PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

De novo protein fold design through sequence-independent fragment assembly simulations

Robin Pearce[a], Xiaoqiang Huang[a], Gilbert S. Omenn[a,b,c,d], and Yang Zhang[a,e,f,g,1]

Edited by William DeGrado, University of California San Francisco, San Francisco, CA; received May 17, 2022; accepted December 22, 2022

De novo protein design generally consists of two steps, including structure and sequence design. Many protein design studies have focused on sequence design with scaffolds adapted from native structures in the PDB, which renders novel areas of protein structure and function space unexplored. We developed FoldDesign to create novel protein folds from specific secondary structure (SS) assignments through sequence-independent replica-exchange Monte Carlo (REMC) simulations. The method was tested on 354 non-redundant topologies, where FoldDesign consistently created stable structural folds, while recapitulating on average 87.7% of the SS elements. Meanwhile, the FoldDesign scaffolds had well-formed structures with buried residues and solvent-exposed areas closely matching their native counterparts. Despite the high fidelity to the input SS restraints and local structural characteristics of native proteins, a large portion of the designed scaffolds possessed global folds completely different from natural proteins in the PDB, highlighting the ability of FoldDesign to explore novel areas of protein fold space. Detailed data analyses revealed that the major contributions to the successful structure design lay in the optimal energy force field, which contains a balanced set of SS packing terms, and REMC simulations, which were coupled with multiple auxiliary movements to efficiently search the conformational space. Additionally, the ability to recognize and assemble uncommon super-SS geometries, rather than the unique arrangement of common SS motifs, was the key to generating novel folds. These results demonstrate a strong potential to explore both structural and functional spaces through computational design simulations that natural proteins have not reached through evolution.

de novo protein design | structural design | novel fold of protein | structural motif | replica-exchange Monte Carlo simulation

Significance

Despite the vast sequence space, only a tiny fraction of possible folds and functions achievable by proteins have been realized in nature, probably due to the selective pressures by environmental constraints during evolution. There is considerable interest in de novo protein design to engineer artificial proteins with novel structures and functions beyond those created by nature. However, the success rate of computational de novo design remains low and frequently requires extensive user intervention and large-scale experimental optimization. To address this issue, we developed an automated open-source program, FoldDesign, which shows improved performance in creating high-fidelity stable folds compared to other state-of-the-art methods. The success of FoldDesign should enable the creation of desired protein structures with promising clinical and industrial potential.

Proteins are important biological molecules that perform the majority of cellular functions in living organisms. Their unique and varied functions are made possible by the diverse structural folds adopted by different protein molecules. However, despite the enormous conformational space available, only a tiny portion appears in nature following billions of years of evolution, probably due to the selective pressures exerted by environmental constraints upon organisms (1). For example, there have been just under 1,500 protein folds classified in the SCOPe database (2), and studies have indicated that the current PDB is nearly complete, representing the vast majority of natural folds (3, 4). Given the vital importance of proteins to living organisms, there has been growing interest in designing artificial proteins with enhanced functionality beyond their native counterparts. However, many of the attempts have focused on generating new protein sequences starting from the structures of experimentally solved proteins (5–8). While this may be effective in certain cases, protein design starting from solved structures is severely limited as nature has essentially sampled from an insignificant portion of the possible structure and function space, thereby greatly limiting the number of design applications.

Given these limitations, de novo protein design, which aims to create not only artificial protein sequences, but also novel structures tailored to specific design applications, e.g., with specific fold types or binding pockets, has gained considerable traction in recent years. For instance, approaches such as Rosetta have been applied to design proteins with promising therapeutic potential (9–11), novel ligand-binding activity (12, 13), and complex logical interactions (14). The core protocol that has enabled Rosetta to design new protein folds is fragment assembly, which involves the identification of small structural fragments from experimentally solved structures that match a desired fold definition and the assembly of the identified fragments to produce full-length structural folds (15–17). Notably, fragment assembly was adapted from the related field of protein structure prediction, where it has been among the most successful classical approaches to template-free structure modeling (17–20). Despite the successes, de novo protein design remains somewhat of an art form, where large-scale experimental optimization is often required to generate successful designs (9, 11). In particular, extensive user intervention during scaffold

Author contributions: Y.Z. designed research; R.P. performed research; X.H. contributed new reagents/analytic tools; R.P. analyzed data; and R.P., X.H., G.S.O., and Y.Z. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: zhang@zhanggroup.org.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2208275120/-/DCSupplemental.

Published January 19, 2023.

creation and selection is frequently necessary (12, 21). Nevertheless, automated fold design tailored to specific applications is highly non-trivial because traditional homologous structure assembly programs often create folds that are similar to the template structures even when distracted with strong external spatial restraints (22, 23). Although ab initio fragment assembly approaches, such as QUARK (19) and Rosetta (17), can create template-free models, they need to start from specific natural sequences and often create conformations that either converge to specific folding clusters or are not protein-like (24).

Most recently, Anishchenko et al. performed an interesting study that combined deep neural-network training with structural refinement simulations to "hallucinate" proteins; it could create novel protein sequences but the structural folds were generally close to PDB structures (with an average TM-score = 0.78) (25). Meanwhile, the resulting protein folds were largely randomized depending on the stochastic process of the design iterations, where the method was further extended to allow for the incorporation of specific functional sites or structural motifs (Smotifs) (26). In another recent approach, Huang et al. combined a neural network-derived, sidechain-independent potential (SCUBA) with stochastic dynamics simulations and demonstrated an impressive ability to generate successfully folded designs (27). Notably, the method should be used in tandem with 3D backbone sketches adapted from a 'periodic table' of protein structures (28) through manual manipulation and thus the conformational space of the final structures is limited to the topological area defined by the initial backbone sketches. Similarly, extensions of the Rosetta fragment assembly protocol such as TopoBuilder require pre-definition of a target fold in the form of sketches that specify the 3D arrangement of the desired secondary structure (SS) elements, where the sketches are first parametrically optimized based on matching the desired fold with analogous structures in the PDB and then assembled from fragments that match the fold definition using Rosetta (29). Other methods like SEWING (30) have been successful at producing stable designs by reassembling relatively large helical substructures identified from the PDB; however, the approach is limited to the conformations adopted by large substructures present in the PDB and has been benchmarked only on helical folds (30, 31). Additionally, most of the successful de novo designs have highly idealized structural folds with optimized SS compositions that lack the complex irregularities often present in native proteins, where a significant portion of the designed folds are well represented in nature or may be described through ideal parametric geometries (32–36). Thus, the development of automated algorithms capable of precisely designing any required fold type, including those without structure analogs in the PDB or idealized SS compositions, with limited human intervention is critical to improve the scope and success rate of de novo protein design.

Toward this goal, we proposed an automated pipeline, FoldDesign, to create desired protein folds starting from user-specified restraints, such as the SS topology and/or inter-residue contact and distance maps, through sequence-independent replica-exchange Monte Carlo (REMC) simulations. Since the designed folds do not necessarily have experimental counterparts, we designed several objective assessment criteria based on the satisfaction rate of the input requirements and the folding stability of the designs, as outlined in *SI Appendix*, Fig. S1. The results showed that FoldDesign is capable of producing protein-like structural folds that closely recapitulate the input features with enhanced folding stability, significantly outperforming other start-of-the-art approaches on the large-scale benchmark tests. Importantly, this was demonstrated on a set of non-idealized, complex SS topologies and roughly 1/4 of the designs possessed novel folds that were not represented in the PDB, illustrating an important ability of the program to explore the areas of protein fold space unexplored by natural evolution. The online server, which presently supports fold design targets up to 1,500 residues long, and the standalone package for FoldDesign are freely available to the community at https://zhanggroup.org/FoldDesign/ and https://github.com/robpearc/FoldDesign, respectively.

Results and Discussion

FoldDesign is an automated algorithm for sequence-independent, de novo protein fold design, where the flowchart is outlined in Fig. 1. The program takes as input the SS topology for a designed structure scaffold, which includes the length, order, and composition of the SS elements. A set of structural fragments with lengths between 1 and 20 residues is then collected from the PDB library by scoring the similarity between the input SS and the SS of the PDB fragments. These fragments are finally reassembled by REMC folding simulations to generate protein-like structural scaffolds that satisfy the input constraints, where the lowest energy structure is subjected to further atomic-level refinement to produce the final structural design (see *Methods*).

Auxiliary Movements Improve the Folding Simulation Efficiency and Ability to Identify Low-Energy States. Fragment substitution is the predominant movement used by FoldDesign, which involves the replacement of a selected region of a decoy structure with the structure from one of the identified fragments collected from the PDB. However, fragment substitution may cause large conformational changes that prevent the movement from being accepted. To improve the simulation efficiency, FoldDesign introduces 10 auxiliary movements, including bond length and angle perturbations, segment rotations, torsion angle substitutions, and those that form packing interactions between specific SS elements (*SI Appendix*, Text S1 and Fig. S2).

Fig. 2*A* displays the FoldDesign energies of the lowest energy structures produced for each of the 354 test SS topologies (see *Methods*), either using the full set of 11 conformational movements or only using fragment substitution. Of note, the 354 test SS sequences were derived from native proteins, which include irregularities and non-ideal compositions, making it a rigorous test set to determine if a method can design stable structures given non-ideal SS definitions. It can be observed that the auxiliary movements enabled the simulations to find structures with significantly lower energies than those found using fragment substitution alone. Overall, the average FoldDesign energy of the best structures produced using the full movement set was −529.5 $k_B T$ compared to −449.7 $k_B T$ when using only fragment substitution, where the difference was statistically significant with a *P*-value of 2.1E-66 as determined by a paired two-sided Student's *t* test. In addition to the improved ability to sample low-energy states, the auxiliary movements reduced the simulation times required to fold the proteins. Fig. 2*B* plots the simulation time versus the protein length for each of the test topologies. From the figure, a clear reduction in the simulation time required can be seen across all protein lengths, where the average time for the simulations with the full movement set was 9.6 h compared to 22.8 h for the simulations that used only fragment substitution. This reduction in simulation time is due to the fact that fragment substitution is computationally expensive and requires additional loop closure to ensure that it does not cause large downstream perturbations, while the auxiliary movements are comparatively fast.

