nature structural & molecular biology

ARTICLES

https://doi.org/10.1038/s41594-018-0141-6

De novo design of a non-local β-sheet protein with high stability and accuracy

Enrique Marcos[1,2,3,10*], Tamuka M. Chidyausiku[1,2,10], Andrew C. McShan[4], Thomas Evangelidis[5], Santrupti Nerli[4,6], Lauren Carter[1,2], Lucas G. Nivón[1,2,7], Audrey Davis[1,2,8], Gustav Oberdorfer[1,2,9], Konstantinos Tripsianes[5], Nikolaos G. Sgourakis[4] and David Baker[1,2*]

β-sheet proteins carry out critical functions in biology, and hence are attractive scaffolds for computational protein design. Despite this potential, de novo design of all-β-sheet proteins from first principles lags far behind the design of all-α or mixed-αβ domains owing to their non-local nature and the tendency of exposed β-strand edges to aggregate. Through study of loops connecting unpaired β-strands (β-arches), we have identified a series of structural relationships between loop geometry, side chain directionality and β-strand length that arise from hydrogen bonding and packing constraints on regular β-sheet structures. We use these rules to de novo design jellyroll structures with double-stranded β-helices formed by eight antiparallel β-strands. The nuclear magnetic resonance structure of a hyperthermostable design closely matched the computational model, demonstrating accurate control over the β-sheet structure and loop geometry. Our results open the door to the design of a broad range of non-local β-sheet protein structures.

β-sheet protein domains are ubiquitous in nature, carrying out a wide range of functions: transporting hydrophobic molecules, recognition and enzymatic processing of carbohydrates, and scaffolding of virus capsids and antibodies, among others. Although β-sheet protein scaffolds are well suited for incorporating new functions, their design from first principles remains an outstanding challenge. Recent progress in de novo protein design has enabled the accurate design of many hyperstable and structurally diverse proteins, but so far, other than short β-sheet peptides[1–3], all exhibit either all-α or mixed-αβ folds[4]. The design of these last has been considerably facilitated by the derivation of a set of rules describing constraints on the backbone geometry of the loops connecting secondary structure elements[5], but all-β proteins contain additional features that are less well understood. All-β-sheet structures are particularly challenging to design from scratch[6] because a larger fraction of the interactions are non-local (that is, between residues distant along the linear sequence), leading to slower folding rates[7], and because β-strands, particularly at the edges of β-sheets, can aggregate into amyloid-like structures. Hence, few β-sheet protein design studies have sought to generate new backbone structures[8,9] and, except for a recent β-barrel structure with primarily local strand pairings[10], those designs confirmed by high-resolution structure determination have relied heavily on sequence information[11,12] and backbone structures[13,14] from naturally occurring β-sheet proteins.

So far, the de novo design of β-sheet loop connections has been limited to β-hairpins (two antiparallel β-strands interacting via backbone hydrogen bonding and connected through a loop), which is the most local strand pairing possible and, in principle, the fastest to fold. However, these structures lack a critical feature of non-local globular all-β structures: loops connecting β-strands not paired to each other, also known as β-arches[15]. These loops connect distinct β-sheets and pair β-strands with larger sequence separation, and they are essential for enabling the protein fold complexity observed in antibodies, β-solenoids, jellyrolls and Greek key-containing structures generally. Here we set out to identify the general principles for designing non-local β-sheet structures.

Results

Constraints on β-arch geometry. We undertook investigation of the constraints on the backbone geometry of β-strands and connecting loops that arise from hydrogen bonding and the requirement for a compact hydrophobic core. We studied side chain directionality patterns of the two β-strand residues adjacent to β-arch loops (Fig. 1a, left) in naturally occurring protein structures, defining the side chain orientation of the β-strand residue preceding the loop as 'concave' (represented by ↓) if its CαCβ vector is parallel to the vector *d* from the first to the second β-strand, and 'convex' (represented by ↑) if the CαCβ vector is antiparallel to *d*. For the residue following the loop, the side chain pattern is described in the same way, but instead using the vector from the second to the first β-strand (*–d*) as a reference (Fig. 1a). This results in four possible β-arch loop side chain orientation patterns: ↑↑, ↑↓, ↓↑ and ↓↓. We analyzed the side chain patterns and the local backbone geometry (as described with ABEGO torsion bins[16]) of 5,061 β-arch loops from a non-redundant database of natural protein structures (torsion bins A and B are the α-helix and extended regions; G and E regions are the positive φ angle equivalents of A and B; and O is the *cis*-peptide bond conformation; Supplementary Fig. 1). We found that all four side chain orientation patterns frequently occur, and, in contrast to other types of loop connections (that is, αβ, βα and

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Spain. [4]Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA, USA. [5]CEITEC—Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [6]Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA, USA. [7]Present address: Cyrus Biotechnology, Seattle, WA, USA. [8]Present address: Amazon, Seattle, WA, USA. [9]Present address: Institute of Biochemistry, Graz University of Technology, Graz, Austria. [10]These authors contributed equally: Enrique Marcos, Tamuka M. Chidyausiku. *e-mail: emarcos82@gmail.com; dabaker@uw.edu

