PROTEIN DESIGN

Expanding the space of protein geometries by computational design of de novo fold families

Xingjie Pan[1,2]*, Michael C. Thompson[1]†, Yang Zhang[1], Lin Liu[1], James S. Fraser[1,3], Mark J. S. Kelly[4], Tanja Kortemme[1,2,3,5]*

Naturally occurring proteins vary the precise geometries of structural elements to create distinct shapes optimal for function. We present a computational design method, loop-helix-loop unit combinatorial sampling (LUCS), that mimics nature's ability to create families of proteins with the same overall fold but precisely tunable geometries. Through near-exhaustive sampling of loop-helix-loop elements, LUCS generates highly diverse geometries encompassing those found in nature but also surpassing known structure space. Biophysical characterization showed that 17 (38%) of 45 tested LUCS designs encompassing two different structural topologies were well folded, including 16 with designed non-native geometries. Four experimentally solved structures closely matched the designs. LUCS greatly expands the designable structure space and offers a new paradigm for designing proteins with tunable geometries that may be customizable for novel functions.

The design of proteins with new and useful architectures and functions requires precise control over molecular geometries (*1*, *2*). In nature, proteins adopt a limited set of protein fold topologies (*3*–*5*) that are reused and adapted for different functions. Here we define "topology" as the identity and connectivity of secondary structure elements (Fig. 1A). Within a given topology, geometric features including length and orientations of secondary structure elements are often highly variable (*3*, *4*). These considerable geometric differences between proteins with the same topology are necessary, as they define the exquisite shape and physicochemical complementarity characteristic of protein functional sites. Creating proteins with new functions de novo therefore requires the ability to design proteins with not only different topologies, but also distinct custom-shaped geometries within these topologies optimal for each function (Fig. 1A).

Computational design has been successful in mimicking the ability of evolution to generate diverse protein structures spanning α-helical (*6*–*10*), α-β (*11*–*13*), and β-sheet (*14*, *15*) fold topologies, including novel folds (*16*). However, most design methods do not include explicit mechanisms to vary geometric features within a topology. For instance, successful design methods assemble protein structures from peptide fragments using a definition of the desired fold and topological rules derived from naturally occurring structures (*12*). Subsequent iterative cycles of fixed-backbone sequence optimization and fixed-sequence structure minimization (*16*) refine atomic packing interactions but do not create substantial changes in geometry. An exception are methods that use parametric equations to sample backbone variation (*17*) or take advantage of modular protein elements, but these methods are restricted to helical bundles (*6*, *8*, *10*) or repeat protein (*18*) architectures, respectively.

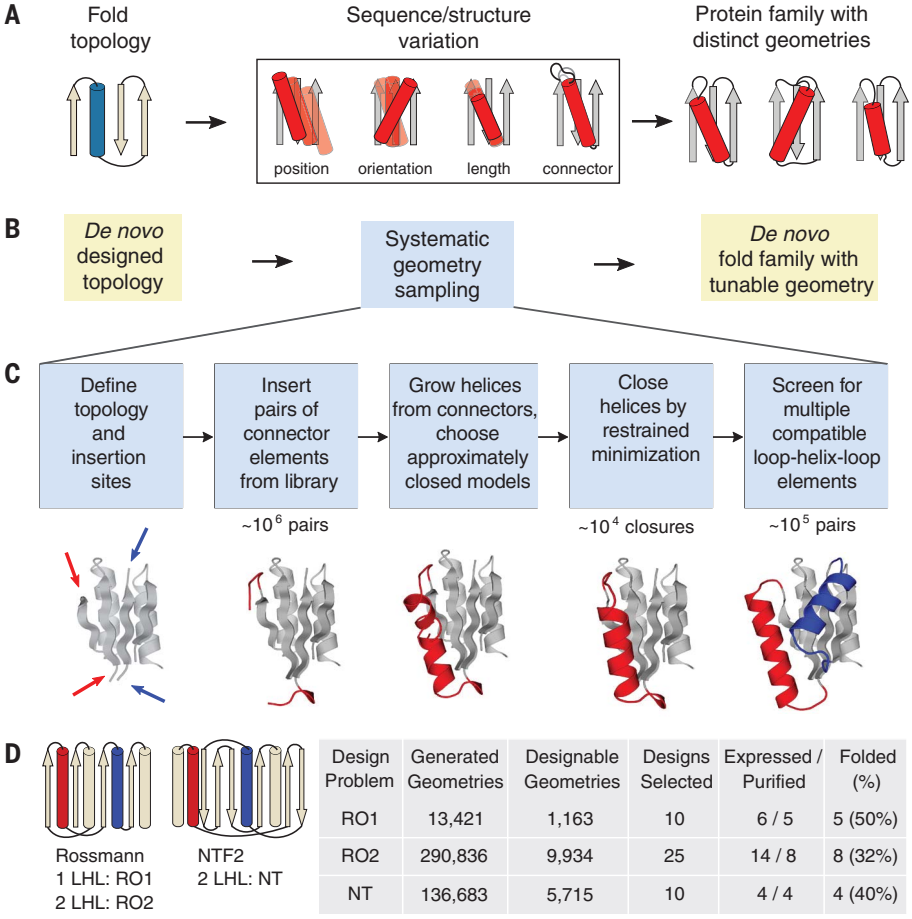
Here, we sought to develop a generalizable computational design approach that mimics the ability of evolution to create geometric variation within a given fold topology (Fig. 1). When analyzing geometric variation in protein fold families, we found that 84% of naturally occurring fold families contain variations in loop-helix-loop (LHL) elements (fig. S1). We hence reasoned that a method that systematically samples geometric variation in these units would be able to recapitulate a large fraction of geometric diversity in naturally occurring structures while also creating fold families of de novo designed proteins with tunable geometries (Fig. 1B).