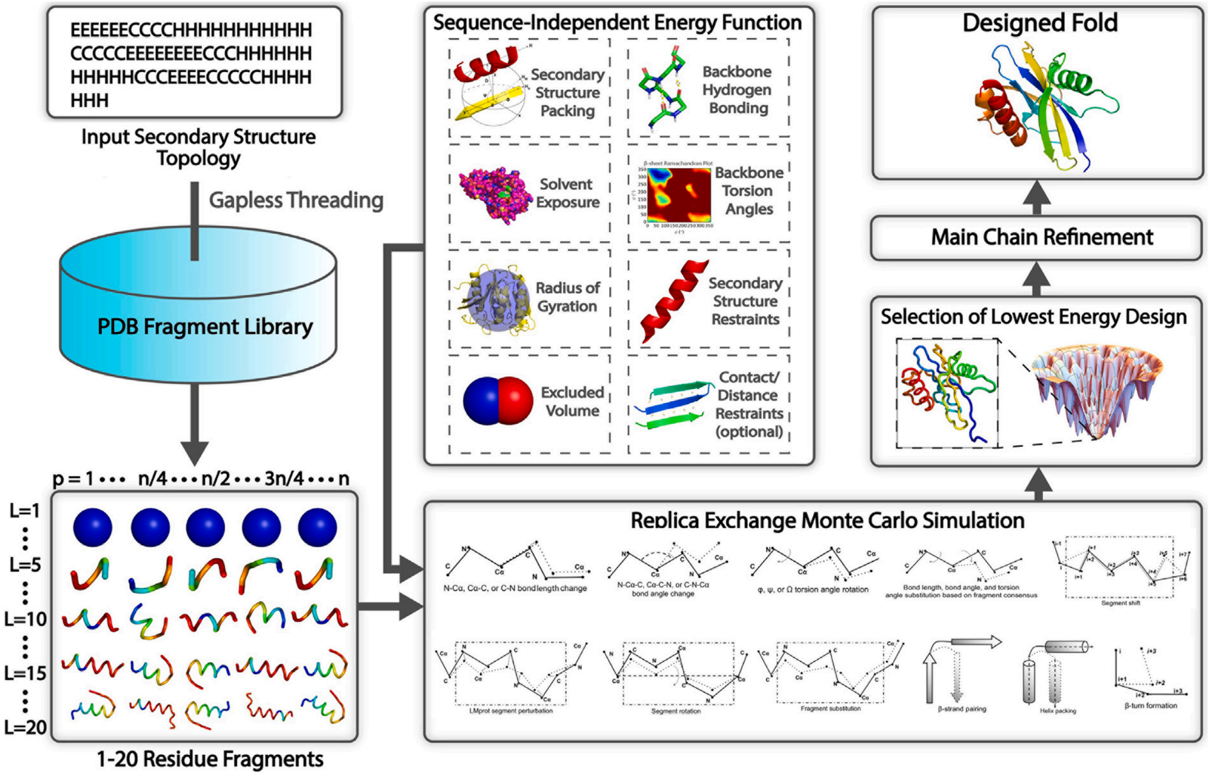
EEEEEECCCCHHHHHHHHHHH
CCCCCEEEEEEEECCCHHHHHH
HHHHHCCCEEEECCCCHHHH
HHH
Input Secondary Structure Topology
Gapless Threading
PDB Fragment Library
p = 1 ••• n/4 ••• n/2 ••• 3n/4 ••• n
L=1
L=5
L=10
L=15
L=20
1-20 Residue Fragments
Sequence-Independent Energy Function
Secondary Structure Packing
Backbone Hydrogen Bonding
Solvent Exposure
Backbone Torsion Angles
Radius of Gyration
Secondary Structure Restraints
Excluded Volume
Contact/ Distance Restraints (optional)
Designed Fold
Main Chain Refinement
Selection of Lowest Energy Design
Replica Exchange Monte Carlo Simulation
N-Cα, Cα-C, or C-N bondlength change
N-Cα-C, Cα-C-N, or C-N-Cα bond angle change
φ, ψ, or Ω torsion angle rotation
Bond length, bond angle, and torsion angle substitution based on fragment consensus
Segment shift
LMprot segment perturbation
Segment rotation
Fragment substitution
β-strand pairing
Helix packing
β-turn formation

Fig. 1. Overview of the FoldDesign pipeline. Starting from a user-defined SS topology as well as any further design constraints such as inter-residue contacts or distances, FoldDesign identifies 1 to 20 residue structural fragments from the PDB with SSs that match the input constraints. These fragments are then assembled together along with 10 other conformational movements during the REMC folding simulations under the guidance of a sequence-independent energy function that accounts for the fundamental forces that underlie protein folding. The lowest energy structure produced during the folding simulations is selected for further atomic-level refinement by ModRefiner to produce the final designed structure.

In Fig. 2 *C* and *D*, we further present a representative case study for the topology from the PDB protein 1ec6A, which adopts an α/β fold. Fig. 2*C* shows the conformational dynamics of the decoys produced during the lowest-temperature replica of the simulations using only the fragment substitution movement, while Fig. 2*D* uses the full movement set. Specifically, the figures plot the TM-score between the decoy at REMC cycle *i* compared to cycle *i-1* from cycles 50 to 100. In Fig. 2*C*, there are several plateaus where no movement could be accepted, leading to identical conformations between a number of cycles, where the most notable plateau lasted for 12 cycles (cycles 59 through 70). On the other hand, with the full movement set in Fig. 2*D*, no such plateaus were observed. Although several cycles had very similar folds, which may be caused by subtle conformational refinements such as bond length perturbation, none of the cycles had identical structures. As a result, the simulations using the full movement set generated a structure with an energy of −346.2 k_BT in 4.7 h compared to a structure with an energy of −224.3 k_BT in 14.2 h using only fragment substitution.

As a comparison, *SI Appendix*, Fig. S3 depicts the native 1ec6A structure, which had a higher FoldDesign energy (−145.5 k_BT) than either of the simulated designs in Fig. 2 *C* and *D*. This is expected as de novo protein design methods optimize the structure of a design with respect to their own energy functions and the native proteins from which an SS topology was derived will most likely never be the lowest energy conformation that the sampling procedures could/should achieve. Moreover, since many natural proteins with divergent global folds may adopt similar SS types, a given natural protein, such as 1ec6A, may not necessarily represent the most optimal fold or the lowest energy structure for a given SS composition, even with a perfect energy force field. In fact, it has been shown that many de novo designed proteins have increased stability compared to their native counterparts (35, 36). This is a departure from the scenario of protein structure prediction, in which the native structure, with some caveats, should lie at the global free energy minimum for a given protein sequence following Anfinsen's thermodynamic hypothesis (37); however, the same is not necessarily true for protein structure design given just the SS composition.

FoldDesign Scaffolds Closely Match the Input Constraints. To assess its ability to design structural folds that possess the desired SS topologies, we list in Table 1 a summary of the FoldDesign results in terms of the average Q3 scores on the 354 test topologies. As a comparison, we also list the results from the state-of-the-art Rosetta method (32), which similarly starts from the desired SS of a designed scaffold, where a detailed description of the procedures used to run Rosetta is given in *SI Appendix*, Text S3. Here, the Q3 score is defined as the fraction of positions with SS elements that are identical to that of the input topology. Following fold generation, the SSs of the designed scaffolds for both FoldDesign and Rosetta were assigned using DSSP (38) and compared to the input for each protein.

Overall, FoldDesign achieved an average Q3 score of 0.877 compared to 0.833 for Rosetta with a *P*-value of 1.7E-08. When considering the Q3 scores for α-proteins, β-proteins, and α/β-proteins separately, FoldDesign achieved Q3 scores of 0.934, 0.863, and 0.875, compared to 0.828, 0.829, and 0.835, respectively, for Rosetta. Thus, across all fold types, FoldDesign was able to generate structures that more closely matched the input topologies than Rosetta. This partially reflects the advanced dynamics of the folding simulations as well as the effectiveness of the optimized energy function in FoldDesign.

