

## Design of a Novel Globular Protein Fold with Atomic-Level Accuracy

Brian Kuhlman,<sup>1\*</sup>† Gautam Dantas,<sup>1\*</sup> Gregory C. Ireton,<sup>4</sup> Gabriele Varani,<sup>1,2</sup> Barry L. Stoddard,<sup>4</sup> David Baker<sup>1,3,†</sup>

A major challenge of computational protein design is the creation of novel proteins with arbitrarily chosen three-dimensional structures. Here, we used a general computational strategy that iterates between sequence design and structure prediction to design a 93-residue  $\alpha/\beta$  protein called Top7 with a novel sequence and topology. Top7 was found experimentally to be folded and extremely stable, and the x-ray crystal structure of Top7 is similar (root mean square deviation equals 1.2 angstroms) to the design model. The ability to design a new protein fold makes possible the exploration of the large regions of the protein universe not yet observed in nature.

There are a large but finite number of protein folds observed thus far in nature, and it is not clear whether the structures not yet observed are physically unrealizable or have simply not yet been sampled by the evolutionary process or characterized by a structural biologist. Methods for de novo design of novel protein structures provide a route to resolving this question and, perhaps more importantly, a possible route to novel protein machines and therapeutics.

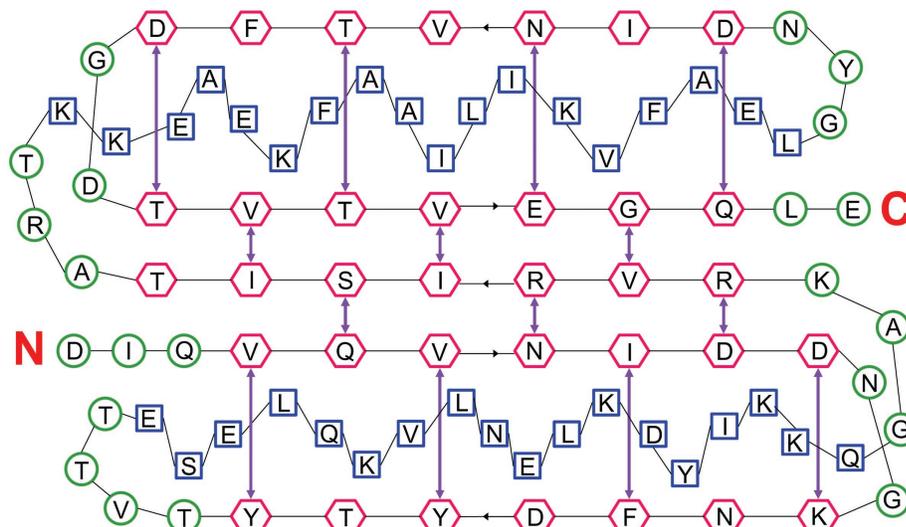
There has been considerable progress in the development of computational methods for identifying amino acid sequences compatible with a target structure (1–3), notably the pioneering complete redesign of a zinc finger protein by Mayo and co-workers (1). In general, these methods have not been used to create new protein structures but rather to redesign naturally occurring proteins so that they have enhanced stability or new functionality (4–6). Because of the strong steric restrictions in the cores of globular proteins, the packing of side chains in redesigned proteins is often quite similar to that in the original native protein (1, 7), and hence high-resolution protein backbone coordinates contain some memory of the original native sequence (8–12). When creating a new protein from scratch,

there is no such sequence memory to aid the process, and it is not even known whether the target backbone is designable. Thus, the computational design of novel protein structures is a more rigorous test of current force fields and optimization methodology than the redesign of naturally occurring proteins.

Because it is unlikely that any arbitrarily chosen protein backbone will be designable, it is essential that the design procedure include a search of nearby conformational space in addition to sequence space. With the exception of the method used by Desjarlais and Handel (2) to redesign the hydrophobic core of a small naturally occurring protein, most previous approaches have either opti-

mized the amino acid sequence for a large number of fixed backbone conformations (4, 12–14) or, as in the landmark design by Harbury and colleagues of coiled coil oligomers with a right-handed superhelical twist (15), refined the backbone conformation for a large number of fixed amino acid sequences (15, 16). The range of sequence-structure pairs that can be searched with the use of these approaches is restricted by the need to specify, in advance, a limited number of backbone conformations or amino acid sequences.

We have developed a general procedure for identifying very low free energy sequence-structure pairs that iterates between sequence optimization and structure prediction and can be applied to the design of any desired target structure. The same energy function is used to guide the search at all stages, and at each stage only the lowest energy sequence or structure identified in the previous iteration is optimized, thereby avoiding the large-scale and computationally expensive enumeration of alternative backbones or alternative sequences. Unlike the genetic algorithm of Desjarlais and Handel (2) in which randomly selected torsion angles and residue identities were simultaneously perturbed, our procedure iterates between full-scale optimization of sequence for a fixed backbone conformation and gradient-based optimization of the backbone coordinates for a fixed sequence. We used this approach to create a 93-residue  $\alpha/\beta$  protein with a topology not present in the Protein Structure Database (PDB).



**Fig. 1.** A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

<sup>1</sup>Department of Biochemistry, <sup>2</sup>Department of Chemistry, <sup>3</sup>Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. <sup>4</sup>Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA.

\*These authors contributed equally to this work.

†Present address: Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA.

‡To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

**Generation of starting models.** The target structure for the de novo design process can range from a detailed backbone model to a back-of-the-envelope sketch. Because we aimed to create a novel protein fold, we selected a topology not present in the PDB according to the Topology of Protein Structure (TOPS) server (17). A rough two-dimensional diagram was created of the target structure (Fig. 1), and constraints were identified that define the topology (Fig. 1, arrows). Three-dimensional models satisfying the constraints were then generated by assembling three- and nine-residue fragments from the PDB with secondary structures consistent with the diagram using the Rosetta de novo structure prediction methodology (18), leading to 172 backbone-only models that had the desired topology and secondary structure content and had root mean square deviations (RMSDs) from each other of 2 to 3 Å.

