De novo protein fusions as platforms for enzyme design

Ian C. Haydon

A thesis
submitted in partial fulfillment of
the requirements for the degree of

Master of Science

University of Washington
2018

Committee:
David Baker
Frank DiMaio
Dustin Maly
Barry Stoddard
Jesse Zalatan

Program Authorized to Offer Degree:
Biochemistry

University of Washington

Abstract

De novo protein fusions as platforms for enzyme design

Ian C. Haydon

Chair of the Supervisory Committee:

David Baker

Biochemistry

Control over enzymatic catalysis is a central goal of biotechnology. Recent advances in computational protein design are beginning to allow for the *de novo* creation of arbitrary protein structures, but the design of arbitrary functions that match or exceed those found in nature remains extremely challenging. To help advance towards this demanding goal, I have designed *de novo* multidomain proteins based on one of the most common and catalytically diverse enzyme architectures known. These TIM-barrel fusion proteins remain folded at high temperatures and concentrations of denaturant and are highly mutable, and can thus serve as platforms for rational enzyme design.

Acknowledgements

To David Baker: Thank you for taking my interests and those of all of your graduate students seriously, and for letting us work hard in a place where people clearly matter. It was an honor to do research in a lab that always felt two steps ahead.

To my labmates: The Institute for Protein Design is a truly exceptional place to work, and not just because of the in-house espresso. David's leadership has created a healthy community of productive scientists who are free to be themselves. Thanks to everyone I worked with over the years — to those who advised me, shared code and brainstormed.

To my 2015 cohort and UW friends: Thank you for keeping UW vibrant and interesting.

To Taylor: Graduate school is hard; you made it easy. Thank you for supporting me.

Background

Enzymes enable life by catalyzing chemical reactions. They digest food, contract muscles, regulate neurons, repair DNA and much more. Modifying natural enzymes to solve new problems is a pillar of modern biotechnology.

One of the most successful protein folds in the evolution of natural enzymes is the TIM barrel, also known as $(\beta/\alpha)_8$ barrel. First observed in 1976 with the crystal structure of the glycolytic enzyme triosephosphate isomerase (TIM)[1], the fold is comprised of a circular β-sheet surrounded by eight α-helices. Thought to be the products of gene duplication[2,3], TIM barrels are now among the most common proteins in the Protein Data Bank. By virtue of the ubiquity of the carbon-fixing protein RuBisCo, they are also likely the most abundant proteins on Earth [4].

The TIM barrel fold is among the most structurally and functionally diverse, though the vast majority of natural TIM barrels are enzymes. TIMs are known to catalyze reactions in five of the six Enzyme Commission classes. What accounts for this striking catalytic diversity? The TIM barrel architecture is inherently amenable to evolution — a well-packed hydrophobic core which is physically distant from an active site permits active site mutations, which are essential for innovation, while maintaining overall stability[5].

The standard method of computational enzyme design — beginning with a natural TIM barrel or other protein and making mutations such that a new activity becomes possible — is fundamentally limited in that natural proteins have already been extensively optimized by evolution for their own native function(s), both at the backbone and sidechain level. Not only can most natural proteins not tolerate extensive mutation, their original function is unlikely to be conducive to or even compatible with the new desired function. This has forced enzyme designers to rely on minimally defined active sites and subsequent directed evolution to produce new enzymes with high levels of activity[6]. *De novo* enzyme design offers an alternative.

The Rosetta molecular modeling suite was recently used to design a minimal, four-fold symmetric TIM barrel from scratch[7]. The structure of this thermostable 184-residue protein, termed **sTIM11**, was verified by X-ray crystallography [T_m = 89° C; $EC_{50\text{-}GdmCl}$ = ~2M]. Sequence-based searches indicate that the designed amino acid sequence of sTIM11 is distant

from all naturally occurring TIM-barrel superfamilies, suggesting nature has sampled only a subset of the sequence space available for the TIM-barrel fold. Natural TIM-barrel scaffolds have been used in several prior enzyme design efforts, but sTIM11 was designed without catalytic function. It thus represents a promising starting point for fully *de novo* rational enzyme design.

I and others have worked to convert sTIM11 into a highly stable starting point for subsequent enzyme design based on the premise that protein function comes at the cost of thermodynamic stability[8]. Enhanced thermostability of this non-catalytic scaffold was achieved via extensive redesign of hydrophobic packing (yielding **reTIM26**), introduction of terminal (N-to-C) disulfide bonds, and homotrimerization. Combining these independently stabilizing features yielded the most stable *de novo* TIM barrels to date [T_m > 105 °C; $EC_{50\text{-}GuHCL}$ > 5M, see **Figure 1**].