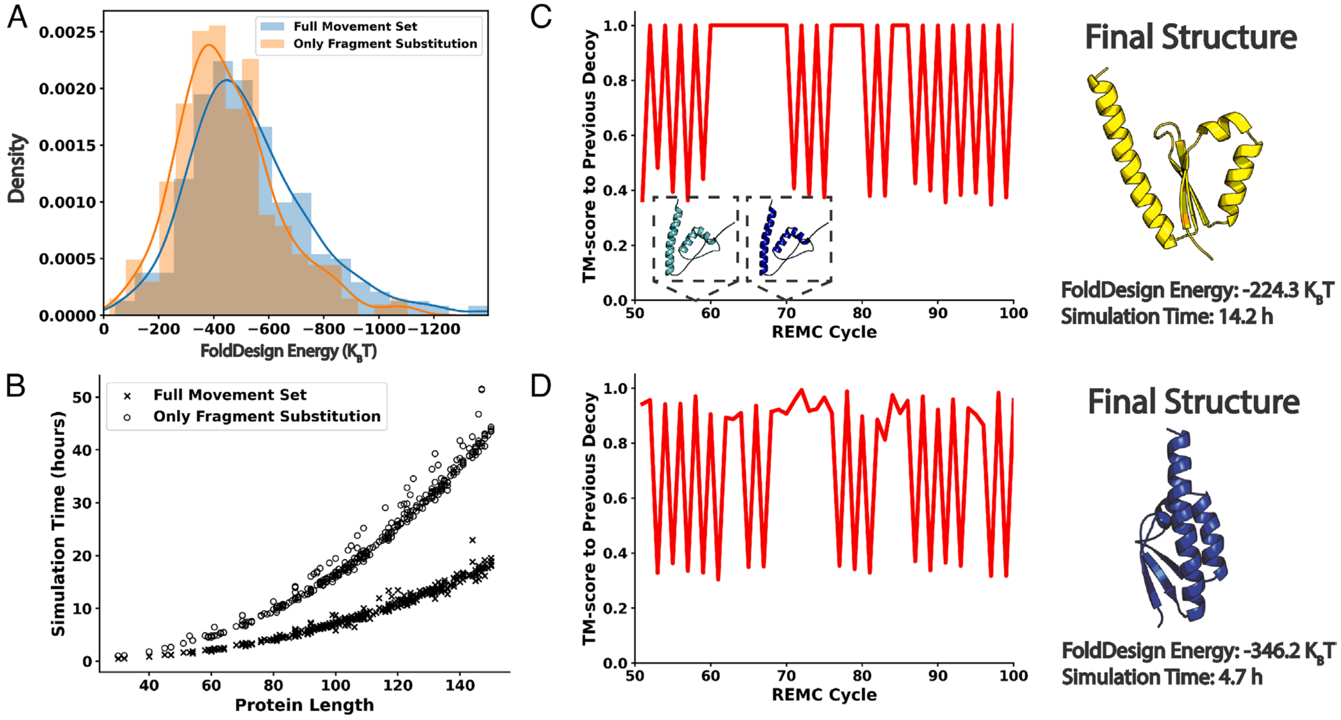

Fig. 2. Importance of the auxiliary conformational movements. (*A*) Energy distributions for the designs produced by the FoldDesign simulations using the full movement set and using only fragment assembly. (*B*) Simulation time required versus protein length for FoldDesign using the full movement set and fragment assembly alone. (*C* and *D*) Two representative case studies that demonstrate the dynamics of the folding simulations without (*C*) and with (*D*) the auxiliary movements. The y-axis displays the TM-score between the decoy at REMC cycle *i* compared to the decoy at cycle *i-1*.

Although no user-defined distance restraints were included in the above tests, these are still important in many design cases where recapitulation of specific folds is desired. In *SI Appendix*, Table S1, we extracted the pairwise Cα distances from the native structures in the test set and used them as restraints during the design simulations. From the table, it can be seen that FoldDesign was able to generate designs that closely matched the native structures with average TM-scores/RMSDs of 0.993/0.31 Å, 0.993/0.27 Å, 0.992/0.32 Å, and 0.994/0.31 Å for all, α, β, and α/β topologies, respectively. Here, TM-score (39) is a structure comparison metric that takes a value in the range (0, 1], where a value of TM-score =1 indicates an identical match between two structures and a TM-score ≥0.5 signifies that two proteins share the same global fold (40). Therefore, the FoldDesign structures nearly perfectly recapitulated the desired folds when guided by user-defined distance restraints. Additionally, the mean absolute errors between the Cα distance maps extracted from the designed folds and native structures were 0.148, 0.115, 0.130, and 0.154 Å for all, α, β, and α/β topologies, respectively, confirming that the generated structures closely satisfied the given distance restraints.

Table 1. Comparison of the Q3 scores for the structures produced by FoldDesign and Rosetta on the 354 test SS topologies

Method	Q3 Score All (*P*-value)	Q3 Score α-proteins (*P*-value)	Q3 Score β-proteins (*P*-value)	Q3 Score α/β proteins (*P*-value)
FoldDesign	0.877 (*)	0.934 (*)	0.863 (*)	0.875 (*)
Rosetta	0.833 (1.7E-08)	0.828 (5.4E-05)	0.829 (0.10)	0.835 (4.5E-06)

Here, the Q3 score is defined as the fraction of positions in the designed structures whose SSs were identical to the input SSs. The results are further separated based on the fold type (α, β, and α/β) and the *P*-values were calculated using paired, two-sided Student's *t* tests.

FoldDesign Generates Low-Energy, Native-Like Protein Structures. While an important metric, the Q3 score is unable to provide a complete picture of the physical quality of the designs. In theory, a method could produce trivial or even unfavorable folds that satisfy the desired SS definitions. Thus, a more detailed analysis of the energetics and physical characteristics of the produced structures had to be performed (*SI Appendix*, Fig. S1). As the designed scaffolds for FoldDesign and Rosetta are both sequence-independent and many of the traditional scoring and assessment tools are sequence-specific, the sequence for each scaffold had to be designed before further quantitative analysis could be conducted. To design the sequences for each scaffold, two sequence design methods were used, namely EvoEF2 (41) and RosettaFixBB (42), where the backbone structures of the designed scaffolds were kept fixed during the sequence design to ensure a fair comparison of the scaffolds that were directly output by FoldDesign and Rosetta. Here, RosettaFixBB and EvoEF2 are sequence design methods that perform Monte Carlo sampling in sequence space guided by combined physics- and knowledge-based energy functions. A total of 100 sequences were designed for each scaffold, and the average results from the 10 lowest energy sequences were reported for both FoldDesign and Rosetta in the following analyses.

First, Fig. 3*A* shows that the percent of buried residues for the FoldDesign scaffolds closely resembled the native protein structures from which the input SSs were extracted. For example, in the native structures, 19.2% of the residues were buried in the hydrophobic core, compared to 20.2% and 17.2% for the FoldDesign scaffolds whose sequences were designed by EvoEF2 and RosettaFixBB, respectively. However, for Rosetta,

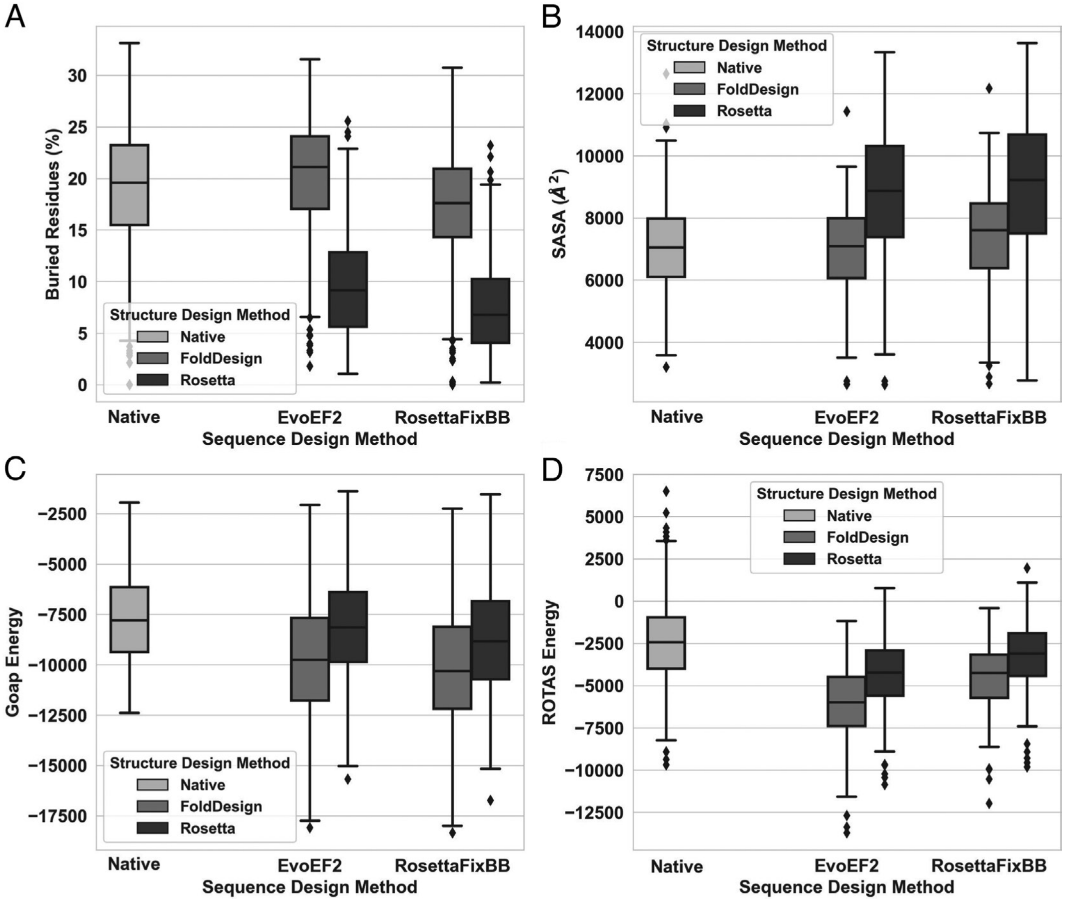

Fig. 3. Comparison of the physical characteristics and energies for the designed folds by FoldDesign and Rosetta on the 354 test proteins, where the sequence for each scaffold was designed by EvoEF2 and RosettaFixBB, respectively. The native designation represents the proteins from which the SSs of the designed folds were derived. (*A*) Percent of buried residues is plotted for each protein, where a buried residue was defined as having a relevant SASA <5%. (*B*) SASA for each protein. (*C* and *D*) Energies for each protein calculated by GOAP and ROTAS.

only 9.8% and 7.5% of the residues were buried in the hydrophobic core. Additionally, the solvent accessible surface area (SASA) for the native proteins was 7081.8 Å^2 compared to 6964.9 Å^2 and 7376.3 Å^2 for the FoldDesign scaffolds whose sequences were designed by EvoEF2 and RosettaFixBB, while the average SASA for the corresponding Rosetta scaffolds was 8721.2 Å^2 and 8944.2 Å^2, respectively. These results suggest that the FoldDesign scaffolds possessed more compact hydrophobic cores and less solvent-exposed area than the Rosetta scaffolds and shared a higher similarity to the native structures for these characteristics. The difference is in part due to the fact that FoldDesign includes a number of energy terms that promote the formation of well-packed SS elements; these include specific fragment-derived distance and solvation potentials, generic backbone atom distance energy terms, and SS-specific fragment packing terms (*SI Appendix*, Text S2). In addition, the energy weights were carefully optimized using the results of the design simulations to ensure the formation of well-folded globular proteins (see *Methods*).

