De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy

Po-Ssu Huang[1,2,6]*, Kaspar Feldmeier[3,6], Fabio Parmeggiani[1,2], D Alejandro Fernandez Velasco[4], Birte Höcker[3]* & David Baker[1,2,5]*

Despite efforts for over 25 years, *de novo* protein design has not succeeded in achieving the TIM-barrel fold. Here we describe the computational design of four-fold symmetrical $(\beta/\alpha)_8$ barrels guided by geometrical and chemical principles. Experimental characterization of 33 designs revealed the importance of side chain–backbone hydrogen bonds for defining the strand register between repeat units. The X-ray crystal structure of a designed thermostable 184-residue protein is nearly identical to that of the designed TIM-barrel model. PSI-BLAST searches do not identify sequence similarities to known TIM-barrel proteins, and sensitive profile-profile searches indicate that the design sequence is distant from other naturally occurring TIM-barrel superfamilies, suggesting that Nature has sampled only a subset of the sequence space available to the TIM-barrel fold. The ability to design TIM barrels *de novo* opens new possibilities for custom-made enzymes.

There has been progress in *de novo* design of protein structures[1–8], but designing all-β and α/β barrels has proven very challenging. For designing novel catalysts, the $(\beta/\alpha)_8$-barrel (or TIM-barrel, for triosephosphate isomerase) fold is of particular interest because it is the most common topology for enzymes and one of the most diverse superfolds[9]. The TIM-barrel fold is structurally and functionally diverse, consisting of 33 superfamilies in the Structural Classification of Proteins (SCOP) database and covering five of the six Enzyme Commission reaction classes[9]. As many as 10% of known enzymes may adopt this fold[10], and it has been the focus of intensive enzyme engineering and design efforts[11–16]. The TIM-barrel fold was one of five design targets in a European Molecular Biology Organization workshop on protein design in 1987, but no *de novo* design efforts to date have yielded proteins with clearly defined tertiary structure[17–22]. In particular, the latest designs in the octarellin series have circular dichroism (CD) spectra consistent with an α/β structure, are stable to temperature denaturation and have near-UV CD and fluorescence spectra suggesting that the aromatic residues are in somewhat well-defined environments. However, there is no evidence from crystallography, multidimensional NMR or any other method indicating the formation of a specific tertiary structure, let alone a TIM-barrel fold[20,21]. Here we take a bottom-up design approach using an idealized symmetrical barrel geometry and created a thermostable and reversibly folding TIM barrel.

RESULTS
TIM-barrel design principles
We begin by deriving design principles for an ideal TIM barrel. The TIM-barrel fold consists of an inner eight-stranded parallel β-barrel ($n = 8$) surrounded by eight α-helices on the periphery. The barrel is a closed structure with a shear number of 8 ($s = 8$); a shift of eight residues is required to return to the same starting point when following a hydrogen-bonded path perpendicular to the strands around the barrel[23]. Native TIM barrels, which often have constituent strands of different lengths, achieve the net $s = 8$ through a variety

of complex structural mechanisms (triosephosphate isomerase is shown as an example in **Fig. 1a**). For simplicity, we sought to design the highest-symmetry barrel possible. An eight-fold-symmetric structure is not feasible because of the alternating pleat of paired β-strands (**Fig. 1**); hence the highest symmetry attainable is four-fold. There are three possible ways to align β-strands with four-fold symmetry to achieve the $s = 8$ shear (depicted in **Fig. 1b–d**). Two of these arrangements (**Fig. 1b,c**) have strands that start with helix-facing residues (open circles); these are unfavorable because loops from helices to strands with this geometry are strained unless the loop is quite long (the "α/β rule")[4]. These considerations dictate that the simplest topology for an idealized TIM barrel is a repeat protein of four identical βαβα units with no strand register-shift within a unit, the first residue in the β-strands pointing into the barrel, and a shift of two residues between units (**Fig. 1d**).

We next sought to determine the lengths of the helices and loops appropriate for the strand arrangement in **Figure 1d**. As illustrated in **Figure 1e**, the helix between two strands with a register shift of 0 must be longer and more tilted than the helix between two strands with a register shift of 2; this requirement is masked by the irregularity of naturally occurring barrels. A sequence specific for such a fold must precisely define the ends of the helices, their packing onto the sheet and the change in chain direction brought about by the loops. We imposed three rules in the sequence design process to meet those requirements: (i) all helices are capped on the N and C termini, (ii) the sheet-facing side of the α-helices cannot be all alanines, and (iii) all backbone hydrogen-bonding groups in the loops must be satisfied. We also decoupled the side chains on the β/α loops from the rest of the core by restricting the amino acids to either be polar or alanine, so that catalytic features could later be introduced into the structure[24,25].

