De novo protein design by citizen scientists

Brian Koepnick, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J. Bick, Aaron Bauer, Gaohua Liu, Yojiro Ishida, Alexander Boykov, Roger D. Estep, Susan Kleinfelter, Linda Wei, Foldit Players, Gaetano T. Montelione, Zoran Popovic, Firas Khatib, Seth Cooper, David Baker*

Abstract

Online citizen science projects such as GalaxyZoo[1], Eyewire[2] and Phylo[3] have been very successful for data collection, annotation, and processing, but for the most part have harnessed human pattern recognition skills rather than human creativity. An exception is the game EteRNA[4], in which game players learn to build new RNA structures by exploring the discrete two-dimensional space of Watson-Crick base pairing possibilities. Building new proteins, however, is a more challenging task to present in a game, as both the representation and evaluation of a protein structure are intrinsically three-dimensional. We posed the challenge of *de novo* protein design in the online protein folding game Foldit[5]. Players were presented with a fully extended peptide chain and challenged to craft a folded protein structure with an amino acid sequence encoding that structure. After many iterations of player design, analysis of the top scoring solutions, and subsequent game improvement, Foldit players can now, starting from an extended polypeptide chain, generate a diversity of protein structures and sequences which encode them *in silico*. 146 Foldit player designs with sequences unrelated to naturally occurring proteins were encoded in synthetic genes; 58 were found to be expressed in *E. coli* with good solubility and to adopt stable monomeric folded structures in solution. The diversity of these structures is unprecedented in *de novo* protein design, representing 21 different folds—including a new fold that is unobserved in natural proteins. High resolution structures were determined for three of the designs, and are nearly identical to the player models. This work makes explicit the considerable implicit knowledge contributing to success in *de novo* protein design, and shows that citizen scientists can discover creative new solutions to outstanding scientific challenges, such as the protein design problem.

Main Text

The principle underlying *de novo* protein design is that proteins fold to their lowest free energy state[6]; hence, designing a new protein structure requires finding an amino acid sequence whose lowest energy state is the prescribed structure. In practice, this challenge can be divided into two subproblems: first, crafting a protein backbone that is designable (i.e. that could be the lowest energy state of some sequence); and second, finding a sequence whose lowest energy state is the crafted structure. One of the challenges of protein design is the exponentially increasing number of conformations available to a polypeptide chain, which is astronomical even for a modestly-sized protein of 60-100 residues. Thus, the first subproblem of crafting a plausible backbone is extremely open-ended, and the second subproblem is difficult because it is not tractable to explicitly check that a designed sequence has lower energy in the crafted structure than in any other structure. There has been considerable progress in *de novo* protein design in recent years[7-10], but it is unclear whether all of the contributions to this success have been made explicit in the protocols used to design proteins, and how much implicit knowledge

resides in the expertise of the designers. Disentangling the role of expert knowledge is particularly difficult for the extremely open-ended challenge posed by the first subproblem (i.e. crafting a plausible backbone), for which there are a practically unlimited number of solutions. Because full computer enumeration of backbones is not possible, there is considerable room for human creativity and intuition in generating and designing new protein structures.

To investigate how crowd-based creativity could contribute to solving the long-standing protein design problem, we incorporated de novo protein design tools into the protein folding game Foldit. Foldit is a free online computer game developed to crowdsource problems in protein modeling, and offers full control over the three-dimensional structure of a protein model[5] (Figure 1). Players compete to build a model with the lowest free energy, as calculated by the Rosetta energy function[11]. In the past, Foldit has been primarily applied to protein structure prediction problems, in which Foldit players were presented with an unstructured amino acid sequence and challenged to determine its native conformation[5,12]. Foldit players in one case redesigned a loop region of an already folded structure[13], but the *de novo* design of an entire protein is a far more expansive challenge.

We repeatedly challenged Foldit players to design stably folded proteins from scratch, and iteratively improved the game based on their results. In each challenge, players were provided with a poly-isoleucine backbone in a fully extended conformation (60-100 residues in length), and were given seven days to fold the backbone into a compact structure and identify a sequence specifying this backbone. Initially, most top-scoring (low energy) Foldit player designs were highly extended, lacked a solvent-inaccessible core, and were composed entirely of polar residues (Figure 1F). Such extended, fully α-helical structures have more favorable hydrogen bonding, electrostatic, and local torsional energies than collapsed structures, which must contort to create a buried core. While poly-lysine and other extended polar sequences resembling these initial Foldit solutions are often α-helical in solution[15,16], the lack of long-range interactions precludes specific folding into a single stable structure[17]. This highlights a limitation of using absolute energy as an optimization criterion for protein design: a low energy design does not guarantee structural specificity, which arises only if all other alternative conformations have higher energy. To favor the design of globular solvent-excluding protein folds, with sequences that uniquely encode them, we introduced three supplementary design rules into Foldit: a “Core Exists” rule that requires a minimum proportion of residues (e.g. 30%) to be solvent-inaccessible in the designed structure; a “Secondary Structure Design” rule that prohibits glycine and alanine in all secondary structure elements; and a “Residue Interaction Energy” rule to penalize large residues that make insufficient intramolecular interactions in the designed structure. With the addition of these rules to Foldit, subsequent top-scoring designs from Foldit players were compact globular proteins.

We obtained custom synthetic genes encoding 12 player designs for which structure prediction calculations converged on the player designed conformation[14]. The sequences of these proteins have no homology to any known protein (Sup. Table 1). The *de novo* designs were expressed in *E. coli* and purified by metal affinity and size exclusion chromatography. Chromatography and circular dichroism (CD) spectroscopy indicated that 7 of the 12 designs were monomeric and

folded in solution, with helical secondary structure consistent with the players' models (Sup. Fig. 1). All of the experimentally tested proteins described in this paper, are entirely the work of Foldit players.

During gameplay, the Foldit application uploads the player's latest model to the Foldit server every 2-5 minutes; from these snapshots we can reconstruct the process by which a Foldit player develops a protein design (Figure 2). Foldit players employ more varied and complex exploration strategies than standard Rosetta automated design protocols, and frequently revert to a previous iteration of their model to explore an alternative path, resulting in a highly-branched search tree. A typical automated design protocol, by contrast, includes only two branch points[18]. In addition, Foldit players regularly sample much higher energy states than the automated protocol (Figure 2), which has only a limited ability to escape local energy minima.

Encouraged by the success of Foldit players in designing stable proteins from scratch, we made additions to the game encouraging players to explore more diverse protein structures. Up until this point, all top-scoring Foldit designs had consisted of either three or four α-helices connected by minimal loops. Indeed, Foldit players had determined that designs with β-sheets did not score on par with α-helical bundles, and competitive players had abandoned any attempt to design more varied folds. (This has an interesting parallel to protein design by practicing scientists, which has also focused much more on helical bundles than other classes of protein folds[19-22].) To encourage the design of a wider variety of folds, we introduced a "Secondary Structure" rule stipulating that no more than 50% of residues may form α-helices. Foldit players responded by designing a multitude of mixed α/β proteins, which were indistinguishable from expert designs upon visual inspection. However, structure prediction calculations for these α/β design sequences showed poor sampling close to the target design structure, suggesting that the designed sequences did not strongly encode their local structure[14]. Further analysis showed that these player designs contained many residues with locally strained backbone conformations (backbone phi and psi torsions in unfavored regions of the Ramachandran plot[23,24]). That such designs had very low energies revealed a problem in the Rosetta energy function at the time: since Rosetta users typically sampled backbones starting from fragments of native proteins, unfavorable local conformations were rarely encountered—hence it had not been discovered that the energies associated with local backbone strain were being underestimated. We addressed this flaw in the Rosetta model by increasing the steepness of the energetic penalties associated with strained local backbone geometry; this is now standard in the latest Rosetta energy function[11]. We also added to Foldit an "Ideal Loops" rule restricting players to a set of 19 unstrained reverse-turn conformations[7], and incorporated new tools to aid generation of unstrained backbones: a fragment lookup-based loop-closure tool, an interactive Ramachandran map, and a protein Blueprint scheme for drag-and-drop assembly of secondary structure elements and common loop conformations. Together, these upgrades brought about a marked improvement in the local backbone quality of Foldit player-designed proteins (Sup. Fig. 3).

The importance of reducing local backbone strain was borne out in experimental characterization. Prior to the backbone modeling improvements described in the previous

paragraph, only 5 of 37 Foldit α/β designs tested (14%) were monomeric and structured in solution. Following the backbone modeling additions, 46 of 97 (47%) were monomeric and exhibited the expected secondary structure in solution. Most showed exceptional stability in thermal and chemical denaturation experiments, with free energies of unfolding (ΔG_{unf}) up to >15 kcal/mol; indeed, 32 designed proteins remained completely folded at 95ºC (Figure 3; Extended Data Fig. 1). This success rate surpasses previous reports of designed α/β proteins[7,12].

Overall, the 58 successful Foldit designs are diverse in structure, representing 21 different protein folds (Figure 3; Extended Data Fig. 2)—one of which is a new fold previously unobserved in natural proteins. The success of Foldit designs is not attributed to just one or two exceptional Foldit players, but is shared broadly by the Foldit community. The 58 successful designs were created by 37 different Foldit players (the most prolific player authored 11 successful designs); 19 designs were created collaboratively by at least two cooperating players, and 5 designs were not top-scoring, but regardless were flagged by players as personal favorites.

We succeeded in solving structures of three Foldit player-designed proteins. X-ray crystal structures of two designed proteins, named by their designers as foldit1 and Peak6, closely match the designed conformations, with Cα-RMSD of 1.1 and 0.9 Å, respectively (Figure 4). Well-resolved electron density in the protein core shows that most sidechains adopt the intended rotamers and preserve the designed packing interactions. The solution NMR structure of a third design, foldit3, also closely matches the design conformation, with a with a Cα-RMSD of 1.1 Å between the design model and a representative structure (i.e., the medoid conformer[25]) of the ensemble.

We can draw several general conclusions about scientific models, citizen science, and the interplay between the two. First, a scientific model which holds within the domain space considered by practicing scientists may not hold outside of this domain. This is most vividly illustrated by the highly extended structures generated by Foldit players in their first *de novo* design efforts, and later by the structures with strained local geometry not previously sampled by Rosetta users. Second, for citizen scientists to make essential and creative scientific contributions through online gaming, the scoring function of the game must be an accurate representation of the science. In our initial iterations, Foldit did not present to players a sufficiently accurate and general model to allow them to robustly design new proteins, even though the underlying Rosetta software had been used for protein design by practicing scientists. Third and most important, citizen science offers a powerful way to systematically improve a scientific model, through iterations of model trial and model improvement. Human game players are exceptionally capable at finding and exploiting unanticipated solutions that are otherwise unexplored by experienced scientists, whose focus is not on getting a high score, but rather on solving their specific scientific problem.

