De novo protein design by citizen scientists

Brian Koeppnick1,2, Jeff Flatt3, Tamir Husain3, Alex Ford1,2, Daniel–Adriano Silva1,2, Matthew J. Bick1,2, Aaron Bauer3, Gaohua Liu4,5, Yojiro Ishida6, Alexander Boykov11, Roger D. Estep11, Susan Kleinfelter11, Toke Nørgård-Solano11, Linda Wei11, Foldit Players10, Gaetano T. Montelione4,6, Frank DiMaio1,2, Zoran Popović3, Firas Khatib5, Seth Cooper6 & David Baker1,2,9

Online citizen science projects such as GalaxyZoo1, Eyewire2 and Phylo3 have proven 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 EteRNA4, 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 Foldit5. Players were presented with a fully extended peptide chain and challenged to craft a folded protein structure and 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 that encode them in silico. One hundred forty-six Foldit player designs with sequences unrelated to naturally occurring proteins were encoded in synthetic genes; 56 were found to be expressed and soluble in Escherichia coli, and to adopt stable monomeric folded structures in solution. The diversity of these structures is unprecedented in de novo protein design, representing 20 different folds—including a new fold not observed in natural proteins. High-resolution structures were determined for four of the designs, and are nearly identical to the player models. This work makes explicit the considerable implicit knowledge that contributes 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.

The principle underlying de novo protein design is that proteins fold to their lowest free-energy state6; hence, designing a new protein structure requires finding an amino acid sequence with its lowest energy state in the prescribed structure. In practice, this challenge can be divided into two subproblems: first, crafting a protein backbone that is designable (that is, that could be the lowest energy state of some sequence); and second, finding a sequence with its lowest energy state in the crafted structure. One of the challenges of protein design is the exponentially increasing number of conformations available to a polypeptide chain, which is huge even for a modestly sized protein of 60–100 residues. Thus, the first subproblem of crafting a plausible backbone 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 years7–10, 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 (that is, 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 de novo protein design problem, we incorporated de novo design tools into the protein-folding game Foldit. Foldit is a free online computer game developed to crowdsource problems in protein modelling, and provides full control over the three-dimensional structure of a protein model11 (Fig. 1). Players compete to build a model with the lowest free energy, as calculated using the Rosetta energy function11. In the past, Foldit has been primarily applied to protein structure prediction problems, in which players are presented with an unstructured amino acid sequence and challenged to determine its native conformation5,12. In one case, Foldit players redesigned a loop region of an already folded structure13, 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 on the basis of their results. In each challenge, players were provided with a polylysine backbone in a fully extended conformation (60–100 residues in length) and were given 7 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 (Extended Data Fig. 1); such extended, fully α-helical structures have more favourable hydrogen bonding, electrostatic and local torsional energies than collapsed structures, which must contort to create a buried core. Whereas polylysine and other extended polar sequences resembling these initial Foldit solutions are often α-helical in solution13,14, the lack of long-range interactions precludes specific folding into a single stable structure15. 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 favour 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 (for example, 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.
Fig. 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 identity of the amino-acid residue at any position of the model. c, The ‘pull’ tool allows players to manipulate the 3D 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 or rotate model, secondary structure assignment, idealize secondary structure, manually design sidechains, delete residues, insert residues, insert cutpoint and idealize peptide bond geometry. f, Foldit players explore diverse structures that have no sequence or structural homology to natural proteins.

Fig. 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 (Ferredox-Diesel); and design trajectories for four Rosetta-designed proteins in grey. 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; by contrast, the Rosetta protocol 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 polyisoleucine; (ii) development of secondary structure; (iii) development of folded tertiary structure; (iv) sequence design of folded structure, with inset showing favourable packing 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 renewed favourable interactions 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 after 17 branch points. Node colours correspond to five different cooperating Foldit players, and the final design is marked with a star. d, Similar representation of a Rosetta design trajectory—there are only two branch points.
initio calculations show that the sequence for each design has an energy higher energy states than the automated protocol, which has only a
done any attempt to design more varied folds. This is an interesting
bundles (Extended Data Fig. 2), and competitive players had aban-
–sheet did not score as well as
α
–
β
proteins.