De novo design of TIM-barrel structures
To generate ideal TIM-barrel backbones satisfying the above principles, we fixed the β-strand length at five residues and sampled different possible lengths for each of the two unique helices and for
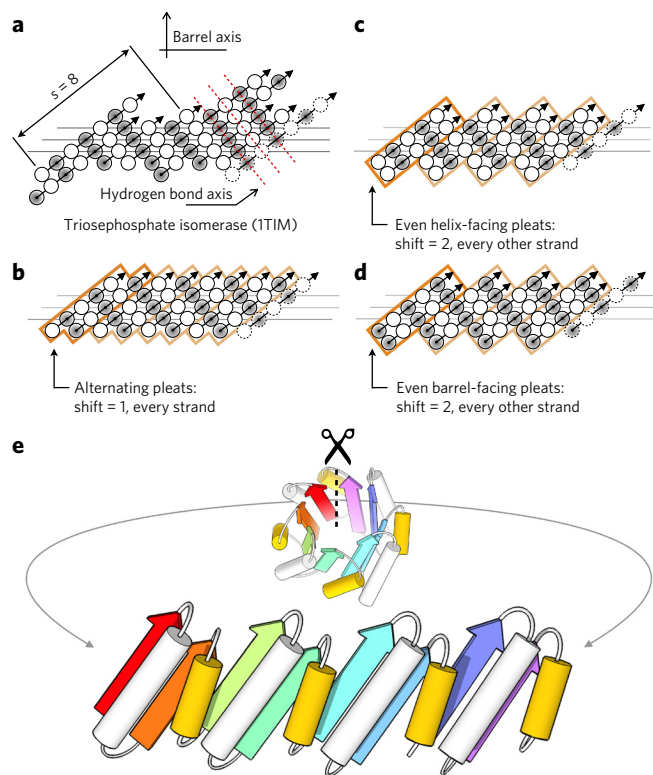
Figure 1 | Geometric constraints on the secondary structure arrangement in an ideal four-fold-symmetric TIM barrel. (a) The asymmetric β-strand arrangement of the classic triosephosphate isomerase barrel from chicken muscle (PDB code: 1TIM). The strands, as defined by DSSP[40], are viewed from the outside of the barrel, with shaded and open circles representing amino acid residues pointing into and out of the barrel, respectively. The first strand is shown at the left and right. Three horizontal lines indicate the residues lining the interior of the barrel. Sheet hydrogen bonds follow diagonal directions indicated by dashed red lines. **(b–d)** The three solutions for achieving the $s = 8$ shear with four-fold symmetry. Orange boxes represent four individual repeat units; the first unit is highlighted for clarity. Because residues represented by open and shaded circles are structurally nonequivalent and the diagonal-running hydrogen bonds must connect residues pointing in the same direction, an eight-fold symmetric barrel is not possible. **(b)** Register shift of one residue between every strand. **(c)** Register shift alternating between 0 and 2 residues around the barrel, with strands starting with residues pointing toward the helices (open circles). **(d)** Alternating (0,2) register shift as in **c**, with strands starting with residues pointing into the barrel (shaded circles). **(e)** To achieve the shear pattern in **d**, the helix spanning the offset strands (yellow) must be shorter than the helix within the repeat unit.

the four unique loops in the repeat unit (one-quarter of the barrel). For each choice of helix and loop lengths, we carried out 2,000 independent Rosetta *de novo* fragment-assembly calculations guided by the secondary structure assignment, propagating the structure of the first repeat unit into a total of four successive tandem repeats[26,27]. The length combination that most strongly converged to form a closed barrel structure for the repeat unit was found to be $5_{strand\ 1} + 3_{\beta/\alpha\ loop\ 1} + 13_{helix\ 1} + 3_{\alpha/\beta\ loop\ 1} + 5_{strand\ 2} + 3_{\beta/\alpha\ loop\ 2} + 11_{helix\ 2} + 3_{\alpha/\beta\ loop\ 2}$. The structure with the most extensively hydrogen-bonded cylindrical sheet was selected as the starting point for sequence design and structure refinement calculations (see Online Methods for details).

We designed sequences for this starting backbone using iterative cycles of side chain placement and all-atom energy minimization, generating an ensemble of structures with different sequences in each cycle. At each iteration, the sequence space was restricted to

that spanned by the top-ranking solutions from the previous cycle (**Supplementary Results**, **Supplementary Fig. 1**). No information from known structures was used. We experimented with several different ways of constraining the solutions to be consistent with the three sequence-design rules described above as follows. α/β loop 2: the register shift between one repeat unit and the next causes a carbonyl and an amide group in the preceding strand to be solvent exposed. We therefore constrained the identity of the third position in the loop to aspartate to allow it to hydrogen bond to the exposed amide (**Fig. 2a**). In addition, position 21 on helix 1 was constrained to be an arginine to interact with the otherwise unsatisfied carbonyl on the first residue in α/β loop 2 (Gly44) and the Asp1 side chain (**Fig. 2a**). β/α loop 2: the two residues flanking the loop (32 and 16) were set to serine and glutamine to form hydrogen bonds with the loop backbone (**Fig. 2b**). α/β loop 1: position 26 was set to threonine to hydrogen bond with Trp42. Strand 2: position 30 was set to tyrosine to generate a more featured surface for the helices to pack on. Helices: the helices were required to have at least one valine or leucine pointing toward the β-barrel (**Fig. 2c**), and the spacing between the helices was set by placing tryptophans at positions 14, 35 and 42 (**Fig. 2d**).

Experimental characterization of designed TIM barrels

We obtained synthetic genes encoding 22 low-energy designs with perfect four-fold repeats that satisfied different subsets of the above criteria. All 22 were expressed at high levels in *Escherichia coli* and could be readily purified, but only five showed cooperative thermal denaturation in CD experiments (**Supplementary Figs. 2** and **3**). The cooperatively folded designs all have Asp1, Trp35 and Trp42, suggesting that side chain–backbone interactions in α/β loop 1 and α/β loop 2 are important. Individual replacements of Asp1 with lysine, Trp35 with alanine, and Trp42 with histidine or leucine all resulted in poor CD spectra, indicating that all three residues are required for folding (**Supplementary Table 1**). Incorporation of an arginine at position 21 on sTIM-1 increased the melting temperature from ~54 °C to 72 °C, perhaps as a result of electrostatic interactions with the helix dipoles or hydrogen-bonding interactions with the α/β loop 2 (ref. 28) (**Supplementary Table 1** and **Supplementary Fig. 4**).

We further explored the sequence determinants of folding using asymmetric and dimeric designs. In the design, the interior of the β-barrel is hydrophobic, with a ring of arginine-aspartate salt bridges crowning the top side (the 'catalytic' side, C-terminal to the β-strands). To explore the contribution of these features to stability, we tested several asymmetric sequences that differ in the salt-bridge ring and in the first layer of hydrophobic residues from the bottom of the barrel (**Fig. 3**). We found that the hydrophobic residues contributed significantly to stability (with the original design being most stable), whereas variations in the salt-bridge ring were neutral (**Supplementary Fig. 5**). Despite the differences in stability, all variants still exhibited α/β CD spectra and cooperative unfolding, establishing that the two-layered toroid structure of a TIM barrel is tolerant to modifications in the barrel interior, which is desirable for evolving catalytic function. A half-barrel construct based on the original interior was found to self-associate into a monodisperse full barrel (**Supplementary Fig. 6**), as has been described for a natural as well as a designed TIM barrel assembled from native templates[16,29]. All sequences are reported in **Supplementary Table 2**.