We have demonstrated that non-expert citizen scientists, playing the online computer game Foldit, can accurately design completely new protein structures from scratch. Locally, players'

solutions are physically plausible and resemble natural proteins, but globally, they are creative and diverse. Proteins designed by citizen-scientist Foldit players are by no measure inferior to those of expert protein designers: they fold accurately to the intended conformation, show exceptional folding stability, and span a wide diversity of structures. This result is all the more impressive given that *de novo* protein design was an almost completely unsolved problem just a few years ago, and the diversity in protein folds spanned by the successful Foldit players' models considerably exceeds that in any previous protein design report. The sustained success of Foldit players over a wide diversity of protein folds highlights the power of human creativity when guided by scientific understanding presented in a readily comprehensible form.

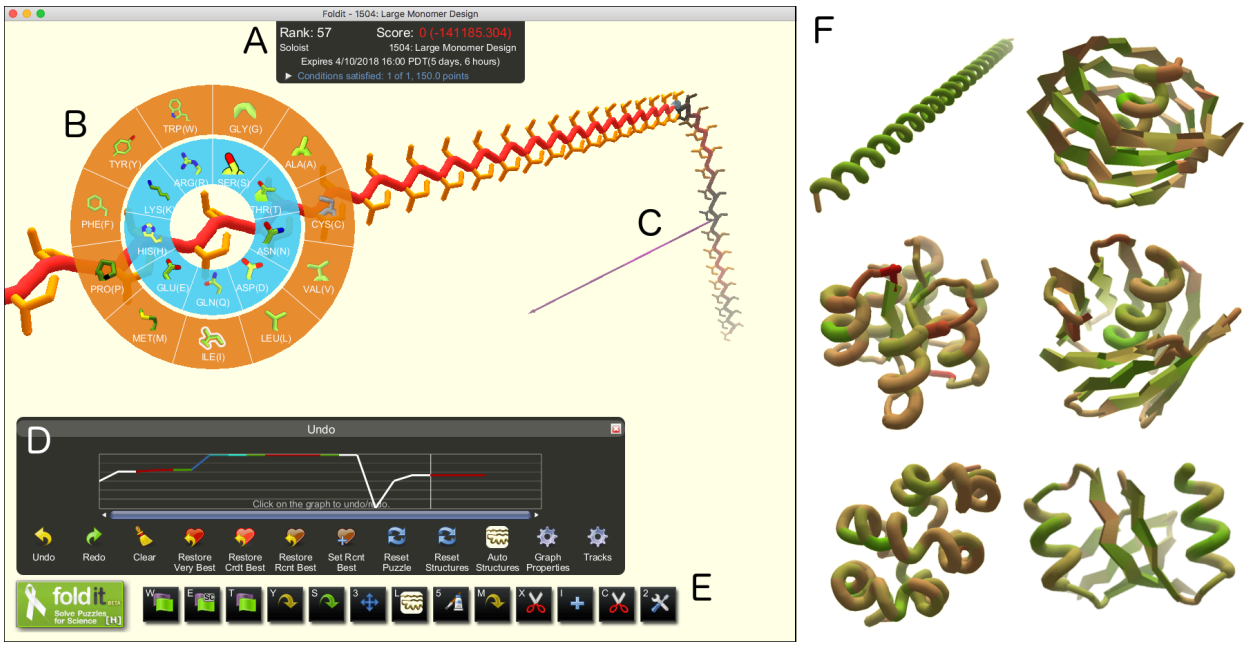


Figure 1. The Foldit user interface. (A) The Foldit score is the Rosetta energy with a negative multiplier, so that better models yield higher scores. (B) The design palette allows players to change the amino-acid residue identity at any position of the model. (C) The Pull tool allows players to manipulate the three-dimensional structure of the model. (D) The Undo graph tracks the score as a model is developed, and allows players to backtrack and load previous versions of a model. (E) Additional Foldit tools (from left to right): full structure minimization, sidechain minimization, backbone minimization, auto-design sidechains, repack sidechains, translate/rotate model, secondary structure assignment, idealize secondary structure, manually design sidechains, delete residues, insert residues, insert cutpoint, idealize peptide bond geometry. (F) Foldit players explore diverse structures that have no sequence or structural homology to natural proteins.

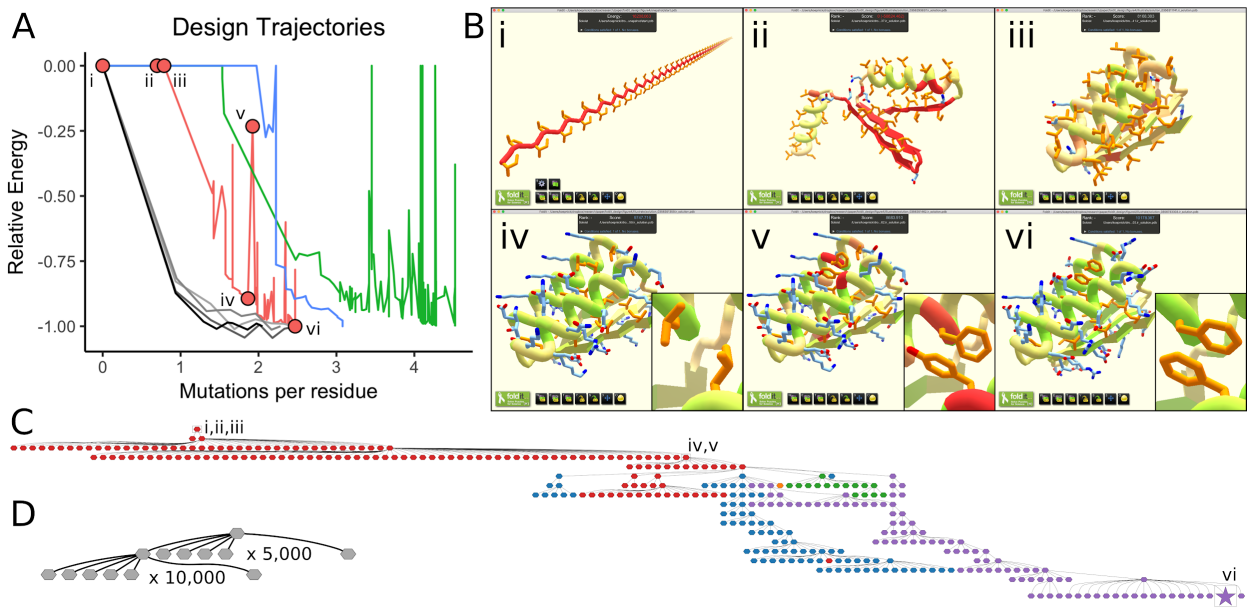


Figure 2. Comparison of Foldit player and automated design sampling strategies. (A) Single trajectories (ignoring abandoned branches) for three Foldit player-designed proteins in red (foldit1), blue (Peak6), and green (2003169_S953); and design trajectories for four Rosetta-designed proteins in gray. The y-axis is the Rosetta energy rescaled so that the final design has a value of -1.00, and positive energies are shown as zero. Foldit players are willing to undergo large increases in energy to explore new regions; the Rosetta protocol in contrast has a limited ability to escape local energy minima. Red circles correspond to structures shown in (B). (B) Snapshots from the design trajectory of foldit1: (i) the initial extended chain of poly-isoleucine; (ii) development of secondary structure; (iii) development of folded tertiary structure; (iv) sequence design of folded structure, with inset showing favorable packing between two Leu sidechains at positions 13 and 45; (v) high-energy intermediate design, with inset showing redesign at positions 13 and 45, which results in steric clashes with the protein backbone; (vi) the final refined design, with inset showing favorable interactions between two Phe sidechains at positions 13 and 45. (C) The design strategy for foldit1 represented as a graph, showing all branch points where multiple design trajectories were spawned from a single intermediate. The final design was reached only after 17 branch points. Node colors correspond to five different cooperating Foldit players, and the final design is marked as a star. (D) Similar representation of a Rosetta design trajectory; there are only two branch points.