We obtained custom synthetic genes encoding 12 player designs
for which structure prediction calculations converged on the play-
er-designed conformation17. The sequences of these proteins have no
homology to any known protein (Supplementary Table 1). The de novo
designs were expressed in E. coli and purified by metal-affinity and
size-exclusion chromatography. Analysis by chromatography and
 circular dichroism indicated that 6 of the 12 designs were monomeric
and folded in solution, with helical secondary structure consistent with
all top-scoring Foldit designs had consisted of either three or four
players to explore more-diverse protein structures. Up until this point,
limited ability to escape local energy minima.

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 (Fig. 2). Foldit players use more-varied and complex
exploration strategies than standard Rosetta automated design pro-
tocols, 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 points18. In addition, Foldit players regularly sample much
higher energy states than the automated protocol, 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 to encourage
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 as well as α-helical
bundles (Extended Data Fig. 2), and competitive players had aban-
doned any attempt to design more varied folds. This is an interesting
parallel to protein design by practicing scientists, which has also
focused much more on helical bundles than on other classes of protein
taks20–22. To encourage the design of a wider variety of folds, we intro-
duced 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 on visual inspection. However, structure prediction cal-
culations for these α/β design sequences showed poor sampling close to
the target design structure, which suggests that the designed sequences
did not strongly encode their local structures23. Further analysis showed
that these player designs contained many residues with locally strained
backbone conformations (backbone φ and ψ torsions in unfavoured
regions of the Ramachandran plot23,24). That such designs had very
low energies revealed a problem in the Rosetta energy function at the
time: because Rosetta users typically sampled backbones starting from
fragments of native proteins, unfavourable local conformations were
rarely encountered—therefore, it had not been discovered that the
energies associated with local-backbone strain were being under-
estimated. We addressed this flaw in the Rosetta model by increas-
ing the steepness of the energetic penalties associated with strained
local-backbone geometry; this is now implemented in the latest Rosetta
energy function11. We also added to Foldit an ‘ideal loops’ rule that
restricted players to a set of 19 unstrained reverse-turn conforma-
tions2, 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 conforma-
tions (Extended Data Fig. 3). Together, these upgrades brought about
a marked improvement in the local–backbone quality of Foldit player-
designed proteins (Extended Data Fig. 4).
The importance of reducing local-backbone strain was borne out in experimental characterization. Before the backbone modelling improvements described in the previous paragraph, only 4 of 37 Foldit \(\alpha/\beta\)-designs tested (11%) were monomeric and structured in solution. Following the backbone modelling 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 some free energies of unfolding (\(\Delta G^{\text{unf}}\)) exceeding 20 kcal mol\(^{-1}\); indeed, 32 designed proteins remained completely folded at 95 °C (Fig. 3, Supplementary Fig. 1). This success rate surpasses that in previous reports of designed \(\alpha/\beta\)-proteins\(^2\text{-}^{12}\).

Overall, the 56 successful Foldit designs are diverse in structure, representing 20 different protein folds (Fig. 3, Extended Data Fig. 5), one of which is a new fold that is 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 (Supplementary Table 1). The 56 successful designs were created by 36 different Foldit players (the most prolific player created 10 successful designs); 19 designs were created collaboratively by at least 2 cooperating players; and 5 successful designs were not top-scoring, but were nevertheless flagged by players as personal favourites. Foldit players lack formal expertise in protein modelling (Extended Data Fig. 6, Supplementary Notes), but knowledge and intuition gained from playing protein structure prediction puzzles in Foldit translated to success in de novo protein design (Extended Data Fig. 7).