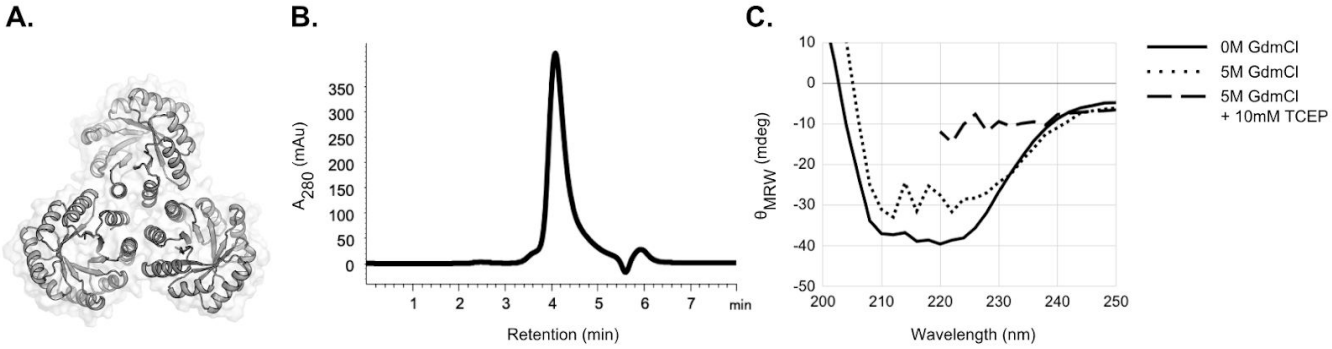

Figure 1: Hyperstable *de novo* TIM barrel scaffolds. A. Design model of homotrimeric reTIM26 with N-to-C disulfides. **B.** A size exclusion chromatogram of the purified protein shown in *A* with a dominant elution peak at 4.08 min, the expected retention for an assembled trimer. **C.** Circular dichroism reveals a folded protein which remains mostly folded in high concentrations of the denaturat guanidinium hydrochloride; upon co-incubation with the reducing agent TCEP, the protein completely unfolds.

Design of TIM-barrel fusions

Natural TIM barrels have structural features that reTIM26 lacks, including extra domains and extended β/α-loops which together make up the ligand binding sites of many natural enzymes[9]. Using RosettaRemodel[10], I fused small *de novo* designed protein domains onto sTIM11 initially and later onto reTIM26 to mimic these features. This is an alternative strategy to

loop design, which remains one of the greatest challenges in computational protein design. (Loop design on the reTIM26 scaffold is being pursued by Hanlun Jiang, a graduate student I helped train.)

Design of monomeric fusions

I fused multiple designed miniproteins[11] onto reTIM26 to create a set of fusions proteins that can serve as stable and readily modifiable enzyme scaffolds. These fusions have a cavity into which small-molecule binding sites can be engineered (see **Figure 2**). *De novo* designed proteins with N- and C-termini in close proximity (<6Å) were selected for insertion. The four-fold symmetry of the TIM barrel allowed for four equivalent points of insertion (**Figure 2A**); in an effort to preserve stability, the insertion point most distal to the TIM barrel's termini was selected (**Figure 2B**).