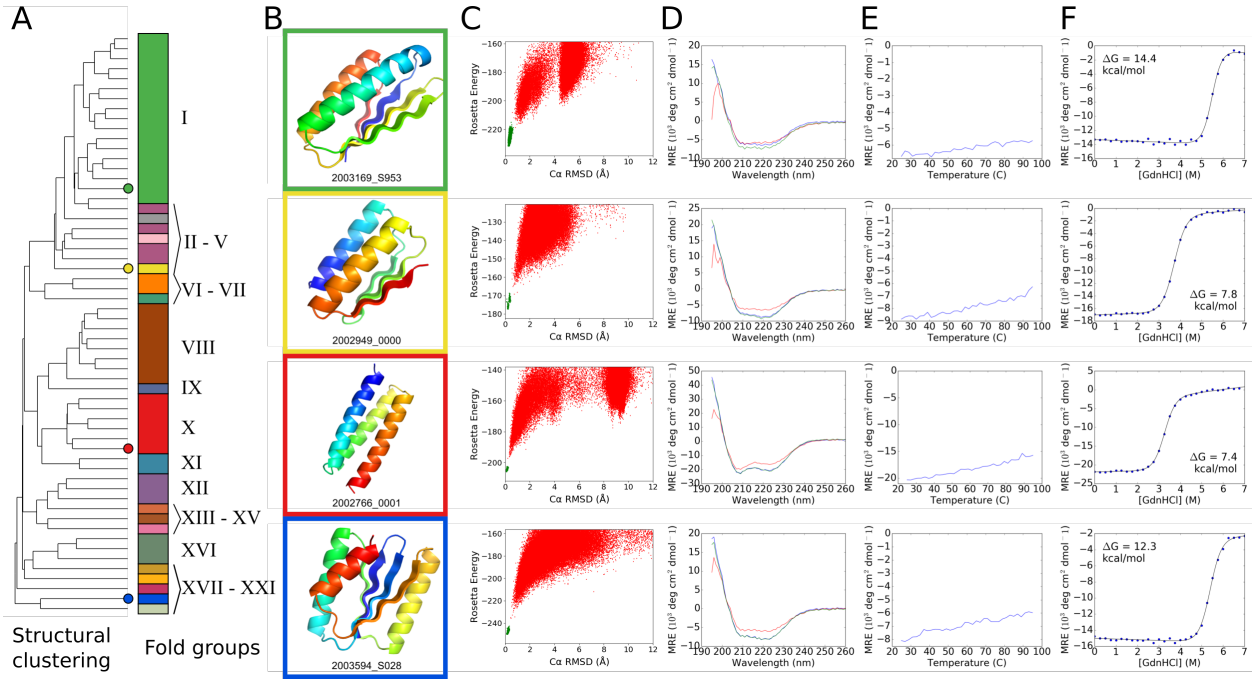


Figure 3. Structural characterization of Foldit player designed proteins. (A) Dendrogram showing all 58 folded Foldit player designs clustered by structural similarity (TM-align[26]), with colored circles highlighting the four designs characterized in (B-F). The stacked bars show the 21 different folds among the clustered designs (Extended Data Fig. 2). Fold XX (see design 2003594_S028) is a new fold, previously unobserved in natural proteins. (B) Cartoon depiction of four select Foldit designs. (C) Rosetta@home *ab initio* calculations show that the sequence for each design has an energy landscape that is strongly funneled toward the design structure. Rosetta energy is on the y-axis and Cα-RMSD to the designed structure on the x-axis; points represent lowest energy structures sampled starting from an extended chain (red points), and starting from the FoldIt design model (green points). (D) Circular dichroism (CD) spectra indicate that the designs adopt the expected secondary structure content in solution at 25ºC (blue trace), when heated to 95ºC (red trace), and when cooled again to 25ºC (green trace). (E) CD mean residue ellipticity at 220 nm as temperature is increased from 25ºC to 95ºC; the designs do not temperature denature. (F) Cooperative unfolding during titration with guanidinium hydrochloride. Blue circles show CD mean residue ellipticity at 220 nm with increasing concentration of denaturant, and the black curve shows a two-state unfolding model fit to the data. ΔG_{unf} values were determined by linear extrapolation using the fit model parameters[27].

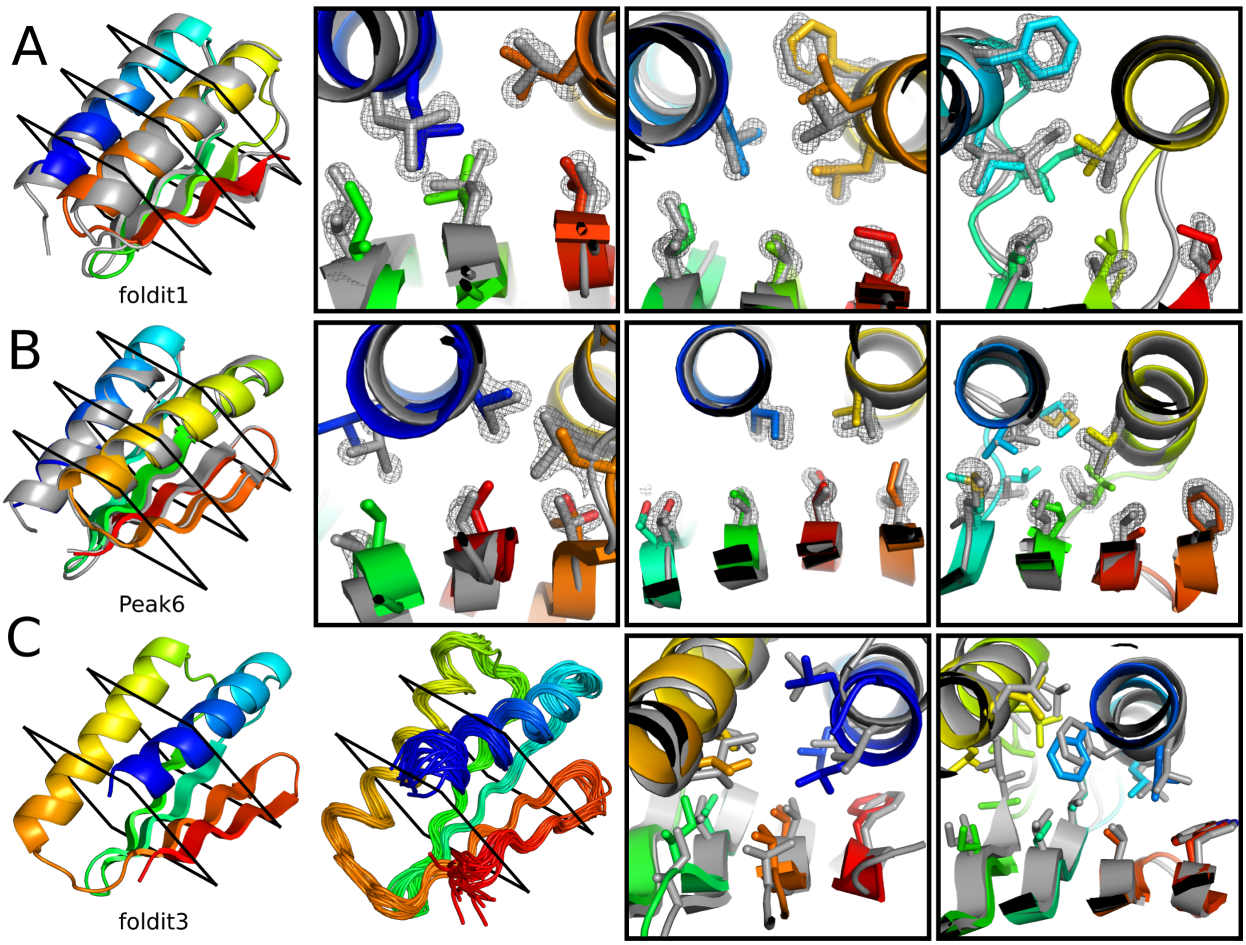


Figure 4. High-resolution structures of Foldit player designed proteins. (A) The foldit1 design (fold V in Fig 3: 3 β-strands with sheet order 1-2-3) model backbone (rainbow) aligns to the crystal structure (gray) with Cα-RMSD of 1.1 Å. (B) The Peak6 design (fold III: 4 strands, sheet order 1-2-4-3) model backbone (rainbow) aligns to the crystal structure (gray) with Cα-RMSD of 0.9 Å. Cross-sections show core residue sidechains, with the composite omit 2mFo-DFc map contoured at 2.0 σ. (C) The foldit3 design model (fold XVIII: 4 strands, sheet order 2-1-3-4) and NMR ensemble. The design model aligns to the representative (medoid) NMR model with a Cα-RMSD of 1.1 Å. Cross sections compare core side chains in the design model (rainbow) and a representative NMR model (gray).

Methods

Foldit protein design puzzles

Foldit puzzles were set up with a model poly-isoleucine in fully extended conformation, with fixed length ranging from 60 to 100 residues. Each puzzle was posted online for seven days, during which Foldit players competed to develop a protein model with the lowest energy, as calculated by the Rosetta energy function. Foldit puzzles used the talaris2013_cart scorefunction with the following modifications: (1) the cart_bonded scoreterm was upweighted (increased from 0.5 to 2.0) to ensure realistic bond lengths and angles as players cut and splice the backbone chain; (2) a penalty-only envsmooth scoreterm (weighted at 2.0) was added to supplement the Rosetta solvation treatment, and to discourage the design of buried polar and exposed nonpolar residues; (3) the reference energy of alanine was modified (increased to 3.0) to discourage the excessive design of alanine. See Supplementary Information for configuration files for all Foldit puzzles. Each Foldit puzzle was accompanied by a brief description, along with an explanation of any supplementary rules enforced in the puzzle. Design puzzles were accessible to all Foldit users; Foldit user registration is free and open to the public, at http://fold.it. Models were collected continuously as Foldit players worked on the puzzles, since the Foldit application automatically uploads the user's latest model to a server every 2-5 minutes. This study was approved by the University of Washington Institutional Review Board, and informed consent for research participation was obtained from all Foldit users at the time of user registration.

Protein design selection

After the end of each puzzle, we selected player models for further analysis as follows: First we selected the lowest-energy model from each of the 10 top-ranked groups, where independent players were treated as individual groups (designs named with suffix "0000-9"). Second, we selected the lowest-energy model from the 10 top-ranked solo players, which includes independent players as well as group members that developed a model without assistance from their group (suffix "s000-9"). Third, we visually inspected models that were flagged by Foldit players for special consideration, and selected any models that appeared plausible (suffix "S***"). Last, we ranked and pruned the set of remaining models, by removing any models that align to a better-scoring model with Cα-RMSD less than 2.5 Å. We visually inspected the 50 top-ranked models in the pruned set and selected any models that appeared plausible (suffix "1001-50"). Models deemed "implausible" typically lacked secondary structure, contained buried polar residues, or included long stretches of completely polar residues. The sequences of selected models were subjected to Rosetta *ab initio* structure prediction[14], using the distributed computing platform Rosetta@home. If *ab initio* predictions identified any decoy structures with energy comparable to (or lower than) the designed structure, or if *ab initio* predictions were unable to sample the designed structure, the design was rejected. All other designs were selected for experimental characterization. The majority of experimentally tested designs (96 of 146) were top-ranked group or solo designs, which were selected "blindly" (without visual inspection). See supplementary data for structures and FASTA sequences of all tested designs.

Protein expression and purification

A 6x-His tag with TEV-cleavable linker (sequence 'MGHHHHHHGWSENLYFQGS') was prepended to the N-terminus of each design selected for experimental characterization. Plasmids containing the encoded genes were ordered from Genscript in pET15 (designs with prefix between 997258 and 1998925), or in pET21 (1998555-2002990), or from Twist in pET28 (2003048-2003594) vectors. Plasmids were transformed into *E. coli* BL21 Star (DE3) cells (Invitrogen), and grown overnight in 4 mL Luria-Bertani medium (LB) with 50 µg/mL carbenicillin (for pET15, pET21 vectors) or 30 µg/mL kanamycin (for pET29). Overnight cultures were used to inoculate 0.5 L auto-induction media, and grown at 37 ºC for 18 hours. Cultures were pelleted and resuspended in 25 mL lysis buffer (20 mM Tris pH 8.0, 300 mM NaCl, 1 mg/mL lysozyme, 0.1 mg/mL DNase, 1 mM PMSF), and lysed by microfluidization. The cell lysate was pelleted and supernatant was filtered with a 0.22 µm filter before loading onto a 2 mL nickel affinity gravity column. Protein bound to the column was washed with 20 mL wash buffer (20 mM Tris pH 8.0, 500 mM NaCl, 30 mM imidazole) and eluted in 10 mL elution buffer (20 mM Tris pH 8.0, 500 mM NaCl, 250 mM imidazole). Purified protein was dialyzed into TBS (20 mM Tris pH 8.0, 300 mM NaCl) at 4ºC overnight to remove imidazole and further purified by gel filtration on an AKTAxpress (GE Healthcare) with a Superdex S75 10/300 GL column (GE Healthcare). For proteins containing cysteine, dialysis and gel filtration were carried out in TBS with 1 mM TCEP. Protein expression and solubility was confirmed by SDS-PAGE of samples from lysate pellet and supernatant. All purified proteins were verified by mass spectrometry.