We succeeded in solving high-resolution structures of four Foldit player-designed proteins. X-ray crystal structures of three designed proteins (named by their designers Foldit1, Peak6 and Ferredog-Diesel) closely match the designed conformations, with \(C_{\alpha}\) root mean square deviations (r.m.s.d.) of 1.1, 0.9 and 1.7 Å, respectively (Fig. 4). Well-resolved electron density in the protein core of Foldit1 and Peak6 shows that most sidechains adopt the intended rotamers and preserve the designed packing interactions. The electron density of Ferredog-Diesel is less clear, but the protein backbone adopts the designed fold, and many core sidechains appear to pack as intended. The solution nuclear magnetic resonance (NMR) structure of a fourth design, Foldit3, also closely matches the design conformation, with a \(C_{\alpha}\) r.m.s.d. of 1.1 Å between the design model and the medoid conformer\(^25\) of the ensemble.

From these results, we can draw several general conclusions about scientific models, citizen science and the interplay between the two. First, a scientific model that 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 importantly, 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, to our knowledge. 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.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1274-4.

Received: 19 October 2018; Accepted: 14 May 2019; Published online: 05 June 2019

1. Lin, S. et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Mon. Not. R. Astron. Soc. 389, 1179–1189 (2008).
2. Kim, J. et al. Space-time wiring specificity supports direction selectivity in the retina. Nature 509, 331–336 (2014).
3. Kawrykow, A. et al. Phyloto: a citizen science approach for improving multiple sequence alignment. PloS ONE 7, e31362 (2012).
4. Lee, J. et al. RNA design rules from a massive open laboratory. Proc. Natl Acad. Sci. USA 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 structure. Cold Spring Harb. Symp. Quant. Biol. 28, 439–449 (1963).
7. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. Proc. Natl Acad. Sci. USA 112, E5478–E5485 (2015).
8. Huang, P.-S., Boyleen, 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 α sheets. Science 355, 201–206 (2017).
10. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. Nature 561, 485–491 (2018).
11. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theory Comput. 13, 3031–3047 (2017).
12. Khabib, 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. Elben, C. B. et al. Increased Diels–Alderase activity through backbone remodeling guided by Foldit players. Nat. Biotechnol. 30, 190–192 (2012).
14. Blout, E. R. & Iseidson, M. Compositional effects on the configuration of water-soluble polypeptide copolymers of l-glutamic acid and l-lysine. J. Am. Chem. Soc. 80, 4909–4913 (1958).
15. Droy, P., Imahori, K. & Klemperer, E. The solution properties and configurations of a polyampholytic polypeptide: copoly-l-lysine-l-glutamic acid. Proc. Natl Acad. Sci. USA 44, 424–431 (1958).
16. Ghosh, K. & Dill, K. A. Theory for protein folding cooperativity: helix bundles. J. Am. Chem. Soc. 131, 2306–2312 (2009).
17. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. Methods Enzymol. 383, 66–93 (2004).
18. Koga, N. et al. Principles for designing ideal protein structures. Nature 491, 222–227 (2012).
19. Regan, L. & DeGrado, W. F. Characterization of a helical protein designed from first principles. Science 241, 976–978 (1988).
20. Hartbury, 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 α-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. Adv. Protein Chem. 23, 283–438 (1968).
24. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D 66, 12–21 (2010).
25. Montellione, 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. & Bolten, D. W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethane sulfonyl-α-chymotrypsin using different denaturants. Biochemistry 27, 8063–8068 (1988).

Acknowledgements
We thank all Foldit players for their gameplay contributions, and for feedback offered on the Foldit website (https://fold.it). We thank A. Kang, S. A. Rettie, C. Chow and L. Carter for help with experiments; D. Alonso, L. Goldschmidt, P. Vecchiato and D. Kim for computer support; and Rossetta@home (https://boinc.bakerlab.org) volunteers for computing resources. We thank G. Rocklin, V. Mulligan and other members of the Baker laboratory for discussions. This material is based on work supported by the National Science Foundation (NSF) Graduate Research Fellowship under grant no. DGE-1256082, NSF grant no. 1629879, National Institutes of Health (NIH) grant 1U1H2CA203780, and NIH grants 1S10 OD018207 and 5R01 GM120574 (to G.T.M.) and HHMI (D.B.). The ALS-ENABLE beamlines are supported in part by the NIH, 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. Foldit is a nominated target of the CASP COMMONS Community Outreach program.

