Protein design by citizen scientists

Brian Koepnick

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

David Baker (chair)

Frank DiMaio

Ethan Merritt

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Protein design by citizen scientists

Brian Koepnick

Chair of Supervisory Committee:

David Baker

Department of Biochemistry

Proteins are a class of molecule best known for their tendency to fold into well-defined 3-dimensional structures. The structure of a protein is determined by the sequence of amino acid units that make up the protein. Our understanding of the sequence-structure relationship has recently reached the point that we can design new proteins *de novo* (i.e. without reference to existing protein sequences). However, this understanding is only partially encoded in protein design software, which still requires a user with considerable expertise in protein engineering. Here, I use citizen science to identify and resolve limitations of protein design software, by crowdsourcing protein design tasks to non-experts playing the computer game Foldit. Using the output of Foldit players as feedback, I iteratively trialed and improved protein design software to the point that non-experts can now use the software to successfully design proteins from scratch. This work reveals implicit assumptions of expert protein engineers, corrects errors in the Rosetta protein structure energy function, and shows how citizen science can be used to improve a scientific model.

Table of Contents

I. Introduction

Protein structure

Proteins are a class of molecule found in all forms of life, and are essential to virtually every biological process. A single organism can express 10^4-10^5 different proteins, each of which plays a specific biological role, from molecular assembly to metabolism to signal transduction. Proteins achieve a wide diversity of functions by adopting very specific 3-dimensional structures with chemical groups arranged in precise positions and orientations. This precise positioning of chemical groups allows highly specific interactions with other molecules based on their atomic structure.

At a high level, protein structures can be categorized into *folds*, which describe the overall architecture of a protein backbone. According to the SCOPe protein classification system[1], there are over 1000 unique protein folds currently represented among all known protein structures in the Protein Databank (PDB)[2]. However, the PDB contains only a subset of naturally-occuring folds (estimated at 10^5 [ref 3]), and the set of all possible folds is expected to be greater still[3].

Despite the structural complexity of proteins, the construction of a protein is relatively simple. A protein is a poly-peptide—an unbranched polymer of amino acid units, ranging in length from 10^1 to 10^4 amino acids. There are only 20 canonical amino acids, but the number of sequence permutations scales exponentially with protein length, to yield a vast space of possible amino acid sequences. For example, for a protein of 65 amino acids (a small protein) there are 20^{65} (about 10^{84}) possible sequences—for reference, the number of electrons in the observable universe is estimated at only 10^{80}.

The 3-dimensional shape of a protein is dictated entirely by its sequence of amino acids. Each of the 20 amino acids has different chemical properties, and the amino acids of a poly-peptide chain will interact with one another and with the surrounding aqueous solvent, according to those properties. Relative to the unfolded state, these interactions can have favorable (negative) or disfavorable (positive) energy, depending on their precise atomic arrangement and local environment. A protein will naturally tend to fold into the structure with the lowest energy[4].

However, there is an entropic barrier to protein folding, due to the degrees of freedom that are constrained when a flexible poly-peptide is folded into a rigid structure. A poly-peptide will only fold if this loss of entropy can be overcome by the interactions in the folded structure (including entropy gained by solvent due to the hydrophobic effect). The vast majority of amino acid permutations yield poly-peptides that are disordered in solution, and only a miniscule subset of possible amino acid sequences encode well-folded proteins.

Protein modeling

Since the function of a protein arises from its 3-dimensional structure, the structure of proteins is essential for understanding biological processes at the molecular level. Unfortunately, the 3-dimensional structure of a protein is very difficult to observe experimentally. (The structure of a protein is resolved at about 0.1 nm, so it cannot be observed by light microscopy with wavelengths on the order of 100 nm.) However, due to advances in genome sequencing, the amino acid sequences of many proteins are easily obtained. And because the fold of a protein is encoded in its amino acid sequence, in theory it should be possible to predict the 3-dimensional structure of a protein from its sequence alone.

Given that a protein naturally folds to its lowest-energy conformation, protein structure prediction is therefore an optimization problem. The solution domain consists of all possible protein conformations; the objective function to be optimized is the energy of the protein conformation. This problem is often discussed in terms of an "energy landscape," which spans the domain of all possible conformations and must be searched to identify the lowest-energy conformation (i.e. the global minimum of the energy landscape).

Protein structure prediction requires two components:

1. Energy calculations: An energy function is needed to calculate the potential energy of a protein conformation.
2. Sampling strategy: A method is needed to search the domain of possible protein conformations.

1. Energy calculations

It is possible to calculate the potential energy of a molecule from physical "first principles," using quantum mechanics theory. However, quantum energy calculations are computationally expensive even for small molecules (10^1 atoms), and intractable for large macromolecules such as proteins (10^{3+} atoms)[5].

Instead, protein energy functions typically use molecular mechanics theory, with a combination of energy terms to account for different forces in protein folding (e.g. van der Waals repulsion, electrostatics, hydrogen bonding, etc.)[6]:

$$E_{total} = \sum E_{vdw} + E_{elec} + E_{Hbond} \ldots$$

These energy terms are often derived from empirical observations. For example, certain phi/psi torsions in the protein backbone are energetically unfavorable and rarely observed. Rather than calculate torsional energies from first principles, it is more convenient to measure the statistical distribution of phi/psi torsions in known protein structures[7], and fit an energy function that describes the observed distribution.

In modern protein energy functions, like the Rosetta energy function[6] and others[8–10], the parameters of these energy terms are fit to reproduce empirical observations, and are sufficiently accurate for most protein modeling tasks[11].

2. Sampling strategy

An ideal sampling strategy would explore every possible protein conformation for a given sequence. However, due to the large number of degrees of freedom in a protein structure (i.e. two torsions for every amino acid), this is computationally intractable.

Therefore it is necessary to reduce the number of possible protein conformations to a subset of "most-likely" conformations. The most effective sampling strategies use fragments of observed proteins to construct protein-like models[12]. Using many computers in parallel[13], protein-like conformations can often be sampled in sufficiently large numbers (10^5) to yield high-confidence structure predictions.

> Sampling is currently the limiting factor in protein structure prediction. That is, when structure prediction fails, it usually fails because the native configuration was not sampled—*not* because the energy of the native structure was incorrectly calculated[14].

The accuracy of a predicted structure cannot be confirmed without an experimentally determined, high-resolution structure of the protein. Unfortunately, experimental structure determination is difficult and time-consuming, and successful structure determination is not guaranteed from an experiment. X-ray crystallography typically produces the highest resolution data, but protein crystallization is notoriously temperamental, and many proteins fail to crystallize at all. Other experiments, like NMR spectroscopy and cryo-electron microscopy are more assured to produce structural data at lower resolution, but cannot be applied to all proteins. Nevertheless, protein structure prediction can be useful for developing hypotheses about a protein's function, even if the accuracy of the prediction is unknown.

Protein design

The problem of protein design is related to the problem of protein structure prediction, and is sometimes referred to as the "inverse folding problem"[15,16]: given a target protein structure, a protein designer wishes to find an amino acid sequence encoding that structure. As the inverse folding problem, the solution domain comprises protein sequences rather than protein conformations. Yet, the same energy functions used for protein structure prediction can be used to optimize the energy of the target structure, with only slight modification[17].

However, protein design is confounded by the need to consider alternative low-energy states for a designed sequence. This is not an issue in structure prediction, where we can typically assume a biological protein sequence folds with high specificity—into a single, well-defined structure. A designed protein sequence, on the other hand, does not guarantee folding specificity, and a protein engineer must take care that a designed sequence does not encode off-target conformations (decoy states).

A designed protein sequence must therefore have an energy landscape where the target conformation is lower in energy than all decoy conformations. Decoy states can include misfolded states (in which the protein folds into an off-target low-energy conformation), and aggregation states (in which multiple copies of a protein interact with lower energy than the designed state).

In 1992, Yue and Dill used a simplified lattice model for protein design to show how optimizing the energy of the target state can produce sequences with high degeneracy (i.e. sequences that are compatible with many low-energy states)[15]. Although there has been some success with such a protein design strategy[18,19], in other cases this has led to poorly-structured "molten globule" proteins (indicating exchange between multiple low-energy conformations)[20,21], or misfolding into an off-target structure[22].

This suggests that protein design is not as simple as optimizing the potential energy of the designed state (so-called "positive design"), and that a designed sequence should be modified to disfavor competing decoy states ("negative design"). In theory, this can be accomplished by explicitly considering decoy conformations, and choosing sequences with greater preference for the design conformation over all decoy conformations. In practice, however, explicit negative design is frustrated by the inability to enumerate all relevant decoy conformations, and can only be carried out for a limited number of off-target states[23,24].

More frequently, protein engineers carry out implicit negative design by applying heuristics intended to confer folding specificity. This can include: incorporating sequences known to favor or disfavor secondary structures[25,26]; restricting overall amino acid composition[16]; restricting amino acids in regions[27–29]; and restricting amino acids at privileged positions[30–32].

There are numerous protein design algorithms for optimizing the energy of the designed state[18,33–37]. But the application of implicit negative design heuristics varies wildly, and often relies on the subjective judgment of human protein engineers with considerable expertise[28,32,38,39]. It is difficult to determine the exact role of this expertise, and which implicit design heuristics—if any—are necessary for successful protein design.

The role of expert protein engineers is further implicated in the choice of target structure. Despite the number of possible protein folds, most protein design efforts have focused on the redesign of natural protein backbones[18,31,40] or the generation of backbones that resemble well-studied natural protein folds[27,28,32,39,41]. In over 30 years of protein design research, there is only one instance of a designed fold that is unlike any natural protein (the Top7 protein designed by Kuhlman et al. is composed of a unique arrangement of α-helices and β-sheets)[19]. This raises the concern that protein design software may be "over-fit" to design only native-like protein backbones.

Citizen science

Citizen science engages volunteers from the general public to assist scientific research. Historically, amateurs and enthusiasts have contributed observations in easily-accessible fields like bird-watching and astronomy[42,43]. However, the internet age has allowed researchers to recruit citizens to help research in more niche fields (like genome analysis[44]), and with more engaging research tasks (like RNA design[45]).

Today, there are hundreds of citizen science projects spanning the full breadth of research fields[46,47]. The large majority of these citizen science projects draw on volunteers for simple, "bite-sized" tasks that demand little or no critical thinking[48]. Many projects rely on human volunteers for rote pattern-recognition tasks that are difficult for computational algorithms, such as classification of images or audio samples[49,50]. In some cases, citizen scientists' output can be used to train computational algorithms[51], although these efforts suffer from the "opacity" problem common to many machine learning techniques[52], and provide little insight about the precise advantages of citizen scientists' contributions.

Citizen science projects that demand critical thinking or creativity are less common[53–56]. As two notable exceptions, EteRNA and Quantum Moves have drawn on the human ingenuity of citizen scientists to discover new problem-solving strategies, which ultimately could be incorporated into improved scientific models of RNA design[45] and quantum computing[57].

Foldit

Foldit is a citizen science online game in which participants help to predict protein structures, and was launched in 2008 by the UW Center for Game Science in collaboration with the Baker Lab in the UW Biochemistry Department[58].

Foldit players are presented with a virtual model of a protein structure. Players can manipulate the 3-dimensional structure of the protein model, and compete with one another to find the lowest-energy

structure. The energy of a player's model is reflected in the Foldit score (after applying a negative multiplier so that low, negative energies are displayed as high, positive scores). The score is updated in real-time, giving the player immediate feedback about the quality of their model as they continue to manipulate its structure.

The Foldit software includes basic tools for manipulating the model, such as simple click-and-drag to "pull" on segments of the protein. Players have access to basic Rosetta protocols for optimizing protein structure, including LBFGS minimization and rotamer packing optimization[59]. Players can also write their own algorithms in the form of Lua macros, which can be edited and distributed to other players[60].

Foldit supports social interactions between players, with in-game chat and online forums. Players can collaborate in groups to develop models together, and some group players will specialize in modeling subtasks—for example, working mainly on initial model construction before passing the model off to a teammate for refinement[61].

Since its launch in 2008, Foldit players have contributed to a number of successes in protein modeling. In 2011, Foldit players correctly predicted the structure of a retroviral protease from the Mason-Pfizer monkey virus[61]. The Foldit players' model was accurate enough to solve a crystal structure of the protein that had eluded crystallographers for over a decade. Foldit players have also innovated effective algorithms for optimizing protein models[60], and in 2012, Foldit players helped to redesign the active site of a designed Diels-alderase, leading to an 18-fold increase in enzyme activity[62].

Outline

Given the recent citizen science-driven discoveries of EteRNA and Quantum Moves, and the proven track record of Foldit players, we resolved to challenge Foldit players with the problem of *de novo* protein design.

Since Foldit players lack protein design expertise, we hypothesized that the designs created by Foldit players (using the underlying Rosetta software) would reveal any expertise-related factors contributing to previous success of protein design. The creativity of citizen scientists is especially appropriate for exploring the vast domain of protein backbones for viable design targets.

We proposed to use an iterative process to test individual hypotheses about protein design expertise, with the ultimate aim of improving protein design software by removing the requirement of user expertise. Compared with more sophisticated methodologies for model improvement (e.g. convolutional neural networks), where the underlying features may be opaque or difficult to interpret[52,63], our hypothesis-driven approach offers advantages of transparency. By iteratively exposing and correcting software limitations in a piecewise fashion, we can identify and address specific factors of "expertise" that are necessary for successful protein design.

In Section II, we describe our methodology and how we use the Foldit program to crowdsource protein design. We identify three separate problems with proteins designed by citizen scientists, which reveal implicit assumptions of expert protein designers. We devise and implement rules in Foldit to address these problems, and show that these rules guide Foldit players to better protein designs. Foldit player-designed proteins are characterized experimentally, and are found to adopt monomeric, stable structures in accordance with the design models.

In Section III, we seek to expand the structural diversity of proteins designed by Foldit players. We discover that protein designs by Foldit players have significantly worse backbone geometry than proteins designed by automated methods, revealing an error in the Rosetta protein energy function. We correct this error, develop new backbone-modeling tools for Foldit, and introduce a new rule to encourage the design of stable backbone conformations. With these improvements, experimental characterization of Foldit player designs shows a drastic increase in design success rate. High-resolution experimental data confirms the accuracy of Foldit players designs, and an analysis of Foldit player design methods suggests Foldit players use much different protein design strategies than automated methods.

In Appendix A, we use Foldit to crowdsource the task of protein model-building into an electron density map. Foldit players are provided with an electron density map of an experimentally-phased x-ray diffraction dataset, and challenged to build a protein model that matches the density. Compared with a model built by expert crystallographers, Foldit players develop a model with a better fit to the experimental data and with fewer structural outliers. Although this work is not directly applicable to protein design, these results further demonstrate the ability of citizen scientists to develop physically-realistic protein models using Foldit.

In conclusion, our results suggest the predominant role of expertise in de novo protein design is in the choice of protein backbone design target. This was revealed by Foldit players' repeated discovery of low-energy design models with unrealistic protein backbones. When we introduce rules guiding Foldit players away from unrealistic backbone targets, and correct the energy function to properly penalize strained backbones, then non-expert citizen scientists can successfully design a wide diversity of protein folds, with sequences that fold stably and accurately to the designed conformation.

The following sections II and III comprise material that has been submitted for publication as:
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, Toke Nørgård-Solano, Foldit Players, Gaetano T. Montelione, Frank DiMaio, Zoran Popovic, Firas Khatib, Seth Cooper, David Baker

Appendix A has previously been published as:
Horowitz S, Koepnick B, Martin R, Tymieniecki A, Winburn AA, Cooper S, Flatten J, Rogawski DS, Koropatkin NM, Hailu TT, Jain N, Koldewey P, Ahlstrom LS, Chapman MR, Sikkema AP, Skiba MA, Maloney FP, Beinlich FR, Foldit Players, UMich students, Popovic Z, Baker D, Khatib F, Bardwell JC. **Determining crystal structures through crowdsourcing and coursework.** *Nat. Commun.* 7, 12549 (2016).

II. Identifying limitations in protein design software using the work of citizen scientists

Abstract

Online citizen science projects such as GalaxyZoo[49], Eyewire[64] and Phylo[44] 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[45], 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[58]. 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 several iterations of player design, analysis of the top scoring solutions, and subsequent game improvement, Foldit players can now, starting from an extended polypeptide chain, design well-folded proteins with sequences that encode them. Foldit player-designed proteins were encoded in custom genes and expressed in *E. coli*, and were found to be soluble, monomeric, and structured in solution.

Introduction

The principle underlying *de novo* protein design is that proteins fold to their lowest free energy state[4]; 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[28,29,39,41], 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.

Protein design in Foldit

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[58] (**Figure 2.1**). Players compete to build a model

with the lowest free energy, as calculated by the Rosetta energy function[6]. 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[58,61]. Foldit players in one case redesigned a loop region of an already folded structure[62], 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. As in structure prediction challenges, Foldit player rankings were determined by the Rosetta energy of their models.

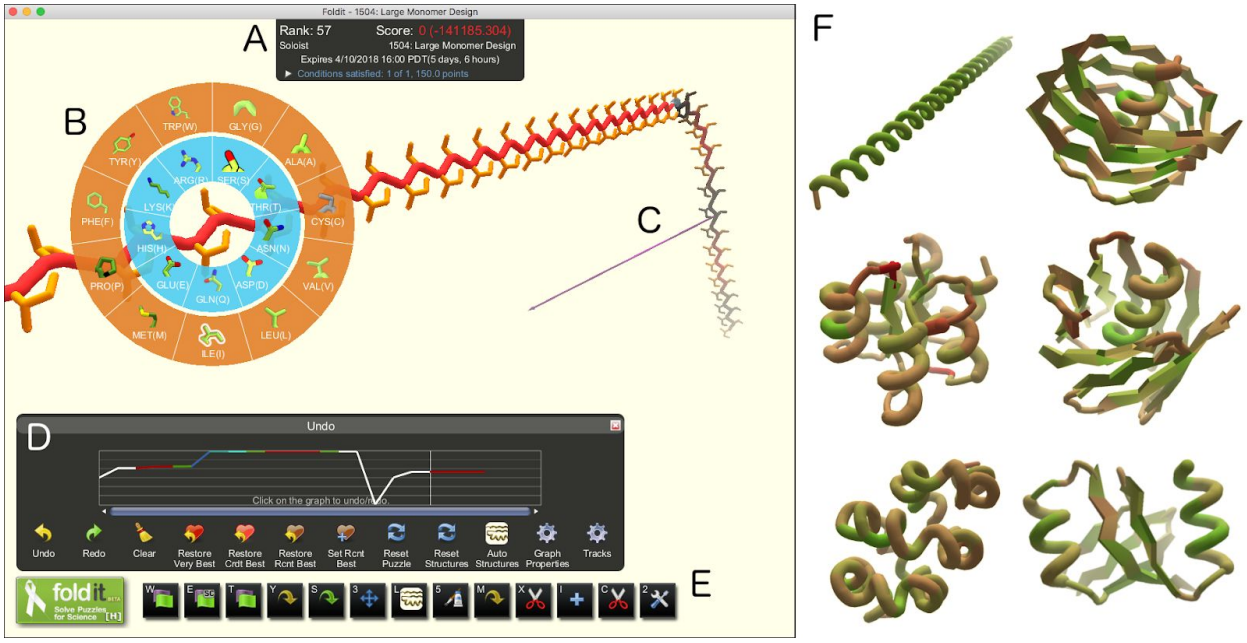


Figure 2.1. The Foldit user interface. (A) The Foldit score is the Rosetta energy with a negative multiplier, so that better models yield higher scores. (B) The design palette allows players to change the amino-acid residue identity at any position of the model. (C) The Pull tool allows players to manipulate 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.

Additional protein design rules

Core Existence Rule

Initial protein designs by Foldit players lacked a solvent inaccessible core and were composed entirely of polar residues (**Figure 2.2**). A significant driver of folding in native proteins is the hydrophobic effect[65], which is thought to result from the entropic cost of solvating a non-polar surface[66,67]. If the folded and unfolded states present the same amount of non-polar surface area, then the benefits of the hydrophobic effect are lost. However, the extended, fully α-helical Foldit designs have more favorable hydrogen bonding, electrostatic, and local torsional energies than collapsed structures (which must contort to create a buried core) and are therefore favored by the Rosetta energy function. Early studies of protein structure showed that poly-lysine and other extended polar sequences resembling these initial Foldit solutions are indeed largely α-helical in solution[68,69], although the lack of long-range interactions precludes specific folding into a single stable structure[70].

We introduced the "Core Existence" rule in Foldit, requiring a minimum 30% of residues to be buried in the designed structure. The threshold of 30% was determined by examining the proportion of buried residues in other successful de novo designed proteins[27,29]. A residue was designated as buried using the "sidechain neighbors" method, which counts the number of neighboring residues in the area around its sidechain[71]. Initially, we tried measuring "buriedness" with direct calculations of solvent-accessible surface area, but these calculations proved too slow for responsive gameplay (the Foldit model must be re-evaluated in real time as a player manipulates the structure, at 30 frames per second).

Secondary Structure Design Rule

After introducing the Core Existence rule, we observed excessive amounts of alanine and glycine in the core of top-ranked Foldit designs (**Figure 2.2**). Foldit players had discovered that an alanine- and glycine-saturated core allows close packing of the protein backbone atoms, with very favorable van der Waals energies and a competitive Foldit score. However, alanine-saturated protein cores are problematic for two reasons. First, in the unfolded state alanine presents much less surface area than larger hydrophobic residues, and hence there is less driving force for folding associated with the hydrophobic effect[66]. Second, interdigitation of sidechains is thought to be important for specific folding of the protein core[72,73]. In helical bundles especially, sidechain interdigitation prevents "sliding" along the z-axis[74]. Glycine is furthermore undesirable rigid secondary structure because it incurs a greater entropic penalty than other amino acids during folding.

Since this appeared problematic mainly when alanine and glycine residues were regularly-spaced in secondary structure elements, we introduced the "Secondary Structure Design" rule to prohibit alanine and glycine from alpha-helices and beta-sheets (as determined by DSSP[75]). This rule is poorly reflected in natural proteins, where alanine is commonly found (in moderation) in α-helices[76], and glycine is often necessary to relieve backbone strain in highly-curved β-sheets[39]. However, the rule is fast to evaluate, which is important for responsive gameplay. And, unlike a global requirement to limit amino acid composition (i.e.. capping the alanine or glycine content of the entire model), this rule is pairwise-decomposable, which is a prerequisite for the automated rotamer packing algorithm available to Foldit players.

Residue Interaction Energy Rule

We also observed that many high-scoring Foldit designs contained numerous large, aromatic residues making insufficient packing interactions (**Figure 2.2**). Because of their large size, these residues are expected to make more interactions in off-target decoy states (e.g. aggregation states), so they should be especially well-packed in the designed state. However, even when under-packed or solvent-exposed,

such large residues can make more interactions than smaller aliphatic residues and are favored when optimizing the energy of the design state.

The “Residue Interaction Energy” rule was introduced to penalize PHE, TYR, and TRP residues that do not make enough interactions in the design structure. The threshold for each residue type was determined from mean interaction energy of residues in native protein structures.

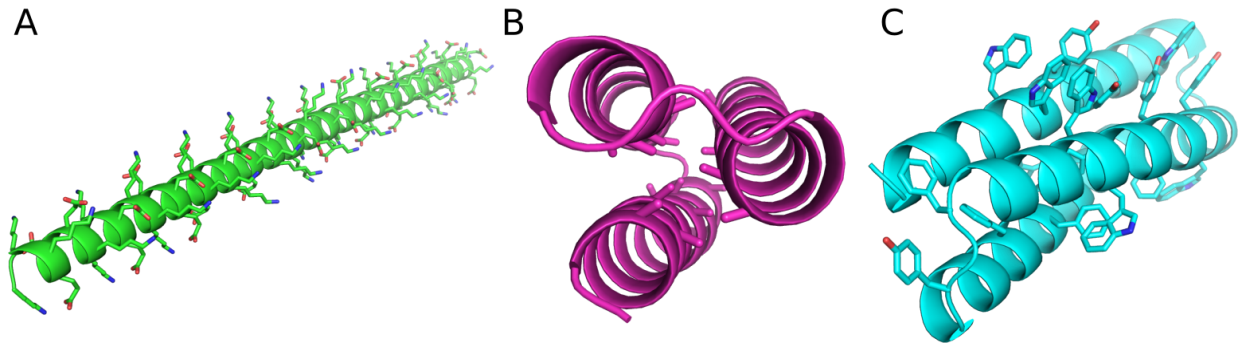


Figure 2.2. Initial top-ranking Foldit player designs. When challenged to design a protein with only the talaris2013 score function (and no additional rules), Foldit players discovered low-energy models that are unlikely to fold as designed. (A) An extended α-helix, composed entirely of lysine and glutamate, has very favorable energies for hydrogen-bonding, electrostatic, and backbone torsions, but is unlikely to fold cooperatively into a single stable structure. These designs are discouraged with the “Core Exists” rule. (B) A design with an alanine- and glycine-saturated core can make favorable 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 (i.e. “molten globule” behavior). These designs are discouraged with the “Secondary Structure Design” rule. (C) Due to their large size, even under-packed or solvent-exposed aromatic sidechains can make more interactions than smaller aliphatic sidechains. These designs are discouraged with the “Residue Interaction Energy” rule.

Experimental characterization of Foldit player designs (Round 1)

With the addition of these three rules to Foldit, subsequent top-scoring designs from Foldit players resembled compact, solvent-excluding folds visually indistinguishable from expert designs. We obtained custom synthetic genes encoding 12 player designs for which structure prediction calculations converged on the player designed conformation[77]. All designs were either 3- or 4-helix bundles, and the sequences of these proteins have no homology to any known protein (**Table 1**).

The designs were expressed in *E. coli*, purified, and characterized by size exclusion chromatography (SEC) and circular dichroism (CD) (**Figure 2.3**). All twelve proteins were expressed solubly; seven elute from SEC as monomers, while the others elute as higher-order oligomers or aggregates. Six of the monomeric designs have CD spectra reflecting the expected secondary structure content. Several of the proteins are extremely thermo-stable, and do not unfold when heated to 95ºC. We carried out chemical denaturation experiments with titration of guanidinium chloride, and four proteins show sigmoidal denaturation indicative of cooperative, two-state unfolding. Fitting a two-state unfolding model to the titration data[78], we calculated unfolding free energies of up to 17.8 kcal/mol.