In Fig. 3 *C* and *D*, we further display the energies of the designed scaffolds by FoldDesign and Rosetta, as assessed by two leading third-party atomic-level statistical energy functions, GOAP (43) and ROTAS (44). For the sequences designed by EvoEF2 and RosettaFixBB, the FoldDesign scaffolds had average GOAP energies of −9736.9 and −10166.7, which were significantly lower than the GOAP energies of −8174.5 and −8838.8 for the Rosetta scaffolds with *P* values of 3.4E-13 and 4.3E-10, respectively. Similar trends were observed for ROTAS. For the sequences designed by EvoEF2 and RosettaFixBB, the FoldDesign scaffolds had average ROTAS energies of −6110.3 and −4446.5 compared to −4360.8 and −3281.5 for the corresponding Rosetta designs; the differences were statistically significant with *P* values of 6.8E-27 and 1.3E-15. Overall, the FoldDesign scaffolds possessed more tightly packed hydrophobic cores and were energetically more favorable than the Rosetta scaffolds, with GOAP energies that were 19.1% and 15.0% lower than the Rosetta scaffolds and ROTAS energies that were 40.1% and 35.5% lower than the Rosetta scaffolds depending on the sequence design method that was used. Importantly, neither FoldDesign nor Rosetta used any of the third-party energy functions for optimization.

It is noted that introduction of ABEGO bias (45) during the Rosetta fragment selection protocol and enabling sub-rotamer sampling during the RosettaFixBB sequence design did not alter the above conclusions (*SI Appendix*, Text S4 and Figs. S4 and S5). Furthermore, despite the fact that Valine was used as the generic center of mass in FoldDesign and Rosetta (see *Methods*), neither method demonstrated a bias toward scaffolds that favored Valine as described in *SI Appendix*, Text S5 and Fig. S6, and all allowable regions of the Ramachandran plot were well represented in the FoldDesign scaffolds (*SI Appendix*, Fig. S7).

The FoldDesign Force Field Plays an Important Role in Promoting the Structural Design Performance. As shown in Eq. **1** in the *Methods* section, FoldDesign utilizes a number of newly introduced energy terms, including fragment-derived distance and solvation potentials ($E_{frag\_dist\_profile}$ and E_{frag_solv}) and detailed SS-specific packing potentials (E_{hhpack}, E_{sspack}, and E_{hspack}), as well as generic atomic contact- and distance-based terms that promote the formation of compact, globular structures ($E_{generic\_dist}$ and $E_{contact\_num}$). Moreover, these terms were optimally combined with other more routine energy terms using an extensive weight optimization protocol based on the 107 training proteins (see *Methods*).

To examine the impact of the FoldDesign force field and to probe the reason for the performance difference from the control method, we present in *SI Appendix*, Fig. S8 the comparative results for the physical characteristics of the Rosetta-designed scaffolds when the final models were selected using either the Rosetta or FoldDesign energy functions. It is noted that for this test we had to disable the fragment-derived distance and solvation potentials for FoldDesign as these are specific to the fragments generated by the FoldDesign program, which were not used to assemble the Rosetta designs given the differences in the fragment databases and identification protocols for the two methods. The data showed that selecting the Rosetta decoys according to their FoldDesign energies led to a significant improvement in the compactness of the folds as well as the GOAP and ROTAS energies compared to the designs selected using their original Rosetta energies. For example, the selection using the FoldDesign energy function increased the percent of buried residues by 31.5% for the EvoEF2 sequence designs and 39.3% for the RosettaFixBB sequence designs, compared to selection by the Rosetta centroid energy function, where the differences were statistically significant with *P*-values of 1.6E-13 and 1.5E-14, respectively. Similarly, improvements were observed in the third-party energies of the designed scaffolds. For example, the average GOAP energy improved by 9.2% and 7.6% for the EvoEF2 and RosettaFixBB sequence designs, respectively, where the differences were significant with *P* values of 3.1E-04 and 1.5E-03.

In *SI Appendix*, Fig. S9, we present a similar comparative result for the FoldDesign scaffolds when the final designs were selected by either the Rosetta or FoldDesign energy functions. For this test, an opposite trend was observed, where the selection of the FoldDesign scaffolds using the alternative force field from Rosetta resulted in a reduced performance compared to the original FoldDesign force field. For instance, the Rosetta energy-based selection led to a 43.2% and 49.4% decrease in the percent of buried residues for the EvoEF2 and RosettaFixBB sequence designs, compared to the models selected using the original FoldDesign energy function; these differences were statistically significant with *P* values of 8.2E-79 and 5.8E-86, respectively. Furthermore, the GOAP energies were 26.7% and 25.2% worse for the EvoEF2 and RosettaFixBB sequence designs with *P* values of 5.8E-35 and 9.8E-34, respectively. Based on the data shown in the above section, apart from the extensive REMC searching simulations, the optimized force field of FoldDesign, with newly introduced energy features, plays another critical role in creating compact and physically sound structure designs that outperform those from other state-of-the-art design methods.

FoldDesign Generates Stable Structures with Novel Folds. To further assess the stability of the designed structures, molecular dynamics (MD) simulations were run starting from the designed scaffolds produced by FoldDesign and Rosetta. MD is a useful tool as it allows for the study of protein motion and stability beyond static measurements such as energy calculations, where 20 ns unconstrained MD simulations were carried out using GROMACS (46) with the CHARMM36 force field (see *Methods*). Following the simulations, the final MD structures were obtained by clustering the 1,000 trajectories from the last nanosecond of each simulation using the GROMOS method with an RMSD cutoff of 2 Å, where the representative structure for each design was taken from the largest cluster center. To determine the stability of the structures, the TM-scores (39) between the initially designed scaffolds and the final clustered MD structures were calculated, where the results are depicted in Fig. 4 *A* and *B* for the structures whose sequences were designed by EvoEF2 and RosettaFixBB, respectively.

From the figures, it can be seen that the TM-scores between the initial FoldDesign scaffolds and the final MD structures were higher than those for the Rosetta scaffolds, indicating a closer match and thus more stable conformations for the FoldDesign scaffolds against MD-based perturbations. For instance, the average TM-score between the FoldDesign scaffolds and final MD structures for the EvoEF2 sequence designs was 0.645 compared to 0.584 for the corresponding Rosetta scaffolds (Fig. 4*A*), where the difference was statistically significant with a *P* value of 7.4E-19. A similar trend was observed for the scaffolds whose sequences were designed by RosettaFixBB, where the average TM-score between the initial FoldDesign structures and the final MD structures was 0.602 compared to 0.525 for the Rosetta scaffolds with a *P*-value of 4.6E-26 (Fig. 4*B*). Furthermore, when considering a cutoff TM-score of 0.5, 93.7% and 87.9% of the FoldDesign scaffolds whose sequences were designed by EvoEF2 and RosettaFixBB, respectively, shared the same global folds as their final MD structures, compared to 77.1% and 54.8% of the corresponding Rosetta structures. Fig. 5*A* shows three examples selected from among the most stable FoldDesign scaffolds, where the TM-scores were all greater than 0.8 and the RMSDs were less than 2Å, indicating a close atomic match between the designed scaffolds and the final MD structures. Overall, the vast majority of the FoldDesign scaffolds possessed stable global folds, outperforming the state-of-the-art Rosetta protocol across the test set.

Interestingly, despite the high fold stability with local structural features that were highly similar to the native proteins, a large portion of the FoldDesign scaffolds adopted novel folds that were different from what exists in the PDB. In Fig. 4*E*, we present a histogram distribution of the TM-scores between the FoldDesign scaffolds and the closest structures identified by TM-align (47) from the PDB, where the average TM-score of 0.551 was relatively low given the searching power of TM-align and the near completeness of the PDB (3, 47). Of the 354 designs, 79 had a TM-score below 0.5 to any structure in the PDB, indicating they possessed novel folds, while the remaining 275 designs had analogous structures in the PDB with the same global folds (TM-scores ≥ 0.5). Furthermore, 74 of the 79 novel structures whose sequences were designed by EvoEF2 had stable folds with TM-scores ≥0.5 to their final structures output by the MD simulations. Moreover, there was no obvious difference between the novel folds and other folds in terms of stability, as the TM-score distributions between the designs and the final MD structures were quite similar (*SI Appendix*, Fig. S10), where their average TM-scores were 0.647 and 0.645, respectively. These results demonstrate that FoldDesign is capable of producing compact and stable scaffolds, while allowing for the exploration of novel areas of protein fold space.