Reviewer information
Nature thanks Jérôme Waidspühl and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions
B.K. designed the study. B.K., Z.P., F.K., S.C. and D.B. performed the experiment. 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., S.M., D.A.S. and B.K. designed all proteins. B.K., F.K., A.F. and A. Bauer analysed Foldit player designs. B.K. performed biophysical characterization. B.K., M.J.B. and F.D. determined crystal structures. G.L. and G.T.M. determined the NMR structure. B.K. and D.B. wrote the manuscript with input form all authors. Foldit players contributed extensively through their feedback and gameplay, which generated the data for this paper.

Competing interests
G.T.M. is a co-founder of Nexomics Biosciences.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1274-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1274-4.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.B.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019
All circular dichroism data were collected on an AVIV Model 420 spectrometer.

Circular dichroism.

Circular dichroism data were collected on an AVIV Model 420 spectrometer.

Circular dichroism. See Extended Data Table 1 for summary statistics on design selection.

Circular dichroism. Prior to X-ray crystallography, the N-terminal 6×His tag was cleaved from protein samples by incubation with 250 μg TEV protease at 25 °C for 4 h in 20 mM Tris pH 8.0, 300 mM NaCl, 1 mM DTT. The reaction product was dialysed into TBS overnight at 4 °C 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 sparse matrix suites available from Qiagen or Hampton Research. A Mosquito Crystal nanolitre 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.

Foldit (2002949_0000) 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 from the supernatant without further cryo-protection. X-ray diffraction was collected to a resolution of 1.18 Å.

PK6 (200333_0006) was crystallized at 40 mg/ml in 0.1 M sodium acetate pH 4.5, 0.2 M lithium sulphate, 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 was collected to a resolution of 1.54 Å.

Ferredog-Diesel (2003169_5953) was crystallized with 6×His tag intact, at 80 mg/ml in 0.1 M citrate pH 4.0, 3.0 M NaCl. Crystals were dehydrated by soaking in 5 μl mother liquor in open air for 10 min, then flash-frozen in liquid nitrogen. X-ray diffraction was collected to a resolution of 1.92 Å.

X-ray diffraction datasets were collected at the Advanced Light Source (Berkeley, CA). Data was processed with HKL2000. Crystal structures were solved by molecular replacement with Phaser, using the backbone of the original designed model with sidechains truncated to the β-carbon (Foldit and Peak6), or using the backbone of a model predicted ab initio from the design sequence (Ferredog-Diesel). Models were built and refined in iterative cycles using Coot and PHENIX. Diffraction data and refinement statistics are listed in Extended Data Table 2.