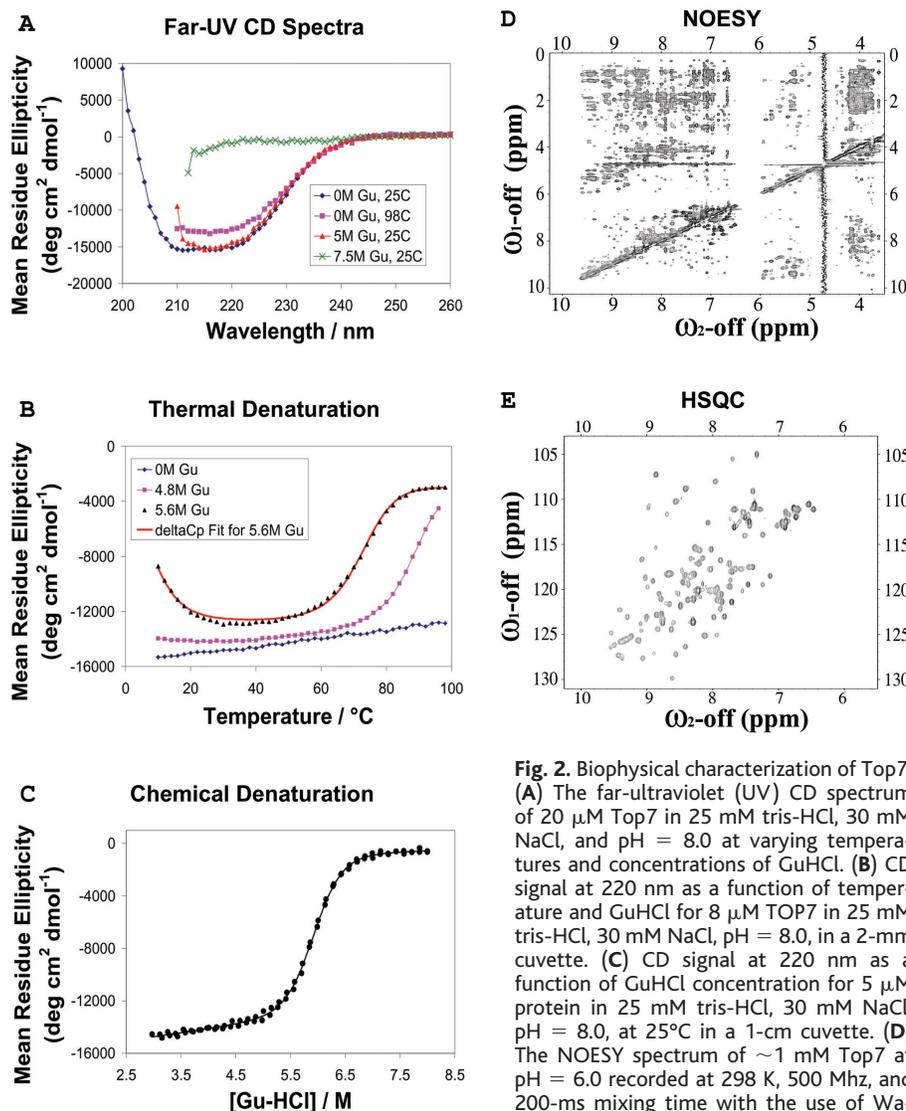
**Generation of starting sequences.** A sequence was designed for each model with the use of the RosettaDesign (9) Monte Carlo search protocol and energy function, which is dominated by a 12-6 Lennard-Jones potential, an orientation-dependent hydrogen bonding term (19), and an implicit solvation model (20, 21). All amino acids except for cysteine were allowed at 71 of the 93 positions [ $\sim 110$  rotamers from Dunbrack's library (22) per position]; the remaining 22 surface  $\beta$ -sheet positions were restricted to polar amino acids ( $\sim 75$  rotamers per position). The search through the  $110^{71} \times 75^{22}$  ( $>10^{186}$ ) rotamer combinations took  $\sim 10$  min for each model on a Pentium III (Intel) processor.

Because the starting backbone conformations were generated without regard to side-chain packing, it was anticipated that sequences with very low free energies might

not exist (i.e., the structures would not be designable). Indeed, the lowest energy sequences selected for the starting structures had energies considerably higher than those of native proteins of roughly the same size. In particular, the Lennard-Jones interaction energies for core residues were on average  $0.8 \text{ kcal mol}^{-1}$  less favorable than the interaction energies for the same residues in native protein cores. The finding that low-energy sequences do not exist for protein backbones generated without regard to side-chain packing emphasizes the need to couple sequence design with backbone flexibility for general protein design.

**Simultaneous optimization of sequence and structure.** The critical feature of the design protocol is the cycling between sequence design, as described above, and backbone optimization. The goal of the backbone optimization step, to identify the lowest free energy backbone conformation for a fixed amino acid sequence, is formally analogous to the high-resolution structure prediction problem, and we used the Rosetta program (23), which we developed for structure prediction. The backbone torsion angles were optimized with the use of a Monte Carlo minimization protocol (24) in which each move has the following parts. (i) An initial perturbation, consisting of either a small random change in the torsion angles of one to five randomly selected residues or a substitution of the backbone torsion angles of one to three consecutive residues with torsion angles from a structure in the PDB. In the latter case, the torsion angles of neighboring residues were varied to minimize the displacement of the downstream portion of the chain. (ii) A rapid optimization of side-chain conformation for all residue positions that had a higher energy after step 1 by cycling through each rotamer at each position in turn and replacing the current side-chain conformation with the lowest energy rotamer conformation. (iii) Optimization of the backbone torsion angles in a 10-residue window surrounding the site of insertion by energy minimization using a quasi-Newton method (25). Moves were accepted or rejected on the basis of the energy difference between the final minimized structure and the starting structure according to the Metropolis criterion. The same energy function was used for backbone optimization and sequence design. Each round of backbone relaxation consisted of several thousand such Monte Carlo minimization moves; a full combinatorial optimization of side-chain rotamer conformations was carried out with the use of a Monte Carlo procedure every 20 moves.

For each starting structure, five independent simulations, each with 15 cycles of sequence design and backbone optimization, were used to obtain low-energy structure sequence pairs. Final energies were comparable to those observed for naturally occurring proteins. Proteins de-



**Fig. 2.** Biophysical characterization of Top7. (A) The far-ultraviolet (UV) CD spectrum of 20  $\mu\text{M}$  Top7 in 25 mM tris-HCl, 30 mM NaCl, and pH = 8.0 at varying temperatures and concentrations of GuHCl. (B) CD signal at 220 nm as a function of temperature and GuHCl for 8  $\mu\text{M}$  TOP7 in 25 mM tris-HCl, 30 mM NaCl, pH = 8.0, in a 2-mm cuvette. (C) CD signal at 220 nm as a function of GuHCl concentration for 5  $\mu\text{M}$  protein in 25 mM tris-HCl, 30 mM NaCl, pH = 8.0, at 25°C in a 1-cm cuvette. (D) The NOESY spectrum of  $\sim 1$  mM Top7 at pH = 6.0 recorded at 298 K, 500 Mhz, and 200-ms mixing time with the use of Watergate suppression.  $\omega$ , frequency. (E) The

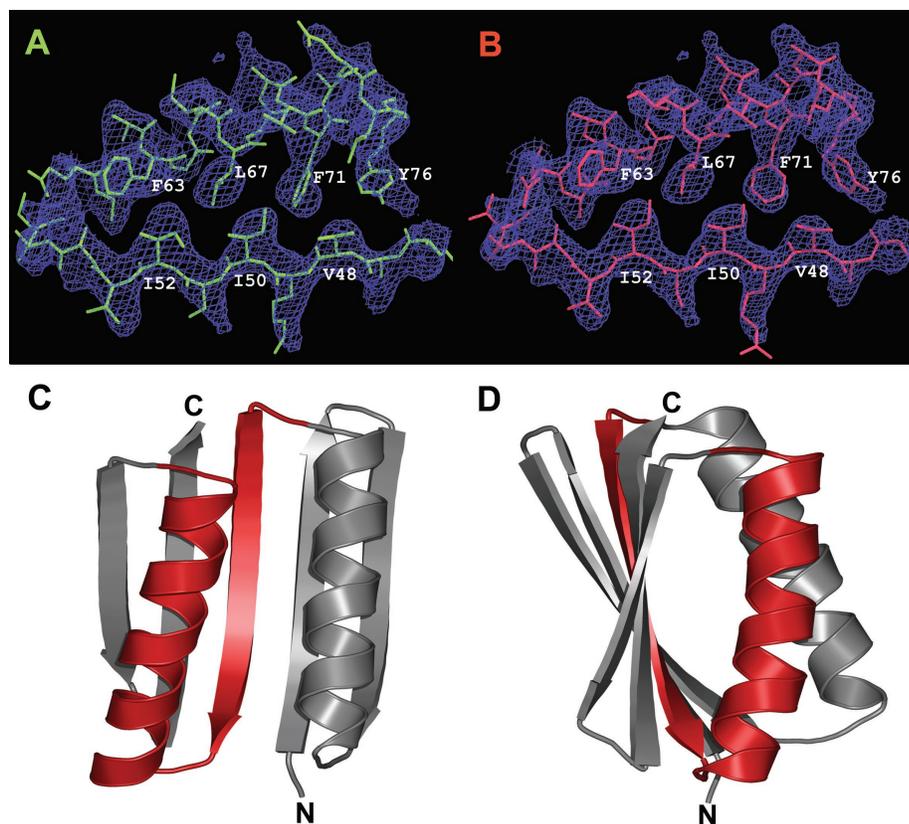
HSQC spectrum of  $\sim 1$  mM 15N-Top7 at pH = 6.0 recorded at 298 K and 500 Mhz with the use of the fast HSQC scheme of Mori *et al.* (43).

