ARTICLE

De novo design of a fluorescence–activating β-barrel

Jiayi Dou[1,2,11], Anastassia A. Vorobieva[1,2,11], William Sheffler[1,2], Lindsey A. Doyle[3], Hahnbeom Park[1,2], Matthew J. Bick[1,2], Binchen Mao[1,7], Glenna W. Foight[4], Min Yen Lee[4], Lauren A. Gagnon[4], Lauren Carter[1,2], Banumathi Sankaran[5], Sergey Ovchinnikov[1,2,8], Enrique Marcos[1,2,9], Po-Ssu Huang[1,2,10], Joshua C. Vaughan[4], Barry L. Stoddard[3] & David Baker[1,2,6]*

The regular arrangements of β-strands around a central axis in β-barrels and of α-helices in coiled coils contrast with the irregular tertiary structures of most globular proteins, and have fascinated structural biologists since they were first discovered. Simple parametric models have been used to design a wide range of α-helical coiled-coil structures, but to date there has been no success with β-barrels. Here we show that accurate de novo design of β-barrels requires considerable symmetry-breaking to achieve continuous hydrogen-bond connectivity and eliminate backbone strain. We then build ensembles of β-barrel backbone models with cavity shapes that match the fluorogenic compound DFHBI, and use a hierarchical grid-based search method to simultaneously optimize the rigid-body placement of DFHBI in these cavities and the identities of the surrounding amino acids to achieve high shape and chemical complementarity. The designs have high structural accuracy and bind and fluorescently activate DFHBI in vitro and in *Escherichia coli*, yeast and mammalian cells. This de novo design of small-molecule binding activity, using backbones custom-built to bind the ligand, should enable the design of increasingly sophisticated ligand-binding proteins, sensors and catalysts that are not limited by the backbone geometries available in known protein structures.

There have been considerable recent advances in designing protein folds from scratch[1,2], as well as redesigning already existing native scaffolds to bind small molecules[3–5], but two outstanding unsolved challenges remain. The first is the de novo design of all-β proteins, which is complicated by the tendency of β-strands and sheets to associate intermolecularly to form amyloid-like structures if their register is not perfectly controlled[6]. The second is the design of proteins customized to bind small molecules of interest, which requires precise control over backbone and side chain geometry[5], as well as the balancing of the often opposing requirements of protein folding and function[7]. Success in developing such methods would reduce the long-standing dependency on natural proteins by enabling protein engineers to craft new proteins optimized to bind chosen small-molecule targets, and lay a foundation for de novo design of proteins customized to catalyse specific chemical reactions.

Principles for designing β-barrels

β-barrels are single β-sheets that twist to form a closed structure in which the first strand is hydrogen-bonded to the last[8]. Anti-parallel β-barrels are excellent scaffolds for ligand binding, as the base of the barrel can accommodate a hydrophobic core to provide overall stability, and the top of the barrel can provide a recessed cavity for ligand binding[9], often flanked by loops that can contribute further binding affinity and selectivity[10]. However, all β-sheet topologies are notoriously difficult to design from scratch—to our knowledge, there has been no reported success to date—although several descriptive parametric models of β-barrels have been proposed[11–13]. We first set out to address this challenge by parametrically generating regular arrangements of eight anti-parallel β-strands using the equations for an elliptic hyperboloid of revolution (adapted from previously published work[14], Fig. 1a). β-barrels are characterized by their shear number (S)—the total shift in strand registry between the first and last strands—which determines the hydrophobic packing arrangement and the diameter of the barrel[15,16] (Supplementary Methods). We selected a shear number of $S = 10$ because it is difficult to achieve good core packing for $S = 8$ (the barrel has a smaller diameter and the C_α–C_β vectors point directly at each other), and $S = 12$ results in a cavity that is too large to fill with side chains (Extended Data Fig. 1b–d). We generated ensembles of hyperboloids by sampling the elliptical parameters and the tilt of the generating lines with respect to the central axis around ideal values computed for $S = 10$, and then placed C_α atoms on the hyperboloid surface (Fig. 1a, Supplementary Methods). As found in earlier simulation work[17], backbones generated with constant angles between strands could not achieve perfectly regular hydrogen bonding. To resolve this problem, we introduced force-field guided variation in local twist by gradient-based minimization. We selected the backbones with the most extensive inter-strand hydrogen bonding, connected the strands with short loops and carried out combinatorial sequence optimization to obtain low-energy sequences (Extended Data Fig. 1a). Synthetic genes encoding 41 such designs were produced and the proteins were expressed in *E. coli*. Almost all were found to be insoluble or oligomeric; none of this first set of 41 designs were monomeric with an all-β circular dichroism spectrum (Supplementary Table 2).

In considering the possible reasons for the failure of the initial designs, we noted that many of the backbone hydrogen-bond interactions on the top and bottom of the barrels were distorted or broken (Extended Data Fig. 1e, f). To investigate the origins of this distortion, we experimented with three alternative approaches to generating
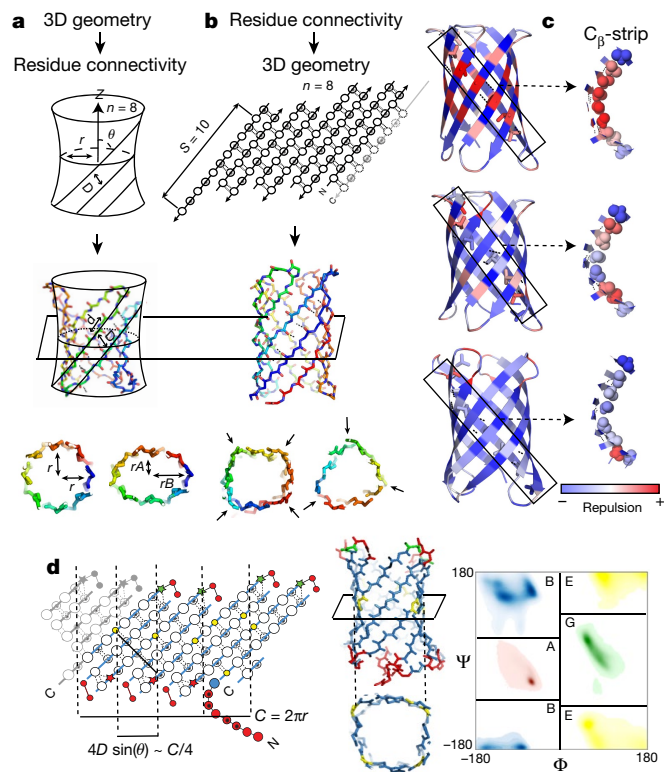
Fig. 1 | Principles for designing β-barrels. a, b, Two methods for β-barrel backbone generation. **a,** Parametric generation of 3D backbones on the basis of the hyperboloid model. The cross-section of the barrel is controlled on the global level with parameters (r, r_A and r_B). **b,** Specification of residue connectivity in a 2D map and assembly of 3D backbones with Rosetta. The cross-section geometry is controlled on the local level with torsion angle bins specified for each residue. **c,** Incorporation of glycine kinks and β-bulges reduces Lennard–Jones repulsive interactions in β-barrels. Full backbones are shown on the left and one C_β-strip is shown on the right. Top, no β-bulge, no glycine kink; middle, one glycine kink in the middle of each C_β-strip, no β-bulge; bottom, one glycine kink in the middle of each C_β-strip, and β-bulges placed near the β-turns. **d,** Blueprint used to generate a β-barrel of type (strand number $n = 8$; shear number $S = 10$) with a square cross-section suitable for ligand binding. The values of the barrel radius (r) and tilt of the strands (θ) used to place glycines are determined by the choice of n and S. The residues in the 2D blueprint (left) and the 3D structure (middle) are coloured by backbone torsion bins (right, ABEGO types nomenclature from Rosetta). Shaded and open circles represent residues facing the barrel interior and exterior, respectively. Glycine positions are shown as yellow circles and β-bulges as stars. The 'corners' in the β-sheet resulting from the presence of glycine kinks are shown as vertical dashed lines. C, barrel circumference; D, distance between strands.

uniform β-barrel backbones lacking loops and with valine at every position as a place-holder (Supplementary Methods). In all cases, we observed breaking of hydrogen-bond interactions after structure minimization with the Rosetta relaxation protocol (Extended Data Fig. 2a), suggesting that there is strain inherent to the closing of the curved β-sheet on itself. To identify the origin of this strain, we repeated the relaxation after imposing strong constraints on the hydrogen-bond interactions to prevent them from breaking. The strain manifested in two places. First, steric clashes build up along strips of side chains in the directions of the hydrogen bonds, perpendicularly to the direction of the β-strands ('C_β-strips', Fig. 1c). Second, a number of residues acquired unfavourable left-handed twist (Extended Data Fig. 2b, c; the chirality of the peptide backbone favours right-handed twist). To reduce the strain arising from steric clashes between C_β atoms, and from the local left-handed twist, we replaced the central valine residue of each C_β-strip with glycine residues (which are normally disfavoured in β-sheets[18]). The achiral glycine can have

a left-hand twist without disrupting the β-sheet hydrogen-bond pattern[15,19] and lacks a C_β atom, reducing the steric clashes within C_β-strips (Fig. 1c, middle). The backbones of most of these glycine residues shifted to the positive Φ torsion bin after minimization, forming torsional irregularities in the β-sheet ('glycine kinks'[15], Extended Data Fig. 2d–e).

On the basis of these observations, we hypothesized that large local deviations in the ideal β-strand twist are necessary to maintain continuous hydrogen-bond interactions between strands in a closed β-barrel, and hence that a parametric approach assuming uniform geometry was not well-suited to building such structures. Therefore, we chose to build β-barrel backbones starting from a 2D map specifying the peptide bonds, the backbone torsion angle bins[20] and the backbone hydrogen bonds (Fig. 1b). In contrast to parametric backbone design, which may be viewed as a '3D-to-2D' approach as a 3D surface is generated and then populated with residues, this alternative strategy proceeds from 2D to 3D and can readily incorporate local torsional deviation (Fig 1a). We generated 3D protein backbones using Rosetta Monte Carlo structure generation calculations starting from an extended peptide chain[21], guided by torsional and distance constraints from the 2D map.

We found that we could control the volume and the 3D shape of the β-barrel cavity by altering the placement of glycine kinks in the 2D map. Such kinks increase local β-sheet curvature, forming corners in an otherwise roughly circular cross-section (Extended Data Fig. 2f, g). We chose to design a square barrel shape and created four corners in the β-sheet by placing five glycine kinks to eliminate strain in the five C_β-strips, and one glycine kink to adjust the curvature of the longest hairpin (Fig. 1d, Extended Data Fig. 3a, Supplementary Methods). With this choice, the resulting 3D backbones have a large interior volume suitable for a ligand-binding cavity. When such backbones were built with canonical type I′ β-turns connecting each β-hairpin, we observed steric strain at the extremities of the C_β-strips (Fig. 1c, centre) and disruption of hydrogen-bond interactions after structure relaxation (Extended Data Fig. 3e). This probably arises because the considerable curvature at the glycine kinks requires that the β-hairpins paired with it (dashed vertical line in Extended Data Fig. 3b) have greater right-handed twist than can be achieved with canonical β-turns. We reasoned that accentuated right-handed twist could be achieved by incorporating β-bulges—disruptions of the regular hydrogen-bonding pattern of a β-sheet[2,22,23]. Indeed, we found that strategic placement of β-bulges on the bottom of the barrel (defined as the side of the N and C termini) and bulge-containing β-turns[22] on the top of the barrel eliminated steric strain and stabilized the hydrogen bonds between the β-strand residues flanking the turns (Fig. 1c, bottom, Extended Data Fig. 3e, f). To tie together the bottom of the barrel, we introduced a 'tryptophan corner'[24,25] by placing a short 3–10 helix, a glycine kink and a Trp at the beginning of the barrel, and an interacting Arg at the C terminus (Extended Data Fig. 3g–j).

Five hundred backbones were generated from the 2D map incorporating the above features, and Rosetta flexible-backbone sequence design calculations were carried out to identify low-energy sequences for each backbone. Four designs with low energy and backbone hydrogen bonding throughout the barrel were selected for experimental characterization (Extended Data Fig. 4a). The sequences of these designs are not related to those of known native proteins (BLAST E values greater than 0.1), and fold into the designed structure in silico (Fig. 2a). Synthetic genes encoding the designs were expressed in *E. coli*. Three of the designs were expressed in the soluble fraction and purified; two had characteristic β-sheet far-ultraviolet circular dichroism (CD) signal (Fig. 2, Extended Data Fig. 4b). Size-exclusion chromatography (SEC) coupled with multi-angle light scattering (MALS) showed that one was a stable monomer (BB1) and the other (BB2) a soluble tetramer (Extended Data Fig. 4c).

BB1 exhibited a strong near-ultraviolet CD signature, which suggests an organized tertiary structure (Fig. 2d). The design was stable at 95 °C, and cooperatively unfolded in guanidine-denaturation experiments (Fig. 2e). The crystal structure of BB1 solved at 1.6 Å resolution was
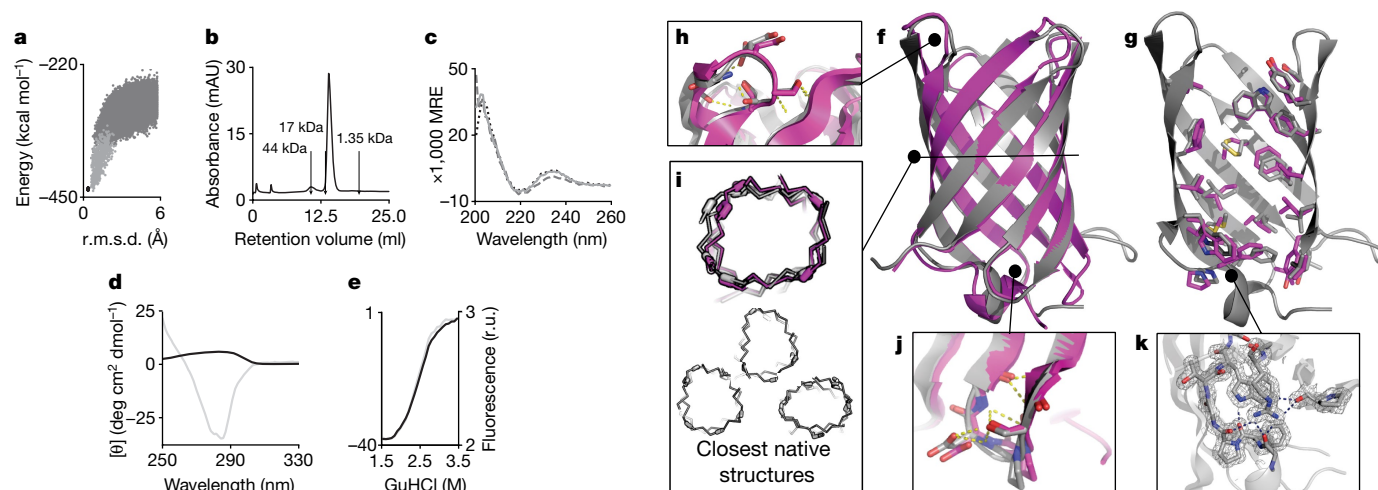
Fig. 2 | Folding, stability and structure of design BB1. a, In silico folding energy landscape. Each grey dot indicates the result of an independent ab initio folding calculation; black dots show results of refinement trajectories starting from design model and light grey dots from the lowest-energy ab initio models. **b**, Size-exclusion chromatogram of the purified monomer (14 kDa). **c**, Far-ultraviolet CD spectra at 25 °C (grey line), 95 °C (black dashed line) and cooled back to 25 °C (black dotted line). **d**, Near-ultraviolet CD spectra in Tris buffer (grey line) and 7 M GuHCl (black line). **e**, Cooperative unfolding in GuHCl monitored by near-ultraviolet CD signal at 285 nm (grey line) and tryptophan fluorescence (black line). **f–j**, Superpositions of the crystal structure (grey)

and the design model (pink): overall backbone superposition (**f**); section along the β-barrel axis showing the rotameric states of core residues (**g**); one of the top loops with a G1 β-bulge (**h**); and equatorial cross-section of the β-barrel, showing the geometry of the interior volume (**i**). The glycine kinks are shown as sticks. The bottom of the panel shows the cross-sections of the three closest native β-barrel structures on the basis of TM-score[33] (RCSB Protein Data Bank (PDB) IDs: 1JMX (0.77); 4IL6 (chain O) (0.73); 1PBY (0.71)). **j**, One of the bottom loops with a classic β-bulge. **k**, Crystal structure and $2mF_o − DF_c$ electron density of the tryptophan corner, contoured at 1.5σ.

very close to the design model (1.4 Å backbone root mean squared deviation (r.m.s.d.) over 99 of 109 residues; Extended Data Fig. 4d–f). Essentially all of the key features of the design model are found in the crystal structure (Fig. 2f–k). The barrel cross-section in the crystal structure is very similar to that of the design model, with an overall square shape with corners at the glycine kinks. Natural β-barrel crystal structures do not have this shape; the cross-sections of the closest structure matches in the PDB are shown in Fig. 2i. All seven designed β-turns and β-bulges are correctly recapitulated in the crystal structure (Fig. 2h, j), along with the 3–10 helix and tryptophan corner (Fig. 2k).