NMR spectroscopy. NMR studies were performed using uniformly 15N,13C-enriched protein samples. A synthetic gene for Foldit (2003265_0008) was obtained from GenScrypt incorporated into plasmid PET15EV_NESG, which includes a N-terminal 6×His purification tag, followed by a TEV protease cleavage site (sequence MGHHHHHHHGGWSENLYFQQGS). E. coli BL21(DE3) cells (Invitrogen), and grown overnight in 4 ml Luria–Bertani (LB) medium 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 medium, and grown at 37 °C for 18 h. 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 dialysed into TBS (20 mM Tris pH 8.0, 300 mM NaCl) at 4 °C overnight to remove imidazole and further purified by size-exclusion chromatography on an AKTAexpress (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 tris-(2-carboxyethyl)phosphine (TCEP). Protein expression and solubility was determined from SDS–PAGE and mass spectrometry. Oligomeric state was determined by size-exclusion chromatography and analysed using the programs SPARKY and XEASY. Spectra were referenced to external DSS. Sequence-specific resonance assignments were determined using AutoAssign software together with interactive manual analysis, as previously described. Backbone dihedral angle...
constraints were derived from the chemical shifts using the program TALOS-N\textsuperscript{36} for residues located in well-defined secondary structure elements. The programs ASDP\textsuperscript{40} and CYANA\textsuperscript{18,39} were used to automatically assign NOEs and to calculate structures. RPF analysis\textsuperscript{57,41} was used in parallel to guide iterative cycles of noise and artefact peak removal, peak picking, and NOEpeak assignments. The 20 conformers with the lowest target CYANA function value were then refined in explicit water\textsuperscript{41} using the program CNS\textsuperscript{34}. The structural statistics and global structure quality factors (Extended Data Table 3) including Verify3D\textsuperscript{43}, ProsaII\textsuperscript{44}, PROCHECK\textsuperscript{45}, and MolProbity\textsuperscript{46} raw and statistical Z-scores were computed using the PSVS\textsuperscript{57} v.1.5 and PDBStat\textsuperscript{48} software packages. The global goodness-of-fit of the final structure ensembles with the NOE peak list data, the NMR DP score, was determined using the RPF analysis program\textsuperscript{40}. Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinates of Foldit1, Peak6 and Ferredog-Diesel crystal structures and the Foldit3 NMR structure have been deposited in the RCSB Protein Data Bank (PDB) with accession numbers 6MRR, 6MRS, 6NUK and 6MSP, respectively. Chemical shift and NOESY peak list data for Foldit3 were deposited in the Biological Magnetic Resonance Data Bank with accession number 30527.

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 modelling 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.