Structure and folding thermodynamics of sTIM-11

By circularly permuting the barrel to start the chain from the N-terminal end of the long α-helix (**Supplementary Fig. 7**) and introducing cysteines at positions 7 and 180 using the disulfide modeling protocol in RosettaRemodel[26], we obtained a design, sTIM-11, that crystallized within 2 months in several conditions and yielded crystals diffracting up to 2.0 Å. The X-ray structure, solved
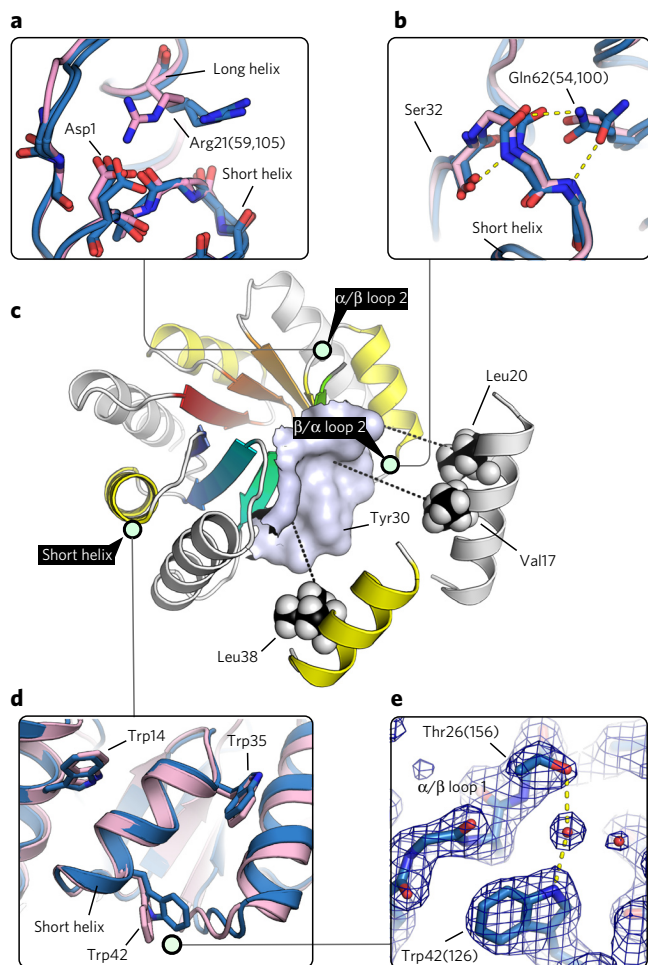
Figure 2 | Sequence determinants of *de novo*–designed TIM barrel. The designed TIM-barrel model is depicted with light green circles tagging regions shown in the insets, where the design models are shown in pink and X-ray structures in blue. Residues are numbered according to the design model; X-ray structure residue numbers are also given, in parentheses. (**a**) The α/β loop at the interface between the repeat units with a register shift of 2. Asp1 was designed to satisfy the hydrogen-bonding requirement for the backbone amide group on the neighboring strand. In the crystal, Arg21 makes lattice contacts rather than the designed interaction; in solution the designed hydrogen bond may be formed. (**b**) Features stabilizing the β/α loop backbone. Ser32 was designed to interact with the amide group that points toward the hydrophobic core, and Gln62 to interact with the carbonyl similarly to Arg21 (in **a**). Alternative conformations of Gln62 are observed in two different repeats in the crystal structure. (**c**) Packing of the sheet facing side of the helices (long helix, white; short helix, yellow) against the surface of the sheet. (**d**) The wedges between the helices are filled by tryptophans. Trp42 was found to adopt a different conformation in the crystal structure. (**e**) Trp42 was designed to interact with Thr26 on the neighboring loop directly, but crystallographic evidence suggests that the interaction is mediated by water, as shown by the clear electron density bridging the two residues.

by molecular replacement using the design model backbone as the template (R_{work} 0.22, R_{free} 0.26), reveals a compact four-fold-symmetric $(β/α)_8$ barrel. The B factors are high in the first three helices, which are probably less well defined in solution (**Supplementary Fig. 8**). The overall Cα root mean square deviation to the model is 1.28 Å, with deviations mostly in the β/α loops and the termini (**Fig. 4a**). Nearly all of the side chains in the refined crystal structure are in perfect agreement with the design, and the internal repeat units
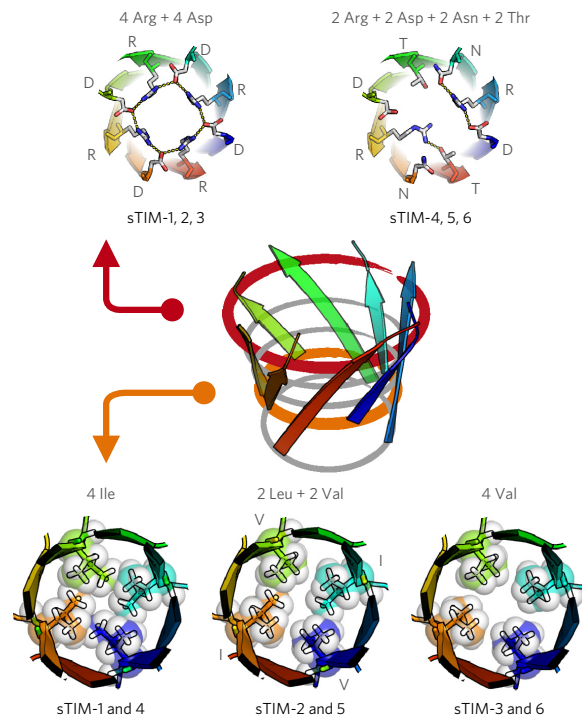