## RESEARCH ARTICLES

signed with the use of an initial version of the protocol with a damped Lennard-Jones repulsive term and a Monte Carlo optimization without the minimization step were observed experimentally to be quite stable but appeared to have somewhat molten cores (26). To optimize steric packing, the atomic radii were reparameterized on the basis of the distances of closest approach of atom pairs in high-resolution protein structures, explicit protons were included on all atoms, the penalty for atom-atom overlaps was greatly steepened, and the full Monte Carlo minimization protocol was used for varying the backbone, resulting in the generation of much lower energy sequence-structure pairs (20% of the final 860 models had more favorable Lennard-Jones energies than an average protein in the PDB). With these improvements, the protocol was used to design a protein sequence called Top7 (27).

The average Lennard-Jones energies for the buried residues in Top7 become favorable during the relaxation process (table S1), and, although the structural changes during the iterative refinement process are modest (the final protein backbone model has an RMSD of 1.1 Å from the starting model), they bring about dramatic changes in the designed sequence: Only 31% of the Top7 residues are identical to those in the starting sequence. Neither the Top7 sequence nor the sequence before the iterative sequence-structure refinement process have significant similarity to any naturally occurring protein sequence; the closest match to the Top7 sequence found with the use of PSI-BLAST (28) in the Non-Redundant protein sequence database is weaker than would be expected by random chance ( $E$  value = 1.6).

**Biophysical and structural characterization of Top7.** The folding, stability, and structure of the Top7 protein (29) were analyzed with the use of a variety of biophysical methods. The Top7 protein is highly soluble (at 25 to 60 mg ml<sup>-1</sup>) and is monomeric as determined by gel filtration chromatography. The circular dichroism (CD) spectrum of Top7 is characteristic of  $\alpha/\beta$  proteins (Fig. 2A), and the protein is remarkably thermally stable: The CD spectrum at 98°C is nearly indistinguishable from that at 25°C. At intermediate concentrations (~5 M) of guanidine hydrochloride (GuHCl), Top7 unfolds cooperatively with an increase in temperature and exhibits cold denaturation (Fig. 2B). Fitting these data to the Gibbs-Helmholtz equation gave a change in heat capacity at constant pressure ( $\Delta C_p$ ) per residue associated with unfolding of about 10 cal deg<sup>-1</sup> mol<sup>-1</sup>, a typical value for well-folded proteins of this size (30). The GuHCl-induced chemical denaturation of Top7 is cooperative, and the steep transition is characteristic of the two-state unfolding expected for a small, monomeric, single-domain protein (Fig. 2C). Fitting the chemical denaturation data to a two-state unfolding model yields a free energy of unfolding of 13.2 kcal mol<sup>-1</sup> at 25°C, indicating that Top7 is more stable than most



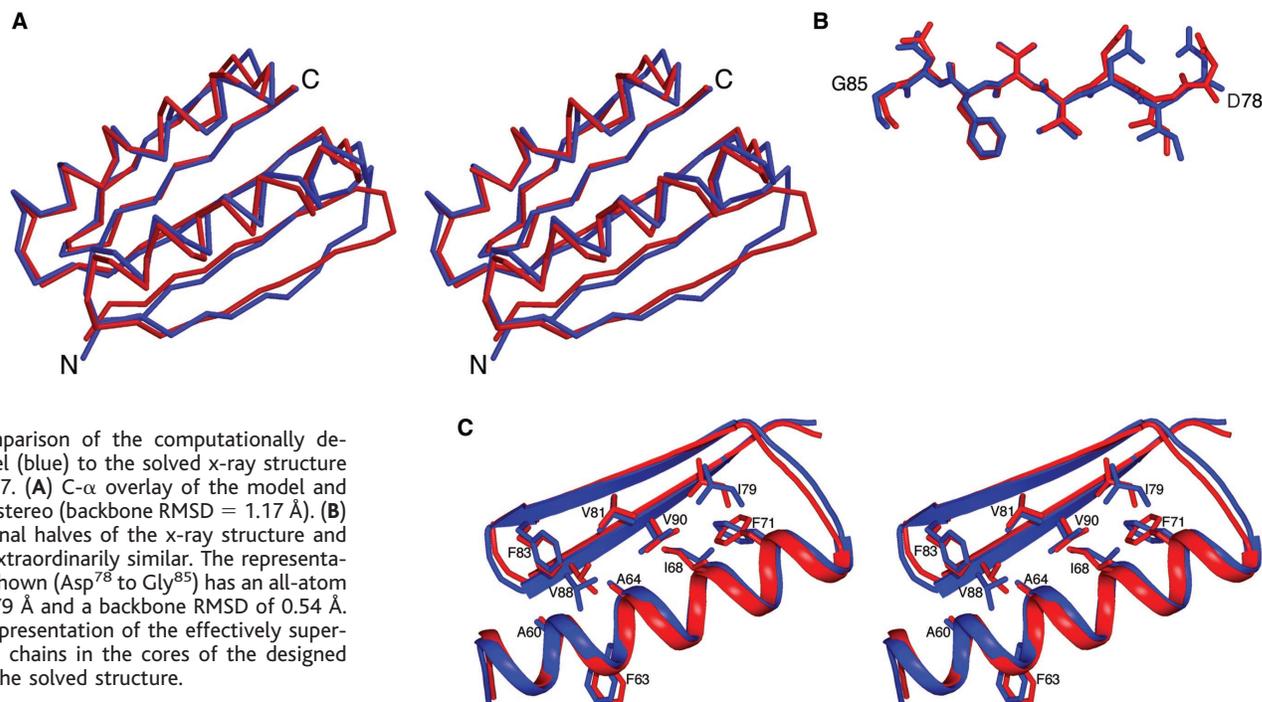
**Fig. 3.** Schematic representation of Top7 in unbiased SAD density. (A and B) Stick representations of residues 46 to 76 from the computationally designed Top7 (left, green) and from the 2.5 Å x-ray structure (right, red) are shown in unbiased density (blue). The map was generated from SAD phasing from a single SeMet-substituted variant of Top7, followed by density modification. (C and D) Ribbon diagrams of Top7 with residues 46 to 76 highlighted in red. The two diagrams are related by a 90° rotation around the vertical axis.

proteins of similar size (31). The nuclear Overhauser effect spectroscopy (NOESY) and heteronuclear single-quantum coherence (HSQC) spectra of Top7 (Fig. 2, D and E) exhibit features characteristic of a folded protein with substantial  $\beta$ -sheet content. The HSQC spectrum contains the expected number of cross peaks, and the dispersion is comparable to that of  $\alpha/\beta$  proteins of similar size. Strong backbone NH-H $\alpha$  cross peaks and the observation of H $\alpha$  resonances downfield of the water signal (to 6 parts per million) indicate the presence of a  $\beta$  sheet, whereas NH-NH peaks are consistent with a partial helical character for the protein.