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 66, 486–501 (2010).
31. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr. D 68, 352–367 (2012).
32. Jansson, M. et al. High-level production of uniformly $^{15}$N- and $^{13}$C-enriched fusion proteins in Escherichia coli. J. Biolog. 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., Güntert, P. & Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J. Biomol. NMR 6, 1–10 (1995).
35. Liu, G. et al. NMR data collection and analysis protocol for high-throughput protein structure determination. Proc. Natl Acad. Sci. USA 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 (2009).
37. Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOE data. Proteins 62, 587–603 (2006).
38. Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. 273, 283–298 (1997).
39. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. 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. J. Am. Chem. Soc. 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 (2003).
42. Brünger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D 54, 905–921 (1998).
43. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356, 83–85 (1992).
44. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355–362 (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. J. Appl. Crystallogr. 26, 282–291 (1993).
46. Word, J. M., Bateman, R. C., Jr, Presley, B. K., Lovell, S. C. & Richardson, D. C. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci. 9, 2251–2259 (2000).
47. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. Proteins 66, 778–795 (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).
49. Trifonov, E. N. in Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci. 9, 2251–2259 (2000).
Extended Data Fig. 1 | Initial top-ranking Foldit player designs. When challenged to design a protein with only the talaris2013 score function (and no additional rules), Foldit players discovered low-energy models that are unlikely to fold as designed. **a**, An extended α-helix, composed entirely of lysine and glutamate, has very favourable energies for hydrogen-bonding, electrostatic and backbone torsions, but is unlikely to fold cooperatively into a single stable structure. This type of design is discouraged with the ‘core exists’ rule. **b**, Owing to their greater surface area, large aromatic sidechains can make more interactions than smaller aliphatic sidechains, even when underpacked or solvent-exposed. This type of design is discouraged with the ‘residue interaction energy’ rule. **c**, A design with an alanine- and glycine-saturated core can make favourable van der Waals interactions between closely packed backbone atoms; however, the burial of these small sidechains is associated with a weaker hydrophobic effect, and the lack of interdigitation allows exchange between multiple conformations with similar core packing energies (that is, molten globule behaviour). These designs are discouraged with the ‘secondary structure design’ rule.
Extended Data Fig. 2 | Rosetta energy of top Foldit player designs.
Rosetta energy of top-ranking designs was calculated with the talaris2013 score function and normalized by residue count. **a**, Energy of top-ten-ranked designs from: initial Foldit puzzles (round 0; \( n = 30 \) designs), round 1 puzzles (\( n = 170 \)), round 2 puzzles (\( n = 510 \)) and round 3 puzzles (\( n = 250 \)). The introduction of supplementary rules in round 1 and round 2 resulted in higher-energy designs (\( P < 10^{-6} \) and \( P < 0.01 \), respectively; Wilcoxon rank-sum test). The backbone modelling improvements in round 3 resulted in lower-energy designs (\( P < 10^{-15} \); Wilcoxon rank-sum test). **b**, Energy of top-ten-ranked designs from round three all-\( \alpha \) puzzles (\( n = 30 \)) or \( \alpha/\beta \)-puzzles using the ‘secondary structure’ rule (\( n = 220 \)). All-\( \alpha \) designs tend to have lower energy than \( \alpha/\beta \)-designs (\( P < 10^{-10} \); Wilcoxon rank-sum test). Box plots show: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 \( \times \) interquartile range; points, outliers.
Extended Data Fig. 3 | New backbone-modelling tools in Foldit. a, The ‘remix’ tool allows players to select a region of the model and search a library of backbone fragments for a conformation that can be substituted. b, An interactive Ramachandran map allows players to easily identify residues with outlier backbone conformations. Players can also click and drag points on the Ramachandran map to set the backbone torsions of individual residues. c, A ‘blueprint’ panel shows the primary sequence and secondary structure content of the model. Residues are coloured according to the ABEGO quadrants of the Ramachandran plot. d, Players can drag-and-drop modular building blocks onto the blueprint panel to insert common turn conformations into their model.
Extended Data Fig. 4 | Improvement of backbone quality in round 3 Foldit designs. MolProbity24 was used to calculate the proportion of residues with unfavored or outlier backbone torsions in: high-resolution crystal structures of native proteins (n = 6,342), de novo design models from a previous study7 (n = 72), and top-ranking Foldit player designs from before (n = 680) and after (n = 250) improvements to Foldit backbone-modelling tools. Initial Foldit player designs contained significantly more unfavoured torsions than native proteins or other de novo designs from a previous study7 (P < 10^{-15}, two-tailed t-test). Improvements to Foldit’s backbone-modelling tools led Foldit players to produce designs with fewer unfavoured torsions (P < 10^{-15}, two-tailed t-test). Box plots show: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers.
Extended Data Fig. 5 | 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 labelled with Roman numerals as in Fig. 3. Fold XX is a new fold, previously unobserved in natural proteins; TM-align\textsuperscript{26} and DALI\textsuperscript{10} alignments of design 2003594_S028 against the entire PDB found no structural homologues with this fold.
Extended Data Fig. 6 | Foldit player demographics. All players who participated in Foldit protein design puzzles and who had not opted out of Foldit-related email were solicited for survey questions. Data are shown for \( n = 324 \) responding Foldit players.
Extended Data Fig. 7 | Category rankings of Foldit players. Foldit player rankings are strongly correlated in the design and prediction categories (Spearman's rank correlation coefficient of 0.84). This suggests that skills developed playing Foldit structure prediction puzzles carry over to design puzzles and vice versa.
Extended Data Table 1 | Success rates of Foldit player-designed proteins

|                      | Foldit player designs |        |        |        |        |
|----------------------|-----------------------|--------|--------|--------|--------|
|                      | Round 0               | Round 1| Round 2| Round 3| Lin et al.† |
| Sequence complexity* | 0.20                  | 0.35   | 0.44   | 0.21   | 0.20   |
| Rosetta energy† (per residue) | -2.6 ± 0.1         | -2.2 ± 0.5 | -2.1 ± 0.2 | -2.2 ± 0.1 | -1.9 ± 0.1 |
| Total puzzles        | 3                     | 17     | 51     | 25     |         |
| Avg. players per puzzle | 123 ± 19              | 212 ± 34 | 189 ± 36 | 151 ± 16 |         |
| Raw model count      | 140,273               | 2,887,213 | 10,556,093 | 4,124,471 |         |
| Top models           | 60                    | 340    | 1020   | 500    |         |
| Shared models        | 53                    | 214    | 726    | 342    |         |
| Clustered models     | 150                   | 850    | 2550   | 1250   |         |
| Total models considered‡ | 263                   | 1404   | 4296   | 2092   |         |
| Models selected for ab initio | 0                  | 100    | 1141   | 612    | (Not reported) |
| Ab initio convergence| NA                    | 12     | 12%    | 37     | 99     | 16%    | 72     |
|                      | NA                    | 12     | 100%   | 37     | 97     | 72     |
| Expresseed           | NA                    | 12     | 100%   | 23     | 86     | 89%    | 70     | 97%    |
| ... and soluble      | NA                    | 12     | 100%   | 18     | 71     | 73%    | 64     | 89%    |
| ... and monomeric    | NA                    | 7      | 58%    | 7      | 52     | 54%    | 39     | 54%    |
| ... and structured   | NA                    | 6      | 50%    | 4      | 46     | 47%    | 29     | 40%    |
| Number of unique folds | NA                   | 3      | 4      | 19     | 2      |