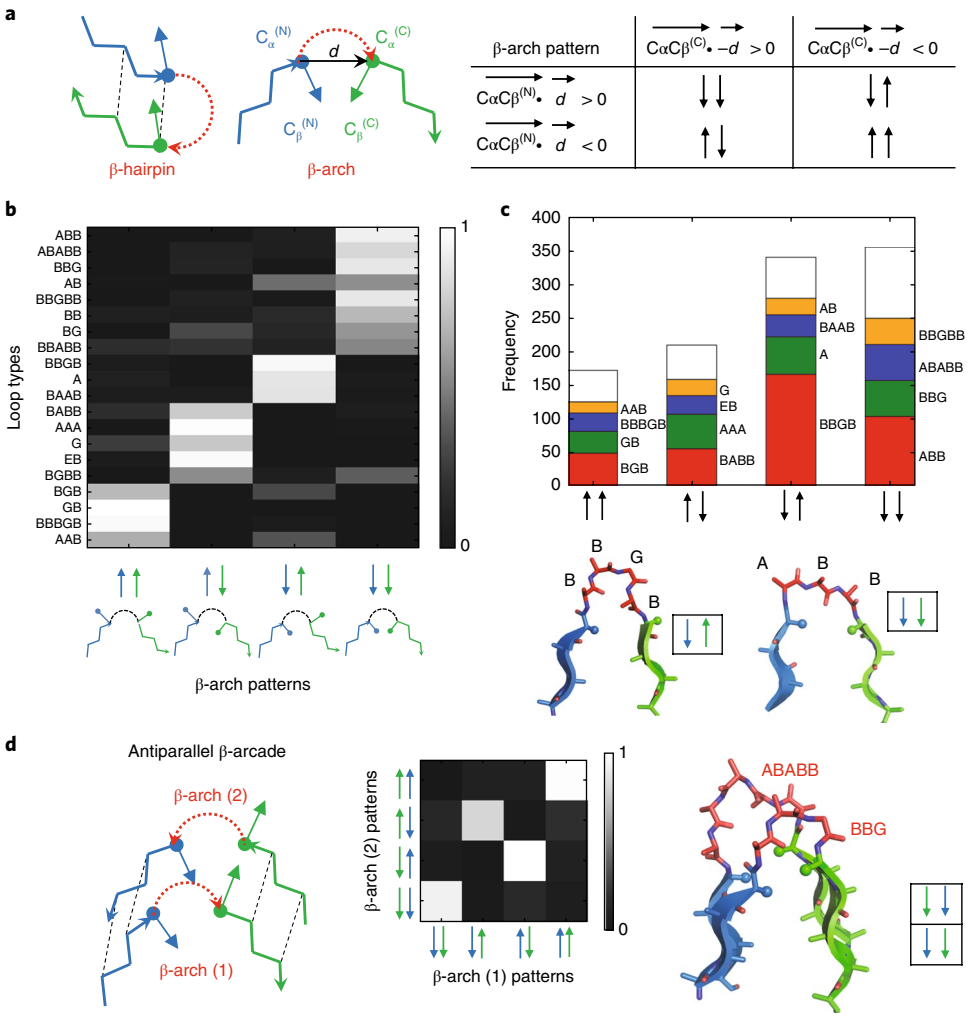

Fig. 1 | Constraints on β-arch geometry. a, Side chain directionality in the β-arch. Comparison between β-hairpin and β-arch (left); the CαCβ and *d* vectors used to define the orientation of the two adjacent side chains are indicated. The four possible side chain directionality patterns are on the right. **b**, Turn type dependence of β-arch side chain patterns. Loops on the *y* axis are described by their ABEGO torsion bins (Supplementary Fig. 1). Most of the loops adopt only one of the four possible side chain patterns. **c**, Frequency of the most common loops for each of the four β-arch side chain patterns. There are strong preferences: for example, BBGB is strongly associated with the ↓↑ pattern, whereas ABB is strongly associated with the ↓↓ pattern (bottom). Only loops with bending <120° (see Methods) and containing between 1 and 5 amino acids were considered in this analysis. **d**, β-arcades consist of two stacked β-arches with in-register strand pairing (left). Because strand pairs of the β-arcade are in register, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains that are adjacent to the second β-arch loop, and therefore not all loop pairs are allowed (middle). Example of a β-arcade formed by two common β-arches with compatible side chain patterns (right).

β-hairpins)[5], there was no correlation between β-arch loop length and side chain pattern. Instead, each loop ABEGO type, because of the way in which it twists and bends the polypeptide chain[16], is associated with a specific flanking residue side chain pattern (Fig. 1b). The most frequently observed turn types (between 1 and 5 amino acids) for each side chain pattern are listed in Fig. 1c; for example, ABB, BBGB, BABB and BGB are the most frequent loop types for the patterns ↓↓, ↓↑, ↑↓ and ↑↑, respectively.

The next level of non-local interaction complexity in all-β folds involves strand pairing (parallel or antiparallel) between two β-arches forming a β-arcade (Fig. 1d), a common structural motif in naturally occurring β-solenoids[15,17]. Because the β-arch loops are stacked in register, the side chains adjacent to one β-arch loop are likely to have the same orientation as the side chains adjacent to the second β-arch loop; analysis of naturally occurring β-arcades confirms that the side chain patterns of the two β-arch loops are indeed correlated (Fig. 1d, middle).

Jellyroll design principles. The double-stranded β-helix can be regarded as a long β-hairpin wrapped around an axis perpendicular to the direction of β-strands, with β-helical turns formed by the pairing between β-arcades (Fig. 2a). In the compact folded structure, two antiparallel β-sheets pack against each other in a sandwich-like arrangement, with the first strand paired to the last, and all β-strands are connected through β-arch loops except for the central β-hairpin. We aimed to design β-helices with three β-arcades forming two antiparallel four-stranded β-sheets, with the eight β-strands connected through six β-arches and one β-hairpin. The non-local character of the structure grows from the first β-arcade, which starts from the central β-hairpin, to the last one, where the N and C termini are paired.

The analysis from Fig. 1 leads to strong constraints on the construction of β-sheet backbone structures, as the side chain directionality patterns of the β-strands and loops are coupled in several ways. First, the directionality patterns of the loops preceding and following

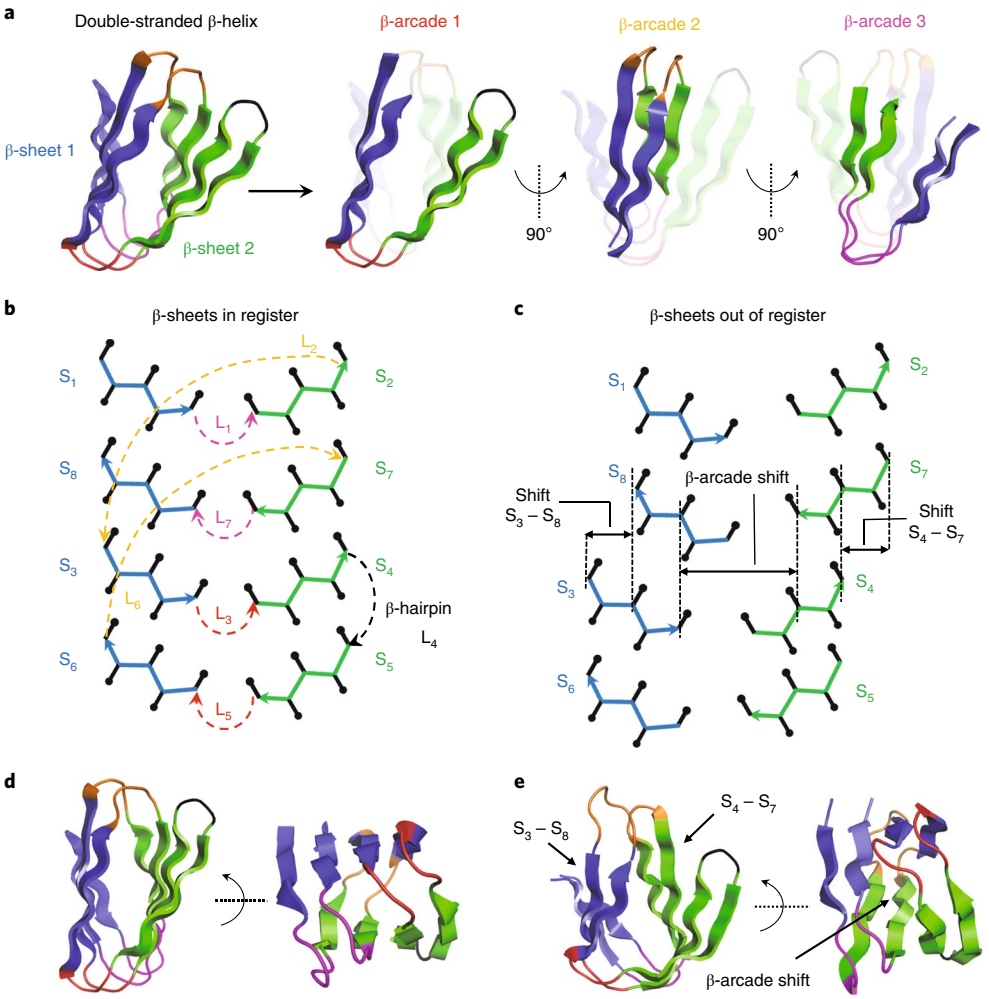

Fig. 2 | Double-stranded β-helix topology specification. a, The double-stranded β-helix fold consists of two four-stranded antiparallel β-sheets (in blue and green) with six β-arch and one β-hairpin connection. Pairs of β-arches forming the three β-arcades are highlighted (right). β-arch loops belonging to the same β-arcade are displayed with the same color throughout the figure (β-arcades 1, 2 and 3 in red, orange and magenta, respectively). **b**, Topology diagram of a designed double-stranded β-helix with all β-strand pairs in register. The Cα traces of the first and second β-sheets are colored in blue and green, respectively. Side chain Cβ positions oriented toward the inner and outer faces of the β-helix are represented with up and down black arrows with rounded tips, respectively. β-arch loops are colored as in **a**. **c**, Definition of β-arcade register shift varied during conformational sampling. The β-arcade register shift (between β-arcades 1 and 3) is determined by the register of β-strand pairs S_3/S_8 and S_4/S_7, and the lengths of β-strands S_3, S_4, S_8 and S_7 (see Methods). In this example, β-strand pairs S_3/S_8 and S_4/S_7 each have a two-residue register shift, resulting in an overall β-arcade register shift of four residues. Loops are omitted to facilitate visualization. **d**, Example of a design model with all β-strand pairs in register forming a sandwich-like structure. **e**, Example of a design model with register shifts between β-arcades 1 and 3 (magenta and red) forming a barrel-like structure.