Design of small-molecule-binding β-barrels

Having determined principles for de novo design of β-barrels, we next sought to design functional β-barrels with binding sites tailored for a small molecule of interest. We chose DFHBI (Fig. 3a, left, green), a derivative of the intrinsic chromophore of GFP, to test the computational design methods. Owing to its internal torsional flexibility in solution, DFHBI does not fluoresce unless it is constrained in the planar Z conformation[26,27]. We sought to design protein sequences that fold into a stable β-barrel structure with a recessed cavity lined with side chains to constrain DFHBI in its fluorescent planar conformation. We chose to take a three step approach: (1) de novo construction of β-barrel backbones, (2) placement of DFHBI in a dedicated pocket, and (3) energy-based sequence design. For the first step, we stochastically generated 200 β-barrel backbones on the basis of the 2D map described above (Extended Data Fig. 5b–d).

The placement of the ligand in the binding pocket requires sampling of both the rigid-body degrees of freedom of the ligand, and the sequence identities of the surrounding amino acids that form the binding site. Because of the dual challenges associated with optimization of structure and sequence simultaneously, most approaches to designing ligand-binding sites to date have separated sampling into two steps: rigid-body placement of the target ligand in the protein-binding pocket and then design of the surrounding sequence[4,5,28]. This two-step approach has the limitation that the optimal rigid-body placement cannot be determined independently of knowledge of the possible interactions with the surrounding amino acids. The RosettaMatch method[29] can identify rigid-body and interacting-residue placements simultaneously,

but is limited to a small number of pre-defined ligand-interacting residues[3]. We addressed these challenges with a new 'rotamer interaction field (RIF)' docking method that simultaneously samples over rigid-body and sequence degrees of freedom. RIF docking first generates an ensemble of billions of discrete amino acid side chains that make hydrogen-bonding and non-polar hydrophobic interactions with the target ligand (Fig. 3a, right). Then, scaffolds are docked into this pre-generated interacting ensemble using a grid-based hierarchical search algorithm (Extended Data Fig. 5a). We used RIF docking to place DFHBI into the upper half of the β-barrel scaffolds, resulting in 2,102 different ligand–scaffold pairs with at least four hydrogen-bonding and two hydrophobic interactions (Fig. 3a).

To identify protein sequences that not only buttress the ligand-coordinating residues from the RIF docking but also have low intra-protein energies to drive protein folding, we developed and applied a Monte Carlo-based sequence design protocol that iterates between (1) fixed-backbone design around the ligand-binding site to optimize the ligand-interaction energy and (2) flexible-backbone design for the rest of protein, optimizing the total complex energy (Fig. 3b). Forty-two designs with large computed folding-energy gaps and low-energy intra-protein and protein–ligand interactions were selected for experimental characterization, plus an additional 14 disulfide-bonded variants (Extended Data Fig. 5e). Ligand-docking simulations after extensive structure refinement revealed that owing to the approximate symmetry of the hydrogen-bonding pattern of DFHBI, many of the designed binding pockets could accommodate the ligand in two equally favourable orientations (Extended Data Fig. 5f).

Synthetic genes encoding the 56 designs were obtained and the proteins were expressed in E. coli. Thirty-eight of the proteins were well-expressed and soluble; SEC and far-ultraviolet CD spectroscopy showed that 20 were monomeric β-sheet proteins (Supplementary Table 3). Four of the oligomer-forming designs became monomeric upon incorporation of a disulfide bond between the N-terminal 3–10 helix and the barrel β-strands. The crystal structure of one of the monomeric designs (b10) was solved to 2.1 Å, and was found to be very close to the design model (0.57 Å backbone r.m.s.d., Fig. 3c). The upper barrel of the crystal structure maintains the designed pocket, which is filled with multiple water molecules (Fig. 3c, Extended Data Fig. 6b). Thus,
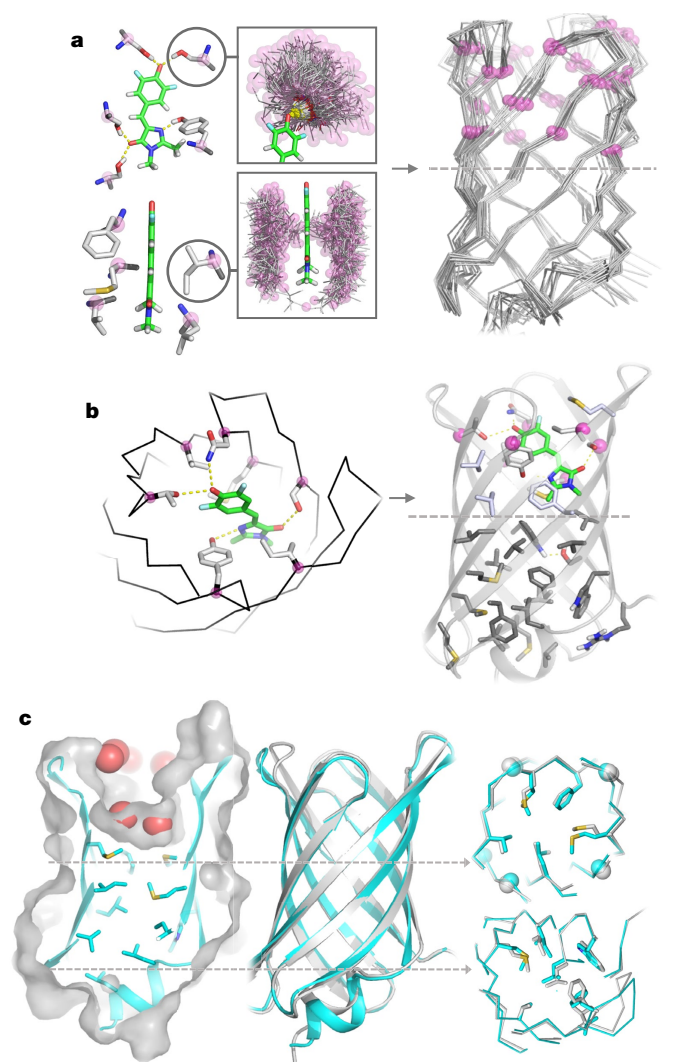
Fig. 3 | Computational design and structural validation of β-barrels with recessed cavities for ligand binding. a, Left, ensembles of side chains generated by the RIF docking method making hydrogen-bonding (upper left) and hydrophobic interactions (lower left) with DFHBI (green); pre-generated interacting rotamers are shown in grey with backbone C_α highlighted by magenta spheres. Right, ensemble of 200 β-barrel backbones with C_α atoms surrounding the binding cleft indicated by magenta spheres. **b**, Each ligand–scaffold pair (left) with multiple ligand-coordinating interactions from RIF docking is subjected to Rosetta energy-based sequence design calculations (right): positions around the ligand (light purple, above the dashed line) are optimized for ligand binding; the bottom of the barrel (dark grey, below the dashed line), for protein stability. **c**, Left, crystal structure (cyan cartoon with grey surface) of b10 with a recessed binding pocket filled with water molecules (red spheres). Middle, b10 design model backbone (silver) superimposed on the crystal structure (cyan). Right, comparison of crystal structure and design model for two different barrel cross-sections (indicated by dashed lines); glycine C_α atoms are indicated by spheres in the upper layer.

the design principles described above are sufficiently robust to enable the accurate design of potential small-molecule-binding pockets.

Two of the 20 monomeric designs—b11 and b32—were found to activate DFHBI fluorescence by 12- and 8-fold with binding dissociation constants (K_D) values of 12.8 and 49.8 μM, respectively (Extended Data Fig. 6f). Knockout of interacting residues in the designed binding pocket eliminated fluorescence (Extended Data Fig. 6g). The ligand-binding activity comes at a substantial stability cost as almost half of the barrel is carved out to form the binding site: although the nonfunctional BB1 design does not temperature-denature, both b11

and b32 undergo reversible thermal melting transitions (Extended Data Fig. 6e), even though b11 contains a stabilizing disulfide bond (the parent design that lacks the disulfide (b38) is not a monomer; Extended Data Fig. 6c, d). We sought to improve the binding interactions by redesigning β-turns around the ligand-binding site (Supplementary Table 6). b11L5F, a 110-residue protein with a five-residue fifth turn, activated DFHBI fluorescence by 18-fold with a K_D value of 7.5 μM (Extended Data Fig. 6f, h).

The sequence determinants of b11L5F fold and function were investigated by assaying the effect of each single amino acid substitution ($19 \times 110 = 2,090$ in total) on both protein stability[30] and DFHBI activation on the yeast cell surface. The function (fluorescence activation) and stability (proteolysis resistance) landscapes have similar overall features consistent with the design model, with residues buried in the designed β-barrel geometry being much more conserved than surface-exposed residues (Fig. 4a, Extended Data Fig. 7a, b). The function landscape suggests that the geometry of the designed cavity is critical to activating DFHBI fluorescence: the key sequence features that specify the geometry of the cavity—the glycine kinks and the tryptophan corner—are strictly conserved (Fig. 4). Among all substitutions of the seven coordinating residues from RIF docking, only a single substitution (V103L) increased fluorescence (Fig. 4c, upper panel). Whereas the structure and function landscapes were very similar at the bottom of the barrel (Fig. 4b), there was a notable trade-off between stability and function at the top of the barrel around the designed binding site (Fig. 4c): many substitutions that stabilize the protein markedly reduce fluorescence activation (Fig. 4c, right). This bottom–top contrast indicates that success in de novo design of fold and function requires a substantial portion of the protein (in our case, the bottom of the barrel) to provide the driving force for folding as the functional site will probably be destabilizing.

Guided by the comprehensive protein stability and fluorescence activation maps, we combined substitutions at three positions that improved function without compromising stability (V103L, V95AG and V83ILM; Extended Data Fig. 8a, b), and obtained variants with ten-fold higher DFHBI fluorescence that form stable monomers without a disulfide bond (b11L5F.1; Extended Data Fig. 8c). The crystal structure of one of the improved variants (b11L5F_LGL; mutant 83L/95G/103L in Extended Data Fig. 8b) was solved to 2.2 Å and was very close to the design model with the majority of the buried side chains adopting the designed conformation (Extended Data Fig. 9a–d). However, the electron density in the binding site could not be resolved, consistent with the multiple DFHBI binding modes suggested by the docking calculations (Extended Data Fig. 9e–g; Extended Data Fig. 5f). A second round of computational design calculations was carried out to favour a specific binding mode by optimizing the protein–ligand interactions in the lowest-energy docked conformation, and to rearrange the hydrophobic packing interactions in the bottom of the barrel now freed from the disulfide bond. Five designs predicted by ligand-docking calculations to have a single ligand-binding conformation were experimentally tested and three showed increased fluorescence activity, the best of which increased the fluorescence by approximately 1.4-fold (b11L5F.2; Extended Data Fig. 8d–e). Screening of two combinatorial libraries (based on b11L5F.1 and b11L5F.2), incorporating additional beneficial substitutions identified in the b11L5F stability and function maps, yielded variants with another 1.5–2-fold increased fluorescence and improved protein stability (Extended Data Figs. 8f–h, 10a, b). We refer to these mini-fluorescence-activating proteins as mFAPs in the remainder of the text; mFAP0 and mFAP1 are variants of b11L5F.2, and mFAP2 of b11L5F.1. mFAP1 and mFAP2 activate 0.5 μM DFHBI fluorescence by 80- and 60-fold with K_D values of 0.56 μM and 0.18 μM, respectively (Fig. 5d).

The 1.8 Å and 2.3 Å crystal structures of mFAP0 and mFAP1 in complex with DFHBI were virtually identical to the design models with an overall backbone r.m.s.d. of 0.91 Å and 0.64 Å (Fig. 5a–c, Extended Data Fig. 9h, i). DFHBI is in the planar Z conformation with unambiguous electron density in both structures (Fig. 5a, Extended Data
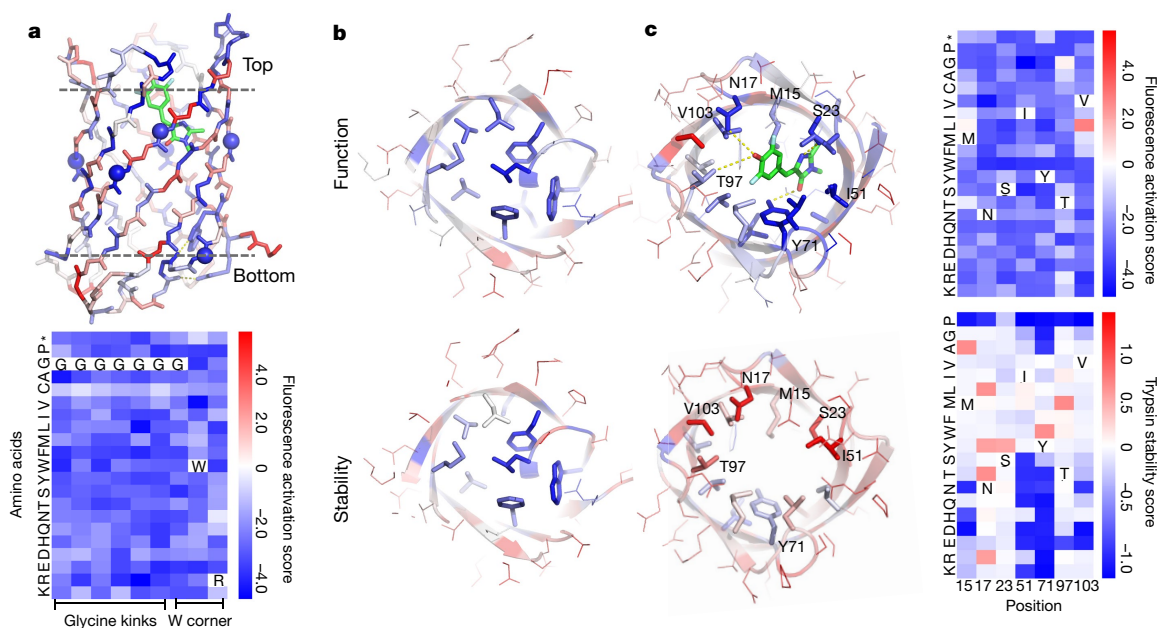
Fig. 4 | Sequence dependence of fold and function. Each position was mutated one at a time to each of the other 19 amino acids, and the resulting library was subjected to selection for fluorescence or stability to proteases. **a**, Top, b11L5F backbone model coloured according to relative fluorescence activation score at each position. Blue positions are strongly conserved during yeast selection; red positions are frequently substituted by other amino acids. Residues buried in the design model are much more conserved than solvent-exposed residues. Bottom, all the mutations to the glycine kinks (spheres in the top panel) and tryptophan (W) corner considerably reduced fluorescence; the shape of the designed structure is critical for the designed function. **b**, **c**, Bottom (**b**) and top (**c**) comparisons of b11L5F side chains coloured by relative fluorescence activation scores (top row) and stability scores (bottom row). In the bottom of the barrel, core residues were strongly conserved in both the function and stability selections (**b**); in the top barrel there is a clear function–stability trade-off with the key DFHBI-interacting residues being critical for function but far from optimal for stability (**c**, substitution patterns at these positions are shown on the right). Fluorescence activation and stability scores were derived from two biologically independent experiments with a greater than tenfold sequencing coverage. Standard deviation and confidence intervals are provided in Extended Data Fig. 7.