Circular dichroism

Purified protein was dialyzed into 50 mM sodium phosphate pH 7.4 at 4ºC overnight (plus 500 µM TCEP for proteins containing cysteine). All circular dichroism data were collected on an AVIV Model 420 spectrometer. Far UV spectra and temperature melts were measured with 11-62 µM protein in a quartz cuvette with path length of 1 mm. Protein concentration was determined by absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Scientific), using predicted extinction coefficients. Wavelength spectra were measured between 195 and 260 nm at 25ºC, 95ºC, and again after cooling to 25ºC. For temperature melts, ellipticity at 220 nm was monitored as temperature increased from 25ºC to 95ºC, in increments of 2ºC. Chemical titrations were carried out with 1.0-21 µM protein in a quartz cuvette with path length of 10 mm. Ellipticity at 220 nm was monitored at concentrations of guanidinium chloride increasing from 0 to 7 M, in increments of 0.25 M. Denaturation curves were fitted with non-linear regression to two-state unfolding model with six parameters: the folding free energy, m-value, and slope and y-intercept for baseline curves[28].

X-ray crystallography

Prior to x-ray crystallography, the N-terminal 6x-His tag was cleaved from protein samples by incubation with 250 µg TEV protease at 25ºC for four hours in 20 mM Tris pH 8.0, 300 mM NaCl, 1 mM DTT. The reaction product was dialyzed into TBS overnight at 4ºC to remove DTT and flowed over a 2 mL metal affinity gravity column to remove TEV protease and residual histidine tag. The cleaved protein was further purified by gel filtration as described above. Purified protein was concentrated to 20-100 mg/mL in 20 mM Tris pH 8.0, 300 mM NaCl. Crystallization screening was carried out with a variety of 96-condition spare matrix suites available from Qiagen or Hampton Research. A Mosquito Crystal nanoliter robot (TTP Labtech)

was used to prepare screens in 3-well sitting drop plates, with 200 nL drops and protein:precipitant ratios of 1:1, 1:2, and 2:1.

foldit1 was crystallized at 20 mg/mL in 50 mM HEPES pH 7.5, 0.2 M potassium chloride, 35% v/v pentaerythritol propoxylate. Crystals were flash-frozen in liquid nitrogen without further cryo-protection.

Peak6 was crystallized at 40 mg/mL in 0.1 M sodium acetate pH 4.5, 0.2 M lithium sulfate, 50% w/v PEG 400. Crystals were briefly soaked in mother liquor plus 20% PEG 200, then flash frozen in liquid nitrogen.

X-ray diffraction datasets were collected at the Advanced Light Source (Berkeley, CA). Data was processed with HKL2000[31]. Crystal structures were solved by molecular replacement with Phaser[29], using the backbone of the original designed model with sidechains truncated to the beta carbon. Models were built and refined in iterative cycles using Coot and PHENIX[30,31]. Diffraction data and refinement statistics are listed in Extended Data Table 2.

NMR spectroscopy

NMR studies were performed using uniformly ^{15}N,^{13}C-enriched protein samples. Synthetic genes were obtained from Genscript already incorporated into plasmid pET15TEV_NESG, which includes a N-terminal 6xHis purification tag, followed by a TEV protease cleavage site (sequence 'MGHHHHHHGWSENLYFQGS'). *E. coli* BL21(DE3) cells harboring plasmid pET15TEV_NESG-foldit3 were grown in 1L MJ9 minimal media[32], supplemented with 100 µg/ml ampicillin at 37 ºC. In order to produce uniformly ^{15}N and ^{13}C enriched protein samples, 1g / L $^{15}NH_4$-salts and 2g / L U-^{13}C glucose were added as sole a nitrogen and a carbon sources, respectively. When $O.D._{600}$ reached around 0.5 units, the culture was transferred to 18 ºC, and the protein production was induced by addition of 1 mM IPTG. After overnight incubation, the cells were collected and resuspended in 20 ml binding buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl and 20 mM imidazole). After passing the cells through 900-1000 psi French press twice, cell debris were removed by 10,000 rpm for 30 min. The supernatant was further spun down at 40,000 rpm for 1hr. The obtained supernatant (soluble fraction) was mixed with 1 ml of Ni-resin and incubated at 4 ºC for 1 hr. The non-specific binding proteins were removed by 20 mL binding buffer and washing buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl and 50 mM imidazole) and the target protein was eluted by 5 mL elution buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl and 300 mM imidazole). The protein was dialyzed against GF buffer (20 mM Tris-HCl pH 8.0, 100 mM NaCl) for overnight and gel filtration was carried out using AKTA express with high-load 26/600 Superdex 200 pg column. Homogeneity (> 97%) was validated by SDS polyacrylamide gel electrophoresis. The purified protein was dialyzed against 20 mM potassium phosphate (pH 6.5), and the protein concentration was adjusted to between 0.3-0.4 mM for NMR studies.

All NMR spectra were recorded at 25 ºC using cryogenic NMR probes. All NMR data were collected on the Bruker AVANCE III 600 MHz spectrometers and processed using the program NMRPipe[33], and analyzed using the programs SPARKY and XEASY[34]. Spectra were referenced to external DSS. Sequence-specific resonance assignments were determined using AutoAssign

software together with interactive manual analysis, as described previously[35]. Backbone dihedral angle constraints were derived from the chemical shifts using the program TALOS_N[36] for residues located in well-defined secondary structure elements. The programs ASDP[37] and CYANA[38,39] were used to automatically assign NOEs and to calculate structures. RPF analysis[37,40] was used in parallel to guide iterative cycles of noise/artifact peak removal, peak picking, and NOESY peak assignments. The 20 conformers with the lowest target CYANA function value were then refined in explicit water[41] using the program CNS[42]. The structural statistics and global structure quality factors (Extended Data Table 3) including Verify3D[43], ProsaII[44], PROCHECK[45], and MolProbity[46] raw and statistical Z-scores were computed using the PSVS[47] 1.5 and PDBStat[48] software packages. The global goodness-of-fit of the final structure ensembles with the NOESY peak list data, the NMR DP score, was determined using the RPF analysis program[40].

Code Availability

Because Foldit crowdsourcing relies on regulated, fair competition between participants, the source code of the Foldit user interface is not open. The underlying Rosetta macromolecular modeling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users, and commercial licenses are available via the University of Washington CoMotion Express License Program. Analysis scripts used in this paper are available in the Supplementary Information.

Data Availability

The atomic coordinates of foldit1 and Peak6 crystal structures, and the foldit3 NMR structure, have been deposited in the RCSB Protein Database with accession numbers 6MRR, 6MRS, and 6MSP, respectively. Chemical shift and NOESY peak list data for foldit3 were deposited in the Biological Magnetic Resonance Bank (BMRB ID 30527).

References

1. Lintott, C. J. *et al.* Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* **389,** 1179–1189 (2008).
2. Kim, J. S. *et al.* Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509,** 331–336 (2014).
3. Kawrykow, A. *et al.* Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS ONE* **7,** (2012).
4. Lee, J. *et al.* RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* **111,** 2122–2127 (2014).
5. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466,** 756–760 (2010).
6. Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. The Genetic Control of Tertiary Protein Structure: Studies With Model Systems. *Cold Spring Harbor Symposia on Quantitative Biology* **28,** 439–449 (1963).
7. Lin, Y.-R. *et al.* Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences* (2015).

8. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537,** 320–327 (2016).
9. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355,** 201–206 (2017).
10. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* (2018).
11. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13,** 3031–3048 (2017).
12. Khatib, F. *et al.* Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol* **18,** 1175–1177 (2011).
13. Eiben, C. B. *et al.* Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology* **30,** 190–192 (2012).
14. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Meth. Enzymol.* **383,** 66–93 (2004).
15. Blout, E. R. & Idelson, M. Compositional Effects on the Configuration of Water-soluble Polypeptide Copolymers of L-Glutamic Acid and L-Lysine. *Journal of the American Chemical Society* **80,** 4909–4913 (1958).
16. Doty, P., Imahori, K. & Klemperer, E. The solution properties and configurations of a polyampholytic polypeptide: copoly-L-lysine-L-glutamic acid. *Proceedings of the National Academy of Sciences* **44,** 424–431 (1958).
17. Ghosh, K. & Dill, K. A. Theory for Protein Folding Cooperativity: Helix Bundles. *J. Am. Chem. Soc.* **131,** 2306–2312 (2009).
18. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2013).
19. Regan, L. & DeGrado, W. Characterization of a helical protein designed from first principles. *Science* **241,** 976–978 (1988).
20. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **282,** 1462–1467 (1998).
21. Thomson, A. R. *et al.* Computational design of water-soluble alpha-helical barrels. *Science* **346,** 485–488 (2014).
22. Jacobs, T. M. *et al.* Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352,** 687–690 (2016).
23. Ramachandran, G. N. & Sasisekharan, V. Conformation of Polypeptides and Proteins. *Advances in Protein Chemistry* **23,** 283–437 (1968).
24. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 12–21 (2010).
25. Montelione, G. T. *et al.* Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21,** 1563–1570 (2013).
26. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33,** 2302–2309 (2005).
27. Santoro, M. M. & Bolen, D. W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry* **27,** 8063–8068 (1988).
28. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* **276,** 307–326 (1997).

29. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40,** 658–674 (2007).
30. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).
31. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68,** 352–367 (2012).
32. Jansson M, Li YC, Jendeberg L, Anderson S, Montelione GT, Nilsson B. High-level production of uniformly ¹□N- and ¹³C-enriched fusion proteins in Escherichia coli. *J Biomol NMR* **7**, 131-141 (1996).
33. Delaglio, F. et al. Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *J Biomol NMR* **6**, 277-293 (1995).
34. Bartels, C., Xia, T. H., Billeter, M., Guntert, P. & Wuthrich, K. The Program Xeasy for Computer-Supported Nmr Spectral-Analysis of Biological Macromolecules. *J Biomol NMR* **6**, 1-10 (1995).
35. Liu, G. H.et al. NMR data collection and analysis protocol for high-throughput protein structure determination. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10487-10492 (2005).
36. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* **44**, 213-223, doi:10.1007/s10858-009-9333-z (2009).
37. Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **62**, 587-603, doi:10.1002/prot.20820 (2006).
38. Guntert, P., Mumenthaler, C. & Wuthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology* **273**, 283-298 (1997).
39. Herrmann, T., Guntert, P. & Wuthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology* **319**, 209-227 (2002).
40. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society* **127**, 1665-1674 (2005).
41. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins* **50**, 496-506, doi:10.1002/prot.10299 (2003).
42. Brunger, A. T.et al.Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D-Biological Crystallography* **54**, 905-921 (1998).
43. Luthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature **356**, 83-85, doi:10.1038/356083a0 (1992).
44. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362, doi:10.1002/prot.340170404 (1993).

45. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography* **26**, 283-291 (1993).
46. Word, J. M., Bateman, R. C., Presley, B. K., Lovell, S. C. & Richardson, D. C. Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Science* **9**, 2251-2259 (2000).
47. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778-795, doi:10.1002/prot.21165 (2007).
48. Tejero, R., Snyder, D., Mao, B., Aramini, J.M., Montelione, G.T. PDBStat: A universal restraint converter and restraint analysis software package for protein NMR. *J. Biomol. NMR* **56**, 337-351 (2013).

Acknowledgements

We thank all Foldit players for their gameplay contributions, and for feedback offered on the https://fold.it website. We thank A. Kang, S.A. Rettie, C. Chow, and L. Carter for help with experiments; D. Alonso, L. Goldschmidt, P. Vecchiato, D. Kim for computer support; and Rosetta@home (https://boinc.bakerlab.org) volunteers for computing resources. We thank G. Rocklin, V. Mulligan, and other members of the Baker lab for discussions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082, and National Institutes of Health Grant 1R01 GM120574 (to G.T.M.). The ALS-ENABLE beamlines are supported in part by the National Institutes of Health, National Institute of General Medical Sciences, grant P30 GM124169-01. The Advanced Light Source is a DOE User Facility under Contract No. DE-AC02-05CH11231. This material is based upon work supported by the National Science Foundation under grant no. 1629879. This work was supported by National Institutes of Health grant 1UH2CA203780.

Author contributions

B.K., Z.P., F.K., S.C., and D.B. designed the study.
B.K., J.F., T.H., A.F., D.A.S., and S.C. developed Foldit software tools.
A. Boykov, R.D.E., S.K., L.W., and Foldit Players designed all proteins.
B.K., F.K., A.F., and A. Bauer analyzed Foldit player designs.
B.K. performed biophysical characterization.
B.K. and M.J.B. determined crystal structures.
G.L., Y.I., and G.T.M. determined the NMR structure.
B.K. and D.B. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests. G.T.M. is a co-founder of Nexomics Biosciences, Inc.

Extended Data Table 1. Foldit player-designed proteins selected for experimental testing

		Nearest Sequence Homolog			Nearest Structural Homolog		Experimental Characterization			
Design ID	Designers	BLAST Score	BLAST E-value	BLAST Hit	TM-align Score	TM-align Hit	Expressed	Soluble	Monomeric	Structured
997258_0001	PLAYER_2,MaartenDesnouck,MurloW	-	-	-	0.806	5cwoA	Yes	Yes	Yes	No
997258_0004	**Timo van der Laan**	**-**	**-**	**-**	**0.814**	**5cwpA**	**Yes**	**Yes**	**Yes**	**Yes**
997383_S346	frood66,PLAYER_16,PLAYER_10	35.4	5.5	KFV13184.1	0.685	3vf0A	Yes	Yes	No	-
997523_0000	PLAYER_13,MurloW	-	-	-	0.877	4tqlA	Yes	Yes	No	-
997523_0003	**caglar**	**48.9**	**2.00E-04**	**XP_015806396.1**	**0.844**	**4e40A**	**Yes**	**Yes**	**Yes**	**Yes**
997523_0005	vakobo,Grom,PLAYER_14	38.9	0.67	XP_015806396.1	0.811	5j0lE	Yes	Yes	No	-
997523_0006	Timo van der Laan	40	0.25	GBE60965.1	0.819	4tqlA	Yes	Yes	No	-
997523_0008	**eikem**	**35.4**	**8.7**	**WP_044163677.1**	**0.782**	**5xqiA**	**Yes**	**Yes**	**Yes**	**Yes**
997523_1003	**nemo7731**	**43.5**	**0.014**	**WP_055520104.1**	**0.740**	**5j0lA**	**Yes**	**Yes**	**Yes**	**Yes**
997523_1040	**PLAYER_6**	**-**	**-**	**-**	**0.811**	**5k7vA**	**Yes**	**Yes**	**Yes**	**Yes**
997791_1027	**PLAYER_6**	**-**	**-**	**-**	**0.820**	**2ojqA**	**Yes**	**Yes**	**Yes**	**Yes**
997915_0000	**Galaxie,PLAYER_2,karstenw**	**36.2**	**2.4**	**YP_009282838.1**	**0.813**	**5k7vA**	**Yes**	**Yes**	**Yes**	**Yes**
1998469_0000	PLAYER_15,retiredmichael	-	-	-	0.686	4p2fA	No	-	-	-
1998520_0002	Galaxie,PLAYER_23	35.8	2.8	XP_011658777.1	0.703	3w40A	Yes	Yes	No	-
1998555_1041	frood66	35	7.4	WP_102588688.1	0.631	2kptA	Yes	No	-	-
1998925_s005	PLAYER_9	-	-	-	0.642	3dyjB	No	-	-	-
1998925_s008	**PLAYER_18**	**36.6**	**1.4**	**WP_013842421.1**	**0.789**	**4fsxA**	**Yes**	**Yes**	**Yes**	**Yes**
2000240_0002	retiredmichael,PLAYER_15,LociOiling,smilingone,PLAYER_21	37.4	0.78	WP_047150276.1	0.719	1vdwA	No	-	-	-
2000240_s003	spvincent	33.9	5.8	XP_003761380.1	0.600	4x00A	Yes	Yes	No	-
2000485_1070	spvincent	-	-	-	0.595	4dlqA	No	-	-	-
2000518_0000	MurloW,PLAYER_9	-	-	-	0.634	4xevD	No	-	-	-
2000518_0003	PLAYER_22	-	-	-	0.610	m572A	Yes	No	-	-
2000518_S468	eusair	-	-	-	0.593	4g0hA	No	-	-	-
2000665_1003	mbinfield,Bruno Kestemont	-	-	-	0.569	4acjA	Yes	Yes	No	-
2001044_0000	**MurloW**	**35.8**	**2.2**	**XP_018018081.1**	**0.746**	**3d6kA**	**Yes**	**Yes**	**Yes**	**Yes**
2002089_0001	Galaxie,Susume	35	5.7	WP_117139598.1	0.705	4p1xF	No	-	-	-
2002089_1029	Galaxie,Susume	-	-	-	0.686	4ok4A	Yes	Yes	No	-
2002243_1016	**Susume**	**36.2**	**2.1**	**PKR98267.1**	**0.656**	**2qzgA**	**Yes**	**Yes**	**Yes**	**Yes**
2002290_S122	fiendish_ghoul	79.7†	1.00E-17†	XP_001631727.1†	0.914†	2pigB†	No	-	-	-
2002308_0000	**Mark-,PLAYER_3,Bletchley Park,PLAYER_1**	**34.3**	**4.2**	**KKU76692.1**	**0.654**	**3rkoC**	**Yes**	**Yes**	**Yes**	**Yes**
2002308_0005	frood66,actiasluna,Mike Lewis	36.2	2.6	XP_022288518.1	0.622	2kt9A	Yes	No	-	-
2002308_S695	Susume	38.9	0.29	XP_012556696.1	0.689	3pg5A	Yes	No	-	-
2002334_0005	fiendish_ghoul	37.4	1	XP_016366754.1	0.637	5f1cB	Yes	Yes	No	-
2002376_0000	Mark-,Bletchley Park	36.2	4.3	CCX16261.1	0.683	1fsaA	No	-	-	-
2002376_0003	PLAYER_21,LociOiling	45.4	0.002	XP_013073713.1	0.650	4he8G	Yes	Yes	Yes	No
2002469_0001	Bruno Kestemont,gloverd,Scopper	37	1.8	PIW76571.1	0.657	1xioA	No	-	-	-
2002469_S848	fiendish_ghoul	45.1	0.003	WP_075688920.1	0.695	5kilA	No	-	-	-
2002486_0006	fiendish_ghoul	-	-	-	0.685	2q1fA	Yes	No	-	-
2002486_1012	Mark-	34.7	7.9	WP_068417976.1	0.646	4zg4E	No	-	-	-
2002486_1048	fiendish_ghoul	35	4.8	WP_096386863.1	0.617	2bkaA	Yes	Yes	No	-
2002544_0000	Mark-,PLAYER_3	-	-	-	0.645	5ms2A	Yes	Yes	Yes	No
2002553_0000	Mark-,Bletchley Park	-	-	-	0.628	3dcpA	Yes	Yes	No	-
2002553_s003	**Susume**	**36.2**	**2**	**PYT68698.1**	**0.744**	**2g0iA**	**Yes**	**Yes**	**Yes**	**Yes**
2002565_0002	Galaxie,tokens	35.4	8.9	XP_012770361.1	0.666	3m1cA	Yes	Yes	-	-
2002590_0001	Galaxie,tokens	34.3	9	WP_077438279.1	0.726	3dodA	Yes	Yes	-	-
2002590_S567	tokens	36.2	1.8	WP_074948256.1	0.738	4wyaB	No	-	-	-
2002613_0004	PLAYER_25	36.2	2.7	XP_021195829.1	0.752	5iz3A	No	-	-	-
2002713_0000	**retiredmichael,smilingone**	**-**	**-**	**-**	**0.862**	**3rh3A**	**Yes**	**Yes**	**Yes**	**Yes**
2002713_0004	**Mark-,gitwut,Bletchley Park**	**-**	**-**	**-**	**0.742**	**3fajA**	**Yes**	**Yes**	**Yes**	**Yes**
2002713_1006	**Mark-,PLAYER_3**	**38.9**	**0.19**	**WP_044473091.1**	**0.799**	**5nxqA**	**Yes**	**Yes**	**Yes**	**Yes**
2002745_0000	**Mark-,Bletchley Park,georg137**	**34.7**	**5.8**	**GBB96165.1**	**0.819**	**3ripA**	**Yes**	**Yes**	**Yes**	**Yes**
2002745_0001	**PLAYER_4**	**-**	**-**	**-**	**0.873**	**5aqtB**	**Yes**	**Yes**	**Yes**	**Yes**
2002745_0003	**Galaxie,PLAYER_2**	**-**	**-**	**-**	**0.831**	**5cwiA**	**Yes**	**Yes**	**Yes**	**Yes**
2002745_0004	**mirp,Bruno Kestemont,Paulo Roque**	**-**	**-**	**-**	**0.807**	**4uy3A**	**Yes**	**Yes**	**Yes**	**Yes**
2002745_0008	**Madde,kabubi**	**-**	**-**	**-**	**0.849**	**2okuA**	**Yes**	**Yes**	**Yes**	**Yes**
2002766_0000	**Mark-,Bletchley Park**	**37.7**	**0.2**	**WP_116244417.1**	**0.806**	**1y4cA**	**Yes**	**Yes**	**Yes**	**Yes**
2002766_0001	**LociOiling**	**36.2**	**2**	**XP_013096440.1**	**0.839**	**1s94A**	**Yes**	**Yes**	**Yes**	**Yes**
2002766_0002	**PLAYER_25**	**-**	**-**	**-**	**0.813**	**4tqlA**	**Yes**	**Yes**	**Yes**	**Yes**
2002766_0003	Galaxie,tokens	-	-	-	0.846	4iv6A	Yes	Yes	Yes	No
2002766_0004	actiasluna,PLAYER_16,Blipperman	-	-	-	0.903	5cwmA	Yes	Yes	No	-
2002766_0006	PLAYER_17	-	-	-	0.853	4hwhA	Yes	Yes	No	-
2002787_0005	fiendish_ghoul	35	6.2	WP_083480128.1	0.813	4kyzA	Yes	Yes	Yes	No
2002877_s005	gitwut	35.8	5.2	KKS40574.1	0.693	3tv9A	Yes	Yes	No	-
2002922_1013	Mark-,PLAYER_11,Bletchley Park	-	-	-	0.690	5ms2A	Yes	No	-	-
2002922_1018	Hollinas,Bruno Kestemont,Scopper	-	-	-	0.701	6eqtA	Yes	No	-	-
2002922_s004	**tokens**	**36.2**	**1.3**	**EXM13361.1**	**0.710**	**2ejxA**	**Yes**	**Yes**	**Yes**	**Yes**
2002949_0000 a.k.a. foldit1	**Galaxie,Susume,PLAYER_12**	**35.4**	**1.6**	**WP_019366325.1**	**0.712**	**3rf0A**	**Yes**	**Yes**	**Yes**	**Yes**
2002949_0007	fiendish_ghoul	34.3	8.4	PJE68342.1	0.859	2ebbA	Yes	No	-	-
2002990_0006	**fiendish_ghoul**	**37.4**	**1.7**	**RDI84383.1**	**0.803**	**4hhuA**	**Yes**	**Yes**	**Yes**	**Yes**
2002990_1031	fiendish_ghoul	-	-	-	0.734	4clfA	Yes	Yes	No	-
2002990_1039	**fiendish_ghoul**	**37**	**2**	**WP_113960216.1**	**0.713**	**4zhqD**	**Yes**	**Yes**	**Yes**	**Yes**
2002990_s006	retiredmichael	-	-	-	0.717	3ejoA	Yes	No	-	-
2003048_0003	**PLAYER_11,gitwut**	**-**	**-**	**-**	**0.738**	**5adxK**	**Yes**	**Yes**	**Yes**	**Yes**
2003048_0005	Bruno Kestemont,Scopper	35.4	3.8	WP_054386055.1	0.749	2bjiA	No	-	-	-
2003048_1024	actiasluna,PLAYER_5	-	-	-	0.689	3lg0A	Yes	Yes	No	-
2003048_1050	fiendish_ghoul	-	-	-	0.715	1k8kF	Yes	Yes	No	-
2003048_S697	tokens	37.4	0.62	WP_071355392.1	0.683	5izvA	No	-	-	-
2003048_s009	ZeroLeak7	-	-	-	0.744	5i3sD	No	-	-	-
2003169_S953	**tokens**	**-**	**-**	**-**	**0.767**	**4wjbA**	**Yes**	**Yes**	**Yes**	**Yes**
2003169_s001	**tokens**	**-**	**-**	**-**	**0.721**	**1e2sP**	**Yes**	**Yes**	**Yes**	**Yes**
2003169_s008	Susume	-	-	-	0.694	2l25A	Yes	No	-	-
2003205_0000	**PLAYER_12,tokens**	**36.6**	**2.3**	**WP_107152233.1**	**0.735**	**5cw9A**	**Yes**	**Yes**	**Yes**	**Yes**
2003205_0002	**fiendish_ghoul**	**-**	**-**	**-**	**0.888**	**4kyzA**	**Yes**	**Yes**	**Yes**	**Yes**
2003205_0003	actiasluna,PLAYER_26,PLAYER_5,Blipperman	37	1.4	XP_020607139.1	0.660	2b8wA	No	-	-	-
2003205_0006	Vinara	-	-	-	0.701	4p0eA	No	-	-	-