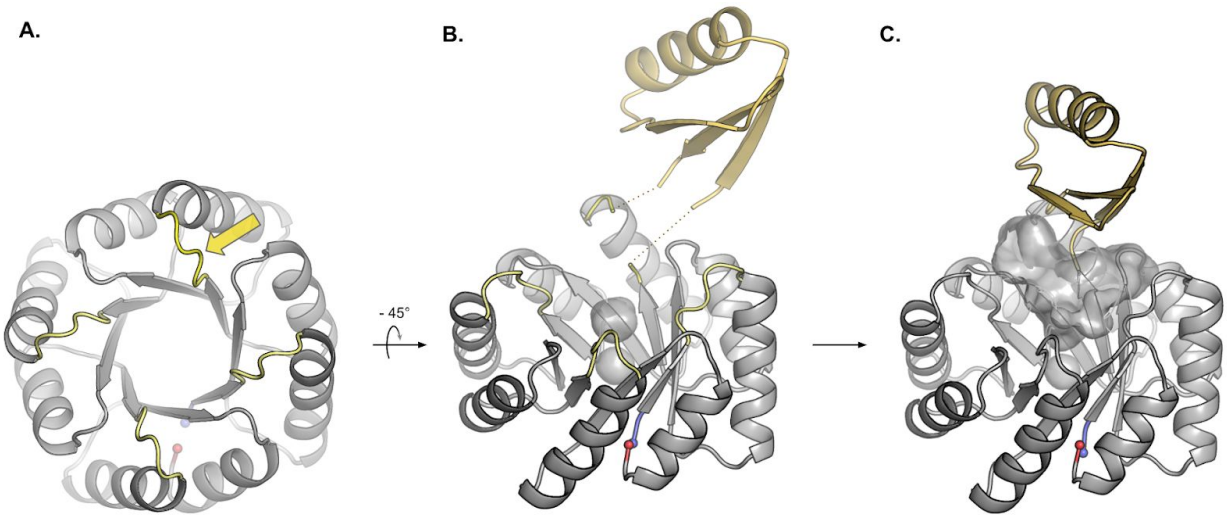

Figure 2: Overview of monomeric fusions. **A**. The four symmetrically equivalent positions for insertion (*yellow*); arrow indicates the selected position. **B.** A single β/α-loop of the *de novo* designed TIM barrel was replaced with a *de novo* designed miniprotein (*yellow*). **C.** A fully connected monomeric fusion with a cavity (*grey surface*) suitable for ligand binding.

RosettaRemodel was used to design the the connections (or 'linkers') between the TIM barrel and inserted domain (see **Figure 3**). Various linker lengths (0–6 amino acids) were sampled, though only a subset produced fully connected fusion proteins. To assess the position of the connected domains relative to the TIM barrel, the center of mass of the TIM barrel was set as a three-dimensional reference point (0,0,0), then the center of mass of the inserted domain was

calculated and compared to this reference (**Figure 3A**). Linker lengths that (i) were short, (ii) produced fully connected fusions on more than 50% of RosettaRemodel trajectories and (iii) placed the inserted domain directly "above" the TIM barrel were selected for further optimization (**Figure 3B**, grey box).

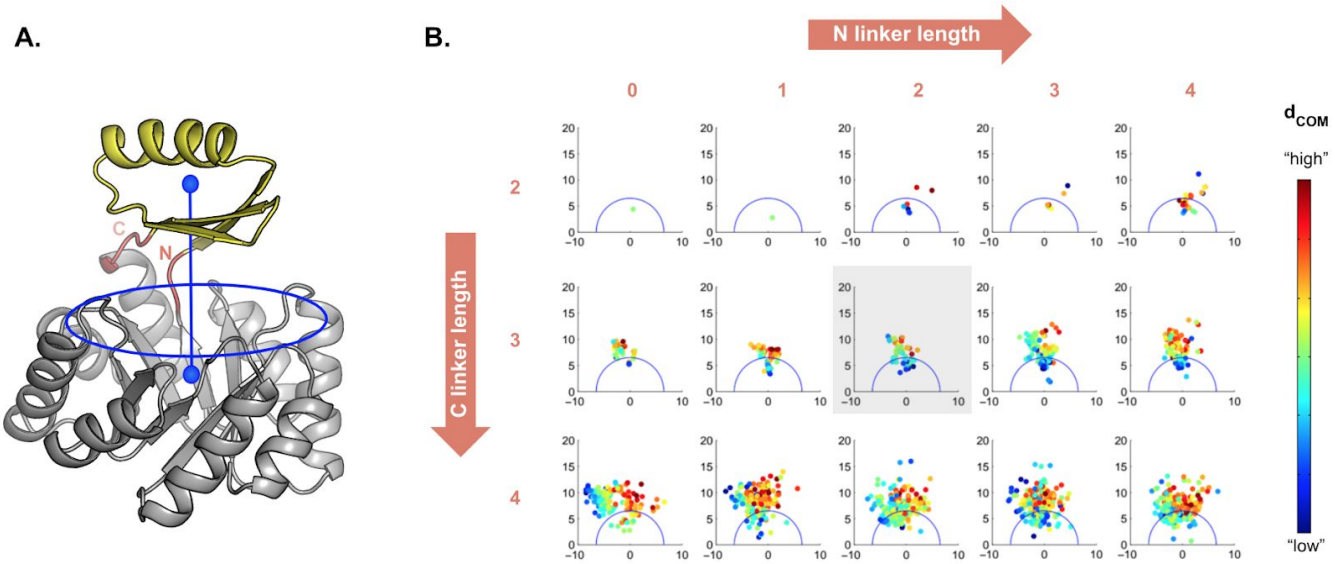

Figure 3: Assigning linker lengths of monomeric fusions. A. To assess the position of the inserted domain (*yellow*) relative to the TIM barrel (*grey*), the center of mass of each was compared (blue spheres); fusions with inserted domains "on top of" the TIM barrel (*inside the blue circle*) were preferred. **B.** A center-of-mass analysis of multiple fusions with various linker lengths (*units on graphs = Å*). 1000 design trajectories for each linker length combination were performed; each successfully closed solution is shown. Fusions with linkers length 2 and 3 (*N- and C-, respectively; grey box*) were selected for further optimization.