Figure 3 | Effect of two-fold- versus four-fold-symmetrical barrel interior on stability. Starting from the sTIM-1 sequence, we explored variations in the interior of the β-barrel. The different packing layers of the barrel are shown with circles in the central barrel figure. Models for two of the layers where asymmetric designs were made are shown in red and orange. Two configurations were tested for the polar red layer and three for the hydrophobic orange layer. The six possible combinations were tested with variants sTIM-1–sTIM-6. For example, sTIM-1 is four-fold symmetrical with four arginine-aspartate salt bridges crowning the top (red layer) and has four isoleucines in the interior of the barrel (orange layer). Strands are sequentially colored from blue to red, and for the orange layer configurations, side chain packing is shown with space-fill spheres. The stabilities of the six different variants correlate strongly with the configurations in the hydrophobic packing layer, in the order sTIM-1 (71.9 °C) ≈ sTIM-4 (71.8 °C) > sTIM-2 (66.5 °C) ≈ sTIM-5 (65.8 °C) > sTIM-6 (59.3 °C) ≈ sTIM-3 (56.9 °C). (CD melting curves of these variants are provided in **Supplementary Fig. 5**.)

are nearly identical to each other (**Fig. 4b**). An exception is the cysteine side chains, which did not form the intended disulfide bond in the crystal structure and might therefore contribute to the observed flexibility in the N-terminal α-helix. Most of the other key design features described are recapitulated in at least one of the repeating units in the crystal structure (**Fig. 4a** and superpositions in **Fig. 2a,b,d**). Trp42, however, forms a water-mediated hydrogen bond with Thr26 (**Fig. 2e**) instead of the direct hydrogen-bonding interaction as designed.

The folding thermodynamics of sTIM-11 were characterized by chemical and temperature denaturation experiments monitored by CD and fluorescence spectroscopy. Both guanidinium chloride (GdmCl) and temperature denaturation were cooperative and fully reversible (**Fig. 4c–e**). Reversibility of the temperature-induced unfolding of TIM barrels is uncommon[30]; the shorter loops and overall more ideal structure of our design likely contribute to folding robustness. The computed ΔG of unfolding is ~4.2 kcal/mol (17.6 kJ/mol), and the melting temperature ~88 °C (the initial decrease of the CD signal at 222 nm is probably due to increased flexibility of α-helices upon heating, as observed in the crystal structure) (**Fig. 4c,d**).

Comparison with naturally occurring TIM-barrel structures

As expected, structure searches[31,32] with the sTIM-11 crystal structure return many global hits to natural TIM barrels (the local structure in the design model is, however, not TIM-barrel specific (**Supplementary Fig. 9**)). To explore how different the newly designed sTIM-11 sequence is from those of known TIM barrels, we carried out PSI-BLAST[33] searches with three iterations. No TIM-barrel sequences were found, indicating that the *de novo* design is distinct from known TIM-barrel superfamilies. In more sensitive profile-profile comparisons with HHsearch[34,35], we found similarities between sTIM-11 and a variety of different TIM-barrel sequences that are distributed over a number of superfamilies: for example, the ribulose phosphate binding barrel (c.1.2), the nicotinate/quinolinate PRTase C-terminal domain–like (c.1.17) and the dihydropteroate synthetase–like superfamilies (c.1.21) (see Online Methods). To locate the new sTIM-11 barrel within the highly connected landscape of TIM-barrel relationships[36], we created a cluster map[37] of all TIM barrels within the Astral SCOPe 2.04 database and the sTIM-11 design (**Fig. 5**). sTIM-11 is clearly distinct from existing TIM-barrel superfamilies.

DISCUSSION

De novo design of TIM-barrel structures has proven difficult, as evidenced by decades of unsuccessful attempts; even the shortest such structure must be nearly 200 residues long and requires precise meeting of N-terminal and C-terminal α/β elements to form a closed toroid. We have succeeded in the *de novo* design of a four-fold-symmetric TIM barrel based on geometric constraints arising from the $n = 8$, $s = 8$ barrel topology and our previously described design rules for connections between secondary structures. Focusing on the four-fold symmetrical case greatly reduced the complexity of the sequence and structure spaces that were searched in the design calculations. Symmetry also facilitated the experimental testing of key interactions. The design principles developed here can potentially extend to β/α barrel arrangements not observed in nature. An idealized leucine-rich repeat, when built out to a full circle, has a barrel topology of $n = 20$, $s = 0$ (ref. 38), and there may be other stable structures in between these 8- and 20-stranded barrels. Key to exploring such arrangements would be mechanisms for ensuring strand register analogous to the side chain–backbone hydrogen bonds found to be important here.