To develop a generalizable method that systematically samples geometries of LHL elements, we first examined the individual connecting loop elements in native LHL units. For all LHL elements from all CATH superfamilies (*3*) of nonredundant structures, 72.8% of the loops contained five or fewer residues (fig. S2A). We extracted 313,072 loops of length 2 to 5 connecting to helices from the Rosetta nonredundant fragment database (*19*) and sorted loops into 12 libraries according to loop length and type of adjacent secondary structure (table S1). For each library, only nonredundant loops were retained (*20*); this procedure yielded between 224 and 5826 loops per library. The loop libraries had degeneracies (total number of loops divided by number of nonredundant loops in each library) ranging from 4.4 to 202 (fig. S2B), indicating that evolution frequently uses similar loop structures in different proteins. This observation suggests that the identified loop element libraries could also be used to computationally sample novel protein structures that have not been explored by nature.

We developed a protocol called loop-helix-loop unit combinatorial sampling (LUCS; Fig. 1C and fig. S3). LUCS starts with an input protein fold, which can be naturally occurring or (as in our case) de novo designed (*20*), and a definition of gaps in which to insert LHL units. The first step systematically samples all individual loop elements from our libraries (table S1). For each gap, loops are inserted at each end of the gap and any loops that clash with the input structure are removed. In a second step, all pairs of remaining loops are tested for supporting LHL units by growing helices from each loop. If helices grown from the two ends meet in the middle, excess residues are removed in the third step and the gap is closed by energy minimization with a chain-break penalty and hydrogen bond restraints. Closed LHL units with distorted hydrogen bond geometries, steric clashes, or suboptimal interactions between designed backbones and the environment are discarded (*20*). In a fourth step, combinations of LHL units at different positions can be screened to yield final structures that have multiple compatible LHL units with systematically sampled lengths and orientations.

To validate the ability of LUCS to generate distinct geometries within given fold topologies, we applied the method to three design problems (Fig. 1D). In the first two design problems, we varied one (RO1) or two (RO2) LHL units of a de novo designed protein (*12*) (PDB ID 2LV8) with a Rossmann fold topology. In the third problem, we varied two LHL units of a de novo designed protein (*21*) (PDB ID 5TPJ) with a nuclear transport factor 2 (NTF2) fold topology (NT). In principle, LUCS can sample topologies with an arbitrary number of LHL units. For the systems we tested, systematic geometry sampling generated ~10^4 LHL elements for each gap. To limit the required computing power, we screened 10^6 random combinations of LHL units and generated 10^4 to 10^5 final backbone structures for each design problem (table S2). We then applied the Rosetta FastDesign protocol (*20*) to optimize sequences for all residue positions within 10 Å from the new LHL elements. The number of designed residues for each backbone was between 33 and 87. We note that Rosetta FastDesign also introduces structural changes outside the reshaped LHL elements of the designed fold through gradient-based torsion minimization, although these changes are small [backbone heavy-atom root mean square deviation (RMSD) < 1 Å]. After sequence design, we filtered the design models computationally using a set of quality criteria

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. [2]UC Berkeley–UCSF Graduate Program in Bioengineering, University of California, San Francisco, CA, USA. [3]Quantitative Biosciences Institute, University of California, San Francisco, CA, USA. [4]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA. [5]Chan Zuckerberg Biohub, San Francisco, CA, USA.
*Corresponding author. Email: xingjiepan@gmail.com (X.P.); tanjakortemme@gmail.com (T.K.)
†Present address: Department of Chemistry and Chemical Biology, University of California, Merced, CA, USA.