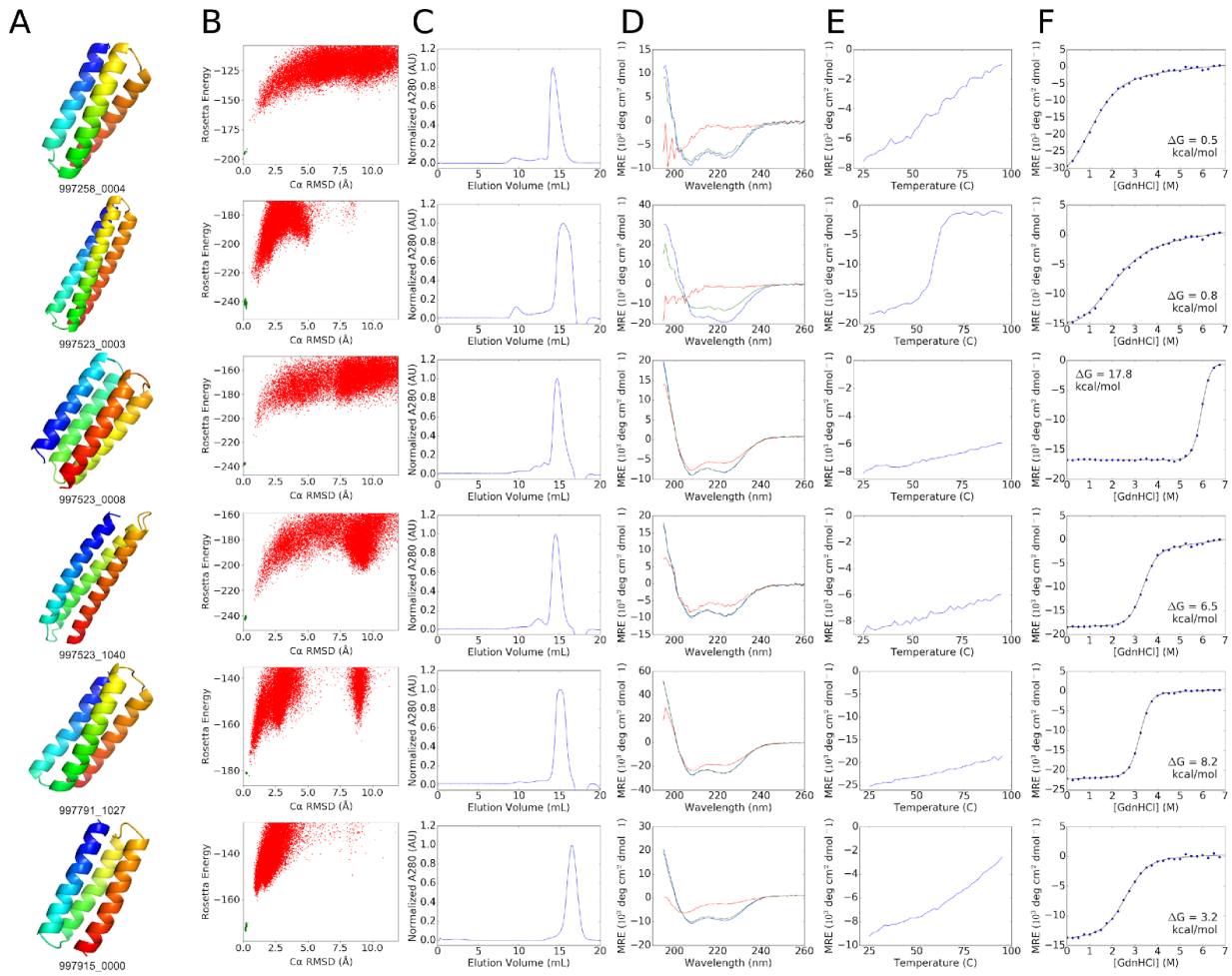


Figure 2.3. **Experimental characterization of successful protein designs by Foldit players.** (A) Cartoon depiction of Foldit design backbones. (B) Rosetta@home *ab initio* calculations show that the sequence for each design has an energy landscape that is strongly funneled toward the design structure. 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). (C) Size-exclusion chromatography (SEC) traces of elution absorbance at 280 nm show that designs are monomeric in solution. (D) Circular dichroism (CD) spectra indicate that the designs are predominantly α-helical in solution at 25ºC (blue trace). Some secondary structure is lost when heated to 95ºC (red trace), but most of these proteins refold completely 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[78].

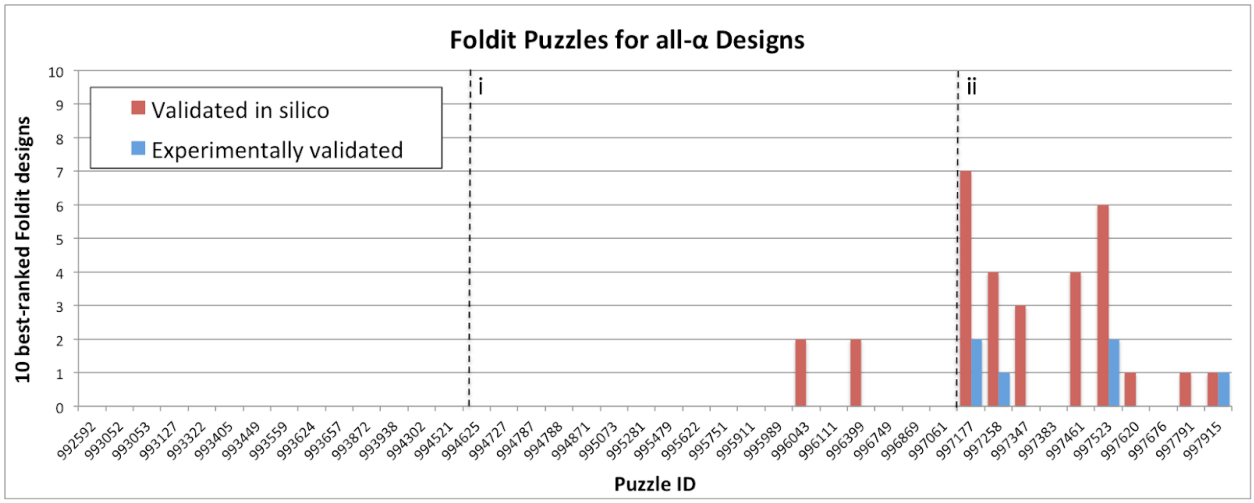


Figure 2.4 Chronology of Foldit design puzzles. Considering only the best-scoring design from the 10 top-ranked groups in each Foldit puzzle, bars show the number of designs validated in silico (i.e. Rosetta structure prediction converges on the designed structured; red), and the number of designs that were experimentally validated (blue). The dashed lines mark (i) the introduction of the "Core Existence" and "Residue Interaction Energy" rules, and (ii) the introduction of the "Secondary Structure Design" rule.

Discussion

Citizen scientists lack the expertise of formally trained scientists. Their solutions to the protein design problem revealed key elements of protein engineering that are not encoded in protein design software, but reside as implicit knowledge in trained protein engineers. We identified three rules that should be enforced when designing proteins with the Rosetta energy function:

1. **A protein should have a non-polar, solvent-inaccessible core.** The burial of hydrophobic residues is critical for decreasing the energy of the design conformation relative to the fully-solvated, unfolded state. Extensive tertiary (non-local) interactions in a globular fold help to ensure folding specificity.
2. **The protein core should be packed with interdigitated sidechains.** Small, poorly-interdigitated sidechains allow degenerate packing interactions. Close packing of large, interdigitated sidechains help to ensure that there is a single, preferred configuration for sidechains in the protein core.
3. **Large residues should make sufficiently strong interactions.** Large sidechains can make more interactions than smaller amino acids, and strict optimization of the design conformation can lead to under-packed aromatic sidechains that are likely to favor alternative states.

When these rules are encoded in the Foldit protein design software, players are guided toward designs that resemble those of expert protein engineers. Experimental testing of Foldit player designs suggests that they are monomeric and well-folded in solution, and likely adopt the intended conformation.

Our protein design rules emphasize a limitation of using absolute energy as the sole 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. Foldit players were very effective at optimizing the

energy of their designs. However, this optimization resulted in unrealistic protein models, with designed sequences that are expected to permit other conformations with similarly low energies.

Although these principles may be obvious to a trained structural biologist, it was not previously clear that they were unaccounted for in the Rosetta energy function. The first two rules in particular are related to the choice of target conformation, and were realized only when Foldit players explored regions of the solution domain (i.e. fully-extended backbones and super-compact backbones) that were not previously considered by professional protein engineers. This result highlights the advantages of citizen scientist creativity for solving problems with large solution domains.

Materials and Methods

Foldit protein design puzzles

Foldit puzzles used the talaris2013_cart scorefunction with the following modifications: (1) the cart_bonded scoreterm was upweighted 4x to ensure realistic bond lengths and angles as players cut and splice the backbone chain; (2) a penalty-only envsmooth scoreterm 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 increased 5x to discourage the excessive design of alanine, which has been problematic in previous Rosetta design efforts[27,29]. 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, as 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 this research 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. At each step, we used TM-align[26] to eliminate duplicate models (TM-score > 0.98) that had already been selected (e.g. models that were top-ranking *and* flagged by players). In Rounds 2 and 3, the top-ranked group and solo models were automatically selected for further analysis, without visual inspection. The sequences of selected models were subjected to Rosetta *ab initio* structure prediction[77], using the distributed computing platform Rosetta@home[13]. 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.

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. 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. 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 size exclusion chromatography (SEC) 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 determined from SDS-PAGE and mass spectrometry. Oligomeric state was determined by size exclusion chromatography.

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[78].

III. Increasing structural diversity in proteins designed by citizen scientists

Abstract

Natural proteins adopt a variety of distinct folds, each of which has different properties and characteristics. We previously showed that, when challenged to design proteins from scratch, citizen scientist Foldit players are capable of designing well-folded 3- and 4-helix bundles. To reap the full benefits of crowdsourced creativity, we sought to expand the structural diversity of Foldit player designed proteins by encouraging the design of more diverse and difficult protein folds. Using the same approach of iterative trial and improvement, we analyzed player designs to identify shortcomings in Foldit related to protein backbone modeling, and then modified the Foldit software to address these shortcomings, which led to much-improved Foldit player designs. 134 Foldit player designs with sequences unrelated to naturally occurring proteins were encoded in synthetic genes; 50 were found to be expressed in *E. coli* with good solubility and to adopt stable monomeric folded structures in solution. Foldit player designs span 20 distinct folds, a diversity that is unprecedented in *de novo* protein design. One design represents a new fold that is 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 shows that citizen scientists can discover creative new solutions to outstanding scientific challenges, such as the protein design problem.

Introduction

Natural proteins adopt a diversity of folds. The SCOPe database has catalogued over 1000 distinct protein folds among known protein structures[2]. This structural diversity is key to the functional diversity of proteins, and different protein folds appear to be better-suited for different functions[79–82]. A recent report by Golinski et al. suggests that proteins of different folds are also more- or less-evolvable in a laboratory setting[83]. The ability to design diverse protein folds will be important for the success of functional protein design in the future.

Encouraged by the success of Foldit players in designing stable proteins from scratch, we sought 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 on par with α-helical bundles (**Figure 3.1**), 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[24,84–86]. Folds with β-sheets are often considered more difficult design targets, because they are stabilized more by long-range tertiary interactions[31,40,87], which expands the number of relevant decoy states that must be considered. β-rich proteins are especially prone to sheet edge-to-edge aggregation and amyloid misfolding[88]. However, proteins with β-sheets may be better suited for some peptide- and protein-binding modes, making them attractive targets for functional protein design[89–91].

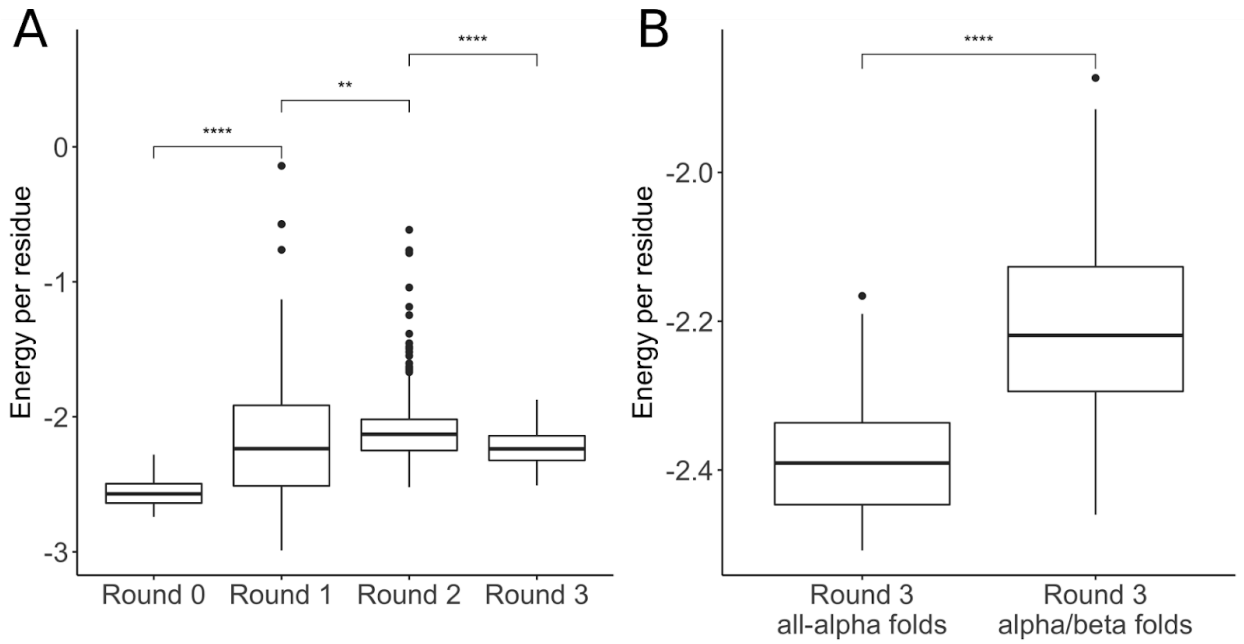


Figure 3.1. Rosetta energy of top Foldit player designs. Rosetta energy of top-ranking designs was calculated with the talaris2013 score function and normalized by residue count. (A) Energy of top 10-ranked designs from: initial Foldit puzzles (Round 0; *n* = 30), Round 1 puzzles (*n* = 170), Round 2 puzzles (*n* = 510), Round 3 puzzles (*n* = 250). The introduction of supplementary rules in Round 1 and Round 2 resulted in higher-energy designs ($p < 1e\text{-}6$ and $p < 0.01$, respectively; Wilcoxon rank-sum test). The backbone modeling improvements in Round 3 resulted in lower-energy designs ($p < 1e\text{-}15$; Wilcoxon rank-sum test). (B) Energy of top 10-ranked designs from Round 3 all-α puzzles (*n* = 30) or α/β puzzles using the Secondary Structure rule (*n* = 220). All-α designs tend to have lower energy than α/β designs ($p < 1e\text{-}10$; Wilcoxon rank-sum test). Boxplots show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

Foldit player-designed α/β proteins

Secondary Structure Rule

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, as determined by DSSP[75]. Foldit players responded by designing a multitude of mixed α/β proteins (**Table 2**), which were indistinguishable from expert designs upon visual inspection.

However, Rosetta structure prediction calculations for initial 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. This is a feature of Rosetta's fragment-based sampling algorithm, which relies on short backbone fragments that are selected from known protein structures based on similarity to the query sequence and its predicted secondary structure. These short protein fragments are assembled stochastically in combination to form a library of complete, protein-like backbone conformations, which are then relaxed with side-chain atoms and scored[77]. If the local structure of a protein design is strongly encoded by its sequence, then the fragment selection algorithm is expected to find many protein fragments with similar structure, and fragment assembly is likely to reproduce the complete designed

structure. However, if the local structure is poorly encoded by the design sequence, then Rosetta will fail to select fragments with the intended structure; and, without accurate fragments, the subsequent fragment assembly step will fail to generate the complete designed structure.

Protein backbone geometry

Backbone torsion energy function

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[7,92]; **Figure 3.2**). 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 torsions; this is now standard in the latest Rosetta energy function[6].

Ideal Loops rule

The “Ideal Loops” rule was introduced together with the improved backbone score function and tools, in an effort to limit poor geometry in the protein backbone. This rule restricts players to a set of 19 turn conformations that are frequently observed in natural proteins[29]. The backbone torsions of each turn conformation are coarsely binned by Ramachandran plot quadrant, allowing players some flexibility in the exact backbone torsions and loop conformation. Initially, we attempted to use a large library of backbone fragments from native proteins, with very fine torsional binning. However, we could only include a fragment library of 4-residue fragments (4-mers) or smaller, since larger libraries (e.g. 9-mers) demanded too much computer memory. With this library, players were still able to build longer backbone conformations that were poorly represented in native proteins, even when the constituent 4-mers were well-represented in the library.

Added backbone modeling tools

We also incorporated new Foldit 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 (**Figure 3.3**). Together with the Rosetta energy correction and Ideal Loops Rule, these upgrades brought about a marked improvement in the local backbone quality of Foldit player-designed proteins (Round 3; **Figure 3.2**).

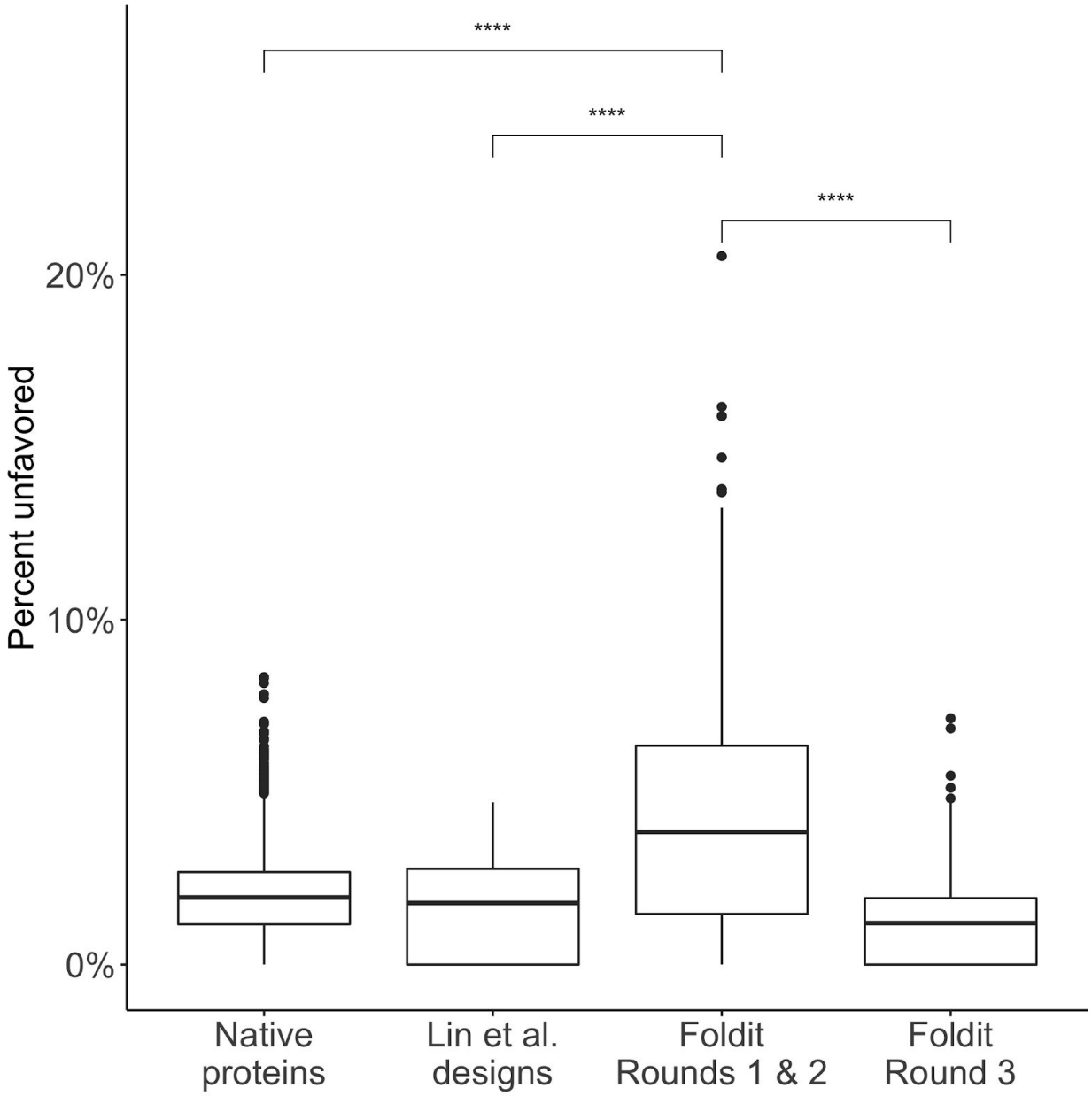


Figure 3.2. Improvement of backbone quality in Round 3 Foldit designs. MolProbity[92] was used to calculate the proportion of residues with "unfavored" or "outlier" backbone torsions in: high-resolution crystal structures of native proteins (*n* = 6342), *de novo* design models by Lin et al.[29] (*n* = 72), and top-ranking Foldit player-designs from before (*n* = 680) and after (*n* = 250) improvements to Foldit backbone modeling tools. Initial Foldit player designs contained significantly more unfavored torsions than native proteins or other *de novo* designs by Lin et al. ($p < 1e\text{-}15$, two-tailed t-test). Improvements to Foldit's backbone modeling tools led Foldit players to produce designs with fewer unfavored torsions ($p < 1e\text{-}15$, two-tailed t-test). Boxplots show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

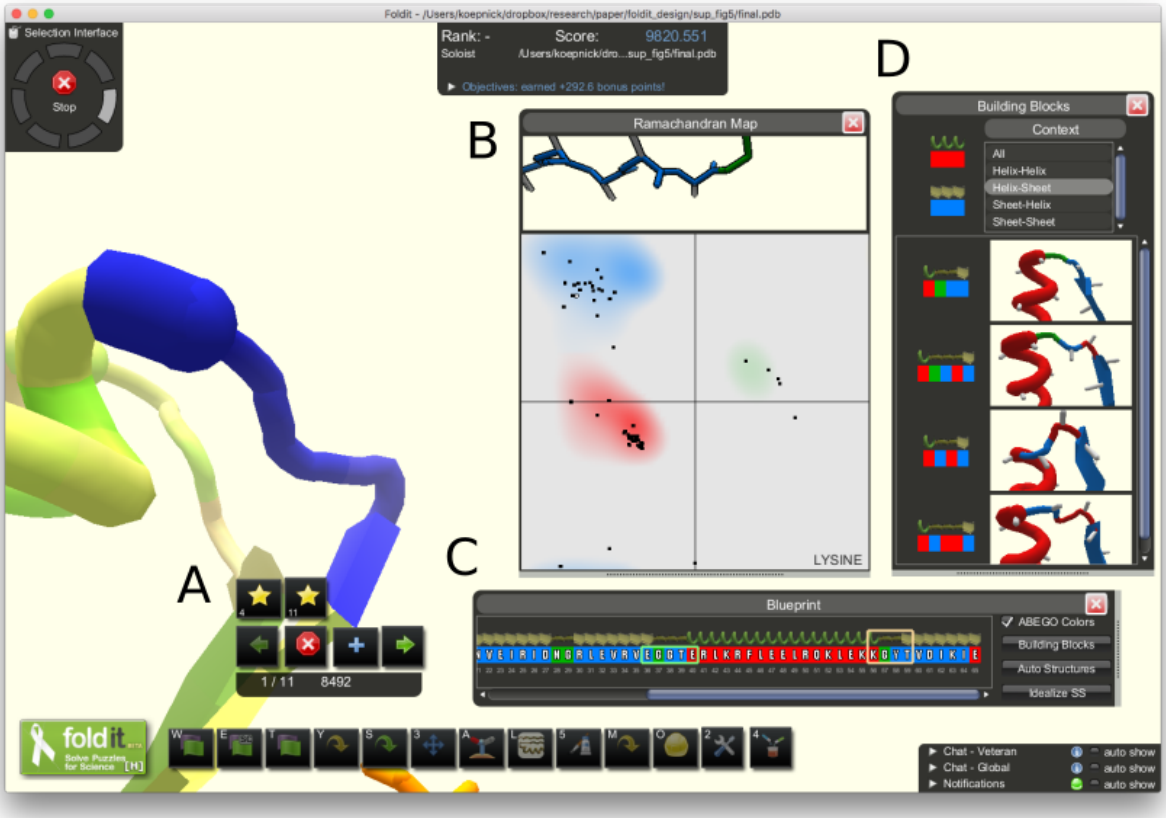


Figure 3.3. New backbone modeling tools in Foldit. (A) The Remix tool allows players to select a region of the model and search a library of backbone fragments for a conformation that can be substituted. (B) An interactive Ramachandran map allows players to easily identify residues with outlier backbone 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 colored according to the ABEGO quadrants of the Ramachandran plot[29]. (D) Players can drag-and-drop modular Building Blocks onto the Blueprint panel to insert common turn conformations into their model.

Experimental characterization of α/β protein designs

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 4 of 37 Foldit α/β designs tested (11%; Round 2) were monomeric and structured in solution. Following the backbone modeling additions, 46 of 97 (47%; Round 3) were monomeric and exhibited the expected secondary structure in solution (**Table 2**). Most showed exceptional stability in thermal and chemical denaturation experiments, with free energies of unfolding (ΔG_{unf}) approaching 20 kcal/mol; indeed, 32 designed proteins remained completely folded at 95ºC (**Figure 3.4; Supplementary Figure S2**). This success rate surpasses previous reports of designed α/β proteins[27,29] (**Table 2**). Along with the designs from Section II, the 56 successful Foldit designs are diverse in structure, representing 20 different protein folds (**Supplementary Figure S1**).

One Foldit player design (2003594_S028) represents a new fold that has not been observed in natural proteins. We used TM-align[93] and DALI[94] structure alignment tools to search the entire PDB for structural homologs of each protein (**Table 1**), and manually examined the top hits for proteins with the same fold. For most designs, we were able to find at least one protein or protein subdomain with the same backbone architecture. However, we found no existing structures in the PDB that contain fold number XX (**Figure 3.4**), with the same arrangement and connectivity of α-helices and β-strands.

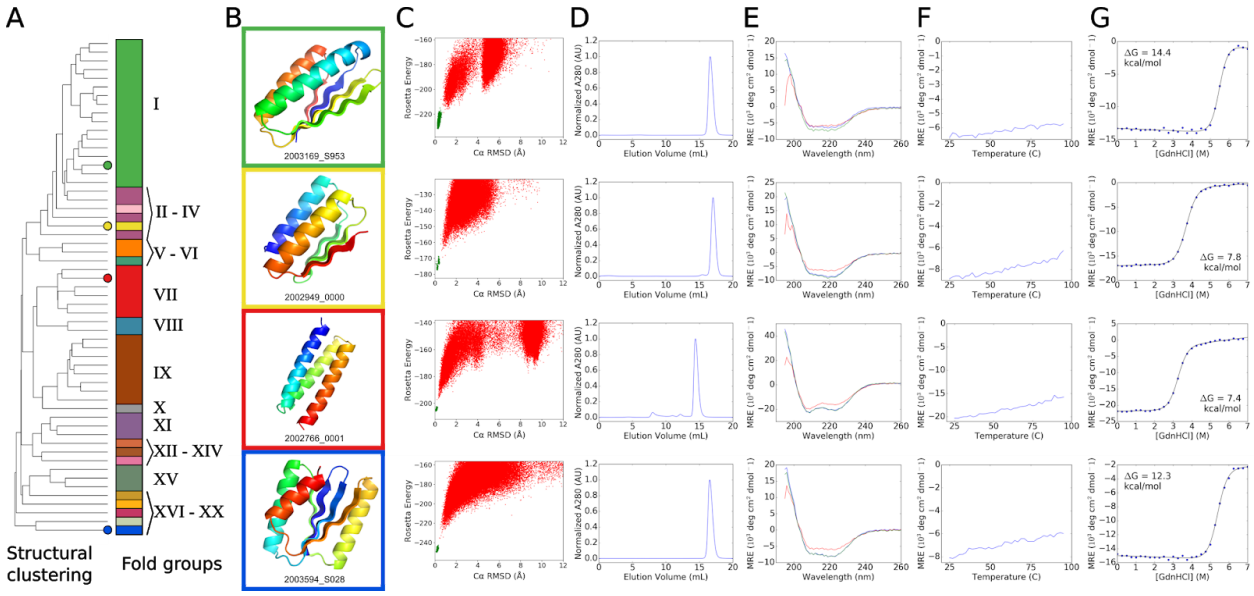


Figure 3.4. Structural characterization of Foldit player designed proteins. (A) Dendrogram showing all 56 folded Foldit player designs clustered by structural similarity (TM-align[93]), with colored circles highlighting the four designs characterized in (B-F). The stacked bars show the 20 different folds among the clustered designs (Extended Data Figure 3). 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) Size-exclusion chromatography (SEC) traces of elution absorbance at 280 nm show that designs are monomeric in solution. (E) 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). (F) CD mean residue ellipticity at 220 nm as temperature is increased from 25ºC to 95ºC; the designs do not temperature denature. (G) 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[78].