It is instructive to compare our results with previous efforts to design TIM barrels. In the octarellin series[20,21], equal-length helices were used, and as outlined in **Figure 1**, with this choice it is difficult to set the strand register. A recent effort aimed at a topology similar to that of **Figure 1** but lacked mechanisms for loop stabilization and for further specifying the strand register[22]. The series of designed variants described here clearly show the importance of specific side chain–backbone hydrogen-bonding interactions for achieving a highly ordered structure. Previous *de novo* designs of α/β protein structures focused on hydrophobic packing[1,4], but for TIM barrels both our results and the comparison with the previous studies suggest that polar interactions are critical for specifying the fold. The difference may in part be one of size. The number of alternative hydrophobic packing arrangements increases rapidly with size, and because TIM barrels are significantly larger than previously designed
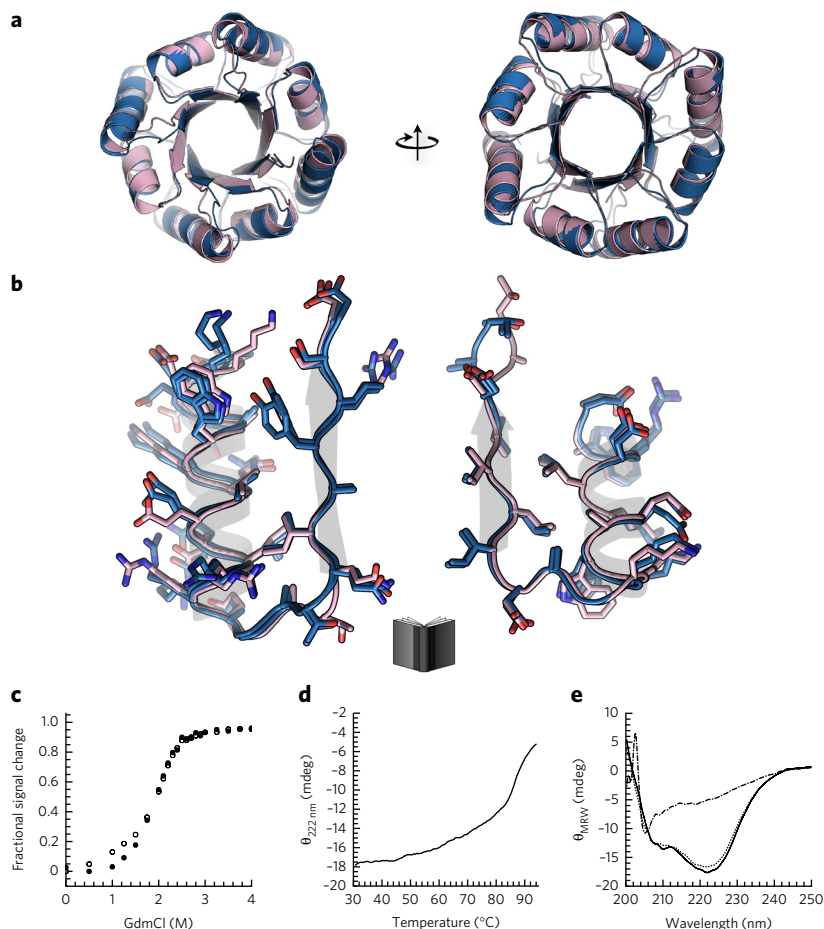

Figure 4 | Stability and structure of sTIM-11. (**a**) Superposition of the X-ray crystal structure (in blue) and the design model (in pink). (**b**) Comparison of the crystal structure and design model over two internal repeat units (residues 47–92 and 93–138). The internal repeats are nearly identical, and their side chains are in perfect agreement with the model. (**c**) Chemical denaturation with guanidinium chloride (GdmCl) followed by CD (open circles) and fluorescence (closed circles). Signals for the secondary (CD) and tertiary (fluorescence) structures are lost simultaneously at ~2 M GdmCl. (**d**) Thermal melt followed by CD, measured as degrees of ellipticity at 222 nm. (**e**) CD wavelength spectra of sTIM-11 measured as degrees of ellipticity at 30 °C (line), when melted at 95 °C (dashed) and after cooling at 30 °C (dotted).

α/β proteins, additional hydrogen-bonding interactions may be required to resolve this degeneracy and specify the overall topology.

The TIM-barrel scaffold offers numerous advantages for catalytic site placement because residues from all eight strands and the adjoining loops point into the region at the top of the barrel that typically contains the active site. The large number of active site geometries this enables is likely to account at least in part for the proliferation of TIM-barrel proteins in nature. Previous enzyme design work has also sought to take advantage of the TIM-barrel scaffold by placing designed active sites onto the backbones of naturally occurring TIM-barrel structures. Although active enzymes have been designed, crystal structures of designed enzymes have shown that long loops adjacent to the active site undergo unexpected reconfigurations in some cases[39]. sTIM-11, with its simple regular structure and minimalist loops, is a robust platform for engineering of new activities, and now that the key design principles and determinants of the fold are understood, large ensembles of TIM-barrel structures can be generated *in silico* as starting points for enzyme design calculations. More generally, the principles identified in this work allow the *de novo* design of custom-made catalysts or binders without the need to negotiate the structural complexity of naturally occurring proteins.
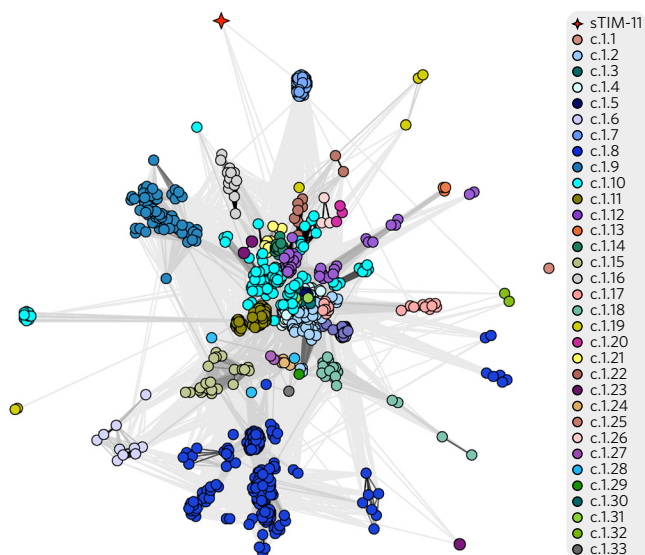
Figure 5 | Sequence relationships between natural (β/α)$_8$-barrels and sTIM-11. HHsearch calculations with sTIM-11 against the Astral SCOPe 2.04 database filtered for 95% sequence identity were carried out, and clustering was performed at a P-value cutoff of 1.0×10^{-2}. Connections are shown in different shades of gray with a linear scaling between P values of 1.0×10^{-2} and 1.0×10^{-56}.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The crystal structure of sTIM-11 has been deposited in the RCSB Protein Data Bank under the accession code 5BVL.