To optimize the amino acid composition of the linkers, blueprint files were used to redesign the amino acids at certain positions (e.g. in some cases, prolines were manually assigned). Features such as helix capping, hydrophobic barrel-domain interactions, the presence of proline and the absence of glycine (see **Figure 4**) were manually assigned to a subset of linkers and were selected for experimental evaluation.

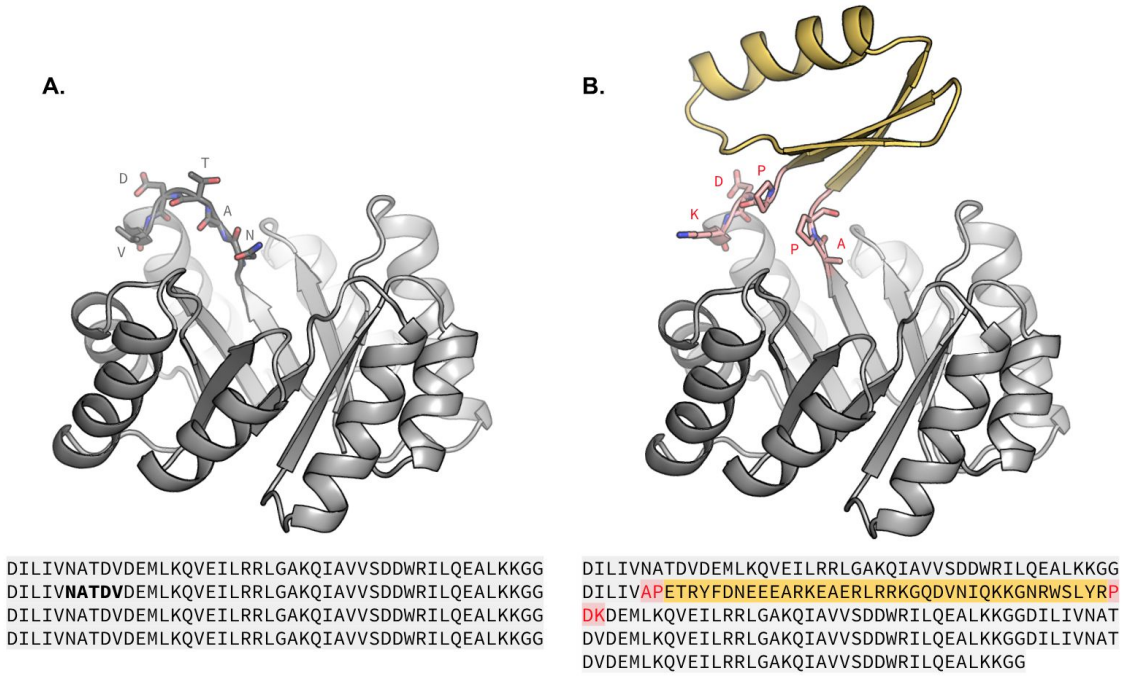

Figure 4. Sequence and structural view of a designed monomeric fusion. A. The sequence and structure of reTIM26 repeats four times (*grey*); an β/α-loop from the second repeat was selected for replacement (*in bold, shown as sticks*). **B.** The sequence and structure of the inserted domain (*yellow*) and the new designed connections (*red, shown as sticks*); These linkers feature helix capping and two prolines, intended to minimize interdomain flexibility.

Design of dimeric fusions

Some designs relied on symmetry. For these, *de novo* designed protein homodimers were fused onto a split homodimeric TIM barrel such that one half of reTIM26 was covalently connected to one of the inserted dimer subunits. Upon assembly, these homodimeric enzyme scaffolds should contain a highly mutable pocket large enough to house complex enzyme active sites (see **Figure 5**). *De novo* designed ferredoxin and CfR homodimers were selected for insertion[12,13].