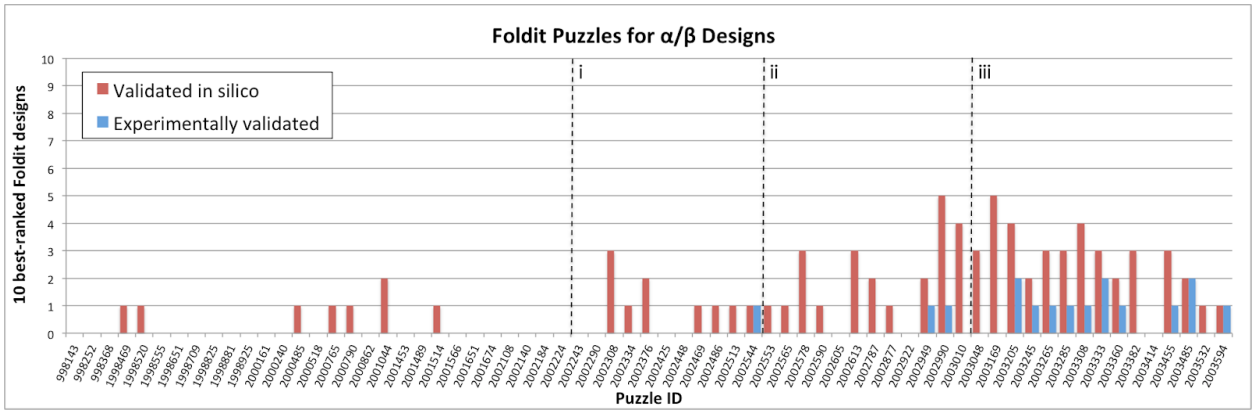


Figure 3.5 Chronology of Foldit puzzles for α/β designs. Considering only the best-scoring design from the 10 top-ranked groups in each Foldit puzzle, bars show the number of designs validated in silico (i.e. Rosetta structure prediction converges on the designed structured; red), and the number of designs that were experimentally validated (blue). The dashed lines mark (i) the correction of the Rosetta backbone torsion potential, (ii) the introduction of the "Ideal Loops" rule, and (iii) the introduction of new backbone modeling tools. Together, these improvements yielded a marked increase in the success rate of top-ranking Foldit designs.

High resolution protein structures

We succeeded in solving high-resolution structures of four Foldit player-designed proteins (**Figure 3.6**). X-ray crystal structures of three designed proteins (named by their designers Foldit1, Peak6, and Ferredog-Diesel) closely match the designed conformations, with Cα-RMSD of 1.1, 0.9, and 1.7 Å, respectively. 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 NMR structure of a fourth design, Foldit3, also closely matches the design conformation, with a Cα-RMSD of 1.1 Å between the design model and a representative structure (i.e., the medoid conformer[95]) of the ensemble.

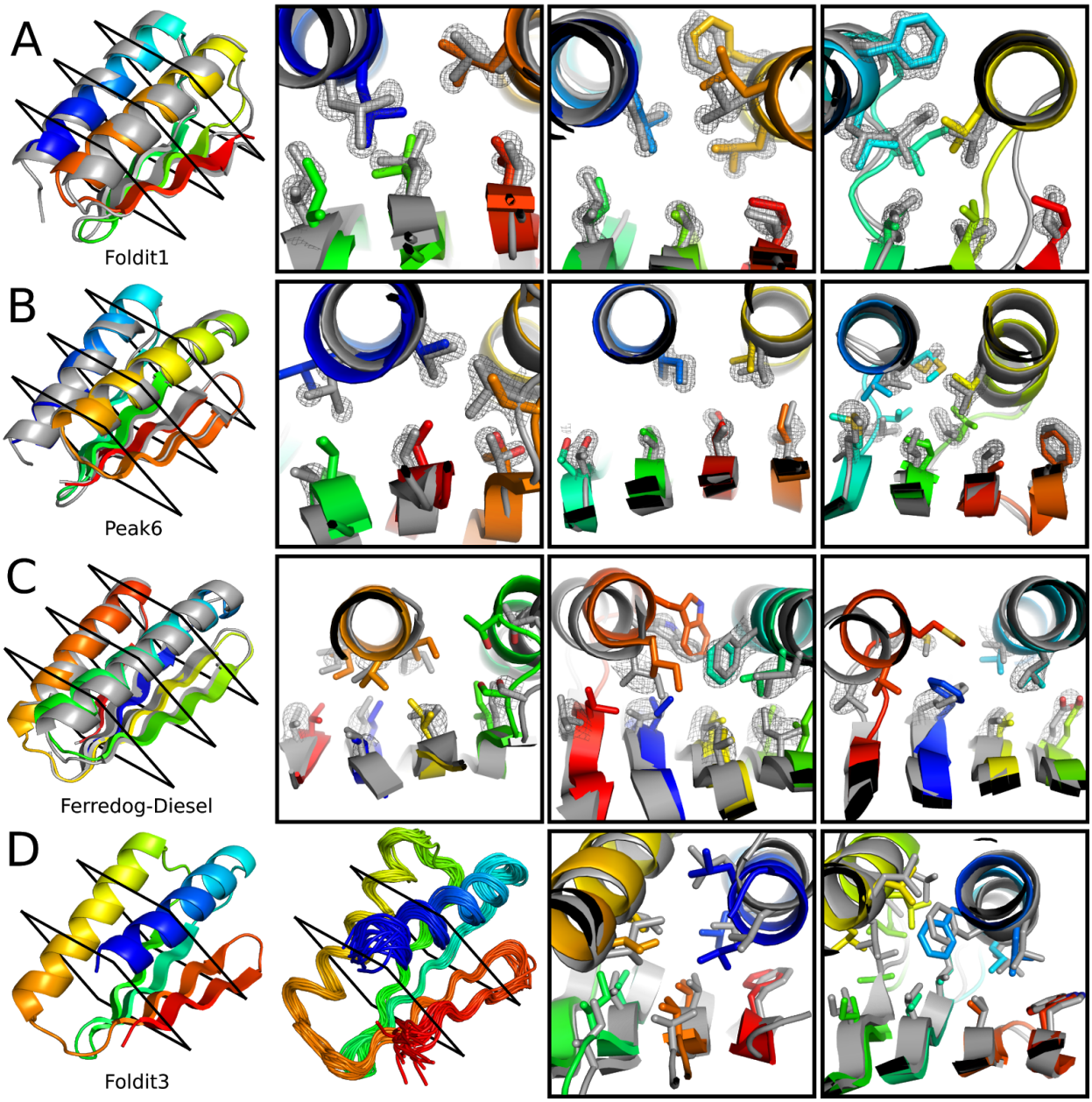


Figure 3.6. High-resolution structures of Foldit player-designed proteins. (A) The Foldit1 design (fold V in Fig 3: 3 β-strands with sheet order 1-2-3) model backbone (rainbow) aligns to the crystal structure (gray) with Cα-RMSD of 1.1 Å. (B) The Peak6 design (fold III: 4 strands, sheet order 1-2-4-3) model backbone aligns to the crystal structure with Cα-RMSD of 0.9 Å. (C) The Ferredog-Diesel design (fold I: 4 strands, sheet order 4-1-3-2) model backbone aligns to the crystal structure with Cα-RMSD of 1.7 Å. Cross-sections show core residue sidechains, with the composite omit 2mFo-DFc map contoured at 2.0 σ (A, B) or 1.0 σ (C). (D) 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 representative NMR model (gray).

Foldit player experience

The success of Foldit designs is not attributed to just one or two exceptional Foldit players, but is shared broadly by the Foldit community (**Table 1**). The 56 successful designs were created by 36 different Foldit players (the most prolific player authored 10 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.

The Foldit website features separate leaderboards for Prediction and Design puzzles, although most players participate in both categories. There is a strong correlation between player rankings in these categories, that skills developed playing Foldit structure prediction puzzles carry over to design puzzles, and vice versa (**Figure 3.7**).

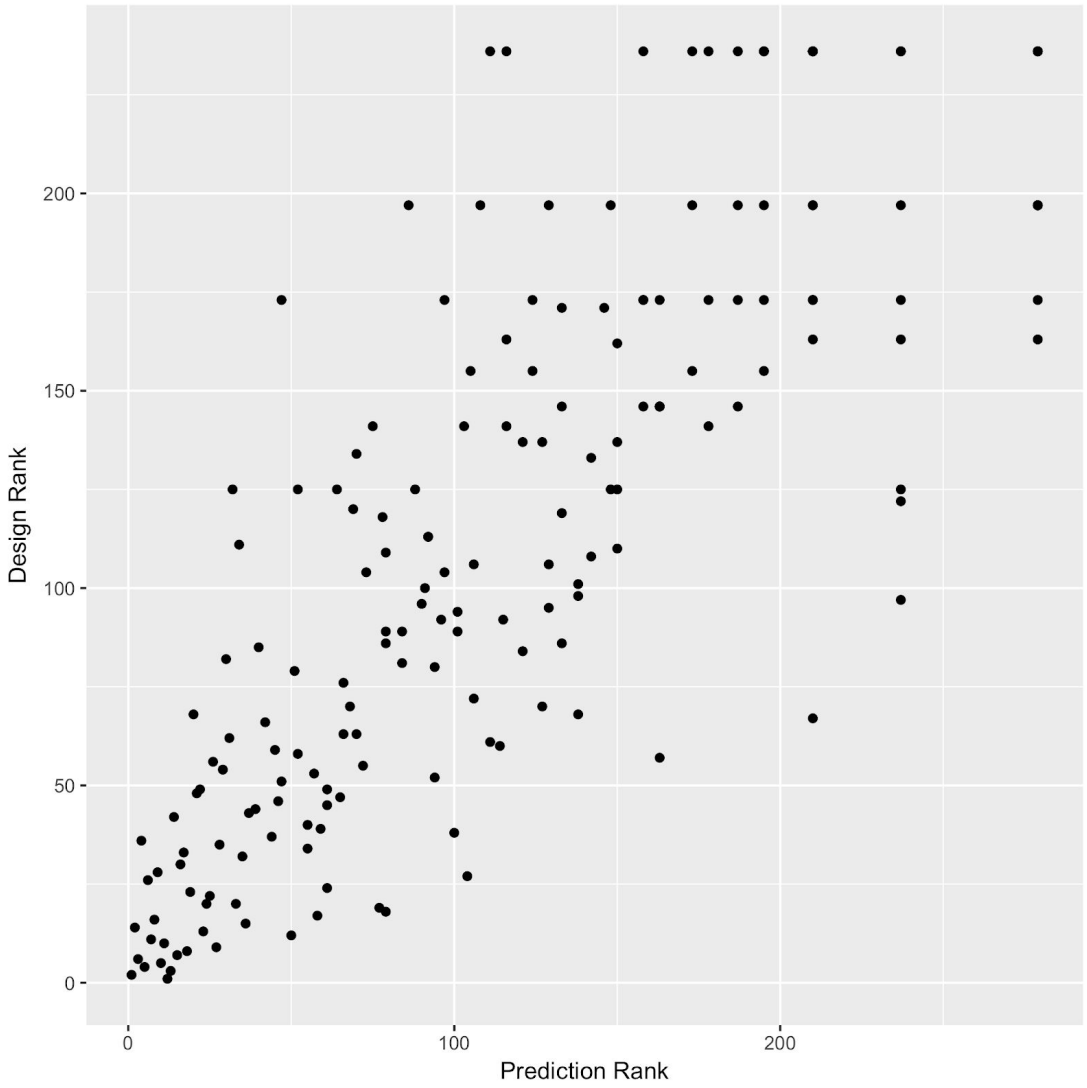


Figure 3.7. Foldit player rankings in "design" and "prediction" categories. Foldit player rankings are strongly correlated categories (Spearman's rank correlation coefficient of 0.84).

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 3.8**). 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[29]. 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.

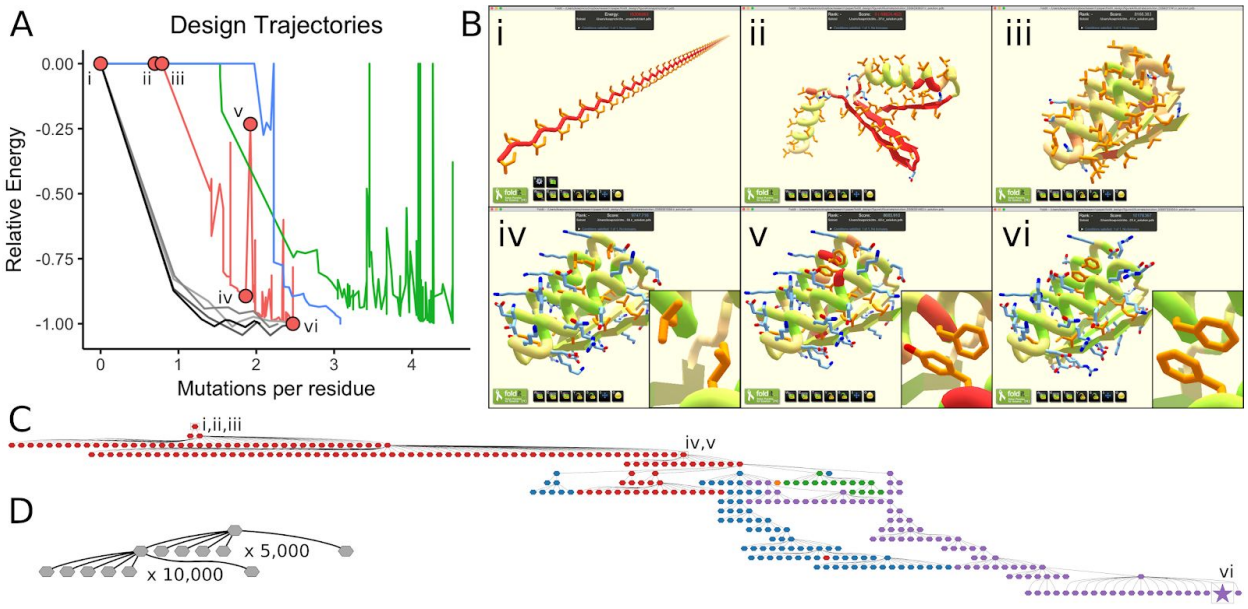


Figure 3.8. Comparison of Foldit player and automated design sampling strategies. (A) Single trajectories (ignoring abandoned branches) for three Foldit player-designed proteins in red (2002949_0000, a.k.a. Foldit1), blue (2003333_0006, a.k.a. Peak6), and green (2003169_S953, a.k.a. Ferredog-Diesel); 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.

We solicited feedback from all 3302 players who participated in Foldit protein design puzzles. The majority of responding Foldit players (61%) have completed no education beyond a college bachelor's degree, and 43% report that most of their knowledge of proteins comes from Foldit (**Figure 3.9**). A

number of players volunteered to share a testimonial of their design strategy which is included in the **Supplementary Information** for players who contributed successful designs. It is clear from the testimonials that players with a wide range of abilities and understanding are capable of using Foldit to design well-folded proteins. Many players have developed sophisticated strategies for protein design, although their descriptions and suggest their “expertise” comes almost entirely through playing FoldIt rather than any outside formal training in these areas.

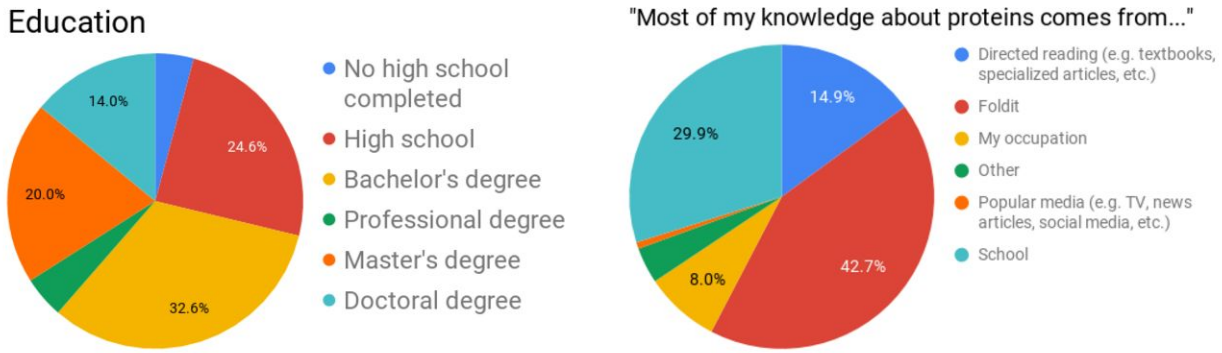


Figure 3.9. Foldit player expertise. All players who participated in Foldit protein design puzzles and who had not opted out of Foldit-related email were solicited for survey questions. Data is shown for *n* = 324 responding Foldit players.

Discussion

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 in Section II, 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.

Materials and Methods

Foldit puzzles and design selection

Foldit puzzles were set up and Foldit player designs were selected as described in Section II.

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 pET21 (designs 1998555*-2002990*), or from Twist in pET29 (designs 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 pET21 vectors or 30 µg/mL kanamycin for pET29 vectors. Proteins were expressed and purified as described in Section II.

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[78].

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. X-ray diffraction datasets were collected at the Advanced Light Source (Berkeley, CA). Data was processed with HKL2000[96]. Crystal structures were solved by molecular replacement with Phaser[97], 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[98,99]. Diffraction data and refinement statistics are listed in **Table 3**.

Crystallization screens were set up for 27 successful designs, and we obtained diffracting crystals for 7 different proteins. Three datasets were solved with satisfactory refinement statistics (2002949_0000, 2003333_0006, 2003169_S953; Figure 4). For two additional proteins, we could not find molecular

replacements solutions. For two other proteins, we were able to find molecular replacement solutions using Rosetta *ab initio* models of the designed sequence, however we could not refine these structure with acceptable crystallographic statistics.

Design 2002949_0000 (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. X-ray diffraction was collected to a resolution of 1.18 Å, and the structure was solved with R_{work}/R_{free} of 0.15/0.18. The refined structure closely matches the design model, with Cα-RMSD of 1.1 Å and all-atom RMSD of 1.8 Å.

Design 2003333_0006 (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 was collected to a resolution of 1.54 Å, and the structure was solved with R_{work}/R_{free} of 0.17/0.20. The refined structure closely matches the design model, with Cα-RMSD of 0.9 Å and all-atom RMSD of 1.9 Å.

Design 2003169_S953 (Ferredog-Diesel) was crystallized with 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 minutes, then flash frozen in liquid nitrogen. X-ray diffraction was collected to a resolution of 1.92 Å, and the structure was solved with R_{work}/R_{free} of 0.25/0.29. R_{meas} increases sharply in the highest-resolution shells, which may account for our difficulty improving the R-factors in model refinement. The refined structure closely matches the design model, with Cα-RMSD of 1.7 Å and all-atom RMSD of 2.3 Å.

Design 2002990_0006 was crystallized at 20 mg/mL in 0.1 M Bis-Tris pH 6.5, 0.2 M sodium chloride, 25% w/v PEG 3350. Crystals were briefly soaked in mother liquor plus 20% PEG 400, then flash frozen in liquid nitrogen. A diffraction dataset was collected at 1.5 Å resolution, but diffraction spots are heavily smeared in one direction and we could not find a molecular replacement solution.

Design 2003360_s000 was crystallized with His tag intact, at 15 mg/mL in 0.1 M phosphate/citrate pH 4.2, 0.2 M sodium chloride, 20% PEG 8000. Crystals were dehydrated by soaking in 5 µL mother liquor in open air for 10 minutes, then flash frozen in liquid nitrogen. A diffraction dataset was collected at 2.0 Å resolution, but unit cell dimensions suggest there are 3 copies in the asymmetric unit, and we could not find a molecular replacement solution.

Design 2003594_S028 was crystallized with His tag intact, at 85 mg/mL in 0.1 M Bis-Tris pH 5.5, 2 M ammonium sulfate. Crystals were briefly soaked in mother liquor plus 20% ethylene glycol, then flash frozen in liquid nitrogen. A diffraction dataset was collected at 2.76 Å resolution, and a molecular replacement solution was found (LLG=195) using a Rosetta *ab initio* model of the design sequence, suggesting the crystal structure is very close to the design. However, the diffraction quality of this dataset was poor (split and smeared diffraction spots), and we could not refine the structure.

Design 2002766_0002 was crystallized at 120 mg/mL in 0.1 M Tris pH 7.5, 27% MPD. Crystals were flash frozen in liquid nitrogen without further cryo-protection. A diffraction dataset was collected at 2.46 Å resolution, and a molecular replacement solution was found (LLG=107) using a Rosetta *ab initio* model of the design sequence. The molecular replacement solution suggests one α-helix of the 3-helix bundle is domain-swapped with a neighboring copy, extending as a fiber of interacting neighbors in the crystal lattice. Because SEC experiments indicate this protein behaves as a monodisperse monomer in solution

with no aggregation (Extended Data Figure 2), we conclude that this domain-swapped, fiber conformation must be a crystal artifact and does not reflect the protein conformation in solution. The diffraction quality of this dataset was poor (split and smeared diffraction spots) and we could not refine the structure.

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, 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[100], and analyzed using the programs SPARKY and XEASY[101]. Spectra were referenced to external DSS. Sequence-specific resonance assignments were determined using AutoAssign software together with interactive manual analysis, as described previously[102]. Backbone dihedral angle constraints were derived from the chemical shifts using the program TALOS_N[103] for residues located in well-defined secondary structure elements. The programs ASDP[104] and CYANA[105,106] were used to automatically assign NOEs and to calculate structures. RPF analysis[104,107] 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[108] using the program CNS[109]. The structural statistics and global structure quality factors (Extended Data Table 3) including Verify3D[110], ProsaII[111], PROCHECK[112], and MolProbity[113] raw and statistical Z-scores were computed using the PSVS[114] 1.5 and PDBStat[115] 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[107].

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, Peak6, and Ferredog-Diesel crystal structures, and the Foldit3 NMR structure, have been deposited in the RCSB Protein Database 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).

IV. References

1. Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* **266**, 635–643 (1996).
2. Chandonia, J.-M., Fox, N. K. & Brenner, S. E. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res.* **47**, D475–D481 (2019).
3. Woolfson, D. N. *et al.* De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **33**, 16–26 (2015).
4. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
5. Boas, F. E. & Harbury, P. B. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* **17**, 199–204 (2007).
6. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
7. Ramachandran, G. N. & Sasisekharan, V. Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283–438 (1968).
8. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **5**, 17 (2007).
9. Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5482–5485 (1999).
10. Zhang, Y., Kolinski, A. & Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145–1164 (2003).
11. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3054–3059 (2018).
12. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
13. Das, R. *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69 Suppl 8**, 118–128 (2007).
14. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
15. Yue, K. & Dill, K. A. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 4163–4167 (1992).
16. Koehl, P. & Levitt, M. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181 (1999).
17. Leaver-Fay, A. *et al.* Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. in *Methods in Enzymology* (ed. Keating, A. E.) **523**, 109–143 (Academic Press, 2013).
18. Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).
19. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
20. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S. & Richardson, D. C. Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 8747–8751 (1994).
21. Handel, T. M., Williams, S. A. & DeGrado, W. F. Metal ion-dependent modulation of the dynamics of a designed protein. *Science* **261**, 879–885 (1993).

22. Figueroa, M. *et al.* The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure prediction tools. *J. Struct. Biol.* **195**, 19–30 (2016).
23. Jin, W., Kambara, O., Sasakawa, H., Tamura, A. & Takada, S. De Novo Design of Foldable Proteins with Smooth Folding Funnel: Automated Negative Design and Experimental Verification. *Structure* **11**, 581–590 (2003).
24. 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).
25. Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10 (1996).
26. Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science* **249**, 884–891 (1990).
27. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
28. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
29. Lin, Y.-R. *et al.* Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5478–85 (2015).
30. Kortemme, T., Ramírez-Alvarado, M. & Serrano, L. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science* **281**, 253–256 (1998).
31. Hu, X., Wang, H., Ke, H. & Kuhlman, B. Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design. *Structure* **16**, 1799–1805 (2008).
32. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
33. Samish, I. Achievements and Challenges in Computational Protein Design. in *Computational Protein Design* (ed. Samish, I.) 21–94 (Springer New York, 2017).
34. Kono, H. & Saven, J. G. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.* **306**, 607–628 (2001).
35. Chowdry, A. B. *et al.* An object-oriented library for computational protein design. *J. Comput. Chem.* **28**, 2378–2388 (2007).
36. Grigoryan, G., Reinke, A. W. & Keating, A. E. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859–864 (2009).
37. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
38. Chen, Z. *et al.* Programmable design of orthogonal protein heterodimers. *Nature* **565**, 106–111 (2019).
39. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
40. Kraemer-Pecore, C. M., Lecomte, J. T. J. & Desjarlais, J. R. A de novo redesign of the WW domain. *Protein Sci.* **12**, 2194–2205 (2003).
41. Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
42. Zelt, J., Courter, J., Arab, A., Johnson, R. & Droege, S. Reviving a Legacy Citizen Science Project to Illuminate Shifts in Bird Phenology. *Int. J. Zool.* **2012**, (2012).
43. Saladyga, M. The 'Pre-Embryonic' State of the AAVSO: Amateur Observers of Variable Stars in the United States from 1875 to 1911. *Journal of the American Association of Variable Star Observers (JAAVSO)* **27**, 154–170 (1999).