Fig. 9j). In addition to three designed hydrogen bonds, a water molecule was found to interact with the solvent-exposed phenol group in DFHBI (Fig. 5b). The DFHBI-binding modes in the crystal structures are nearly identical to the lowest-energy docked conformations used in the second round of design calculations, with all-atom r.m.s.d. values of 0.12 Å and 0.35 Å, respectively (Fig. 5c, Extended Data Fig. 9k). Three mutations shared by mFAP0 and mFAP1 in the bottom barrel (P62D, M65L and L85M or L85Y, Extended Data Fig. 10b) probably stabilize the protein by helical capping and subtle hydrophobic rearrangements (Extended Data Fig. 9l). The M27W mutation in mFAP1 introduced an additional hydrogen bond to DFHBI that probably produces the 5-nm red-shift in its fluorescence spectra (Fig. 5d, Extended Data Fig. 10c, e). mFAP2, which is based on b11L5F.1, has a six-residue insertion in the seventh β-turn that is predicted to form multiple intra-loop hydrogen bonds (Extended Data Fig. 10b, right).

In vivo fluorescence activation

To determine whether the designed DFHBI-binding fluorescence-activating proteins function in living cells, we imaged mFAP1 and mFAP2 in *E. coli*, yeast and mammalian cells by conventional wide field epifluorescence microscopy and confocal microscopy. Both mFAP1 and mFAP2 activated fluorescence less than 5 min after addition of 20 μM DFHBI. Cytosolic expression of mFAPs in *E. coli* and mammalian cells resulted in clear fluorescence throughout the cells (Fig. 5e, Extended Data Fig. 10f). Yeast cells with mFAPs targeted to the cell surface displayed fluorescence in a thin region outside of the plasma membrane (Fig. 5f, Extended Data Fig. 10g). Fusion of the mFAPs to a mitochondria-targeting signal peptide and to the endoplasmic reticulum-localized protein sec61β resulted in fluorescence tightly localized to these organelles in both fixed (Fig. 5g, h) and living cells (Supplementary Videos). The quantum yields of mFAP1 and mFAP2 in complex with DFHBI are 2.0% and 2.1%, respectively (Extended

Data Fig. 4g, comparable with Y-FAST:HBR[31]). The brightness of de novo mFAPs in complex with DFHBI is about 35-fold lower than that of eGFP; there is still considerable room for improving their fluorescence activity.

Conclusion

It is instructive to compare the structures of our designed fluorescence-activating proteins with those of natural fluorescent proteins (Fig. 6). Both are β-barrels, and have similar chromophores, but our designs have less than half the residues and narrower barrels connected with short β-turns (Fig. 6a). In both cases, specific protein–chromophore interactions reduce energy dissipation from intramolecular motions[32], but the hydrogen bonding and hydrophobic packing around DFHBI is different from GFP and tailored to the smaller and simpler β-barrel (Fig. 6b). The precise structural control enabled by computational design, together with the greater exposure of the chromophore, may prove useful for fluorescence-based imaging and sensing applications.

The comparison in Fig. 6 highlights the two primary advances in this paper: the first successful de novo design of a β-barrel, and the first full de novo design of a small-molecule-binding protein. The first advance required the elucidation of general principles for designing β-barrels, notably the requirement for systematic symmetry-breaking to enable hydrogen bonding throughout the barrel structure. These principles, identified by pure geometric considerations, coupled with computer simulations after the failure of the initial parametric design approach, are borne out by both the crystal structures and the sequence fitness landscapes. The second advance goes considerably beyond the design of ligand-binding proteins and catalysts to date, which has relied on repurposing naturally occurring scaffolds. The three-step approach described in this paper—first, identifying the basic principles required for specifying a general fold class; second, using these principles to generate a family of backbones with pocket geometries
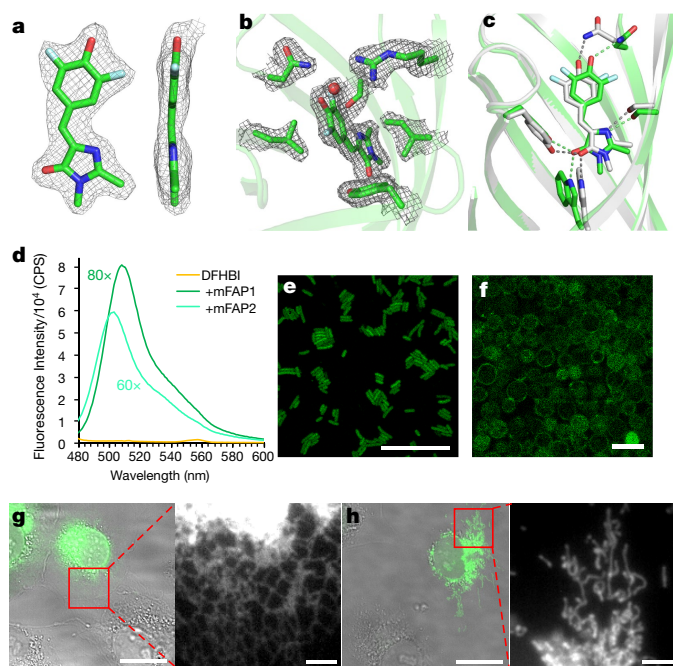
Fig. 5 | Structure and function of mFAPs. a, **b**, $2F_o - F_c$ omit electron density in the mFAP1–DFHBI complex crystal structure contoured at 1.0σ. **a**, DFHBI is clearly in the planar Z conformation rather than the non-fluorescent twisted conformations. **b**, The planar conformation is stabilized by closely interacting residues. **c**, Superposition of mFAP1 design model (silver) and the crystal structure (green). Hydrogen bonds coordinating DFHBI are indicated with dashed lines. **d**, Fluorescence emission spectra of $0.5\,\mu M$ DFHBI with or without $5\,\mu M$ mFAPs, excited at 467 nm. **e**, **f**, Confocal micrographs of *E. coli* cells expressing mFAP2 in the presence of DFHBI (**e**) and yeast cells displaying Aga2p–mFAP2 fusion proteins on the cell surface (**f**). Scale bars, $20\,\mu m$ (**e**), $10\,\mu m$(**f**). **g**, **h**, Overlay of widefield epifluorescence (green) and brightfield (grey) images of fixed COS-7 cells expressing sec61β–mFAP1 (**g**) and mito–mFAP2 (**h**; mito, mitochondrial targeting sequence) with expanded views of the fluorescence in the boxed regions. Scale bars, $20\,\mu m$ (**g**, **h**), $3\,\mu m$ (expanded). Two biological replicates were performed with similar observation.

matched to the ligand or substrate of interest and third, designing complementary binding pockets buttressed by an underlying hydrophobic core—provides a general solution to the problem of de novo design of ligand-binding proteins. This generative approach enables the exploration of an effectively unlimited set of backbone structures with shapes customized to the ligand or substrate of interest and provides a test of our understanding of the determinants of folding and binding that goes well beyond descriptive analyses of existing protein structures.

Online content

Fig. 6 | Comparison of structures of GFP and mFAP1. a, Surface mesh and ribbon representations of structures of GFP (left, PDB ID: 1EMA) and the computationally designed mFAP1 (right) with the chromophores (spheres) embedded in the protein. GFP, a product of natural evolution, has more than twice the number of residues, and a taller (top) and wider (bottom) barrel. Resolved water molecules in the crystal structures are shown as light purple spheres. **b**, Close-up of chromophore binding interactions in GFP (left) and mFAP1 (right).

1. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
2. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
3. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
4. Bick, M. J. et al. Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **6**, e28909 (2017).
5. Dou, J. et al. Sampling and energy evaluation challenges in ligand binding protein design. *Protein Sci.* **26**, 2426–2437 (2017).
6. Richardson, J. S. & Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
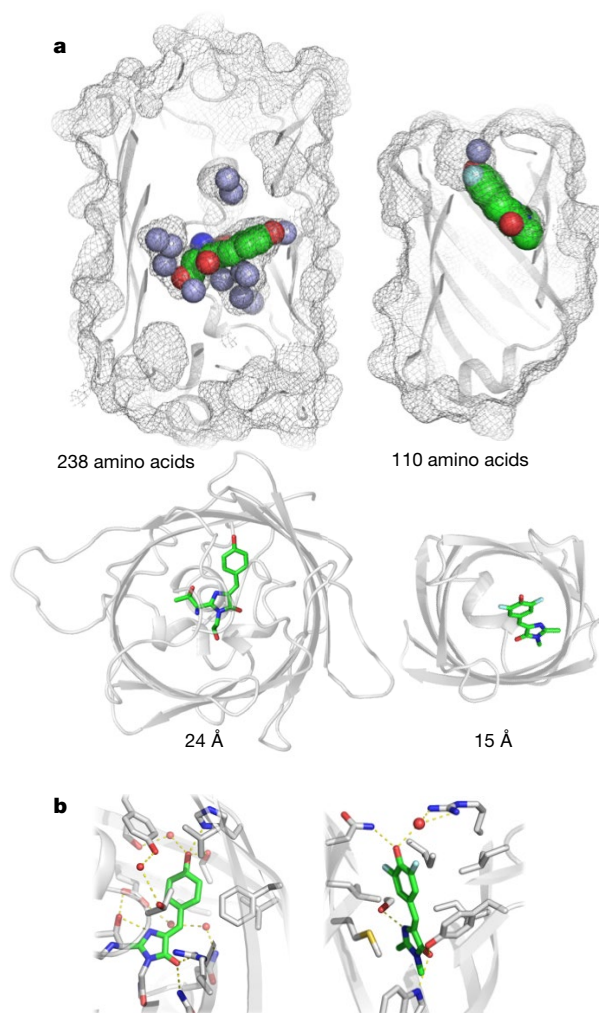7. Polizzi, N. F. et al. De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nat. Chem.* **9**, 1157–1164 (2017).
8. LaLonde, J. M., Bernlohr, D. A. & Banaszak, L. J. The up-and-down beta-barrel proteins. *FASEB J.* **8**, 1240–1247 (1994).
9. Flower, D. R. Structural relationship of streptavidin to the calycin protein superfamily. *FEBS Lett.* **333**, 99–102 (1993).
10. Richter, A., Eggenstein, E. & Skerra, A. Anticalins: exploiting a non-Ig scaffold with hypervariable loops for the engineering of binding proteins. *FEBS Lett.* **588**, 213–218 (2014).
11. Toda, M., Zhang, F. & Athukorallage, B. Elastic surface model for beta-barrels: geometric, computational, and statistical analysis. *Proteins* **86**, 35–42 (2018).
12. Novotný, J., Bruccoleri, R. E. & Newell, J. Twisted hyperboloid (Strophoid) as a model of β-barrels in proteins. *J. Mol. Biol.* **177**, 567–573 (1984).
13. Naveed, H., Xu, Y., Jackups, R., Jr & Liang, J. Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. *J. Am. Chem. Soc.* **134**, 1775–1781 (2012).
14. Lasters, I., Wodak, S. J., Alard, P. & van Cutsem, E. Structural principles of parallel beta-barrels in proteins. *Proc. Natl Acad. Sci. USA* **85**, 3338–3342 (1988).
15. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* **236**, 1369–1381 (1994).
16. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* **236**, 1382–1400 (1994).
17. Salemme, F. R. Conformational and geometrical properties of β-sheets in proteins. III. Isotropically stressed configurations. *J. Mol. Biol.* **146**, 143–156 (1981).

18. Minor, D. L., Jr & Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367**, 660–663 (1994).
19. Fujiwara, K., Ebisawa, S., Watanabe, Y., Toda, H. & Ikeguchi, M. Local sequence of protein β-strands influences twist and bend angles. *Proteins* **82**, 1484–1493 (2014).
20. 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).
21. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
22. Richardson, J. S., Getzoff, E. D. & Richardson, D. C. The beta bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl Acad. Sci. USA* **75**, 2574–2578 (1978).
23. Chan, A. W., Hutchinson, E. G., Harris, D. & Thornton, J. M. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.* **2**, 1574–1590 (1993).
24. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S. & Richardson, D. C. The tyrosine corner: a feature of most Greek key β-barrel proteins. *Protein Sci.* **3**, 1927–1937 (1994).
25. Greene, L. H., Hamada, D., Eyles, S. J. & Brew, K. Conserved signature proposed for folding in the lipocalin superfamily. *FEBS Lett.* **553**, 39–44 (2003).
26. Paige, J. S., Wu, K. Y. & Jaffrey, S. R. RNA mimics of green fluorescent protein. *Science* **333**, 642–646 (2011).
27. Warner, K. D. et al. Structural basis for activity of highly efficient RNA mimics of green fluorescent protein. *Nat. Struct. Mol. Biol.* **21**, 658–663 (2014).
28. Allison, B. et al. Computational design of protein–small molecule interfaces. *J. Struct. Biol.* **185**, 193–202 (2014).
29. Zanghellini, A. et al. New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794 (2006).
30. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
31. Plamont, M.-A. et al. Small fluorescence-activating and absorption-shifting tag for tunable protein imaging in vivo. *Proc. Natl Acad. Sci. USA* **113**, 497–502 (2016).
32. Meech, S. R. Excited state reactions in fluorescent proteins. *Chem. Soc. Rev.* **38**, 2922–2934 (2009).
33. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

METHODS

Computational design of non-functional β-barrels. De novo design of non-functional β-barrels can be divided into two main steps: backbone construction and sequence design. For backbone construction, two different approaches were presented: parametric backbone generation and fragment-based backbone assembly. Example scripts and command lines for each method are available in Supplementary Data.

Parametric backbone generation and sequence design on the basis of hyperboloid models. β-strand arrangements were generated using the equation of a hyperboloid of revolution with an elliptic cross-section, sampling the elliptic radii around the ideal value of β-barrel radius with number of strands (*n*) and the shear number (*S*) (see Supplementary Methods). Eight β-strands were arranged as equally spaced straight lines running along the surface of the hyperboloid. A reference C_α atom was defined as the intersection between the first strand and the cross-section ellipse. The other C_α atoms were systematically populated along the eight strands from this reference residue. The peptide backbone was generated from the C_α coordinates using the BBQ software[34]. The arrangements of discrete β-strands were minimized with geometric constraints to favour backbone hydrogen bonds. One round of fixed-backbone sequence design calculation was carried out to pack the barrel cavity with hydrophobic residues. The resulting β-strand arrangements with the best hydrogen-bond connectivity and the tightest hydrophobic packing were selected to be connected by short (two to four residues) β-turns. Two iterations of the loop hashing protocol implemented in RosettaRemodel[35] were performed to close the strands and refine the turns. The sequence design of those β-turns was constrained to sequence profiles derived from natural proteins. Low-energy amino acid sequences were obtained for the connected backbones using a flexible-backbone design protocol (see Supplementary Data). Designs with high sequence propensity for forming β-strands, reasonable peptide-bond geometry and tight-packed hydrophobic cores are selected for experimental test (see Supplementary Table 2).

Backbone assembly from fragments guided by a 2D map. The presented 2D map (Fig. 1d) was designed with the longest strand length observed in soluble β-barrel structures to obtain a β-barrel tall enough to accommodate a hydrophobic core and a binding cavity. The length of each strand depends on its specific position and the shear number of the barrel (see Supplementary Methods). Glycine kinks and β-bulges were placed on the map as described in the main text. Specific β-turn types were used to connect the β-strands on the basis of their relative positions to β-bulges (see Supplementary Methods). On the basis of this 2D map, we generated a constraint file and a blueprint file to guide the assembly of the barrel using peptide fragments from Rosetta fragments library. In the constraint file, each backbone hydrogen bond was described as a set of distance and angle constraints (Extended Data Fig. 5b). A set of distance and torsion constraints specific to the tryptophan corner were added to the constraint file (Extended Data Fig. 3g–j, Supplementary Methods). In the blueprint file, a torsion angle bin was attributed to every residue in the peptide chain, according to the ABEGO nomenclature of Rosetta. After minimizing the assembled backbones using Rosetta centroid-scoring function with imposed constraints, our protocol output an ensemble of poly-valine β-barrel backbones with defined glycine kinks, β-bulges, β-turns and the backbone of the tryptophan corner. The main challenge of building scaffolds with this protocol is to properly balance structure diversity and reasonable backbone torsion angles with the strong geometric constraints imposed during minimization. For this work, we circumvented this problem by performing two additional rounds of sequence design calculation to regularize and prepare scaffolds for designing ligand-binding β-barrels (Extended Data Fig. 5b–d, Supplementary Methods).