each β-strand are coupled to the length of the strand (Fig. 2b): for example, a β-strand with an even number of residues that is preceded by a ↑↑ loop must be followed by a ↓↑ or a ↓↓ loop, but not a ↑↑ or ↑↓ loop, owing to the alternating pleating of β-strands. Second, because the β-arcades of the β-helix have paired β-strands and β-arch loops, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains adjacent to the second β-arch loop (Fig. 1d). Owing to the antiparallel orientation of the β-arcades, ↓↓ and ↑↑ loops are compatible with loops of the same type, but ↑↓ loops are only compatible with ↓↑ loops (Fig. 1d). Third, the twist and curvature of the two β-sheets of the β-helix are constrained by the hydrogen-bonding register between β-arcades 1 and 3 (herein called β-arcade register), and within β-strand pairs S_3/S_8 and S_4/S_7, as shown in Fig. 2c.

De novo design of protein structures. We constructed double-stranded β-helix protein backbones by Monte Carlo fragment assembly using blueprints (representations of the target protein topologies specifying the ordering, lengths and backbone torsion bins of secondary structure elements and loop connections[5]) in conjunction with backbone hydrogen-bonding constraints specifying all pairings between β-strands. We explored strand lengths between 5 and 7 residues and the most commonly observed β-arch loops between 3 and 5 residues (Fig. 1c). The central β-hairpin was designed with two-residue loops following the ββ rule[5]. The register shifts between pairs of β-strands from different β-arcades (1 and 3) were allowed to range from 0 to 2 and the β-arcade register shifts between 0 and 4; strand pairs within the same β-arcade were kept in register. A total of 3,673 combinations were enumerated, of which 1,853 had mutually compatible strand lengths and loop types consistent with the constraints summarized in the previous paragraph. For each of these internally consistent blueprints, we used Rosetta to build thousands of protein backbones. The resulting ensemble of backbone structures has considerable structural diversity; those

with all strands in register had narrow, sandwich-like structures (Fig. 2d), while those with large register shifts had wider, barrel-like structures (Fig. 2e).

For each generated backbone, we carried out flexible-sequence design calculations[18,19] to identify low-energy amino acid identities and side chain conformations providing close complementary packing, side chain–backbone hydrogen bonding in β-arch loops (to pre-organize their conformation and facilitate folding), and high sequence–structure compatibility. We favored inward-pointing charged or polar amino acids at the four edge strands to minimize aggregation propensity[20]. Loop sequences were designed with consensus profiles obtained from fragments with the same backbone ABEGO torsion bins[21]. Because the very large size of the space sampled by our design procedure limits convergence on optimal sequence–structure pairs, we carried out a second round of calculations starting from the blueprints yielding the lowest-energy designs, intensifying sampling at both the backbone and sequence level. For a subset of designs, we introduced disulfide bonds between paired β-strand positions with high sequence separation (for example, between the first and last β-strands) and optimal orientation (see Methods): disulfide bonds distant in primary sequence decrease the entropy of the unfolded state and therefore enhance the thermodynamic stability of the native state. To assess compatibility of the top-ranked designed sequences with their structures, we characterized their folding energy landscape with biased forward folding simulations[21], and those with substantial near-native sampling were subsequently assessed by Rosetta ab initio structure prediction calculations[22,23]. Designs with funnel-shaped energy landscapes (where the designed structure is at the global energy minimum and has a substantial energy gap with respect to alternative conformations) were selected for experimental characterization. Ab initio structure prediction of natural β-sheet proteins tends to oversample local contacts[24,25] (that is, favors β-hairpins over β-arches), but we succeeded in designing sequences with the β-arches sufficiently strongly encoded that they folded in silico to near the designed target structure.

Experimental characterization. For experimental characterization, we chose 19 designs with funnel-shaped energy landscapes ranging between 70 and 94 amino acids (Supplementary Table 1). BLAST searches[26,27] indicated that the designed sequences had little or no similarity with native proteins (lowest Expect (*E*) values ranging from 0.003 to >10; Supplementary Table 2). Synthetic genes encoding the designs (design names are BH_n, where BH stands for β-helix, n stands for the design number and the _ss suffix is used if disulfide bonds are present) were obtained; the proteins were expressed in *Escherichia coli* and purified by affinity chromatography. Sixteen of the designs expressed well and were soluble, and two (BH_10 and BH_11) were monomeric (Supplementary Fig. 2) by size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) (most of the non-monomeric designs were either dimers or soluble aggregates). Both monomeric designs had far-ultraviolet circular dichroism spectra (CD) at 25 °C characteristic of β proteins, a melting temperature (T_m) above 95 °C, and well-ordered structures according to 2D ^{1}H-^{15}N heteronuclear single-quantum coherence (HSQC) spectra (Fig. 3a–c and Supplementary Fig. 3). For both designs, the number of NMR peaks matched the number of expected amide resonances based on the protein sequence, but the higher stability of BH_10 under the conditions of the NMR experiments made it a better candidate for NMR structure determination.

The two monomeric designs with well-ordered structures were among those with better-packed cores and a larger proportion of β-arch loops containing prolines and with backbone polar atoms making hydrogen bonds (Supplementary Table 3). β-arch loops that are structurally preorganized with the polar groups making internal hydrogen bonding likely favor folding to the correct topology and contribute to stability by compensating for the loss of interactions with water of polar groups in the side chains and backbone. These interactions likely disfavor the competing local strand-pairing arrangement in which the two strands form a β-hairpin; this is a very common pathology in ab initio structure prediction[25]. For the most stable dimeric design (BH_6), we introduced disulfide bonds to stabilize protein regions having contacts with large sequence separation (for example, between the N- and C-terminal strands), but this did not succeed in yielding stable monomers. Addition of an α-helix to the C terminus (one of the two extremes of the β-helix), as a capping domain protecting the strand edges from intermolecular pairing, also failed to yield stable monomers, even in combination with disulfide bonds. This suggests that the sequence of the core β-sheet must strongly encode its structure independently of disulfide bonds or protecting domains aimed at increasing stability.

NMR structure of a de novo–designed β-helix. We succeeded in solving the structure of BH_10 by 4D NMR spectroscopy (Fig. 3d, Table 1 and Supplementary Fig. 4), using the 4D-CHAINS/AutoNOE-Rosetta automated pipeline for resonance assignments and structure calculation[28], and found it to be in very close agreement with the computational model (Cα r.m.s. deviation (r.m.s.d.) of 0.84 Å, averaged over 10 NMR models). The overall topology is accurately recapitulated, including all strand pairings, register shifts and loop connections, as supported by 132 long-range nuclear Overhauser effects (NOEs) between backbone amide and side chain protons (Supplementary Fig. 5). The designed aliphatic and aromatic side chain packing in the protein core, as well as salt bridge interactions across the two β-sheet surfaces, was also accurately reproduced; three salt bridges between the two paired β-arcades and one within the third β-arcade are well supported by the observed NOEs (Supplementary Fig. 6). The agreement in both the backbone conformation and hydrogen-bonding interactions of the loops forming the three β-arcades is remarkable given that these elements are the most flexible parts of the structure and therefore difficult to design owing to sampling bottlenecks. The β-arcades were designed with pairs of β-arch loops that interact via backbone-backbone hydrogen bonds (owing to the complementarity between their backbone conformations) stabilizing loop pairing and avoiding burial of polar backbone atoms (see Supplementary Fig. 7 for the BH_10 loop sequences and side chain patterns). For example, β-arcade 1 is formed by BBG and ABB loops, and the buried backbone NH group of the G position in the former makes a hydrogen bond with the buried backbone C=O of the neighboring loop (Fig. 3e). The other two β-arcades were designed with one β-arch loop containing buried and fully hydrogen-bonded asparagines (four hydrogen bonds in total) that stabilize both loop pairing and the local β-arch conformation (of ABABB loops). By design, the asparagine side chain geometry was further stabilized with hydrophobic stacking interactions from the two β-arch loops of the same arcade. The high degree of convergence of the designed rotamer in the NMR ensemble illustrates the high structural preorganization of this particular motif (Fig. 3f).