44. Kawrykow, A. *et al.* Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One* **7**, e31362 (2012).
45. Lee, J. *et al.* RNA design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2122–2127 (2014).
46. Simpson, R., Page, K. R. & De Roure, D. Zooniverse: Observing the World's Largest Citizen Science Platform. in *Proceedings of the 23rd International Conference on World Wide Web* 1049–1054 (ACM, 2014).
47. CitSci.org. *CitSci.org* Available at: https://www.citsci.org/. (Accessed: 6th February 2019)
48. Franzoni, C. & Sauermann, H. Crowd science: The organization of scientific research in open collaborative projects. *Res. Policy* **43**, 1–20 (2014).
49. Lintott, C. J. *et al.* Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *arXiv [astro-ph]* (2008).
50. Bello, J. P. *et al.* SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution. *arXiv [cs.SD]* (2018).
51. Sullivan, D. P. *et al.* Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).
52. Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* **3**, 2053951715622512 (2016).
53. Burnett, S., Furlong, M., Melvin, P. G. & Singiser, R. Games that Enlist Collective Intelligence to Solve Complex Scientific Problems. *J. Microbiol. Biol. Educ.* **17**, 133–136 (2016).
54. Yi, S. K. M., Steyvers, M., Lee, M. D. & Dry, M. J. The wisdom of the crowd in combinatorial problems. *Cogn. Sci.* **36**, 452–470 (2012).
55. Mugar, G. Preserving the Margins: Supporting Creativity and Resistance on Digital Participatory Platforms. *Proceedings of the ACM on Human-Computer Interaction* **1**, (2017).
56. Jennett, C. *et al.* Creativity in Citizen Cyberscience. *HC* **3**, (2016).
57. Sørensen, J. J. W. H. *et al.* Exploring the quantum speed limit with computer games. *Nature* **532**, 210–213 (2016).
58. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
59. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
60. Khatib, F. *et al.* Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18949–18953 (2011).
61. 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).
62. Eiben, C. B. *et al.* Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* **30**, 190–192 (2012).
63. Filip Karlo Dosilovic, Mario Brcic, Nikica Hlupic. Explainable Artificial Intelligence: A Survey. *MIPRO 2018 - 41st International Convention Proceedings* 210–215 (2018).
64. Kim, J. S. *et al.* Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331–336 (2014).
65. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. Forces contributing to the conformational stability of proteins. *The FASEB Journal* **10**, 75–83 (1996).
66. Callaway, D. J. E. Solvent-induced organization: A physical model of folding myoglobin. *arXiv [cond-mat]* (1994).
67. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

68. Blout, E. R. & Idelson, 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).
69. 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).
70. Ghosh, K. & Dill, K. A. Theory for Protein Folding Cooperativity: Helix Bundles. *J. Am. Chem. Soc.* **131**, 2306–2312 (2009).
71. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
72. Crick, F. H. C. The packing of α-helices: simple coiled-coils. *Acta Crystallogr.* **6**, 689–697 (1953).
73. Chothia, C., Levitt, M. & Richardson, D. Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215–250 (1981).
74. Richmond, T. J. & Richards, F. M. Packing of α-helices: Geometrical constraints and contact areas. *J. Mol. Biol.* **119**, 537–555 (1978).
75. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
76. Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* **13**, 222–245 (1974).
77. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
78. Santoro, M. M. & Bolen, D. W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* **27**, 8063–8068 (1988).
79. Skerra, A. Engineered protein scaffolds for molecular recognition. *J. Mol. Recognit.* **13**, 167–187 (2000).
80. Gebauer, M. & Skerra, A. Engineered protein scaffolds as next-generation antibody therapeutics. *Curr. Opin. Chem. Biol.* **13**, 245–255 (2009).
81. Stern, L. A., Case, B. A. & Hackel, B. J. Alternative Non-Antibody Protein Scaffolds for Molecular Imaging of Cancer. *Curr. Opin. Chem. Eng.* **2**, (2013).
82. Vazquez-Lombardi, R. *et al.* Challenges and opportunities for non-antibody scaffold drugs. *Drug Discov. Today* **20**, 1271–1283 (2015).
83. Golinski, A. W., Holec, P. V., Mischler, K. M. & Hackel, B. J. Biophysical Characterization Platform Informs Protein Scaffold Evolvability. *ACS Comb. Sci.* (2019). doi:10.1021/acscombsci.8b00182
84. Regan, L. & DeGrado, W. Characterization of a helical protein designed from first principles. *Science* **241**, 976–978 (1988).
85. Thomson, A. R. *et al.* Computational design of water-soluble α-helical barrels. *Science* **346**, 485–488 (2014).
86. Jacobs, T. M. *et al.* Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
87. Karanicolas, J. & Brooks, C. L., 3rd. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3954–3959 (2003).
88. Richardson, J. S. & Richardson, D. C. The de novo design of protein structures. *Trends Biochem. Sci.* **14**, 304–309 (1989).
89. Remaut, H. & Waksman, G. Protein-protein interaction through beta-strand addition. *Trends Biochem. Sci.* **31**, 436–444 (2006).
90. Watkins, A. M. & Arora, P. S. Anatomy of β-strands at protein-protein interfaces. *ACS Chem. Biol.* **9**, 1747–1754 (2014).

91. Cheng, P.-N., Pham, J. D. & Nowick, J. S. The supramolecular chemistry of β-sheets. *J. Am. Chem. Soc.* **135**, 5477–5492 (2013).
92. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
93. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
94. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
95. Montelione, G. T. *et al.* Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563–1570 (2013).
96. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. in **276**, 307–326 (Elsevier, 1997).
97. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
98. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
99. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
100. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
101. 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).
102. Liu, G. *et al.* NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10487–10492 (2005).
103. 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).
104. 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 (2006).
105. 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).
106. 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).
107. 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).
108. Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins* **50**, 496–506 (2003).
109. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
110. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
111. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362 (1993).
112. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. \it PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
113. 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).
114. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).
115. 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).

V. Supplementary information

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 (NCBI RefSeq ID)	TM-align Score	TM-align Hit (PDB ID)	Expressed	Soluble	Monomeric	Structured
Round 1										
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,dbuske,PLAYER_9	35.4	5.5	KFV13184.1	0.685	3vf0A	Yes	Yes	No	-
997523_0000	PLAYER_11,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,Znaika	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	No	-
997523_1040	**johnmitch**	**-**	**-**	**-**	**0.811**	**5k7vA**	**Yes**	**Yes**	**Yes**	**Yes**
997791_1027	**johnmitch**	**-**	**-**	**-**	**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**
Round 2										
1998469_0000	reefyrob,retiredmichael	-	-	-	0.686	4p2fA	No	-	-	-
1998520_0002	Galaxie,gmn	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_8	-	-	-	0.642	3dyjB	No	-	-	-
1998925_s008	**PLAYER_14**	**36.6**	**1.4**	**WP_013842421.1**	**0.789**	**4fsxA**	**Yes**	**Yes**	**Yes**	**Yes**
2000240_0002	retiredmichael,reefyrob,LociOiling,smilingone,PLAYER_16	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_8	-	-	-	0.634	4xevD	No	-	-	-
2000518_0003	PLAYER_17	-	-	-	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_16,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	No	-
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	dcrwheeler	36.2	2.7	XP_021195829.1	0.752	5iz3A	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	-
Round 3										
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	**dcrwheeler**	**-**	**-**	**-**	**0.813**	**4tqlA**	**Yes**	**Yes**	**Yes**	**Yes**
2002766_0003	Galaxie,tokens	-	-	-	0.846	4iv6A	Yes	Yes	Yes	No
2002766_0004	actiasluna,dbuske,Blipperman	-	-	-	0.903	5cwmA	Yes	Yes	No	-
2002766_0006	PLAYER_13	-	-	-	0.853	4hwhA	Yes	Yes	No	-
2002922_1013	Mark-,Cyberkashi,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 (Foldit1)	**Galaxie,Susume,PLAYER_10**	**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	Cyberkashi,gitwut	-	-	-	0.738	5adxK	Yes	Yes	Yes	No
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 (Ferredog-Diesel)	**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_10,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_19,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 (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_10	-	-	-	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_13	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_7**	**-**	**-**	**-**	**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_15**	**35**	**7.4**	**WP_006978974.1**	**0.813**	**4nezA**	**Yes**	**Yes**	**Yes**	**Yes**
2003333_0006 (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	reefyrob	-	-	-	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,Aotearoa**	**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_6**	**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_6	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[51] 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.

Table 2. Success rates of Foldit player-designed proteins

	Foldit player designs				
	Round 0	Round 1	Round 2	Round 3	Lin et al.[26]
Sequence complexity[a]	0.20	0.35	0.44	0.21	0.20
Rosetta energy[b] (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[c]	263	1404	4296	2092	
Models selected for ab initio	0	100	1141	612	*(Not reported)*
Ab initio convergence	NA	12 12%	37 3%	99 16%	72
Models tested	NA	12	37	97	72
Expressed	NA	12 100%	23 62%	86 89%	70 97%
… and soluble	NA	12 100%	18 49%	71 73%	64 89%
… and monomeric	NA	7 58%	7 19%	52 54%	39 54%
… and structured	NA	6 50%	4 11%	46 47%	29 40%
Number of unique folds	NA	3	4	19	2

[a]Linguistic sequence complexity was calculated from the top 10-ranked models in all puzzles, using word lengths of 1, 2, and 3.

[b]Rosetta energy is the talaris2013 energy normalized by residue count. Values shown are mean and standard deviation for top 10-ranked models in all puzzles. See Figure 3.1 for sample sizes.

[c]Includes redundant models, since very similar models can appear in two or more categories (top, shared, and clustered). See Methods for details on model selection.

Table 3. Crystallographic data collection and refinement statistics

	Foldit1	Peak6	Ferredog-Diesel
Wavelength	1	1	1
Resolution range	28.92 - 1.18 (1.222 - 1.18)	26.21 - 1.541 (1.596 - 1.541)	45.29 - 1.916 (1.985 - 1.916)
Space group	P 1 21 1	P 31 2 1	P 42 21 2
Unit cell	24.045 43.584 29.276 90 98.998 90	52.414 52.414 56.086 90 90 120	69.21 69.21 90.59 90 90 90
Total reflections	60389 (6169)	129411 (4118)	203164 (19299)
Unique reflections	18574 (1830)	12866 (860)	17438 (1682)
Multiplicity	3.3 (3.4)	10.1 (4.8)	11.7 (11.5)
Completeness (%)	92.67 (88.38)	94.86 (65.00)	99.06 (97.65)
Mean I/sigma(I)	25.65 (9.97)	18.52 (1.34)	16.94 (0.86)
Wilson B-factor	10.36	17.88	39.95
R-merge	0.02508 (0.1209)	0.0872 (0.7896)	0.08947 (3.164)
R-meas	0.03015 (0.1439)	0.09186 (0.878)	0.09364 (3.31)
R-pim	0.01654 (0.07738)	0.02847 (0.3694)	0.02721 (0.9595)
CC1/2	0.999 (0.991)	0.999 (0.714)	1 (0.385)
CC*	1 (0.998)	1 (0.913)	1 (0.746)
Reflections used in refinement	18574 (1749)	12861 (860)	17376 (1663)
Reflections used for R-free	1829 (174)	1282 (85)	1732 (166)
R-work	0.1464 (0.1278)	0.1682 (0.2761)	0.2477 (0.3965)
R-free	0.1819 (0.1755)	0.1975 (0.3091)	0.2907 (0.3789)
CC(work)	0.963 (0.982)	0.967 (0.830)	0.959 (0.616)

CC(free)	0.956 (0.956)	0.953 (0.806)	0.910 (0.602)
Number of non-hydrogen atoms	690	755	1709
macromolecules	574	646	1672
ligands		20	
solvent	116	89	37
Protein residues	68	77	200
RMS(bonds)	0.008	0.007	0.005
RMS(angles)	0.83	1.03	1.01
Ramachandran favored (%)	100.00	100.00	98.97
Ramachandran allowed (%)	0.00	0.00	1.03
Ramachandran outliers (%)	0.00	0.00	0.00
Rotamer outliers (%)	0.00	0.00	1.72
Clashscore	2.60	3.75	8.54
Average B-factor	16.37	24.96	68.81
macromolecules	14.54	22.82	69.09
ligands		47.36	
solvent	25.39	35.49	55.90
Number of TLS groups		3	9

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

Table 4: 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[110] and PSVS 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.

Protein design strategies of Foldit players

We asked Foldit players to describe their protein design strategy when playing Foldit. Below are Foldit player responses to the following prompt:

"When we discuss this work with other researchers, they always want to know more about Foldit player strategies! Would you tell us more about how you design proteins in Foldit? Your response may be published in an online supplement to the research paper.

Some possible prompts:
Does your strategy change between the early and late stages of a design puzzle?
Which Foldit tools do you find most useful?
How much time is spent hand-folding vs. running recipes?
Do you use your own custom recipes or do you use common recipes shared by the Foldit community?
Do you use tools or resources outside of Foldit?
When you play Foldit design puzzles, is your sole objective to achieve the highest score, or do you pursue other objectives?
What would you say are the most important considerations for designing a high-scoring protein in Foldit?"

<u>Responses from designers of successful proteins</u>

Aotearoa (collaborating designer of 2003360_1013)
Random Chaos is my game play along with my wiki strategy - then an Aotearoa's Romance script Blue Fuse and Ravens scripts, Pletsch and A few more.

Bletchley Park (collaborating designer of 2002308_0000, 2002713_0004, 2002376_0000, 2003308_0004, 2002766_0000, 2002745_0000, 2002553_0000, 2002922_1013)
manually set up raw structure, process using scripts using a mix of public and private scripts and external data processing tools (not using PDB). My goal is to learn how to achieve better models (higher scores in Foldit) using different strategies in order to help find cures for diseases.

Blipperman (collaborating designer of 2003205_0003, 2002766_0004)
My approach depends upon the puzzle. After a variable amount of hand folding to align those segments I desire, I use a series of scripts to optimize the structure at low wiggle power. Most of the scripts are written by others with about 50% being open scripts and the rest developed by present or past group members. A few of the scripts have been modified slightly by me either for my own purposes or at request by group members. I make extensive use of banded structures to hold units in place until the structure has settled sufficiently. I often will start three or four different folds before the puzzle session has expired. Some of these are just mid-game tweaks, but others are grossly different in layout. I generally do not seek out external information, but sometimes tidbits will be suggested by other group members that influence my fold.

Bruno Kestemont (designer of 2003455_S886, 2003485_0002; collaborating designer of 2003382_0002, 2002745_0004, 2003048_0005, 2000665_1003, 2003360_1013, 2002469_0001, 2002922_1018)

I start with a self made shared to group recipe ('Design Models') that proposes several optional macro-structures depending on the length of the protein (e.g. HHH for 3 helices, HHSSS, HSSSH, HHSSSS, SSSS_SSSS etc.). This avoids me to calculate again and again - I did it once in an excel file and I implemented it in the recipe.

Depending on the bonuses (seeking more or less helices and sheets) I design 3-4 different models in 3-4 different tracks. These are the 3-4 definitive candidate separate paths ('B1, B2, B3'). For each of them, I do the following when I find time (all in auto wiggle power).

1) 'Ideal SS' all structures one by one. I might here add some segments in loops in order to get preferably loops of 4-5 (they are more flexible for later design) replacing some loops of 2 segments proposed by the recipe in order to keep sheets with pair number of segments and avoid loops of 3 segments. Then I use BluePrint tool on all loops. Views are showing bonded sheets and 'view non ideal loops'.

I then add bands between non-bonded sheets. I freeze all H and bonded sheet parts as well as the ideal loops. I try to close the sheet bonding (and idealize the loops) with small clash importance (about 0.7) local and/or global wiggling, or pushing-pulling with the mouse. If it doesn't work, I use the Remix tool on the non ideal loops. If it's not ok, I 'Auto Structures' and I remix again the non ideal loops. I might switch to selection interface in order to select and remix. I might add or delete segments in loops or helices in order to get a nice aggregated overall picture.

When it's more or less ok I freeze all sheets, helices and sheet-sheet loops. Using some bands with clash importance 0.4, I wiggle all and I stop immediately when the Monomer Core Bonus is ok.

Clash importance 1 mutate all (2 or 3 or infinite if I'm afk). Shake all (1 or 2). Wiggle Sidechains. Mutate (2-3). Save.

From here I can divide the main track (e.g. B1) in subtracks B11 B12 ... I might keep one of them for further hand fold (playing with as in sentence above, with a unique goal to maximize all bonuses, not points). On parallel tracks, or when I'm off, I run recipes.

2) Recipes starting with self made & shared 'Design Stabilize' and/or 'Find Starting Filter'. The first one 'breathes' the protein playing with wiggle power mutate and wiggles as in the lines above, selecting optionally the best filter score or/and best filters. The 'Find Starting Filter' one remixes the unideal loops on various lengths and only keeps the best filter bonuses (with options to give more or less importance to score or filters). I intend in the future to merge these two recipes.

Save and share best of tracks B1 B2 ... to myself, and the very best one (best filter bonus AND score) to group.

3) Some hand fold to try to idealize bad loops. At the end I only select 1 best scoring track and several (B1, B2, B3) best filter score designs.

4) Night or day 12 hours recipes on subtracks B1 B11, B12, B2, B21, etc. using following recipes in parallel and in different orders: 'JET' (a local wiggle strategy recipe), 'quickfix' (a worst

segments local wiggle and idealize recipe) then 'Random Idealize' (a macro idealizing recipe including mutating).
5) Again hand fold & mutating in order to get a better visually 'beautifull' structure with all filters ok.
6) Untill Day-3 using recipes one after the other during 8 hours of my absence (work 8 hours, select best result for each design, run other recipe, same before to go to sleep and after night, sharing the best scoring one to group). Recipes: Gab Bands In Space (incl. mutate) filters enabled the first day, disabled on wiggle other days. JET. Latest days: recipes including cuts. When it's stiffed: Medium wiggle power with a dedicated recipe ('fuzing' with cuts). The low wiggle power and again JET, GAB BIS etc.
7) 3 latest days on low wiggle power: recipes with local bands and local wiggles: Local Quake (includes mutating) Banded Worm Pairs Infinite Filters, JET etc.
8) Latest hours: Medium wiggle power & same recipes as in 7.
Conclusion: the start game motivation is filter bonus and esthetics with 15-30 minutes hand fold per design. After 1 day (by recipes 2-3 parallel tracks per design), some 15-30 minutes hand fold to repair the designs. Then almost only recipes run changing recipes and selecting best scores/filters every 8 hours (3 minutes per design). Total 98% time recipes, 2% time human intervention.

dbuske (designer of 997383_S346, 2002766_0004)
A few hours rather then days. Law of deminishing returns.

dcrwheeler (designer of 2002613_0004, 2002766_0002)
Q1-Not really. Q2- Shake, Wiggle, Rebuild, Freeze, Remix, Mutate, Bands, all wheel menu all modes, Q3- 10% hand - 90% recipe. Q4- Common recipes. Q5- Just Foldit. Q6- Goal is something useful. Q7- I wish I knew :)

georg137 (collaborating designer of 2002745_0000)
I use a combination of intuition and knowledge of what has folded (scored) well previously. I don't have a roadmap. My most used command is 'undo'.

fiendish_ghoul (designer of 2003333_0006 a.k.a. 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. 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.

fiendish_ghoul, cont. (concerning the design of 2002290_S122, which has high sequence- and structure-homology to a family of bacterial transferases)
Overall shape is definitely inspired by natural beta helices, but i was not trying to reproduce any particular protein structure or sequence. I observed some natural proteins to decide on helix handedness and beta strand length, and core residues and loops were designed to fit the overall helix shape.

Galaxie (collaborating designer of 2002949_0000 a.k.a. 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.

gitwut (designer of 2002877_s005, 2003382_1021; collaborating designer of 2003048_0003, 2002713_0004, 2003308_0004)
Strategies vary depending on the type of puzzle. But generally I hand fold to begin with and use differing scripts (written by others, sometimes modifying a few parameters myself) according to

the puzzle stage (early, middle, late) and guidelines. If early attempts don't scoring well, I go with standard designs that have worked in the past.

johnmitch (designer of 997523_1040, 997791_1027)
I do hand-folding if necessary. I use public recipes with a few modifications for most of the work. The sequence of running different recipes is very important. I have little knowledge of protean science but I know what a protean looks like.

LociOiling (designer of 2003360_s004, 2002766_0001; collaborating designer of 2000240_0002, 2002376_0003)
Design puzzles usually involve a protein that's presented as a straight 'extended chain', usually all isoleucine. Good hand folding at the start is important, it can be difficult or impossible to correct problems later on. A section of non-ideal loop is likely to stay non-ideal at a certain point. When possible I like to try multiple designs. I often go with one proven design, and one inspired by real proteins. Proven designs score well, but are unlikely to fold up on their own. A realistic design at least offers hope that a protein will fold on its own, a rare achievement.
Whatever the design recipes are required for finishing. Most of the recipes my group uses are variations on public recipes. Good recipes don't give up easily, and they often try small random variations, such as changing the clashing importance setting a little, or freezing the backbone of a few random segments.
Running multiple instances of Foldit a key strategy in general. This lets me work on several puzzles or different designs at the same time. Being willing to let your computers run 24x7 is another element of Total Folding.

smilingone (collaborating designer of 2002713_0000, 2000240_0002)
I use scripts/recipes more than hand folding and I don't use any tools outside of foldit. My early stage script usage is very different from late stage scripts used.

Susume (designer of 2003265_s008 a.k.a. Foldit3; collaborating designer of 2002949_0000 a.k.a. 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, 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. 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 or psipred 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 (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.

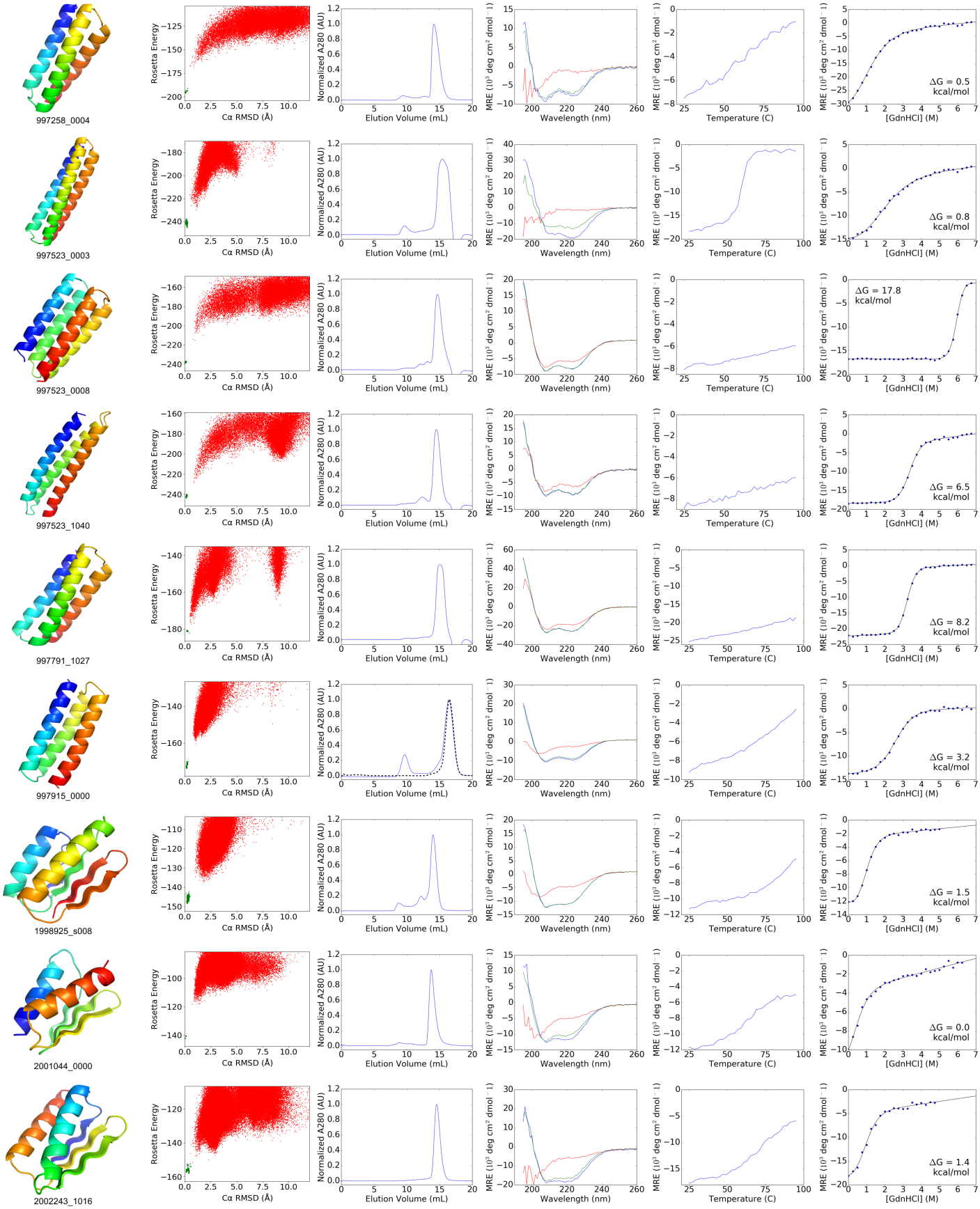
(See also Susume's instructional screencast at https://youtu.be/-nizMbICCM0)

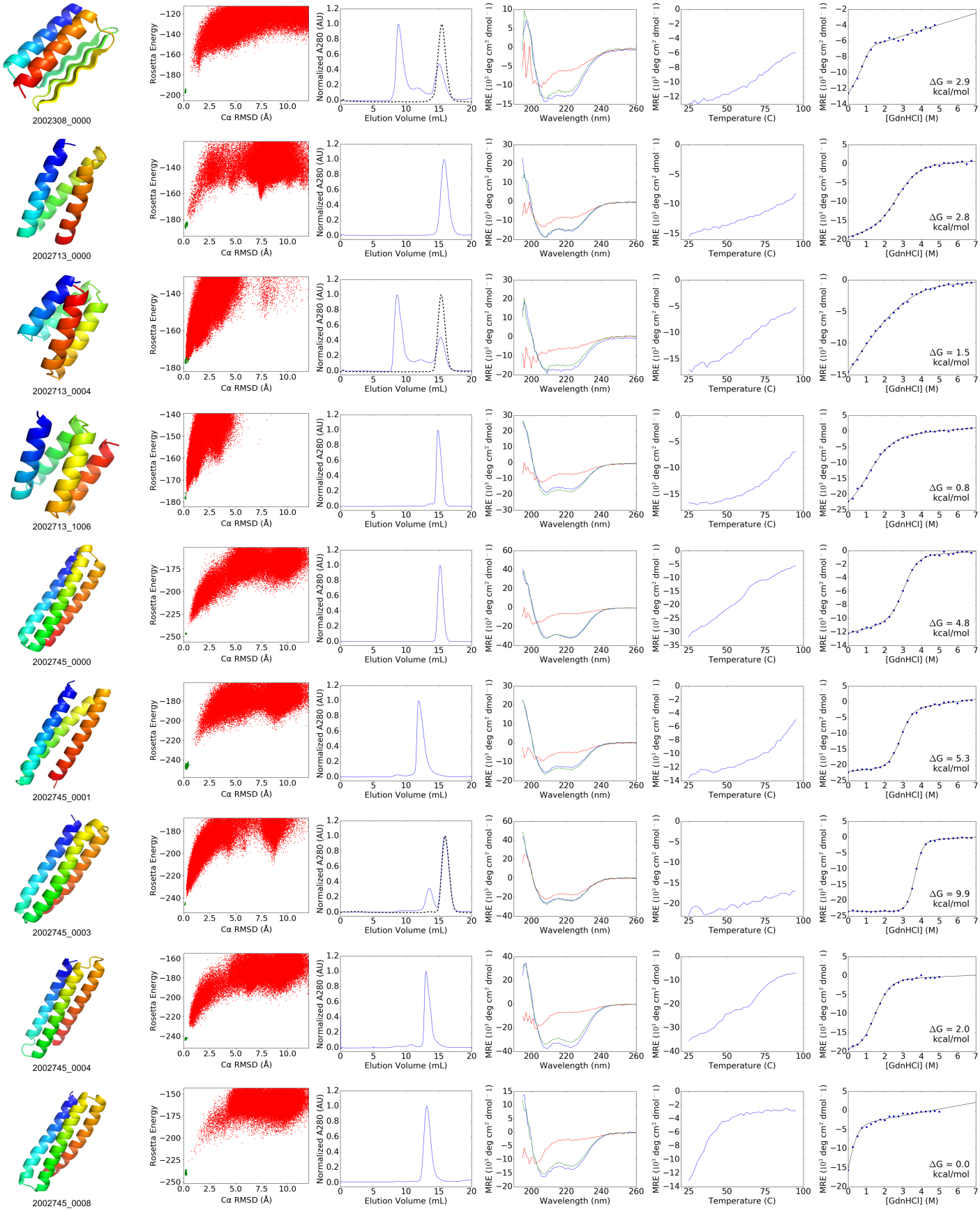
Timo van der Laan (designer of 997523_0006, 997258_0004)