Sequence design of nonfunctional β-barrels. Five hundred poly-valine backbones with good hydrogen bonds and torsion angles were selected as input for Rosetta sequence design. Low-energy sequences for the desired β-barrel fold were optimized over several rounds of flexible-backbone sequence design. We used a genetic algorithm to effectively search the sequence space: each parent backbone was used as input to produce ten designs through individual Monte Carlo searching trajectory. The best ~10% of the output designs were selected on the basis of the evaluation for total energy, backbone hydrogen bonds, backbone omega and Φ/Ψ torsion angles and hydrophobic-packing interactions. The selected models were used as inputs for the next round of design calculation. After 12 rounds of design and selection, no more improvements on the backbone quality metrics were observed (an indication of searching convergence). We then performed a backbone refinement by minimization in Cartesian space and a final round of design calculation (backbone flexibility was limited in torsion space for all the design calculation). The final top designs converged to the offspring of three initial backbones, sharing 36% to 99% sequence identity. For every parent backbone, one or two designs with the best hydrophobic packing interactions were selected for experimental characterization. The four designs (BB1–4) share 46% to 72% sequence identity.

Computational design of DFHBI-binding fluorescence-activating β-barrels. DFHBI is short for the chemical name ((Z)-4-(3,5-difluoro-4-hydroxybenzylidene)-1,2-dimethyl-1H-imidazol-5(4H)-one). De novo design of DFHBI-binding β-barrels consists of three steps: (1) generation of ensembles β-barrel scaffolds (see above), (2) ligand placement by RIF docking and (3) sequence design. Two hundred input scaffolds were generated in step 1 and used in the following steps. Example scripts and command lines are available in Supplementary Data.

RIF docking. The RIF docking method performs a simultaneous, high-resolution search of continuous rigid-body docking space as well as a discrete sequence-design space. The search is highly optimized for speed and in many cases, including the application presented here, is exhaustive for given scaffold–ligand pair and design criteria. RIF docking comprises two steps. In the first step, ensembles of interacting discrete side chains (referred to as 'rotamers') tailored to the target are generated. Polar rotamers are placed on the basis of hydrogen-bond geometry whereas apolar rotamers are generated via a docking process and filtered by an energy threshold. All the RIF rotamers are stored in ~0.5 Å sparse binning of the six-dimensional rigid body space of their backbones, allowing extremely rapid lookup of rotamers that align with a given scaffold position. To facilitate the next docking step, RIF rotamers are further binned at 1.0 Å, 2.0 Å, 4.0 Å, 8.0 Å and 16.0 Å resolution. In the second step, a set of β-barrel scaffolds is docked into the produced rotamer ensembles, using a hierarchical branch-and-bound search strategy (see Extended Data Fig. 5a). Starting with the coarsest 16.0 Å resolution, an enumerative search of scaffold positions is performed: the designable scaffold backbone positions are checked against the RIF to determine whether rotamers can be placed with favourable interacting scores. All acceptable scaffold positions (up to a configurable limit, typically ten million) are ranked and promoted to the next search stage. Each promoted scaffold is split into 2^6 child positions in the six-dimensional rigid-body space, providing a finer sampling. The search is iterated at 8.0 Å, 4.0 Å, 2.0 Å, 1.0 Å and 0.5 Å resolutions. A final Monte Carlo-based rotamer packing step is performed on the best 10% of rotamer placements to find compatible combinations.

Sequence design of DFHBI-binding β-barrels. A total number of 2,102 DFHBI–scaffold pairs from RIF docking were continued for Rosetta sequence design. Our design protocol iterated between a fixed-backbone binding-site design calculation and a flexible-backbone design for the rest of scaffold positions. Three variations of this design protocol were used during the sequence optimization. In the initial two rounds of design calculation, RIF rotamers (interacting residues placed during RIF docking) were fixed to maintain the desired ligand coordination. Repacking of RIF rotamers was allowed in the final round of design calculation, assuming that the binding sites have been optimized enough to retain these interactions. A Rosetta mover that biases aromatic residues for efficient hydrophobic packing was added after the first round of design. A similar selection approach and Cartesian minimization as described for non-functional sequence design were used to propagate sequence search and refine the design models. Evaluations on ligand-binding interface energy and shape complementarity were added to the selection criteria. The final set of designs was naturally separated into clusters on the basis of their original RIF docking solutions. For each cluster, a sequence profile was generated to guide an additional two rounds of profile-guided sequence design. Forty-two designs from 22 RIF docking solutions (20 input scaffolds) were selected for experimental characterization (see Supplementary Table 3).

Post-design model validation and ligand-docking simulation. To validate the protein and ligand conformations of the selected designs, we performed ligand-docking simulations on refined apo-protein models. Protein model refinement was carried out on the unbound model of the designs by running five independent 10-ns molecular dynamics simulations with structural averaging and geometric regularization[36]. Then ligand-docking simulation was performed on this refined unbound model using RosettaLigand[37] using Rosetta energy function[38], allowing rigid-body orientation and intra-molecular conformation of the ligand as well as surrounding protein residues (both on side chains and backbones) to be sampled. The ligand-binding energy landscapes were generated by repeating 2,000 independent docking simulations.

Design of disulfide bonds. The disulfide bonds were designed between the N-terminal 3–10 helix and a residue on one of the β-strands on the opposite side to the tryptophan corner. The first six residues of the designs model were rebuilt with RosettaRemodel[35] and checked for disulfide bond formation using geometric criteria. Once a disulfide bond was successfully placed, the N-terminal helix was redesigned.

Redesign of β-turns for b11. Three β-turns (loops 3, 5 and 7) surrounding the DFHBI-binding site of b11 were redesigned to make additional protein–ligand contacts. A set of 'pre-organized' loops with high content of intra-loop hydrogen bonds and low B-factors were collected from natural β-barrel structures, and used as search template to build individual loop fragment library. Those custom libraries were used as input for RosettaRemodel to build an ensemble of loop

insertions in the b11 design model bound to DFHBI. Two rounds of flexible-backbone design calculation were carried out to optimize ligand interface energy and shape complementarity using sequence profiles to maintain the template backbone hydrogen bonds. Designed loop sequences were validated in silico by kinematic loop closure[39] (KIC). Five hundred loop conformations were generated by independent KIC sampling and scored by Rosetta energy function. Thirty-six designs with improved ligand-interface energy, shape complementarity and converging conformation sampling were selected for experimental characterization (see Supplementary Data, Supplementary Table 6).

Redesign of β-barrel core and DFHBI-binding site for b11L5F.1. After releasing the disulfide bond in b11L5F, with ligand modelled in the lowest-energy docked conformation for b11L5F (see Extended Data Fig. 5f, right), we performed another round of design calculation to further optimize the β-barrel core packing and ligand-binding interactions. The design protocol was very similar to the one used before with fixed ligand–hydrogen-bonding residues from RIF docking. Five designs with 9–15 mutations after manual inspection were selected for experimental characterization.

Protein expression and purification. Genes encoding the non-functional β-barrel designs (41 from parametric design and 4 from fragment-base design) were synthesized and cloned into the pET-29 vector (GenScript). Plasmids were then transformed into BL21*(DE3) E. coli strain (NEB). Protein expression was induced either by 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) at 18 °C, or by overnight 37 °C growth in Studier autoinduction medium. Cells were lysed either by sonication (for 0.5–1-l cultures) or FastPrep (MPBio) (for 5–50-ml cultures). Soluble designs were purified by Ni-NTA affinity resin (Qiagen) and monomeric species were further separated by Akta Pure fast protein liquid chromatography (FPLC)(GE Healthcare) using a Superdex 75 increase 10/300 GL column (GE Healthcare). Fifty-six designs encoding DFHBI-binding designs were synthesized and cloned into pET-28b vector (Gen9). Protein expression and purification were carried out in the same way.

Circular dichroism (CD). Purified protein samples were prepared at 0.5 mg/ml in 20 mM Tris buffer (150 mM NaCl, pH 8.0) or PBS buffer (25 mM phosphate, 150 mM NaCl, pH 7.4). Wavelength scans from 195 nm to 260 nm were recorded at 25 °C, 75 °C, 95 °C and cooling back to 25 °C. Thermal denaturation was monitored at 220 nm or 226 nm from 25 °C to 95 °C. Near-ultraviolet wavelength scan from 240 nm to 320 nm and tryptophan fluorescence emission were recorded in the absence and presence of 7 M guanidinium chloride (GuHCl). Chemical denaturation in GuHCl was monitored by both tryptophan fluorescence and near-ultraviolet CD signal at 285 nm. The concentration of the GuHCl stock solution was measured with a refractometer (Spectronic Instruments). Far-ultraviolet CD experiments were performed on an AVIV model 420 CD spectrometer (Aviv Biomedical). Near-ultraviolet CD and tryptophan fluorescence experiments were performed on a Jasco J-1500 CD spectrometer (Jasco). Protein concentrations were determined by 280 nm absorbance with a NanoDrop spectrophotometer (Thermo Scientific). Melting temperatures were estimated by smoothing the sparse data with a Savitsky–Golay filter of order 3 and approximating the smoothed data with a cubic spline to compute derivatives. Reported T_m values are the inflection points of the melting curves.

Size-exclusion chromatography with multi-angle light scattering. Protein samples were prepared at 1–3 mg/ml and applied to a Superdex 75 10/300 GL column (GE Healthcare) on a LC 1200 Series HPLC machine (Agilent Technologies) for size-based separation, and a miniDAWN TREOS detector (Wyatt Technologies) for light-scattering signals.

Fluorescence binding assay. Protein-activated DFHBI fluorescence signals were measured in 96-well plate format (Corning 3650) on a Synergy neo2 plate reader (BioTek) with $\lambda_{ex} = 450$ nm or 460 nm and $\lambda_{em} = 500$ nm or 510 nm. Binding reactions were performed at 200 μl total volume in PBS pH 7.4 buffer. Protein concentrations were determined by 280 nm absorbance as described above. DFHBI (Lucerna) was resuspended in DMSO as instructed to make 100 mM stock and diluted in PBS to 0.5–10 μM.

Library construction. Deep mutational scanning library for b11L5F were constructed by site-directed mutagenesis as described[40]. One hundred and eleven PCR reactions were carried out using DNA oligonucleotides directed to each position in two 96-well polypropylene plates (USA Scientific, 1402-9700), and products were pooled and purified by gel extraction kit (Qiagen) for yeast transformation. Combinatorial libraries for b11L5F.1 and b11L5F.2 were assembled using synthesized DNA oligonucleotides (Integrated DNA technologies) as described[41]. Selected positions were synthesized with 1–2% mixed bases to control mutation rate and library size. Full-length assembled genes were amplified and purified for yeast transformation as described[42].

Yeast surface display and fluorescence-activated cell sorting. Transformed yeast cells (strain EBY100)[42] were washed and re-suspended in PBSF (PBS plus 1 g/l of BSA). DFHBI in DMSO stock was diluted in PBSF for labelling yeast cells at various concentrations. PBSF-treated cells were incubated with DFHBI for 30 min to 1 h

at room temperature on a benchtop rotator (Fisher Scientific). Library selections were conducted using the GFP fluorescence channel at 520 nm with 488-nm laser on a SH800 cell sorter (Sony). Proteolysis treatment and fluorescence labelling were performed in the same way as described[30]. Cell-sorting parameters and statistics for all selections are given in Supplementary Table 16.

Deep sequencing and data analysis. Pooled DNA samples for b11L5F deep mutational scanning library were transformed twice to obtain biological replicates. Two libraries were treated and sorted in a parallel fashion. Yeast cells of naive and selected libraries were lysed and plasmid DNA was extracted as described[43]. Illumina adaptor sequences and unique library barcodes were appended to each library by PCR amplification using population-specific primers (see Supplementary Table 8). DNA was sequenced in paired-end mode on a MiSeq Sequencer (Illumina) using a 300-cycle reagent kit (Catalogue number: MS-102-3003). Raw reads were first processed using the PEAR program[44] and initial counts analysed with scripts adapted from Enrich[45]. Stability scores were modelled using sequencing counts from proteolysis sorts as described[30]. Unfolded states were modelled without disulfide bonds (cysteine replaced with serine). Function scores were modelled using sequencing counts from DFHBI fluorescence sorts. A simple meta-analysis statistical model with a single random effect was applied to combine two replicates using the framework developed in Enrich2[46].

BB1 crystal structure. BB1 protein was concentrated to 20 mg/ml in an AMICON Ultra-15 centrifugation device (Millipore), and sequentially exchanged into 20 mM Tris pH 8.0 buffer. The initial screening for crystallization conditions was carried out in 96-well hanging drop using commercial kits (Hampton Research and Qiagen) and a mosquito (TTP LabTech). With additional optimization, BB1 protein crystallized in 0.1 M BIS-Tris pH 5.0 and 2 M ammonium sulfate at 25 °C by hanging drop vapour diffusion with 2:1 (protein:solution) ratio. Diffraction data for BB1 was collected at 200° with 1° oscillations, 5-s exposures, at the Advanced Light Source (Berkeley) beamline 5.0.1 on an ADSC Q315R area detector, at a crystal-to-detector distance of 180 mm. The data was processed in space group P2$_1$ to 1.63 Å using Xia2[47]. The BB1 design model was used as a search model for molecular replacement using the program Phaser[48], which produced a weak solution (TFZ 6.5). From this, a nearly complete model was built using the Autobuild module in Phenix[49]. This required the rebuild-in-place function of autobuild to be set to 'False'. Iterative rounds of model building in the graphics program Coot[50] and refinement using Phenix.refine[51] produced a model covering the complete BB1 sequence. Diffraction data and refinement statistics are given in Supplementary Table 18.

b10, b11L5F_LGL crystal structure and mFAPs–DFHBI co-crystal structures. b10 was initially tested for crystallization via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech). Crystallization conditions were then optimized in larger 24-well hanging drops. b10 crystallized in 100 mM HEPES pH 7.5 and 2.1 M ammonium sulfate at a concentration of 38 mg/ml. The crystal was transferred to a solution containing 0.1 M HEPES pH 7.5 with 3.4 M ammonium sulfate and flash-frozen in liquid nitrogen. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

b11L5F_LGL was concentrated to 19.6 mg/ml (1.58 mM), incubated at room temperature for 30 minutes with 1 mM TCEP then mixed with an excess of DFHBI (re-suspended in 100% DMSO). b11L5F_LGL complexed with DFHBI was screened via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech) and crystallized in 100 mM Bis-Tris pH6.5 and 45% (v/v) Polypropylene Glycol P 400. The crystal was flash-frozen in liquid nitrogen directly from the crystallization drop. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

mFAP0 and mFAP1 were mixed with excess DFHBI (re-suspended in 100% DMSO), while keeping the final DMSO concentration at less than 1%. The mFAP0 and mFAP1 complexes were then concentrated to approximately 41 mg/ml and 64 mg/ml, respectively, and initially tested for crystallization via sparse matrix screens in 96-well sitting drops using a mosquito (TTP LabTech). Crystallization conditions were then optimized in larger 24-well hanging drops macroseeded with poor-quality crystals obtained in sitting drops. mFAP0 complexed with DFHBI crystallized in 200 mM sodium chloride, 100 mM HEPES pH 7.5 and 25% (w/v) polyethylene glycol 3350. The crystal was transferred to the mother liquor plus 2 mM DFHBI and 10% (w/v) polyethylene glycol 400 then flash-frozen in liquid nitrogen. Data were collected at the Berkeley Center for Structural Biology at the Advanced Light Source (Berkeley), on beamline 5.0.2 at a wavelength of 1.0 Å. and processed in HKL2000[52]. mFAP1 complexed with DFHBI crystallized in 100 mM MES pH6.5 and 12% (w/v) polyethylene glycol 20,000. The crystal was transferred to the mother liquor plus 2 M DFHBI and 15% glycerol then flash-frozen in liquid nitrogen. Data were collected with a home-source rotating anode on a Saturn 944+ CCD and processed in HKL2000[52].

Structures were solved by molecular replacement with Phaser[48] via phenix[49] using the Rosetta design model with appropriate residues cut back to C$_\alpha$ and DFHBI removed. The structure was then built and refined using Coot[50] and

phenix[51], respectively, until finished. Diffraction data and refinement statistics are given in Supplementary Table 18.