The amino acid sequence of BH_10 is unrelated to any sequence in the NCBI nr database (BLAST found one hit with insignificant sequence similarity; *E* value 6.3). We searched the Protein Data Bank (PDB) for similarities in structure (using the Dali server[29] with the lowest-energy NMR model as the query structure) or sequence (with HHpred[30] for sensitive profile-based sequence search), and identified matches similar in fold but containing additional and irregular secondary structures and longer loops. These matches are all homodimers with sheet-to-sheet interface packing (Supplementary Fig. 8) or domains integrated in larger structures, in sharp contrast to the BH_10 monomer.

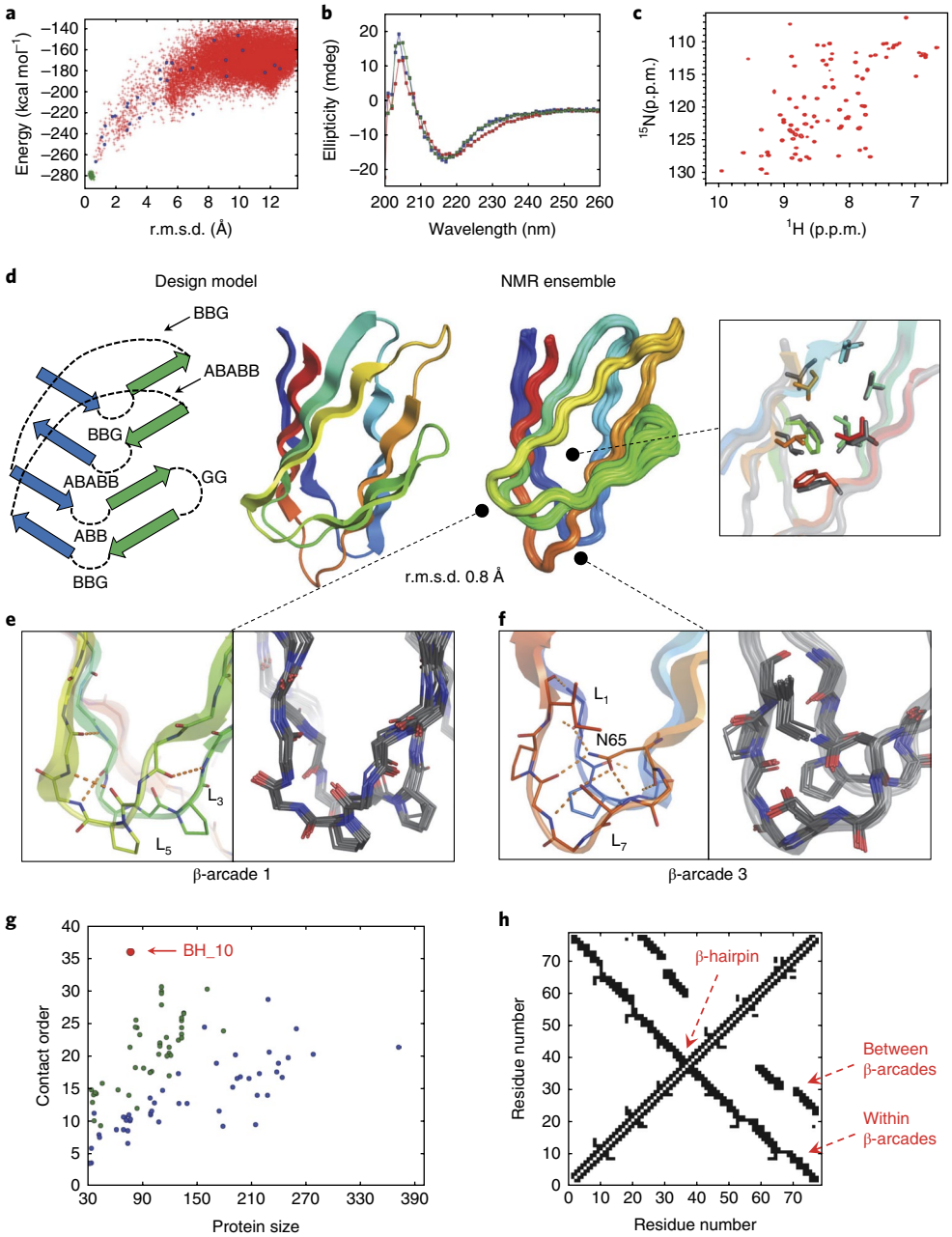

Fig. 3 | The NMR structure of BH_10 is nearly identical to the design model. a, Calculated BH_10 folding energy landscape. Each dot represents the lowest-energy structure obtained from ab initio folding trajectories starting from an extended chain (red dots), biased forward folding trajectories (blue dots) or local relaxation of the designed structure (green dots); the *x* axis is the Cα r.m.s.d. from the designed model and the *y* axis is the Rosetta all-atom energy. **b**, Far-ultraviolet CD spectra (blue, 25 °C; red, 95 °C; green, 25 °C after cooling). **c**, ^{1}H-^{15}N HSQC spectra obtained at 37 °C at a ^{1}H field of 800 MHz. **d**, NMR structure in comparison with the design model. Comparison of core side chain rotamers (NMR structure in gray and design in rainbow) (inset). Topology diagram showing the ABEGO torsion bins of all loop connections (left). Atomic coordinates for the design model are in Supplementary Dataset 1. **e**, Backbone hydrogen bonding of β-arcade 1 is well preserved across the NMR ensemble. **f**, Side chain interactions of N65 with backbone and side chains form a hydrogen-bonded network in β-arcade 3 that is well recapitulated in the NMR ensemble. **g**, Contact order of de novo–designed protein domains confirmed by high-resolution structure determination; all-α (blue), αβ (green) and all-β (red). The BH_10 design stands out with a contact order of 35.8 for a chain length of 78 residues. The domains are listed in Supplementary Tables 4 and 5. **h**, Contact map illustrating the large sequence separation of the contacts present in the BH_10 topology.