2003205_0008	kabubi	-	-	-	0.674	3cniA	Yes	No	-	-
2003205_1035	actiasluna	34.7	9.2	RHR75271.1	0.715	3c6kD	Yes	No	-	-
2003205_S506	fiendish_ghoul	45.8	2.00E-04	2MQ8_A	0.766	2ddzA	Yes	Yes	No	-
2003205_S722	Susume	36.6	2.4	OVF09966.1	0.722	4ky3A	Yes	No	-	-
2003205_s002	**markm457**	**37**	**1.7**	**WP_095044620.1**	**0.820**	**2kl8A**	**Yes**	**Yes**	**Yes**	**Yes**
2003245_0004	**fiendish_ghoul**	**36.6**	**3.3**	**OQR79684.1**	**0.827**	**4wjbA**	**Yes**	**Yes**	**Yes**	**Yes**
2003245_S383	**fiendish_ghoul**	**39.3**	**0.36**	**WP_113146696.1**	**0.860**	**4wjbA**	**Yes**	**Yes**	**Yes**	**Yes**
2003245_S385	**fiendish_ghoul**	**35**	**9.9**	**WP_008166005.1**	**0.844**	**4pxbA**	**Yes**	**Yes**	**Yes**	**Yes**
2003265_0007	**fiendish_ghoul**	**-**	**-**	**-**	**0.916**	**4neyB**	**Yes**	**Yes**	**Yes**	**Yes**
2003265_1034	fiendish_ghoul	35.8	4.8	OBT67522.1	0.818	4pxdA	Yes	Yes	No	-
2003265_S115	**fiendish_ghoul**	**-**	**-**	**-**	**0.808**	**6c0dA**	**Yes**	**Yes**	**Yes**	**Yes**
2003265_S714	Susume	-	-	-	0.695	5zbrB	Yes	Yes	No	-
2003265_s003	MurloW	-	-	-	0.635	5i4mA	No	-	-	-
2003265_s005	**markm457**	**37.4**	**1.2**	**XP_018401807.1**	**0.802**	**4pxdA**	**Yes**	**Yes**	**Yes**	**Yes**
2003265_s008 a.k.a. foldit3	**Susume**	**41.2**	**0.057**	**WP_005505876.1**	**0.748**	**6g6yA**	**Yes**	**Yes**	**Yes**	**Yes**
2003285_0000	**Galaxie,tokens**	**35.8**	**2.3**	**XP_013764557.1**	**0.794**	**4cllA**	**Yes**	**Yes**	**Yes**	**Yes**
2003285_s000	tokens	43.9	0.003	ATA67140.1	0.791	4cllA	‡	‡	‡	‡
2003285_s005	**Galaxie**	**35.4**	**3.1**	**WP_105677077.1**	**0.668**	**2pozA**	**Yes**	**Yes**	**Yes**	**Yes**
2003285_s008	PLAYER_12	-	-	-	0.718	3jyyA	Yes	Yes	No	-
2003308_0004	**spvincent,gitwut,Bletchley Park**	**36.6**	**2.9**	**OAA53803.1**	**0.678**	**5n9jW**	**Yes**	**Yes**	**Yes**	**Yes**
2003308_0005	PLAYER_17	38.9	0.46	OJT21624.1	0.712	3ejoA	Yes	No	-	-
2003308_0009	kabubi	36.2	3.6	XP_012773543.1	0.674	6eqtA	Yes	No	-	-
2003308_1010	**actiasluna,PLAYER_8**	**-**	**-**	**-**	**0.704**	**3ejoA**	**Yes**	**Yes**	**Yes**	**Yes**
2003333_0000	markm457	37.7	0.71	XP_018574486.1	0.758	4cllA	Yes	No	-	-
2003333_0005	**PLAYER_19**	**35**	**7.4**	**WP_006978974.1**	**0.813**	**4nezA**	**Yes**	**Yes**	**Yes**	**Yes**
2003333_0006 a.k.a. Peak6	**fiendish_ghoul**	**-**	**-**	**-**	**0.819**	**4neyB**	**Yes**	**Yes**	**Yes**	**Yes**
2003333_1006	**MurloW**	**34.7**	**2.3**	**WP_075637384.1**	**0.651**	**5hzyA**	**Yes**	**Yes**	**Yes**	**Yes**
2003333_1013	retiredmichael	35	6.4	WP_011364167.1	0.663	1ukxA	Yes	Yes	No	-
2003333_s001	PLAYER_15	-	-	-	0.739	5lywA	Yes	No	-	-
2003333_s003	Susume	35.8	3.9	WP_027956627.1	0.718	4znlB	Yes	Yes	No	-
2003360_0005	**fiendish_ghoul**	**35**	**4.2**	**WP_044185747.1**	**0.885**	**5tp4B**	**Yes**	**Yes**	**Yes**	**Yes**
2003360_1013	**Bruno Kestemont,PLAYER_20**	**37**	**2.3**	**XP_003196708.1**	**0.644**	**2fhyA**	**Yes**	**Yes**	**Yes**	**Yes**
2003360_s000	**markm457**	**36.2**	**5**	**ATY58566.1**	**0.725**	**3n5fA**	**Yes**	**Yes**	**Yes**	**Yes**
2003360_s002	**PLAYER_7**	**36.2**	**3.1**	**WP_057511752.1**	**0.744**	**4wjbA**	**Yes**	**Yes**	**Yes**	**Yes**
2003360_s003	**MurloW**	**-**	**-**	**-**	**0.635**	**2ixnA**	**Yes**	**Yes**	**Yes**	**Yes**
2003360_s004	**LociOiling**	**38.9**	**0.58**	**XP_019623596.1**	**0.762**	**5eq7A**	**Yes**	**Yes**	**Yes**	**Yes**
2003382_0002	Bruno Kestemont,ZeroLeak7	34.7	8.9	WP_051171461.1	0.708	3lwtX	Yes	Yes	No	-
2003382_0004	Vinara	-	-	-	0.672	5cqcA	Yes	Yes	No	-
2003382_0009	fiendish_ghoul	35.8	3.1	ABW09484.1	0.884	2n3zA	Yes	Yes	No	-
2003382_1021	gitwut	33.1	9.7	WP_037858059.1	0.719	4nogB	Yes	No	-	-
2003382_s005	markm457	35.8	3.1	XP_018821081.1	0.766	2ln3A	Yes	Yes	Yes	No
2003382_s008	**tokens**	**39.3**	**0.16**	**WP_086637674.1**	**0.720**	**2nzcA**	**Yes**	**Yes**	**Yes**	**Yes**
2003414_1018	PLAYER_7	37.7	1.1	WP_056892421.1	0.671	1k8kF	Yes	Yes	No	-
2003455_0001	**Galaxie,tokens**	**40**	**0.085**	**XP_010698604.1**	**0.752**	**4hhuA**	**Yes**	**Yes**	**Yes**	**Yes**
2003455_0002	ZeroLeak7	-	-	-	0.625	1pp0A	No	-	-	-
2003455_0009	Vinara	-	-	-	0.698	5ov5A	Yes	Yes	Yes	No
2003455_1023	PLAYER_2	-	-	-	0.795	1yz7A	No	-	-	-
2003455_S886	Bruno Kestemont	34.3	9.1	WP_039196110.1	0.667	3c66B	No	-	-	-
2003455_S943	Susume	35	5.4	WP_067214428.1	0.667	4nezA	Yes	Yes	Yes	No
2003455_s008	Susume	36.2	1.9	XP_020230923.1	0.709	5o85C	‡	‡	‡	‡
2003485_0000	**Galaxie,markm457**	**-**	**-**	**-**	**0.864**	**4pxdA**	**Yes**	**Yes**	**Yes**	**Yes**
2003485_0002	**Bruno Kestemont**	**38.9**	**0.35**	**WP_006459835.1**	**0.667**	**4zivB**	**Yes**	**Yes**	**Yes**	**Yes**
2003485_1017	ZeroLeak7	-	-	-	0.681	1pp0A	Yes	Yes	No	-
2003485_1029	fiendish_ghoul	35.4	7	WP_091182063.1	0.832	6c0dA	Yes	Yes	No	-
2003485_1036	Vinara	-	-	-	0.704	4akrA	No	-	-	-
2003485_S412	Susume	36.6	2.5	WP_067248652.1	0.712	2vcgA	Yes	No	-	-
2003532_0000	actiasluna	35	1.1	YP_007675131.1	0.651	2kt9A	No	-	-	-
2003532_1020	actiasluna	35.8	4.4	WP_012982902.1	0.785	5ae2D	Yes	Yes	No	-
2003532_1022	kabubi	35	8.3	CCA74651.1	0.680	5nj5A	Yes	No	-	-
2003594_0000	**Galaxie,tokens**	**36.6**	**4.4**	**SDB06526.1**	**0.682**	**4wjbB**	**Yes**	**Yes**	**Yes**	**Yes**
2003594_S028	**tokens**	**37.4**	**2.1**	**CCX30340.1**	**0.672**	**4irxA**	**Yes**	**Yes**	**Yes**	**Yes**
2003594_S603	tokens	38.1	1.4	PYN54536.1	0.638	4zhqD	Yes	Yes	Yes	No
2003594_s008	Susume	-	-	-	0.635	6cfwK	Yes	Yes	No	-