Design Problem	Generated Geometries	Designable Geometries	Designs Selected	Expressed / Purified	Folded (%)
RO1	13,421	1,163	10	6 / 5	5 (50%)
RO2	290,836	9,934	25	14 / 8	8 (32%)
NT	136,683	5,715	10	4 / 4	4 (40%)

Fig. 1. LUCS sampling strategy to create de novo designed protein fold families with tunable geometries. (**A**) In nature, protein fold topologies (left) are diversified to create families of proteins with distinct geometries (right) optimized for function; α helices are shown as cylinders and β strands as arrows. The box shows schematic representations of common types of geometric variation. (**B**) The LUCS computational design protocol seeks to mimic the ability of evolution to diversify protein geometries to generate de novo designed fold families. (**C**) Schematic of the LUCS protocol for sampling LHL geometries. The reshaped LHL units are colored in red and blue. Typical numbers of models generated at major stages of the protocol are indicated. (**D**) Designed fold families. Schematic shows fold topologies and design problems: Rossmann fold with one or two reshaped LHL units (RO1 and RO2) and NTF2 fold with two reshaped LHL units (NT). Also shown are numbers for geometries generated by LUCS, designed models that passed quality filters, and experimentally characterized designs for three design problems. The rightmost column indicates the fraction of experimentally tested designs that adopted folded structures.

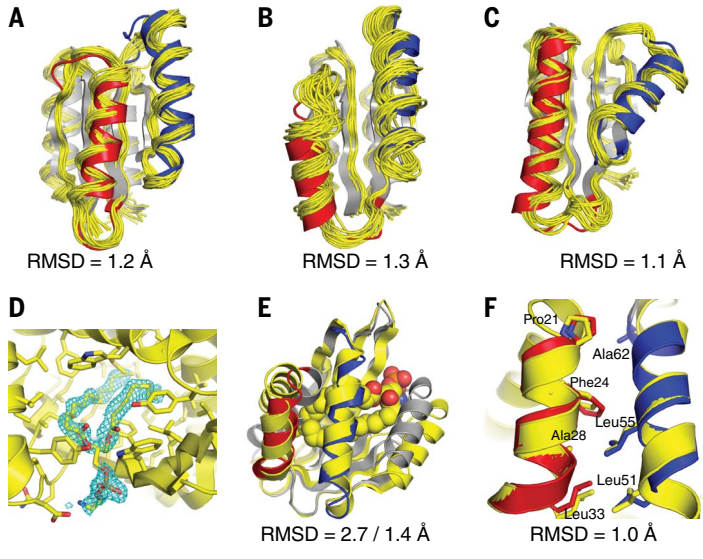


Fig. 2. Close agreement between models and experimentally determined structures of designed proteins. Experimentally determined structures are shown in yellow and design models in gray, with the reshaped LHL elements highlighted in red and blue. (**A** to **C**) Designs for the Rossmann fold topology. Computational models and NMR structures are compared for designs RO2_1 (A), RO2_20 (B), and RO2_25 (C). Also shown are the backbone heavy-atom RMSDs calculated using the lowest-energy structure from the NMR ensemble. (**D** to **F**) Design for the NTF2 fold topology. (D) The binding pocket of a phosphatidylethanolamine ligand. The $2F_{obs} - F_{calc}$ electron density map (cyan) for the ligand molecule is shown at 1.0σ level. (E) Comparison between computational model and x-ray crystal structure for design NT_9. The phosphatidylethanolamine ligand is shown in space-filling representation (carbon atoms in yellow, oxygen atoms in red, phosphorus atoms in orange, and nitrogen atoms in blue). Also shown are the backbone heavy-atom RMSDs calculated including or excluding the terminal helices, respectively. (F) Alignment between the designed helices in the computational model and the experimentally solved structure for design NT-9. The hydrophobic residues at the packing interface are shown in stick representation. The RMSD shown includes the helix backbone heavy atoms and side-chain heavy atoms displayed as sticks.

that included a minimal number of buried unsatisfied hydrogen bond donors and acceptors, tight atomic packing interactions in the protein core, and compatibility between sequences and local structures (*20*).