Statistics and reproducibility. In Fig. 1c, the models were coloured on the basis of the mean values of repulsion energy by position (Rosetta fa_rep) derived from a set of poly-valine backbones relaxed with constraints ($n = 189$ independently generated models); relaxed with constraints with a glycine in the middle of each C_β-strip ($n = 186$ independently generated models) and relaxed without constraints with glycines and β-bulges ($n = 194$ independently generated models). This experiment has been performed twice on different sets of backbones and produced similar results. In Fig. 2b–e, BB1 was purified and sized with SEC at least five times independently, yielding different ratios of monomeric-to-oligomeric species (20–75%). The fraction of monomer could be increased by heat-shocking the cells at 42 °C shortly before induction. Two biological replicates of the far- and near-ultraviolet CD and tryptophan fluorescence spectra acquisition of BB1 were done with similar results, and the chemical denaturation experiment performed once. Extended Data Fig. 4a–c, the analysis of BB1 with SEC–MALS was repeated twice on independently prepared protein samples and similar molecular masses were obtained. Additionally, the experiments were repeated on one sample stored at 4 °C at different time points ($t = 0$, $t = 7$ days and $t = 30$ days); all experiments had similar results and confirmed the stability of the monomeric species. BB2, 3 and 4 were purified once. The molecular mass (with SEC–MALS) and the far-ultraviolet CD spectra of the purified proteins were tested once. The sizing of purified BB1 mutants was performed once, with wild-type BB1 as an internal control.

Cell culture and transfection. COS-7 cells (ATCC CRL-1651, confirmed negative for mycoplasma) were grown in DMEM supplemented with $1\times$ NEAA, 100 units/ml penicillin, 100 μg/ml streptomycin, and 10% FBS; and collected using 0.25% trypsin EDTA. Per transfection, approximately one million cells were transfected with 2 μg of plasmid using 18 μl of Lonza s.e. cell supplement, 82 μl of Lonza s.e. nucleofection solution and pulse code DS-120 on a Lonza 4D X Nucleofector system. After nucleofection cells were immediately seeded into ibidi μ-Slide eight-well glass bottom chambers at a density of ~30,000 cells/well and incubated overnight at 37 °C.

Cell fixation. Cells were fixed at 37 °C for 10 min in PFA/GA fixation solution containing 100 mM aqueous PIPES buffer pH 7.0, 1 mM EGTA, 1 mM $MgCl_2$, 3.2% paraformaldehyde, 0.1% glutaraldehyde; reduced for 10 min with freshly prepared 10 mM aqueous sodium borohydride; then rinsed with PBS for 5 min.

Microscopy. Conventional widefield epifluorescence imaging was performed on an inverted Nikon Ti-S microscope configured with a 60×1.2 NA water-immersion objective lens (Nikon), a light emitting diode source (LED4D120, Thorlabs), a multiband filter set (LF405/488/532/635-A-000, Semrock) and images were captured with a Zyla 5.5 sCMOS camera (Andor). The samples were illuminated 470 nm light at an intensity of ~2 W/cm² and with 200 ms exposures. For live-cell experiments, samples were incubated at 37 °C with Gibco CO_2 Independent Medium containing 50 μM DFHBI for 10 min before imaging. Time-lapse movies were acquired over a period of 5 min with a 200-ms exposure every 5 s. For fixed cell imaging, samples were incubated at room temperature (~22 °C) in PBS containing 50 μM DFHBI for 10 min before imaging.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The Rosetta macromolecular modelling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available via the University of Washington Technology Transfer Office. Design protocols and analysis scripts used in this paper are available in the Supplementary Information and at https://doi.

org/10.5281/zenodo.1216229. The source code for RIF docking implementation is freely available at https://github.com/rifdock/rifdock.

Data availability. The atomic coordinates and experimental data of BB1, b10, b11L5F_LGL, mFAP0–DFHBI, and mFAP1–DFHBI crystal structures have been deposited in the RCSB Protein Database with the accession numbers of 6D0T, 6CZJ, 6CZG, 6CZH and 6CZI respectively. All the design models, Illumina sequencing data, sequencing analysis and source data (Figs. 2, 4, Extended Data Figs. 6e, 7, 8a, h) are available at https://dx.doi.org/10.5281/zenodo.1216229.

34. Gront, D., Kmiecik, S. & Kolinski, A. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.* **28**, 1593–1597 (2007).
35. Huang, P.-S. et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *Nat. Protocols* **6**, e24109 (2011).
36. Park, H., DiMaio, F. & Baker, D. The origin of consistent protein structure refinement from structural averaging. *Structure* **23**, 1123–1128 (2015).
37. Davis, I. W. & Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* **385**, 381–392 (2009).
38. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
39. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6**, 551–552 (2009).
40. Procko, E. et al. Computational design of a protein-based enzyme inhibitor. *J. Mol. Biol.* **425**, 3563–3575 (2013).
41. Thyme, S. B. et al. Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Res.* **42**, 2564–2576 (2014).
42. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nat. Protocols* **1**, 755–768 (2006).
43. Whitehead, T. A. et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).
44. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
45. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).
46. Rubin, A. F. et al. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).
47. Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).
48. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
49. Adams, P. D. et al. in *International Tables for Crystallography, Volume F*, 2nd edition (eds Arnold, E. et al.) Ch. 18.11, 539–547 (Wiley, Hoboken, 2012).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
51. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
52. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276 (Ed. Carter, W. C.) Ch. 3, 307–326 (Academic, Cambridge, 1997).
53. Merkel, J. S. & Regan, L. Aromatic rescue of glycine in β sheets. *Fold. Des.* **3**, 449–456 (1998).
54. Shaner, N. C., Steinbach, P. A. & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nat. Methods* **2**, 905–909 (2005).
55. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).
56. Hauser, C. A. E. et al. Natural tri- to hexapeptides self-assemble in water to amyloid β-type fiber aggregates by unexpected α-helical intermediate structures. *Proc. Natl Acad. Sci. USA* **108**, 1361–1366 (2011).
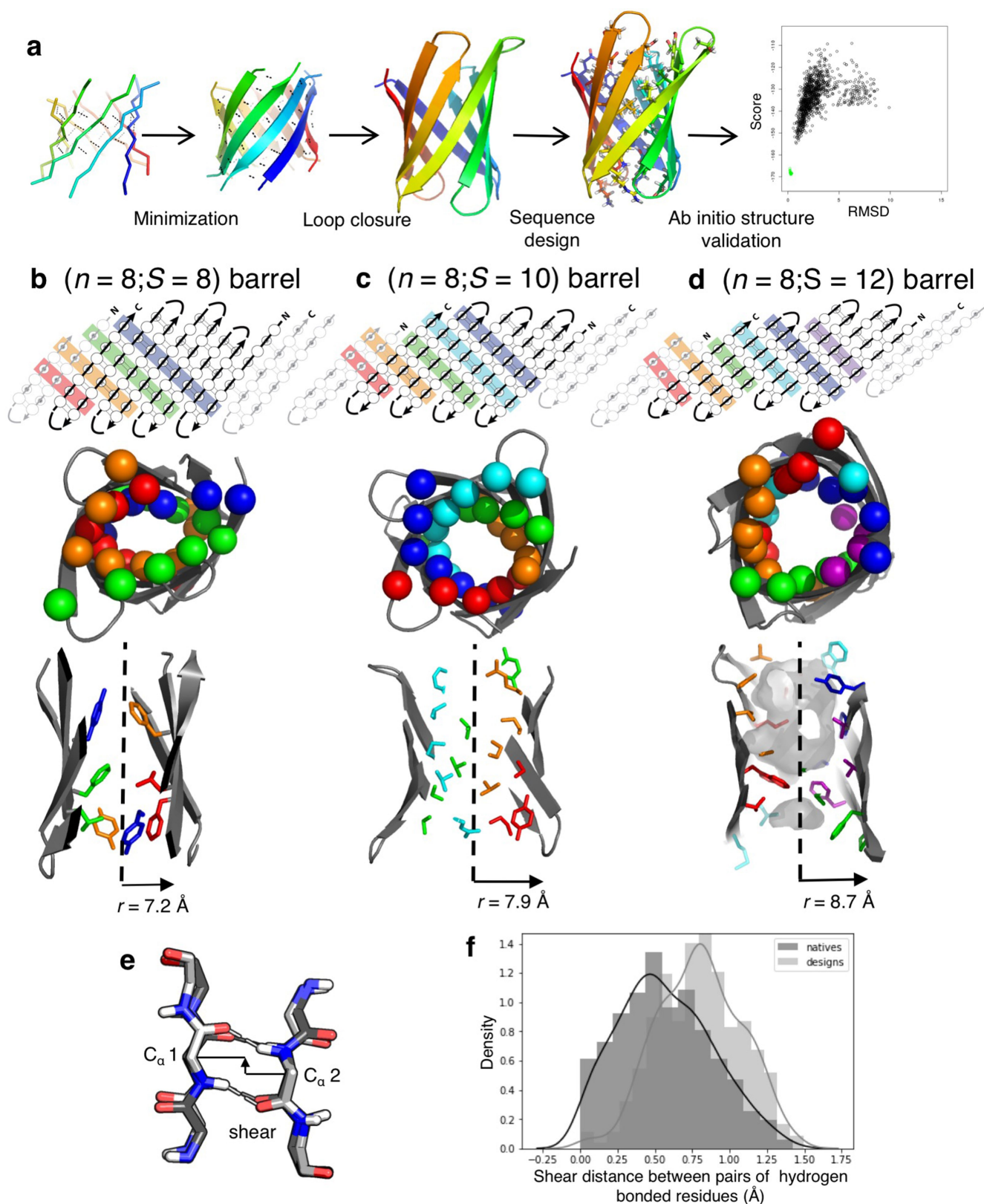
Extended Data Fig. 1 | Parametric design: workflow and shortcomings.
a, Schematic of the parametric approach to generate β-barrel designs.
b–d, Comparison between β-barrels of type ($n = 8$, $S = 8$) (**b**), type ($n = 8$, $S = 10$) (**c**) and type ($n = 8$, $S = 12$) (**d**); showing an example of 2D map with residue connectivity (top), the arrangement of the C_β atoms in the C_β-strips (middle) and the packing pattern of the core side chains (bottom). Although the three β-barrels have the same number of strands

n, the difference in shear number S translates into different overall strand staggering, barrel radii (r) and number of C_β-strips. The number of core C_β-strips (top, middle) results in different arrangements of side chains in the core of the barrel. **e, f**, The parametric designs exhibited distorted hydrogen bonds, reflected by the shear distance (defined in **e**) between C_α atoms of paired antiparallel β-strand residues. The shear distance in the designs deviates from the distribution observed in native β-sheet proteins (**f**).
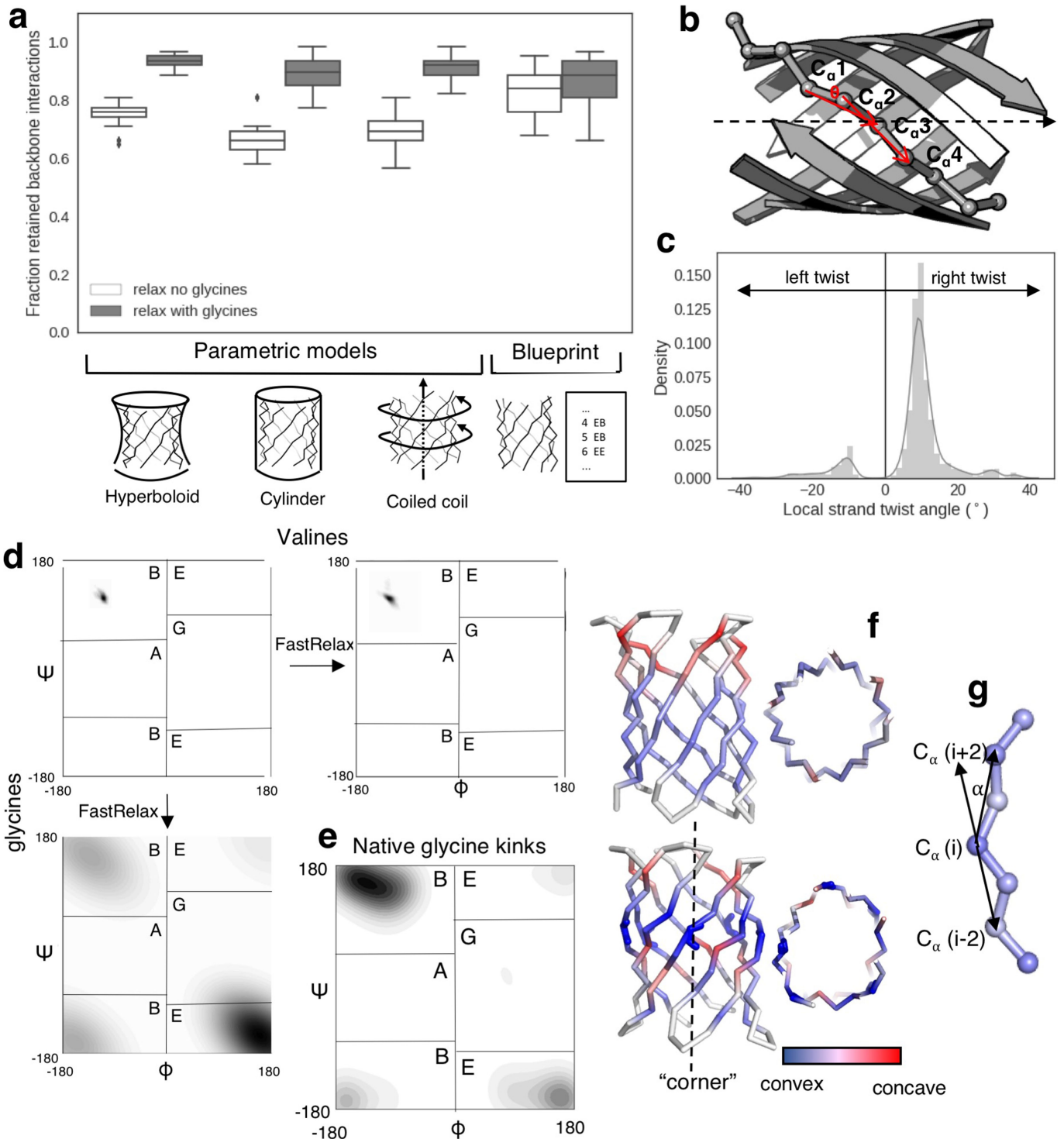
Extended Data Fig. 2 | Glycine kinks release strain in β-barrel backbones. a, Fraction of retained hydrogen-bond interactions after relaxation with Rosetta ('relax') of uniform poly-valine backbones (white) and poly-valine backbones with a glycine in the middle of each C_β-strip (grey). We compare disconnected strand arrangements generated with the parametric hyperboloid model ($n = 225$ independently generated models), the cylindric model ($n = 36$ independently generated models), the coiled-coil model ($n = 150$ independently generated models) and assembled on the basis of a 2D map ($n = 144$ independently generated models). For all box plots: centre line, median; box limits, upper and lower quartiles; whiskers, minimum and maximum values; points, outliers. **b, c**, In poly-valine backbones ($n = 189$ independently generated models) relaxed with constraints to maintain hydrogen bonds between strands, several residues have unfavourable left-handed twist (**c**). The local strand twist is calculated on a sliding window of four residues along β-strands, as the angle between the vectors $C_\alpha 1$–$C_\alpha 3$ and $C_\alpha 2$–$C_\alpha 4$. The handedness of the twist is defined as the triple scalar product between these two vectors and the central axis of the barrel. Positive and negative values denote right-handed and left-handed twist, respectively (**b**). **d**, After relaxation ('FastRelax'), the valine positions in the middle of each C_β-strip remained in the β-sheet-specific ABEGO space (right); or were shifted towards the positive Φ space (E ABEGO) if mutated to glycines (bottom). **e**, A similar torsion angle distribution was observed for glycines in the β-strands of native β-barrels ($n = 35$ high-resolution crystal structures). **f**, In comparison with regular β-strands (top), the presence of glycine kinks (bottom) increases the local bending of the strands and creates corners in an otherwise circular barrel cross-section. **g**, The bending angle α is calculated on a sliding window of three residues.
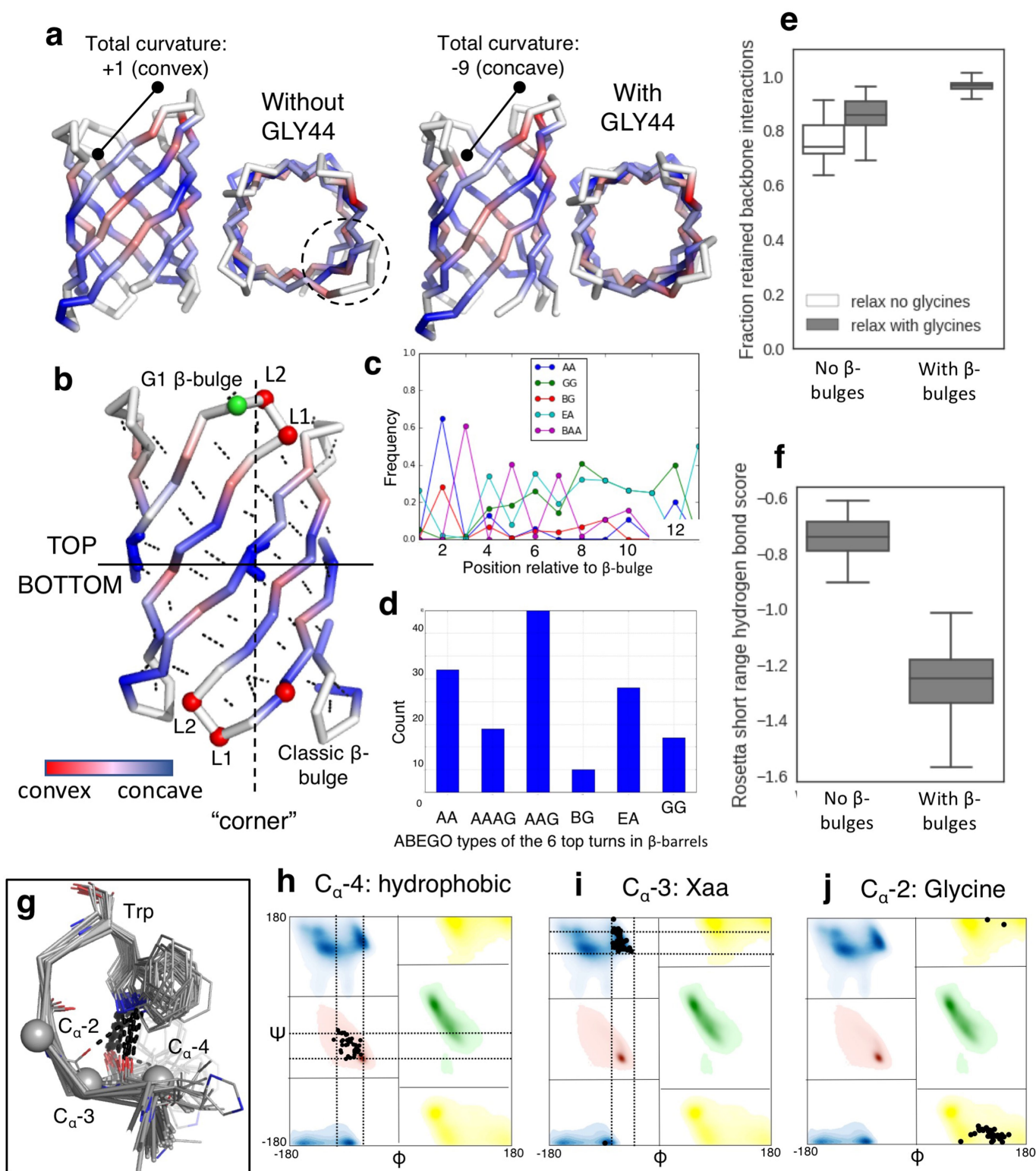
Extended Data Fig. 3 | Placement of β-bulges, β-turns and the tryptophan corner. a, Change of curvature (from convex to concave) and protrusion (dashed circle) of the longest hairpin associated with the placement of a glycine kink at position 44. **b**, Relationship between the 'corners' in the β-sheet (dashed line) generated by the glycine kinks and the type and position of the β-bulges and β-turns (Supplementary Methods). C_α are shown as spheres and coloured by ABEGO type. The bottom of the barrel was defined as the side of the N and C termini. **c**, The type I β-turn ('AA' ABEGO type) is frequently found at the second position relative to a β-bulge in native proteins and was selected to connect bottom hairpins. **d**, This choice is further supported by the enrichment of type I