References

1. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
2. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
3. Joh, N.H. *et al.* De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle. *Science* **346**, 1520–1524 (2014).
4. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
5. Smadbeck, J. *et al.* De novo design and experimental characterization of ultrashort self-associating peptides. *PLOS Comput. Biol.* **10**, e1003718 (2014).
6. Bellows-Peterson, M.L. *et al.* De novo peptide design with C3a receptor agonist and antagonist activities: theoretical predictions and experimental validation. *J. Med. Chem.* **55**, 4159–4168 (2012).
7. Khoury, G.A., Smadbeck, J., Kieslich, C.A. & Floudas, C.A. Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* **32**, 99–109 (2014).
8. Correia, B.E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206 (2014).
9. Sterner, R. & Höcker, B. Catalytic versatility, stability, and evolution of the (βα)$_8$-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055 (2005).
10. Gerlt, J.A. New wine from old barrels. *Nat. Struct. Biol.* **7**, 171–173 (2000).
11. Höcker, B. Directed evolution of (βα)$_8$-barrel enzymes. *Biomol. Eng.* **22**, 31–38 (2005).
12. Kiss, G., Çelebi Ölçüm, N., Moretti, R., Baker, D. & Houk, K.N. Computational enzyme design. *Angew. Chem. Int. Ed. Engl.* **52**, 5700–5725 (2013).
13. Höcker, B., Claren, J. & Sterner, R. Mimicking enzyme evolution by generating new (βα)$_8$-barrels from (βα)$_4$-half-barrels. *Proc. Natl. Acad. Sci. USA* **101**, 16448–16453 (2004).
14. Höcker, B., Lochner, A., Seitz, T., Claren, J. & Sterner, R. High-resolution crystal structure of an artificial (βα)$_8$-barrel protein designed from identical half-barrels. *Biochemistry* **48**, 1145–1147 (2009).
15. Claren, J., Malisi, C., Höcker, B. & Sterner, R. Establishing wild-type levels of catalytic activity on natural and artificial (βα)$_8$-barrel protein scaffolds. *Proc. Natl. Acad. Sci. USA* **106**, 3704–3709 (2009).
16. Fortenberry, C. *et al.* Exploring symmetry as an avenue to the computational design of large protein domains. *J. Am. Chem. Soc.* **133**, 18026–18029 (2011).
17. Goraj, K., Renard, A. & Martial, J.A. Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the α/β-barrel Proteins. *Protein Eng.* **3**, 259–266 (1990).
18. Houbrechts, A. *et al.* Second-generation octarellins: two new de novo (β/α)$_8$ polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability. *Protein Eng.* **8**, 249–259 (1995).
19. Tanaka, T. *et al.* Characteristics of a de novo designed protein. *Protein Sci.* **3**, 419–427 (1994).
20. Offredi, F. *et al.* De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol.* **325**, 163–174 (2003).
21. Figueroa, M. *et al.* Octarellin VI: using Rosetta to design a putative artificial (β/α)$_8$ protein. *PLoS ONE* **8**, e71858 (2013).
22. Nagarajan, D., Deka, G. & Rao, M. Design of symmetric TIM barrel proteins from first principles. *BMC Biochem.* **16**, 18 (2015).
23. 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).
24. Ochoa-Leyva, A. *et al.* Exploring the structure-function loop adaptability of a (β/α)8-barrel enzyme through loop swapping and hinge variability. *J. Mol. Biol.* **411**, 143–157 (2011).
25. Ochoa-Leyva, A. *et al.* Protein design through systematic catalytic loop exchange in the (β/α)$_8$ fold. *J. Mol. Biol.* **387**, 949–964 (2009).
26. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6**, e24109 (2011).
27. Parmeggiani, F. *et al.* A general computational approach for repeat protein design. *J. Mol. Biol.* **427**, 563–575 (2015).
28. Yang, X., Kathuria, S.V., Vadrevu, R. & Matthews, C.R. βα-hairpin clamps brace βαβ modules and can make substantive contributions to the stability of TIM barrel proteins. *PLoS ONE* **4**, e7179 (2009).
29. Höcker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner, R. Dissection of a (βα)$_8$-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36 (2001).
30. Romero-Romero, S., Costas, M., Rodríguez-Romero, A. & Fernández-Velasco, D.A. Reversibility and two state behaviour in the thermal unfolding of oligomeric TIM barrel proteins. *Phys. Chem. Chem. Phys.* **17**, 20699–20714 (2015).
31. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
32. Minami, S., Sawada, K. & Chikenji, G. MICAN: a protein structure alignment algorithm that can handle multiple-chains, inverse alignments, C(α) only models, alternative alignments, and non-sequential alignments. *BMC Bioinformatics* **14**, 24 (2013).
33. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
34. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
35. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
36. Farías-Rico, J.A., Schmidt, S. & Höcker, B. Evolutionary relationship of two ancient protein superfolds. *Nat. Chem. Biol.* **10**, 710–715 (2014).
37. Alva, V., Remmert, M., Biegert, A., Lupas, A.N. & Söding, J. A galaxy of folds. *Protein Sci.* **19**, 124–130 (2010).
38. Rämisch, S., Weininger, U., Martinsson, J., Akke, M. & Andre, I. Computational design of a leucine-rich repeat protein with a predefined geometry. *Proc. Natl. Acad. Sci. USA* **111**, 17875–17880 (2014).
39. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
40. Kabsch, W. & Sander, C. Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

Author contributions

P.-S.H., K.F., D.A.F.V., B.H. and D.B. designed the research. P.-S.H. wrote program code and designed structures with help from B.H. and D.A.F.V. P.-S.H., F.P. and D.A.F.V. built the clones. K.F. and B.H. solved the crystal structure of sTIM-11 and collected thermodynamic data for sTIMs. F.P. characterized the first designs. P.-S.H., K.F., B.H. and D.B. collected and analyzed sequence and structure comparison data. K.F. and B.H. generated the cluster map. P.-S.H., K.F., B.H. and D.B. wrote the manuscript with help from all the authors. All authors discussed the results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Any supplementary information, chemical compound information and source data are available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to D.B., B.H. or P.-S.H.

ONLINE METHODS

Rosetta modeling suite. The Rosetta software suite is freely available to academic and government laboratories and require commercial licenses for business use. It can be obtained through the RosettaCommons website: https://www.rosettacommons.org.

