DOI 10.1002/prot.24538

WeFold: A Coopetition for Protein Structure Prediction

George A. Khoury[1], Adam Liwo[2], Firas Khatib[3‡], Hongyi Zhou[4], Gaurav Chopra[5,6], Jaume Bacardit[7], Leandro O. Bortot[8], Rodrigo A. Faccioli[9], Xin Deng[10], Yi He[11], Pawel Krupa[2,11], Jilong Li[10], Magdalena A. Mozolewska[2,11], Adam K. Sieradzan[2], James Smadbeck[1], Tomasz Wirecki[2,11], Seth Cooper[12], Jeff Flatten[12], Kefan Xu[12], David Baker[3], Jianlin Cheng[10], Alexandre C. B. Delbem[9], Christodoulos A. Floudas[1], Chen Keasar[13], Michael Levitt[5], Zoran Popović[12], Harold A. Scheraga[11], Jeffrey Skolnick[4], Silvia N. Crivelli*[14], and Foldit Players[15]

Affiliations

[1] Department of Chemical and Biological Engineering, Princeton University, USA
[2] Faculty of Chemistry, University of Gdansk, Poland
[3] Department of Biochemistry, University of Washington, USA
[‡]Present Address: Department of Computer and Information Science, University of Massachusetts Dartmouth, USA
[4] Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, USA
[5] Department of Structural Biology, School of Medicine, Stanford University, USA
[6] Diabetes Center, School of Medicine, University of California San Francisco (UCSF), USA
[7] School of Computing Science, Newcastle University, United Kingdom
[8] Laboratory of Biological Physics, Faculty of Pharmaceutical Sciences at Ribeirão Preto, University of São Paulo, Brazil
[9] Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
[10] Department of Computer Science, University of Missouri, USA
[11] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA
[12] Center for Game Science, Department of Computer Science & Engineering, University of Washington, USA
[13] Departments of Computer Science and Life Sciences, Ben Gurion University of the Negev, Israel
[14] Department of Computer Science, University of California, Davis, USA
[15] Worldwide

2/15/2014

*Author to whom all correspondence should be addressed
Silvia N. Crivelli, Ph.D.
Department of Computer Science
University of California, Davis
1 Shields Avenue
Davis, CA 95616
Emails: SNCrivelli@ucdavis.edu, SNCrivelli@lbl.gov
Phone: 925-367-5900

Running Title: WeFold Protein Structure Prediction Coopetition

Received: Aug 23, 2013; Revised: Jan 25, 2014; Accepted: Feb 08, 2014

Abstract: The protein structure prediction problem continues to elude scientists. Despite the introduction of many methods, only modest gains were made over the last decade for certain classes of prediction targets. To address this challenge, a social-media based worldwide collaborative effort, named WeFold, was undertaken by thirteen labs. During the collaboration, the labs were simultaneously competing with each other. Here, we present the first attempt at "coopetition" in scientific research applied to the protein structure prediction and refinement problems. The coopetition was possible by allowing the participating labs to contribute different components of their protein structure prediction pipelines and create new hybrid pipelines that they tested during CASP10. This manuscript describes both successes and areas needing improvement as identified throughout the first WeFold experiment and discusses the efforts that are underway to advance this initiative. A footprint of all contributions and structures are publicly accessible at http://www.wefold.org.

Keywords: Coopetition; protein structure prediction; CASP; structure refinement; Foldit;

Introduction

The complexity of current scientific research requires broad and open collaboration among researchers. Recently, the scope of these collaborations has expanded significantly to include individuals with no expertise in the specific field, known as citizen scientists. A noteworthy example of this approach uses computer games to engage participants. Launched in May 2008, Foldit is the first computer game designed to harness the natural human ability to recognize the 3D shape of proteins. More than 300,000 people have participated to date producing significant results[1]. These projects illustrate a shift in how scientists collaborate, as well as in the relationship between science and society.

A different sociological approach to tackle science is CASP[2] (Critical Assessment of techniques for protein Structure Prediction). Started in 1994 by Moult et al.[3], CASP is a community-wide, worldwide experiment to assess and advance the protein structure prediction field by helping identify where efforts should be directed[4]. CASP, which recently completed its 10th experiment[5], has challenged computational scientists to accurately and consistently predict protein structures using only the sequence of amino acids of soon to be or newly experimentally determined but unpublished structures. More recently, it has introduced other categories such as refinement[6-8] that challenges participants to improve the accuracy of a given protein model by submitting 5 new models.