Protein Structure Prediction Indicates FoldDesign Produces Well-Folded Structures. As additional proof of the foldability of the designed structures, we examined the structural similarity between the designed scaffolds and the predicted models generated by the state-of-the-art AlphaFold2 program (48) starting from the designed sequences for each scaffold. As protein structure prediction is essentially the inverse problem of protein design, it

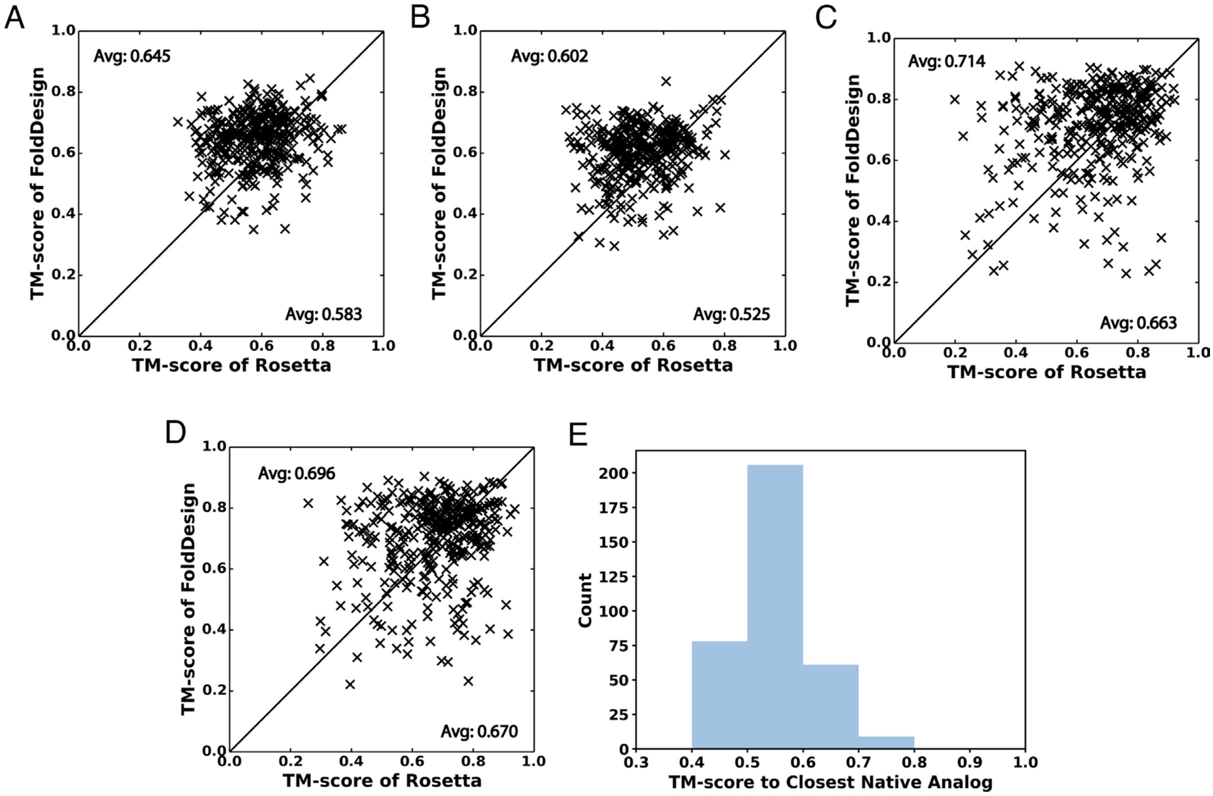

Fig. 4. Analysis of the FoldDesign and Rosetta scaffolds using MD (*A* and *B*) and protein structure prediction by AlphaFold2 (*C* and *D*). (*A* and *B*) TM-scores of the FoldDesign and Rosetta scaffolds relative to their final structures following 20 ns MD simulations, where the sequence for each scaffold was designed by EvoEF2 (*A*) and RosettaFixBB (*B*). (*C* and *D*) TM-scores of the FoldDesign and Rosetta scaffolds relative to the structures predicted by AlphaFold2 starting from the EvoEF2 (*C*) and RosettaFixBB (*D*) sequences designed for each scaffold. (*E*) TM-score distribution between the FoldDesign structures and their closest native analogs obtained by searching the designed scaffolds through the PDB using TM-align.

would stand to reason that well-formed structure designs should be able to be recapitulated starting from their corresponding designed sequences.

However, given that AlphaFold2 is a deep learning-based modeling program, its performance largely depends on collecting meaningful MSAs (48), yet de novo designed proteins almost always lack natural sequence homologs. To illustrate this, in *SI Appendix*, Fig. S11 we plot the number of Blast hits that were detected from the nr sequence database (E-value < 1E-5) when starting from either a single designed sequence or from jumpstarting the Blast search using an alignment of all 100 designed sequences for each FoldDesign scaffold. As shown in *SI Appendix*, Fig. S11*A*, no Blast hits were detected when starting from a single EvoEF2 sequence design and jumpstarting the Blast search from the alignment of designed sequences only picked up 1 to 2 hits for 4 of the 354 designs. For the RosettaFixBB designs, neither the single-designed sequence searches nor the jumpstarted Blast searches yielded any detectable homologs (*SI Appendix*, Fig. S11*B*).

In *SI Appendix*, Table S2, we also list the structure prediction results by AlphaFold2 for the 354 native protein structures starting from the MSAs generated by the DeepMSA program (49) compared to the results starting from the single designed sequences. As expected, AlphaFold2 created excellent models with an average TM-score of 0.913 when starting from the native MSAs; but starting from the single designed sequences by either EvoEF2 or RosettaFixBB produced significantly less accurate models, where the average TM-scores were only 0.506 and 0.482 for the EvoEF2 and RosettaFixBB sequence designs, respectively, and nearly (or more than) half of the cases had TM-scores below 0.5. This result is in line with previous studies that have indicated that single sequence-based modeling using deep learning approaches for non-ideal folds is significantly less accurate than that for idealized de novo designed folds (50). This is likely due to the fact that most of the computationally designed structures have relatively simple global folds with optimized SS compositions that lack the irregularities that exist in native proteins (33, 35, 36). Since the 354 SS topologies in the benchmark dataset were derived from native protein structures, which contain numerous irregularities, the above results indicate that single sequence-based AlphaFold2 modeling may not be reliable for the FoldDesign and Rosetta scaffolds. Interestingly, when starting from artificial MSAs collected from the 100 designed sequences for the native structures, AlphaFold2 could generate reasonable folding results, where more than 97% of the cases had TM-scores >0.5, which was close to the modeling results obtained when starting from the DeepMSA MSAs (*SI Appendix*, Table S2). This demonstrates that the MSAs collected from sequence design simulations contain some level of evolutionary information that can facilitate deep learning-based structure prediction.

Thus, given the lack of natural sequence homologs and the difficulty of AlphaFold2 to model complicated folds from single sequence designs, we constructed the input MSAs for AlphaFold2 by taking the 100 sequences designed by EvoEF2 and RosettaFixBB for each of the FoldDesign/Rosetta scaffolds. As shown in *SI Appendix*, Table S3, when starting from the sequences designed by EvoEF2, the average TM-score between the AlphaFold2 models and the FoldDesign scaffolds was 0.714 compared to 0.663 for the Rosetta scaffolds, where the difference was statistically significant with a *P*-value of 4.6E-09. In Fig. 4*C*, we present a head-to-head TM-score comparison, where the FoldDesign scaffolds had higher TM-scores than the corresponding Rosetta scaffolds

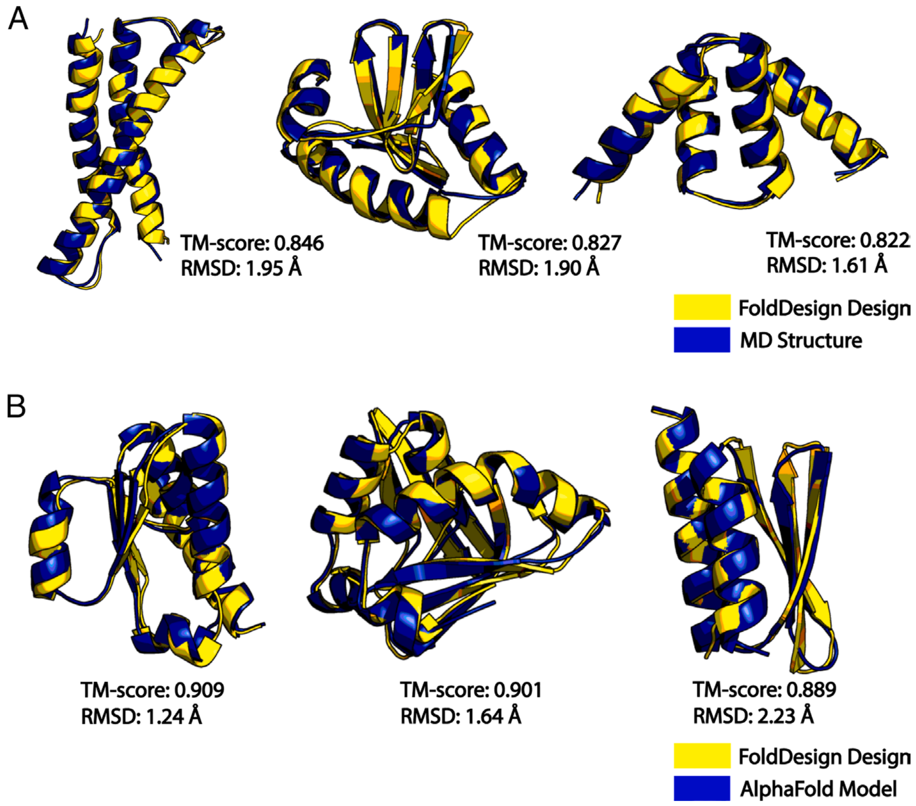

Fig. 5. Examples of stable, well-folded FoldDesign scaffolds as assessed by MD (*A*) and AlphaFold2 (*B*), where the sequences for each scaffold were designed by EvoEF2. (*A*) The initial FoldDesign structures (yellow) superposed with the final MD structures (blue). (*B*) The FoldDesign scaffolds (yellow) superposed with the AlphaFold2 models (blue).

for 211 cases, while Rosetta did so for 133 of the 354 cases. If we consider the number of designs with TM-score ≥0.5, 324 (or 91.5%) of the FoldDesign scaffolds shared the same global folds as the AlphaFold2 models compared to 301 (or 85.0%) of the scaffolds by Rosetta. These results demonstrate that the FoldDesign scaffolds more closely resembled the AlphaFold2 models than the Rosetta scaffolds did, indicating their enhanced stability/foldability. Similar patterns were observed for the sequences designed by RosettaFixBB, where the average TM-score between the FoldDesign scaffolds and AlphaFold2 models was 0.696 compared to 0.670 for Rosetta with a *P*-value of 3.0E-04 (*SI Appendix*, Table S3). Moreover, 208 of the 354 FoldDesign scaffolds had higher TM-scores than the Rosetta scaffolds and 315 (or 89.0%) of the designs had TM-scores ≥0.5 (Fig. 4*D*).