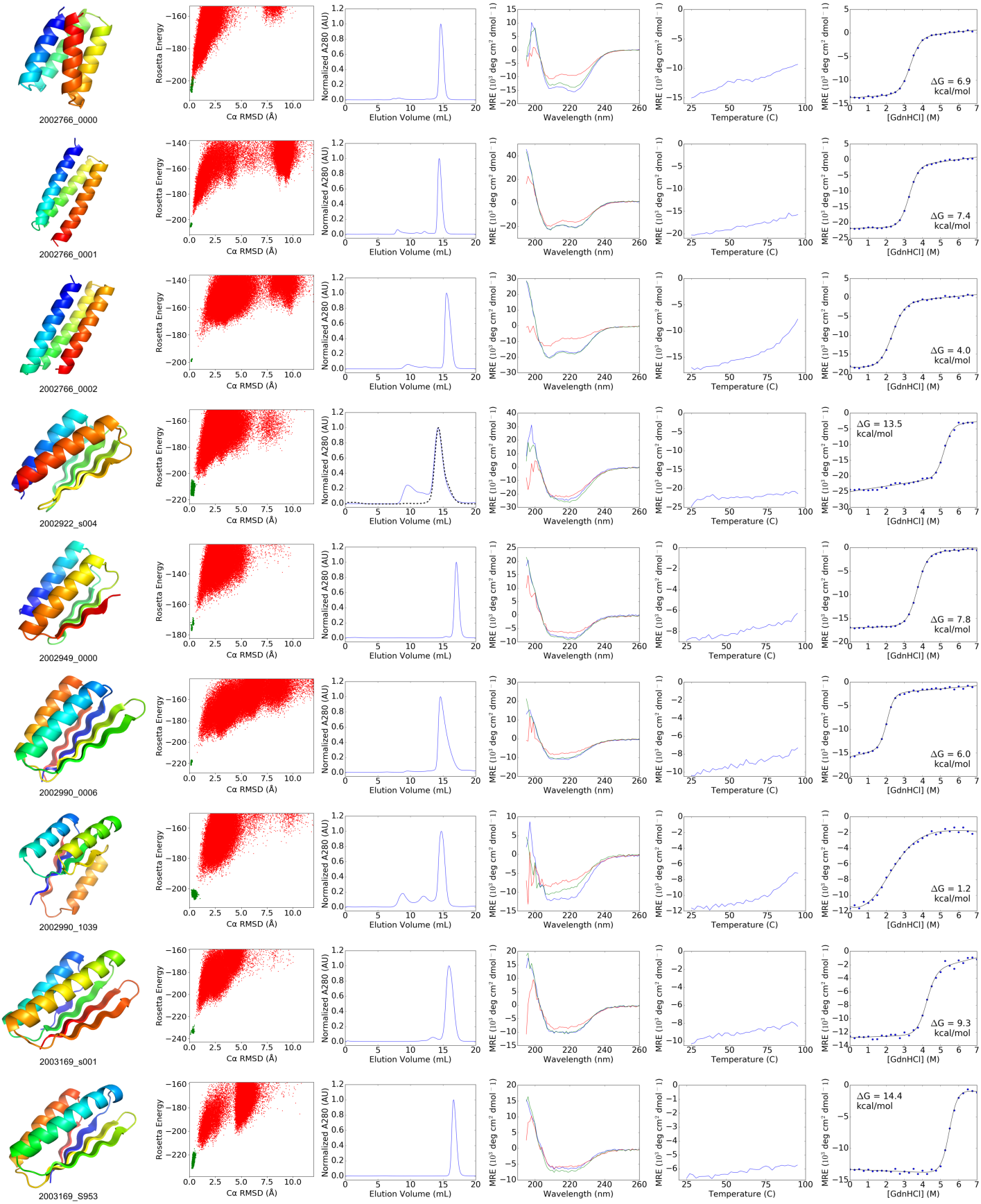
First using a simple custom script of my own making the structures, depending on the constraints given,I will use. Then using cut and move to implement the major structure or the blueprint tool to do that. After that using some bands or cut and move to create a principle core. Then using a special script of my own (bandmaker) to prevent the created structure to get destroyed while running Wiggle at low CI and later on higher CI.
Now it is time to mutate all either by using a special script or the standard mutate tool. Some wiggle one run of my DRemix script, running a compressor script, one run of a Mutate Combo script, running Rav GAB Bis, All the above on low Wiggle power. Maybe not in this sequence. If needed at the beginning one run of Voidcrusher using the all option to get more in the core if that is needed.
Then it is time to go to wiggle power medium using a simple idealize script.
After that going back to GAB Bis on low wiggle power followed by idealize on medium.
Finally in the end using several scripts that are usefull in the endgame. Jet a Settle script, an old Acid tweeker script etc.
In a game where hydrogen networks are asked I will use the Hbond script.
So I have a strategy for the initial phase for the medium phase and the end game using different scripts, mostly common scripts.
The initial part is about 4 to 6 hours mostly hand work. The rest is recipes running a long time.
For design initial most useful is Blueprint and Cut and Move.
No outside tools are used by me for this type of puzzle.
Getting the most points is what I am after and that is my only objective.

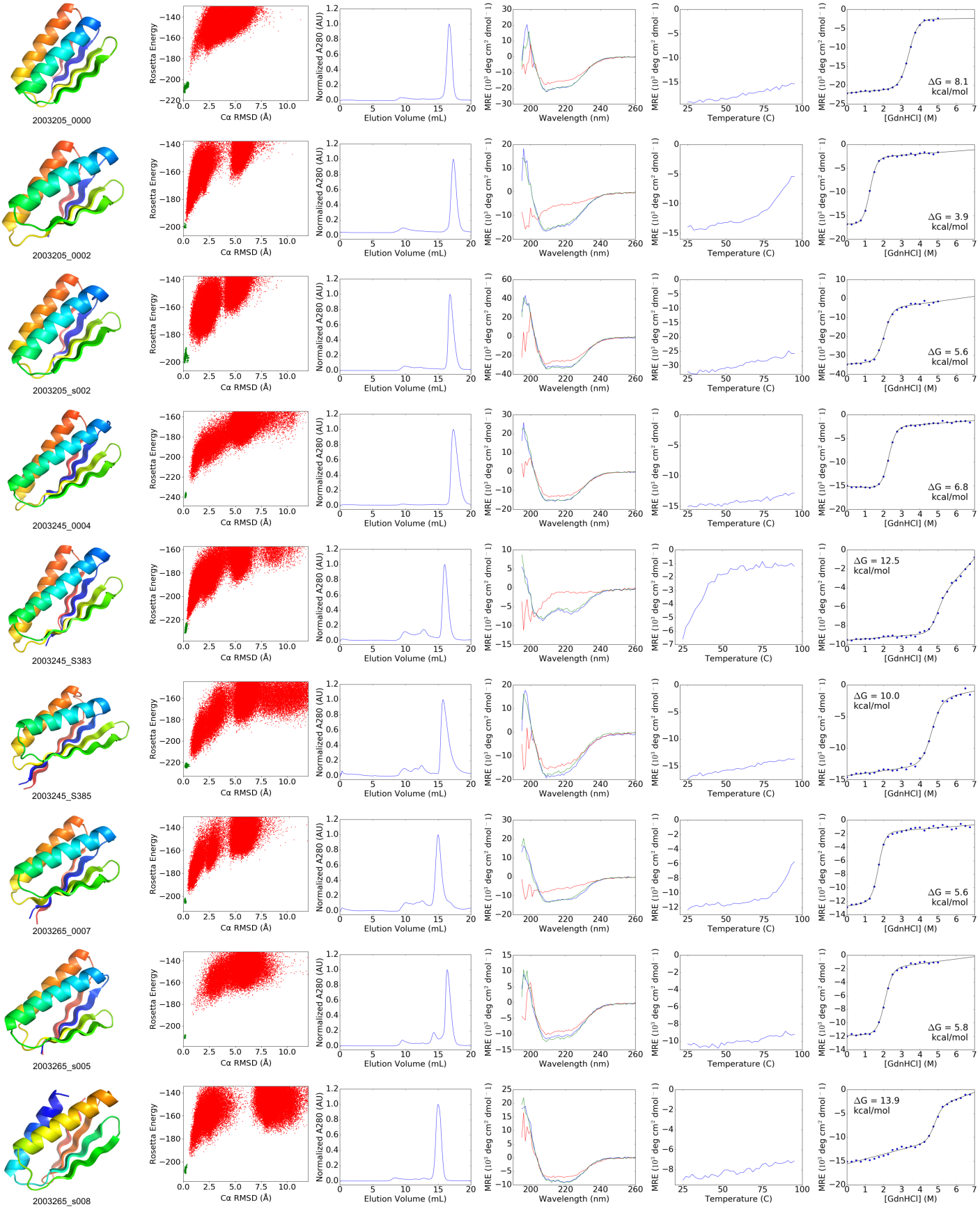
Znaika (collaborating designer of 997523_0005)
Shake, wiggle, idealize in low, medium and high behaviour. If needed, use rubbers, freeze and mutate.

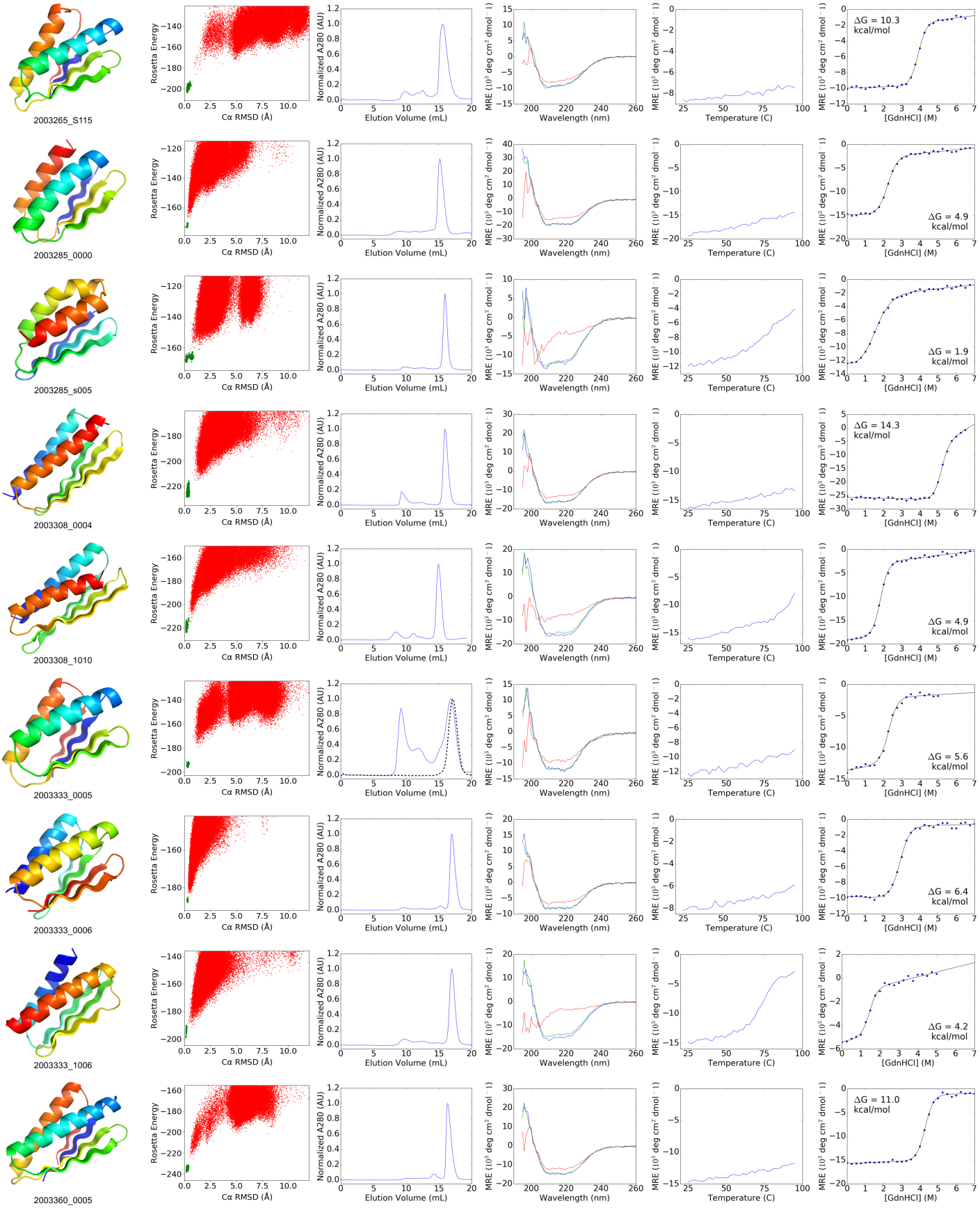
Figure S1. Protein folds represented by successful Foldit player designs. Each fold has a unique arrangement and connectivity of secondary structure elements, depicted in cartoon diagrams. Diagrams are labeled with Roman numerals as in **Figure 3.4**. Fold XX is a new fold, previously unobserved in natural proteins; TM-align[89] and DALI[111] alignments of design 2003594_S028 against the entire PDB found no structural homologs with this fold.

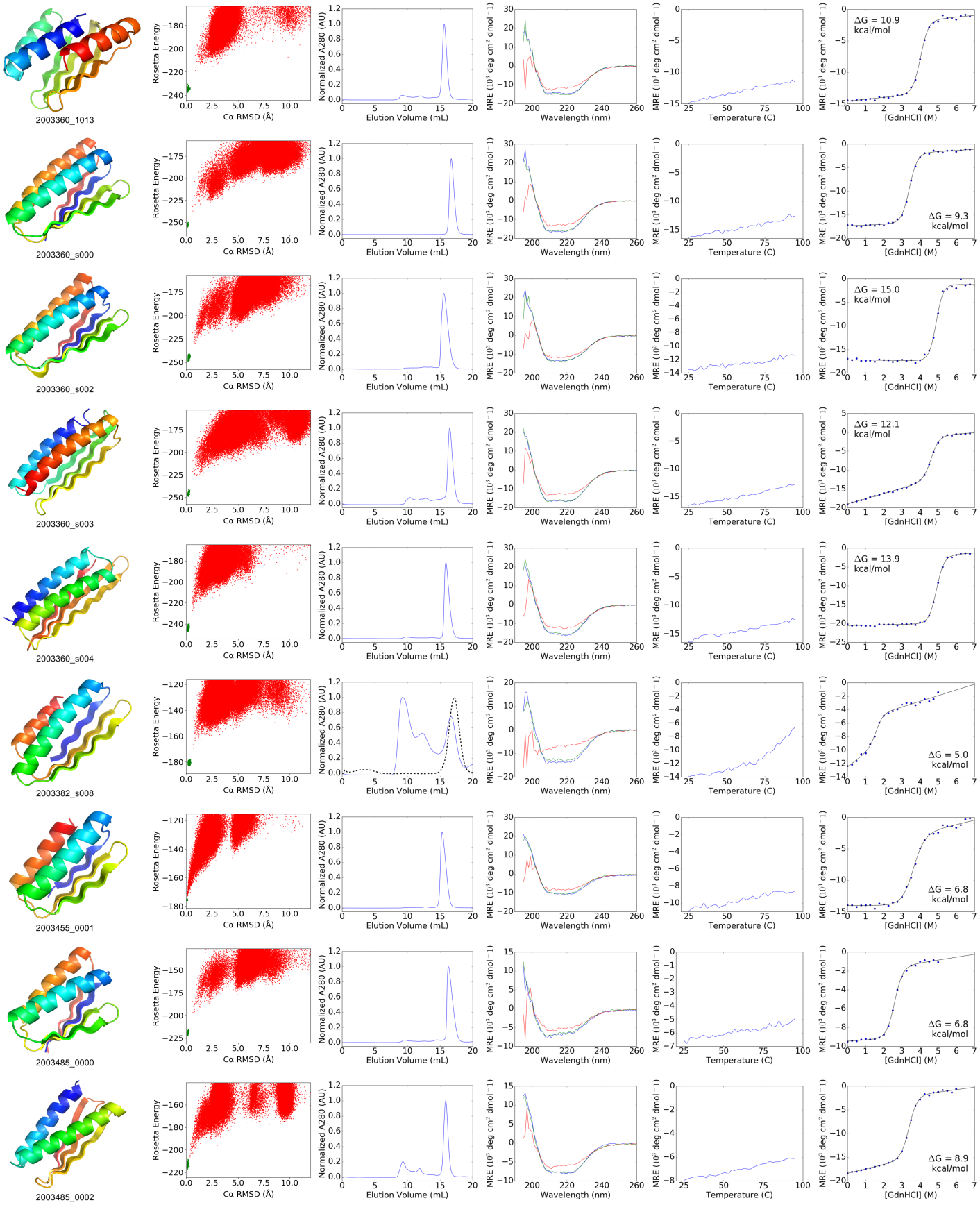

997258_0004
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE (10^3 deg cm^2 dmol^{-1})
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 0.5 kcal/mol
997523_0003
ΔG = 0.8 kcal/mol
997523_0008
ΔG = 17.8 kcal/mol
997523_1040
ΔG = 6.5 kcal/mol
997791_1027
ΔG = 8.2 kcal/mol
997915_0000
ΔG = 3.2 kcal/mol
1998925_s008
ΔG = 1.5 kcal/mol
2001044_0000
ΔG = 0.0 kcal/mol
2002243_1016
ΔG = 1.4 kcal/mol


2002308_0000
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE (10^3 deg cm^2 dmol^{-1})
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 2.9 kcal/mol
2002713_0000
ΔG = 2.8 kcal/mol
2002713_0004
ΔG = 1.5 kcal/mol
2002713_1006
ΔG = 0.8 kcal/mol
2002745_0000
ΔG = 4.8 kcal/mol
2002745_0001
ΔG = 5.3 kcal/mol
2002745_0003
ΔG = 9.9 kcal/mol
2002745_0004
ΔG = 2.0 kcal/mol
2002745_0008
ΔG = 0.0 kcal/mol


2002766_0000
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE (10^3 deg cm^2 dmol^{-1})
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 6.9 kcal/mol
2002766_0001
ΔG = 7.4 kcal/mol
2002766_0002
ΔG = 4.0 kcal/mol
2002922_s004
ΔG = 13.5 kcal/mol
2002949_0000
ΔG = 7.8 kcal/mol
2002990_0006
ΔG = 6.0 kcal/mol
2002990_1039
ΔG = 1.2 kcal/mol
2003169_s001
ΔG = 9.3 kcal/mol
2003169_S953
ΔG = 14.4 kcal/mol


2003205_0000
2003205_0002
2003205_s002
2003245_0004
2003245_S383
2003245_S385
2003265_0007
2003265_s005
2003265_s008
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE (10^3 deg cm^2 dmol^{-1})
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 8.1 kcal/mol
ΔG = 3.9 kcal/mol
ΔG = 5.6 kcal/mol
ΔG = 6.8 kcal/mol
ΔG = 12.5 kcal/mol
ΔG = 10.0 kcal/mol
ΔG = 5.6 kcal/mol
ΔG = 5.8 kcal/mol
ΔG = 13.9 kcal/mol


2003265_S115
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE ([10³ deg cm² dmol⁻¹)
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 10.3 kcal/mol
2003285_0000
ΔG = 4.9 kcal/mol
2003285_s005
ΔG = 1.9 kcal/mol
2003308_0004
ΔG = 14.3 kcal/mol
2003308_1010
ΔG = 4.9 kcal/mol
2003333_0005
ΔG = 5.6 kcal/mol
2003333_0006
ΔG = 6.4 kcal/mol
2003333_1006
ΔG = 4.2 kcal/mol
2003360_0005
ΔG = 11.0 kcal/mol


2003360_1013
Rosetta Energy
Cα RMSD (Å)
Normalized A280 (AU)
Elution Volume (mL)
MRE (10^3 deg cm^2 dmol^{-1})
Wavelength (nm)
Temperature (C)
[GdnHCl] (M)
ΔG = 10.9 kcal/mol
2003360_s000
ΔG = 9.3 kcal/mol
2003360_s002
ΔG = 15.0 kcal/mol
2003360_s003
ΔG = 12.1 kcal/mol
2003360_s004
ΔG = 13.9 kcal/mol
2003382_s008
ΔG = 5.0 kcal/mol
2003455_0001
ΔG = 6.8 kcal/mol
2003485_0000
ΔG = 6.8 kcal/mol
2003485_0002
ΔG = 8.9 kcal/mol

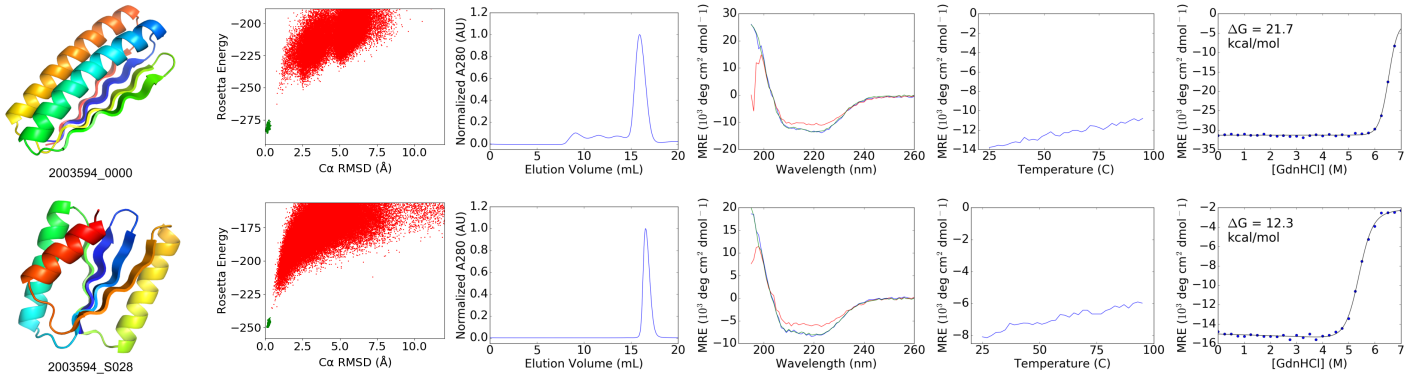


Figure S2. Biophysical characterization of successful designs. For each of the 56 successful Foldit designs is shown (left to right): a cartoon depiction the design; Rosetta@home ab initio calculations for the design; size-exclusion chromatography (SEC) traces of elution absorbance at 280 nm; circular dichroism (CD) spectra at 25ºC (blue trace), when heated to 95ºC (red trace), and when cooled again to 25ºC (green trace); CD mean residue ellipticity at 220 nm as temperature is increased from 25ºC to 95ºC; CD mean residue ellipticity at 220 nm with increasing concentration of guanidinium chloride 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].

Appendix A: Determining crystal structures through crowdsourcing and coursework

Scott Horowitz[1,2]*, Brian Koepnick[3]*, Raoul Martin[1,4]*, Agnes Tymieniecki[1,2], Amanda A. Winburn[5,6], Seth Cooper[7], Jeff Flatten[8], David S. Rogawski[9], Nicole M. Koropatkin[10], Tsinat T. Hailu[1,11], Neha Jain[1], Philipp Koldewey[1,2], Logan S. Ahlstrom[1,2], Matthew R. Chapman[1], Andrew P. Sikkema[12], Meredith A. Skiba[12], Finn P. Maloney[13], Felix R. M. Beinlich[1,14], Foldit players[15], University of Michigan students[16], Zoran Popović[8], David Baker[3,17,18], Firas Khatib[19] & James C. A. Bardwell[1,2]

[1]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. [2]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. [4]Biophysics Graduate Group, University of California, Berkeley, California 94720, USA. [5]Center for Complex Networks and Systems Research, Department of Informatics, Indiana University, Bloomington, Indiana 47408, USA. [6]Program in Cognitive Science, Indiana University, 1900 E 10th St, Bloomington, IN 47406, USA. [7]Northeastern University, College of Computer and Information Science, Boston, Massachusetts 02115, USA. [8]Department of Computer Science and Engineering, Center for Game Science, University of Washington, Seattle, Washington 98195, USA. [9]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109, USA. [10]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan 48109, USA. [11]QProQR Therapeutics N.V., Leiden, The Netherlands. [12]Department of Biological Chemistry and Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109, USA. [13]Chemical Biology Doctoral Program, University of Michigan, Ann Arbor, Michigan 48109, USA. [14]Institute of Complex Systems, Zelluläre Biophysik (ICS-4), Forschungszentrum Jülich, Germany. [15]Worldwide. [16]University of Michigan, Ann Arbor, Michigan 48109, USA. [17]Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. [18]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [19]Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, Massachusetts 02747, USA.

*These authors contributed equally to this work.

Abstract

We show here that computer gamers can produce crystal structures of exceptional quality. This was apparently facilitated by our introduction of a feature of into the computer game Foldit that allows players to build structures into electron density maps. To assess the usefulness of this new feature, we held a crystallographic model-building competition between trained crystallographers, biology undergraduate students, Foldit players and automatic model-building algorithms. After removal of disordered residues, a team of Foldit players achieved the most accurate structure. This Foldit structure proved to be of exceptional quality when compared with structures in the Protein Data Bank (PDB) of similar resolution. Analyzing the target protein of the competition, YPL067C, uncovered a new family of histidine triad proteins apparently involved in the prevention of amyloid toxicity. From this study, we conclude that crystallographers can utilize crowdsourcing to circumvent experimental phasing, and to produce structure solutions of the highest quality.

Main Text

"Macromolecular refinement against high-resolution data is never finished, only abandoned"[1]. George Sheldrick's statement on the labor-intensive nature of model building and refining crystal structures reflects the difficulty in producing highly accurate models. As a result, approximately 85% of deposited protein crystal structures contain discernable errors[2]. Unfortunately, as crystal structures are frequently used as the basis of further studies, inaccurate crystal structures can cause significant harm to the scientific process. Continued improvement of crystal structure accuracy therefore remains an important goal within the biology community.

Recently, in a class assignment, we asked 57 undergraduate students to build the structure of a protein, lectin scytovirin[3], using only the model-building program Coot and an electron density map downloaded from the electron density server[4]. Students were not given the amino acid sequence of the protein, but were provided with the position of the N-terminal amino acid. The students were instructed to build the structure of the protein, residue-by-residue, into the $2F_o$-F_c electron density map. Many students expressed appreciation for the puzzle-like quality of the assignment. In addition to learning about protein structure, ~10% of these students improved upon the previously published model[4]. One student even generated a structure that ranked in the 100th percentile in both Molprobity clashscore and total score when compared to other structures in its resolution range. These results raised the intriguing possibility that even a relatively small group of amateur model builders could collectively build higher quality models than a single trained crystallographer. This concept was remarkably reminiscent of ideas recently championed by the online protein-folding computer game Foldit.