For each of the three design problems, we selected 50 low Rosetta energy (*22*) designs from models that passed the quality filters and had diverse conformations for further computational characterization. The Rosetta FastDesign simulations optimized low-energy sequences given a desired structure. To determine the converse—whether the desired structure is also a low-energy conformation, given the sequence—we conducted ab initio protein structure prediction simulations in Rosetta (*23*). For the Rossmann fold designs, we required the lowest-energy predicted structure to be within 1 Å Cα RMSD of the design model. For the NTF2 fold designs, we used a less strict criterion requiring a number of low-energy models to be close to the design model; this criterion enabled us to account for the more difficult problem of sampling native-like structures

for proteins larger than 100 amino acids. For each of the three design problems, 10, 25, and 10 designs that passed these tests, respectively, were chosen for experimental characterization (Fig. 1D and data S1 and S2). The designed proteins were recombinantly expressed in *Escherichia coli* and purified using His-tag affinity and size exclusion chromatography. We found that 5/10, 8/25, and 4/10 designs were monomeric and well folded for each of the three design problems, respectively, as determined by far-ultraviolet circular dichroism and one-dimensional ^{1}H and two-dimensional ^{15}N heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectroscopy (Fig. 1D, fig. S4, and table S3).

To assess whether the designed structures adopted their intended geometries, we solved structures for three designs (RO2-1, RO2-20, and RO2-25) that sampled two LHL units in the Rossmann fold topology by NMR spectroscopy, and one structure for the NTF2 fold topology designs (NT-9) by x-ray crystallography (*20*) (fig. S5 and tables S4 and S5). The experimentally solved RO2 design structures closely matched the computational models (Fig. 2, A to C), with backbone heavy-atom RMSDs between models and solved structures within 1.3 Å, core hydrophobic side chains in good agreements with the designed models (fig. S6), and five of the loops in designed LHL units well converged (fig. S7). In the crystallographic electron density map obtained at 1.5 Å resolution for the NTF2 fold design (NT-9), strong signal was clearly identifiable inside a surface pocket (Fig. 2D), which was interpreted as a bound phospholipid [1,2-diacyl-*sn*-glycero-3-phosphoethanolamine (*20*)]. The two N- and C-terminal helices (residues 1 to 20 and 113 to 128), which had not been reshaped by LUCS, were pushed apart to accommodate the ligand, leading to an overall backbone heavy-atom RMSD between design and model of 2.7 Å. However, when excluding the N- and C-terminal helices and aligning the remainder of the design, the backbone heavy-atom RMSD between the model and the solved structure was 1.4 Å (Fig. 2E). Moreover, the designed side-chain packing interactions between the reshaped helices were in excellent agreement with the design (Fig. 2F). Taken together, our structural analysis confirmed the designed geometry in the reshaped regions for all four designs. The presence of a ligand in the NT-9 design is consistent with the known ability of the NTF2 fold to bind to diverse hydrophobic small molecules. Such a result implies the possibility of introducing new functions such as ligand binding by reshaping protein geometries.

We next analyzed the magnitude of the geometric differences between our designs. We first compared the backbone heavy-atom RMSDs between the reshaped helices of all well-folded designs (Fig. 1D) after aligning the non-reshaped regions using both the design models and experimentally solved structures (Fig. 3A and fig. S8). For the designs with one LHL unit reshaped, 18 of 20 off-diagonal differences were more than 3 Å (Fig. 3A, left). For the designs with two LHL units reshaped, 55 of 68 off-diagonal differences were more than 4 Å (Fig. 3A, center and right). This scale of variation exceeds the backbone changes generated by existing flexible backbone design methods (*24*, *25*), which are typically smaller than 2 Å RMSD. For each well-folded design, we also identified the closest structures in the Protein Data Bank (PDB) using TM-align (*26*). Of the designed LHL regions in the 17 well-folded LUCS designs, 15 were significantly different (RMSD > 3 Å for designs with one LHL unit reshaped; RMSD > 4 Å for designs with two LHL units reshaped) from their closest match in the PDB (Fig. 3A and fig. S9), indicating that the design protocol not