Successful designs are shown in bold. Foldit player usernames are shown only for players who consented to be named in print; non-consenting players are listed anonymously as PLAYER_1, PLAYER_2, etc.

BLAST[49] search was conducted against the nr database of non-redundant protein sequences; scores are omitted where BLAST was unable to find a significant sequence alignment.

TM-align[26] search was conducted against all non-redundant protein chains in the PDB.

† Design 2002290_S122 has high sequence homology with, and is structurally similar to, a family of bacterial transferases with an unusual β-solenoid fold. Unfortunately, the design failed to express.

‡ Because the DNA was never delivered, designs 2003285_s000 and 2003455_s008 were not tested experimentally.

Extended Data

Extended Data Table 1. Foldit player-designed proteins selected for experimental testing
(See Extended Data Table 1 PDF)

Extended Data Table 2. Crystallographic data collection and refinement statistics

	foldit1	**Peak6**
Wavelength	1	1
Resolution range	28.92 - 1.18 (1.222 - 1.18)	26.21 - 1.541 (1.596 - 1.541)
Space group	P 1 21 1	P 31 2 1
Unit cell	24.045 43.584 29.276 90 98.998 90	52.414 52.414 56.086 90 90 120
Total reflections	60389 (6169)	129411 (4118)
Unique reflections	18574 (1830)	12866 (860)
Multiplicity	3.3 (3.4)	10.1 (4.8)
Completeness (%)	92.67 (88.38)	94.86 (65.00)
Mean I/sigma(I)	25.65 (9.97)	18.52 (1.34)
Wilson B-factor	10.36	17.88
R-merge	0.02508 (0.1209)	0.0872 (0.7896)
R-meas	0.03015 (0.1439)	0.09186 (0.878)
R-pim	0.01654 (0.07738)	0.02847 (0.3694)

CC1/2	0.999 (0.991)	0.999 (0.714)
CC*	1 (0.998)	1 (0.913)
Reflections used in refinement	18574 (1749)	12861 (860)
Reflections used for R-free	1829 (174)	1282 (85)
R-work	0.1464 (0.1278)	0.1682 (0.2761)
R-free	0.1819 (0.1755)	0.1975 (0.3091)
CC(work)	0.963 (0.982)	0.967 (0.830)
CC(free)	0.956 (0.956)	0.953 (0.806)
Number of non-hydrogen atoms	690	755
macromolecules	574	646
ligands		20
solvent	116	89
Protein residues	68	77
RMS(bonds)	0.008	0.007
RMS(angles)	0.83	1.03
Ramachandran favored (%)	100.00	100.00
Ramachandran allowed (%)	0.00	0.00
Ramachandran outliers (%)	0.00	0.00

Rotamer outliers (%)	0.00	0.00
Clashscore	2.60	3.75
Average B-factor	16.37	24.96
macromolecules	14.54	22.82
ligands		47.36
solvent	25.39	35.49
Number of TLS groups		3

Statistics for the highest-resolution shell are shown in parentheses.

Extended Data Table 3: NMR data and refinement statistics for foldit3[a]

Distance restraints	
Total NOE-based restraints	2012
Intra-residue	553
Inter-residue	
Sequential (|i-j| = 1)	505
Medium-range (|i-j| ≤ 4)	301
Long-range (|i-j| > 5)	653
Hydrogen bonds restraints	66
Dihedral angle restraints	118
phi	59
psi	59
Restricting restraints / restrained residue	23.0
Restricting long range restraints / restrained residue	6.2
Structure quality statistics	
Restraint Violations	
RMS of distance violation / restraint[b] (Å)	0.01
RMS of dihedral angle violation / restraint (°)	0.88

Max distance restraint violation (Å)	0.66
Max dihedral angle violation (degrees)	7.80
Average r.m.s.d. to medoid conformer[c] (Å)	
Backbone (N. Cα, C')	0.71 ± 0.11
Heavy atoms (all N, C, S, and O)	1.52 ± 0.11
RPF Scores	
Recall	0.912
Precision	0.936
F-measure	0.924
NMR DP-score	0.786
Structure quality factors (raw score / Z-scores[d])	
Procheck G-factor (phi / psi only)	-0.09 / -0.04
Procheck G-factor (all dihedral angles)	-0.14 / -0.83
Verify3D	0.45 / -0.16
ProsaII (-ve)	0.91 / 1.08
MolProbity clashscore	17.51 / -1.48
Ramachandran plot summary (Richardson statistics)	
Most favored regions (%)	97.3
Allowed regions (%)	2.5

Disallowed regions (%)	0.1

[a]Analyzed for the ensemble of 20 lowest-energy structures, residues 1-97, using PDBStat[50] and PSVS[51] ver 1.5 software.

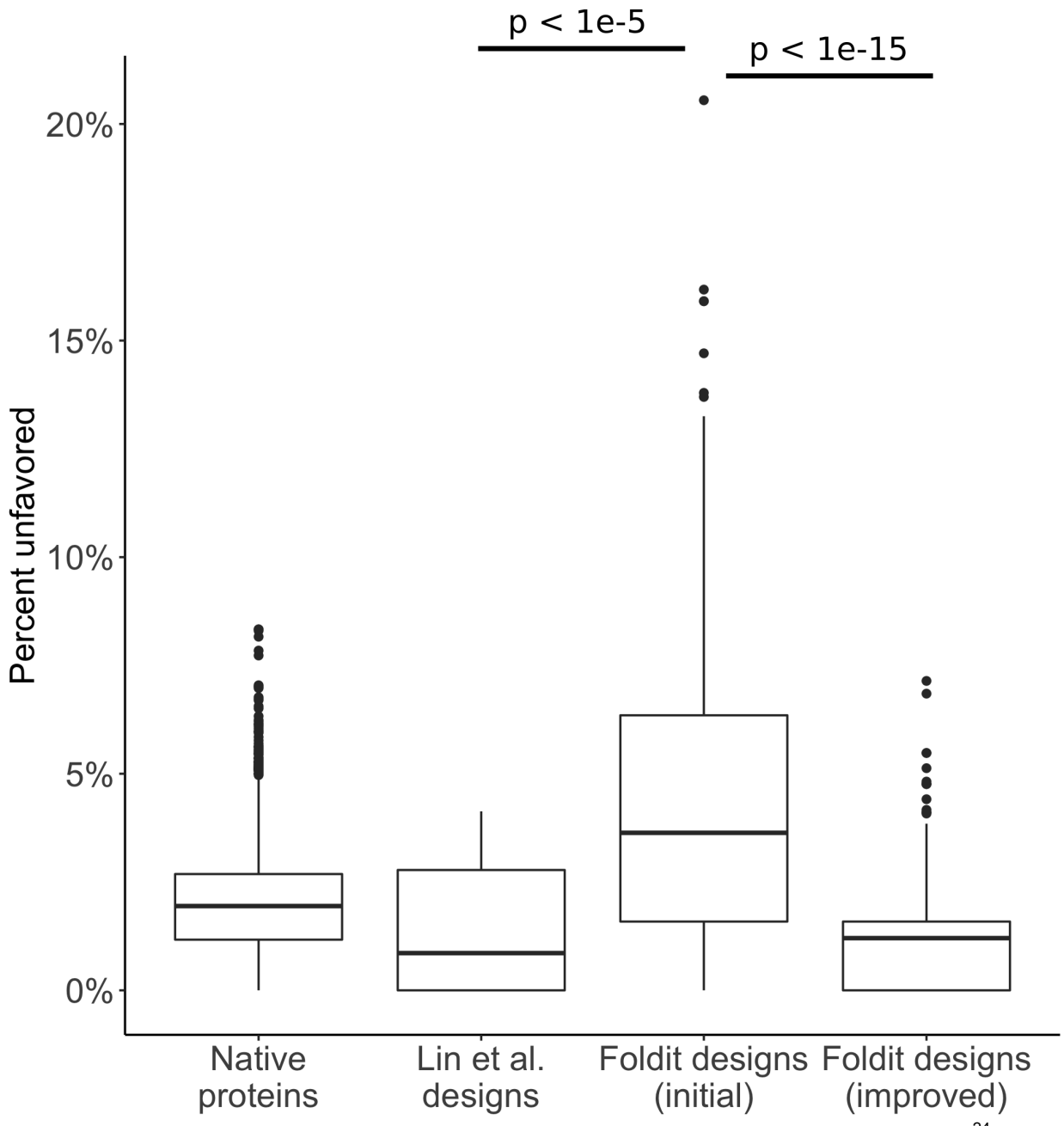
[b]Calculated by using sum over r^{-6} averaging method.