Contact order and sequence determinants of the BH_10 fold. The non-local character of BH_10 is of particular note: a large fraction of the contacting residues are distant along the linear sequence, with extensive strand pairing between the N- and C-terminal β-strands. The contact order of the structure (that is, the average separation along the linear sequence of residues in contact in the 3D structure) is higher than that for any previous single-domain protein designed de novo (Fig. 3g,h). Proteins with high contact order fold more slowly than those with low contact order as there is a greater loss in chain entropy for forming the first native interactions, and they tend to form long-lived non-native structures that can oligomerize or aggregate[31]. We have overcome the challenges in designing

Table 1 | NMR and refinement statistics for BH_10

	BH_10 (PDB 6E5C)
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	659
Intraresidue	272
Inter-residue	387
Sequential ($\|i-j\|=1$)	222
Medium range ($2\leq\|i-j\|\leq4$)	33
Long range ($\|i-j\|\geq5$)	132
Intermolecular	0
Hydrogen bonds	0
Total dihedral angle restraints	156
φ	78
ψ	78
Structure statistics	
Violations (mean ± s.d.)	
Distance constraints (Å)[a]	0.30 ± 0.46
Dihedral angle constraints (°)[b]	9.30 ± 2.49
Max. dihedral angle violation (°)[b]	47.59
Max. distance constraint violation (Å)[a]	1.32
Deviations from idealized geometry	
Bond lengths (Å)	0.00 ± 0.00
Bond angles (°)	0.00 ± 0.00
Impropers (°)	0.00 ± 0.00
Average pairwise r.m.s.d. (Å)[c]	
Heavy	0.61 ± 0.13
Backbone	0.51 ± 0.11

[a]Distance constraint violations in the structural ensemble were calculated using a 7-Å universal upper distance bound for the NOE restraints assigned by AutoNOE-Rosetta. [b]Dihedral angle restraints were derived from TALOS-N. The violations were calculated for the core secondary structural regions of the ten lowest-energy models using a 15° cutoff beyond TALOS-N-predicted dihedral angles. [c]Pairwise r.m.s.d. was calculated among ten refined models for a core secondary structural region defined by residues 2–8, 11–18, 21–28, 32–36, 39–43, 46–53, 59–65 and 71–75.

non-local structures by focusing on backbones lacking internal strain and having maximal internal coherence, and programming β-strand orientation with highly structured loops.

One of the challenges in achieving high contact order through β-arches is to disfavor formation of more sequence-local β-hairpins. To evaluate in silico how each of our design features contributes to favoring β-arches over β-hairpins, we generated folding energy landscapes for a series of mutants of BH_10 in which loop hydrogen bonding, side chain packing of loop neighbors and loop local geometry were disrupted one at a time. For all conformations generated, we classified all the β-strand connections as β-arch or β-hairpin depending on strand-pairing formation, and calculated the overall frequency of β-hairpin formation for each pair of consecutive β-strands. As shown in Supplementary Fig. 9, disruption of packing within or between β-arch loops, removal of side chain–backbone hydrogen-bonding interactions and reducing loop geometry encoding by eliminating prolines all increase sampling of competing β-hairpin conformations, and thus substantially decrease sampling of β-arches and the target designed structure.

Discussion

The design of all-β globular proteins from first principles has remained elusive for two decades of protein design research. We have successfully designed a double-stranded β-helix de novo, as confirmed by the NMR structure of the design BH_10, based on a series of rules describing the geometry of β-arch loops and their interactions in more complex β-arcades. Our work also achieves two related milestones: the first accurate design of an all-β globular protein with exposed β-sheet edges, and the most non-local structure yet designed from scratch. Comparison of successful and failed designs suggests that folding and stabilization of the monomeric structure (and implicitly disfavoring competing topologies with more local strand pairings) is bolstered by loops containing side chain–backbone and backbone-backbone hydrogen bonds together with well-packed mixed aliphatic/aromatic side chains in the protein core, inward-pointing polar amino acids at strand edges and salt bridges between paired strands. Previous design studies on β-propellers[11] or parallel β-helices[12] have used naturally occurring backbone structures and consensus sequence information on the target fold families; this approach, while powerful, sheds less light on the key principles underlying β-sheet structure construction and does not allow the programming of new backbone geometries. The β-helix fold designed here is well suited for incorporating metal, ligand-binding and active sites, as illustrated by the broad functional diversity of cupin protein domains, which are the closest naturally occurring structural analogs. With the basic design principles now understood, our de novo design strategy should enable the construction of a wide range of β-helix structures tailored to a broad range of target ligands.

Initial advances in protein design were from algorithms that allowed rapid identification of a very-low-energy sequence for a given backbone structure. In recent years, progress has come from the realization that the requirements of burying hydrophobic residues in a core away from solvent while avoiding the burial of backbone polar groups without compensating hydrogen bonds, together with torsional restrictions on the peptide backbone, considerably constrain overall globular protein backbone geometry, particularly for β-sheet-containing proteins: it is much harder than originally expected to construct new backbones that have these properties. The de novo design of β-sheet-containing proteins advanced considerably following the elucidation of β-sheet design principles for construction of backbones meeting the above constraints while having desired geometries: for example, principles for controlling the chirality of β-hairpins[5], reducing strain in β-strands with glycine kinks[10], and combining β-bulges and register shifts to curve β-sheets[21]. The design rules described here are a considerable further advance as they provide control over β-arch connections between distinct β-sheets, and should enable the design of a broad range of β-protein families beyond the β-barrel and β-helix, with considerable medical and biotechnological potential; for example, the immunoglobulin fold widely utilized for binding and loop scaffolding in nature is topologically very similar to the double-stranded β-helices designed here, with a larger proportion of β-hairpins over β-arches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41594-018-0141-6.

Received: 16 July 2018; Accepted: 11 September 2018;
Published online: 29 October 2018

References

1. Kortemme, T., Ramírez-Alvarado, M. & Serrano, L. Design of a 20-amino acid, three-stranded β-sheet protein. *Science* **281**, 253–256 (1998).
2. Searle, M. S. & Ciani, B. Design of β-sheet systems for understanding the thermodynamics and kinetics of protein folding. *Curr. Opin. Struct. Biol.* **14**, 458–464 (2004).

3. Hughes, R. M. & Waters, M. L. Model systems for β-hairpins and β-sheets. *Curr. Opin. Struct. Biol.* **16**, 514–524 (2006).
4. Marcos, E. & Adriano-Silva, D. Essentials of de novo protein design: methods and applications. *WIREs Comput. Mol. Sci.* **8**, e1374 (2018).
5. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
6. Hecht, M. H. De novo design of β-sheet proteins. *Proc. Natl Acad. Sci. USA* **91**, 8729–8730 (1994).
7. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
8. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S. & Richardson, D. C. Betadoublet: de novo design, synthesis, and characterization of a β-sandwich protein. *Proc. Natl Acad. Sci. USA* **91**, 8747–8751 (1994).
9. Nanda, V. et al. De novo design of a redox-active minimal rubredoxin mimic. *J. Am. Chem. Soc.* **127**, 5804–5805 (2005).
10. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
11. 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).
12. MacDonald, J. T. Synthetic β-solenoid proteins with the fragment-free computational design of a β-hairpin extension. *Proc. Natl Acad. Sci. USA* **113**, 10346–10351 (2016).
13. Ottesen, J. J. & Imperiali, B. Design of a discretely folded mini-protein motif with predominantly β-structure. *Nat. Struct. Biol.* **8**, 535–539 (2001).
14. Hu, X., Wang, H., Ke, H. & Kuhlman, B. Computer-based redesign of β sandwich protein suggests that extensive negative design is not required for de novo β sheet design. *Structure* **16**, 1799–1805 (2008).
15. Hennetin, J., Jullian, B., Steven, A. C. & Kajava, A. V. Standard conformations of beta-arches in β-solenoid proteins. *J. Mol. Biol.* **358**, 1094–1105 (2006).
16. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–E5485 (2015).
17. Kajava, A. V., Baxa, U. & Steven, A. C. β arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils. *FASEB J.* **24**, 1311–1319 (2010).
18. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97**, 10383–10388 (2000).
19. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
20. Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
21. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
22. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
23. Bradley, P. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
24. Kuhn, M., Meiler, J. & Baker, D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins* **54**, 282–288 (2004).
25. Bradley, P. & Baker, D. Improved β-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* **65**, 922–929 (2006).
26. 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).
27. Camacho, C. et al. BLAST: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
28. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
29. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016).
30. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
31. Clark, P. Protein folding in the cell: reshaping the folding funnel. *Trends Biochem. Sci.* **29**, 527–534 (2004).