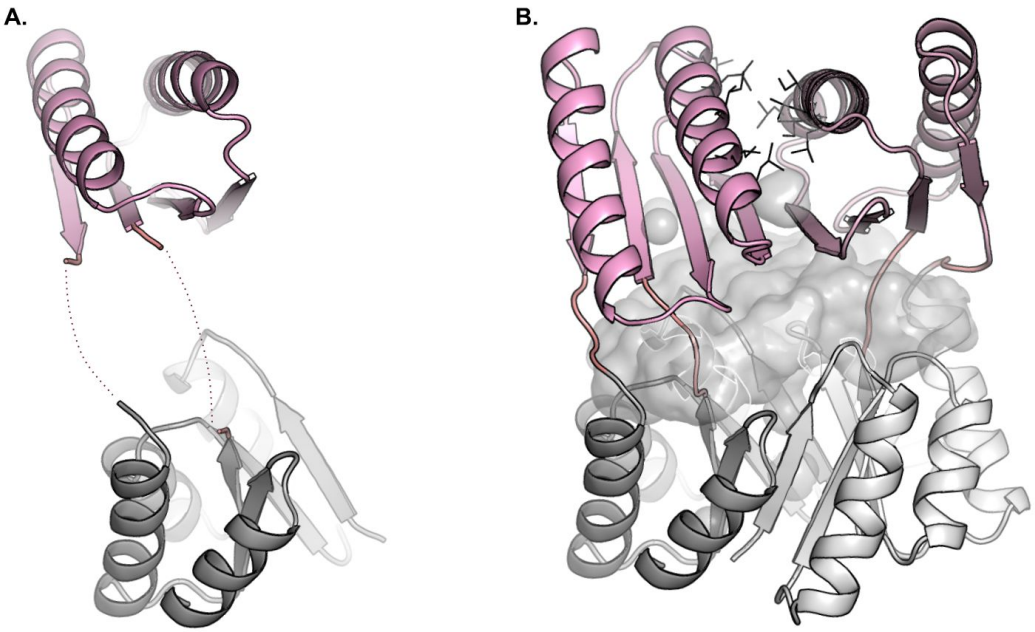

Figure 5. Overview of the design of dimeric fusions. A. A single *de novo* designed ferredoxin domian (*pink*) was fused to half of reTIM26 (*grey*). **B.** Upon homodimerization, a large cavity (*grey surface*) forms between the reconstituted TIM barrel and ferredoxin dimer (*ferredoxin-ferredoxin interface shown as lines*).

Connections were designed as described except that C2 symmetry was imposed throughout the insertion and design process via the following symmetry definition file :

```
symmetry_name reTIM26_C2_input_2
E = 2*VRT0_base + 1*(VRT0_base:VRT1_base)
anchor_residue COM
virtual_coordinates_start
xyz VRT0 0.2343902,-0.9719931,-0.0170486  0.9721236,0.2342398,0.0103677  0.0096813,0.0910492,-0.3335388
xyz VRT0_base 0.2343902,-0.9719931,-0.0170486  0.9721236,0.2342398,0.0103677  -1.9238571,8.1092418,-0.1929011
xyz VRT1  -0.2343902,0.9719931,0.0170486  -0.9721236,-0.2342398,-0.0103677  0.0096813,0.0910492,-0.3335388
xyz VRT1_base  -0.2343902,0.9719931,0.0170486  -0.9721236,-0.2342398,-0.0103677  1.9432197,-7.9271433,-0.4741765
xyz VRT  0.0000000,-0.9998194,-0.0190038  0.9999815,-0.0001156,0.0060828  0.0096813,-0.9087517,-0.3525422
virtual_coordinates_stop
connect_virtual JUMP0_to_com VRT0 VRT0_base
connect_virtual JUMP0_to_subunit VRT0_base SUBUNIT
connect_virtual JUMP1_to_com VRT1 VRT1_base
connect_virtual JUMP1_to_subunit VRT1_base SUBUNIT
connect_virtual JUMP0 VRT VRT0
connect_virtual JUMP1 VRT0 VRT1
set_dof JUMP0_to_com x(8.24922789260257) angle_x
```

set_dof JUMP0_to_subunit angle_x angle_y angle_z
set_jump_group JUMPGROUP2 JUMP0_to_com JUMP1_to_com
set_jump_group JUMPGROUP3 JUMP1_to_subunit JUMP0_to_subunit

For the initial round of dimeric designs, the sequences of the inserted domains and TIM barrel were unchanged. To optimize expression, a second round of design was performed in which select interfacial residues were changed from charged (D, E, K and R) to polar (N, S, T and Q).