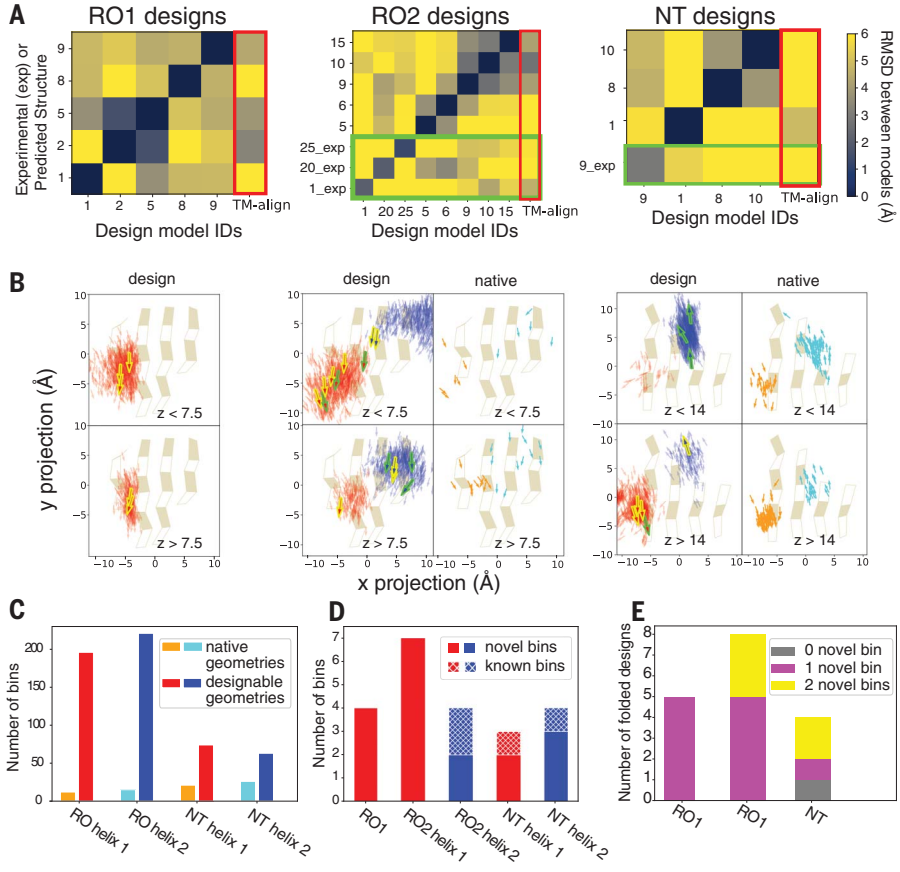


Fig. 3. Geometry space sampled by de novo designed fold families. In (A) and (B), the columns show the three design problems: Left, Rossmann fold with one designed LHL unit (RO1); center, Rossmann fold with two designed LHL units (RO2); right, NTF2 fold with two designed LHL units (NT). (**A**) Heat maps showing backbone RMSDs between the reshaped LHL regions of well-folded designs, comparing design models (*x* axis) with experimentally determined structures (_exp) or lowest-scoring models from Rosetta structure prediction (*y* axis). Green boxes show RMSDs calculated using experimentally solved structures. Red boxes (rightmost column in each grid) show the RMSDs between designs and the closest known structures found by TM-align. (**B**) Projection of centers and directions of designed helices (arrows) onto the underlying β sheets. For the RO2 (center) and NT (right) panels, distributions in designable models (Fig. 1D) are shown on the left (helices colored red and blue), and in known naturally occurring structures on the right (corresponding helices in orange and cyan). The two rows show helices on two *z*-level planes based on their distances from the β-sheet projection plane. For planes that have more than 1000 sampled structures, only 1000 randomly selected helices are shown. For the designs, experimentally confirmed folded designs are represented as bold arrows with yellow boundaries; designs with experimentally solved structures as bold arrows with green boundaries. For the natural proteins, the Rossmann fold structures are from the CATH superfamily 3.40.50.1980 and the NTF2 fold structures are from the CATH superfamily 3.10.450.50. (**C**) Number of structure bins occupied by known structures (orange, cyan) and sampled by designable models generated by LUCS (red, blue). (**D**) Structure bins occupied by well-folded designs. (**E**) Classification of the well-folded structures by the number of novel structure bins they occupy.

only generates stable structures with considerable conformational divergence, but also geometries not observed in known structures.