Acknowledgements

We thank S. Rettie for mass spectrometry assistance and the rest of the Baker laboratory members for discussion. We acknowledge computing resources provided by the Hyak supercomputer system funded by the STF at the University of Washington, and Rosetta@Home volunteers in ab initio structure prediction calculations. Work carried out at the Baker laboratory was supported by the Howard Hughes Medical Institute, Open Philanthropy, and the Defense Threat Reduction Agency. E.M. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2011-IOF 298976). G.O. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2012-IOF 332094). This research was financially supported by the Ministry of Education, Youths and Sports of the Czech Republic within the CEITEC 2020 (LQ1601) project, the Grant Agency of Masaryk University (to K.T.), an R35 Outstanding Investigator Award to N.G.S. through NIGMS (1R35GM125034-01), and the Office of the Director, NIH, under High End Instrumentation Grant S10OD018455, which funded the 800-MHz NMR spectrometer at UCSC. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the Ministry of Economy, Industry and Competitiveness (government of Spain).

Author contributions

E.M. designed the research, carried out the loop structural analysis, set up the design method and performed design calculations. T.M.C. carried out design calculations, protein expression, purification and CD experiments. A.C.M. collected 4D NMR data. T.E. performed 4D-CHAINS analysis. S.N. carried out AutoNOE-Rosetta calculations. L.C. expressed isotopically labeled proteins and performed SEC-MALS analysis. L.G.N. designed the research and carried out design calculations. A.D. and G.O. helped in protein expression and characterization. K.T. and N.G.S. supervised NMR structure determination. D.B. designed and supervised the research. E.M., and D.B. prepared the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-018-0141-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.M. or D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Loop analysis. Loop connections between β-strands were collected from a non-redundant database of PDB structures obtained from the PISCES server[32] with sequence identity <30% and resolution ≤2 Å. We discarded those loops connecting β-strands with hydrogen-bonded pairing (β-hairpins), and the remaining 5,061 β-arch loops were subsequently analyzed. The ABEGO torsion bins of each residue position were assigned based on the definition shown in Supplementary Fig. 1, and the side chain directionality pattern of neighboring residues was defined according to Fig. 1a. The secondary structure of all residue positions was assigned with DSSP[33], and the last β-strand residue preceding and the first β-strand residue following the β-arch loop were chosen as the critical neighboring residues determining the side chain pattern of the loop. The loop bending was defined as the angle between the loop center of mass and the two strand positions adjacent to the loop. Those loops with bending angles larger than 120° were discarded from the analysis to correctly identify those loops producing a substantial change in the direction of the two connected β-strands. The loop dataset is available in Supplementary Dataset 2.

Backbone generation. We used the Blueprint Builder mover[5] of RosettaScripts[34] to build protein backbones by Monte Carlo fragment assembly using nine- and three-residue fragments compatible with the target secondary structure and torsion bins (ABEGO), as specified in the blueprints of every target topology. We used a polyvaline centroid representation of the protein and a scoring function accounting for backbone hydrogen bonding, van der Waals interactions (namely to avoid steric clashes), planarity of the peptide bond (omega score term), and compactness of structures (radius of gyration). Thousands of independent folding trajectories were performed and subsequently filtered. Owing to the non-local character of β-sheet contacts, we used distance and angle constraints to favor the correct hydrogen-bonded pairing between β-strand main chain atoms. For every target topology, we automatically set all pairs of residues involved in β-strand pairing to generate all constraints for backbone building. Protein backbones were filtered based on their match with the blueprint specifications (secondary structure, torsion bins and strand pairing), and subsequently ranked based on backbone hydrogen-bonding energy (lr_hb score term) and the total energy obtained from one round of all-atom flexible-sequence design (see next section).

Flexible-sequence design. Generated protein backbones were subjected to flexible-sequence design calculations with RosettaDesign[18,19] using the Rosetta all-atom energy function Talaris2014[35] to favor amino acid identities and side-chain conformations with low energy and tight packing. We performed cycles of fixed backbone design followed by backbone relaxation using the FastDesign mover[36] of RosettaScripts[34]. Designed sequences were filtered based on total energy, side-chain packing (measured with RosettaHoles[37], packstat and core side-chain average degree[21]), side chain–backbone hydrogen bond energy, and secondary structure prediction (match between the designed secondary structure and that predicted by PSPIRED[38] based on the designed sequence). Amino acid identities were restricted based on the solvent accessibility of protein positions, ensuring that hydrophobic amino acids were located in the core and polar ones in the surface. Further restrictions were imposed to improve sequence-structure compatibility in loop regions. Sequence profiles were obtained for naturally occurring loops with the same ABEGO string sequence, as previously[21].

For those blueprints that yielded the lowest energy designs we performed a second round with ten times more backbone samples. Backbones generated in this second round were subjected to more exhaustive sequence design by running multiple Generic Monte Carlo trajectories optimizing total energy and side chain average degree simultaneously, and then applied all filters described above.

Design of disulfide bonds and helix capping domain. We used the Disulfidize mover of RosettaScripts[34] to identify pairs of residue positions able to form disulfide bonds with a good scoring geometry. We searched for disulfide bonds between residues distant in primary sequence and with a disulfide score <−1.0. We designed a C-terminal helix capping domain (followed with a β-strand pairing with the first β-strand) using the backbone-generation protocol described above but starting from design BH_6. The structure of BH_6 was kept fixed during fragment assembly and the C-terminal domain was generated. Then, sequence design was performed for the C-terminal domain and those neighboring residues within 10 Å.

Sequence–structure compatibility. For assessing the local compatibility between designed sequences and structures we picked 200 naturally occurring fragments (9- and 3-mers) with sequences similar to the design, and evaluated the structural similarity (by r.m.s.d.) between the ensemble of picked fragments and the local designed structure. Those with overall low r.m.s.d. fragments, and therefore with high fragment quality, were subsequently assessed by Rosetta folding simulations using the Rosetta energy function 'ref2015'[39]. First, biased forward folding simulations[21] (using the three lowest r.m.s.d. fragments and 40 folding trajectories) were used to quickly identify those designs more likely to have funnel-shaped energy landscapes. Those designs achieving near-native sampling (r.m.s.d. to target structure below 1.5 Å) were then assessed by standard Rosetta ab initio structure prediction[22,23].

To evaluate the amount of β-hairpin sampling in each loop connection during ab initio structure prediction, we first detected all strand pairings formed in each generated decoy and then mapped the residues involved in those strand pairings to the secondary structure elements of the designed structure. After secondary structure mapping, pairings between strands consecutive in the sequence were counted as β-hairpins. The total count of β-hairpins sampled in each loop over the total number of generated decoys is a relative quantity of hairpin sampling allowed us to compare the β-hairpin propensity of different loops and mutants (see Supplementary Fig. 9).

Contact order. To evaluate the non-local character of protein structures we computed 'contact order' as the average sequence separation between pairs of Cα atoms within a distance of 8 Å and with a sequence separation of three residues at least.