RosettaRemodel *de novo* repeat modeling procedure. Two new features were implemented in RosettaRemodel[26] in order to carry out *de novo* sequence designs and refinements in the context of repeat structures. The most convenient setup for handling tandem repeat designs is to allow all description for the task, including both backbone modeling and sequence optimization, to be specified in a blueprint file that spans only a single repeat unit and let RosettaRemodel automatically handle the mirroring of all the duplicated copies of a repeat unit. To build *de novo* structures, we need to (1) construct a *de novo* backbone from fragments and propagate it into a repeat protein, (2) simultaneously design the sequences for all repeats and (3) refine the models while maintaining the symmetry. Generally the backbone building and refinement steps are treated separately. RosettaRemodel can handle repeat construction from fragments already[27]. We revamped the sequence design optimization steps and the iterative refinement setup to better handle *de novo* repeat structures.

The sequence optimization algorithm, which is based on Monte Carlo searches, was improved to handle "rotamer links" that can be created for a list of equivalent positions in a structure. During Monte Carlo sampling, a perturbation step flips all of the linked residues to the same query state before the energy of the system is evaluated. There are other possible mechanisms to handle this design step. For example, with the symmetry machinery already in Rosetta, one can treat each repeat unit as the asymmetrical unit in a global symmetry definition to achieve the same effect. However, by setting up links independent of the global symmetry, symmetrical assemblies of repeat proteins (for example, collagen fibers), which require both a linear repeat and a global symmetry, can be designed without modifying code.

The functionality built for the all-atom refinement steps are essentially the mechanisms for RosettaRemodel to use the information provided by a blueprint and understand how the definitions are mapped to an input repeat structure. We had previously relied on constraints derived from native repeat families for this step, but the same type of information is not available to *de novo* models. Once the blueprint-to-structure relationship is established, existing protocols are used for the refinement calculations, but an additional set of non-crystallographic symmetry (NCS) constraints are automatically generated and applied to maintain the symmetry between repeat units; the NCS constraint links the torsion angles of a specific pair of residues. This allows synchronized backbone perturbations over either the entire repeat structure or defined sections within a repeat—a loop can be sampled, redesigned, extended, or shortened in all the repeat units simultaneously while the rest of the repeat structure is kept untouched. As part of this feature, any repeat protein, once given the definition of a repeat unit by blueprint, can be extended to copies.

The features were set up to be automatically enabled when the -repeat_structure [number of repeats] flag is given to the program together with an input PDB-formatted coordinate file carrying at least two repeating units. This requires the PDB to span longer than a single unit because the torsional angles at the junction will be used to define the positions of the downstream repeat units.

For TIM-barrel designs, where the repeats form a closed toroid, a cyclic peptide mode can be enabled by issuing -cyclic_peptide flag. An automated constraint setup will drive the N and C termini to join as if making a planar peptide bond and not clashing.

Conformational sampling for TIM-barrel topology. RosettaRemodel *de novo* building protocol was used to find the secondary structure length combinations that can fold into a TIM barrel.

We set up sampling runs for smaller units first to estimate the lengths of the secondary structure elements. Based on the geometric description we derived for the TIM-barrel fold, the first two strands have to pair up evenly in an β-α-β unit. We set up sampling runs for a β-α-β unit, keeping the lengths of the β-strands at 5 residues, and sampled the two loop lengths between 2 and 3 and the helix length between 10 and 14. Approximately 50 models were generated

for each setup. The loop lengths were found to be 2 for the β-α loop and 3 for the α-β loop, and the optimal helix length was found to be 13—other lengths either changed the β-strand register shifts or prevented strand pairing. To approximate the shorter helix length for setting the $s = 2$ strand register shift, a β-α-β-α-β unit was sampled using the best definition of β-α-β and a new α-β unit with varying lengths. We kept the same connecting loop lengths as ones found previously for the additional α-β unit and varied the length of the additional helix between 10 and 11. The lengths for the shorter helix did not converge as cleanly as the β-α-β unit alone, so both lengths were used in the next step, in which the units are built into four repeats.

Using the results above, more sampling units were built as four-copy repeats with the *de novo* repeat protein machinery to sample loop lengths between 2 and 3 and helix lengths of 13 for the long helix and 10 to 11 for the short helix. Fragment-only sampling was used at first, and approximately 2,000 models were built for each length combination. The optimal length was determined by the number of barrel-like structures produced. Only about 1% of the structures are TIM-barrel like. Models from the calculations were selected based on their backbone hydrogen bonding energy scores (hb_lrbb). However, satisfying hydrogen bonding does not guarantee a toroidal structure. A "flatness" measure that calculated the deviations of the central Cα positions (the third residue in each strand) from a plane was used for this work for identifying toroidal structures, but RosettaRemodel now reports helical fitting parameters—the superhelical properties of a repeat protein can be described by three parameters, radius, twist and helical rise—directly from sampling results, and a toroid can easily be identified if the helical rise parameter is near zero.

Our final choice of a 46 residue repeat unit was identified. The secondary structure length combination that produced this result was then resampled with full-atom refinement steps (controlled by the -use_pose_relax flag) and cyclic peptide constraints (controlled by the -cyclic_peptide flag) enabled to produce backbones for further sequence optimization. These settings generated perfectly four-fold symmetrical models. The structure best satisfying the hydrogen bonding pattern and toroidal shape was chosen for sequence design. The flags for running the sampling steps and the blueprint file that produced the final structure are given in **Supplementary Table 3**.

Sequence design and iterative backbone adjustments. The single model chosen from the fragment sampling stage for refinement has a complete β-barrel, but its backbone conformation and sequence were improbable for folding. To obtain a set of sequences for experimental testing, ten cycles of iterative refinement steps were carried out, with each cycle generating 2,000 models from the same starting structure. In each refinement step, the backbone and the sequence of the starting model were iteratively perturbed to explore the conformational space, making an ensemble of similar structures of different sequences. The refinement cycles were thus controlled by the blueprint definitions that gradually reduce the sequence search space. Between each consecutive refinement cycle, the amino acid choices available for each position is reduced manually—based on both enrichment ratios in all the models (as illustrated in **Supplementary Fig. 1**) and chemical intuition—until they converge to a single amino acid, except for when there was no strong preference for a position, the degeneracy was carried forward. Backbone conformations drifted in the first cycle but quickly converged when positions were locked into certain amino acid types in subsequent cycles. The sequences largely converged by the 8th cycle. (Command lines and blueprint files in **Supplementary Table 4**.)