(AA) turns over the canonical type I′ turn (GG) in native β-barrels ($n = 35$ high-resolution crystal structures). **e**, **f**, Poly-valine backbones built with β-bulges and the corresponding β-turns ($n = 194$ independently generated models) retain more hydrogen bonds after relaxation than backbones built without β-bulges and with canonical type I′ β-turns ($n = 186$ independently generated models) (**e**) and exhibit better-scored hydrogen bonds per β-strand residue flanking the β-turns (**f**). **g**, Superposition of tryptophan corner motifs ($n = 41$ high-resolution crystal structures) extracted from native β-barrels. **h–j**, Amino acid preference and torsional constraints derived from the set and used to model the tryptophan corner. Bounded constraints limits are shown as dashed lines.

a

Design ID*	E-value†	*E.coli* Expression	Solubility	SEC	β CD spectrum
BB1	0.1100	yes	yes	**monomer**	yes
BB2	1.1000	yes	yes	tetramer	yes
BB3	1.9000	yes	no		
BB4	0.5600	yes	yes		no

b

Design = BB2   Design = BB4

c

Design = BB1

	Peak 1	Peak 2	Peak 3
Mw	1.044×10^5 ($\pm 8.643\%$)	5.739×10^4 ($\pm 3.655\%$)	1.487×10^4 ($\pm 21.399\%$)

Design = BB2

	Peak 1	Peak 2	Peak 3
Mw	1.344×10^5 ($\pm 2.528\%$)	6.380×10^4 ($\pm 1.699\%$)	1.639×10^4 ($\pm 31.873\%$)

d

e

f

MD-refined

Gray: Xtal
Cyan: Design model

g

	λ_{abs}(nm)*	λ_{ex}(nm)*	λ_{em}(nm)*	Extinction Coefficient ($M^{-1}cm^{-1}$)	Brightness (extinction coefficient × QY)	Absolute Quantum Yield†	Relative Quantum Yield†	$K_D(\mu M)$
mFAP1-DFHBI	485	487	506	49,300	986 (34× dimmer than eGFP)	0.020	0.022	0.56 ± 0.07
mFAP2-DFHBI	476	480	500	43,200	907 (37× dimmer than eGFP)	0.021	0.021	0.18 ± 0.05
DFHBI	418	-	-	31,935	-	0.001	0.0007‡	
eGFP§		488	507	56,000	33,600		0.6	
Y-FAST:HBR‖	467		527	44,000	3,960		0.090	0.62 ± 0.05

Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Biochemical and structural characterizations of designs BB1–4. a, Results of experimental characterization of the nonfunctional designs (BB1–4). Reproducibility is described in the Methods. [†]E value is calculated by BLAST, the non-redundant protein database. **b**, Far-ultraviolet CD spectra of designs BB2 and BB3 at 25 °C. **c**, SEC–MALS analysis showed a major monomer peak for BB1 and a major tetramer peak for BB2. **d**, Variants of BB1 with residues of the tryptophan corner and glycine kinks mutated to alanine were purified and sized. SEC traces are superimposed on the SEC trace of wild-type BB1 (WT). The mutations of all residues of the tryptophan corner eliminate the monomeric peak. Most of the glycine kink mutations negatively affect the monomeric species. The exceptions are Gly53 and Gly55, which are next to each other on the fourth strand. One glycine kink per strand might be sufficient to introduce enough negative twist to remove strain in the β-barrel. **e–f**, Deviations between BB1 design model and crystal structure. **e**, One of the three bottom turns of the crystal structure (grey) deviates from the design model (magenta) and forms additional crystal contacts (indicated by a dashed circle). **f**, Three phenylalanine side chains have

different rotameric states. In the crystal structure, Phe41 interacts with Gly53 (which shows the most backbone deviation between the crystal structure and the design) to form an aromatic rescue motif[53]. It is likely that the discrepancies in the Phe rotamers reflect a scoring and sampling challenge to accurately capture such aromatic rescue; molecular dynamics simulation starting from the crystal structure (cyan) was also unable to recover the correct Phe41–Gly53 interaction. **g**, Biophysical properties (absorbance or fluorescence spectra, quantum yield and binding affinity) of mFAP1 and mFAP2 in complex with DFHBI. Mean values from three biological replicates were used for the nonlinear regression to determine the K_D. The error estimates are the standard deviation from the fitting calculation. *λ_{abs} is peak absorbance wavelength, λ_{ex} is peak excitation wavelength and λ_{em} is peak emission wavelength. [†]Absolute quantum yield is measured with an integrating sphere; relative quantum yield is measured using acridine yellow and fluorescein as the standards. [‡]Previously reported value[26]. [§]Taken from previously published work[54]. [‖]Taken from previously published work[31].
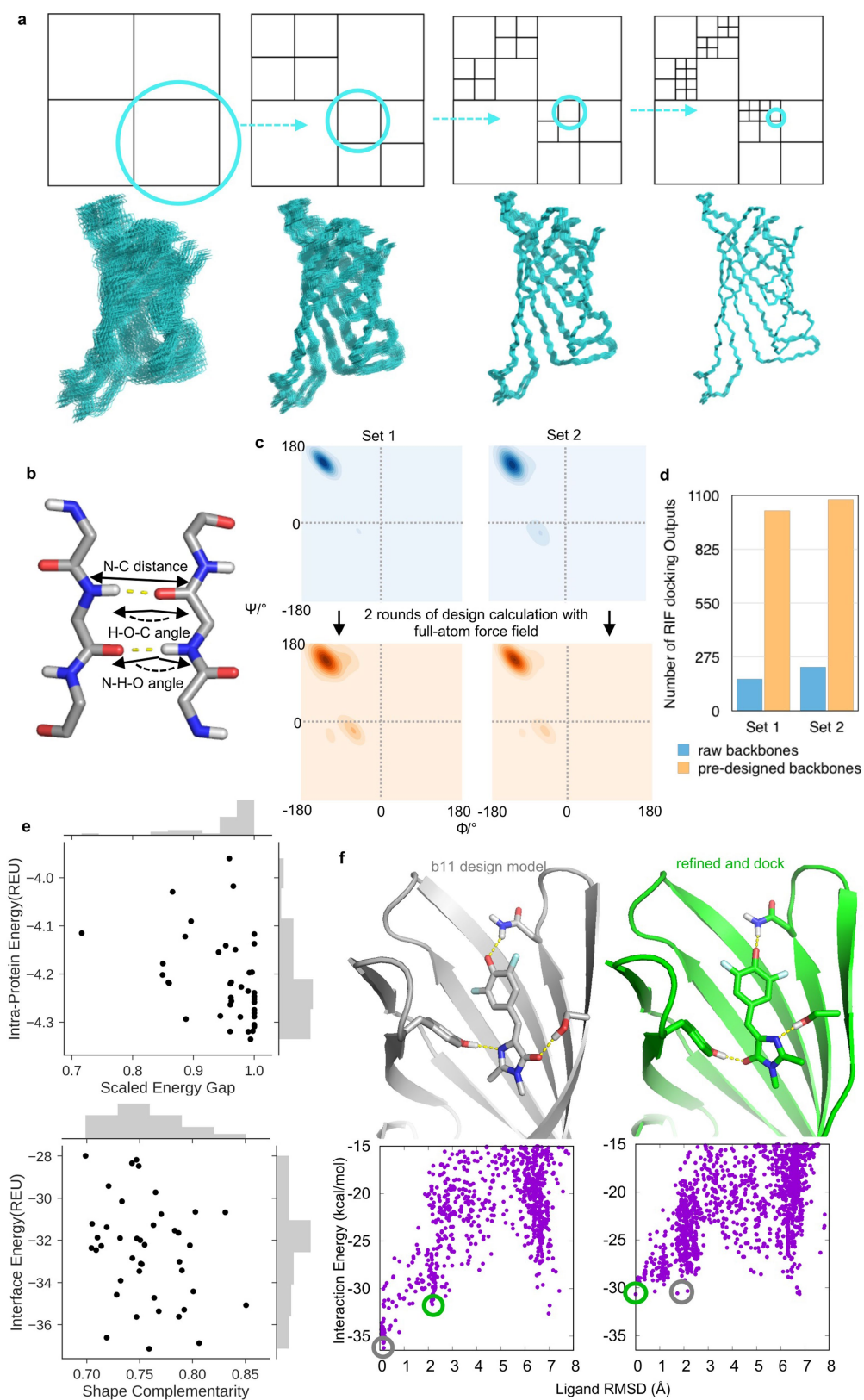
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | RIF docking grid-based search algorithm, β-barrel scaffold construction and post-design ligand-docking simulations. a, Illustration of grid-based hierarchical search strategy in RIF docking. After generating an ensemble of interactions for the target ligand (Fig. 3), each one of the selected scaffold is docked into the fixed RIF using the grid-based hierarchical searching algorithm. This search procedure starts from coarse sampling grids to fine sampling grids in 3D space. An example 2D grid scheme is shown in the upper row, from the lowest resolution (coarse sampling, left) to the highest resolution (fine sampling, right). At each searching stage, the backbone is assigned to different grids on the basis of its relative position and the resulting docking configurations are scored. The top-scored backbone positions (highlighted by cyan circles in the 2D scheme) are shown as 3D structures in the lower row for each searching resolution and are continued for the next grid search and scoring. The 3D structure example shown here was streptavidin structure (PDB ID: 1STP) with grid searching resolutions of 8.0 Å, 4.0 Å, 2.0 Å, and 1.0 Å. **b–d**, β-barrel scaffold construction for small-molecule binding. Three geometric constraints (**b**) were used to describe each backbone hydrogen bond and drive the backbone assembly during Rosetta low-resolution centroid modelling. Backbones generated with all three constraints had a very narrow Φ/Ψ distribution as a result