Expression and biochemical analysis

Synthetic genes encoding the designed amino acid sequences were purchased from Integrated DNA Technologies (IDT) as gBlocks (sequence-verified double-stranded DNA fragments). For the initial set of monomeric fusion designs, the following 5′- and 3′- DNA overhangs homologous to the yeast-surface display vector pETCON were included:

5'-GCGGTAGCGGAGGCGGAGGGTCGGCTTCGCATATG…
...CTCGAGGGTGGAGGTTCCGAACAAAAGCTTATTTC-3'

Transformation of plasmid components into yeast and colony PCR: gBlocks were rehydrated with Milli-Q H_2O (Millipore Corporation) and transformed into yeast cells. Briefly, 1uL of each gBlock resuspension (100ng/uL) and 3uL of doubly digested pETCON3 plasmid (25ng/uL, gel purified after being cut with NdeI and XhoI) were added to standard PCR strip tubes. Then 56uL of a master mix made from 100uL of chemically competent *S. cerevisiae* (strain EBY100), 480uL of 50% PEG3350, 72uL of 1M LiOAc, and 60uL of dH_2O was added to each DNA-containing tube. Tubes were vortexed until mixed, incubated in a thermocycler for 30 minutes at 30°C, then incubated for 20 minutes at 42ºC. Tubes were then centrifuged at low speed, supernatants were discarded by aspiration and pellets were resuspended in 200uL dH2O. This spin-wash step was repeated two more times. After resting for 10 minutes, each final resuspension was pipetted onto a c-Trp-Ura plate and incubated at 30°C for two days, yielding >20 colonies per plate. PCR was used to identify colonies which contained fully assembled plasmids. Briefly, ~3uL of cells lifted from a single colony were added to 50uL of 20mM NaOH in a standard 1.5mL microcentrifuge tube. Tubes were then incubated at 95°C for 30 minutes to rupture the cells, liberating plasmid DNA. The mixture was then centrifuged at low speed and 1uL of each supernatant was transferred to a PCR strip tube containing 0.5uL of a forward primer (T7f: TAATACGACTCACTATAGGG), 0.5uL of a reverse primer (T7r:

GCTAGTTATTGCTCAGCGG), 8uL of H2O, and 10uL of GoTaq® Green Master Mix (Promega Corp). After 20 rounds of amplification, 7uL of each reaction was run via electrophoresis on a 2% agarose gel. The primers were complementary to regions of the plasmid upstream and downstream of the MCS, so colonies which produced amplicons that corresponded to the length of the synthetic gene were selected.

Growth of yeast culture for surface display: Cells from select yeast colonies were used to inoculate 5mL of c-Trp-Ura+glucose media. Cultures were grown overnight at 30°C with shaking at 250rpm. The next day, 100uL of each culture was used to inoculate 5mL aliquots of SGCAA media which were then grown overnight at 30°C with shaking at 250rpm.

Protease stability assay: To assess whether the designed proteins were folded, a yeast surface display protease stability screen was performed as in[11]. In this screen, proteins which fail to fold (and are therefore susceptible to protease) are distinguished from well-folded proteins that resist protease degradation. Cultures were sorted on a BD Accuri™ C6 Flow Cytometer, and titrations of protease trypsin were used to establish relative stability rankings (see **Figure 6**).

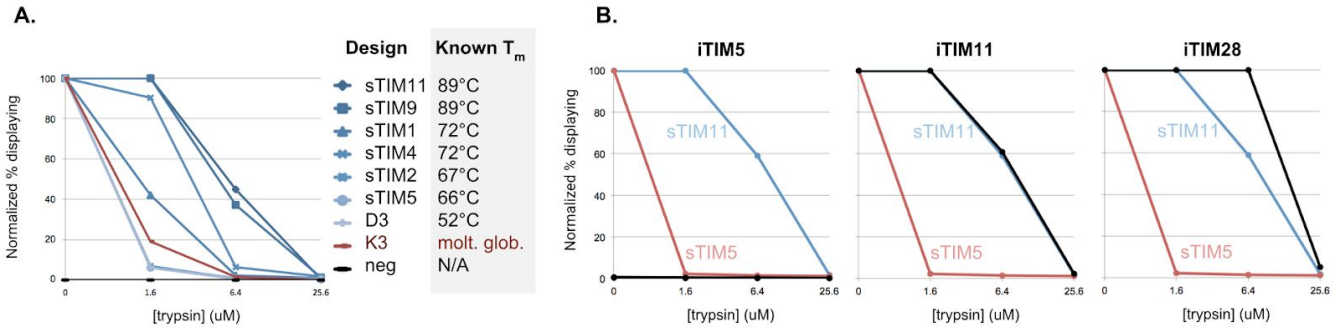

Figure 6. Protease screen reveals stable and unstable fusions. A. To determine whether this screen could accurately discriminate between well-folded and poorly folded *de novo* TIM barrels, I assayed in triplicate a collection of non-inserted *de novo* TIM barrels with known stabilities; the results perfectly recapitulate the known stability ranking of these designs, with sTIM11 found to be the most stable and K3, a known molten globule, among the least. **B.** Monomeric fusions either failed to display (*iTIM5*) or were less stable (*not shown*), as stable (*iTIM11*) or more stable than sTIM11 (*iTIM28*).