Fig. 5*B* presents three examples from some of the closest matches between the FoldDesign scaffolds and AlphaFold2 models, where each had a TM-score greater than or close to 0.9 and RMSDs below 2.25 Å, indicating close atomic matches between the designed scaffolds and predicted models. Notably, these cases came from designs with some level of analogous structural information in the PDB, although the TM-scores between the designed scaffolds and their closest native analogs (0.517 to 0.611, see *SI Appendix*, Fig. S12) were much lower than those between the designed scaffolds and the AlphaFold2 predicted models (0.889 to 0.909, Fig. 5*B*). To further examine the foldability of the novel structures produced by FoldDesign, *SI Appendix*, Fig. S13 plots the AlphaFold2 TM-score distributions for the FoldDesign scaffolds that lacked or possessed native analogs, where the novel designs (with TM-score = 0.723/0.718 for the EvoEF2/RosettaFixBB sequences) were found to be as foldable or even more so than those with native analogs (with TM-score = 0.711/0.689 for the EvoEF2/RosettaFixBB sequences). Overall, these tests demonstrated that the FoldDesign scaffolds more closely matched the predicted models than the Rosetta scaffolds did, and the overwhelming majority of the designs shared the same global folds as the AlphaFold2 models. This structural consistency may suggest that FoldDesign captures some structural characteristics that have been integrated in the AlphaFold2 learning process.

Assembling Uncommon Smotifs Is Essential to Produce Novel Fold Designs. Given the high population of novel folds produced by FoldDesign starting from native SS compositions, it was of interest to quantitatively examine the structural characteristics of these folds and determine how they deviate from native protein structures. Toward this goal, we first examined their local structural quality using MolProbity (MP) (51), where the results are summarized in *SI Appendix*, Table S4. It was observed that the novel designs possessed favorable MP-scores, with an average MP-score of 1.66 compared to 1.57 for the designs that had identifiable native analogs, where both scores were comparable to (or only slightly higher than) those of the corresponding native structures (1.19). Meanwhile, the novel folds had very few Ramachandran outliers, atomic clashes, or deviations in bond lengths and angles, largely comparable to (or slightly better than) the native and analogous designs. This result provides support that the novel folds possessed favorable local geometries and physical realism that resembled native proteins, although they had completely different global folds.

To further probe the source of the distinct structural folds adopted by the novel designs, following the idea of previous studies (52–54), we investigated the local geometries of the associated super-SS elements by decomposing the global folds into their local Smotifs. Briefly, a Smotif is composed of two adjoining regular

SS elements, either helices or strands, that are linked by a loop region (52). As shown in *SI Appendix*, Fig. S14, the geometry of a Smotif is specified by four spatial characteristics, including the distance (D) between the bracing SS elements and the three angles formed between them (hoist δ, packing θ, and meridian ρ). The overall fold of a protein can then be broken down into the basic SS building blocks, where a total of 540 Smotif types can be obtained by splitting the four-dimensional (D-δ-θ-ρ) space into 4-3-3-6 intervals and only ~320 to 330 Smotif geometries can be used to describe all existing protein structures (53). In Fig. 6, we present the relative frequency of Smotifs in the 79 novel folds and 354 native proteins in the test set versus the normalized background frequency of the Smotifs calculated from 51,094 non-redundant full-chain structures in the I-TASSER template library (55, 56), where the relative frequency values were normalized for each protein across the four background frequency bins in the plot (see *SI Appendix*, Eq. **S16** in *SI Appendix*, Text S10).

It can be observed from Fig. 6 that compared to the native proteins, the novel designs by FoldDesign were highly enriched for rare or uncommon Smotifs, where 24.5% and 70.8% of the Smotifs in the novel designs had normalized background frequencies in the range [0, 1E-3] and (1E-3, 1E-2], respectively, compared to just 4.5% and 29.7% for the 354 native proteins. Additionally, 50.6% of the Smotifs from the native folds were common with background frequencies >1E-1, while just 4.1% of the Smotifs from the novel designed folds were commonly found. Of note, the vast majority of the Smotifs in the novel designs were found in nature, with the exception of one geometry that did not appear in the proteins from the PDB as shown in *SI Appendix*, Fig. S15. Thus, the novelty of the designed folds by FoldDesign may largely be a consequence of the combination of rare/uncommon local super-SS geometries, rather than the creation of new local geometries or a unique arrangement of common Smotifs. Furthermore, given the computationally assessed stability of the novel folds, these results support the claim that FoldDesign is able to produce stable designs for non-idealized SS elements, as the majority of the super-SS geometries were rarely observed in nature.

Fig. 7 highlights two design cases with novel folds whose SS compositions were taken from the PDB proteins 1id0A and 2p19A, where the designed scaffolds are shown superposed with their AlphaFold2 models and closest native analogs from the PDB. It can be observed that the AlphaFold2 models closely resembled the designed scaffolds with TM-scores of 0.809 and 0.811 for the 1id0A and 2p19A designs, respectively, indicating they were foldable by the deep learning program. Interestingly, the clusters that these designs were selected from were highly conserved with average TM-scores of 0.769/0.826 between the cluster members and 1id0A/2p19A, pointing to a clear evolutionary relationship between the SS topologies and the native folds. Despite this, FoldDesign generated novel scaffolds for these two topologies, which had low TM-scores (0.467 and 0.451) to their closest structures in the PDB, again demonstrating an ability to explore structure space unexplored by nature even for highly conserved clusters.

In the right column of Fig. 7, we illustrate the Smotifs that the two designs were composed of, where the Smotifs for the two native structures are shown in *SI Appendix*, Fig. S16. For the 1id0A topology design, the global structure was composed of eight Smotifs, where all eight were rare with a background frequency ≤1E-3, while the corresponding native structure was composed of 7 common

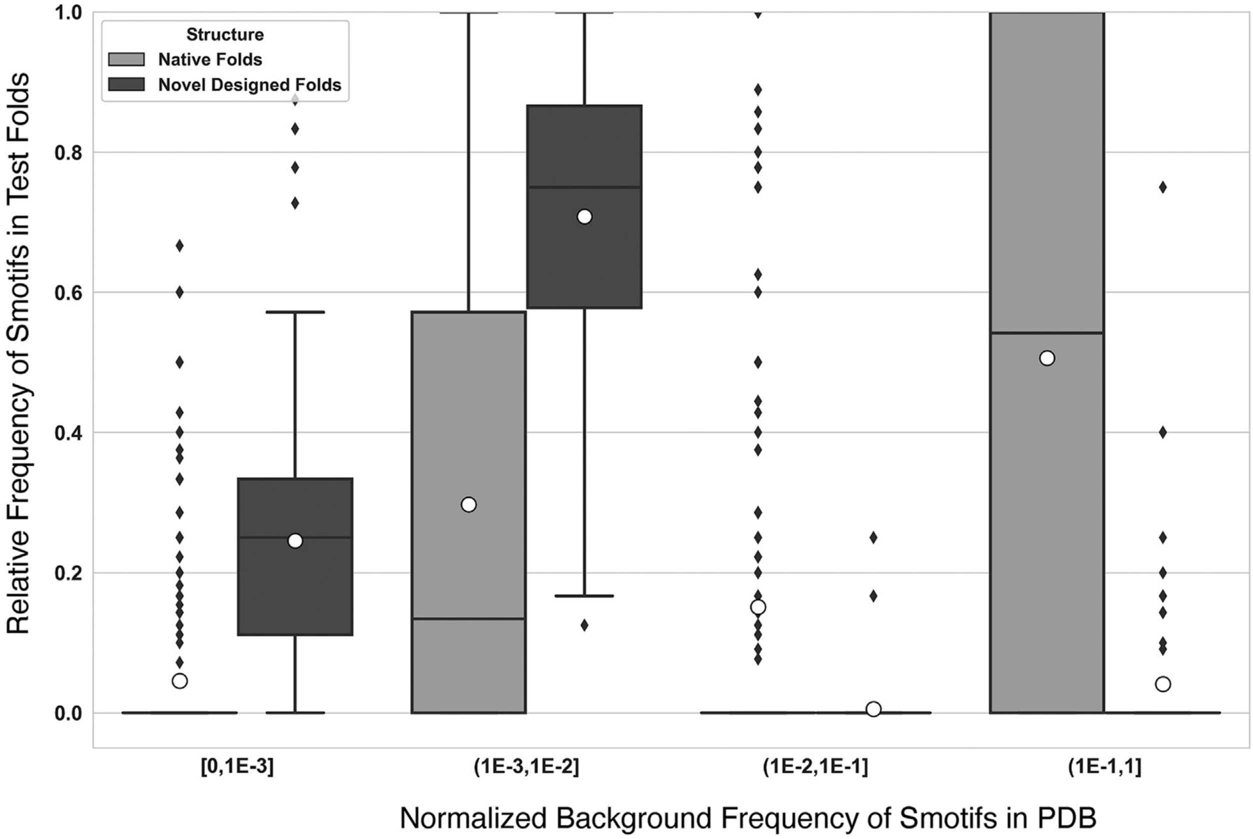

Fig. 6. Relative frequency of Smotifs found in the 354 native protein structures and 79 novel folds produced by FoldDesign vs. the normalized background frequency of the Smotifs calculated from the 51,094 non-redundant full-chain structures in the I-TASSER template library (*SI Appendix*, Text S10). Two motifs are considered as identical if they fall into the same bin in the four-dimensional (D-δ-θ-ρ) space (53). The mean values of the distributions are shown by the white circles, where a point with 0-frequency indicates that a Smotif with the indicated background frequency did not appear in one of the tested structural folds.