General cloning, expression and protein purification information. Genes were obtained from GenScript directly in pET21b or synthesized as gBlocks (Integrated DNA Technologies) and cloned into pET29b. For the clones used for folding assessments, plasmids were transformed into *E. coli* BL21(DE3) strain (Novagen, cat. no. 69450) for IPTG induced expression in LB media (MP Biomedicals, cat. no. 113002056) at 18 °C overnight. Cell pellets were collected by centrifugation at 6,000 g for 20 min, resuspended in Tris buffer (20 mM Tris pH 8, 150 mM NaCl) with protease inhibitors (Roche, cat. no. 11697498001) and lysed by sonication. Ni-NTA beads (Qiagen, cat. no. 30410) or HisTrap FF (GE Life Sciences, cat. no. 17-5255-01) were used to purify the proteins, with imidazole elution concentrations between 10 mM to 500 mM. Size exclusion by gel filtration steps were performed on ÄKTAxpress units (GE Life Sciences)

using Superdex 75 columns (GE Life Science, cat. no. 17-5174-01) with 50 mM sodium phosphate buffer pH 8 with 150 mM NaCl.

Detailed expression and purification protocols for sTIM-11 characterization and crystallization. The sTIM-11 protein was expressed in *E. coli* BL21. The cells were grown at 37 °C in TB medium and—at an OD_{600} of 1—expression was induced with IPTG at 1 mM concentration. The cells expressed protein over 15 h at a temperature of 30 °C. After harvesting by centrifugation, the pellet was washed with 50 mM potassium phosphate buffer pH 8, 150 mM NaCl, centrifuged again and resuspended in 40 mL of the same buffer. The cells were lysed by sonication on ice. The cell debris was removed by centrifugation (18,000 rpm for 45 min) and additional filtration of the supernatant (0.45 μm and 0.22 μm syringe filters). The filtered solution was loaded onto a 1 mL Ni-NTA column and washed with 50 mM potassium phosphate buffer pH 8, 150 mM NaCl, 10 mM imidazole. The protein was eluted from the column with increasing concentrations of imidazole and fractions containing sTIM-11 were pooled and loaded onto a Superdex 75 HiLoad 26/60 column equilibrated with 50 mM potassium phosphate buffer pH 8, 150 mM NaCl, 1 mM DTT. The same buffer was used for the elution with a flow rate of 1.5 mL/min. The protein eluted in one peak and was concentrated. The protein sample was then dialyzed three times against 1 L of 50 mM potassium phosphate buffer pH 8, 150 mM NaCl or—for crystallization—against 50 mM Tris pH 8, 150 mM NaCl.

Qualitative folding assessments with circular dichroism (CD). For variants reported in **Supplementary Figures 2** and **3**, the melting curves were collected on an AVIV-420 CD spectrometer monitored at 220 nm in 50 mM sodium phosphate buffer pH 8 with 150 mM NaCl. Data points were collected at every 2 °C increments from 25 °C to 95 °C with 1 min equilibration time and with 30 s signal averaging time in a 1 mm pathlength cuvette.

Biophysical characterization of sTIM-11. The quality of the purification was determined by both electrophoresis on 15% polyacrylamide gels followed by Coomassie blue staining and analytical gel filtration (Superdex 75 10/300GL, 50 mM potassium phosphate buffer pH 8, 150 mM NaCl). The formation of secondary structure was determined by CD spectra recorded at a spectropolarimeter (Jasco J-810) at a protein concentration of 0.2 mg/ml with a sampling depth of five. Melting curves between 30 and 95 °C were made with the same setup. The increase in temperature was set to 1 °C/min. The changes in secondary structure were recorded at 222 nm. Additionally, complete CD spectra were recorded every 10 °C and at the end of the melting curve at 95 °C. An additional CD spectrum was recorded after the sample cooled back to 30 °C.

Chemical denaturation was measured by setting up parallel protein samples with increasing concentrations of guanidinium chloride (GdmCl) in 50 mM potassium phosphate buffer pH 8.0, 150 mM NaCl. Three days after the addition of GdmCl we recorded no more changes in signal and we measured at 25 °C both changes in secondary structure by CD at 222 nm (five recordings) as well as changes in tertiary structure by Trp fluorescence recorded at a spectrofluorometer (Jasco FP-6500, five recordings; excitation at 280 nm and the fractional change at 344/377 nm was measured and used to determine stability).

X-ray crystallography. Crystallization trials were set up in 96 well hanging drop plates. Crystals were first found after 2 months and used to record spectra at the synchrotron beamline PXII (wavelength = 1 Å) from the Swiss Light Source, Villigen PSI, Switzerland. Data were indexed, integrated and scaled with XDS and converted with XDSCONV[41], followed by molecular replacement with Phenix using a relaxed Rosetta sTIM-11 model as a template. Coot[42] was used for model building and both Phenix-Rosetta[43,44] and Phenix.refine[45] were used for refinement (Ramachandran outliers at 0% and Ramachandran favored at 96% for the final model). The structure coordinates were submitted to the Protein Data Bank under the accession code 5BVL (see **Supplementary Table 5** for details).

Bioinformatic analysis. For sequence comparisons of sTIM-11 with profiles based on (β/α)$_8$-barrels, we ran HHsearch (hhsuite 2.0.16) (ref. 34) against the Astral SCOPe 2.04 database filtered for 95% sequence identity (SCOPe95). The profiles were build with HHblits[35]. We used default parameters, but did not score secondary structure alignment to avoid bias. The cluster map compares sequences of all (β/α)$_8$-barrel structures in SCOPe95 and in addition sTIM-11 and was generated using the pairwise HHsearch *P* values in CLANS, that scales negative log-*P* values into attractive forces in a force field[46]. Clustering was done to equilibrium in 2D at a *P*-value cutoff of 1.0×10^{-2} using default settings.

41. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr* **66**, 125–132 (2010).
42. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics.*Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
43. DiMaio, F. *et al.* Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10**, 1102–1104 (2013).
44. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
45. Afonine, P.V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
46. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).