Crystallization trials with Top7 yielded crystals that diffracted to 2.5 Å. Remarkably, a strong molecular replacement (MR) solution to the phase problem was found with the use of the design model. This suggested immediately that the design model was quite close to the true structure: Even the small deviations of nuclear magnetic resonance (NMR) solution structures from x-ray crystal structures can make molecular replacement searches fail. To obtain unbiased phase information, we produced and crystallized a selenomethionyl (SeMet)-substituted variant of Top7 with a surface lysine at position 37 mutated to methionine, and we solved the

x-ray crystal structure to 2.5 Å by direct rebuilding into an unbiased single-wavelength anomalous difference (SAD) electron density map (Fig. 3B) and residual difference Fourier maps (32). The final  $R_{\text{work}}$  and  $R_{\text{free}}$  were 0.268 and 0.293, respectively (table S2).

The high-resolution crystal structure reveals that the Top7 protein adopts the designed topology (Fig. 4A). Indeed, the structure is strikingly similar to the design model at atomic resolution (1.17 Å RMSD over all backbone atoms). The overall protein structure is very well ordered, with the exception of two turns (comprising residues 11 to 15 and 24 to 31), each of which exhibit elevated B-factors and poor quality electron density. The first of these two turns and the immediately adjoining residues from its neighboring strand deviate the most from the computational model. However, even in this region, the all-atom RMSD between the two models does not exceed 2.8 Å. In contrast, the C-terminal half of the x-ray structure is well ordered and very similar to the computational model; for example, the region from Asp<sup>78</sup> to Gly<sup>85</sup> has an all-atom RMSD of 0.79 Å (Fig. 4B). Many side chains in the core of the solved structure are effectively superposable with those of the designed Top7 (Fig. 4C).



**Fig. 4.** Comparison of the computationally designed model (blue) to the solved x-ray structure (red) of Top7. **(A)** C- $\alpha$  overlay of the model and structure in stereo (backbone RMSD = 1.17 Å). **(B)** The C-terminal halves of the x-ray structure and model are extraordinarily similar. The representative region shown (Asp<sup>78</sup> to Gly<sup>85</sup>) has an all-atom RMSD of 0.79 Å and a backbone RMSD of 0.54 Å. **(C)** Stereorepresentation of the effectively superposable side chains in the cores of the designed model and the solved structure.

Like the design model, the Top7 crystal structure is judged to be a novel topology by the TOPS server. The strongest structural similarity found in a Dali search of the PDB (33) is to a discontinuous portion of the 668-residue protein S-adenosylmethionine decarboxylase, which has a large 68-residue insertion between strands 1 and 2, and the third and fourth strands are connected by an unrefined loop instead of a helix. According to A. Murzin, the curator of the Structural Classification of Proteins (SCOP) database, the Top7 fold is not present in SCOP (34, 35).

**Implications.** The 1.17-Å backbone atom RMSD between the Top7 design model and the crystal structure implies that deep minima in the free energy function used in design correspond to deep minima in the actual free-energy landscape and hence are an important validation of the accuracy of current potential functions. This atomic-level accuracy contrasts sharply with the low accuracy of ab initio structure predictions for naturally occurring sequences: The most accurate structure predictions in the Critical Assessment of Structure Prediction experiments for 90- to 100-residue proteins have RMSDs greater than 4 Å from the experimentally determined structure. Why does the simultaneous optimization of sequence and structure identify the global free energy minimum, whereas the optimization of structure for fixed sequence does not? The answer may involve both of the challenges facing ab initio structure prediction, the vast size and ruggedness of the conformational space to be searched and the limited accuracy of current potential functions. The capability to alter the sequence and hence reconfigure the landscape may greatly facilitate

the search for low-free-energy protein structures as compared to standard ab initio prediction, where the sequence is fixed. In addition, Top7 lacks functional constraints, which can lead to locally suboptimal regions in native structures that are particularly challenging for structure prediction, and the more extensive optimization of the folding free energy may partially compensate for inaccuracies in the potential functions. Finally, it should be noted that the design process focused entirely on minimizing the free energy of the folded monomeric structure; attaining a highly stable new structure did not require extensive negative design against possible alternative conformations (36, 37) nor consideration of the kinetic process of protein folding (38).

The design of Top7 shows that globular protein folds not yet observed in nature not only are physically possible but can be extremely stable. This extends the earlier observation that helical coiled coil geometries not found in nature can be generated by computational protein design (15). The protein therapeutics and molecular machines of the future should thus not be limited to the structures sampled by the biological evolutionary process. The methods used to design Top7 are, in principle, applicable to any globular protein structure and open the door to the exploration and use of a vast new world of protein structures and architectures.

#### References and Notes

1. B. I. Dahiya, S. L. Mayo, *Science* **278**, 82 (1997).
2. J. R. Desjarlais, T. M. Handel, *Protein Sci.* **4**, 2006 (1995).
3. J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **193**, 775 (1987).
4. J. Reina *et al.*, *Nature Struct. Biol.* **9**, 621 (2002).

5. L. L. Looger, M. A. Dwyer, J. J. Smith, H. W. Hellinga, *Nature* **423**, 185 (2003).
6. S. M. Malakauskas, S. L. Mayo, *Nature Struct. Biol.* **5**, 470 (1998).
7. E. C. Johnson, G. A. Lazar, J. R. Desjarlais, T. M. Handel, *Structure Fold. Des.* **7**, 967 (1999).
8. P. Koehl, M. Levitt, *J. Mol. Biol.* **293**, 1161 (1999).
9. B. Kuhlman, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10383 (2000).
10. A. Jaramillo, L. Wernisch, S. Hery, S. J. Wodak, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13554 (2002).
11. K. Raha, A. M. Wollacott, M. J. Italia, J. R. Desjarlais, *Protein Sci.* **9**, 1106 (2000).
12. A. Su, S. L. Mayo, *Protein Sci.* **6**, 1701 (1997).
13. S. M. Larson, J. L. England, J. R. Desjarlais, V. S. Pande, *Protein Sci.* **11**, 2804 (2002).
14. B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. Zhang, D. Baker, *J. Mol. Biol.* **315**, 471 (2002).
15. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, *Science* **282**, 1462 (1998).
16. A. E. Keating, V. N. Malashkevich, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14825 (2001).
17. The TOPS database is available at [www.tops.leeds.ac.uk/](http://www.tops.leeds.ac.uk/).
18. P. M. Bowers, C. E. Strauss, D. Baker, *J. Biomol. NMR* **18**, 311 (2000).
19. T. Kortemme, A. V. Morozov, D. Baker, *J. Mol. Biol.* **326**, 1239 (2003).
20. T. Lazaridis, M. Karplus, *Proteins Struct. Func. Genet.* **35**, 132 (1999).
21. Materials and methods are available as supporting material on Science Online.
22. R. L. Dunbrack, F. E. Cohen, *Protein Sci.* **6**, 1661 (1997).
23. R. Bonneau *et al.*, *Proteins* (suppl. 5), 119 (2001).
24. C. Rohl, C. E. M. Straus, K. M. S. Misura, D. Baker, *Methods Enzymol.*, in press.
25. We used the Davidson-Fletcher-Powell algorithm as described by W. H. Press *et al.*, in *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, ed. 2, 1992), pp. 428–429.
26. B. Kuhlman *et al.*, data not shown.
27. The sequence of Top7 is mgDIQVQVNIIDNGKNFDYTYTVTTESELQKVLNKLKDYIKKQGAKRVRISITARTKKEAEKFAAILIKVFAELGYNDINVTDFGDTVTVEQGLegglehhhhhh; the computationally designed sequence is in uppercase and residues added to allow expression and purification are in lowercase. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H,