Foldit is a popular video game that crowdsources protein structure prediction[5], challenging players to discover low-energy protein models by exploring protein conformational space. The newest version of Foldit provides players with electron density maps and the associated protein sequences, and asks players to use experimental data to guide protein folding (Fig. 1, Supplementary Fig. 1). With these new features, players can trim maps around a model and customize features of the electron density map, such as contour level, rendering style and transparency (Supplementary Fig. 1). The standard Foldit score function is supplemented with a fit-to-density term, allowing in-game structure minimization similar to crystallographic real-space refinement[6]. Foldit players are able to view the fit-to-density score for each residue of a model, providing valuable feedback about specific parts of a model that require more attention. As an preliminary test of this feature, we gave an electron density puzzle to the Foldit players that was nearly identical to the lectin scytovirin classroom assignment mentioned above, and found that

the Foldit players were also able to improve upon the published scytovirin structure (Supplementary Introduction, Supplementary Fig. 2).

Based on the success of both undergraduate students and Foldit players in improving a published crystal structure, we hypothesized that crystallographic model building through crowdsourcing might result in more accurate crystal structures than those resulting from traditional model-building methods. Thus, we held a crystallographic model-building competition. Five groups of competitors took part in our model-building competition: (1) 469 Foldit players worldwide, (2) two trained crystallographers, (3) 58 undergraduate students in the University of Michigan class MCDB411 (Introduction to Protein Structure and Function) who built the structure as a class assignment (Supplementary Introduction), (4) Phenix Autosolve[7, 8] and (5) MR-Rosetta[9]. We chose YPL067C, a yeast protein with no significant sequence similarity to any structure in the PDB (Supplementary Results) as the target for our competition.YPL067C was also chosen because of its biological interest, as it had previously been shown that null mutants in its gene are sensitive to amyloid production, implying that YPL067C is involved in preventing amyloid toxicity[10]. Crystals of this protein diffracted to 1.9 Å resolution (Supplementary Table 1). We asked all human competitors to build the best possible protein structure that they could given the protein sequence, a secondary structure prediction and an experimentally-phased, density refined map of YPL067C. The MCDB411 class assignment was conducted similarly to a previous crystallography assignment discussed above, in which undergraduates improved upon a published crystal structure[4]. As before, the undergraduates lacked previous model-building experience. In contrast, 54% of the participating Foldit players had attempted to solve early electron density puzzles in Foldit and thus had some experience in Foldit-based model building. Both the trained crystallographers and undergraduates used the model-building and real-space refinement program Coot[11], whereas the Foldit players used the Foldit version released on October 14, 2015. None of the competitors were given any starting points for building, and thus they needed to establish the relationship between the electron density and the protein's sequence.

The various groups used different approaches. The students and trained crystallographers worked independently, generally utilizing large aromatic residues to identify the relationship between sequence and electron density features. The best Foldit solutions, in contrast, came from a group of players working collaboratively, with one player serving as the trailblazer, making the majority of the substantive moves (Fig. 1), and other players providing detailed structural tweaks and refinements (Supplementary Movie 1). Although the Foldit players also used a large aromatic residue to initially anchor the sequence similar to the Coot users, their building process sampled conformational space very widely (Supplementary Movie 1), unlike the Coot users. The Foldit players made heavy use of XX moves… (Supplementary Figure 3).

To determine which group produced the best structural models, we used an automated refinement procedure[12] on all the structures, and then compared key crystallographic statistics. These statistics consisted of R_{free}, root-mean-square deviations (RMSDs) of bonds and angles, the number and severity of steric clashes (represented by Molprobity clashscore[13]) and Ramachandran outliers. Analyzing the R_{free} values, we quickly realized that the Foldit players were at a distinct disadvantage (Supplementary Fig. 4). Whereas the Coot users, Phenix Autosolve and MR-Rosetta were able to exclude regions that had poor quality electron density, the current version of Foldit required that players model the entire sequence without gaps, retaining all disordered residues. To correct for this deficiency in Foldit, we pruned the Foldit models afterwards to include only those residues modeled by the top trained crystallographer, so that they contained only the well-ordered regions of the protein. We then re-refined the pruned Foldit structures and compared them to the other models.

The Foldit structures improved considerably after pruning. As a result, the top pruned Foldit structure was the overall highest quality structure produced in the competition (Fig. 2) as measured by geometry, density fit and steric clashes. In addition to the R_{free} value of the pruned Foldit structure becoming marginally better than that achieved by the trained crystallographers and containing zero Ramachandran outliers, the Foldit structure had the lowest level of steric clashes (Fig. 2). According to Molprobity[13], the top Foldit structure is an exceptional model, ranking in the 100th percentile in both its overall Molprobity score and clashscore of all structures in the PDB of similar resolution (1.95 ± 0.25 Å). The superiority of the top Foldit structure can be attributed to better side-chain conformations than those in the top structure produced by the trained crystallographers (Supplementary Fig. 5). Better Foldit scores were associated with lower R_{free} values (Supplementary Figure 6), suggesting that the Foldit model-building strategy and its scoring algorithm could be generalizable as a means of producing high-quality structures.

This surprising win by Foldit suggests that, in at least some cases, this video game can help produce crystallographic models of higher quality than those from trained crystallographers or automated model-building algorithms. The difference in accuracy is likely in part the result of different underlying philosophies behind Coot and Foldit. Whereas Coot primarily uses a real-space refinement system[14] that only respects local geometry, the Rosetta force field used by Foldit is much more extensive, including additional steric, electrostatic and solvation terms, as well as statistical potentials based on observed backbone torsions and side-chain rotamers[6]. That some of the Foldit models were of higher quality than those of trained crystallographers suggests that expert model builders might also benefit from the Foldit force field for real-space refinement. Human intervention either by crystallographers or Foldit players is clearly helpful, as both Phenix Autosolve and MR-Rosetta on their own produced suboptimal structures. The collaborative building process used by the Foldit players could also be a beneficial strategy for professional crystallographers, who could achieve a similar effect by having multiple laboratory members take turns working on model building and refinement. Looking forward, we hope that further analysis of electron density puzzle solutions in Foldit could inform continued improvement of automated structure solution algorithms.

Here we show that Foldit players can build structural models of higher quality than trained crystallographers or automated methods, enabling a novel crowd-powered strategy for solving high-accuracy crystal structures. Citizens hold a tremendous reserve of brainpower that remains largely untapped by the scientific community. The new electron density feature and the highly-accurate model building provided by Foldit players, combined with the ability to generate molecular replacement solutions in Foldit[15] makes it now feasible to obtain a complete structure solution using Foldit given only a native crystallography dataset. Thus, the need for experimental phase determination as well as a substantial portion of the model building and refinement process can be circumvented. We have shown that non-expert Foldit players are capable of using structural data to build first-rate models, and we expect that Foldit will be a powerful tool for crowdsourcing many new high quality structures.

Foldit players might also be tasked with improving structures of questionable quality already in the PDB. These Foldit puzzles would benefit the entire community of scientists that depend upon accurate structural models. Comparing the results here to our previous study, in which students were able to readily improve upon a published crystal structure, suggests that validation and editing of nearly complete structures could form a base of easier Foldit puzzles, allowing players to graduate to more difficult *de novo* model-building puzzles with increased practice. To further improve the capability for Foldit players to aid crystallographers, ongoing development will make it possible for the players to add or remove residues with insufficient electron density, and have these changes accurately reflected in the Foldit score. We envision a future in which professional crystallographers frequently tap the collective

model-building expertise of Foldit players for help in the model-building, refinement, and validation steps of crystallography.

From an educational perspective, the participation of an undergraduate class in this study explored how crystallographic model building can be used not just to teach students the structures and chemistry of proteins in great depth, but in addition, to teach the scientific process. In our previous study, students built into a high-resolution, fully refined map[3]. In the study presented here, students received a lower resolution, unrefined map, and no starting place for building. In dealing with disordered residues for instance, the students were forced to interpret data of varying quality and to decide when the data became too ambiguous to draw firm conclusions. Similar decision-making processes govern the use of scientific data across many disciplines. This assignment thus helped give students a realistic view about the power and limitations of the scientific method. Importantly, the ease with which students and Foldit players were able to interpret and understand density maps suggests that scientists other than crystallographers can very readily interpret electron density maps, which will assist them in designing or analyzing experiments based on crystal structures.

YPL067C's structure yielded unexpected insights into its biological function. Despite the lack of sequence homology to any protein in the PDB, a DALI[16] search of YPL067C (Supplementary Table 2) found that it is structurally similar to members of the widely conserved superfamily of histidine triad (HIT) proteins. These proteins contain three histidine residues with an almost identical spatial organization to that of YPL067C, as well as a ß-sheet core nested inside a similar arrangement of loops and helices (Fig. 3). HIT proteins have been shown to be involved in diverse cellular stress responses, such as DNA damage, oxidative stress and induced apoptosis[17, 18]. However, the specific *in vivo* activity of HIT proteins remains unclear[17]. Although YPL067C bears some resemblance to known HIT proteins, it is sequentially and structurally distinct (Supplementary Fig. 7). Its most notable distinguishing structural characteristic is an open channel not found in other HIT proteins (Supplementary Results). YPL067C's characterization makes it the founding member of a new family we are calling HTC (for histidine triad channel), with YPL067C being the first member, HTC1. The HTC family contains over 900 members found in a wide variety of eukaryotes and viruses (Supplementary Results). As mentioned above, HTC1 null mutants increase the toxicity of amyloid overproduction[10]. We find here that HTC1 is very effective in preventing *in vitro* amyloid formation of three model proteins, $A\beta_{1-40}$, α-synuclein and reduced carboxy-methylated α-lactalbumin (RCMalA) (Fig. 4). Based on docking simulations, HTC1 may bind to unfolded proteins using its conserved channel (Supplementary Results, Supplementary Fig. 8). Our crowdsourcing-enabled discovery of a new family of proteins involved in preventing amyloid formation provides insight into a novel physiological role of the ubiquitous HIT proteins.

METHODS

Electron density in Foldit

To facilitate work on electron density data in Foldit, new visualizations and tools, along with a tutorial puzzle to introduce them, were developed and distributed to Foldit players in periodic software updates. Electron density maps in Foldit are displayed as a visual guide in the form of an isosurface. Players have control over parameters of the density isosurface, such as the contour level, surface texture, transparency and color, and can tag regions of the density with notes. After initial testing, it was clear that density visualization alone was insufficient to improve model building by Foldit players. Players simply ignored the density, finding that their existing, familiar strategies were most competitive on Foldit leaderboards. In response, we adapted the Rosetta fit-to-density score term elec_dens_fast into the Foldit score function[6]. This not only provides competitive incentive to match the density, but also allows Foldit automated tools like structure minimization to be guided by electron density, similar to crystallographic real-space

refinement. Under this configuration, players were able to fit models to several experimental electron density maps with high accuracy. An important feature was added later that allowed players to trim excess density that was distant from the player's model from the visualization. According to Foldit player testimony, this feature has proved invaluable on certain experimental density maps where it allowed a clearer interpretation of relevant density. To protect the integrity of unpublished crystallographic work, electron density data were obfuscated before online distribution to Foldit players.

The competition

Phenix Autosolve[7, 8], with model building disabled, was used to create density-modified maps of selenenomethionine (SeMet) YPL067C. To make the map manageable for the Foldit program, the map was masked beyond 5 Å from the initial solution at the start of the competition. This map was given to Foldit players, MCDB 411 students and the experienced crystallographers for model building. The individual responsible for model building and refinement of the initial solution of YPL067Cbefore the contest was initiated had no contact with any of the competitors.

58 students in the University of Michigan undergraduate class MCDB411 (Introduction to Protein Structure and Function) were introduced to the assignment through a description of the previous iteration of the assignment in class[4], together with a 1.5 h lecture on X-ray crystallography. Students then had two in-class computer laboratory sessions in which features of Coot were presented. In the first lab session, the students were given basic instructions on opening electron density maps and molecules, changing map levels, scrolling and changing map size, finding secondary structure elements, converting C_α representations to all-atom molecules, placing helices and strands, adding terminal residues, real-space refinement, controlling regularization and refinement, rotating and translating atoms and residues, viewing the skeleton, mutating residues, and changing rotamers. The instructors suggested that changing the weighting of the real-space refinement from the default value of 60 to 10 and making subsequent changes to this value as needed could help in the building process. In the second lab session, the students were taught how to merge molecules, look for grouped tryptophans, phenylalanines and/or tyrosines as starting places for building, and use validation tools such as density fit analysis, geometry fit analysis and unmodelled blobs. Four instructors were present in the first lab session and three in the second to answer questions on the operation of Coot. Starting from the initial lab session, students were given a total of one month to complete the assignment. During this period, the instructor held walk-in help sessions twice a week for 1.5 h each and answered questions on the operation of Coot as well as general model-building questions. Common questions included how to identify density for specific sequences, how to correctly merge molecules and how to approach gaps in electron density. Regarding gaps in density, students were told to model through gaps only if they were confident that the modeling would be correct based on the size of the gap and the number of residues they were modeling in. Students were not told what to do in specific cases of building through disordered segments. They were informed that water molecules would not be included in grading. One student asked if there were external validation tools that could help and was told that the Molprobity server might be useful. Students were allowed to discuss the project and ask each other questions, but were required to do their own model building.

A Foldit puzzle was posted online with the masked electron density map and a model of the target polypeptide in fully extended conformation. Players were challenged to fold the extended polypeptide into the electron density map to achieve a good fit to density. Any advice given to MCDB411 students by the instructors as to how to begin model building was also posted on the Foldit messaging board. After four weeks, the puzzle was closed and 900,000 player models were scored and ranked according to the Rosetta energy function. The top scoring models were clustered into a set of 1000 such that no two aligned to < 1.0 Å C_α RMSD. To this clustered set we added the 50 best unique models produced by

Foldit teams or soloists, as well as any models flagged by Foldit players for special consideration—1094 Foldit models in total. Two trained crystallographers were given the same number of days for model building as the students and Foldit players. They were given specific instructions not to use tools outside of Coot or Molprobity and not to interact with each other during model building. The trained crystallographers spent approximately eight and fourteen hours, respectively, working on the puzzle. The trained crystallographers reported using the following approach to the puzzle, which corresponds well with what the instructor observed with many of the undergraduate students. First, they looked for large density blobs that might correspond to large aromatic side chains like Trp, Tyr, or Phe. Working forwards and backwards from the Trp-Phe-Val-Asn sequence proved particularly useful. Modeling in a few of these large residues led to the assignment of density to sequence location. The direction of the polypeptide chain was reversed on a few occasions, but was fixed by looking at the carbonyl density. The Find Secondary Structure tool in Coot was used, especially for regions where the density was poor. Real Space Refine Zone was used with the refinement weight set to 20 or 10, based on the instructor's suggestion for building in an unrefined map. Regions where the density was very poor and decisions had to be made about whether to keep trying to build or not proved to be the hardest part of the task. The trained crystallographers reported that at first they did build in these sections of poor electron density. However, when they realized the extent of the guessing involved, they subsequently removed most of the model in these areas. After modeling in the residues, the trained crystallographers used the validation tools in Coot, including Ramachandran plot, Rotamer analysis and Density Fit analysis, which flagged areas with poor geometry. They also ran the structure through MolProbity, which gave similar results to the Coot validation tools. Finally, the crystallographers fixed problem areas as best as possible with the Coot modeling tools, such as Flip Peptide, Rotamers, Regularize Zone and Real Space Refine Zone. When asked to describe the difficulty level of this assignment, the trained crystallographers rated it as somewhat difficult (on a scale of: very difficult, somewhat difficult, neither easy nor difficult, somewhat easy, very easy).

Phenix Autosolve[7, 8] was run with default parameters (using phase_and_build) to produce the Autosolve model. The MR-Rosetta model was obtained by relaxing and rebuilding the Autosolve model in the same electron density map provided to human groups, using Rosetta mr_protocols[9] with nstruct=10 and selecting the model with the lowest R_{free}. ARP/Warp[19] and Phenix Autobuild[20] did not create models of as high quality as Phenix Autosolve or MR-Rosetta, and were thus not analyzed in the competition.

After completion of the competition, all structures were automatically refined using Phenix to analyze the results. The refinement strategy included XYZ coordinates, temperature factors and updating waters. Notably, the best structures from Foldit, as measured by R_{Free}, came from the group of highest-scoring Foldit structures according to Foldit score.

Bioinformatics

YPL067C sequence conservation was analyzed using a four-iteration PSI-BLAST of the UniRef50 database, with an E-value cutoff of 0.005. No sequences in the PDB were found. Sequence conservation was projected onto the structure of YPL067C using the Consurf server[21]. The top DALI[16] match to the crystal structure of YPL067C was to a HIT protein of unknown function from *Clostridium difficile* (PDB: 4EGU), with a Z-score of 4.9. The top 47 hits were all HIT proteins with Z-scores ranging from 4.9 to 4.2 (Z-scores greater than 2.0 are considered significant). Secondary structure predictions for the competition were generated using PSIPRED[22].

Protein expression and purification

The gene for YPL067C was amplified from yeast strain Y2HGold (Clontech) and cloned into a

pET28-sumo plasmid using primer 1 (5'-AAATATGGATCCATGCAACAAGATATCGTCAACGATCACCAG-3') and primer 2 (5'-AAATATCTCGAGTCAGGCAAGTGGCTCGAAACC-3'). pET28-sumo-ypl067C was transformed into *Escherichia coli* BL21(DE3) cells.

Cells were grown at 37 °C overnight in 100 ml Luria Broth (LB) (containing 100 μg ml^{-1} kanamycin), and 10 ml were used to inoculate 1 liter LB (containing 100 μg ml^{-1} kanamycin). At early log phase, the temperature was reduced to 20 °C and 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce expression of our construct overnight. Cells were harvested by centrifugation and resuspended in 100 ml lysis buffer (40 mM Tris, 10 mM sodium phosphate, 400 mM NaCl, 10% glycerol, 10 mM imidazole, pH 8.0) enriched with 1 mg ml^{-1} DNaseI, 1 mM $MgCl_2$ and 2 tablets of complete EDTA-free protease inhibitor (Roche). Cells were lysed by two French press cycles at 1300 PSI and centrifuged at 37,000*g* for 30 min at 4 °C. The supernatant was run through a Ni-HisTrap 5 ml column (GE Healthcare) pre-equilibrated with lysis buffer at a rate of 1.5 ml min^{-1}. Following binding, the column was washed with 60 ml lysis buffer. The protein was eluted with 20 ml lysis buffer enriched with 500 mM imidazole. To cleave the sumo-His×6 tag, 10 μl ULP1 protease (from stock of 50 mg ml^{-1}) was added to the eluted solution. 10 μl β-mercaptoethanol was added and the solution was dialyzed overnight in 40 mM Tris, 10 mM sodium phosphate, 400 mM NaCl, 10% glycerol, pH 8.0. To remove the tag, the solution was run through a Ni-HisTrap 5 ml column (GE Healthcare) pre-equilibrated with dialysis buffer at a rate of 1.5 ml min^{-1}, and the flowthrough was saved and diluted in 8 volumes of 20 mM Tris, pH 8.0. The protein was then run through a HiTrap Q HP 5 ml column (GE Healthcare), and the flowthrough contained greater than 95% pure YPL067C as measured by SDS-PAGE. Prior to each experiment, YPL067C was exchanged into appropriate buffer. Expression and purification of SeMet YPL067C was performed with the same protocol except a methionine auxotroph variant of *E. coli* BL21(DE3) and SelenoMethionine Medium Complete (Molecular Dimensions) were used.

α-synuclein was expressed and purified using the protocol described previously[23] with minor modifications. Briefly, 1% of the overnight grown culture was transferred in fresh media and induced with 0.8 mM IPTG for 4 h after the O.D. of the culture reached 0.6. The induced cells were pelleted at 4,000 r.p.m. and resuspended in 25 ml lysis buffer (10 mM Tris, 1 mM EDTA, pH 8). The lysed cells were then boiled at 95 °C for 15–20 min and centrifuged at 11,000 r.p.m. for 20 min. The supernatant was thoroughly mixed with 10% streptomycin sulfate (136 μl ml^{-1}) and glacial acetic acid (228 μl ml^{-1}) then centrifuged at 11,000 r.p.m. for 30 min. To the clear supernatant, an equal volume of saturated ammonium sulfate was added, and the solution was incubated at 4 °C for 1 h with intermittent mixing. The precipitated protein was separated by centrifugation at 11,000 r.p.m. for 30 min. The pellet was dissolved in equal volumes of absolute ethanol (chilled) and 100 mM ammonium acetate. Finally, the pellet was washed (twice; optional) with absolute ethanol, dried at room temperatrue and resuspended in 10 mM Tris, pH 7.4. The protein solution was filtered through a 50 kDa cutoff column (AMICON, Millipore) followed by ion-exchange chromatography (Q-sepharose) against a NaCl gradient. The fractions of pure protein eluted at ~300 mM NaCl were checked on SDS-PAGE and the molecular weight was confirmed by mass spectrometry. The pure fractions were pooled and dialysed overnight against buffer (10 mM Tris and 50 mM NaCl, pH 7.4). The concentration of α-synuclein was determined using ε_{280} = 5,600 $M^{-1}cm^{-1}$. The purified α-synuclein was stored at -80 °C at a concentration of ~100 μM until use.

$A\beta_{1-40}$ peptide was purchased from AlexoTech AB (Umeå, Sweden) and prepared as previously described[24]. $A\beta_{1-40}$ peptide was dissolved in 10 mM NaOH to a peptide concentration of 1 mg ml^{-1} and then sonicated for 1 min in an ice bath before dilution in the assay buffer. The preparations were kept on ice. α-lactalbumin (aLA) from bovine milk (cat: L6010) and porcine citrate synthase (cat: C3260-5KU)

were purchased from Sigma Inc. RCMaLA was prepared as previously described[25]. 500 μM aLA (freshly prepared in water) was incubated with 1 mM DTT in 0.5 M Tris, 1 mM EDTA, pH 7.0 for 10 min, then 3 mM iodoacetic acid (out of 1 M stock solution in water) was added and the solution incubated for another 30 min. aLA was then dialyzed into 50 mM phosphate buffer, pH 7.0, 100 mM KCl, 10 mM $MgCl_2$.

Protein crystallization

Native and SeMet YPL067C crystals were grown at 20 °C by vapor diffusion using both sitting (1 μl drops) and hanging drop methods (2 μl drops). Drops were prepared by mixing a 1:1 solution of YPL067C (25 mg ml^{-1}) and reservoir solution (5.6–8.1% glycerol, 1.6–2.1 M ammonium sulfate and 0.1–0.2 M Tris). Crystals were cryoprotected by gradually supplementing the drop with glycerol up to 25% and were flash frozen in liquid nitrogen.

X-ray crystallography

Data were collected at the Life Sciences Collaborative Access Team (LS-CAT) beamlines at the Argonne National Laboratory's Advanced Photon Source at 100 K. The data were integrated and scaled using HKL2000. Phases and initial model building of the SeMet derivative were obtained using Phenix AutoSolve[7, 8]. Native YPL067C was solved by molecular replacement with the initial SeMet structure. Iterative refinement and model building was performed using Phenix Refine[12] and Coot[11]. Channel size was analyzed using the 3V server[26]. Data collection and modeling statistics are shown in Supplementary Table 1.

Fibrillar aggregation assays

Fibrillar aggregation was monitored by a thioflavine T (ThT) fluorescence assay. ThT is a benzothiazole dye that exhibits enhanced fluorescence specifically upon binding to amyloid fibrils. For RCMaLA aggregation experiments, solutions containing 100 μM RCMaLA, YPL067C in varying concentrations and 20 μM ThT were prepared in 50 mM potassium phosphate buffer, pH 7.0, 100 mM KCl and 10 mM $MgCl_2$[25]. The ThT fluorescence assays with $A\beta_{1-40}$ peptide were performed with 2.5 μM $A\beta_{1-40}$ peptide, YPL067C in varying concentrations and 20 μM ThT in PBS, pH 7.4, 1% DMSO. The fibrillar aggregation of α-synuclein was tested in a solution of 70 μM α-synuclein, YPL067C in desired concentrations and 20 μM ThT in PBS, pH 7.4. For α-syuclein assays, 4 glass beads were added in each well to induce aggregation.

ThT fluorescence assays were performed with a final volume of 100 μl of the prepared solution in black 96-microwell plates (costar, UV Plate, 96 well) that were sealed to prevent evaporation. ThT fluorescence was measured in a Synergy HT Multi-Mode Microplate Reader (Biotek) at 37 °C, with constant medium shaking. Excitation and emission wavelengths were 440 nm and 490 nm, respectively. All samples were assayed in triplicate and the assay was repeated twice. Incubation of YPL067C with ThT alone produced no fluorescence increase.

Docking of α-synuclein and HTC1

HTC1 was docked against a 200-member NMR ensemble of α-synuclein[27] using ZDOCK 3.0.2[28]. The top five scoring poses of HTC1 bound to each member of the ensemble were used to generate a contact frequency map of the HTC1:α-synuclein interaction. To determine the contact map, an interaction was assigned to a given residue pair if their Cα-Cα distance was less than or equal to $\lambda \bullet r_{ij}^{min}$, where λ = 1.2 and r_{ij}^{min} is taken from the mean Cα-Cα distance for residue pairs that form intermolecular contacts in the PDB[29]. For each intermolecular residue pair, we reported the contact probability averaged over the

extracted binding poses. To project the contact maps onto the structures of α-synuclein and HTC1 on the same scale, the contact frequency for each residue pair was averaged over all residues.

Analytical ultracentrifugation

Sedimentation velocity experiments of HTC1 were performed using a Beckman ProteomeLab XL-I analytical ultracentrifuge (Beckman Coulter). YPL067C was first dialyzed against 20 mM HEPES, pH 7.5, then diluted to a concentration of 20 or 200 μM using the dialysis buffer. Samples were loaded into cells containing standard sector shaped 2-channel Epon-centerpieces with 1.2 cm path-length (Beckman Coulter) and equilibrated to 22 °C in the centrifuge for at least 1 h prior to sedimentation. All samples were spun at 48,000 r.p.m. in a Beckman AN-50 Ti rotor, and the sedimentation of the protein was monitored continuously using the interference optics. Data analysis was conducted with SEDFIT (version 14.1)[30], using the continuous c(s) distribution model. The confidence level for the ME (maximum entropy) regularization was set to 0.7. Buffer density and viscosity were calculated using SEDNTERP (http://sednterp.unh.edu/).

References

1. Sheldrick GM. A short history of SHELX. *Acta Crystallogr A* **64**, 112-122 (2008).

2. Touw WG, Joosten RP, Vriend G. New Biological Insights from Better Structure Models. *J Mol Biol*, (2016).

3. Moulaei T*, et al.* Atomic-resolution crystal structure of the antiviral lectin scytovirin. *Protein Sci* **16**, 2756-2760 (2007).

4. Horowitz S, Koldewey P, Bardwell JC. Undergraduates improve upon published crystal structure in class assignment. *Biochem Mol Biol Educ* **42**, 398-404 (2014).

5. Cooper S*, et al.* Predicting protein structures with a multiplayer online game. *Nature* **466**, 756-760 (2010).

6. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D. Refinement of protein structures into low-resolution density maps using rosetta. *J Mol Biol* **392**, 181-190 (2009).

7. Adams PD*, et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).

8. Terwilliger TC*, et al.* Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr* **65**, 582-601 (2009).

9. Terwilliger TC*, et al.* phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *J Struct Funct Genomics* **13**, 81-90 (2012).

10. Willingham S, Outeiro TF, DeVit MJ, Lindquist SL, Muchowski PJ. Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein. *Science* **302**, 1769-1772 (2003).

11. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).

12. Afonine PV*, et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D* **68**, 352-367 (2012).

13. Chen VB*, et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).

14. Diamond R. Real-Space Refinement Procedure for Proteins. *Acta Crystall a-Crys* **A 27**, 436-& (1971).

15. 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).

16. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**, W545-W549 (2010).

17. Huebner K, Saldivar JC, Sun J, Shibata H, Druck T. Hits, Fhits and Nits: beyond enzymatic function. *Adv Enzyme Regul* **51**, 208-217 (2011).

18. Trapasso F*, et al.* Fhit interaction with ferredoxin reductase triggers generation of reactive oxygen species and apoptosis of cancer cells. *J Biol Chem* **283**, 13736-13744 (2008).

19. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **3**, 1171-1179 (2008).

20. Terwilliger TC*, et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* **64**, 61-69 (2008).

21. Celniker G*, et al.* ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr J Chem* **53**, 199-206 (2013).

22. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195-202 (1999).

23. Jain N, Bhasne K, Hemaswasthi M, Mukhopadhyay S. Structural and Dynamical Insights into the Membrane-Bound alpha-Synuclein. *Plos One* **8**, (2013).

24. Luo JH, Warmlander SKTS, Graslund A, Abrahams JP. Non-chaperone Proteins Can Inhibit Aggregation and Cytotoxicity of Alzheimer Amyloid beta Peptide. *Journal of Biological Chemistry* **289**, 27766-27775 (2014).

25. Kulig M, Ecroyd H. The small heat-shock protein alpha B-crystallin uses different mechanisms of chaperone action to prevent the amorphous versus fibrillar aggregation of alpha-lactalbumin. *Biochem J* **448**, 343-352 (2012).

26. Voss NR, Gerstein M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res* **38**, W555-562 (2010).

27. Schwalbe M*, et al.* Predictive atomic resolution descriptions of intrinsically disordered hTau40 and alpha-synuclein in solution from NMR and small angle scattering. *Structure* **22**, 238-249 (2014).

28. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771-1773 (2014).

29. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology* **256**, 623-644 (1996).

30. Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. *Biophys J* **78**, 1606-1619 (2000).

31. Moult J. The current state of the art in protein structure prediction. *Curr Opin Biotech* **7**, 422-427 (1996).

32. Stokes-Rees I, Sliz P. Protein structure determination by exhaustive search of Protein Data Bank derived databases. *Proc Natl Acad Sci U S A* **107**, 21476-21481 (2010).

33. Finn RD*, et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285 (2016).

34. Mitchell A*, et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**, D213-221 (2015).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank K. Wan for protein purification, Z. Wawrzak for phasing advice and U. Jakob for reading and editing the manuscript. This work was supported by National Institutes of Health (NIH) grants GM102829 to J.C.A.B. and GM118651 to M.R.C. D.S.R. was supported by the University of Michigan Chemistry-Biology Interface (CBI) training program (NIH grant 5T32GM008597) and the University of Michigan Medical Scientist Training Program (NIH grant 5T32GM007863). M.S. was partially supported by a predoctoral fellowship from the Cellular Biotechnology Training Program (T32GM008353). N.M.K. was supported by the University of Michigan Medical School Host Microbiome Initiative. A.S. was supported by the Molecular Biophysics Training Program (GM008270). B.K. acknowledges support by the National Science Foundation Graduate Research Fellowship Program (DGE-1256082). D.B. and J.C.A.B. are Howard Hughes Investigators. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817).

Author Contributions: The competition was designed by F.K., S.H. and B.K. and was implemented by S.H., B.K., F.K., N.K., D.R., A.S., F.M. and M.S. Foldit design and improvements were carried out by J.F. and S.C. Experiments were designed by J.C.A.B., A.T., T.H., S.H., P.K. and R.M. Experiments were carried out by A.T., P.K., T.H., F.B. and R.M. N.J. and M.R.C. made experimental reagents. Analysis was performed by B.K., L.S., P.K., F.K., A.W., A.T., D.B., S.H. and J.C.A.B. The paper was written by S.H., J.C.A.B., R.M., A.T. and B.K., with assistance from all authors. U.M.S. and F.P. participated in the competition.

Author Information: The final model of HTC1 is deposited in the PDB. Authors declare no conflicts of interest. Reprints and permissions information are available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.H. (horowsah@umich.edu), F.K. (fkhatib@umassd.edu), or J.C.A.B. (jbardwel@umich.edu).

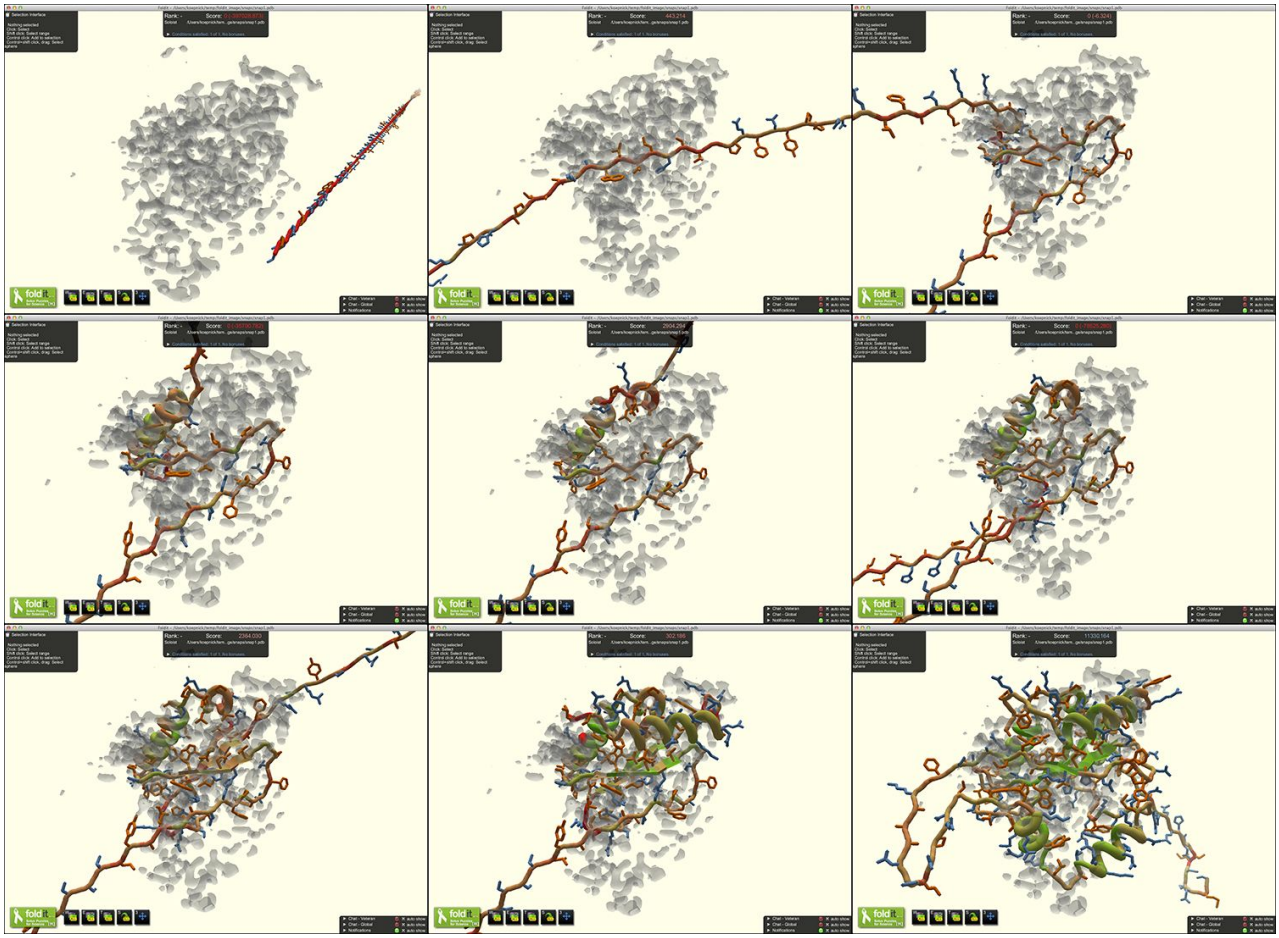


Figure 1 | Snapshots of Foldit players building YPL067C in the Foldit user interface. For the complete path of the Foldit model building, see Supplementary Movie 1. The starting point for the puzzle (top left) presented the electron density map and the protein sequence to the player. The players then used Trp108 to help anchor the sequence in electron density (top middle) before beginning to fold secondary structure elements (top right through bottom left). After many rounds of modification in Foldit (bottom middle and Supplementary Movie 1), the players arrived at a high-scoring solution in which the ordered regions of electron density were well fit by YPL067C (bottom right). Disordered regions were later pruned.

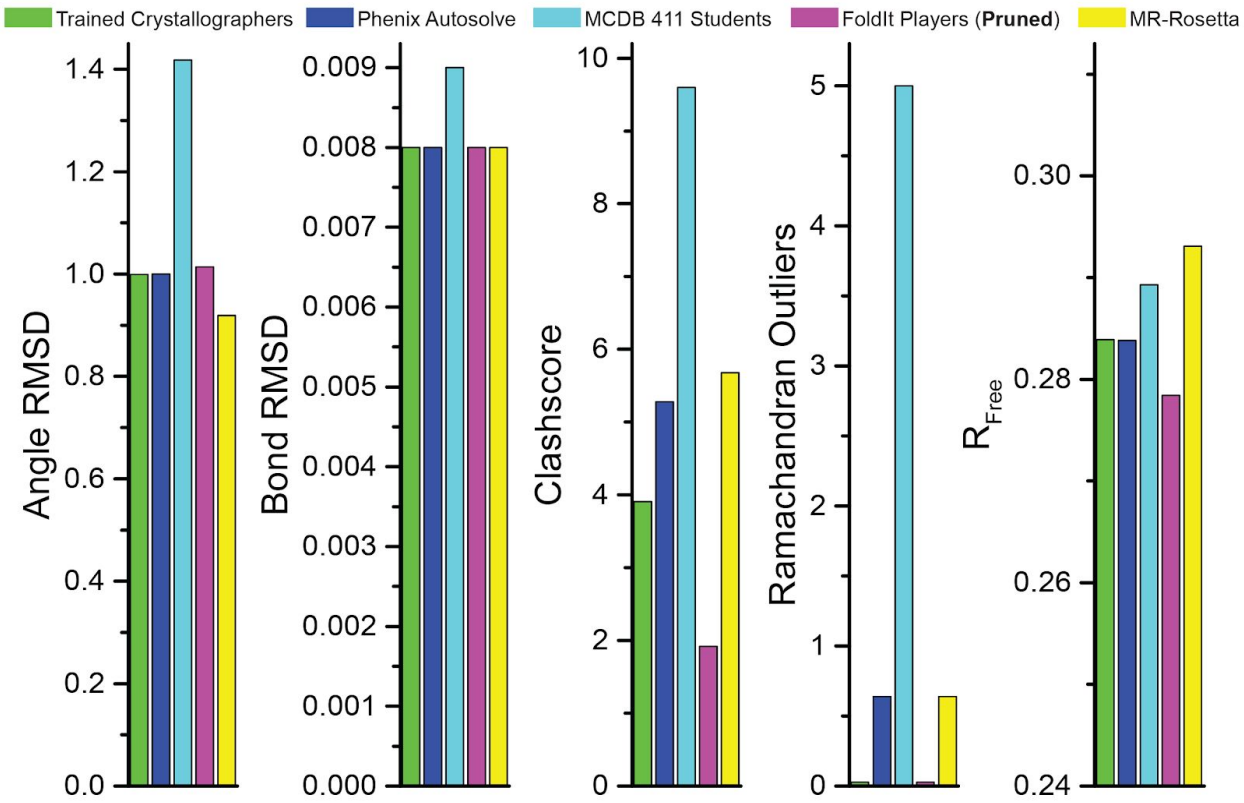


Figure 2 | Model-building competition results. Comparison of key statistics of the best model from each group after pruning disordered residues from Foldit structures. In all cases, lower values represent better scores. Comparison before pruning disordered residues is shown in Supplementary Figure 4.

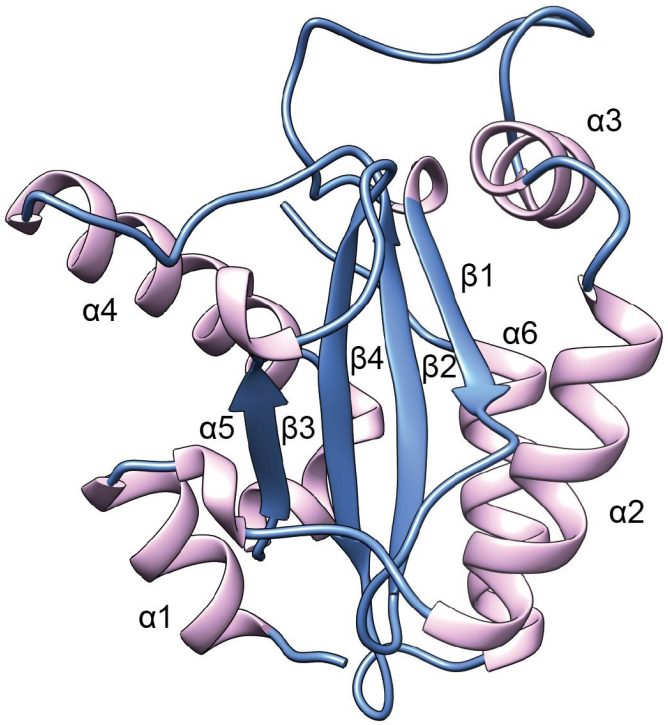


Figure 3 | Overall structure of HTC1. Structural alignment with the top DALI search hit is shown in Supplementary Figure 7.

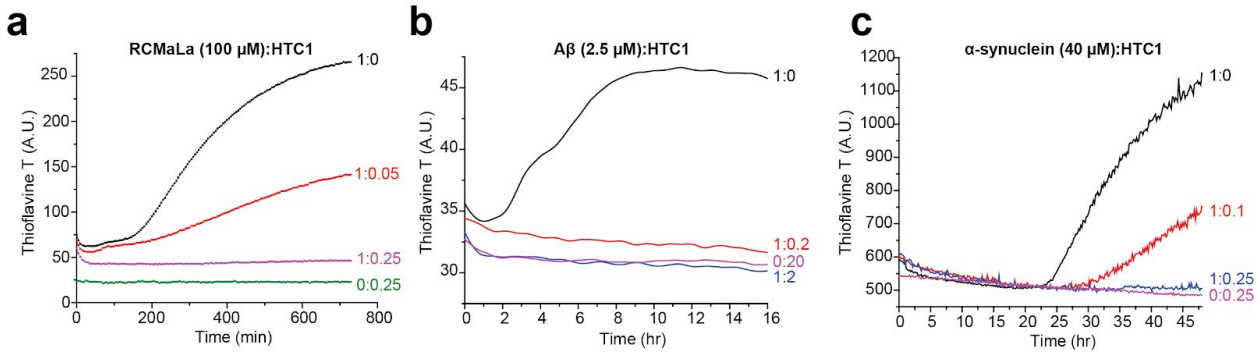
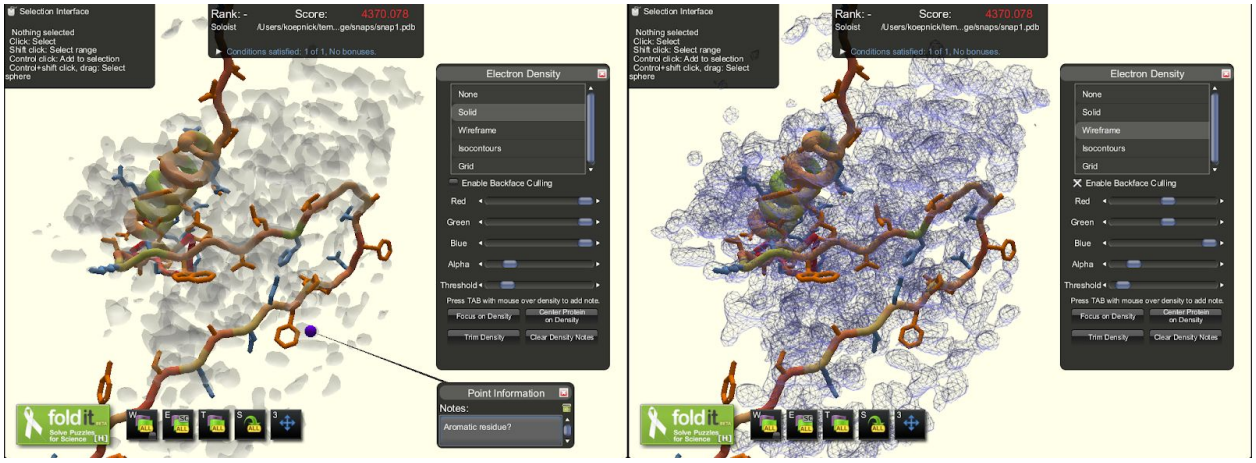
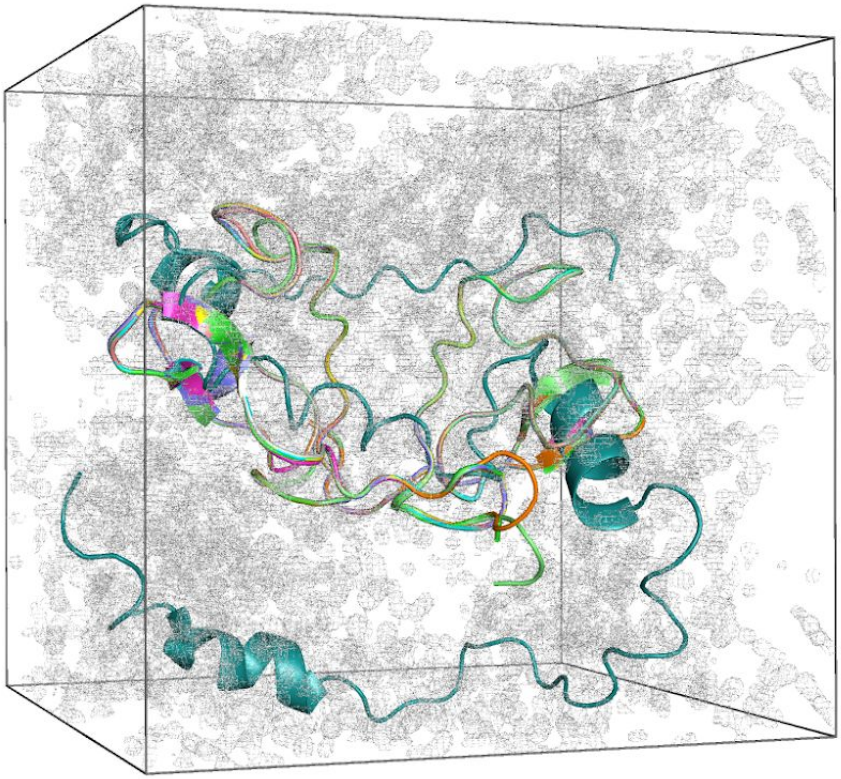


Figure 4 | HTC1 aggregation inhibition. HTC1 prevents amyloid formation of RCMLa (**a**) $A\beta_{1-40}$ **(b)** and α-synuclein (**c**), as measured by thioflavin T (ThT) fluorescence at 490 nm.

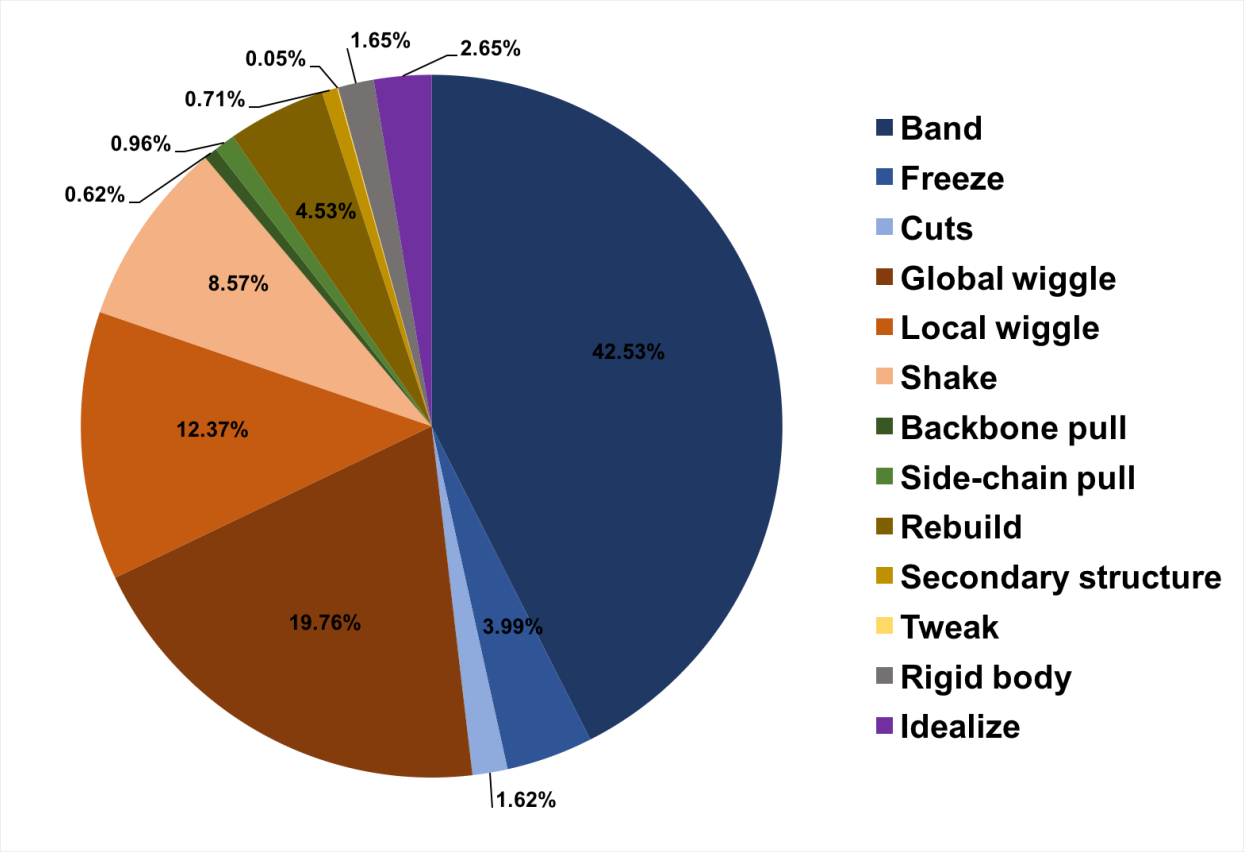
Supplementary Information



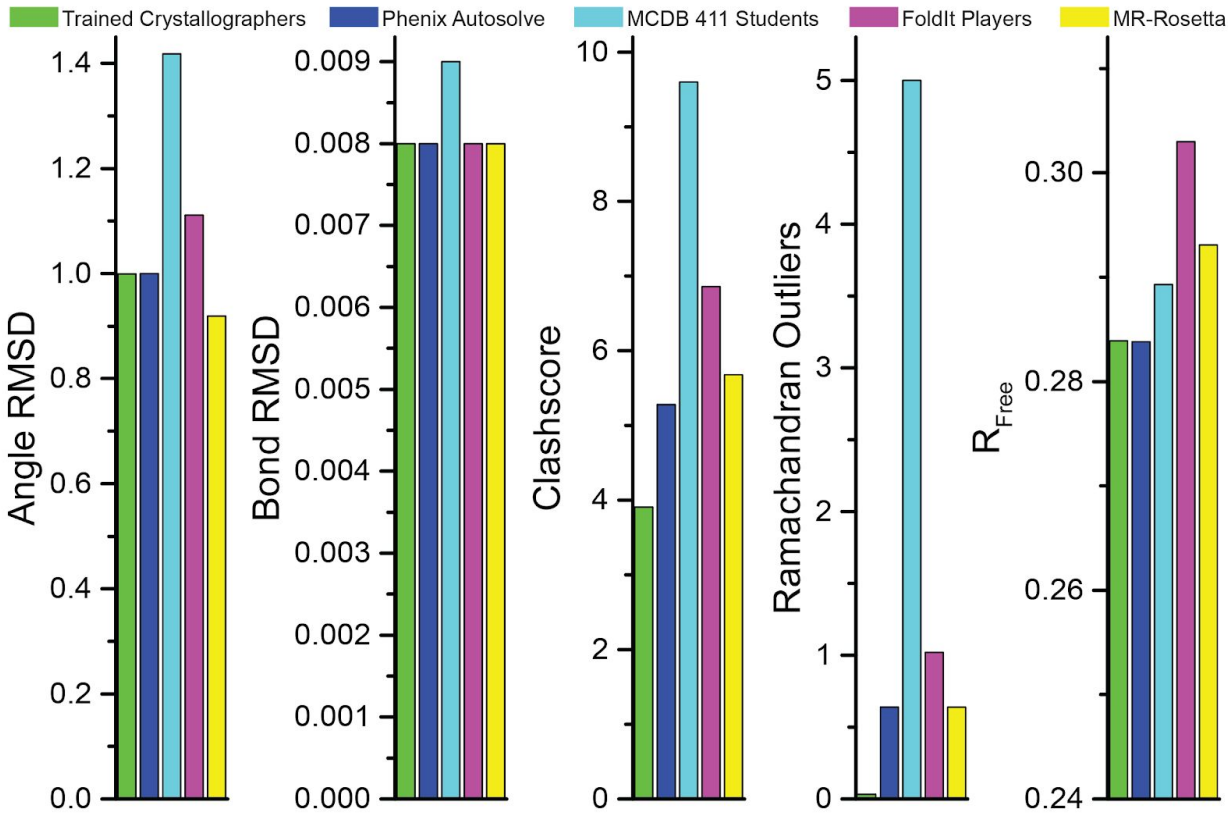
Supplementary Figure 1 | Electron density visualization in Foldit. Electron density rendered as solid (left) or wireframe (right), which can be selected from four different visualization options in the Electron Density menu (right portion of each panel). The Electron Density menu also contains controls for viewing adjustment, including notes to specific points in t



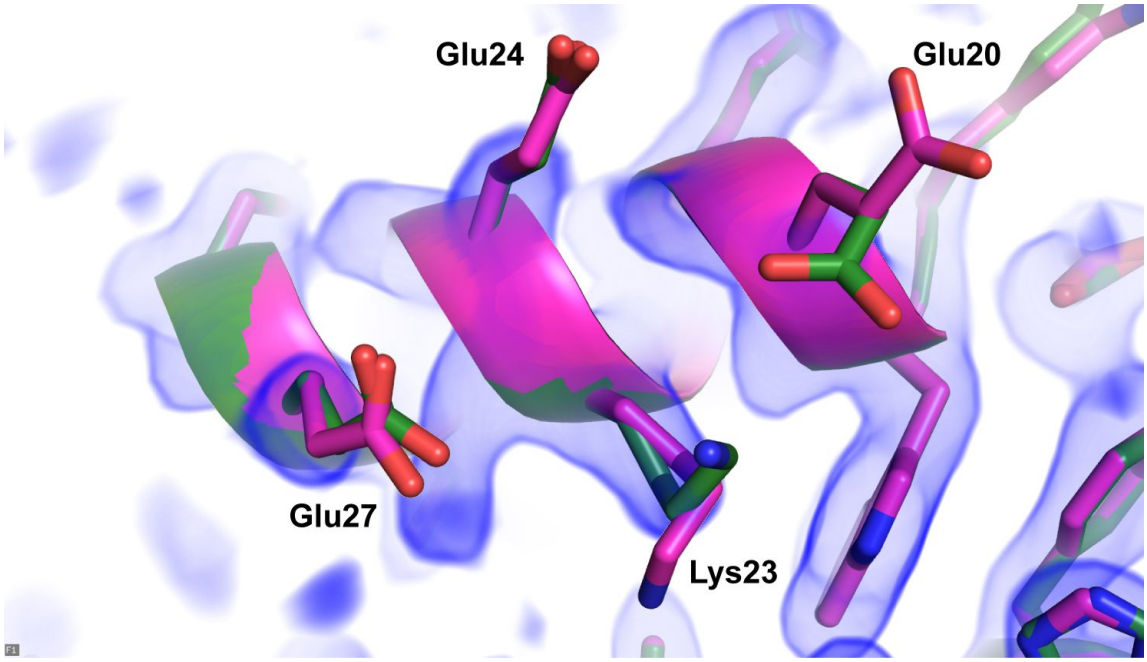
Supplementary Figure 2 | Lectin scytovirin top ten Foldit models overlaid in unit cell density. In a previous challenge, Foldit players were provided with the complete $P2_12_12_1$ unit cell of lectin scytovirin crystal density (PDB ID: 2QT4). The top nine Foldit models (ranked by Foldit score) were all correctly placed within the density of a single monomer in the unit cell. The tenth-ranked structure (deep teal) bridged density from several symmetric copies of protein in the unit cell. In the YPL067C puzzle, the density map was masked to include only a portion of the unit cell comprising a single monomer.