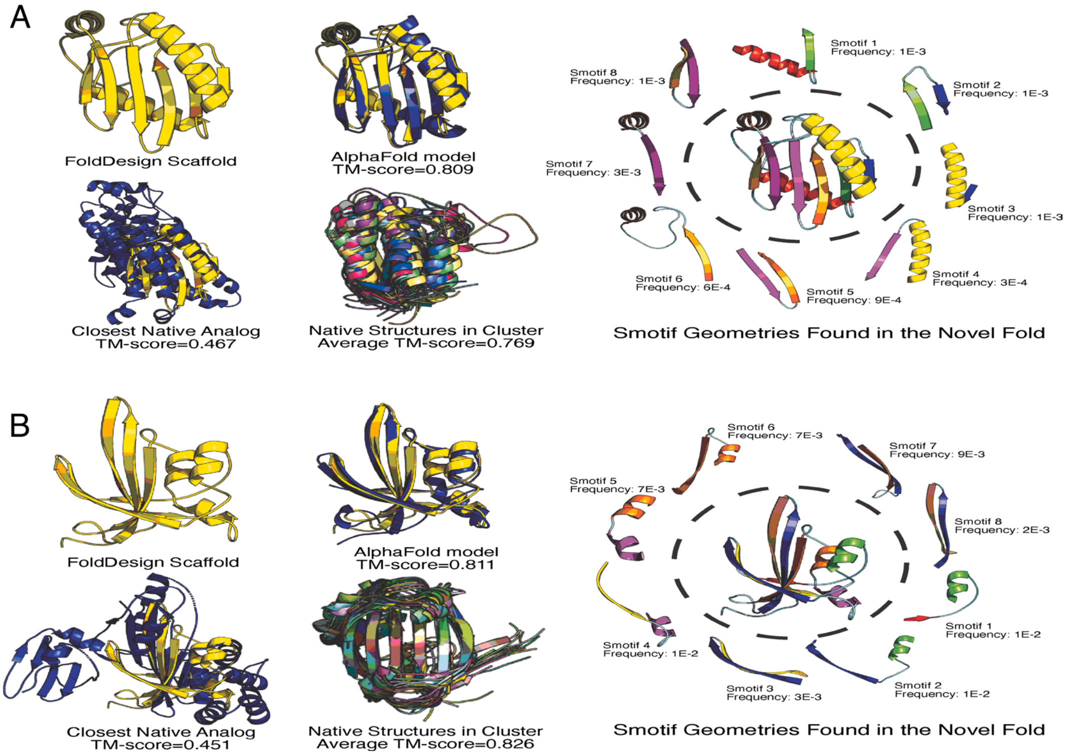

Fig. 7. Case study of two novel designed folds for the SS topologies taken from 1id0A (*A*) and 2p19A (*B*). The designed structures are shown on the left-hand side of the figure in yellow superposed with their AlphaFold2 models and closest native analogs in blue. Additionally, each native structure in the same SS cluster as 1id0A (*A*) and 2p19A (*B*) are shown aligned with their respective cluster centers, where the average TM-scores were calculated based on the alignment of each structure in the cluster to the cluster center. Lastly, the right-hand side of the figure illustrates the Smotif geometries found in the novel folds, where the depicted frequencies for each Smotif represent the relative background frequencies calculated from the representative structures in the PDB.

Smotifs with a high background frequency of ~3E-1 and 1 Smotif that was less common with a background frequency of ~1E-2. Similar trends were observed for 2p19A, where the designed structure was composed of eight uncommon Smotifs with a background frequency ≤1E-2, while the native structure was composed of eight common Smotifs with a background frequency of ~3E-1. Thus, from these cases, it can be seen that the combination of rare or uncommon local super-SS geometries gave rise to new global folds, which was observed across the 79 novel designs.

Concluding Remarks

Protein design generally consists of two steps of structural fold design and sequence design. Many protein design efforts have focused on the second step of sequence design with input scaffolds taken from existing protein structures in the PDB. Despite the success, such experiments constrain design cases to the limited number of folds adopted by natural proteins, while curtailing the exploration of novel areas of protein structure and biological function.

In this work, we developed a pipeline, FoldDesign, for de novo protein fold design. Different from traditional protein folding simulations which start from native sequences and therefore, as expected, often result in folds that are similar to what exists in the PDB library, FoldDesign starts from structural restraints (e.g., SS assignments and/or inter-residue distance restraints) and performs folding simulations under the guidance of an optimized sequence-independent energy function. Large-scale tests on a set of 354 unique, non-ideal fold topologies demonstrated that FoldDesign could create protein-like folds with a closer Q3 score similarity to the desired structural restraints than the state-of-the-art design program, Rosetta. Meanwhile, the FoldDesign scaffolds had well-compacted core structures with buried residue rates and solvent-exposed areas that more closely matched those of native proteins, while MD simulations showed that the folds were more stable than those produced by Rosetta. Importantly, FoldDesign is capable of designing folds that are completely different from the native structures in the PDB, highlighting its ability to explore novel areas of protein structure space despite the high fidelity to the input restraints and the native-like local structural characteristics. Detailed data analyses showed that the major contributions to the success of fold design lie in the optimal energy force field, which contains a balanced set of energy terms that account for fragment and SS packing, as well as the efficient exploration of conformational space through REMC simulations assisted with a composite set of efficient movements. It was also found that the ability to identify and assemble less common super-SS geometries from the PDB, rather than creating new motifs or the unique arrangement of common SS motifs, represents the key for FoldDesign to create novel fold designs.

Although the FoldDesign server outputs both the designed fold and the lowest energy designed sequences when combined with the EvoDesign/EvoEF2 programs (5, 41), the validation of the designed sequences remains to be experimentally examined. However, complete experimental validation requires both designed structures and designed sequences, where the latter is out of scope of the present study, and we leave this important work to future investigation. Nevertheless, the findings presented here have shown that FoldDesign can be used as a robust tool for generating high-quality, stable structural folds when applied to the very challenging task of completely de novo scaffold generation without human-expert intervention. This therefore provides a strong potential for the experimental protein design to effectively explore both structural and functional spaces which natural proteins have not reached despite billions of years of evolution.

Methods

FoldDesign aims to automatically design desired protein structure folds starting from user-specified rules such as SS composition and/or inter-residue contact and distance maps. The pipeline consists of three main steps, including fragment generation, REMC folding simulations, and main chain refinement and fold selection (Fig. 1).

Fragment Generation. Starting from a user-specified SS, high-scoring fragments are identified from a fragment library, which consists of structural fragments collected from a non-redundant set of 29,156 high-resolution PDB structures used by QUARK (19, 57). The fragments were collected from structures deposited on or before 4/3/2014 and shared <30% sequence identity to each other (19, 57). Notably, this library has been extensively validated in the related field of protein structure prediction in the most recent CASP experiments (58, 59). Gapless threading through the library is performed to generate 1 to 20 residue fragments, where the fragments are scored based on the compatibility of their torsion angles and SS similarity to the desired SS at each position. The top 200 fragments are generated for each overlapping 1 to 20 residue window. The information for each fragment includes the backbone bond lengths, bond angles, and torsion angles, as well as other useful data such as the position-specific solvent accessibility and Cα coordinates, which are later used to derive distance and solvation restraints.

REMC Folding Simulations and Refinement. Following fragment generation, REMC folding simulations are performed in order to assemble full-length structural models, where each simulation uses 40 replicas and runs 500 REMC cycles (see *SI Appendix*, Text S1 for a full description of the REMC parameters and movements). The protein conformation in FoldDesign is represented with a coarse-grained model, which specifies the backbone N, Cα, C, H, and O atoms as well as the Cβ atoms and an atom that represents the side-chain center of mass (*SI Appendix*, Fig. S17). To allow for a less biased exploration of structure space, the energy terms used by FoldDesign are sequence-independent, where the side-chain center of mass for Valine is used as the generic center of mass for each residue to minimize steric clashes.

The initial conformations are produced by randomly assembling different high-scoring 9 residue fragments and then minimized using a set of 11 movements. Here, the major conformational movement is fragment substitution, which involves swapping a selected region of a decoy structure with the structure from one of the fragments randomly selected from the fragment library. Next, cyclical coordinate descent loop closure (60) is used to minimize the structural perturbations downstream. Since FoldDesign uses 1 to 20 residues fragments, larger fragment insertions are typically attempted during the initial REMC cycles, while smaller ones are attempted during the later steps of the simulations to improve its acceptance rate when the protein is more globular and well-folded. In addition to fragment insertion, 10 other conformational movements are attempted throughout the course of the simulations, including perturbing the backbone bond lengths, angles or torsion angles, segment rotations, segment shifts, and movements that form specific interactions between different SS elements, where these are described in *SI Appendix*, Text S1 and Fig. S2.