- His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
28. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
  29. A synthetic gene with the Top7 sequence under the control of the T7 promoter, a C-terminal 6X His tag, and codon usage optimized for *Escherichia coli* was obtained from BlueHeron Biotechnologies (Bothell, WA). After expression in *E. coli*, the protein was readily purified to >95% homogeneity with the use of nickel affinity chromatography followed by anion exchange chromatography.
  30. J. K. Myers, C. N. Pace, J. M. Scholtz, *Protein Sci.* **4**, 2138 (1995).
  31. K. W. Plaxco, K. T. Simons, D. Baker, *J. Mol. Biol.* **277**, 985 (1998).
  32. The structure of Top7\_K35M was solved by molecular replacement with the program EPMR (39) and by direct rebuilding into an unbiased SAD electron density map and residual difference Fourier maps. For molecular replacement, 19 large surface residues such as Lys, Arg, Gln, and Glu were truncated to Ala in the search model. The correlation coefficient for the initial MR search, using data to 4.0 Å resolution, was 0.52, compared with background of 0.36. For SAD phasing, the position of SeMet 37 was determined from an anomalous difference Patterson map. The initial phasing power and figure of merit for SAD phasing were 1.99 and 0.24, respectively, before density modification. An interpretable electron density map was obtained after density modification with solvent flipping with a solvent content of 43% with the use of CNS (40). An initial model was built with the use of XtalView (41) and O (42). The model was refined with CNS with the use of the mlhl target (maximum likelihood, Hendrickson-Lattman coefficients) with 5% of the data excluded for the calculation of the cross-validating free R. Of all the built residues, 88% are in the most favorable regions of Ramachandran space, and 12% are in the allowed regions. Statistics from phasing and refinement are shown in table S2.
  33. L. Holm, C. Sander, *Trends Biochem. Sci.* **20**, 478 (1995).
  34. T. J. Hubbard, A. G. Murzin, S. E. Brenner, C. Chothia, *Nucleic Acids Res.* **25**, 236 (1997).
  35. A. Murzin, personal communication.
  36. J. J. Havranek, P. B. Harbury, *Nature Struct. Biol.* **10**, 45 (2003).
  37. W. Jin, O. Kambara, H. Sasakawa, A. Tamura, S. Takada, *Structure (Cambridge)* **11**, 581 (2003).
  38. L. Mirny, E. Shakhnovich, *J. Mol. Biol.* **308**, 123 (2001).
  39. EPMR: A program for crystallographic molecular replacement by evolutionary search (C. R. Kissinger, D. K. Gehlhaar, Agouron Pharmaceuticals, La Jolla, CA).
  40. A. T. Brünger *et al.*, *Acta Crystallogr. D. Biol. Crystallogr.* **54**, 905 (1998).
  41. D. E. McRee, *J. Struct. Biol.* **125**, 156 (1999).
  42. T. A. Jones, J. Y. Zou, S. W. Cowan, M. Kjeldgaard, *Acta Crystallogr. A* **47**, 110 (1991).
  43. S. Mori, C. Abeygunawardana, M. O. Johnson, P. C. van Zijl, *J. Magn. Reson. B.* **108**, 94 (1995).
  44. We acknowledge the expert assistance of B. Shen in crystallographic phasing, modeling, and refinement of the TOP7 structure, C. Rohl for aiding in the incorporation of RosettaDesign into Rosetta, C. Strauss for helping to generate the initial models of Top7, T. Leeper for help with two-dimensional NMR studies, and R. Kleivit and the Kleivit laboratory for help with preliminary NMR characterization of Top7. Academic users can obtain licensing information for RosettaDesign at [www.unc.edu/kuhlmanpg/rosettadesign.htm](http://www.unc.edu/kuhlmanpg/rosettadesign.htm). The coordinates and structure factors for the Top7 x-ray crystal structure have been deposited in the PDB with accession code 1QYS. B.K. was supported by a fellowship from the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation. This work was also supported by NIH grants to G.V., B.L.S., and D.B.

## Supporting Online Material

[www.sciencemag.org/cgi/content/full/302/5649/1364/DC1](http://www.sciencemag.org/cgi/content/full/302/5649/1364/DC1)

Materials and Methods  
Tables S1 to S6

21 July 2003; accepted 25 September 2003

## Targeted Protein Degradation and Synapse Remodeling by an Inducible Protein Kinase

Daniel T. S. Pak\*† and Morgan Sheng†

Synaptic plasticity involves the reorganization of synapses at the protein and the morphological levels. Here, we report activity-dependent remodeling of synapses by serum-inducible kinase (SNK). SNK was induced in hippocampal neurons by synaptic activity and was targeted to dendritic spines. SNK bound to and phosphorylated spine-associated Rap guanosine triphosphatase activating protein (SPAR), a postsynaptic actin regulatory protein, leading to degradation of SPAR. Induction of SNK in hippocampal neurons eliminated SPAR protein, depleted postsynaptic density-95 and Bassoon clusters, and caused loss of mature dendritic spines. These results implicate SNK as a mediator of activity-dependent change in the molecular composition and morphology of synapses.

Synaptic activity can induce a variety of changes within postsynaptic neurons, ranging from transient posttranslational modifications to altered programs of gene expression. Long-lasting forms of synaptic plasticity require new gene expression and protein synthesis (1–3). Some activity-inducible genes may mediate the conversion of short-term responses to long-term changes by altering synaptic structure (4, 5).

Numerous activity-inducible genes have been identified (6, 7). Notably, few protein

kinases are known to be inducible by synaptic activity at the mRNA level; the best-characterized induction profiles are those of SNK and FGF-inducible kinase (FNK) (8). However, the roles of these polo family kinases in plasticity are unknown.

One reasonable expectation is that synaptic remodeling will involve the dismantling and/or reorganization of key cytoskeletal and scaffolding protein complexes. In the postsynaptic density (PSD) of mammalian excitatory synapses, actin is the major cytoskeletal element, and scaffold proteins of the PSD-95 family are important for assembling glutamate receptors with their signaling-cytoskeletal complexes (9, 10). One PSD-95-interacting partner, SPAR [spine-associated Rap guanosine triphosphatase (GTPase) activating protein (GAP)], is a multidomain postsynaptic protein that controls dendritic spine shape by regulating

actin arrangement as well as signaling by the small GTPase Rap (11). Actin dynamics and Rap activity are both regulated by synaptic activity and involved in synaptic plasticity (12–16). With its postsynaptic location in the N-methyl-D-aspartate (NMDA) receptor–PSD-95 complex, SPAR is an attractive candidate for mediating activity-dependent remodeling of synapses.