* Linguistic sequence complexity§ was calculated from the top-ten-ranked models in all puzzles, using word lengths of 1, 2 and 3.
† Rosetta energy is the talaris2013 energy normalized by residue count. Values shown are mean and standard deviation for the ten top-ranked models in all puzzles. See Extended Data Fig. 2 for sample sizes.
‡ Includes redundant models, as very similar models can appear in two or more categories (top, shared and clustered). See Methods for details on model selection.
## Extended Data Table 2 | X-ray crystallography data and refinement statistics

|                    | Foldit1 (6MRR) | Peak6 (6MRS) | Ferredog-Diesel (6NUK) |
|--------------------|----------------|--------------|-----------------------|
| **Data collection**|                |              |                       |
| Space group        | P 1 2 1       | P 3 1 2 1    | P 4 2 1 2             |
| Cell dimensions    |                |              |                       |
| $a$, $b$, $c$ (Å)  | 24.05, 43.58, 29.28 | 52.41, 52.41, 56.09 | 69.21, 69.21, 90.59 |
| $\alpha$, $\beta$, $\gamma$ (°) | 90, 99.0, 90 | 90, 90, 120 | 90, 90, 90 |
| Resolution (Å)     | 28.92 - 1.18 | 26.21 - 1.541 | 45.29 - 1.916 |
|                    | (1.222 - 1.18) | (1.596 - 1.541) | (1.985 - 1.916) |
| $R_{merge}$        | 0.02508 (0.1209) | 0.0872 (0.7896) | 0.08947 (3.164) |
| $I / \sigma I$     | 25.65 (9.97)  | 18.52 (1.34)  | 16.94 (0.86) |
| Completeness (%)   | 92.67 (88.38) | 94.86 (65.00) | 99.06 (97.65) |
| Redundancy         | 3.3 (3.4)    | 10.1 (4.8)   | 11.7 (11.5) |
| **Refinement**     |                |              |                       |
| Resolution (Å)     | 1.18           | 1.541         | 1.916                 |
| No. reflections    | 18574          | 12861         | 17376                |
| $R_{work} / R_{free}$ | 0.146 / 0.182 | 0.168 / 0.198 | 0.248 / 0.291 |
| No. atoms          |                |              |                       |
| Protein            | 574            | 646           | 1672                  |
| Ligand/ion         | 0              | 20            | 0                     |
| Water              | 116            | 89            | 37                    |
| **B-factors**      |                |              |                       |
| Protein            | 14.54          | 22.82         | 69.09                 |
| Ligand/ion         | 0              | 47.36         | 0                     |
| Water              | 25.39          | 35.49         | 55.90                 |
| **R.m.s. deviations** |              |              |                       |
| Bond lengths (Å)   | 0.008          | 0.007         | 0.005                  |
| Bond angles (°)    | 0.83           | 1.03          | 1.01                   |

