25484842]

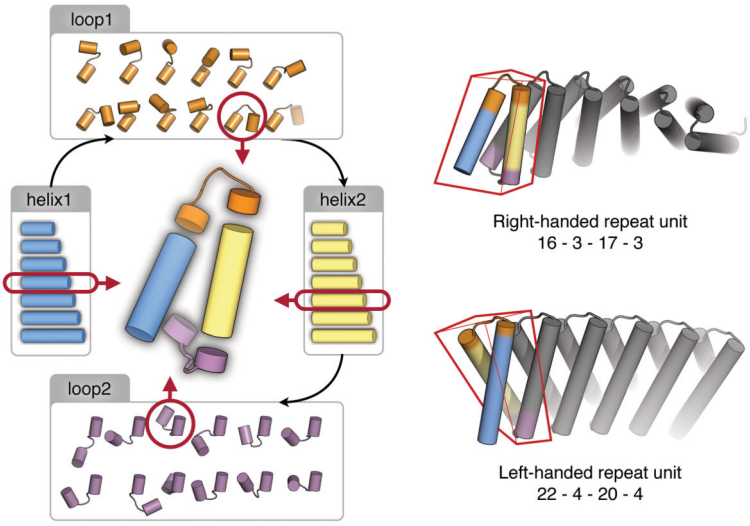


Figure 1. Schematic overview of the computational design method

The lengths of each helix and loop were systematically enumerated. For each choice of helix and loop lengths, individual repeat units (red boxes on right) were built up from fragments of proteins of known structure, and then propagated to generate extended repeating structures (gray) with right-handed or left-handed twist.

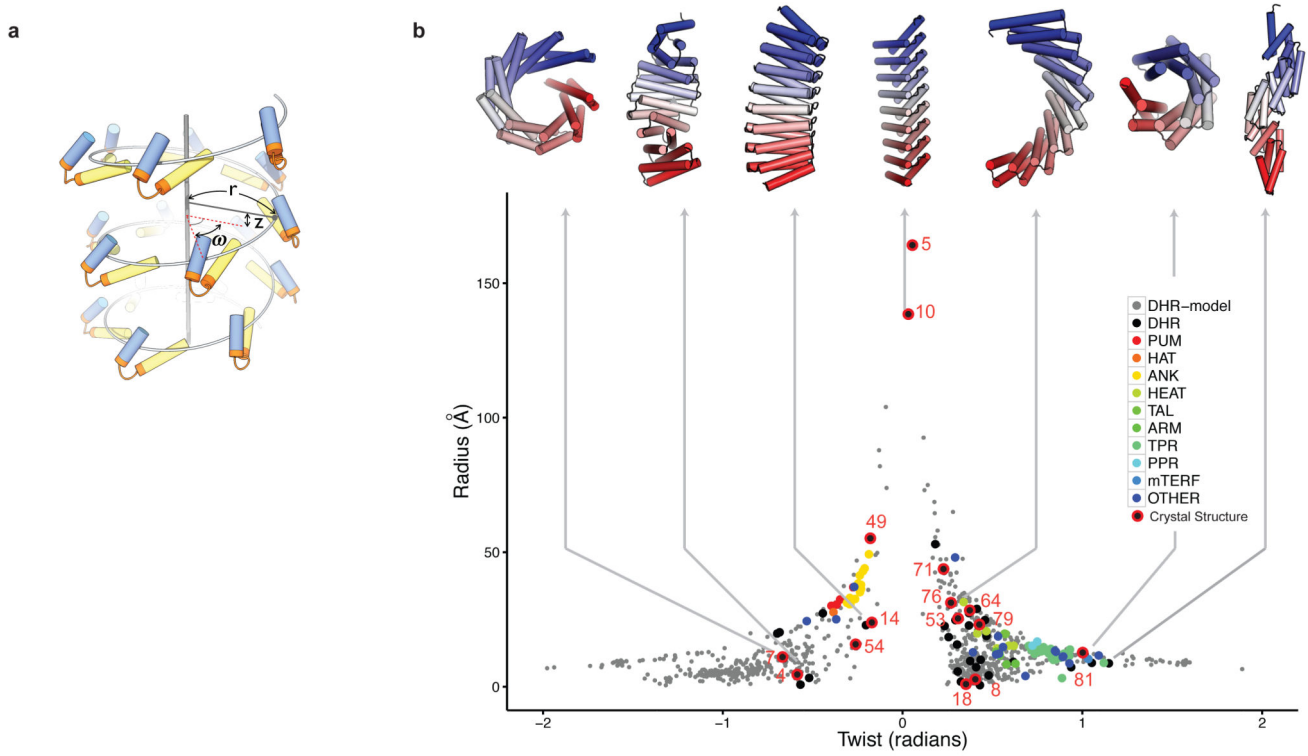


Figure 2. The helical repeat protein universe

a, the geometry of a repeat protein can be described by axial displacement (z), radius of the helix (r) and angular displacement or twist (ω) between repeat units. **b**, designed helical repeat (DHR) proteins (grey) cover radius and twist ranges not found in native repeat protein families (colors). Designs forming right-handed superhelices have positive ω values; left-handed, negative ω values. Native families abbreviations: 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. Designs structurally validated by small angle x-ray scattering (SAXS) (black) or crystallography (black with red circle) are distributed throughout the space. On top, representative experimentally validated designs with a variety of shapes.