Soluble expression and purification: Synthetic genes encoding the most stable monomeric designs as well as all of the dimeric designs were cloned into pET28B, which includes a 21-residue

N-terminal sequence containing a His-tag and thrombin cleavage site to aid purification (full sequence: MGSSHHHHHHSSGLVPRGSHM). Designs were expressed in *Escherichia coli* BL21* (DE3) cells (Invitrogen) and purified as described in[11]. The expression, solubility and purity of the designed proteins were assessed by SDS-PAGE.

Biophysical analysis: Circular dichroism (CD) was used to assess the tertiary structure of the purified samples. An Aviv 62A DS spectrometer (200–260 nm) and 1 mm pathlength cuvettes were used. Samples were transferred into PBS buffer (pH 8.0) and protein concentrations were determined from their absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Scientific). GdmCl denaturation was monitored at 220 nm or 222 nm in PBS buffer at 25 °C in 1 cm pathlength cuvettes. GdmCl concentrations were controlled by an automatic titrator (Hamilton). T_m is the melting temperature where the fraction of folded proteins is equal to the fraction of unfolded proteins during temperature denaturation. CD revealed that several fusions had increased thermo- and chemostability relative to reTIM26.

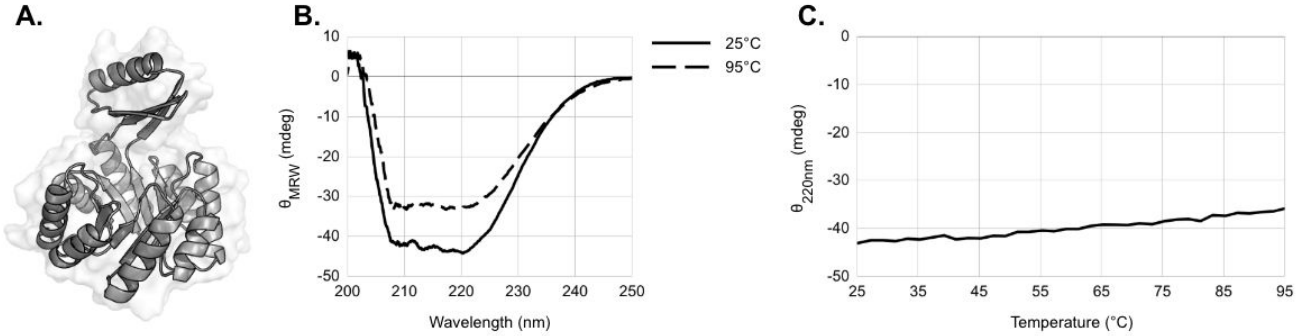

Figure 7. Circular dichroism reveals stable fusions. A. Design model of iTIM54, the most stable monomeric fusion; this design is identical to iTIM28 except that reTIM26 replaced sTIM11. **B–C.** CD spectra of iTIM54 reveal a folded, thermostable protein.

Structural Analysis

X-ray crystallography: Collaborators in the Donald Hilvert Lab at ETH Zurich were able to solve a crystal structure of one of the dimeric fusions at 2.2Å (see **Figure 8**; crystal conditions: 100 mM HEPES pH 7.5, 5-10 % iso-propanol, 10-15 % PEG 4K). Crystals were diffracted at the X06SA (PXI) beam line at the Swiss Light Source (360° with 0.1° wedges).