**Interaction of SPAR and SNK.** We screened for SPAR-binding proteins with the use of individual domains of SPAR (Fig. 1A) as bait in the yeast two-hybrid system (17). SPAR contains two actin regulatory domains, termed Act1 and Act2, a GAP domain specific for Rap, a PDZ domain of unknown function, and a C-terminal region (termed GKBD) that binds specifically to the guanylate kinase domain of PSD-95 (11). When the Act2 domain was used to screen a brain cDNA library, one of the positive clones (clone 19) isolated was SNK, initially identified in fibroblasts as an mRNA transcript induced by mitogenic stimulation (18). Clone 19 encoded roughly the C-terminal half of SNK protein (amino acids 395 to 682; hereon termed SNKc) (Fig. 1A), including a motif characteristic of the polo family of kinases (19, 20). Full-length SNK also bound to Act2 in the yeast two-hybrid assay. Neither full-length SNK nor clone 19 interacted with the GKBD region of SPAR (Fig. 1A).

The interaction between SNK and SPAR was confirmed with the use of an in vitro precipitation assay in which GST fused to SNKc precipitated full-length SPAR expressed in COS-7 cells but not a SPAR construct with an internal deletion of the Act2 domain (SPARΔAct2) (Fig. 1B). The isolated Act2 domain also bound GST-SNKc with greater efficiency than full-length SPAR. GST alone failed to bind any of these SPAR constructs.

Picower Center for Learning and Memory, RIKEN Massachusetts Institute of Technology (MIT) Neuroscience Research Center, Howard Hughes Medical Institute, MIT, Cambridge, MA 02139, USA.

\*Present address: Department of Pharmacology, Georgetown University, Washington, DC 20057, USA.  
†To whom correspondence should be addressed. E-mail: [dtp6@georgetown.edu](mailto:dtp6@georgetown.edu) (D.T.S.P.); [msheng@mit.edu](mailto:msheng@mit.edu) (M.S.)

**Supplementary Online Materials for:****Design of a Novel Globular Protein Fold with Atomic Level Accuracy**

Brian Kuhlman\*, Gautam Dantas\*, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker

\* These authors contributed equally to this work

**Energies and sequence for Top7 before and after alternating cycles of backbone and sequence optimization.**

before DIEITVRINNNGEDYDYKKTATTLSEINAHFEELEKHLKEENGEKITISVKLRNEKEAYW

after DIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKDYIKKQGAKRVRISITARTKKEAEK

before VAAKIKEQALRAGVETIQIDKQSDTMTATLGKQ

after FAAILIKVFAELGYNDINVTFDGDTVTVEGQLE

**Table S1.** Energies for Top7 before and after iterative cycles of backbone and sequence optimization (kcal / mole). Expected Lennard-Jones energies are derived from the average Lennard-Jones energy for each of the twenty amino acids for different degrees of burial.

	Top7 before relaxation	Final Top7 model
Lennard-Jones (LJ) attractive	-370	-385

Lennard-Jones (LJ) repulsive	28	8.6
Hydrogen bonding	-89	-80
Solvation energy	188	175
Total energy	-324	-386
LJ attractive – expected LJ attractive (avg. per buried residue)	0.3	-0.3
LJ repulsive – expected LJ repulsive (avg. per buried residue)	0.2	-0.2

## EXPERIMENTAL METHODS

### Protein Expression and Purification

Synthetic genes which place the computationally selected protein sequences under the control of the T7 promoter, with a C terminal 6X His tag, and a codon usage optimal for *Escherichia coli* (*E. coli*) were obtained from BlueHeron Biotechnologies. The gene constructs were cloned in plasmid pet29b(+) (Novagen) and expressed in the BL21(DE3)pLysS strain of *E. Coli*. Cells were grown in LB media at 37°C to an OD<sub>600</sub> of 0.6, induced with 1mM isopropyl-thio-β-D-galactosidase (IPTG), and cells were harvested after another 5 hours of growth at 37°C. Harvested cells were lysed by three freeze-thaw cycles, and soluble protein collected after centrifugation of cellular debris. Soluble protein was purified on a Ni<sup>+</sup> affinity column (Pharmacia Biotech) followed by 10<sup>4</sup>-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein was further purified on a QFF anion exchange column (Pharmacia) with a 30mM to 500mM NaCl gradient in 25mM TRIS-HCl, pH 8.0, followed by a final 10<sup>4</sup>-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein identity and purity was determined by SDS-PAGE and ESI-MALDI Mass Spectroscopy. Protein concentrations were determined by

UV absorbance at 280nm with extinction coefficients calculated using the ExPASy ProtParam tool (<http://us.expasy.org/tools/protparam.html>).

The following modifications were made to the above procedure for Top7 crystallography. A Lys<sup>37</sup> to Met<sup>37</sup> point mutant of Top7 (Top7\_K37M) was generated using the Single Quikchange Mutagenesis kit (Stratagene). Selenomethionine containing Top7\_K37M was expressed in minimal media from the *E. coli* strain BL21(DE3) adapted for growth with methionine pathway inhibition (*I*). Cells were grown in minimal media at 37°C to an OD<sub>600</sub> of 0.8 and the following amino acids were added to inhibit the methionine biosynthetic pathway: 100 mg/L lysine, threonine, phenylalanine; 75 mg/L selenomethionine; 50mg/L leucine, isoleucine, valine. Following a 15-minute incubation at 37°C, IPTG was added to induce expression and the cultures were harvested after 5 hours of growth at 37°C. Purification was performed as described.

<sup>15</sup>N-labelled Top7 was prepared by expression in M9 minimal media with <sup>15</sup>N-labelled NH<sub>4</sub>Cl. Purification was performed as described for unlabelled protein.

### **Circular Dichroism (CD)**

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260-200nm) at varying protein concentrations (15-25μM), guanidinium hydrochloride (Gu-HCl) concentrations (0-8.3M), and temperatures (0-98°C) were collected in a 1mm pathlength cuvette. Gu-HCl induced protein denaturation was followed by the change in ellipticity at 220nm in a 1cm pathlength cuvette, using a Microlab titrator (Hamilton) for

denaturant mixing. Temperature was maintained at 25°C with a Peltier device. All CD data were converted to mean residue ellipticity. Temperature induced protein denaturation was followed by the change in ellipticity at 220nm in a 2mm pathlength cuvette. To obtain a value for  $DG_U^{H2O}$ , chemical denaturation curves were fit by nonlinear least squares analysis using the linear extrapolation model as applied by Santoro and Bolen. To obtain a value for  $DC_p^\circ$ , thermal denaturation curves were fit using the Gibbs-Helmholtz equation in the form:

$$f = f_f + \frac{(f_u - f_f)}{1 + e^{\frac{-\Delta G^\circ}{RT}}}$$

$$-\Delta G^\circ = \Delta H^\circ \left(1 - \frac{T}{T_m}\right) + \Delta C_p^\circ \left\{T - T_m - T \cdot \ln\left(\frac{T}{T_m}\right)\right\}$$

where  $f$  is CD signal,  $f_f$  and  $f_u$  are the estimated CD signal for the folded and unfolded states, respectively,  $R$  is the gas constant,  $T$  is temperature,  $T_m$  is the temperature where 50% of the protein is folded,  $DG^\circ$  is the change in the Gibbs free energy for the unfolding reaction,  $DH^\circ$  is the change in enthalpy, and  $DC_p^\circ$  is the change in heat capacity.