[c]Calculated among 20 structures for “well defined” residues, defined as those that have sum of phi and psi order parameters S(phi)+S(psi) > 1.8. The “well defined” residues are: 21-45, 48-54, 58-76, 81-87, and 90-96.

With respect to mean and standard deviation for a set of 252 X-ray structures with sequence lengths < 500, resolution ≤ 1.80 Å, R-factor ≤ 0.25, and R-free ≤ 0.28; a positive value indicates a 'better' score.

Extended Data Figure 1. Biophysical characterization of successful designs
(See Supplementary Figure 1 PNG)

Extended Data Figure 2. Protein folds represented by successful Foldit player designs. Each fold has a unique arrangement and connectivity of secondary structure elements, depicted in cartoon diagrams. Diagrams are labeled with Roman numerals as in Figure 3. Fold XX is a new fold, previously unobserved in natural proteins; TM-align[26] and DALI[52] alignments against the entire PDB found no structural homologs for design 2003594_S028 with this fold.



Extended Data Figure 3. Improvement of backbone quality in Foldit designs. MolProbity[24] was used to calculate the proportion of residues with “unfavored” or “outlier” backbone torsions in: high-resolution crystal structures of native proteins (*n* = 6342), successful *de novo* design models by Lin et al.[7] (*n* = 18), and top-ranking Foldit player-designs from before (*n* = 717) and after (*n* = 250) improvements to Foldit backbone modeling tools. Initial Foldit player designs contained significantly more unfavored torsions than successful designs by Lin et al. ($p < 1e-5$, two-tailed t-test). Improvements to Foldit’s backbone modeling tools led Foldit players to produce designs with fewer unfavored torsions ($p < 1e-15$, two-tailed t-test). Boxplots show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

Extended Data Foldit player testimonials:

Susan Kleinfelter, a.k.a. Susume (designer of foldit3, and collaborating designer of foldit1):
I almost always start with a pencil sketch of the protein I am going to build; if I don't make a sketch I look at either an old design of mine that I am modifying, or a protein in the pdb or in some other foldit puzzle that I am modifying. I come up with the sketches either by generating shapes using the rules from the Koga & Koga 2012 paper[18], or by modifying proteins I find in the pdb. The Koga rules all rely on very short loops between the sheets and helices, so adapting a natural fold includes shortening the loops to the few that are built in to foldit, and lengthening or shortening strands or helices to make those loops possible. The foldit1 backbone came from a series of sketches I did of modifications to a ferredoxin fold - drop a strand (foldit1), add a strand, add a helix, etc., and/or reverse the termini of any of those designs, all following the Koga rules. The foldit3 backbone was from an exercise I did to find all the 4-strand 2-helix shapes that follow the Koga rules other than the two featured in the Koga paper (ferredoxin-like and IF3-like).

I hand fold first (1-4 hours), then run scripts (2-4 days). The tools I use most to fold up the backbone are blueprint and dragging dots in the rama map (before blueprint came out I just used rama map). At the start I make all the strands ILE, the helices LEU, and the loop AAs according to Fig. S3 of the Lin 2015 paper[7]. I usually get the whole backbone pretty close to the planned shape without wiggling, then band the sheets and shake and wiggle a couple of times. Then I do a few rounds of mutate tool, manually fix AAs, shake and wiggle at various CI. By "fix" I mean manually fix any AAs that mutate has chosen that I don't like, such as buried serine or other hydrophilics, not enough hydrophobics on edge strands or helices, or non-hydrophobics in ideal loop positions where Lin 2015 says they should be hydrophobic.

At this point I switch to running scripts: a remix and mutate script on all the loops plus a few extra residues on each side of them; one or two banders that alternate random collections of bands with mutate; idealize the backbone and switch to medium wiggle; another bander; and a bands plus local wiggle script at the end. I use only a few scripts and run each one for a long time (several hours or overnight) to let lots of nearby positions and mutations get tried. In between scripts I look for AAs I don't like and manually change them back to ones I like, which always lowers my score but I think makes successful folding more likely. Sometimes I will run all the scripts on one track letting foldit choose all sidechains, in order to get higher on the leaderboard, and run one or more other tracks where I choose some sidechains to share with scientists. When puzzle 1297 (the one in which foldit1 was designed) came out I was still using public versions of scripts, but later I modified some of my favorites to let me mark certain sidechains that I want to stay hydrophobic, since that is the most common manual "fix" I was making.

Once I start running scripts I also take the primary sequence after each script and run it through either jpred[53] or psipred[54] to see if the secondary structure I want is considered likely to form. I often manually change AAs just with the purpose of improving the match between the psipred or jpred prediction and my design. I consider this to be both negative design (making it less likely

the protein will fold up some other way than designed) and gaming the system (because Rosetta uses psipred to inform its prediction, and a design only goes to the wet lab if Rosetta successfully matches the player's design). I try to come up with a compromise between foldit's score function and a good psipred/jpred prediction, but I will sometimes submit a design to scientists that has a good secondary structure prediction even if its foldit score is much lower than what I had before.

High score is not my biggest motivator, although in the long script runs the score function is choosing which modifications to keep. I place higher priority on following the rules from the Koga and Lin papers[7,18] (including the ones built into foldit in the form of the ideal loop filter), getting a good psipred prediction, and doing folds that I think are original and cool. I like the challenge of being creative within the somewhat constrained solution space defined by the ideal design rules and foldit's filters, and I think following those rules gives me a leg up in the contest I really care about - getting proteins to fold successfully in the wet lab and ultimately to be published.

(See also Susume's instructional screencast at https://youtu.be/-nizMbICCM0)

Linda Wei, a.k.a. Galaxie (collaborating designer of foldit1):
To begin a design puzzle I look at the puzzle page, noting the filter requirements and puzzle comments. Usually I try to maximize the number of helix segments allowed and use the remaining segments for sheets and loops. Depending on time and computer usage required by other puzzles, several designs are produced, one with the minimum number of segments and one with the maximum amount. Tools used most often in this type of puzzle include blueprint, rama map, idealize ss, freeze, band and the move tool. Once the basic design is set, group and public shared recipes are used to refine the shape of the protein in low to auto wiggle, progressing from low to full ci. Mid to end game involves scripts that idealize and fine tune the protein in medium wiggle. "hand mutating" is used to correct low scoring residues. Generally I rely on the scoring function as an indication of strategy effectiveness. Often I will go back to an earlier save to try a different strategy in an attempt to achieve a higher score.

Player testimonials (cont.)

fiendish_ghoul (designer of Peak6):

Hand-folding of one protein may take 2-4 hours, and the most of this time is occupied by running mutate tool. Before i start folding i usually draw a layout to help determine SS lengths and types of loops. Hand-folding starts with designation of secondary structures, formation of loops with blueprint tool where it is possible, and pulling the resulting structure close to its intended shape. Most amino acids are initially set to valine, amino acids in loops are assigned according to their surroundings, loop shape and amino acid preferences shown in supplementary material from "Control over overall shape and size in de novo designed proteins" paper[7]. Then the protein is wiggled, shaken, mutated and shaken again, and this procedure is repeated with manual interventions between rounds to fix inappropriate mutation choices and other issues. Clashing importance is gradually increased between rounds from 0.01-0.1 to 1. After hand-folding stage, running recipes takes around 10 hours. TvdL DRemixW is run the majority of time and some idealizing recipes like Microidealize and Random Idealize are used at the end. I use only publicly available recipes since i don't make my own.

One approach to designing proteins i use is to combine structures (mostly βαββ and βαβ units) described in "Control over overall shape and size in de novo designed proteins" paper[7] in various ways. Some modifications can be applied to these standard structures, like replacing sheet-sheet loop with loop-helix-loop-helix-loop sequence.

Another approach is to modify existing folds. For example, proteins with N and C termini positioned near each other (like in ferredoxin-like fold from previously mentioned paper) can be imagined as continuous structure, and then a "cut" can be made in place of one of the loops of the protein to get a protein with similar shape but different SS sequence. Recently validated protein was designed this way.

These approaches have been yielding my best scoring solutions so far and are relatively easy to implement. More experimental folds tend to have much lower score. In general, simpler folds with 2 helices and 3-4 sheets have the highest score, and the more complex the protein gets, the more difficult it is to make it score well.

Regarding foldit tools, i find blueprint tool useful because it speeds up folding process and makes it possible to quickly sample different structures. Rama Map is useful for creating structures like beta bulges and loops that are not in blueprint library.

Extended References

49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).
50. Tejero, R., Snyder, D., Mao, B., Aramini, J.M., Montelione, G.T. PDBStat: A universal restraint converter and restraint analysis software package for protein NMR. *J. Biomol. NMR* **56**, 337-351 (2013).
51. Bhattacharya, A., Tejero, R., Montelione, G.T. Evaluating protein structures determined by structural genomics consortia. *PROTEINS: Struct. Funct. Bioinformatics* **66**, 778 - 795 (2007).
52. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44,** W351–5 (2016).
53. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43,** W389–94 (2015).
54. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16,** 404–405 (2000).