of strong constraints (**c**, Ramachandran plot in upper left, set 1, density coloured in blue); by omitting N–H–O angle constraint, backbone torsion diversity slightly improved (**c**, upper right, set 2). These two raw backbone sets yielded few non-redundant RIF docking solutions (**d**, blue bars). After two rounds of sequence design calculation using Rosetta full-atom force field (Supplementary Methods), regularized backbones (peptide bonds with proper dihedral geometry) and broadened Φ/Ψ distribution (**c**, Ramachandran plot in the lower row, density coloured in orange) yielded more unique RIF docking solutions (**d**, orange bars). **e**, Computed metrics for 42 designs ordered and tested. Results from ab initio folding simulation were scaled to 0.0 to 1.0, in which 1.0 represents a funnel-shaped folding landscape[55]. **f**, Alternative ligand-binding conformations revealed by post-design ligand-docking simulations. The lowest-energy docking conformation using the design model (by simply taking out the ligand from the pocket) was similar to the designed DFHBI-binding mode (top left, grey; designed binding mode was circled in grey in the energy landscape in the lower row). Docking simulations using an apo-protein model refined by molecular dynamics simulations revealed an alternative equal-energy docking conformation (top right, green) that is indicated by a green circle in the docking energy landscapes (bottom). Both binding modes rely on three hydrogen-bonding residues from RIF docking (top).
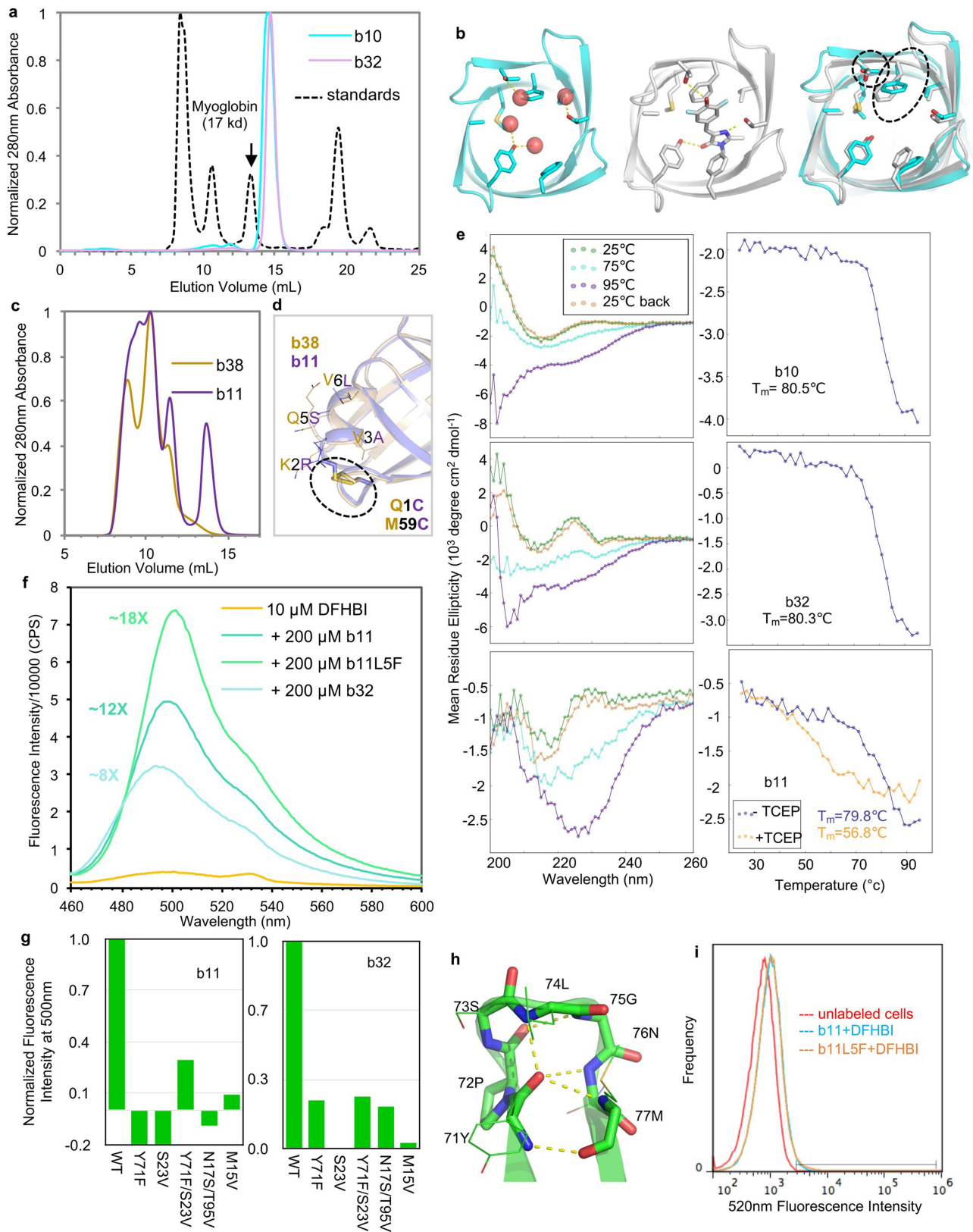
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Biochemical and structural characterization of design b10, b32 and b11. a, Size-exclusion chromatogram of His6-tagged b10 and b32 after Ni-NTA affinity purification. The monodispersed peaks of absorbance at 280 nm of b10 and b32 (cyan and lavender, respectively) have an elution volume compatible with the monomeric β-barrel (14 kDa), on the basis of their relative position to the protein standard peaks (dashed line). Biological replicates were performed with similar observation: $n = 4$ for b10, $n = 5$ for b32. **b**, Comparison of the ligand-binding pocket in the b10 design model (middle, grey) with the crystal structure (left, cyan). The side chain disagreements are highlighted with a dashed black circle on the right panel. **c, d**, The designed disulfide bond as a stabilizing mechanism. SEC curves of His6-tagged b11 (purple line) and b38 (dark yellow line) were overlaid to show the appearance of a monomer peak for b11 (the same standard as in **a** was applied here). A disulfide bond connecting the N-terminal helix to a β-strand (Q1C and M59C, circled in **d**) and four mutations of neighbouring residues were introduced into design b38 (dark yellow) to make design b11 (purple). Biological replicates were performed with similar observation: $n = 3$ for b38, $n = 5$ for b11. **e**, Far-ultraviolet CD spectra of b10, b32 and b11. Left, spectra at different temperatures within one heating–cooling cycle. Right, thermal melting curves (the CD signal of b10 was monitored at 220 nm; b32 and b11 at 226 nm). b11 probably forms an amyloid-like beta structure at 95 °C (left, bottom row) with a negative peak around 226 nm[56] and refolds back after cooling to 25 °C. The thermal stability of b11 decreases when the disulfide was reduced with 1 mM tris(2-carboxyethyl) phosphine (TCEP) (right, bottom). Measurements were performed once for each design ($n = 1$). **f**, Fluorescence emission spectra of b32, b11 and b11L5F in complex with DFHBI. With 200 μM proteins, b32, b11 and b11L5F can activate 10 μM DFHBI fluorescence by 8-, 12- and 18-fold, respectively. Two biological replicates were performed with similar results. **g**, The residues designed to interact with DFHBI contribute to b11 and b32 activity. Single or double knockouts of hydrogen-bonding residues (Y71, S23, N17 and T95) and a hydrophobic-packing residue (M15) showed decreased fluorescence intensity at 500 nm in comparison with the wild-type b11 or b32 (WT). Mutants were purified once for activity measurement. **h, i**, Re-designed five-residue fifth turn in b11L5F. The original bulge-containing 'AAG' β-turn in b11 (Extended Data Fig. 3b) was redesigned into a five-residue turn. b11L5F was detected by yeast surface display and flow cytometry (**i** and Supplementary Data). Yeast cells displaying b11 and b11L5F showed an increased 520-nm fluorescence signal (excited by 488-nm laser, **i**). Three biological replicates were performed with similar observation.
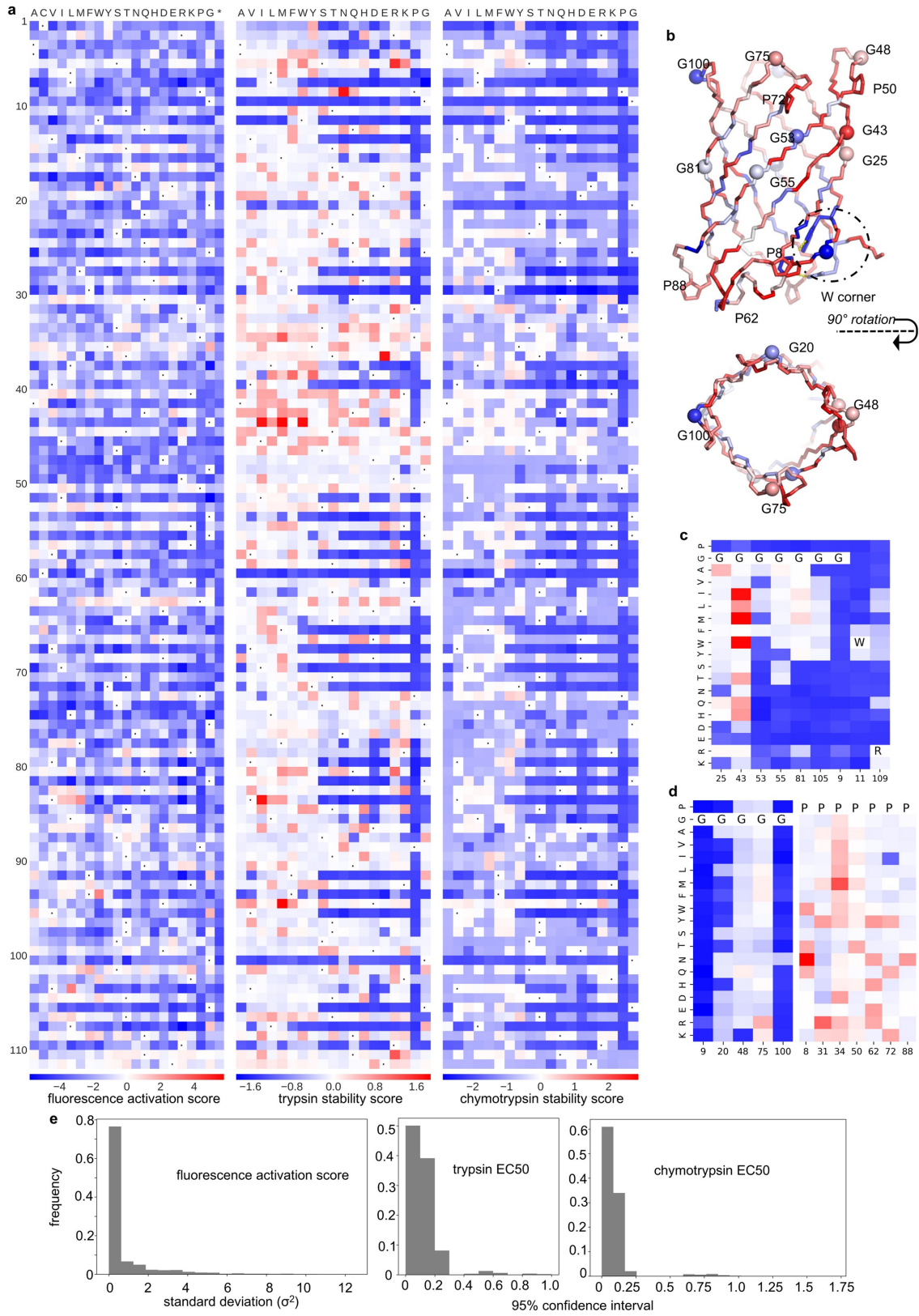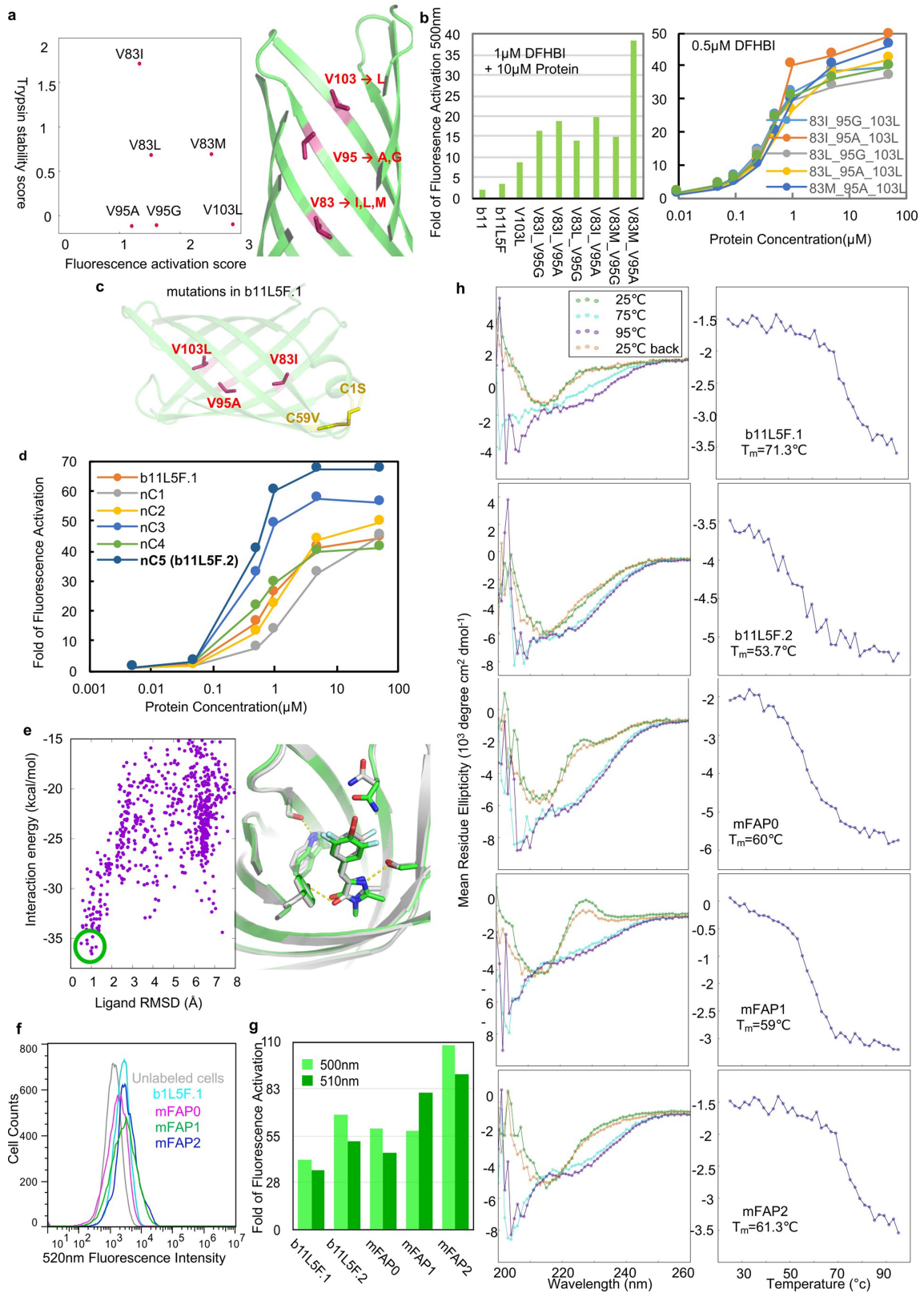
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Deep mutational scanning maps for b11L5F.
a, The complete function (left) and protease stability (middle and right) landscapes of b11L5F. Fluorescence activation scores, trypsin and chymotrypsin stability scores were calculated as described in Supplementary Methods and demonstrated in the Supplementary Data (b11L5F_DMS_analysis.ipy). Data are from two biological replicates with more than tenfold sequencing coverage. Red colour represents beneficial effect whereas mutations coloured in blue are detrimental (relative to the wild-type b11L5F). Wild-type residues at each position are indicated by black dots. **b**, b11L5F backbone model coloured one the basis of the average stability scores. Glycine backbone C_α are shown as spheres. **c**, **d**, Mutational scanning maps of glycine kinks (G25, G43, G53, G55, G81 and G105) and tryptophan corner positions (G9, W9 and R109) (**c**), and of glycines in the β-turns and prolines (**d**). **e**, Statistics of the fluorescence activation and stability scores. The standard deviation between the two replicates used for calculating fluorescence activation scores is smaller than two for most of the data points (left); 95% confidence interval calculated for the proteolysis/stability analysis is less than 0.25 for most the experimental protease half maximal effective concentration (EC_{50}) values (middle and right).

Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Experimental and computational improvement on the basis of b11L5F. a–c, Incorporation of point mutations from deep mutational scanning. Beneficial mutations that improve fluorescence activity without compromising protein stability (positive scores relative to wild-type b11L5F; **a**, left, $n = 2$ biological replicates) were mapped onto b11L5F backbone model (**a**, right). **b,** Purified b11L5F variants incorporating those single, double or triple mutations showed consistently improved fluorescence activity. Binding titration curves were obtained for all six possible triple mutants (right, $n = 1$ biological measurement). **c,** b11L5F with V103L, V95A, V83I, C59V and C1S mutations was renamed as 'b11L5F.1'. **d,** Characterization of five designs from the second round of design calculation. Three of the five designs (nC1–5) that were based on b11L5F showed improved binding activities by titrating purified proteins into 0.5 μM DFHBI ($n = 1$ biological sample was used for the measurement). The best variant (nC5) was renamed 'b11L5F.2'. **e,** Ligand-docking simulations with the molecular dynamics-refined apo b11L5F.2.