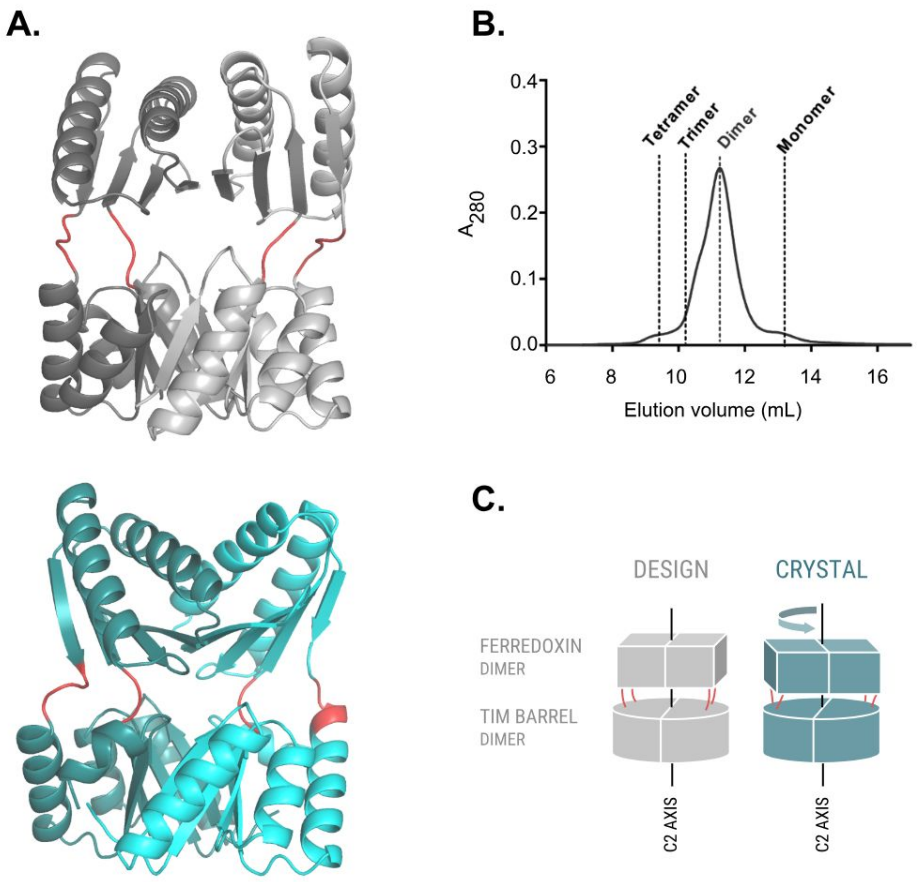

Figure 8. Crystal structure of a dimeric fusion reveals internal twist. A. Design model (*grey*) and crystal structure (*cyan*), shown with TIM barrels aligned ($RMSD_{TIM}$ = 1.7Å). **B.** Size-exclusion chromatogram of the IMAC-purified protein shows a single elution peak corresponding to the predicted dimer. **C.** Cartoon depicting the internal C2 twist of the crystal structure relative to the design model.

This protein, which had its His-tag cleaved, contained two TIM-barrel mutations (N31E and N154E) intended to create a binding site for two divalent ions. Upon recrystallization with europium(III) or terbium(III) salts, anomalous scattering at the designed metal binding site was observed, suggesting this dimeric fusion can be readily redesigned to contain active-site features such as a metal binding site.

1. Banner, D. W., Bloomer, A. c., Petsko, G. A., Phillips, D. C. & Wilson, I. A. Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem. Biophys. Res. Commun.* **72,** 146–155 (1976).
2. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* **289,** 1546–1550 (2000).
3. Höcker, B., Claren, J. & Sterner, R. Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 16448–16453 (2004).
4. Raven, J. A. Rubisco: still the most abundant protein of Earth? *New Phytol.* **198,** 1–3 (2013).
5. Tóth-Petróczy, A. & Tawfik, D. S. The robustness and innovability of protein folds. *Curr. Opin. Struct. Biol.* **26,** 131–138 (2014).
6. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19,** 1817–1819 (2010).
7. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12,** 29–34 (2016).
8. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4,** e1000002 (2008).
9. Kadumuri, R. V. & Vadrevu, R. Diversity in $\alpha\ \beta$ and $\beta\ \alpha$ Loop Connections in TIM Barrel Proteins: Implications for Stability and Design of the Fold. *Interdiscip. Sci.* **10,** 805–812 (2018).
10. Huang, P.-S. *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* **6,** e24109 (2011).

11. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357,** 168–175 (2017).
12. Lin, Y.-R., Koga, N., Vorobiev, S. M. & Baker, D. Cyclic oligomer design with de novo $\alpha\ \beta$ -proteins. *Protein Sci.* **26,** 2187–2194 (2017).
13. Dantas, G. *et al.* Mis-translation of a computationally designed protein yields an exceptionally stable homodimer: implications for protein engineering and evolution. *J. Mol. Biol.* **362,** 1004–1024 (2006).