The movements are accepted or rejected using the Metropolis criterion (61), where the energy for each conformation is assessed by the following energy function:

$$\begin{aligned}E_{FoldDesign} = {} & E_{HB} + E_{ss\_satisfaction} + E_{rama} + E_{hhpack} + E_{sspack} \\ & + E_{\mathrm{hspack}} + E_{ev} + E_{generic\_dist} + E_{frag\_dist\_profile} \\ & + E_{frag\_solv} + E_{rg} + E_{contact\_num}.\end{aligned} \quad [\mathbf{1}]$$

Here, 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 that account for backbone hydrogen bonding, the satisfaction rate of the input SS, 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. A more detailed explanation of these terms is given in *SI Appendix*, Text S2. After the REMC simulations are completed, the design with the lowest energy is selected for further atomic-level refinement, for which sequence design and structural refinement are performed iteratively using EvoDesign (5) and ModRefiner (62), respectively.

Training and Test Dataset Collection. To test FoldDesign's ability to perform de novo protein fold design, we collected a non-redundant set of SS sequences. This was accomplished by extracting the three-state SSs from 76,166 protein domains in the I-TASSER template library (55, 56) using DSSP (38). All of the pairwise SS alignments were obtained using Needleman–Wunsch dynamic programming to align the three-state SS sequences. The target sequences were then clustered based on the distance matrix defined by their SS identities, i.e., the number of identical SSs divided by the total alignment length, where an identity cutoff =70% was used to define the clusters.

The identified clusters were further refined by eliminating atypical SS topologies (clusters with less than 10 members) and by selecting only those clusters where a clear relationship existed between the SS and the tertiary structure adopted by the cluster members. The latter requirement was accomplished by using TM-align (47) to perform structural alignment between each cluster member and the cluster center, where conserved clusters were required to have an average TM-score ≥0.5 between the members and cluster center. Finally, we obtained 461 clusters; 107 and 354 SS sequences were used for the training and test sets, respectively. The training set was composed of 22 α, 25 β, and 60 α/β topologies, while the test set was composed of 24 α, 55 β, and 275 α/β topologies.

FoldDesign Energy Function Optimization. In order to ensure proper structure generation, each energy term must be carefully weighted in the FoldDesign energy function. This was done on the 107 training topologies. Briefly, a grid searching strategy was used to optimize the weights, where all weights were initially assigned as 0, except for the weight for the steric clash term, which was set to 1.0. Then the values for each weight were adjusted one-at-a-time around the grid values and the FoldDesign simulations were run to produce scaffold structures using the new weight set. After structure generation, the sequences for each scaffold were designed using EvoEF2 (41) and the designed structures were assessed based on:

$$E_{accept} = -\Delta EvoEF2 + 100 * \Delta BuriedResidues + 100 * \Delta Q3Score. \quad [\mathbf{2}]$$

where, $\Delta EvoEF2$, $\Delta BuriedResidues$, and $\Delta Q3Score$ are the changes in the average EvoEF2 energy, percent of buried residues, and SS Q3 score, respectively, between the structures produced by the old and new weight sets. If the new weighting parameter increased the value of E_{accept}, the weights were accepted. Once the initial weights for each energy term were determined, many more iterations were conducted to precisely fine-tune their values based on Eq. **2** as well as by hand inspection of the structures. Although time-consuming, the process of directly optimizing the weights based on the results of the folding simulations resulted in high-quality scaffolds with physical characteristics that resembled native proteins.

MD Simulation for Examining Fold Stability. To examine the stability of the FoldDesign scaffolds, we performed MD simulations starting from the designed structures. For each simulation, a dodecahedron box was constructed with a distance of 10 Å from the solute and filled with TIP3P water molecules, where Na^+ and Cl^- ions were used to neutralize the charge of the system. Following this, energy minimization was carried out using the steepest descent minimization with a maximum force of 10 kJ/mol. The system was then equilibrated at 300 K using 100 ps NVT simulations and 100 ps NPT simulations with position restraints (1,000 kJ/mol) on the heavy atoms of the protein. After the two equilibration phases, the system was well-equilibrated at the desired temperature and pressure, and unconstrained MD simulations were performed at 300 K for 20 ns. During the simulations, non-bonded interactions were truncated at 12 Å and the Particle Mesh Ewald methods was used for long-range electrostatic interactions. Lastly, the velocity-rescaling thermostat and Parrinello–Rahman barostat were used to couple the temperature and pressure, respectively. A total of 1,000 structures were collected from the MD trajectories during the final nanosecond of the simulations. This ensemble was then clustered using the GROMOS method with an RMSD cutoff of 2 Å, and the final MD structure for each simulation was collected from the cluster center.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*. The online server, stand-alone program and benchmark data for FoldDesign are available at https://zhanggroup.org/FoldDesign/, while the stand-alone program may also be downloaded from https://github.com/robpearc/FoldDesign.

ACKNOWLEDGMENTS. This work used the Extreme Science and Engineering Discovery Environment , which is supported by the NSF (ACI-1548562). This work was supported in part by the NIGMS (GM136422, S10OD026825), the NIAID (AI134678), the NSF (IIS1901191, DBI2030790, MTM2025426), the NCI (U24CA210967), and the NIEHS (P30ES017885).

Author affiliations: [a]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [b]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109; [c]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109; [d]School of Public Health, University of Michigan, Ann Arbor, MI 48109; [e]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109; [f]Department of Computer Science, School of Computing, National University of Singapore 117417, Singapore; and [g]Cancer Science Institute of Singapore, National University of Singapore 117599, Singapore

1. C. R. Darwin, *The Origin of Species* (John Murry, Landon, 1859).
2. N. K. Fox, S. E. Brenner, J. M. Chandonia, SCOPe: Structural Classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
3. Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, J. Skolnick, On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2605–2610 (2006).
4. Y. Zhang, J. Skolnick, The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1029–1034 (2005).
5. R. Pearce, X. Huang, D. Setiawan, Y. Zhang, EvoDesign: Designing protein-protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function. *J. Mol. Biol.* **431**, 2467–2476 (2019).
6. B. I. Dahiyat, C. A. Sarisky, S. L. Mayo, De novo protein design: Towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789–796 (1997).
7. G. A. Lazar, J. R. Desjarlais, T. M. Handel, De novo design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167–1178 (1997).
8. G. Dantas, B. Kuhlman, D. Callender, M. Wong, D. Baker, A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460 (2003).
9. L. Cao *et al.*, De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426–431 (2020).
10. D.-A. Silva *et al.*, De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
11. A. Chevalier *et al.*, Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
12. J. Dou *et al.*, De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
13. C. E. Tinberg *et al.*, Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
14. M. J. Lajoie *et al.*, Designed protein logic to target cells with precise combinations of surface antigens. *Science* **369**, 1637–1643 (2020).
15. B. Kuhlman *et al.*, Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
16. E. Verschueren *et al.*, Protein design with fragment databases. *Curr. Opin. Struct. Biol.* **21**, 452–459 (2011).
17. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
18. J. U. Bowie, D. Eisenberg, An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4436–4440 (1994).
19. D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).
20. R. Pearce, Y. Zhang, Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.* **68**, 194–207 (2021).
21. P.-S. Huang *et al.*, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
22. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
23. Y. Zhang, J. Skolnick, Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7594–7599 (2004).
24. H. Deng, Y. Jia, Y. Zhang, 3DRobot: Automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **32**, 378–387 (2016).
25. I. Anishchenko *et al.*, De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
26. J. Wang *et al.*, Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
27. B. Huang *et al.*, A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523–528 (2022).
28. W. R. Taylor, A "periodic table" for protein structures. *Nature* **416**, 657–660 (2002).
29. Z. Harteveld *et al.*, A generic framework for hierarchical de novo protein design. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.04.07.487481 (Accessed 22 October 2022).
30. T. M. Jacobs *et al.*, Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
31. S. L. Guffy, F. D. Teets, M. I. Langlois, B. Kuhlman, Protocols for requirement-driven protein design in the rosetta modeling program. *J. Chem. Inf. Model* **58**, 895–901 (2018).
32. N. Koga *et al.*, Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
33. X. Pan, T. Kortemme, Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).
34. W. Zhou, T. Smidlehner, R. Jerala, Synthetic biology principles for the design of protein with novel structures and functions. *FEBS Lett.* **594**, 2199–2212 (2020).
35. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
36. D. Baker, What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* **28**, 678–683 (2019).
37. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
38. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
39. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
40. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
41. X. Huang, R. Pearce, Y. Zhang, EvoEF2: Accurate and fast energy function for computational protein design. *Bioinformatics* **36**, 1135–1142 (2020).
42. R. F. Alford *et al.*, The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
43. H. Zhou, J. Skolnick, GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **101**, 2043–2052 (2011).
44. J. Park, K. Saitou, ROTAS: A rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics* **15**, 307 (2014).
45. Y. R. Lin *et al.*, Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5478–E5485 (2015).
46. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
47. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
48. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
49. C. Zhang, W. Zheng, S. M. Mortuza, Y. Li, Y. Zhang, DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105–2112 (2020).
50. M. Baek, D. Baker, Deep learning and protein structure modeling. *Nat. Methods* **19**, 13–14 (2022).
51. V. B. Chen *et al.*, MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
52. N. Fernandez-Fuentes, B. Oliva, A. Fiser, A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res.* **34**, 2085–2097 (2006).
53. N. Fernandez-Fuentes, J. M. Dybas, A. Fiser, Structural characteristics of novel protein folds. *PLoS Comput. Biol.* **6**, e1000750 (2010).
54. S. Wu, Y. Zhang, Recognizing protein substructure similarity using segmental threading. *Structure* **18**, 858–867 (2010).
55. S. Wu, J. Skolnick, Y. Zhang, Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **5**, 17 (2007).
56. J. Yang *et al.*, The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–18 (2015).
57. D. Xu, Y. Zhang, Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* **81**, 229–239 (2013).
58. W. Zheng *et al.*, Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).
59. W. Zheng *et al.*, Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins* **89**, 1734–1751 (2021).
60. A. A. Canutescu, R. L. Dunbrack Jr., Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963–972 (2003).
61. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
62. D. Xu, Y. Zhang, Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **101**, 2525–2534 (2011).