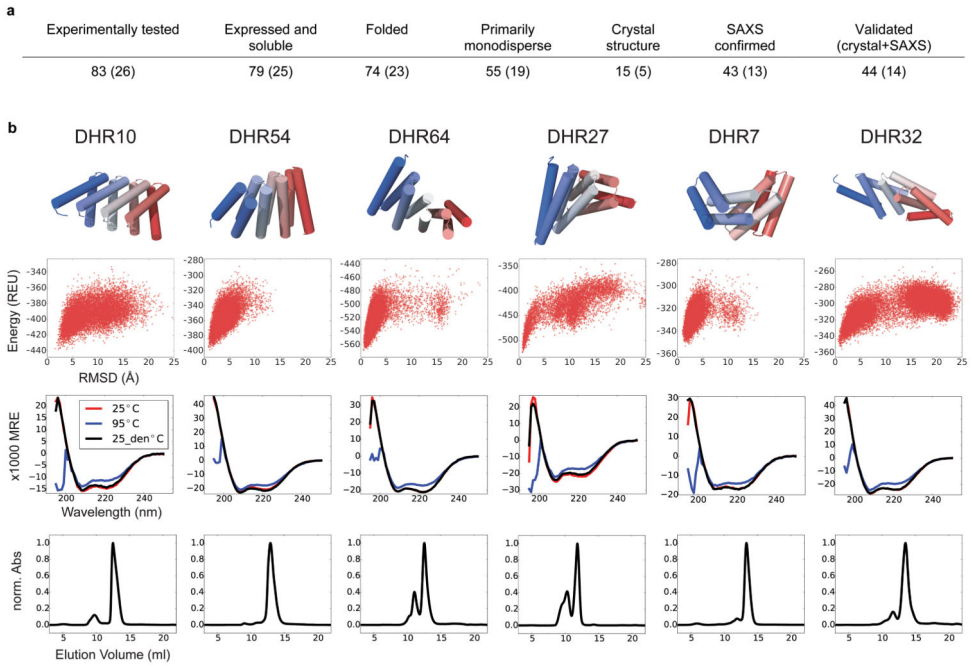


Figure 3. Characterization of designed repeat proteins

a, Design success rate. Values for subset with disulfide bonds are in parentheses. **b**, results on six representative designs. Top row: design models. Second row: computed energy landscapes. Energy is on y axis (REU, Rosetta energy unit) and RMSD from design model on x axis. All six landscapes are strongly funneled into the designed energy minimum. Third row: CD spectra collected at 25°C (red), 95°C (blue) and back to 25°C (black). The proteins do not denature within this temperature range (MRE, mean residue elipticity; $deg \bullet cm^2 \bullet dmol^{-1} \bullet residue^{-1}$). Bottom row: SEC elution profile directly after affinity chromatography purification. The designs are mostly monodisperse. The maximum absorbance at 280 nm was normalized to 1.

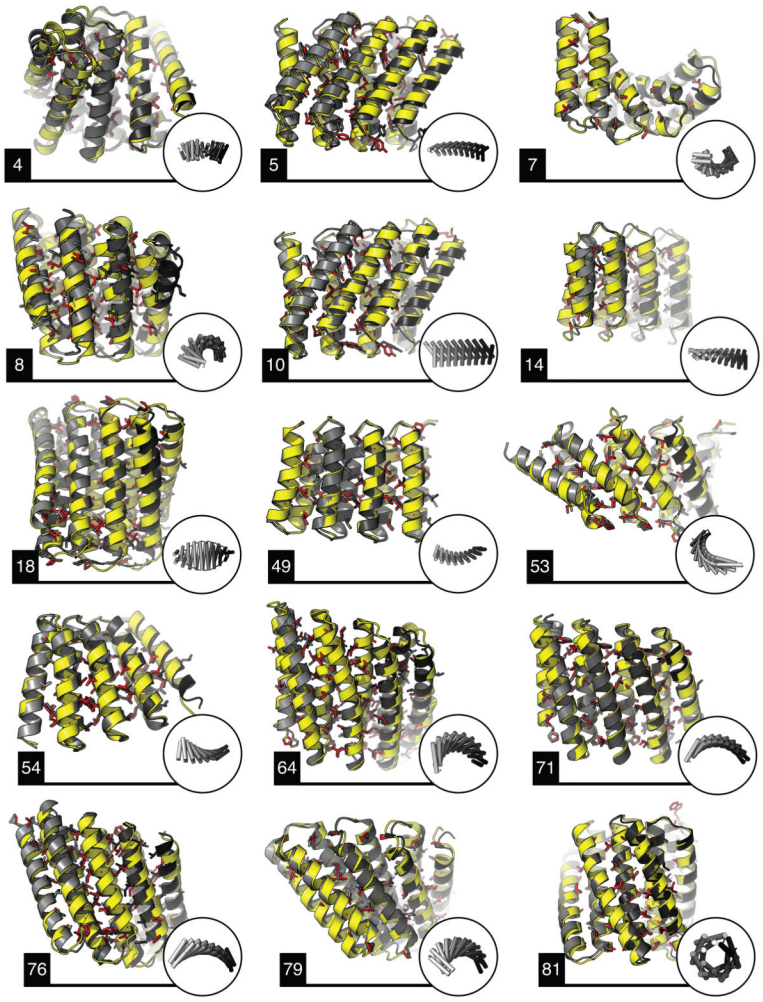


Figure 4. Crystal structures of fifteen designs are in close agreement with the design models

Crystal structures are in yellow, and the design models in grey. Insets in circles show the overall shape of the repeat protein. The RMSD 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 (in red) are largely captured by the designs (Extended Data Fig. 5).