### Nuclear Magnetic Resonance

The 2D NOESY spectrum of ~1mM Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz and 200ms mixing time using Watergate suppression. The 2D HSQC spectrum of ~1mM 15N-labelled Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz using the fast HSQC scheme of Mori et al. (2)

## Crystallization

Selenomethionyl substituted Top7\_K37M was crystallized in hanging drops (1  $\mu$ l of protein solution at 25 mg/ml with 1  $\mu$ l of well solution). The well solutions ranged from 15 - 20% PEG 3350 and 250 mM ammonium formate pH 6.6. The protein crystals grew within a day and were between 50-200  $\mu$ m on a side. They were initially transferred to a cryo-solution of well solution at 25% PEG 3350 plus 25 % (v/v) glycerol in 4 steps of increasing glycerol and flash frozen in liquid nitrogen. With this treatment the crystals diffracted in a trigonal space group (P3<sub>2</sub>21) with unit cell dimensions  $a = 35.9 \text{ \AA}$ ,  $b = 35.9 \text{ \AA}$ ,  $c = 140.6 \text{ \AA}$ . A single wavelength (0.9793  $\text{\AA}$ ) anomalous dispersion (SAD) (3) data set was collected to 2.5  $\text{\AA}$  resolution on beam-line 8.2.1 at the ALS (Advanced Light Source, Lawrence Berkeley Laboratory, Berkeley) using a four panel ADSC CCD area detector. Multiwavelength data collection (MAD phasing) was not possible due to significant radiation decay. Data were processed and scaled using HKL2000 (4).

## Structure Determination

The structure of Top7\_K35M was solved by molecular replacement with the program EPMR (5), and by direct rebuilding into an unbiased SAD electron density map and residual difference Fourier maps. For molecular replacement, 19 surface large surface residues such as Lys, Arg, and Glx were truncated to Ala in the search model. The correlation coefficient for the initial MR search, using data to 4.0  $\text{\AA}$  resolution, was 0.52, vs. background of 0.36. For SAD phasing, the position of SeMet 37 was determined from an anomalous difference Patterson map. The initial phasing power and figure of merit for SAD phasing was 1.99 and 0.24 prior to density modification. An interpretable

electron density map was obtained after density modification with solvent flipping with a solvent content of 43 % (CNS). An initial model was built using XtalView (6) and O (7). The model was refined with CNS using the mlhl target (maximum likelihood, Hendrickson-Lattman coefficients) with 5% of the data excluded for the calculation of the cross-validating free R (8). 88% of all the built residues are in the most favorable regions of Ramachandran space and 12% are in the allowed regions (9). Statistics from phasing and refinement are shown in Table S2. The structure has been deposited in the PDB with the accession code 1QYS. Examples of the experimental electron density map were generated with XtalView and Raster 3D (10). Ribbon diagrams were generated with SwissPDB Viewer (11).

**Table S2.** Crystal Structure Statistics

DATA COLLECTION	
Resolution	50-2.5Å
Space Group	P3 <sub>2</sub> 21 [primitive trigonal]
Unit Cell Dimensions	35.9 Å, 35.9 Å, 140.6 Å
Wavelength	0.9793
Asymmetric Unit	Monomer
V <sub>m</sub>	2.1 Å <sup>3</sup> /dalton
Total Reflections	144,933
Unique Reflections	6,989
Completeness / (2.59-2.5)	99.1 % / (100.0%)
R <sub>merge</sub> / (2.59-2.5)	4.5 / (34.4)
I / σ / (2.59-2.5)	37.8 (5.0)
PHASING	
Phasing Power	1.99

Figure of Merit (before/after DM)	0.24 (0.85)
REFINEMENT	
R <sub>work</sub>	0.268
R <sub>free</sub>	0.293
Number of atoms	693
Number of waters	7
Residues in most-favored regions	75 (88.2%)
Residues in additional allowed regions	7 (8.2 %)
Residues in generously allowed regions	3 (3.5%)
Residues in disallowed regions	0 (0.0%)
r.m.s.d bond length	0.0076
r.m.s.d. bond angles	1.35
Mean B value, mainchain	61.30 Å <sup>2</sup>
Mean B value, sidechain	66.67 Å <sup>2</sup>

### Energy Function

The energy of a protein was computed as a linear sum of the following 11 energy terms.

$$E_{protein} = W_{rot} E_{rot} + W_{aa|phi,psi} E_{aa|phi,psi} + W_{rama} E_{rama} + W_{atr} E_{atr} + W_{solv} E_{solv} + W_{pair} E_{pair} + W_{bb\_hbond} E_{bb\_hbond} + W_{sc\_hbond} E_{sc\_hbond} + W_{sc\_bb\_hbond} E_{sc\_bb\_hbond} + W_{pair} E_{pair} - E_{ref}$$

The weights ( $W$ ) for each term are given in a table at the end of this section. To calculate the solvation energy ( $E_{solv}$ ) and the Lennard-Jones energies ( $E_{atr}$  and  $E_{rep}$ ) the various atoms of the 20 amino acids were binned into types (Table S3).

**Table S3:** Definitions for atom types used in the energy functions

Atom Type Number	Atom type description
1	carbonyl carbon in sidechain of Asn and Gln, and guanidyl carbon in Arg
2	carboxyl carbon in Asp and Glu
3	aliphatic carbon with one hydrogen
4	aliphatic carbon with two hydrogens
5	aliphatic carbon with three hydrogens
6	aromatic ring carbon
7	nitrogen in Trp sidechain
8	nitrogen in His sidechain
9	nitrogen in Asn and Gln sidechain
10	nitrogen in Lys sidechain
11	nitrogen in Arg sidechain
12	nitrogen in Pro backbone
13	hydroxyl oxygen
14	carbonyl oxygen in Asn and Gln sidechains
15	carboxyl oxygen in Asp and Glu
16	sulfur in Cys and Met
17	backbone nitrogen
18	backbone alpha carbon
19	backbone carbonyl carbon
20	backbone oxygen
21	polar hydrogen
22	nonpolar hydrogen
23	aromatic hydrogen
24	backbone HN

### Lennard-Jones Potential ( $E_{atr}$ and $E_{rep}$ )

A standard 12-6 Lennard-Jones potential is used except there is cutoff distance below which the potential is extrapolated linearly. Favorable energies are placed in  $E_{atr}$  and unfavorable energies are placed in  $E_{rep}$ .

$$E_{atr} = \sum_i^{natom} \sum_{j>i}^{natom} \left[ \left( \frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left( \frac{r_{ij}}{d_{ij}} \right)^6 \right] e_{ij} \quad \text{if } \frac{r_{ij}}{d_{ij}} < 1.12$$

$$E_{rep} = \sum_i^{natom} \sum_{j>i}^{natom} \left[ \left( \frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left( \frac{r_{ij}}{d_{ij}} \right)^6 \right] e_{ij} \quad \text{if } 1.33 > \frac{r_{ij}}{d_{ij}} > 1.12 + \sum_i^{natom} \sum_{j>i}^{natom} y_{int\ except} - d_{ij} * slope \quad \text{if } \frac{r_{ij}}{d_{ij}} > 1.33$$

$$slope = -12e_{ij} (1.33^{13} - 1.33^7) * (1/r_{ij})$$

$$y_{int\ except} = -slope * \left( \frac{r_{ij}}{1.33} \right) + e_{ij} (1.33^{12} - 2(1.33)^6)$$

$$r_{ij} = r_i + r_j$$

$$e_{ij} = \sqrt{e_i e_j}$$

**Table S4.** Well depths and radii used for the Lennard-Jones calculations. The well depths are those used in the CHARMM19 parameter set (12). The radii were determined by fitting the Lennard-Jones potential to the distribution of distances observed between the atom types in the PDB.