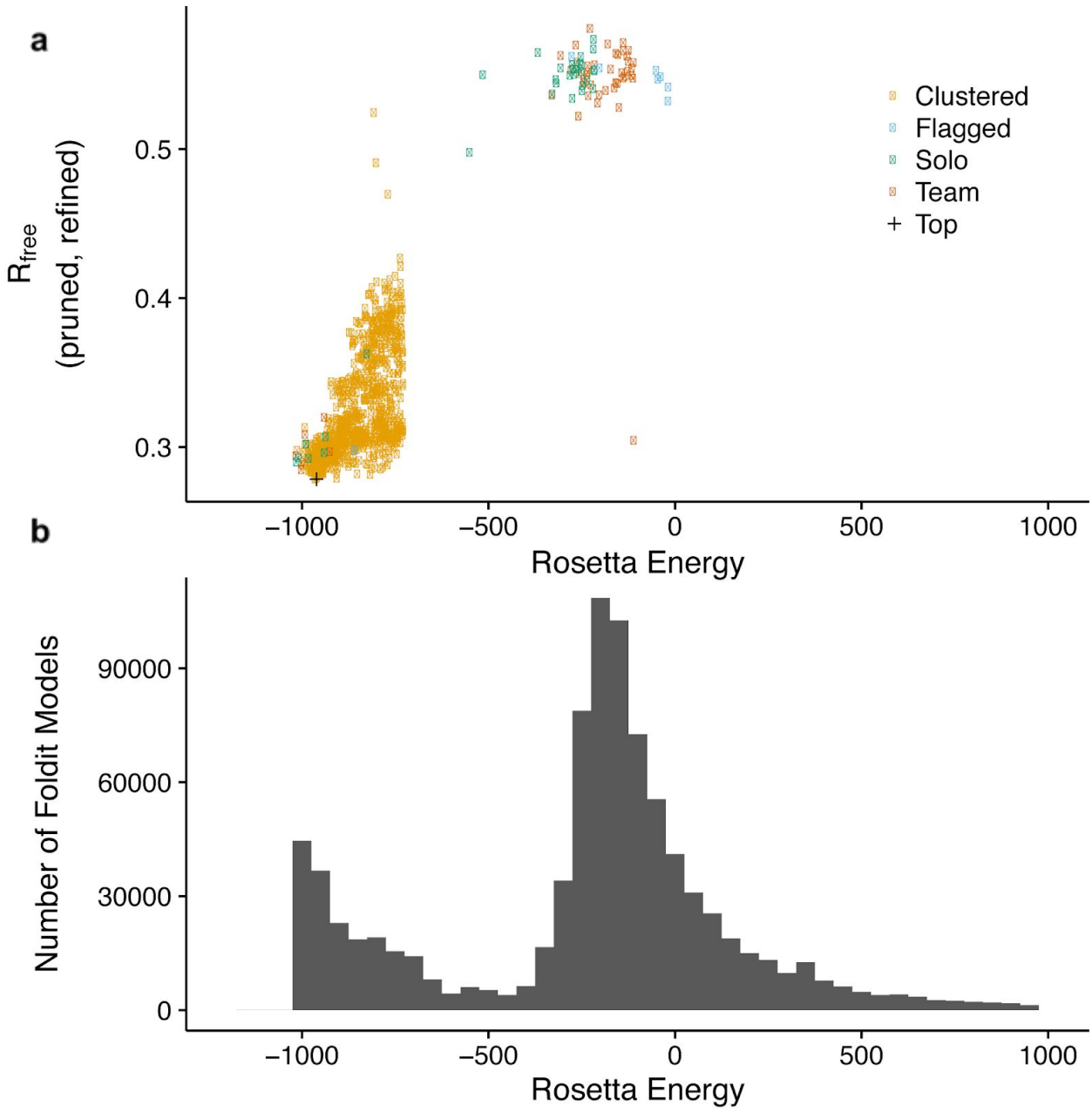
Supplementary Figure 3 | Total counts of move types used by Foldit players while creating the winning structure. Band, freeze, and cut actions (represented by shades of blue) are constraint movements. Global wiggle, local wiggle, shake, rigid body, and mutate auto actions (represented by shades of orange) are automated movements to the protein. Backbone pull and side-chain pull actions (represented by shades of green) are move types in which players make manual adjustments to the protein. Rebuild, tweak, secondary structure and idealize move types (represented by shades of yellow) are tools to modify secondary structures. For amore complete explanation of what the types of moves represent see Foldit documentation (http ref)[5].



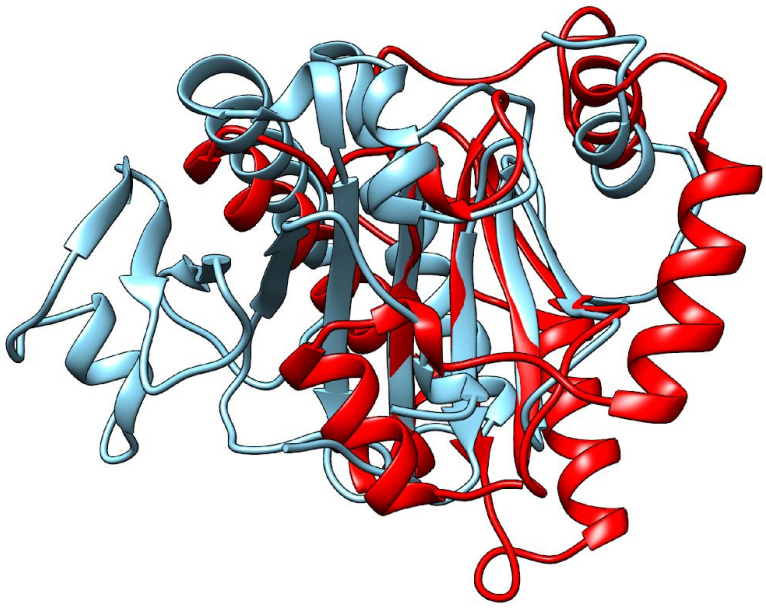
Supplementary Figure 4 | Comparison of key statistics of best models from each group before pruning disordered residues from Foldit structures. See Fig. 2 for key statistics after pruning disordered residues from the Foldit models.



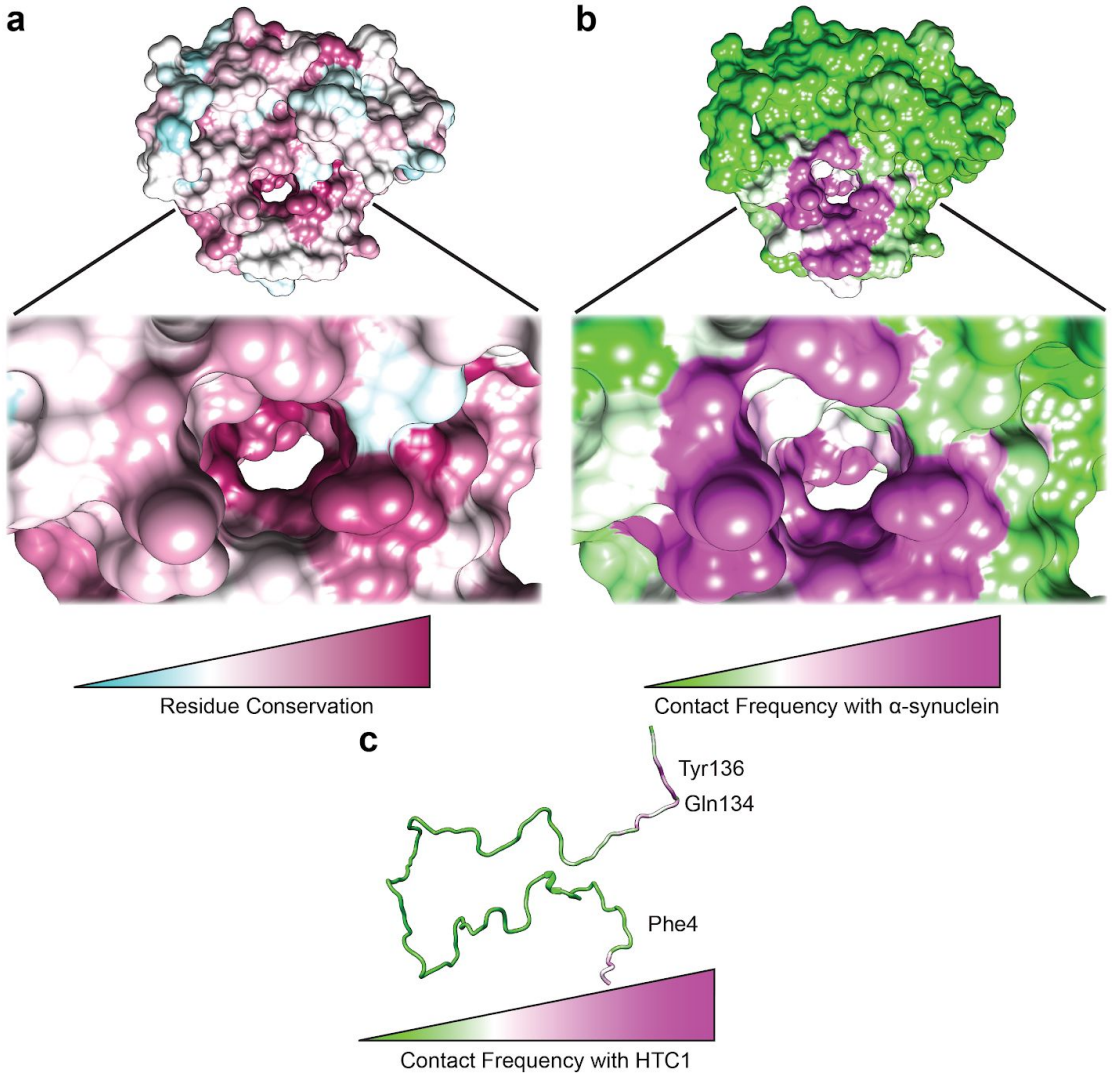
Supplementary Figure 5 | Example side chain conformational differences between the best Foldit model (green) and the model (magenta). Map used in competition rendered as blue volume contoured at 0.8 σ



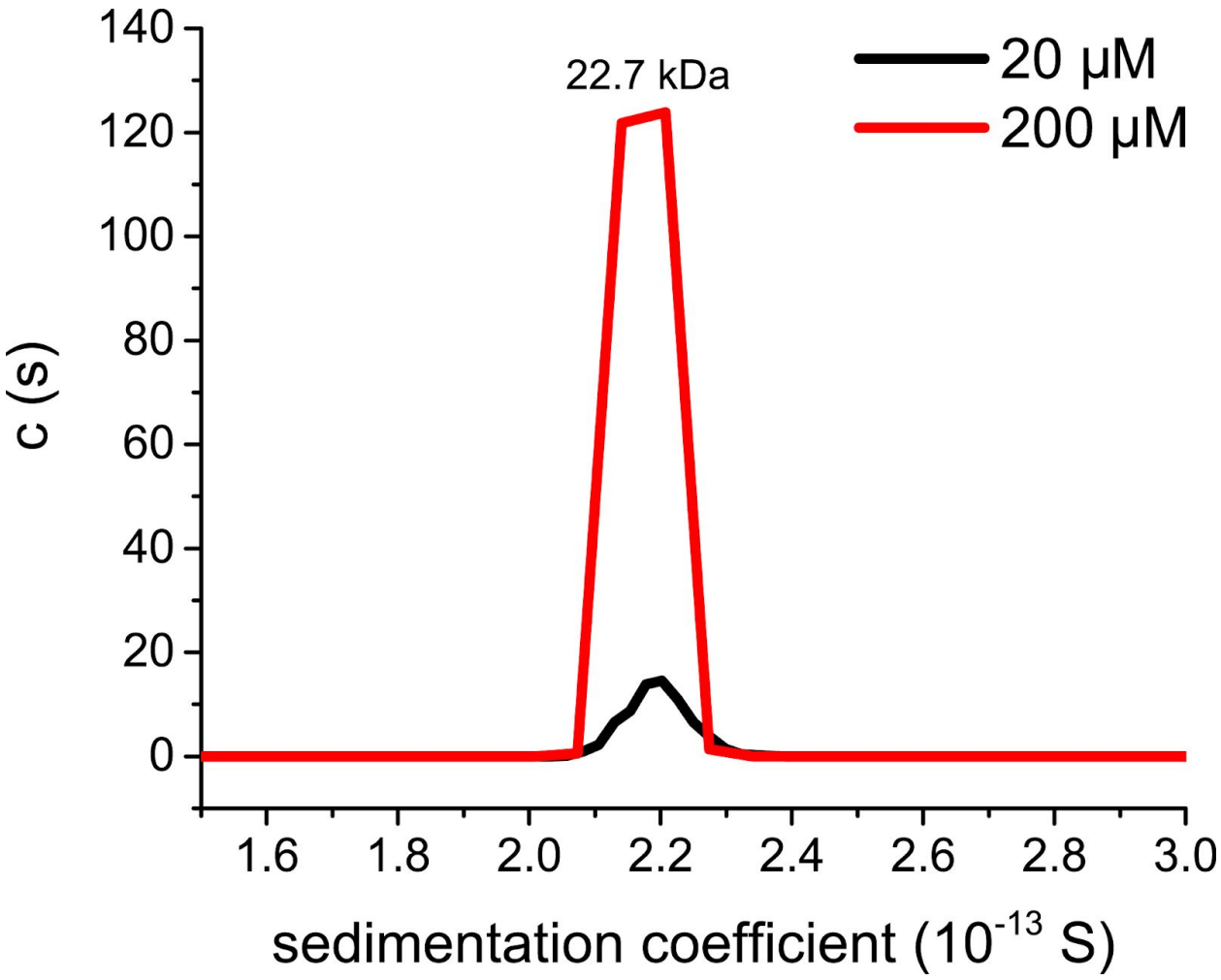
Supplementary Figure 6 | Foldit model score distribution and relation to R_{free}. (a) Of >900,000 Foldit models, Phenix refinement was carried out with the 1000 best-scoring cluster centers, along with the 50 best-scoring solo or team models and any models that were flagged by Foldit players for special consideration—1094 Foldit models in total. The most favorable Foldit scores (lowest Rosetta energy) were associated with low R_{free} values. (**b**) The score distribution of Foldit models is strongly bimodal, suggesting that Rosetta scoring can effectively discriminate models that correctly fit the electron density map. Experienced Foldit players performed best in this puzzle. Among all 472 participants of the Foldit puzzle, each player had previously played an average of 10 Foldit puzzles with electron density; among the seven soloist players that achieved a Rosetta energy less than -700, each player averaged 39 previous electron density puzzles.



Supplementary Figure 7 | Structural alignment of HTC1 (red) with top DALI match, PDB 4EGU (blue).



Supplementary Figure 8 | Docking simulations suggest that the highly conserved channel in HTC binds α-synuclein. Residue conservation (**a**) and projected contact map of docking simulations between HTC1 and α-synuclein onto surface of HTC1 (**b**). (**c**) Projected contact map of docking simulations between HTC1 and α-synuclein onto surface of α-synuclein. Contact frequency is ranked from dark green (lowest) to dark purple (highest)



Supplementary Figure 9 | Sedimentation velocity analytical ultracentrifugation of 20 (black) and 200 (red) µM HTC1. Calculated molecular weight based on sequence: 22.9 kDa.

Supplementary Table 1 | Crystallography statistics for HTC1

	SeMet HTC1 (top pruned Foldit)	Native HTC1
Data collection		
Wavelength (Å)	0.9876	0.97851
Space group	P43212	P43212
Cell dimensions		
a, *b*, *c* (Å)	63.3, 63.3, 117.8	62.5,62.5,117.6
α, *β*, *γ* (°)	90, 90, 90	90, 90, 90
Resolution (Å)	50-1.95 (1.98-1.95)	42.81-1.83 (1.89-1.83)
R_{merge} (%)	0.077 (0.940)	0.074 (0.683)
I/s*I*	59.5 (1.8)	12.7 (2.0)
Completeness (%)	99 (89)	100 (100)
Redundancy	12.4 (7.5)	7.8 (7.7)
Figure of merit	0.31	
CC1/2	0.998 (0.916)	0.998 (0.873)
Refinement		
Resolution (Å)	1.95	1.83
No. of reflections	18107	21161
R_{work}/R_{free}	0.26/0.28	0.20/0.25
No. of non-hydrogen atoms	1343	1661
Protein	1305	1518
Ligand/ion	0	12
Water	38	131
Average B-factors	53.8	43.6
Protein	53.8	43.6
Ligand/ion		52.2
Water	54.3	43.2
R.M.S deviations		
Bond lengths (Å)	0.008	0.009
Bond angles (°)	1.0	0.90
MolProbity percentile		
Clashscore	100th	97th
Overall score	100th	83rd
Ramachandran		
Favored (%)	98.71	97
Allowed (%)	1.29	2.7
Outliers (%)	0	0.5

Supplementary Table 2 | Top matches from DALI search of YPL067C

Chain	Z-score	RMSD	% Sequence identity	PDB description
4egu-A	4.9	3.6	11	HISTIDINE TRIAD (HIT) PROTEIN
4egu-B	4.8	3.6	10	HISTIDINE TRIAD (HIT) PROTEIN
5bv3-C	4.7	3.6	10	M7GPPPX DIPHOSPHATASE
4q61-B	4.6	3.2	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-F	4.6	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-D	4.6	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-D	4.6	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-I	4.6	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-I	4.6	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
1xqu-B	4.5	3.7	12	HIT FAMILY HYDROLASE
1xqu-A	4.5	3.8	12	HIT FAMILY HYDROLASE
4q61-F	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-B	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-H	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-G	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-C	4.5	3.2	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-A	4.5	3.2	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4zgl-J	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-C	4.5	3.2	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-A	4.5	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
5bv3-A	4.5	3.6	12	M7GPPPX DIPHOSPHATASE
4njz-C	4.4	4	17	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
3omf-A	4.4	3.5	14	PUTATIVE HISTIDINE TRIAD FAMILY PROTEIN
3oj7-A	4.4	3.5	15	PUTATIVE HISTIDINE TRIAD FAMILY PROTEIN
4q61-G	4.4	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-J	4.4	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4q61-H	4.4	3.3	16	UNCHARACTERIZED HIT-LIKE PROTEIN HP_0404
4njx-A	4.3	4.1	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
3oxk-A	4.3	3.7	13	PUTATIVE HISTIDINE TRIAD FAMILY PROTEIN
5bv3-B	4.3	3.6	12	M7GPPPX DIPHOSPHATASE
4njx-B	4.3	4	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
3n1s-E	4.3	4.3	17	HIT-LIKE PROTEIN HINT
3n1s-M	4.3	4.1	17	HIT-LIKE PROTEIN HINT
4ini-B	4.3	3.9	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
4njz-B	4.3	4.1	17	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
4inc-B	4.3	3.9	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
4njz-A	4.3	3.9	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2

4nk0-A	4.3	3.9	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
4njz-D	4.3	3.9	18	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 2
3r6f-A	4.2	3.6	14	HIT FAMILY PROTEIN
1xml-A	4.2	3.8	14	HEAT SHOCK-LIKE PROTEIN 1
3n1s-B	4.2	4.1	17	HIT-LIKE PROTEIN HINT
3n1s-F	4.2	4.1	17	HIT-LIKE PROTEIN HINT
3n1t-B	4.2	4	16	HIT-LIKE PROTEIN HINT
3n1t-E	4.2	4	16	HIT-LIKE PROTEIN HINT
3tw2-B	4.2	4	15	HISTIDINE TRIAD NUCLEOTIDE-BINDING PROTEIN 1
3n1t-A	4.2	3.9	16	HIT-LIKE PROTEIN HINT

Supplementary Information

Supplementary Introduction

We gave the Foldit players the same lectin scytovirin puzzle as the students from the previous iteration of the assignment to see if Foldit players could come up with similar solutions. Foldit players were provided with the same electron density map as the students had been and a polypeptide of 5 alanine residues in extended conformation. Players were allowed to add and mutate residues as needed, up to a length of 95 amino acids (the length of the native protein). No sequence information was given, and—unlike the class assignment—the position of the N-terminal amino acid was not provided. Although two weeks were allotted for the Foldit puzzle, players had reproduced the complete backbone of the native protein within 20 h of the puzzle's first posting. At many positions in top-ranked Foldit models, players showed a clear preference for non-native amino acids. Although they often diminished the fit-to-density, these substitutions were reinforced by other score terms in the Rosetta score function. For example, it was common for Foldit players to substitute a histidine at a solvent-exposed position actually occupied by Phe37, offsetting the drop in fit-to-density with a solvation bonus. At least two players produced accurate models with 100% sequence identity and they flagged them for special consideration, presumably having identified the native protein by sequence or structural alignment. Similar to the undergraduate structures, both of these Foldit models ranked in the 100th percentile in both Molprobity clashscore and total score.

Of note, in the lectin scytovirin puzzle, the students and Foldit players were given density covering the whole unit cell, including density from multiple monomer proteins. The students nearly uniformly built into a single monomer chain (as opposed to bridging adjacent monomers), as did the top nine Foldit structures (Supplementary Fig. 1). As such, for planning future assignments, we considered it unnecessary to provide the players and students with electron density outside of a single monomer. Given the relatively large size of the YPL067C Foldit puzzle compared with previous electron density puzzles, we trimmed the map around a single monomer to enhance software performance speed and facilitate player participation.

Supplementary Results

Lack of sequence similarity to YPL067C in PDB

When students were given an electron density map from a protein deposited in the PDB, a small proportion of the students used BLAST to find the structure. These students were then able to download the complete structure and use it as a guide in their model-building efforts. We had observed similar results with Foldit electron density puzzles that used published structures. To eliminate the potential for

this type of cheating in the competition, we avoided using a published electron density map or any structure that was similar enough to any structure in the PDB to be discovered by using BLAST. The Critical Assessment of Structure Prediction (CASP) protein competition, where modellers compete to predict structures, also focuses on unpublished crystal structures for similar reasons[31]. The lowest E-value score of YPL067C to any protein in the PDB, as of February 15th, 2016, was 1.9. As only E-values < 0.005 are generally considered significant, competition participants were unable to find a YPL067C homologue in the PDB to direct modeling efforts. Similarly, a wide-search molecular replacement attempt[32], which uses structures of every known domain as molecular search models, did not successfully find a structure solution for the experimental electron density. The lack of structural information on YPL067C prevented cheating by the various model-building competitors.

Comparison before pruning disordered regions

Based on the comparison criteria before pruning disordered regions from the Foldit structures, the best structure came from the trained crystallographer group. When evaluating the best structure from each group, the trained crystallographers produced the structure that scored either the best or was tied for the best in nearly every category. Most notably, the best trained-crystallographer structure contained zero Ramachandran outliers and the lowest Molprobity clashscore (3.9). Interestingly, Phenix Autosolve produced the second highest quality structure, with the second lowest number of steric clashes, and essentially equivalent R_{free} values and RMSDs for angles and bonds as the trained crystallographer structures.

HTC1 structural details

The N-terminal 15 residues of the HTC1 protein are disordered and could not be modeled in either the SeMet or native datasets. In the native dataset, residues Met35 to Arg50 appear to form a loop and right-handed α-helix that are marked by high B-factors, which are not visible in the SeMet dataset used in the model-building competition. A second disordered region discernable in the native dataset but not the SeMet dataset includes the loop between residues Trp66 and Ala78.

Despite its similarity to HIT protein structure, there are differences between the structure of HTC1 and canonical HIT proteins. The majority of the top DALI search results (Supplementary Table 2) dimerize through an antiparallel ß-sheet interaction (Supplementary Fig. 7). In contrast, HTC1 is monomeric in the asymmetric unit (Fig. 3) and in solution (Supplementary Fig. 9). Additionally, the sequence organization of a standard histidine triad is His-x-His-x-His, where x are hydrophobic residues. Instead of being distributed over a stretch of 5 amino acids, the histidine triad members in this newly recognized HTC family are distributed over a stretch 60 amino acids. It therefore appears that HTC1 is a somewhat atypical member of the HIT protein superfamily. Indeed, the protein family servers Pfam[33] and Interpro[34] currently list YPL067C as a member of a distinct family of unknown function (DUF3605), with over 900 members found in a wide variety of eukaryotes and viruses.

We examined the structure of HTC1 for clues as to the mechanism of its anti-aggregation activity. There are various conserved surface features of HTC1 that could possibly be involved in binding peptides, including a shallow channel running near the surface of the protein (Supplementary Fig. 8a). This shallow channel is ~4.3 Å in width at one end, narrowing as it passes through the protein. This channel is therefore large enough to potentially accommodate an unfolded protein terminus binding to HTC1. Blind docking of the α-synuclein NMR ensemble[27] with HTC1 suggests that this highly conserved channel is the most likely α-synuclein binding site (Supplementary Fig. 8b). Thus, it is possible that HTC1 prevents amyloid aggregation by binding protein termini in its conserved channel. This channel appears to be unique to the HTC family, as none of the top ten unique DALI hits in the PDB (Supplementary Table 2) feature a corresponding channel.

In the final model, a large tetrahedral molecule sits at the center of the channel, near His166 and His168 and the side chain of Gln154. Given the high concentration of sulfate ions in the crystallization

buffer, it is highly likely this density arises from a bound sulfate ion. Several of the top DALI search hits also were crystallized with bound sulfate ions. However, at this resolution, we cannot entirely rule out that this density could be due to a bound phosphate carried over from purification buffers. This ion occupies the only position in the channel that is well-segregated from bulk solution. The closest DALI search hit, 4EGU, was crystallized bound to guanosine monophosphate. Analogous electron density for a potential nucleotide bound in the correlating site of HTC1 is poorly defined, perhaps due to partial occupancy, and also could represent a glycerol molecule from the crystallization solution.