Protein expression and purification. Genes encoding the designed sequences were obtained from Genscript and cloned into the pET-28b+ (with N-terminal 6× His tag and a thrombin cleavage site) expression vectors. Plasmids were transformed into *Escherichia coli* BL21 Star (DE3) competent cells, and starter cultures were grown at 37 °C in LB medium overnight with kanamycin. Overnight cultures were used to inoculate 500 ml LB medium supplemented with antibiotic, and cells were grown at 37 °C and 225 r.p.m. until an optical density (OD_{600}) of 0.5–0.7 was reached. Protein expression was induced with 1 mM IPTG at 18 °C and, after overnight expression, cells were collected by centrifugation (at 4 °C and 4,400 r.p.m. for 10 min) and resuspended in 25 ml of lysis buffer (20 mM imidazole and PBS). Resuspended cells were lysed in the presence of lysozyme, DNase and protease inhibitors. Lysates were centrifuged at 4 °C and 18,000 r.p.m. for 30 min, and the supernatant was loaded onto a nickel affinity gravity column pre-equilibrated in lysis buffer. The column was washed with three column volumes of PBS + 30 mM imidazole and the purified protein was eluted with three column volumes of PBS + 250 mM imidazole. The eluted protein solution was dialyzed against PBS buffer overnight. The expression of purified proteins was assessed by SDS–PAGE and mass spectrometry, and protein concentrations were determined from the absorbance at 280 nm measured on a NanoDrop spectrophotometer (Thermo Scientific) with extinction coefficients predicted from the amino acid sequences using the ProtParam tool (https://web.expasy.org/protparam/). Proteins were further purified by fast protein liquid chromatography size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column.

Circular dichroism. Far-ultraviolet CD measurements were carried out with the AVIV 420 spectrometer. Wavelength scans were measured from 260 to 195 nm at temperatures between 25 and 95 °C, using a 1-mm path-length cuvette. Protein samples were prepared in PBS buffer (pH 7.4) at a concentration of 0.2–0.4 mg ml^{-1}.

Size-exclusion chromatography combined with multiple-angle light scattering. SEC-MALS experiments were performed using a Superdex 75 10/300 GL (GE Healthcare) column combined with a miniDAWN TREOS multi-angle static light scattering detector and an Optilab T-rEX refractometer (Wyatt Technology). One-hundred-microliter protein samples of 1–3 mg ml^{-1} were injected onto the column equilibrated with PBS (pH 7.4) or TBS (pH 8.0) buffer at a flow rate of 0.5 ml min^{-1}. The collected data were analyzed with ASTRA software (Wyatt Technology) to estimate the molecular weight of the eluted species.

Protein expression of isotopically labeled proteins for NMR structure determination. Plasmids were transformed using standard heat-shock transformation into the Lemo21 expression strain of *E. coli* (NEB) and plated onto a minimal M9 medium containing glucose and kanamycin to maintain tight control over expression. A single colony was selected, inoculated into 50 ml of LB containing 50 μg ml^{-1} of kanamycin and grown at 37 °C with shaking overnight. After approximately 18 h, the 50-ml starter culture was removed and 25 ml was used to inoculate 500 ml of Terrific Broth (TB) containing 50 μg ml^{-1} kanamycin and mixed mineral salts[40]. The TB culture was grown at 37 °C with shaking at 250 r.p.m. until the OD_{600} reached a value of 1.0. Then, the culture was removed and the cells were pelleted by centrifugation at 4,000 r.p.m. for 15 min. The TB broth was removed and the pelleted cells were resuspended gently with 50 ml of 20 mM $NaPO_4$, 150 mM NaCl, pH 7.5. The resuspended cells were transferred into minimal labeling media, containing ^{15}N-labeled ammonium chloride at 50 mM and ^{13}C-labeled glucose to 0.25% (w/v), as well as trace metals, 25 mM Na_2HPO_4, 25 mM KH_2PO_4 and 5 mM Na_2SO_4. The culture was returned to 37 °C at 250 r.p.m. for 1 h in order to replace unlabeled nitrogen and carbon with labeled nitrogen and carbon. After 1 h, IPTG was added to 1 mM, the temperature was reduced to 25 °C and the culture was allowed to express overnight. The following morning, the culture was removed and the cells were pelleted by centrifugation at 4,000 r.p.m. for 15 min. The cells were resuspended with 40 ml of lysis buffer (20 mM Tris, 250 mM NaCl, 0.25% CHAPS, pH 8) and lysed with a Microfluidics M110P Microfluidizer at 18,000 p.s.i. The lysed cells were clarified using centrifugation at 24,000*g* for 30 min. The labeled protein in the soluble fraction was purified using immobilized metal affinity chromatography using standard methods (QIagen Ni-NTA resin). The purified protein was then concentrated to 2 ml and purified by FPLC

size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column into 20 mM $NaPO_4$, 150 mM NaCl, pH 7.5. The efficiency of labeling was confirmed using mass spectrometry.

Heteronuclear single-quantum coherence spectra. The designed BH_10 (0.81 mM) and BH_11 (0.64 mM) were exhaustively buffer exchanged into NMR buffer (50 mM NaCl, 20 mM sodium phosphate, pH 6.5, 0.01% (v/v) NaN_3, 4 mM EDTA and 1 U Roche protease inhibitor cocktail) in 95% H_2O/5% D_2O. 2D ^{1}H-^{15}N HSQC experiments were acquired at 37 °C with four scans, acquisition times of 72 ms (^{15}N) in the indirect dimension and recycle delay of 2 s.

Chemical shift assignment. For chemical shift assignment of BH_10, a set of two non-uniformly sampled (NUS) 4D NMR experiments, a 4D HC(CC-TOCSY(CO))NH and 4D ^{13}C-^{15}N edited HMQC-NOESY-HSQC, were acquired at 800 MHz at 37 °C as previously described[28]. For the 4D HC(CC-TOCSY(CO))NH experiment, spectrum widths were set to 12,500 (acquisition dimension) × 2,100 (^{15}N) × 8,000 (^{13}C) × 7,300 (^{1}H) Hz and acquisition times in the indirect dimensions of 60 ms (^{15}N), 8 ms (^{13}C) and 8 ms (^{1}H) using 16 scans and a recycle delay of 1 s. Spectra were acquired with 2,000 hypercomplex NUS points distributed over the indirectly detected dimensions (0.38% sparsity). For the 4D ^{13}C-^{15}N edited HMQC-NOESY-HSQC, spectrum widths were set to 12,500 (acquisition dimension) × 1,000 (^{15}N) × 8,000 (^{13}C) × 10,000 (^{1}H) Hz, respectively, and acquisition times in the indirect dimensions of 38 ms (^{15}N), 10 ms (^{13}C) and 20 ms (^{1}H) using 8 scans, a recycle delay of 1 s and a NOESY mixing time of 120 ms. Spectra were acquired with 4,000 hypercomplex NUS points distributed over the indirectly detected dimensions (0.32% sparsity). 4D NUS spectra were processed in NMRPipe[41] using SMILE reconstruction[42] and analyzed using NMRFAM-SPARKY[43]. For every ^{1}H-^{15}N HSQC peak, the corresponding planes in 4D-HCNH TOCSY and 4D-HCNH NOESY spectra were inspected and peaks were picked manually. The 4D peak lists were used as input for the 4D-CHAINS algorithm[28] to obtain sequence-specific resonance assignments of backbone and side chain atoms automatically. The overall assignment completeness reached 92%. No aromatic resonances were assigned. 4D-CHAINS assignments together with the 4D-HCNH NOESY peak list were used in AutoNOE-Rosetta for structure determination.

NOE assignment and structure determination using AutoNOE-Rosetta. To determine the structural models of the BH_10 target protein, we used CS-Rosetta[44] that provides AutoNOE-Rosetta[45] and RASREC-Rosetta[46] protocols. AutoNOE-Rosetta is an iterative NOE assignment method that utilizes RASREC-Rosetta to model protein structures de novo. These methods make use of valuable information contained within NMR chemical shifts about secondary and tertiary structures, and dynamics of proteins, to model targets of interest accurately[44,47]. The primary aim of AutoNOE-Rosetta is to label proton atoms to the unassigned NOESY cross-peaks by mapping their resonance frequencies to the assigned chemical shift frequencies. The resulting assignments can be utilized to create NOE-based distance restraints that aid the structure calculation process. The method begins by creating an initial mapping between the assigned chemical shift list and the unassigned NOESY cross-peak list. This mapping produces ambiguous assignments due to possible noise in the NOESY spectra[48,49]. These assignments undergo evaluation and filtering. The evaluation criteria rely on the symmetry of cross-peaks, chemical shift compatibility, intermediate structural model compatibility (in the subsequent stages of the protocol) and the participation of any NOE in a network of NOEs (network anchoring)[50]. The cross-peaks are eliminated if they lie along the diagonal in the NOESY spectra or their contribution to some of the evaluation criteria (such as network anchor score) is insignificant. The intensities of high-scoring NOE peaks are calibrated to produce distance restraints. These restraints are used to calculate structures within the highly parallel RASREC-Rosetta, which performs fragment assembly[44] using Monte Carlo methods and additional optimized algorithms[46]. This process of assigning NOEs and calculating structures is carried out iteratively across eight distinct stages. The final stage retains well-converged structural models alongside generated NOE restraints used for their calculation.