Atom Type	Radii(r)	well depth (e)
1	2.00	0.1200
2	2.00	0.1200
3	2.00	0.0486
4	2.00	0.1142
5	2.00	0.1811
6	2.00	0.1200
7	1.75 <sup>1</sup>	0.2384
8	1.75 <sup>1</sup>	0.2384
9	1.75 <sup>1</sup>	0.2384
10	1.75 <sup>1</sup>	0.2384
11	1.75 <sup>1</sup>	0.2384
12	1.75 <sup>1</sup>	0.2384
13	1.55 <sup>1,2</sup>	0.1591
14	1.55 <sup>2</sup>	0.1591
15	1.55 <sup>2</sup>	0.2100
16	1.90	0.1600
17	1.75	0.2384
18	2.00	0.0486
19	2.00	0.1400
20	1.55	0.1591
21	1.00 <sup>3</sup>	0.0500
22	1.20	0.0500

23	1.20	0.0500
24	1.00 <sup>3</sup>	0.0500

<sup>1</sup>These atom types are hydrogen bond donors and when paired with atom types that are hydrogen bond acceptors(13,14,15),  $r_{ij}$  is set to 2.95, the optimal distance for hydrogen bonding. This is to prevent the repulsive portion of the Lennard-Jones term from disfavoring hydrogen bonds.

<sup>2</sup>These atom types are hydrogen bond acceptors and when paired with atom types that are hydrogen bond donors (7,8,9,10,11,12,13)  $r_{ij}$  is set to 2.95.

<sup>3</sup>These are polar hydrogens and when paired with hydrogen bond acceptors (13,14,15),  $r_{ij}$  is set to 1.95.

### Lazaridis-Karplus solvation model ( $E_{solv}$ )

An implicit solvation model developed by Lazaridis and Karplus is used to evaluate the solvation energy of a protein (13).

$$E_{solv} = \sum_i^{natom} \sum_{j>i}^{natom} \left\{ \frac{-2\Delta G_i^{free}}{4\mathbf{p}\sqrt{\mathbf{p}}\mathbf{l}_i r_{ij}^2} \exp(-d_{ij}^2) V_j - \frac{2\Delta G_j^{free}}{4\mathbf{p}\sqrt{\mathbf{p}}\mathbf{l}_j r_{ij}^2} \exp(-d_{ji}^2) V_i \right\}$$

$d_{ij}$  and  $r_{ij}$  are the same as in  $E_{atr}$ ,  $\Delta G^{free}$  is related to the solvation energy of the fully solvated atom,  $\lambda_i$  is a correlation length, and  $V$  is atomic volume. The values for the parameters are taken from Lazaridis and Karplus, except some of the  $\Delta G^{free}$  values have been perturbed to better reproduce the relative frequencies amino acids are placed in the core versus the surface during design experiments (Table S5). We have left out the

intrinsic solvation energy of each atom because the sum of these values is a constant for each amino acid and can be incorporated into the reference energies.

**Table S5.** Parameters for the Lazaridis-Karplus solvation model.

Atom Type	$DG^{free}$	$V$	$I$
1	0.00	14.7	3.5
2	-1.40	8.3	3.5
3	-0.25	23.7	3.5
4	0.52	22.4	3.5
5	1.50	30.0	3.5
6	0.08	18.4	3.5
7	-8.9	4.4	3.5
8	-4.0	4.4	3.5
9	-7.8	11.2	3.5
10	-20.0	11.2	6.0
11	-11.0	11.2	6.0
12	-1.55	0.0	3.5
13	-6.77	10.8	3.5
14	-7.8	10.8	3.5
15	-10.0	10.8	6.0
16	-4.1	14.7	3.5
17	-5.0	4.4	3.5
18	1.00	23.7	3.5
19	1.00	14.7	3.5
20	-5.00	10.8	3.5
21	0.00	0.0	3.5
22	0.00	0.0	3.5
23	0.00	0.0	3.5
24	0.00	0.0	3.5

### Rotamer Self-energy ( $E_{rot}$ )

$$E_{rot} = \sum_i^{nres} -\ln(P(rot(i) | \phi(i), \psi(i)))$$

$E_{rot}$  represents the internal energy of a rotamer and was derived from Protein Data Bank statistics by observing the probability of a particular rotamer and amino acid for a given phi angle and psi angle. These probabilities were taken directly from Dunbrack and Cohen (14). During the final design simulations we also considered rotamers with chi angles perturbed from the most commonly observed chi angles (+/- 0.5 standard deviation). These sub-rotamers were penalized by assuming a gaussian distribution about the mean using tabulated variances from Dunbrack and Cohen.

#### **Amino acid preferences for particular regions of phi, psi space ( $E_{aa|\phi,\psi}$ )**

A non-redundant set of PDB files were used to determine the probabilities for observing each of the 20 amino acids within  $10^\circ \times 10^\circ$  bins in phi,psi space,  $P(aa,|\phi,\psi)$ . The energy was calculated by taking the negative log of the probabilities.

#### **Amino acid dependent torsion potential for phi and psi ( $E_{rama}$ )**

For each of the 20 amino-acid types in each of three secondary structure types (helix, strand, and other as defined by DSSP), the frequency of (phi,psi) pairs was determined for  $10^\circ \times 10^\circ$  bins. Probabilities were calculated using added pseudocounts, and the potential calculated by taking the log of the interpolated probabilities.

#### **Residue pair potential ( $E_{pair}$ )**

$$E_{pair} = \sum_i^{nres} \sum_{j>i}^{nres} \frac{P(aa_i, aa_j | d_{ij}, env_i, env_j)}{P(aa_i | d_{ij}, env_i)P(aa_j | d_{ij}, env_j)}$$

$E_{pair}$  is derived from the probability of seeing two amino acids close together in space in the PDB database after accounting for the intrinsic probabilities of these amino acids to be in that environment (15). Two classes of environments are considered, buried and exposed, and five distance bins were used, 0-4.5, 4.5-6.0, 6.0-7.5, 7.5-9.0 and 9.0-10.5. This term was only evaluated between polar amino acids. The distances were measured between the action centers on each residue, e.g. the nitrogen on the lysine sidechain.

#### **Orientation-dependent hydrogen bonding term ( $E_{bb\_hbonds}$ , $E_{sc\_hbonds}$ , $E_{bb\_sc\_hbond}$ )**

The energy of backbone-backbone, sidechain-backbone and sidechain-sidechain hydrogen bonds were determined using a function derived from the distances and angles observed for naturally occurring hydrogen bonds in the PDB database. This function is described in detail in the supporting material of Kortemme & Baker (16). In this study we did not weight the strength of the hydrogen bonds according to their degree of burial. We removed this weight to encourage hydrogen bonds at positions that are partially buried.

#### **Energy of the unfolded state ( $E_{ref}$ )**

$$E_{ref} = \sum_i^{nres} W_{ref}(aa(i))$$