We further analyzed the distribution of sampled geometries and their coverage of designable backbone structure space, where a structure is defined as designable if at least one sequence folds into that structure. As a computational approximation, we defined the models that passed the quality filters after the first iteration of sequence design (*20*) as designable because they had good core packing, hydrogen bond satisfaction, and local sequence structure compatibility with the designed sequence. We projected the center and directions of the helices onto the underlying β sheets (Fig. 3B). The sampled helices from designable models at each position encompassed the distributions derived from native protein structures in the PDB (Fig. 3B, right panels). For the NTF2 fold, the distributions sampled in the designs were slightly shifted to the upper left when compared to the distributions in known structures (fig. S8). This difference could be a result of the presence of a C-terminal helix in our designs occupying the region shown at the right of the space projection, whereas C-terminal helices were often missing in the ensemble of known structures. Overall, because the number of known protein structures for a given topology is limited, the structure space covered by the known structures is much sparser than the space covered by the sampled structures. We quantified the size of structure space by dividing the six-dimensional space of helix centers and orientations into bins (*20*) (fig. S10). For the geometries sampled in this work, the known structures covered 12 to 26 bins, whereas LUCS-generated structures covered 63 to 221 bins (Fig. 3C); the smaller number of geometries in the NT designs (relative to the RO designs) could be a consequence of the additional C-terminal helix present on our NT designs restricting the accessible space of the two sampled helices. The 17 well-folded designs (Fig. 1D) sampled three to seven bins for each helix, respectively, and the majority (18/22) of these bins were not covered by known structures (Fig. 3D). All but one of the well-folded designs had at least one helix in a novel bin. Five well-folded designs had both helices in novel bins (Fig. 3E). Taken together, these results show that LUCS generates highly diverse geometries encompassing those found in nature but also exceeding known structure space, indicating that a large proportion of designable protein structure space remains unexplored.

We next sought to understand in more detail how the backbone geometries of the designed proteins were defined by the precise details of their noncovalent intramolecular interactions. The three experimentally solved Rossmann fold topology structures had distinct sequence patterns (Fig. 4A), resulting in distinct packing arrangements (Fig. 4, B and C) in their hydrophobic cores. The β sheets favored β-branched residues, as expected, whereas the side-chain sizes varied across different designs and resulted in differential hydrophobic packing. In particular, we observed previously described knob-socket–type packing motifs (*27*) (Fig. 4C and fig. S11) where nonpolar side chains fit into pockets formed by three residues on helices. These arrangements result in matched geometries between the side chains from sheets and helices that likely contribute to specifying the three-dimensional arrangement of the helices (*20*) (fig. S12). We also applied tertiary motif analysis using MASTER (*28*). For all well-folded designs, we were able to match tertiary motifs to both the designed loops and interacting secondary structure elements (fig. S13). Moreover, we identified side chains mediating helix-helix, helix-sheet, and helix-loop interactions that are similar in our designs and the corresponding matched tertiary motifs (Fig. 4D). Despite the close match between the local structures in the design and the tertiary motifs, the source proteins of the motifs had overall structures very different from the designs (fig. S13). Because tertiary motif information was not used directly in LHL backbone sampling or side-chain design, we conclude that recurrent tertiary motifs can be recapitulated solely by our LUCS sampling protocol and the Rosetta energy function (*22*).

Previous key achievements in de novo design (*11*–*15*, *21*) focused on designing one or a

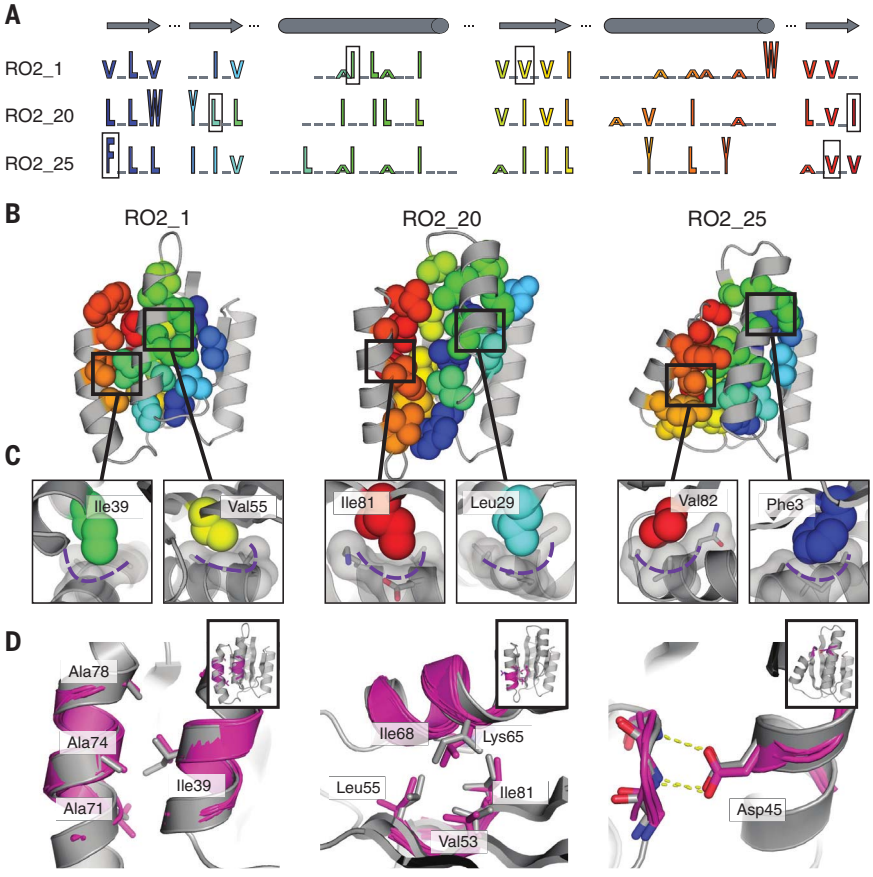


Fig. 4. Structural features encoding distinct protein geometries. (**A**) Sequence patterns of the hydrophobic cores in three designed models for the Rossmann fold, aligned by corresponding secondary structure elements (top). Hydrophobic residues are shown as letters in rainbow colors (A, Ala; F, Phe; I, Ile; L, Leu; V, Val; W, Trp; Y, Tyr) ordered by position in the primary protein sequence and scaled by side-chain size. Gray line segments indicate positions of surface-exposed polar residues. The residues in the boxes are the knob residues shown in (C). (**B**) Atomic packing of hydrophobic cores in the three experimentally determined structures for the Rossmann fold (Fig. 2). The hydrophobic side chains in the designed cores are shown as spheres. (**C**) Knob-socket packing motifs found in the designs. Three residues on a helix (gray sticks and surfaces) form a socket accommodating a knob residue, shown as colored spheres. (**D**) Examples of tertiary motifs matching the designed LHL structures. Designed structures are shown in gray; matched motifs are shown in magenta. Side chains of the best-matched tertiary motifs and design models are shown as sticks. Insets indicate location of the tertiary motif in the structure in the same orientation as in (B).

few structures for diverse non–helical bundle topologies by deriving design rules for specific topologies to identify the most favorable "idealized" geometries. This topology-centric strategy typically finds deep energy minima and thereby succeeds in overcoming errors in energy functions to create highly stable de novo folds. In contrast, natural and LUCS-generated structure families adopt non-ideal geometric features such as diverse helix positions, orientations, lengths, and conformations of connector elements, and exploring these non-ideal regions presents extra challenges (*29*). Nonetheless, we show here that LUCS achieves accurate atom-level control over diverse geometries, and our designs are not notably less stable than their de novo designed starting points (fig. S4). This success could at least partially be explained by the ability of LUCS to recover three-dimensional packing arrangements that are recurrent in nature (Fig. 4D and fig. S13), but without using this information as input.

We envision many applications for LUCS to precisely tune protein geometries for new protein functions that require atom-level control. The generalizable strategy underlying LUCS (Fig. 1C) does not require prior definition of structural variation based on design rules identified in native structures (*21*, *30*). New protocols could exploit this ability to flexibly tune protein geometries during design simulations while simultaneously building new functional sites for ligand binding or protein-protein recognition. The systematic sampling of protein geometries should also enable the design of dynamic proteins (*31*) that can switch between multiple distinct de novo designed conformations. Methods such as LUCS bring control over designable protein geometry space for arbitrary functions within reach.

REFERENCES AND NOTES

1. D. Baker, *Protein Sci.* **19**, 1817–1819 (2010).
2. K. Kundert, T. Kortemme, *Biol. Chem.* **400**, 275–288 (2019).
3. N. L. Dawson *et al.*, *Nucleic Acids Res.* **45**, D289–D295 (2017).
4. N. K. Fox, S. E. Brenner, J. M. Chandonia, *Nucleic Acids Res.* **42**, D304–D309 (2014).
5. J. Hou, S. R. Jun, C. Zhang, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 3651–3656 (2005).
6. P. S. Huang *et al.*, *Science* **346**, 481–485 (2014).
7. T. M. Jacobs *et al.*, *Science* **352**, 687–690 (2016).
8. A. R. Thomson *et al.*, *Science* **346**, 485–488 (2014).
9. R. B. Hill, D. P. Raleigh, A. Lombardi, W. F. DeGrado, *Acc. Chem. Res.* **33**, 745–754 (2000).
10. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, *Science* **282**, 1462–1467 (1998).
11. G. J. Rocklin *et al.*, *Science* **357**, 168–175 (2017).
12. N. Koga *et al.*, *Nature* **491**, 222–227 (2012).
13. P. S. Huang *et al.*, *Nat. Chem. Biol.* **12**, 29–34 (2016).
14. J. Dou *et al.*, *Nature* **561**, 485–491 (2018).
15. E. Marcos *et al.*, *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).
16. B. Kuhlman *et al.*, *Science* **302**, 1364–1368 (2003).
17. F. Crick, *Acta Crystallogr.* **6**, 689–697 (1953).
18. T. J. Brunette *et al.*, *Nature* **528**, 580–584 (2015).
19. D. Gront, D. W. Kulp, R. M. Vernon, C. E. Strauss, D. Baker, *PLOS ONE* **6**, e23294 (2011).
20. See supplementary materials.
21. E. Marcos *et al.*, *Science* **355**, 201–206 (2017).
22. H. Park *et al.*, *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
23. P. Bradley, K. M. Misura, D. Baker, *Science* **309**, 1868–1871 (2005).
24. J. A. Davey, R. A. Chica, *Methods Mol. Biol.* **1529**, 161–179 (2017).
25. N. Ollikainen, C. A. Smith, J. S. Fraser, T. Kortemme, *Methods Enzymol.* **523**, 61–85 (2013).
26. Y. Zhang, J. Skolnick, *Nucleic Acids Res.* **33**, 2302–2309 (2005).
27. H. Joo, A. G. Chavan, J. Phan, R. Day, J. Tsai, *J. Mol. Biol.* **419**, 234–254 (2012).
28. J. Zhou, G. Grigoryan, *Protein Sci.* **24**, 508–524 (2015).
29. D. Baker, *Protein Sci.* **28**, 678–683 (2019).
30. Y. R. Lin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5478–E5485 (2015).
31. J. A. Davey, A. M. Damry, N. K. Goto, R. A. Chica, *Nat. Chem. Biol.* **13**, 1280–1285 (2017).

ACKNOWLEDGMENTS

We thank M. Wu, N. Hoppe, and members of the Kortemme lab for discussion. **Funding:** Supported by NIH grant R01-GM110089 and NSF grant DBI-1564692 (T.K.); the UCSF Program for Breakthrough Biomedical Research, funded in part by the Sandler Foundation; a UCSF Discovery Fellowship (X.P.); and a NIH F32 Postdoctoral Fellowship (M.C.T.). T.K. is a Chan Zuckerberg Biohub Investigator. **Author contributions:** X.P. conceived the idea for the project; X.P. and T.K. conceived the computational and experimental approach; X.P. developed and performed the computational design; X.P. and Y.Z. performed the majority of the experimental characterization; X.P. and M.J.S.K. determined the NMR structures; X.P., M.C.T., L.L., and J.S.F. determined the crystal structure; J.S.F., M.J.S.K., and T.K. provided guidance, mentorship, and resources; and X.P. and T.K. wrote the manuscript with contributions from the other authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Coordinates and structure files have been deposited to the Protein Data Bank (PDB) with accession codes 6VG7, 6VGA, 6VGB, and 6W90. NMR data have been deposited to the Biological Magnetic Resonance Data Bank with accession codes 30706, 30707, and 30708. All other relevant data are available in the main text or the supplementary materials. Rosetta source code is available from rosettacommons.org.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6507/1132/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S14
Tables S1 to S5
Data S1 and S2
References (*32*–*61*)

6 April 2020; accepted 14 July 2020
10.1126/science.abc0881

Science

Expanding the space of protein geometries by computational design of de novo fold families

Xingjie Pan, Michael C. Thompson, Yang Zhang, Lin Liu, James S. Fraser, Mark J. S. Kelly and Tanja Kortemme

Science **369** (6507), 1132-1136.
DOI: 10.1126/science.abc0881

Exploring the design landscape

Protein design typically selects a protein topology and then identifies the geometries (secondary-structure lengths and orientations) that give the most stable structures. A challenge for this approach is that functional sites in natural proteins often adopt nonideal geometries. Pan *et al.* addressed this issue by exploring the diversity of geometries that can be sampled by a given topology. They developed a computational method called LUCS that systematically samples geometric variation in loop-helix-loop elements and applied it to two different topologies. This method generated families of well-folded proteins that include structures with non-native geometries. The ability to tune protein geometry may enable the custom design of new functions.

Science, this issue p. 1132