Values in parentheses are for highest resolution shell. X-ray diffraction data for each protein structure were collected on a single crystal and processed as described in the Methods.
Extended Data Table 3 | NMR and refinement statistics for protein structures

|                         | Foldit3 (6MSP) |
|-------------------------|----------------|
| **NMR distance and dihedral constraints** |                |
| Distance constraints    |                |
| Total NOE               | 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          | 66             |
| **Total dihedral angle restraints** |            |
| \( \phi \)             | 59             |
| \( \psi \)              | 59             |
| **Structure statistics** |                |
| Violations              |                |
| Distance constraints (Å)| 0.01           |
| Dihedral angle constraints (°) | 0.88       |
| Max. dihedral angle violation (°) | 7.80       |
| Max. distance constraint violation (Å) | 0.66       |
| **Structure quality factors (raw score / Z-scores)** |            |
| Procheck G-factor (phi/psi only) | -0.09 / -0.04 |
| Procheck G-factor (all dihedrals angles) | -0.14 / -0.83 |
| Verify3D                | 0.45 / -0.16  |
| ProSali (-ve)           | 0.91 / 1.08   |
| MolProbity clashscore   | 17.51 / -1.48 |
| **Average pairwise r.m.s. deviation* (Å)** |            |
| Heavy                   | 1.52           |
| Backbone                | 0.71           |

*Pairwise r.m.s.d. was calculated among 20 refined structures for 'well-defined' residues 21–45, 48–54, 58–76, 81–87 and 90–96.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection
The pre-compiled Foldit game is freely available for download at https://fold.it for Windows, Linux, and Mac. A standalone version of Foldit is also freely available for academic use; for details visit https://fold.it/standalone. Foldit configuration files for all design puzzles are included in the Supplementary Information. The Rosetta software suite was used to perform ab initio prediction calculations; Rosetta is freely available for academic users on Github, and can be licensed for commercial use by the University of Washington CoMotion Express License Program.

Data analysis
Custom Python scripts written to analyze circular dichroism data are included in the Supplementary Information. Protein structures were analyzed with MolProbity (version 4.2). Crystallographic data were analyzed with PHENIX (release 1.101.1-2155) and Coot (v0.8.7 EL). NMR data were analyzed with SPARKY, XEASY, TALOS_N, ASDP, CYANA, PDBStat and PSVS (version 1.5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic coordinates of Foldit1, Peak6, and Ferredog-Diesel crystal structures, and the Foldit3 NMR structure, have been deposited in the RCSB Protein Data Bank with accession numbers 6MRR, 6MRS, 6NUK and 6MSP, respectively. Chemical shift and NOESY peak list data for Foldit3 were deposited in the Biological Magnetic Resonance Bank (BMRB ID 30527).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences  -  [ ] Behavioural & social sciences  -  [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size: Sample size for protein characterization was determined by estimated work load. In total, 146 protein designs were tested, from 97 Foldit puzzles. This was deemed sufficient due to the high number of successfully folded designs in our testing. For in silico designed-backbone analysis, the sample sizes of (n = 717 or 250) was considered sufficient. The inclusion of additional samples is not expected to affect the distribution of measured values.

Data exclusions: No data were excluded from analysis.

Replication: Puzzle configurations were used repeatedly in replicated Foldit puzzles to ensure reproducibility. The final puzzle configuration was used for 25 replicate Foldit puzzles. Protein expression and solubility was tested in duplicate. Structural characterization were performed once or twice with internal statistical validation. If positive results (e.g. protein expression, solubility, etc.) could not be replicated, they are reported as negative results.

Randomization: There was no randomized sample allocation in this work. All tested protein designs received identical treatment.

Blinding: Blinding was not relevant to this work, since all tested proteins received identical treatment.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a Involved in the study

- [x] Unique biological materials
- [x] Antibodies
- [x] Eukaryotic cell lines
- [x] Palaeontology
- [x] Animals and other organisms
- [ ] Human research participants

Methods

n/a Involved in the study

- [x] ChIP-seq
- [ ] Flow cytometry
- [ ] MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics: Participation was open and free to the public, and we did not control for participant demographics. See Extended Data Fig. 6 for demographic data from a voluntary (non-obligatory) participant survey.

Recruitment: Participation was open and free to the public, at https://fold.it.