The process of setting up AutoNOE-Rosetta calculations is highly automated and accessible via a Python interface within the toolbox available at the CS-Rosetta website (https://csrosetta.chemistry.ucsc.edu). Before setting up NOE assignment and structure calculation runs, (1) NMR chemical shifts and target sequences are used to predict secondary structure (rigid regions) and flexible end regions from TALOS-N[51], (2) the flexible end regions are trimmed from sequence and chemical shift files because they cause deterioration of the performance of structure determination methods by inducing large numbers of degrees of freedom, and (3) the predicted secondary structure, together with trimmed chemical shift files, is used to select 200 structural fragments of amino acid lengths of three and nine, for each position in the target sequence. On completion of the previous steps, AutoNOE-Rosetta calculations are set up with target sequence, structural fragments, chemical shifts and unassigned NOESY cross-peak lists. For the BH_10 target protein, we obtained NMR chemical shifts from 4D-CHAINS[28] and additionally utilized amide to aliphatic (HCNH) unassigned NOESY cross-peak lists. Thereafter, we performed four rounds of AutoNOE-Rosetta calculations, where each round was supplied with a different restraint weight (standard restraint weights of 5, 10, 25 and 50 were used). All the calculation runs were evaluated using a function that assesses the all-atom energies[39] of the structural models and their convergence. After selecting the best-scoring restraint weight run, the ten models that exhibited the lowest energy within this run were selected. Commands to set up the calculations were used exactly as provided in the Supplementary Methods of a previous work[28]. Molprobity[52] was used to compute Ramachandran statistics for the ten lowest-energy structural models (100% of residues in favored regions of Ramachandran space, and 99% in favored regions) and deviations from the ideal geometry (Table 1).

Salt bridges. We used ESBRI[53] to predict salt bridges in the ten lowest-energy structural models. Out of 19 salt bridges predicted using ESBRI, AutoNOE-Rosetta recovers four salt bridges in the form of NOE contacts on the surface of the BH_10 protein. To identify salt bridges, we examined the NOE restraints assigned by AutoNOE-Rosetta between negatively charged glutamic acid or aspartic acid and positively charged arginine or lysine. We further filtered the restraints based on the upper distance bound of 4 Å in the ten lowest-energy structures. From these filters, we found that the salt bridges recapitulated by the NOE assignment module between the residue pairs are (15,62), (23,78), (33,64) and (35,62).

Hydrophobic core of BH_10. Buried residues were selected from the ten lowest-energy structural models using a 10 $Å^2$ solvent-accessible surface area threshold. There are 18 buried residues that contribute up to 70% of the total NOEs assigned by AutoNOE-Rosetta, and two of them are aromatic residues (F34 and F73). While 4D-CHAINS does not assign chemical shifts of side chain groups of aromatic residues (specifically aromatic rings), it provides respective chemical shift assignments of backbone atoms (Cα, Hα, N, H) and the β-carbon and β-proton (Cβ, Hβ) atoms. AutoNOE-Rosetta assigned a total of nine NOEs for the aromatic residues in the hydrophobic core, and the placement of aromatic side chains was optimized by the Rosetta packer algorithm. On close examination of the BH_10 structures, we found that the geometry of the two aromatic side chains was constrained by neighboring residues with methyl groups placed based on NOEs, supporting the accuracy of the aromatic side chain placement.

Visualization of protein structures and image rendering. Images of protein structures were created with PyMOL[54].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The Rosetta macromolecular modeling suite (http://www.rosettacommons.org/) is freely available to academic and non-commercial users. Scripts and protocols used in this article for generating protein backbone blueprints, performing Rosetta design calculations and analyzing protein structures are all available on GitHub (https://github.com/emarcos/beta_sheet).

Data availability

NMR chemical shifts and NOESY cross-peak lists used to determine structures of BH_10 have been deposited in the BMRB with accession code 30495. Coordinates of the ten lowest-energy structures and the restraint lists have been deposited in the wwPDB as PDB 6E5C. The design model of BH_10 is available as Supplementary Dataset 1, and the loop dataset used to analyze the side chain patterns of naturally occurring β-arches is available in Supplementary Dataset 2. Other data are available from the corresponding authors upon reasonable request.

References

32. Wang, G. & Dunbrack, R. L. Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
33. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
34. Fleishman, S. J. et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS ONE* **6**, e20161 (2011).
35. O'Meara, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
36. Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
37. Sheffler, W. & Baker, D. RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. *Protein Sci.* **19**, 1991–1995 (2010).
38. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
39. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
40. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expres. Purif.* **41**, 207–234 (2005).
41. Delaglio, F. et al. NMRPipe: a –spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

42. Ying, J., Delaglio, F., Torchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **68**, 101–118 (2017).
43. Lee, W., Tonelli, M. & Markley, J. L. Nmrfam-Sparky: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
44. Nerli, S., McShan, A. C. & Sgourakis, N. G. Chemical shift-based methods in NMR structure determination. *Prog. Nucl. Mag. Res. Sp* **106-107**, 1–25 (2018).
45. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59**, 147–159 (2014).
46. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* **80**, 884–895 (2012).
47. Berjanskii, M. V. & Wishart, D. S. Unraveling the meaning of chemical shifts in protein NMR. *Biochim. Biophys. Acta* **1865**, 1564–1576 (2017).
48. Nilges, M. A calculation strategy for the structure determination of symmetric dimers by ^{1}H NMR. *Proteins* **17**, 297–309 (1993).
49. Nilges, M. Ambiguous distance data in the calculation of NMR structures. *Fold Des.* **2**, S53–S57 (1997).
50. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227 (2002).
51. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227–241 (2013).
52. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
53. Costantini, S., Colonna, G. & Facchiano, A. M. ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformation* **3**, 137–138 (2008).
54. The PyMOL Molecular Graphics System, Version 1.7.2 (Schrödinger, LLC, 2016).

nature research

Corresponding author(s): Enrique Marcos, David Baker

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Protein structures used for loop geometric analysis were obtained from a non-redundant database of natural proteins from the Protein Data Bank.
Data analysis	Analyses on protein structures were carried out with PyRosetta and all design and folding calculations were done with Rosetta. All NMR analyses were done as described in the methods section with the following programs: NMRPipe, NMRFAM-SPARKY, 4D-CHAINS, CS-Rosetta and TALOS-N. Graphical data representations were done with python and matplotlib. Visualization and image rendering of protein structures was done with Pymol

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

NMR chemical shifts and NOESY cross-peak lists used to determine structures of BH_10 have been deposited under BMRB ID 30495. Coordinates of ten lowest-energy structures and the restraint lists have been deposited under the PDB ID 6E5C.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We selected 19 computationally designed sequences for experimental characterization
Data exclusions	No data was excluded
Replication	Replication was not revelant to our study
Randomization	Randomization was not relevant to our study
Blinding	Blinding was not relevant to our study

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging