Supplemental Information for

De Novo Protein Fold Design Through Sequence-Independent Fragment Assembly Simulations

Robin Pearce, Xiaoqiang Huang, Gilbert S. Omenn, and Yang Zhang

Supplementary Tables

Table S1. Results of FoldDesign starting from distance restraints extracted from the native structures. All metrics were computed between the designed and native structures. Here, MAE is the mean absolute error between the Cα distance maps from the designed and native structures and is calculate by $MAE = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n}$, where x_i is a distance from a designed structure, y_i is the corresponding distance from the native structure, and n is the number of considered distances.

Protein Type	MAE (Å)	TM-score	RMSD (Å)
All	0.148	0.993	0.31
A	0.115	0.993	0.27
B	0.130	0.992	0.32
α/β	0.154	0.994	0.31

Table S2. Results of AlphaFold2 modeling using different MSA generation methods for the 354 native protein structures. P-values were calculated using paired, two-sided Student's t-tests between the results by DeepMSA and the other approaches. In the table, the 'DeepMSA MSA' option refers to the results obtained by AlphaFold2 starting from the MSAs identified by searching the original native sequences using the DeepMSA program, the 'Designed MSA' option refers to the results obtained by AlphaFold2 when starting from the alignment of 100 designed sequences by EvoEF2 or RosettaFixBB, and the 'Single Sequence' option refers to the results for AlphaFold2 modeling starting from the single lowest energy designed sequence produced by EvoEF2 or RosettaFixBB.

AlphaFold2 Input	**TM-score (*p*-value)**	**RMSD Å (*p*-value)**	**#TM-score ≥ 0.5[a]**
Native sequences			
DeepMSA MSA	0.913 (*)	1.99 (*)	350
Sequences designed by EvoEF2			
Designed MSA	0.852 (3.8E-13)	2.48 (1.8E-02)	345
Single Sequence	0.506 (7.7E-113)	12.45 (3.4E-91)	179
Sequences designed by RosettaFixBB			
Designed MSA	0.837 (2.5E-18)	2.72 (3.3E-04)	344
Single Sequence	0.482 (1.3E-120)	12.08 (5.4E-94)	161

[a]This column indicates the number of AlphaFold2 models with correct global folds (i.e., TM-score ≥0.5).

Table S3. Results of AlphaFold2 modeling starting from the designed sequences for the FoldDesign and Rosetta scaffolds. *P*-values were calculated using paired, two-sided Student's t-tests.

Method	TM-score (*p*-value)	RMSD (*p*-value)	# TM-score ≥ 0.5
Sequences designed by EvoEF2			
FoldDesign	**0.714 (*)**	**3.66 (*)**	**324**
Rosetta	0.663 (1.1E-07)	5.10 (4.6E-09)	301
Sequences designed by RosettaFixBB			
FoldDesign	**0.696 (*)**	**4.13 (*)**	**315**
Rosetta	0.670 (0.004)	4.95 (3.0E-4)	310

Table S4. Local structure characteristics of the designed folds by FoldDesign. The table illustrates the overall Molprobity scores (MP-score) and additional structure quality metrics output by the Molprobity program for the 354 native structures (Native) as well as the 354 FoldDesign scaffolds (All Designs), the 79 novel designs (Novel Designs), and the 275 designs with native fold analogs (Analogous Designs).

Structures	MP-Score	Rama Outliers (%)	Rama Favorable (%)	Rotamer Outliers (%)	Clash Score	RMS Bonds	RMS Angles
Native	1.19	1.19	93.95	5.53	0.00	0.01	1.48
All Designs	1.59	0.46	96.91	0.05	0.00	0.04	3.43
Novel Designs	1.66	0.42	96.58	0.06	0.00	0.04	3.43
Analogous Designs	1.57	0.47	97.00	0.04	0.00	0.04	3.43

Table S5. Empirically observed acceptance probabilities for swaps between adjacent replicas during the FoldDesign simulations for the 354 test proteins.

Replica Number	Fraction of Accepted Swaps
1	0.771
2	0.767
3	0.759
4	0.742
5	0.728
6	0.716
7	0.695
8	0.686
9	0.680
10	0.687
11	0.690
12	0.697
13	0.708
14	0.714
15	0.718
16	0.723
17	0.733
18	0.735
19	0.739
20	0.749
21	0.754
22	0.758
23	0.762
24	0.769
25	0.770
26	0.776
27	0.778
28	0.780
29	0.778
30	0.776
31	0.777
32	0.773
33	0.771
34	0.762
35	0.755
36	0.737
37	0.720
38	0.689
39	0.643

Table S6. Feature values μ_{kl}/δ_{kl} for each hydrogen bonding restraint type, T_k, in Eq. S4. The features are presented as averages/standard deviations.

Restraint Type	Secondary Structure	$f_1: D(O_i, H_j)$ (Å)	$f_2: A(C_i, O_i, H_j)$ (degrees)	$f_3: A(C_i, O_i, H_j)$ (degrees)	$f_4: T(C_i, O_i, H_j, N_j)$ (degrees)
T_1	Helix, $j = i + 4$	2.00/0.53	147/10.58	159/11.25	160/25.36
T_2	Helix, $j = i + 3$	2.85/0.32	89/7.70	111/8.98	-160/7.93
T_3	Parallel Strand	2.00/0.30	155/11.77	164/11.29	180/68.96
T_4	Antiparallel Strand	2.00/0.26	151/12.38	163/11.02	-168/69.17

Supplementary Figures

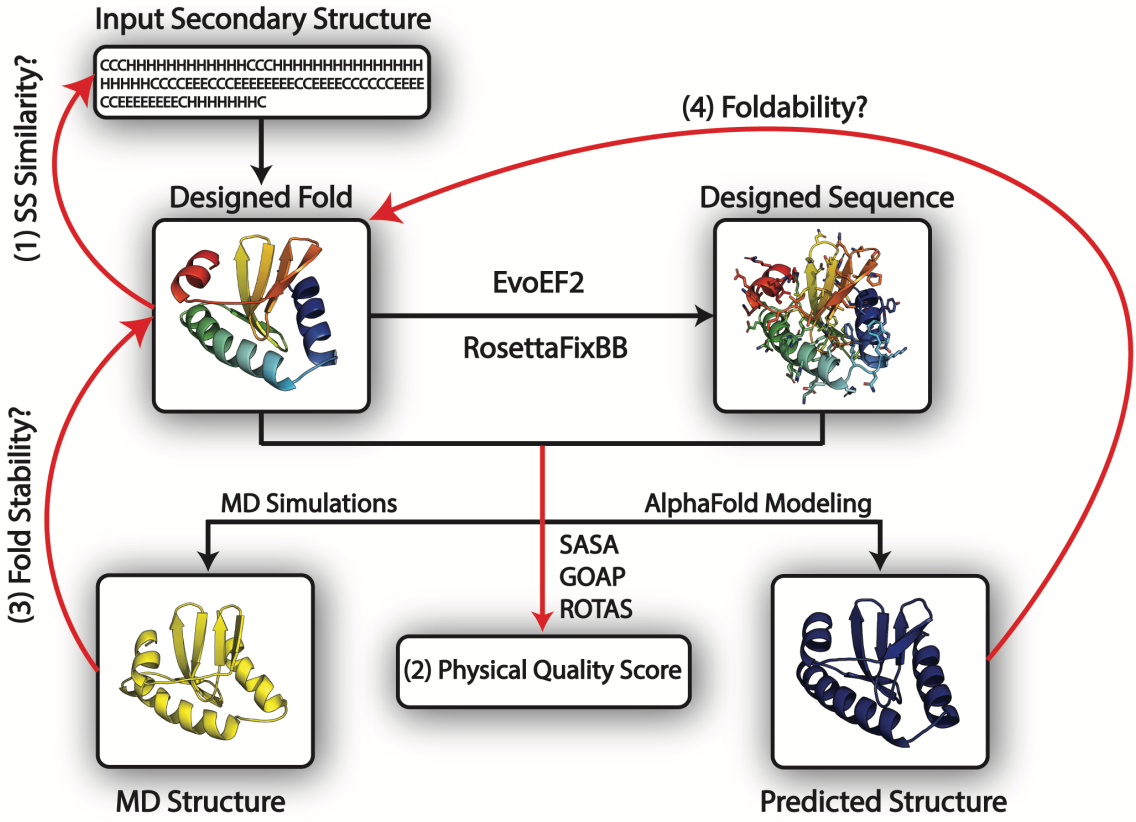

Figure S1. Illustrations of the strategies used to evaluate the quality of the FoldDesign scaffolds. The red lines mark the four criteria used to assess the FoldDesign scaffolds: (1) the secondary structure similarity between the input secondary structures and the secondary structures of the scaffolds designed by FoldDesign; (2) the physical quality score including hydrophobic core formation and statistical energies; (3) the fold stability assessed by the structural similarity (TM-score/RMSD) between the FoldDesign scaffolds and the final models after constraint-free molecular dynamic simulations (MD); (4) the foldability as determined by the structural similarity between the FoldDesign scaffolds and the predicted models by AlphaFold.

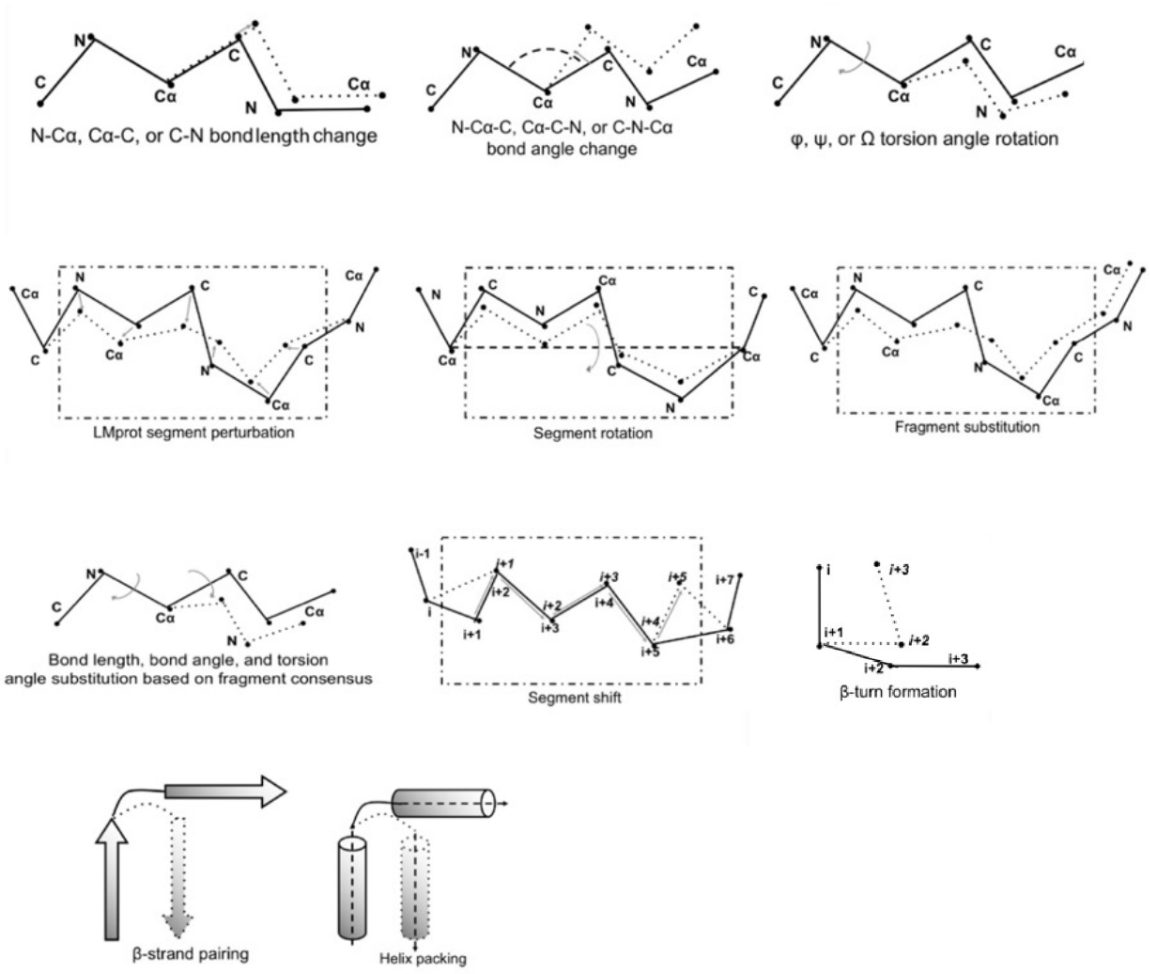

Figure S2. Depiction of the conformational movements used by FoldDesign, with explanations in Supplementary Text S1.

FoldDesign Energy: -145.5 k_BT

Figure S3. Structure and FoldDesign energy for the native 1ec6A fold.

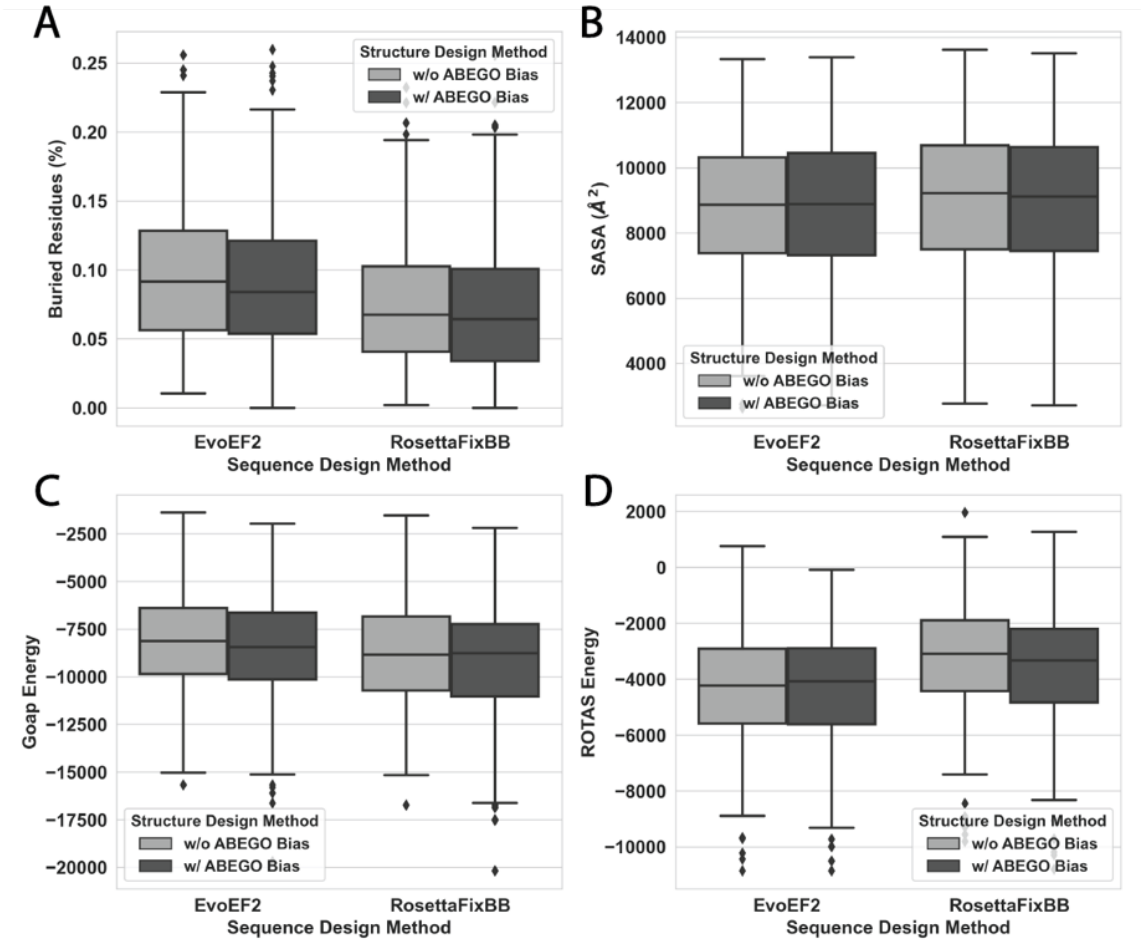

Figure S4. Comparison of the physical characteristics and energies for the designed folds by Rosetta with and without ABEGO bias on the 354 test proteins, where the sequence for each scaffold was designed by EvoEF2 and RosettaFixBB. A) Proportion of buried residues is plotted for each design, where a buried residue was defined as having a relevant solvent accessible surface area <5%. B) Solvent accessible surface area (SASA) for each design. C-D) Energies for each design calculated by GOAP and ROTAS.

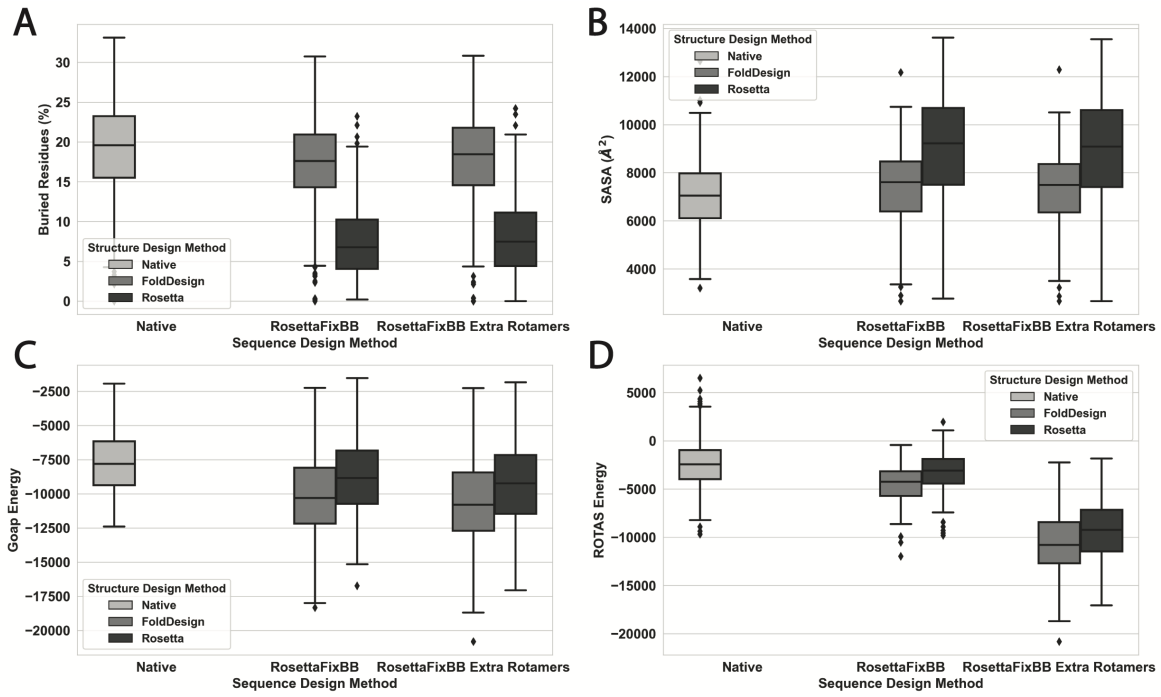

Figure S5. Comparison of the physical characteristics and energies for the designed folds by FoldDesign and Rosetta on the 354 test proteins, where the sequence for each scaffold was designed by RosettaFixBB with (RosettaFixBB Extra Rotamers) or without (RosettaFixBB) sub-rotamer sampling for the χ_1 and χ_2 angles. The native designation represents the proteins from which the secondary structures of the designed folds were derived. A) Proportion of buried residues is plotted for each protein, where a buried residue was defined as having a relevant solvent accessible surface area <5%. B) Solvent accessible surface area (SASA) for each protein. C-D) Energies for each protein calculated by GOAP and ROTAS.

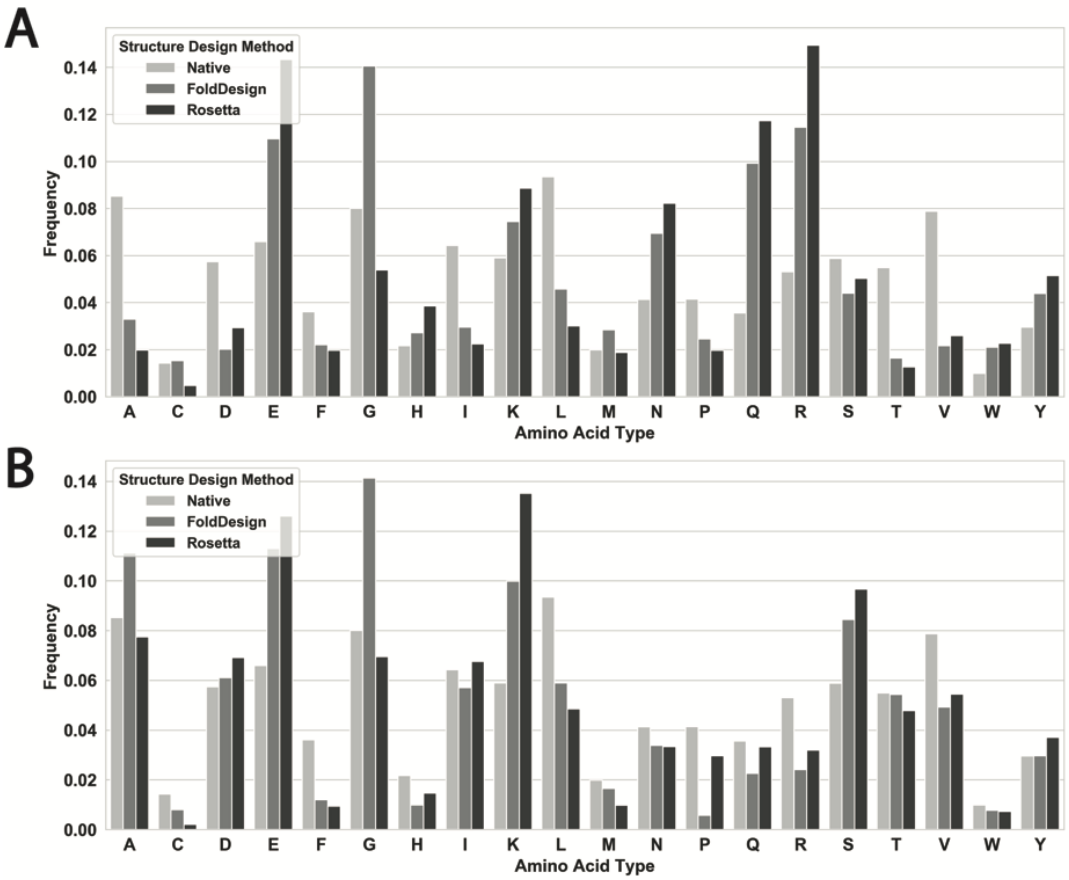

Figure S6. Comparison of the amino acid distributions for the native proteins as well as the FoldDesign and Rosetta scaffolds whose sequences were designed by EvoEF2 (A) and RosettaFixBB (B), respectively. The native designation represents the 354 proteins from which the secondary structures of the designed folds were derived.

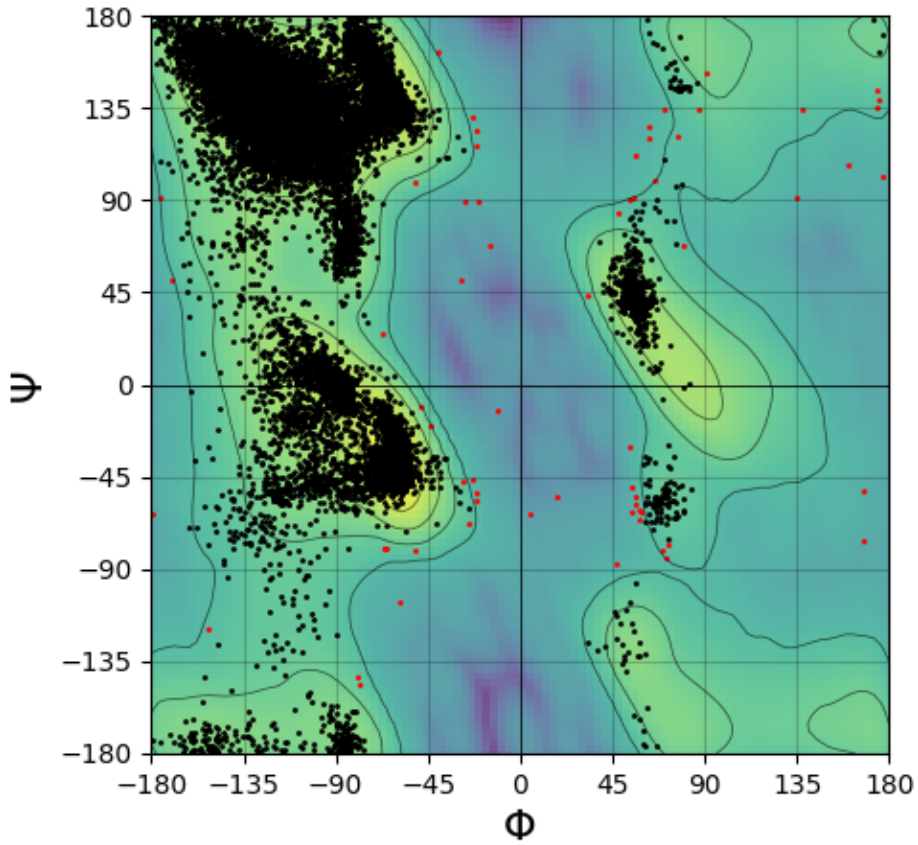

Figure S7. Ramachandran plot derived from the 354 FoldDesign scaffolds, where favored/allowable torsion angles are plotted using black circles and outliers are plotted using red circles.

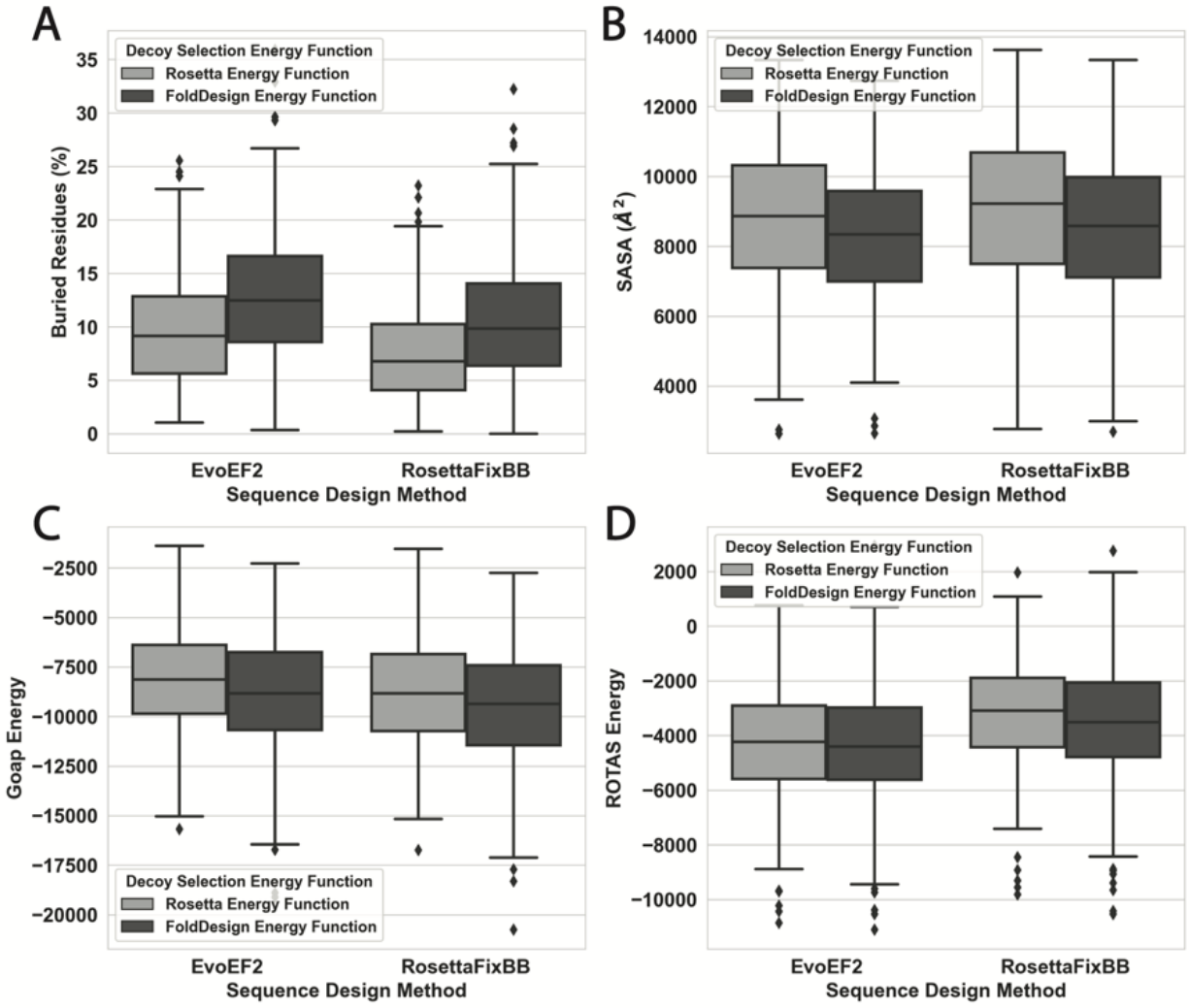

Figure S8. Comparison of the physical characteristics and energies for the designed folds by Rosetta on the 354 test proteins, where the final designs were selected using either the Rosetta centroid energy function or the FoldDesign energy function. A) Proportion of buried residues is plotted for each protein, where a buried residue was defined as having a relevant solvent accessible surface area <5%. B) Solvent accessible surface area (SASA) for each protein in the test set. C-D) Energies for each protein calculated by GOAP and ROTAS respectively.

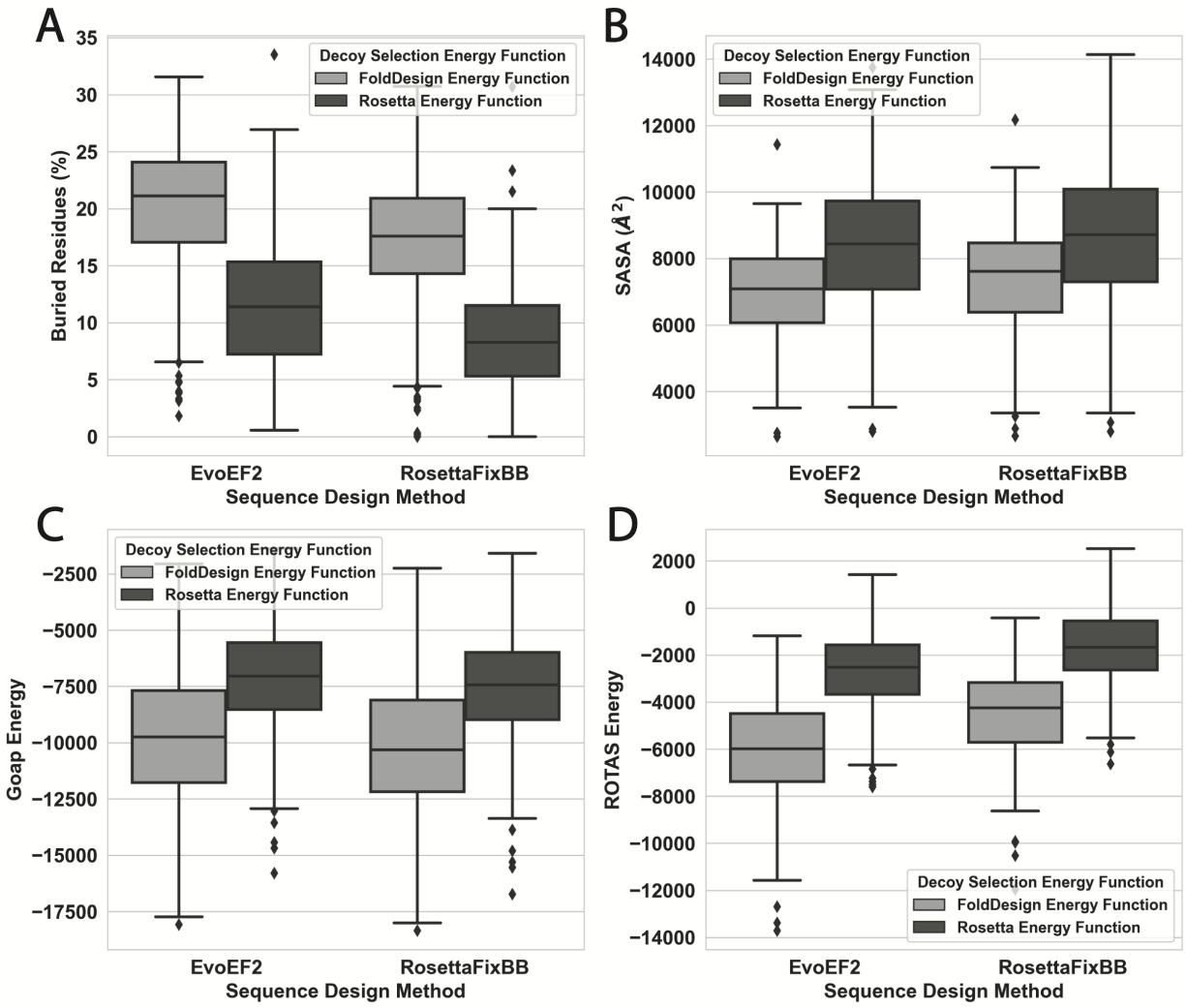

Figure S9. Comparison of the physical characteristics and energies for the designed folds by FoldDesign on the 354 test proteins, where the final designs were selected using either the FoldDesign energy function or the Rosetta centroid energy function. A) Proportion of buried residues is plotted for each protein, where a buried residue was defined as having a relevant solvent accessible surface area <5%. B) Solvent accessible surface area (SASA) for each protein in the test set. C-D) Energies for each protein calculated by GOAP and ROTAS respectively.

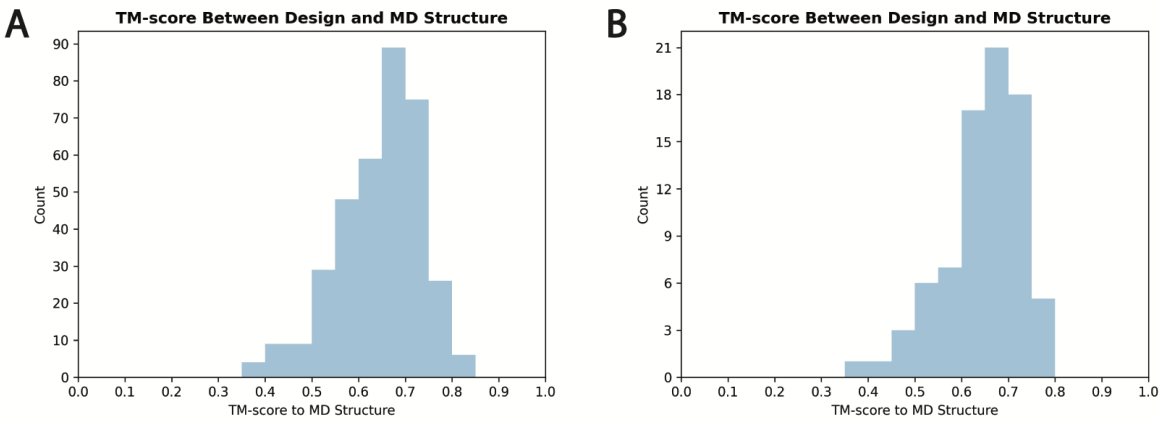

Figure S10. Assessment of the stability of the novel folds generated by FoldDesign. A) TM-score distribution between the FoldDesign scaffolds and their final MD structures on the 354 test topologies. B) TM-score distribution between the 79 novel FoldDesign structures and their final MD structures.

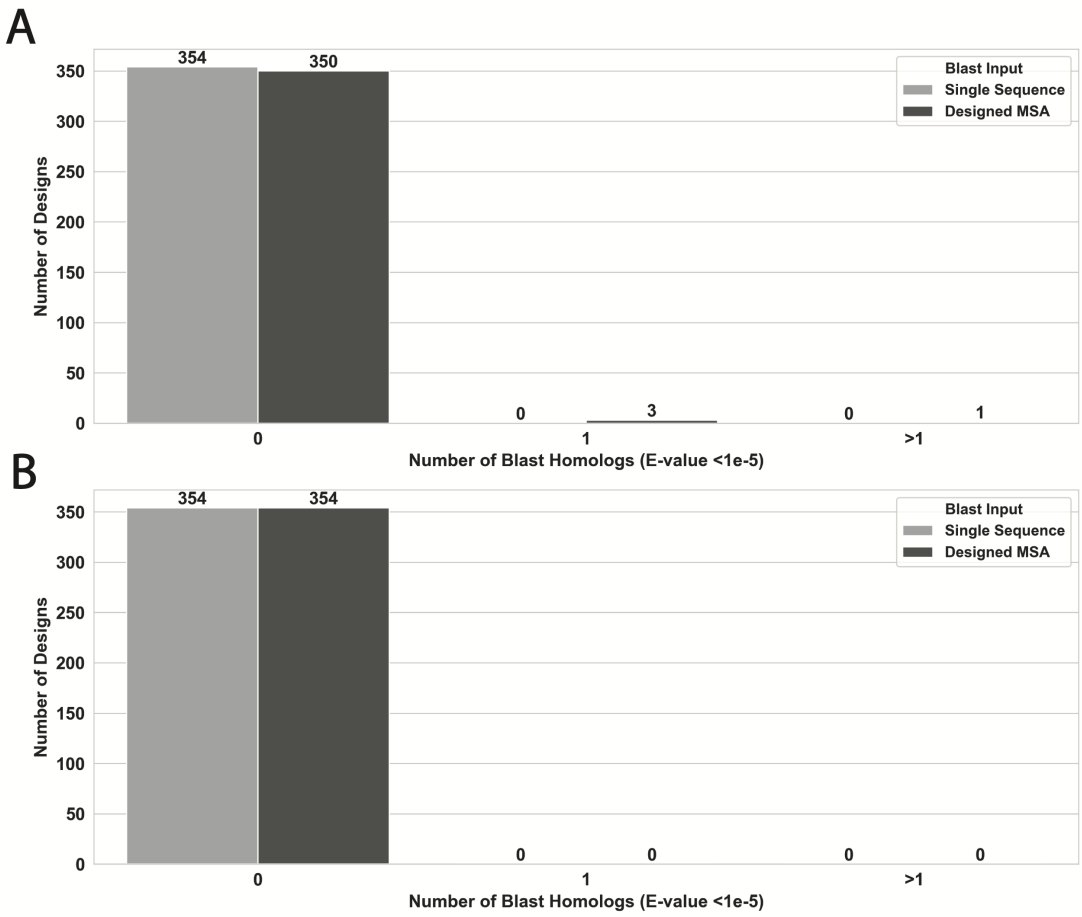

Figure S11. Sequence homologs detected by searching the FoldDesign designs through the nr database using Blast, where the sequences were designed by EvoEF2 (A) or RosettaFixBB (B). Two search strategies were used, either searching the single lowest energy sequence produced by EvoEF2/RosettaFixBB (Single Sequence) or jumpstarting the Blast search from the alignment of all 100 designed sequences (Designed MSA). The x-axis shows the number of Blast hits detected below an E-value threshold of 1e-5, while the y-axis shows the number of FoldDesign designs with the corresponding number of Blast hits.

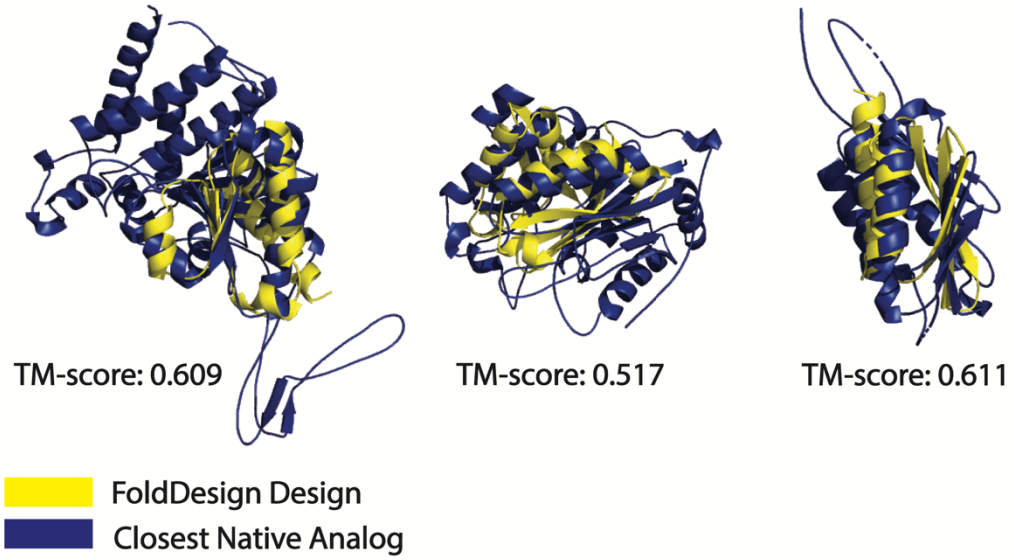

Figure S12. Structural alignment between the designed proteins shown in Fig. 5B and their closest native analogs in the PDB. The FoldDesign structures are shown in yellow, while the closest native analogs are shown in blue.

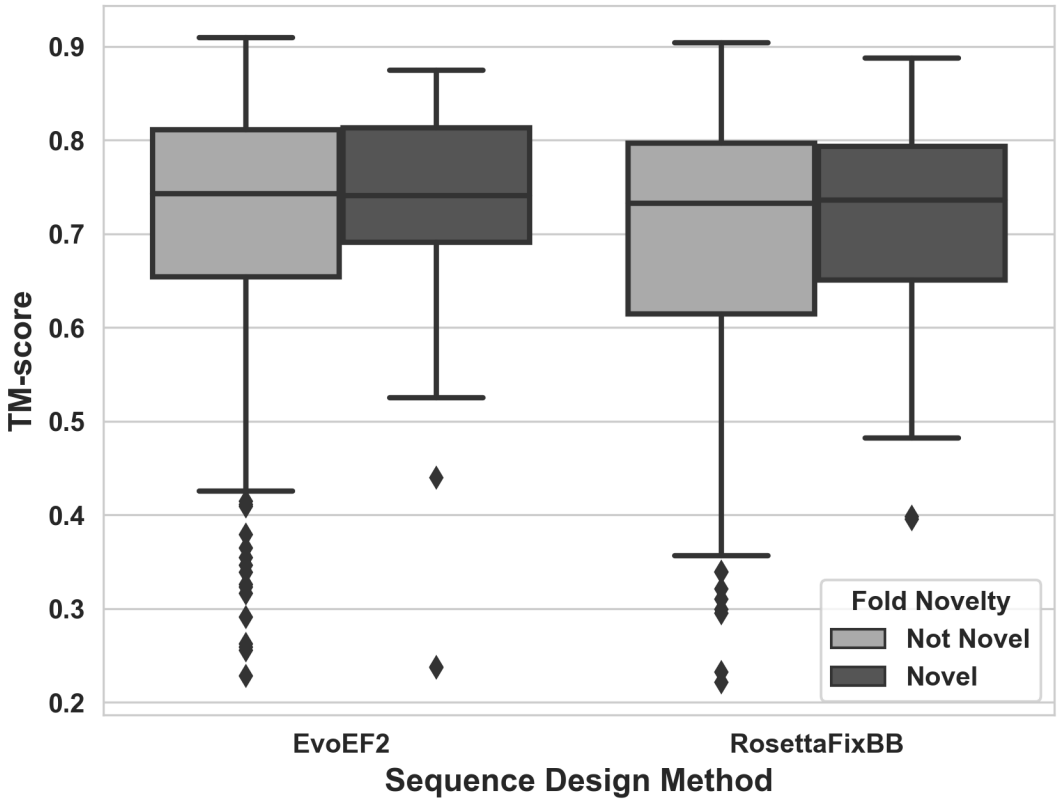

Figure S13. AlphaFold2 structure prediction results for the 79 FoldDesign scaffolds with novel folds (Novel) and the 275 scaffolds with natural analogs (Not Novel). The y-axis depicts the TM-scores between the AlphaFold2 models and the designed scaffolds, while the x-axis separates the EvoEF2 and RosettaFixBB sequence designs.

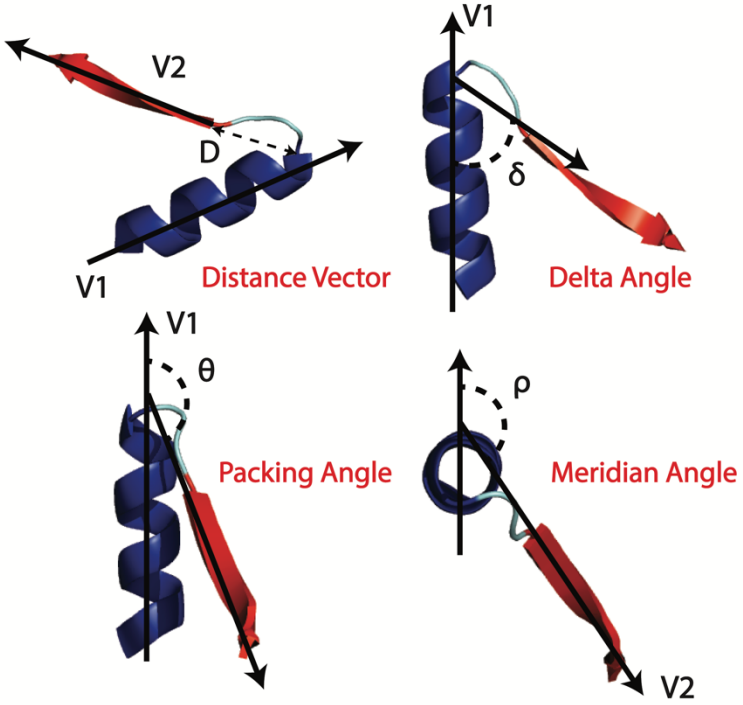

Figure S14. Smotif geometry definition. The axis for an α or β secondary structure is defined as the shortest of the principal moments of inertia of that structure, where V1 and V2 are the axis vectors of the secondary structure. The geometry of each motif is defined by four geometric features: (1) D, the distance between the ending points of the two secondary structure elements, (2) Hoist angle, δ, the angle between axis V1 and vector D; (3) Packing angle, θ, the angle between V1 and V2; and (4) Meridian angle, ρ, the angle between V2 and the plane that contains the vector V1.

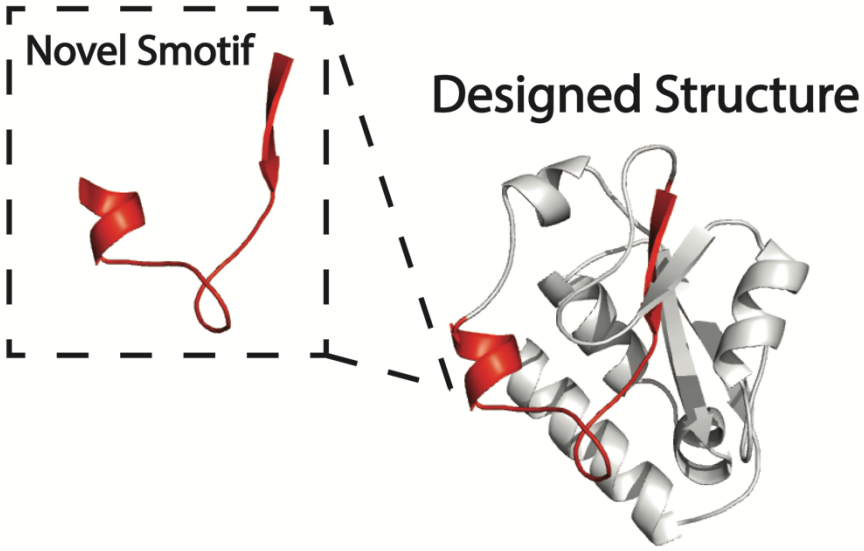

Figure S15. Novel Smotif geometry. The novel Smotif produced by FoldDesign is shown in the inset and highlighted in red, while the remainder of the structure is shown in gray.

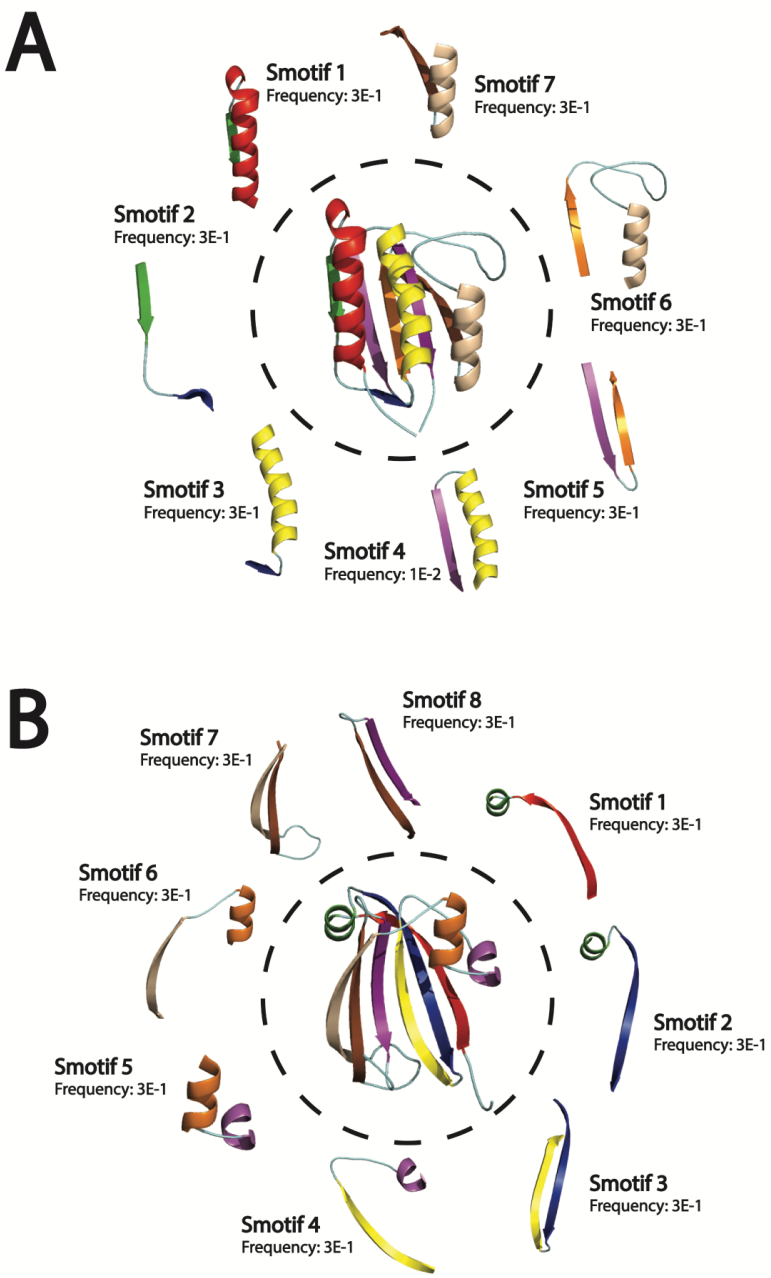

Figure S16. Smotif geometries found in the native folds. A) Native fold for 1id0A as well as each Smotif in the structure. B) Native fold for 2p19A as well as each Smotif in the structure. The frequencies for each Smotif are the background frequencies calculated from the PDB.

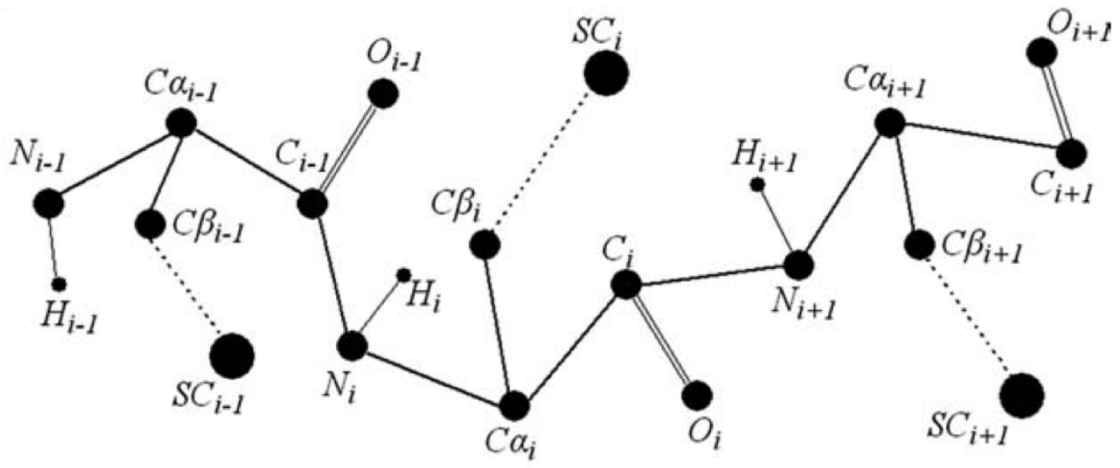

Figure S17. Depiction of the reduced model used to represent protein conformations during the FoldDesign simulations, including the backbone atoms (N, H, Cα, C, and O) as well as the Cβ atoms and side-chain centers of mass (SC). The center of mass for Valine is used in this study to evaluate steric clashes.

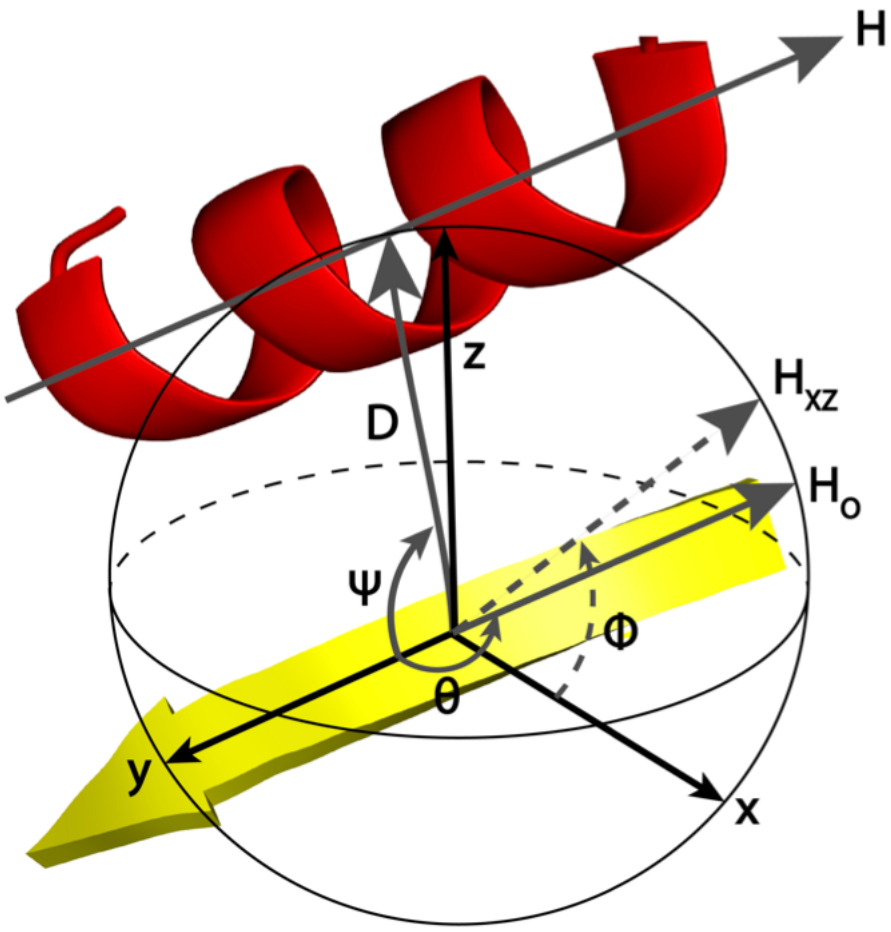

Figure S18. Illustration of the features used to calculate the energy for packing two secondary structure elements. Note, here a helix and strand are used, but the parameters are the same for two helices or two strands. The y-axis is defined along the direction of the strand, where the origin is set at the center. D is a vector that represents the distance between the center of the strand and the center of the helix, and the x-axis is defined as the cross product between the y-axis vector and the D vector. The z-axis is defined as the cross product of the y-axis and the x-axis. H is the helical axis and H_0 is the helical axis translated to the origin. H_{xz} is the projection of H_0 onto the xz-plane. Lastly, ψ, ϕ, and θ are the angles between the y-axis and the D vector, the x-axis and H_{xz}, and the y-axis and H_0, respectively.

Supplementary Texts

Text S1: Replica-exchange Monte Carlo simulation parameters and movements.

The conformational landscape is explored in FoldDesign using replica-exchange Monte Carlo (REMC) simulations. Within REMC, four parameters need to be carefully considered. First, the highest temperature (T_{max}) should be high enough to enable the simulation to overcome energy barriers, while the lowest temperature (T_{min}) should be low enough to ensure the simulation sufficiently scans the low-energy states. Second, the number of replicas (N_{rep}) should be large enough to ensure sufficient chances for the adjacent replicas to communicate with each other. Third, the number of local movements (N_{sweep}) before the global swaps should be selected to make the local Metropolis search achieve satisfactory equilibrium. After successive rounds of optimization, the final parameters were selected as: $T_{max} = min(20 * (1 + (L - 100) * 0.004), 20 * 2.5)$, $T_{min} = max(1 * (1 + (L - 100) * 0.001), 1 * 0.5)$, $N_{rep} = 40$, and $N_{sweep} = 30*\sqrt{L}$, where L is the sequence length and a total of 500 REMC simulation cycles are carried out for each design.

Given the maximum and minimum temperature settings, the temperature at each replica *i* is determined using an inverse linear temperature scheme (1, 2). Briefly, the temperature for replica 1 is set to T_{max}, i.e., $T_1 = T_{max}$, and the temperature for the i^{th} replica ($i > 1$) is determined by the following equation:

$$T_i = \frac{1}{3 * \Delta\beta_{min\_max} + \beta_{i-1} + 12 * \Delta\beta_{\min\_max} * \frac{i}{N_{rep} - 2}} \quad (S1)$$

Here, β refers to an inverse temperature, where $\Delta\beta_{min\_max} = \frac{(\beta_{min} - \beta_{max})}{N_{rep}-1}$, $\beta_{min} = \frac{1}{T_{min}}$, $\beta_{max} = \frac{1}{T_{max}}$, and $\beta_{i-1} = \frac{1}{T_i - 1}$. To illustrate the communication between replicas, Table S5 presents the empirically observed acceptance probabilities for swaps between adjacent replicas during the design simulations for the 354 FoldDesign scaffolds. As can be seen from the table, the fraction of accepted swaps was similar across each of the adjacent replicas, where the average acceptance probability was 0.738, demonstrating a high degree of communication between the replicas.

During the REMC simulations, 11 different conformational movements are used by FoldDesign, as show in Fig. S2, to sample the structural space. Movements are accepted or rejected using the Metropolis Criterion (3) based on the associated changes in energy calculated by the energy function described in Text S2. The major conformational movement is fragment substitution, where the decoy conformation in a selected region of the protein is replaced with the conformation from one of the highest scoring fragments. In order to perform this movement, it is first necessary to identify local fragments from a fragment library that match the input secondary structure topology. The fragment library is composed of 1-20 residue fragments from 29,156 high-resolution PDB structures used by QUARK (4, 5). The fragments were collected from structures deposited on or before 4/3/2014 and shared <30% sequence identity to each other (4, 5). Notably, this library has been extensively validated in the related field of protein structure prediction during even the most recent CASP experiments (6, 7). The information present for each fragment includes the position-wise backbone torsion angles (ϕ, ψ, ω), secondary structure, bond lengths, bond angles, solvent accessibility and Cα coordinates. During the movement, the backbone torsion angles (ϕ, ψ, ω) and backbone bond lengths and angles in the decoy structure are swapped with

those present in the selected fragment. Next, cyclical coordinate descent loop closure (8) is used to connect the anchor points and prevent large downstream perturbations. Larger insertions are attempted at the beginning of the simulation, when the protein is largely unfolded, and smaller insertions are attempted as the protein become more compact.

In addition to fragment assembly, FoldDesign uses 10 auxiliary movements. The first of the auxiliary movements involves changing the length of one of the backbone bonds, including the N-Cα, Cα-C, or C-N bonds, by a random value in the range [-0.24 Å, 0.24Å], which is sampled from using a uniform probability distribution. The second movement involves randomly changing one of the backbone angles by a uniform random value in the range [-10°, 10°], including the N_i-$C\alpha_i$-C_i, $C\alpha_i$-C_i-N_{i+1}, and C_i-N_{i+1}-$C\alpha_{i+1}$ angles, where i corresponds to the residue position. The third auxiliary movement changes one or more of the backbone torsion angles (ϕ, ψ, ω). The ϕ and ψ angles are updated by sampling from the allowed regions in the Ramachandran plots based on the input secondary structure at a given position. The ω angle is changed by a uniform random value selected from the range [-8°, 8°], where the movement is automatically rejected if it would result in the ω angle falling outside of the range of (170°, 190°). The fourth movement is LMProt perturbation (9), which randomly changes the positions of the backbone atoms in a selected region and then attempts to restrict all bond lengths and bond angles to physically allowable values. The fifth movement is segment rotation, which rotates the backbone atoms by a uniform random value in the range of (-90°, 90°) for a 2-12 residue segment along the axis defined by the Cα atoms of the first and last residues of the selected region. The sixth movement is similar to the fragment substitution movement but is based on fragment consensus from the 10 residue long fragments. To perform this movement, the 10 residue long identified fragments are clustered based on the distance matrix defined by their ϕ/ψ angle pairs. Then during the simulations, the ϕ/ψ angle pairs for a 10 residue segment in the decoy structure are swapped for the corresponding angles from the consensus fragments. The seventh movement is a segment shift. It involves shifting the residue numbers in a segment forward or backwards by one residue, which means that the coordinates of each residue are copied from their preceding or subsequent residues in the segment. We then delete the unused coordinates of one residue at the selected terminal region and insert new coordinates for another residue at the other terminal based on physically allowable bond lengths and angles. This movement can easily adjust the β-pairing in two well-aligned β-strands. The eighth auxiliary movement is β-turn formation, which attempts to form a β-turn in regions of the protein whose input secondary structure is defined as coiled. The final two movements are β-strand and α-helix formation. For these two movements, two regions that are defined as β-strands or α-helices are moved closer together based on distance and torsion angle distributions collected from the PDB.

Text S2: FoldDesign energy function.

The energy function used to guide the FoldDesign simulations is a combination of 10 energy terms:

$$E_{DeepFold} = E_{HB} + E_{ss\_satisfaction} + E_{rama} + E_{hhpack} + E_{sspack} + E_{\mathrm{hspack}} + E_{ev} + E_{generic\_dist} + E_{frag\_dist\_profile} + E_{frag\_solv} + E_{rg} + E_{contact\_num} \quad (S2)$$

where E_{HB}, $E_{ss\_satisfaction}$, E_{rama}, E_{hhpack}, E_{sspack}, E_{hspack}, E_{ev}, $E_{generic\_dist}$, $E_{frag\_dist\_profile}$, E_{frag_solv}, E_{rg}, and $E_{contact\_num}$ are terms for backbone hydrogen bonding, secondary structure satisfaction, Ramachandran torsion angles, helix-helix packing, strand-strand

packing, helix-strand packing, excluded volume, generic backbone atom distances, fragment-derived distance restraints, fragment-derived solvent accessibility, radius of gyration, and expected contact number, respectively. The equations for each energy term are detailed below.

E_{HB} is calculated as follows:

$$E_{HB} = \sum_{i,j,T_k} E_{hb\_feat}(i,j,T_k) \quad (S3)$$

where i and j are the residue indices and T_k is the kth type of hydrogen bonding restraint. In FoldDesign, there are 4 types of hydrogen bonding restraints: hydrogen bonds between residues i and $i+4$ in regions defined as helical by the input secondary structure (T_1), virtual hydrogen bonds between residues i and $i+3$ in regions defined as helical by the input secondary structure (T_2), and hydrogen bonds between residues i and j in parallel β-strands (T_3) or antiparallel β-strands (T_4) for regions defined as strands by the input secondary structure. The energy for each type of hydrogen bonding restraint is calculated using the following equation:

$$E_{hb_{feat}}(i,j,T_k) = \sum_{l=1}^{n_k} \frac{(f_l(i,j) - \mu_{kl})^2}{2\delta_{kl}^2}, \quad n_k = \begin{cases} 4 & k = 1,2 \\ 3 & k = 3,4 \end{cases} \quad (S4)$$

where $f_l(i,j)$ is the value of the l^{th} feature from the decoy structure, n_k is the number of features considered for the k^{th} type of hydrogen bond restraint, μ_{kl} is the average value of the l^{th} feature for the k^{th} type of hydrogen bond restraint calculated from the PDB library, and δ_{kl} is the standard deviation of the l^{th} feature for the k^{th} type of hydrogen bond restraint. For hydrogen bonding, we consider four features: the distance, $D(O_i,H_j)$, between backbone atom O_i from residue i and the backbone hydrogen, H_j, from residue j, the angle, $A(C_i,O_i,H_j)$, between backbone atoms C_i and O_i from residue i and the backbone hydrogen, H_j, from residue j, the angle, $A(C_i,O_i,H_j)$, between backbone atom O_i from residue i and the backbone hydrogen, H_j, and nitrogen, N_j, from residue j, and the torsion angle, $T(C_i,O_i,Hj,N_j)$, between atoms C_i and O_i from residue i and the backbone hydrogen, H_j, and nitrogen, N_j, from residue j. Note for hydrogen bonding in strand regions, T_3 and T_4 restraints, $T(C_i,O_i,Hj,N_j)$ is not considered as there is a large standard deviation for this feature in strand regions. The values of μ_{kl} and δ_{kl} are shown in Table S6.

$E_{ss\_satisfaction}$ is calculated as follows:

$$E_{ss\_satisfaction} = -\sum_{i=1}^{i=L} \begin{cases} -2 & if\ ss_i = helix\ and\ input_{ss_i} = strand\ or\ ss_i = strand\ and\ input_{ss_i} = helix \\ 1 & if\ ss_i = coil\ and\ input_{ss_i} = coil \\ 2 & if\ ss_i = helix\ and\ input_{ss_i} = helix\ or\ ss_i = strand\ and\ input_{ss_i} = strand \\ -1 & else \end{cases} \quad (S5)$$

where ss_i is the secondary structure of the decoy at position i and $input\_ss_i$ is the input secondary structure at the corresponding position. If the input secondary structure is defined as helical and the secondary structure of the decoy structure is a strand or if the input secondary structure is defined as a strand and the secondary structure of the decoy structure is helical, then a penalty of -2 is assigned to penalize opposite secondary structure assignments more heavily. Similarly, if the helical or strand regions are correct in the decoy structure, then a stronger bonus is assigned. Mismatches in coiled regions are penalized less heavily, and correctly generated coiled regions are

also rewarded to a lesser degree as they are more flexible and lack regular hydrogen bonding patterns.

E_{rama} is calculated as follows:

$$E_{rama} = -\sum_{i=2}^{i=L-1} \log\left(P(\phi_i, \psi_i) \mid input_{ss_i}\right) \quad (S7)$$

where ϕ_i and ψ_i are the backbone torsion angles at position *i* and $input\_ss_i$ is the input secondary structure at position *i.* The probabilities for each backbone torsion angle pair were determined from the I-TASSER (10) PDB library based on the secondary structure at a given position.

E_{hhpack}, E_{sspack}, and E_{hspack} are calculated as follows:

$$\begin{cases} E_{hhpack} = -\sum_{i,j} \log\left(P_{hh}\left(\psi_{ij}, \theta_{ij}, \Phi_{ij}\right) \mid seq\_sep\right) - \sum_{i,j} \log\left(P_{hh}\left(D_{ij}, \theta_{ij}\right) \mid seq\_sep\right) \\ E_{sspack} = -\sum_{i,j} \log\left(P_{ss}\left(\psi_{ij}, \theta_{ij}, \Phi_{ij}\right) \mid seq\_sep\right) - \sum_{i,j} \log\left(P_{ss}\left(D_{ij}, \theta_{ij}\right) \mid seq\_sep\right) \\ E_{hspack} = -\sum_{i,j} \log\left(P_{hs}\left(\psi_{ij}, \theta_{ij}, \Phi_{ij}\right) \mid seq\_sep\right) - \sum_{i,j} \log\left(P_{hs}\left(D_{ij}, \theta_{ij}\right) \mid seq\_sep\right) \end{cases} \quad (S8)$$

where ψ_{ij}, θ_{ij}, Φ_{ij} are the angles between two secondary structure elements (either two helices, E_{hhpack}, two strands E_{sspack}, or a helix and a strand, E_{hspack}) defined in Fig. S18, D_{ij} is the distance between the centers of the two secondary structure elements, and seq_sep is the number of residues between two secondary structure elements along the sequence. The potential is split into three different groups depending on the sequence separation, including short, medium, and long-range interactions. Here, short, medium, and long-range refers to residue pairs (*i,j*) that fall in the following ranges, respectively: $6 \leq |i-j| < 12$, $12 \leq |i-j| < 24$, and $|i-j| \geq 24$. The secondary structure specific probabilities distributions for the features were derived from PDB structures in the I-TASSER library and were fit using kernel density estimation to smooth the potentials.

For the estimation of $P(\psi, \theta, \Phi)$, the periodic von Mises probability distribution was used as the kernel function (k_{angle}); specifically $k_{angle}(x, \kappa) = \frac{1}{2\pi I_0(k)} \exp\left(\kappa * cos(x)\right)$, where x is an angle value, κ is a tunable concentration parameter, and I_0 is the modified Bessel function of the first kind of order zero. Thus, the probability distribution, $P(\psi, \theta, \Phi)$, for each of the three interaction types and sequence separation categories was estimated by $P(\psi, \theta, \Phi|\kappa) = \frac{1}{N}\sum_{i=1}^{N} k_{angle}(\psi - \psi_i, \kappa) k_{angle}(\theta - \theta_i, \kappa) k_{angle}(\Phi - \Phi_i, \kappa)$. Here, Φ was computed over the range [0°, 360°), while θ and Φ were computed over the range [0°, 180°], where a bin size of 1° was used for each angle. Additionally, i denotes the index of the datapoint derived from the PDB dataset for observed values of ψ, θ, and Φ, where the summation was carried out over the N datapoints in the dataset for each interaction type and sequence separation category. Lastly, the concentration parameter, κ, may be tuned, where the larger the value of κ is, the narrower the kernels will be. To optimize this parameter, the dataset was randomly divided into 10 equal subsets

and the value of κ was varied from 0° to 180° by an increment of 1°, where the value that resulted in the maximum mean log-likelihood for the observed angles across the 10 subsets was used for each interaction type and sequence separation.

For the estimation of $P(D,\theta)$, the same periodic von Mises function was used as the kernel for θ. However, for the distance, D, a non-periodic gaussian distribution was used as the kernel function (k_{dist}), specifically $k_{dist}(D,h) = \frac{1}{\sqrt{2\pi}h}\exp\left(\frac{-D^2}{2h^2}\right)$, where D is a distance and h is the bandwidth parameter. Thus, the probability distribution, $P(D,\theta)$, for each of the three interaction types and sequence separation categories was estimated by $P(\psi,\theta,\Phi|\kappa,h) = \frac{1}{N}\sum_{i=1}^{N} k_{dist}(D - D_i,h)k_{angle}(\theta-\theta_i,\kappa)$. Here, θ was computed over the range [0°, 180°] with a bin size of 1°, while D was computed over the range [0, 20Å] with a bin size of 0.1 Å. As before, i denotes the index of the datapoint derived from the PDB dataset for observed values of D and θ, where the summation was carried out over the N datapoints in the dataset for each interaction type and sequence separation category. Again, κ and h are tunable parameters, where κ was varied from 0° to 180° by an increment of 1°, while h was varied from 0.1 Å to 20 Å using an increment of 0.1 Å. As before, the optimal values of these parameters were determined by randomly splitting the dataset into 10 subsets and selecting the values that resulted in the highest mean log-likelihood across all 10 datasets for the observed values.

E_{ev} is calculated as follows:

$$E_{ev} = \sum_{i=1}^{i=L}\sum_{j=i+1}^{j=L}\sum_{ii}\sum_{jj}\begin{cases}(vdw(i,ii)+vdw(j,jj))^2 - r_{ii,jj}^2 & if\ r_{ii,jj} < vdw(i,ii)+vdw(j,jj)\\ 0 & else\end{cases} \quad (S9)$$

where clashes are calculated between each atom ii from residue i and atom jj from residue j and $r_{ii,jj}$ is the distance between the two atoms. For the side-chain center atoms, the center of mass of valine is used to assess steric clashes. All atoms presented in Fig. S17 are considered except for hydrogen.

$E_{generic\_dist}$ is calculated as follows:

$$E_{generic\_dist} = \sum_{i=1}^{i=L}\sum_{j=i+1}^{j=L}\sum_{ii}\sum_{jj} -RT * \log\left(\frac{N_{obs}(ii,jj,r_{ii,jj})}{r_{ii,jj}^{\alpha}N_{obs}(ii,jj,r_{cut})}\right) \quad (S10)$$

where L is the protein length, i and j are the two residue indices and *ii/jj* are the atoms N, Cα, C, O and Cβ. $N_{obs}(ii,jj,r_{ii,jj})$ is the observed number of pairs between atoms *ii* and *jj* with distance $r_{ii,jj}$ determined from the I-TASSER PDB library. A cutoff, r_{cut}, of 15Å is used and the distances for the observed atom pairs is divided into 0.5Å bins from 0Å to 15Å. The potential is similar to DFIRE, where $\alpha = 1.61$ and $N_{obs}(ii,jj,r_{cut})$ is used to calculate the background probability.

$E_{frag\_dist\_profile}$ is calculated as follows:

$$E_{frag\_dist\_profile} = -\sum_{(i,j)\subseteq S_{dp}} \log\left(N_{ij}(d_{ij})\right) \quad (S11)$$

where d_{ij} is the distance between the Cα atoms of residues *i* and *j* in the decoy structure and N_{ij} is the distance profile for residues *i* and *j* extracted from the 10 residue long fragments where *d* falls in the range [0Å, 9Å] with a bin width of 0.5 Å. S_{dp} is the set of residues that have fragment-derived distance profiles. To derive the distance profiles, we first analyze each of the 10 residue fragments that originate from the same PDB structure and are aligned to different residues, *i* and *j*. Then we calculate the distance between the Cα atoms for the two positions from the fragments based on their corresponding positions in their PDB structure. If the distance between the two residues in the PDB structure is <9Å, then these positions may be encouraged to from contacts in the designed structure. This procedure is repeated for each query residue pair (*i, j*) to construct a histogram of distances. If the histogram for a given pair of residues has a peak <9Å, then the histogram is saved to calculate the distance profile energy and the residue pair is added to the set S_{dp}.

E_{frag_solv} is calculated as follows:

$$E_{frag\_solv} = \sum_{i=1}^{i=L} \left| s_i - s_i^E \right| \qquad (S12)$$

where *L* is the protein length, s_i is the solvent accessibility of residue *i* in the decoy structure, and s_i^E is the expected solvent accessibility derived from the 20 residue fragments. The following formula is used to calculate s_i:

$$s_i = 1 - 0.007 \sum_{d(G_i,G_j)<9\text{Å}} \frac{A_{aa(j)}}{d^2(G_i, G_j)} \qquad (S13)$$

Here, $A_{aa(j)}$ is the maximum solvent accessible surface area for the given residue aa at position *j*. Since polyvaline sequences are used in FoldDesign, the maximum solvent accessible surface area for Valine is used. G_i and G_j are the geometric centers of residues *i* and *j*, $d(G_i, G_j)$ is the distance between the two geometric centers, and $d^2(G_i, G_j)$ is the squared distance. A cutoff of 9Å is used as residues that are further apart contribute little to the solvent accessibility. As mentioned above, s_i^E is the expected solvent accessibility calculated from the overlapping 20 residue fragments. For each fragment, the solvent accessibility of the residue in its native PDB structure is recorded, and the estimated solvent accessibility is calculated by averaging the solvent accessibility of each fragment residue aligned to position *i*.

E_{rg} is calculated as follows:

$$E_{rg} = \begin{cases} 0 & r_{min} \le r \le r_{max} \\ (r_{min} - r)^2 & r < r_{min} \\ (r - r_{max})^2 & r > r_{max} \end{cases} \qquad (S14)$$

where r is the radius of gyration for the decoy structure calculated from the Cα positions produced during the FoldDesign simulations and r_{min}/r_{max} are the estimated minimum and maximum radii of gyration calculated from the PDB based on the protein length and secondary structure

composition. More specifically, the minimum and maximum radii of gyration are estimated following previous work in protein structure prediction by QUARK (4), where $r_{min} = 2.316L^{0.358} - 0.5$ and $r_{max} = max\{r_{min} + 8.0, 0.5\sqrt{3/5}N_{maxh}\}$. Here, N_{maxh} is the length of the longest helix in the structure and L is the protein length. Using these values, 95% of the experimental structures in the PDB have a radius of gyration within $[r_{min}, r_{max}]$ (4).

$E_{contact\_num}$ is calculated as follows:

$$E_{contact\_num} = \left|num_{short\_cont} - expected\_num_{short\_cont}\right| + \left|num_{med\_cont} - expected\_num_{med\_cont}\right| + \left|num_{long\_cont} - expected\_num_{long\_cont}\right| \quad (S15)$$

where num_{short_cont}, num_{med_cont}, and num_{long_cont} are the number of short, medium, and long-range contacts in the decoy structure. Here, short, medium, and long-range contacts refer to residue pairs (*i,j*) that fall in the following ranges, respectively: $6 \le |i-j| < 12$, $12 \le |i-j| < 24$, and $|i-j| \ge 24$. $expected\_num_{short\_cont}$, $expected\_num_{med\_cont}$, and $expected\_num_{long\_cont}$ are the expected short, medium, and long-range contacts calculated from PDB structures in the I-TASSER library based on protein length.

Text S3: Rosetta protocol used to generate designed folds.

The following command was used to generate backbones by Rosetta:

<rosetta_bin>/main/source/bin/rosetta_scripts.static.linuxgccrelease -database <rosetta_bin>/main/database/ -s ./input.pdb -parser:protocol ./backbone_generation.xml -nstruct 250

The contents of the backbone_generation.xml files are detailed below, which were adapted from a representative recent publication (11).

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
      <ScoreFunction name="SFXN1" weights="fldsgn_cen_omega02.wts" />
    </SCOREFXNS>
    <FILTERS>
        <ScoreType name="cen_total" scorefxn="SFXN1" score_type="total_score" threshold="1000000" />
        <ScoreType name="vdw" scorefxn="SFXN1" score_type="vdw" threshold="1000000" />
        <ScoreType name="rg" scorefxn="SFXN1" score_type="rg" threshold="1000000" />
        <ScoreType name="cen_rama" scorefxn="SFXN1" score_type="rama" threshold="1000000" />
        <ScoreType name="sspair" scorefxn="SFXN1" score_type="ss_pair" threshold="1000000" />
        <ScoreType name="rsigma" scorefxn="SFXN1" score_type="rsigma" threshold="1000000" />
    </FILTERS>
    <TASKOPERATIONS>
    </TASKOPERATIONS>
    <MOVERS>
        <Dssp name="dssp"/>
        <SwitchResidueTypeSetMover name="fullatom" set="fa_standard"/>
        <SwitchResidueTypeSetMover name="cent" set="centroid"/>
        <MakePolyX name="polyval" aa="VAl" keep_pro="1" />
        <BluePrintBDR name="bdr1" scorefxn="SFXN1" use_abego_bias="1" blueprint="blueprint.xml"/>
        <MinMover name="min1" scorefxn="SFXN1" chi="1" bb="1" type="dfpmin_armijo_nonmonotone_atol"
tolerance="0.0001"/>
        <ParsedProtocol name="cenmin1" >
```

```
            <Add mover_name="cent" />
            <Add mover_name="min1" />
            <Add mover_name="fullatom" />
          </ParsedProtocol>
          <ParsedProtocol name="bdr1ss" >
            <Add mover_name="bdr1" />
            <Add mover_name="cenmin1" />
            <Add mover_name="dssp" />
          </ParsedProtocol>
    </MOVERS>
    <PROTOCOLS>
          <Add mover_name="bdr1ss" />
          <Add mover_name="fullatom" />
          <Add filter_name="cen_total" />
          <Add filter_name="vdw" />
          <Add filter_name="rg" />
          <Add filter_name="cen_rama" />
          <Add filter_name="sspair" />
          <Add filter_name="rsigma" />
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

The contents of the weights file (fldsgn_cen_omega02.wts) were as follows, which were also adapted from the previous study (11):

```
vdw 1.0
rg 1.0
rama 0.1
hs_pair 1.0
ss_pair 1.0
rsigma 1.0
omega 0.5
hbond_lr_bb 1.0
hbond_sr_bb 1.0

STRAND_STRAND_WEIGHTS 1 11
```

Here, for each input topology, 250 designs were generated using Rosetta, where the final designs were selected from the lowest energy structures as assessed by the Rosetta centroid energy function. In terms of the total number of conformational movements, the average number of movements attempted by Rosetta per design was 8,291,689.9, not including the L-BFGS-based minimization, which was slightly higher than the 6,000,000 movements attempted by FoldDesign for each design. This protocol follows the standard, widely used fragment assembly-based design procedure by Rosetta, where topologies are defined by the BluePrintBDR mover and built using stepwise Monte Carlo fragment assembly simulations guided by the Rosetta centroid energy function (12). Following this, the designs were minimized using L-BFGS optimization of the internal coordinates and filtered using a combination of score thresholds. Since the purpose of the benchmark tests was to perform fully automated *de novo* protein design, no user-provided restraints were utilized other than the 3-state secondary structure sequences. An example of the Rosetta blueprint files without and with ABEGO bias are provided in Texts S8 and S9 (see below), respectively.

Text S4: Analysis of the results with ABEGO bias and sub-rotamer sampling

In the manuscript, Rosetta was run without ABEGO bias, which divides the Ramachandran plot into 4 regions (A,B,E,G) and restricts the fragment selection to the region defined by the specified bias for each residue (13). This bias allows for more control over the fragment selection process and fold definition; however, given that the benchmark dataset was composed of just the 3-state SS sequences from the native proteins, the proper ABEGO definition for each position is ambiguous as the same SS type can be sampled from multiple regions of the Ramachandran plot, e.g., right-handed (ABEGO region A) vs. left-handed alpha helices (ABEGO region G). Nevertheless, given that this bias is often used, we reran Rosetta and restricted helical regions to the A region of the Ramachandran plot and strands to the B region of the Ramachandran plot (13). We then calculated the percent of buried residues/SASA and the GOAP/ROTAS energies for the Rosetta designs that utilized ABEGO bias, where the results are summarized in Fig. S4. This analysis showed that there was not a significant difference in the percent of buried residues/SASA or the GOAP/ROTAS energies between the designs that utilized ABEGO bias and those that did not (with p-values >0.05).

Additionally, similar to EvoEF2, RosettaFixBB was run without sub-rotamer sampling (see Text S6 for the RosettaFixBB protocol). To examine if enabling additional rotameric sampling during the sequence design impacted the results, we reran RosettaFixBB with χ_1 and χ_2 sub-rotamer sampling enabled for the FoldDesign and Rosetta scaffolds (see Text S7 for the RosettaFixBB protocol with sub-rotamers enabled), where the results are depicted in Fig. S5. Overall, only the ROTAS energy improved significantly (p-values <0.05) with the addition of sub-rotamer sampling, which may be expected as ROTAS places special emphasis on the rotameric conformations adopted by the side-chains (14). Nevertheless, the FoldDesign scaffolds still had significantly lower ROTAS energies (-10684.5) than the Rosetta scaffolds (-9446.4) with a p-value of 7.7E-08. Thus, enabling sub-rotamer sampling and including ABEGO bias did not alter the conclusions drawn in the text, where it would be expected that any improvements in the sequence design protocol would benefit both FoldDesign and Rosetta.

Text S5: Analysis of the amino acid compositions of the designed scaffolds

Given that Valine is used as the generic center of mass in FoldDesign and Rosetta (see Methods), one important issue is to examine whether the designed scaffolds exhibited any systematic bias against particular amino acids, such as smaller non-polar residues like Glycine and Alanine as well as bulkier aromatic amino acids or Proline. In Fig. S6, we plot the frequency of each of the 20 amino acids in the EvoEF2/RosettaFixBB designed sequences for the FoldDesign and Rosetta scaffolds compared to the frequency from the corresponding native protein sequences. As expected, the specific amino acid preferences varied depending on the sequence design method that was used; however, it can be observed that there was no bias towards Valine for FoldDesign or Rosetta, and smaller non-polar amino acids such as Glycine and Alanine were well represented in the designed sequences, as well as bulkier amino acids like Tryptophan, Tyrosine, and Proline, with some variation for Proline and Alanine depending on the sequence design method. Quantitatively, the Kullback-Leibler (KL) divergence between the native amino acid distribution and the distributions for the EvoEF2/RosettaFixBB sequence designs for the FoldDesign scaffolds was 0.236/0.122, which was slightly lower than the KL divergence for the Rosetta scaffolds (0.352/0.123). In addition, since FoldDesign does not include any chirality restraints on the backbone torsion angles during the folding simulations, the designed folds contained structures with both right- and left-handed helices and covered the full diversity of the torsion angle space adopted by natural proteins as highlighted in the Ramachandran plot (Fig. S7).

Text S6: RosettaFixBB protocol used to generate designed folds.

The following command was used to generate sequence designs by RosettaFixBB without sub-rotamer sampling:

<rosetta_bin>/main/source/bin/fixbb.static.linuxgccrelease -database <rosetta_bin>/main/database/ -s ./design.pdb -nstruct 100

Text S7: RosettaFixBB protocol with sub-rotamer sampling.

The following command was used to generate sequence designs by RosettaFixBB with χ_1 and χ_2 sub-rotamer sampling:

<rosetta_bin>/main/source/bin/fixbb.static.linuxgccrelease -database <rosetta_bin>/main/database/ -s ./design.pdb -nstruct 100 -ex1 -ex2

Text S8: Example Rosetta blueprint file without ABEGO bias.

The following illustrates the contents of the Rosetta blueprint file without ABEGO bias for the secondary structure topology derived from 2jx8A.

```
1 V L R
2 V L R
0 V H R
0 V H R
0 V H R
0 V H R
0 V H R
0 V H R
0 V L R
0 V L R
0 V L R
0 V E R
0 V E R
0 V E R
0 V E R
0 V L R
0 V L R
0 V L R
0 V L R
0 V E R
0 V E R
0 V E R
0 V E R
0 V E R
0 V E R
0 V L R
0 V L R
0 V L R
0 V L R
0 V E R
0 V E R
0 V E R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
```

```
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
0 V L R
```

Text S9: Example Rosetta blueprint file with ABEGO bias.

The following illustrates the contents of the Rosetta blueprint file with ABEGO bias for the secondary structure topology derived from 2jx8A.

```
1 V L  R
2 V L  R
0 V HA R
0 V HA R
0 V HA R
0 V HA R
0 V HA R
0 V HA R
0 V L  R
0 V L  R
0 V L  R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V EB R
0 V EB R
0 V EB R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
0 V L  R
```

0 V L R

Text S10: Relative frequency of Smotifs for the test protein structures

In Fig. 6, we first split the Smotifs into 4 bins based on the normalized background frequency of the Smotifs that appear in the PDB structures, i.e., [0, 1E-3], (1E-3, 1E-2], (1E-2, 1E-1], and (1E-1, 1], where the normalized background frequency of a Smotif is equal to the number of times that the Smotif appeared in the 51,094 non-redundant full-chain structures in the I-TASSER template library divided by the total number of Smotifs in the structural library.

For a given protein i in the test set of the 79 novel folds or the 354 native structures, the relative frequency of Smotifs for one of the 4 bins, j, is calculated by

$$Relative\ Frequency\ (i,j) = \frac{Num\_Smotif_{i,j}}{\sum_{j=1}^{j=4} Num\_Smotif_{i,j}} \qquad (S16)$$

where $Num\_Smotif_{i,j}$ is the number of Smotifs from the *i*-th protein that fall into the *j*-th bin.

References

1. M. Ulmke, H. Müller-Krumbhaar, Linear scaling of computer time with the inverse temperature for the grand canonical quantum Monte Carlo method. *Zeitschrift für Physik B Condensed Matter* **89**, 239-241 (1992).
2. I. Rozada, M. Aramon, J. Machta, H. G. Katzgraber, Effects of setting temperatures in the parallel tempering Monte Carlo algorithm. *Phys Rev E* **100**, 043311 (2019).
3. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087-1092 (1953).
4. D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-1735 (2012).
5. D. Xu, Y. Zhang, Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* **81**, 229-239 (2013).
6. W. Zheng *et al.*, Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149-1164 (2019).
7. W. Zheng *et al.*, Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins: Structure, Function, and Bioinformatics* **89**, 1734-1751 (2021).
8. A. A. Canutescu, R. L. Dunbrack, Jr., Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein science : a publication of the Protein Society* **12**, 963-972 (2003).
9. R. A. da Silva, L. Degrève, A. Caliri, LMProt: an efficient algorithm for Monte Carlo sampling of protein conformational space. *Biophys J* **87**, 1567-1577 (2004).
10. J. Yang *et al.*, The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7-8 (2015).
11. A. A. Vorobieva *et al.*, De novo design of transmembrane beta barrels. *Science* **371** (2021).
12. L. An, G. R. Lee, De Novo Protein Design Using the Blueprint Builder in Rosetta. *Current Protocols in Protein Science* **102**, e116 (2020).
13. Y. R. Lin *et al.*, Control over overall shape and size in de novo designed proteins. *P Natl Acad Sci USA* **112**, E5478-E5485 (2015).
14. J. Park, K. Saitou, ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC bioinformatics* **15**, 307-307 (2014).