Energy landscape was plotted by comparing all the docking conformations to the design model (left). The lowest-energy docking conformations (highlighted in green circle) match the design model (right, design mode in silver and docking model in green). **f, g,** Characterization of three best variants (mFAP0–2) from combinatorial library selections. **f,** Yeast cells displaying mFAP proteins incubated with 5 μM DFHBI analysed by flow cytometry (excited by 488-nm laser, $n = 1$ biological sample was used for the measurement with proper controls). **g,** Purified proteins showed up to 100-fold fluorescence activation (5 μM protein + 0.5 μM DFHBI, excited at 450 nm and monitored at 500 nm and 510 nm in a plate reader, $n = 1$ biological measurement). **h,** Far-UV CD characterization of b11L5F.1, b11L5F.2, mFAP0, mFAP1 and mFAP2. Left, spectra at different temperatures within one heating–cooling cycle. Right, thermal melting curves (CD signals were monitored at 226 nm, spectra were recorded once ($n = 1$) with internal noise estimation).
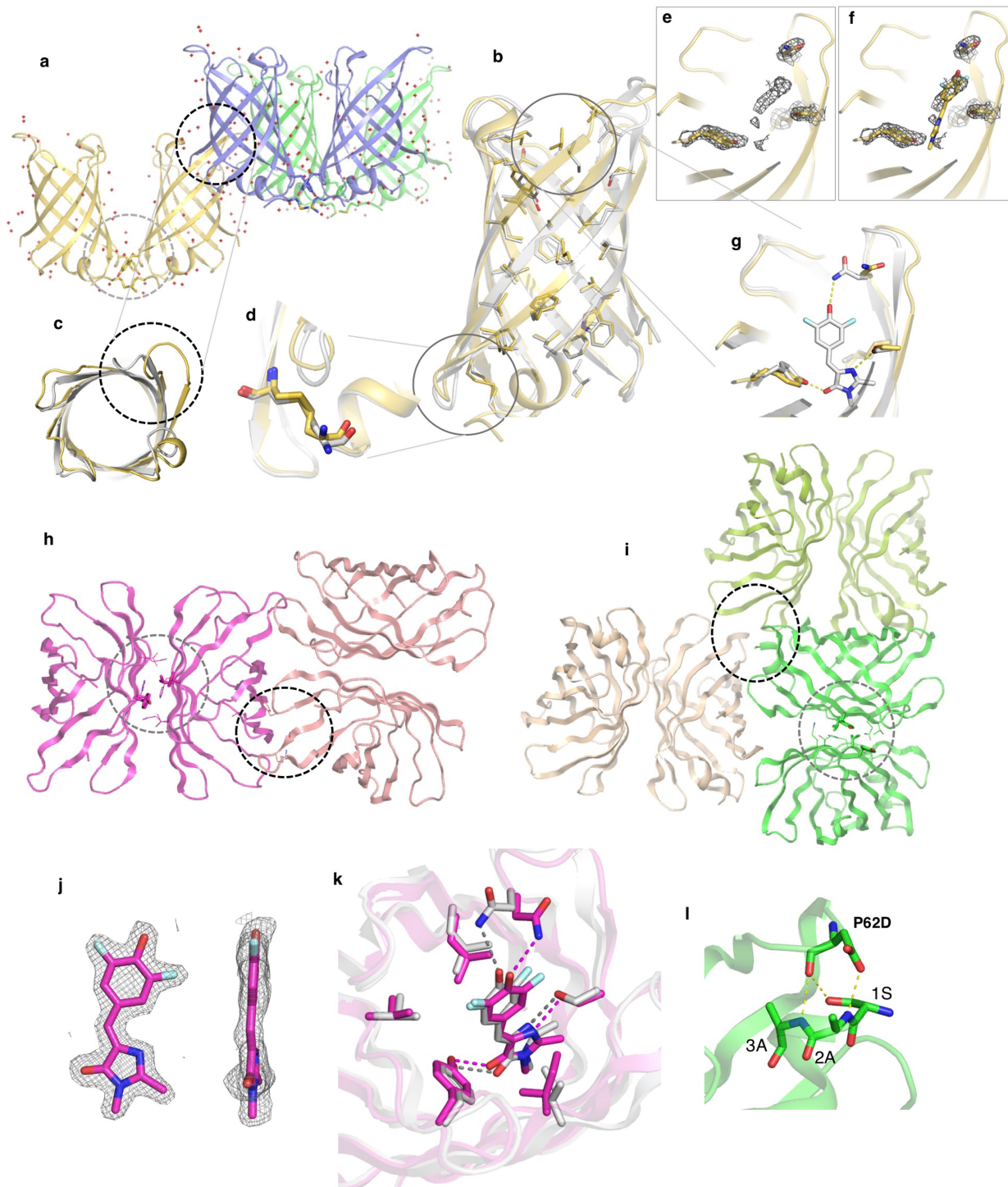
Extended Data Fig. 9 | Crystal structure of b11L5F_LGL, mFAP0 and mFAP1. a–g, b11L5F_LGL crystal structure. Protein samples of all six triple mutants in Extended Data Fig. 8b (right) were prepared for crystallization. b11L5F_LGL with V83L/V95G/V103L combination was successfully crystallized. Crystal contacts between protein copies in one asymmetric unit (yellow) were mediated by two tyrosines (stick representation, grey dashed circle); contacts between three asymmetric units (yellow, blue and green) were formed between β-turns (black dashed circle), which might have displaced one of the top β-turns (**c**). Overall backbone and side chain conformations in the design model matched the crystal structure with a backbone C_α r.m.s.d. of 1.02 Å (**b**, crystal in yellow and design model in silver), and the designed disulfide bond was present in the crystal structure (**d**). Ligand density in the crystal structure was ambiguous: $2F_o - F_c$ omit map showing the electron density after

refinement without placing DFHBI (**e**), the best ligand placement to match the density (**f**), and designed ligand-binding interactions (silver) overlaid with the crystallized binding pocket (**g**). **h, i,** Crystal contacts in the DFHBI-bound structures of mFAP0 (**h**) and mFAP1 (**i**). Contacts between protein copies in one asymmetric unit were formed around 40V and 54Y (grey dashed circle) that were introduced for helping crystallization (Extended Data Fig. 10a). Contacts between asymmetric units were formed between β-turns (black dashed circle). **j,** $2F_o - F_c$ omit electron density of DFHBI in the mFAP0–DFHBI complex crystal structure. DFHBI density contoured at 1.0σ is clear and matches the planar conformation of the ligand (right). **k,** Superposition of mFAP0 design model (silver) and the crystal structure (magenta). Hydrogen bonds are indicated with dashed lines. **l,** Helical capping interactions mediated by P62D mutation in mFAP1 crystal structure.
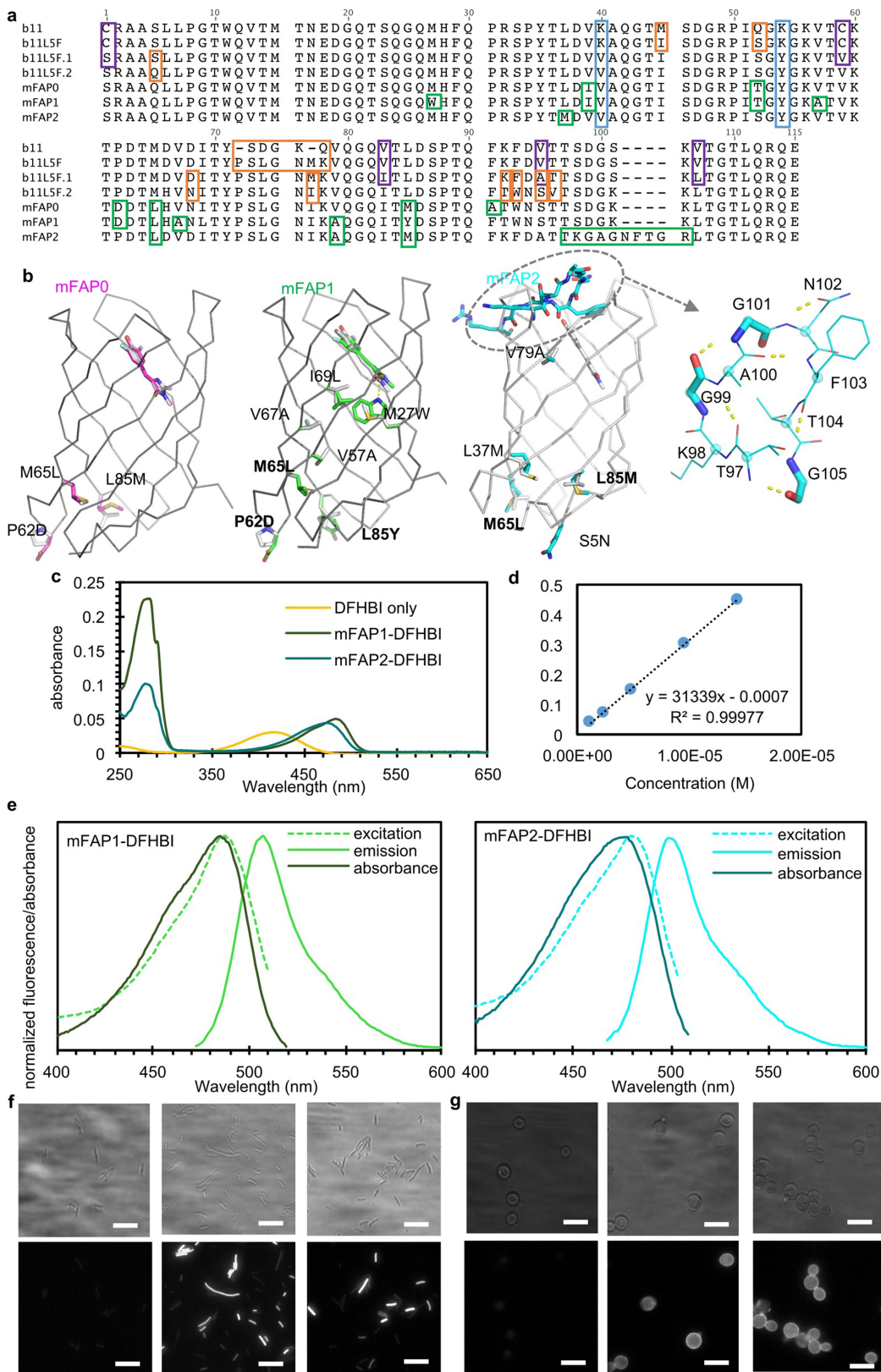
Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Mapping of mutations introduced into b11 to yield the final brighter variants, biophysical characterization of mFAP1 and 2, and epifluorescent images. a, Sequence alignment of b11-based DFHBI-binding fluorescence-activating proteins. Orange boxes indicate mutations or loop insertions introduced by computational design; purple boxes highlight mutations rationally introduced on the basis of the deep mutational scanning maps (Extended Data Figs. 7, 8); green boxes indicate mutations or loop insertions that were incorporated during combinatorial library selections; K40V and K54Y in light blue boxes were introduced to help crystal formation (Extended Data Fig. 9h, i). Despite having hydrophobic residues on the surface, mFAP2 remains soluble at 150 mg ml^{-1}. **b**, Mutations in the mFAPs mapped on the design models. Common mutations in all three mFAPs are highlighted in bold.

c, Absorbance spectra for DFHBI, and the mFAP1–DFHBI and mFAP2–DFHBI complexes ($n = 4$ biological replicates with similar observations). **d**, Extinction coefficient determination for DFHBI at 418 nm. **e**, Normalized absorbance and fluorescence spectra of the mFAP1–DFHBI and mFAP2–DFHBI complex. Data are representative of two biological replicates with similar observations. **f, g**, Widefield epifluorescence (bottom) and brightfield (top) images of *E. coli* and yeast cells with 20 μM DFHBI. Untransformed *E. coli* Lemo21 cells (**f**, left, $n = 2$ biological replicates with similar observation) and yeast EBY100 cells displaying ZZ domain (**g**, left, $n = 2$ biological replicates with similar observation) were treated with the same amount of DFHBI and imaged in the same way (1000 mA 470-nm LED and 200-ms exposure time).

nature research

Corresponding author(s): 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	Rosetta software suite was used to perform protein design calculations; Rosetta is freely available for academic users on GitHub and command lines and scripts are available in Supplementary Information. The custom script written to generate beta-barrel backbones parametrically are available on https://dx.doi.org/10.5281/zenodo.1216229. The BBQ software used to re-constitute backbones from Calpha coordinates is freely available online. The source code for RIF docking implementation is freely available at https://github.com/rifdock/rifdock. FlowJo v8(FlowJo, LLC) was used to read and analyze the flow cytometry data. Open-source softwares PEAR, ENRICH, ENRICH2 were used to analyze the sequencing data.
Data analysis	Custom python and pyrosetta scripts written to analyze the data were included in the Supplementary Data or available on https://dx.doi.org/10.5281/zenodo.1216229. Pyrosetta module is freely available for academic users on GitHub. The commercial FlowJo software (v8) was used to analyze Flow Cytometry results. Crystallographic data were analyzed with Phenix (release 1.101.1-2155) and Coot (v0.8.7 EL).

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

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

The atomic coordinates and experimental data of BB1, b10, b11L5F_LGL, mFAP0-DFHBI, and mFAP1-DFHBI crystal structures have been deposited in the RCSB Protein Database with the accession numbers of 6D0T, 6CZJ, 6CZG, 6CZH, and 6CZI respectively. All the design models, Illumina sequencing data, sequencing analysis and source data (Fig.2 &.4, Extended Data Fig. 6e, 7, 8a&h) are available on https://dx.doi.org/10.5281/zenodo.1216229.

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	Sample size was determined by the estimated work load. In total, 144 designed proteins were experimentally tested. The number of parametric designs experimentally tested (41) was considered sufficient since no significant improvement of the success rate was observed after 4 iteration of design/characterization. The number of characterized designs generated based on the 2D map was considered as sufficient since each round of design calculations yielded folded proteins, and at least some of them had the expected activity or improvement of activity. All computational experiments were carried out on sets of 100-250 models. The size of these sets was chosen to be in the same order of magnitude as the sets used as starting point for design (200-500), while small enough to generate data on our local clusters. In only one case, the size of the set was limited by the possibility to generate enough distinct backbones with that particular parametric model.
Data exclusions	All the data points were reported, for both experimental and computational studies.
Replication	Basic biochemical and structural characterization were performed once or twice with internal statistical validation. The replicated experiments (CD spectra, fluorescence spectra, fluorescence imaging, SEC-MALS, computational experiments) produced similar results. Variability was observed between some purified protein samples in terms of the relative ratio of monomeric/aggregated proteins. Yeast library generation and selection for Miseq sequencing were performed twice.
Randomization	No randomized samples allocation in this work. Because of the complexity of the designs (the design of beta-barrels has been unsuccessfully attempted several times by different groups) and the success rate (20-25% folded proteins) and diversity of successful designs, it is unlikely that the folding amino acids sequences reported in this paper are due to chance rather than design calculation. Negative and positive controls were well defined and validated.
Blinding	Blinding was not relevant to this work since the experiments were well-defined by the computational design.

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

Antibodies

Antibodies used	FIT-conjugated anti-cMyc antibody (chicken) purchased from Immunology Consultants Labs (catalog number: CMYC-45F).
Validation	efficient cell labeling was confirmed by flow cytometry.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	NIH3T3 cells (Flp-In-3T3, Thermo Fisher Scientific); COS-7 cells (ATCC CRL-1651)
Authentication	Not authenticated
Mycoplasma contamination	Confirmed negative for mycoplasma
Commonly misidentified lines (See ICLAC register)	No common misidentified lines were used in this study.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://dx.doi.org/10.5281/zenodo.1216229 (This doi is reserved and will be activate and become public once the publication is finalized.)
Files in database submission	b11L5F_DMS_analysis.zip, b11L5F_DMS_EnrichProcessed.tar.gz, b11L5F_DMS_PearAssembled.tar.gz, b11L5F_DMS_IlluminaRead.tar.gz (detailed descriptions in "List.docx" file.)
Genome browser session (e.g. UCSC)	no longer applicable.

Methodology

Replicates	2 biological replicates
Sequencing depth	300 cycles paired-end reads
Antibodies	FIT-conjugated anti-cMyc antibody (chicken) purchased from Immunology Consultants Labs (catalog number: CMYC-45F); fluoreogenic fluorophore DFHBI from Lucerna.
Peak calling parameters	Illumina's on-instrument data analysis MiSeq Reporter was use to process the fluorescence data.
Data quality	a statistic confidential interval is assigned to each mutation.
Software	open-source softwares PEAR, ENRICH, ENRICH2 were used to analyze the data.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	see Supplementary Methods and Supplementary Table 15
Instrument	Sony SH800
Software	FlowJo v8

Cell population abundance | see Supplementary Table 15

Gating strategy | FSC/SSC was used to gate the size of yeast cells; FSC width signal/ FSC height signal was used to gate the singlet cells.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.