During the CASP season, which occurs every other summer, each participating group applies a series of methods (some publicly shared, others secretly guarded) to the prediction pipeline and submits models for more than 100 different protein sequences or targets. After the experiment is over, the true experimental structures are published, the submitted models are examined by independent assessors, and the results are discussed in a subsequent meeting. Consecutive editions of CASP have shown substantial improvements in the category of "easy" proteins where

high sequence similarity to known proteins in the Protein Data Bank exists and such information is used to predict protein structures[4,9,10]. However, no single group has yet been able to consistently predict the structure of "hard" proteins with even moderate accuracy. Reviews on structure prediction in protein folding include those by Zhang[11], Dill and MacCallum[12], Khoury et al.[13], and Floudas[14]. CASP was not designed as a competition and participants are encouraged to focus on new ways of addressing the problem. However, although not intended, this ranking induces an atmosphere that is inherently competitive[15]. Because of the success of CASP, similar experiments were started[16-19] _ENREF_12_ENREF_13_ENREF_14. CASP remains the most-participated one among these to date with 95 manual prediction groups and 122 prediction servers submitting models in CASP10[2].

An approach that has not been tried in a scientific context until recently is "coopetition," which refers to cooperative competition. Coopetition is a common business practice. Companies sometimes engage their competitors in their product development process. Their goal is to create products of higher quality/extended functionality than the original products, resulting in an increased competitive advantage. The WeFold experiment is the first attempt at using coopetition, both open collaboration and competition among research scientists and citizen scientists, by generating methods that combine elements of the participating teams. WeFold took place during the CASP10[5] experiment with the goal of shaking up the field of protein structure prediction. It brought together thirteen labs worldwide (see Supplementary Table S1), ranging from purely bioinformatics to physics-based approaches that, for the first time, collaborated and competed in search for methodologies that are better than their individual parts. The size of the collaboration was unprecedented in the history of CASP, with participants contributing a superset of almost 8.8 million structures to WeFold from which a small fraction were submitted

to CASP10[5]. This paper describes the WeFold experiment. It analyzes the performance of the combined methods with respect to their base methods in the context of blind structure predictions during CASP10[5], describes the challenges faced, and reports those which still remain. The lessons learned from this experiment could be useful to other coopetition efforts that may be attempted in the context of other CASP-like competitions.

Materials and Methods

A unique aspect of the WeFold experiment is that the mechanism for the collaboration was largely unknown until the CASP10[5] experiment started. Therefore, on the first day of CASP10[5], WeFold participants logged into the WeFold gateway to discuss how to best combine the different components they were contributing to the project. Five branches resulted from that discussion. Their names and group numbers in CASP10[5] are: wfFUIK (149), wfFUGT (260), wfCPUNK (287), WeFold Branch (101), and WeFoldMix (441). The first three branches were named based on the first letter of their component methods. wfCPUNK and WeFold Branch were applied to the prediction of human tertiary structure prediction targets whereas wfFUIK, wfFUGT, and WeFoldMix applied to both human tertiary structure prediction and refinement targets. Neither branch competed as a server. There were 46 human targets (53 total, of which 7 were cancelled) and 27 refinement targets (28 total, of which 1 was cancelled). Figure 1 illustrates the organization of the different branches. Please refer to Supplementary Material for a description of the science gateway.

The wfFUIK branch. This branch starts with a set of structures produced by Foldit[20] players and then applies a selection process based on state-of-the-art computational methods. It was applied to the prediction of 15 human and 23 refinement CASP10[5] targets, which are all of the targets attempted by Foldit. Team members from the contributing labs adapted their methods to

work within the context of this branch. For example, some methods that were originally designed to operate on smaller datasets had to be modified to work on the large sets of Foldit-generated structures. Other methods had to be adapted to handle systems containing structural symmetry as these systems were attempted by Foldit.

Supplementary Figure S1 represents the combined methodology. (A) Foldit[20] players generate an ensemble of protein models on the order of 10^5 models per target. (B) Structural filtering is performed to eliminate very similar structures (RMSD ≤ cutoff), those with unrealistic solvent accessible surface areas[21](SASA), and those lacking secondary structure elements. This yields an enriched set consisting of 10^3-10^4 structures, called the Unique/Filtered set. (C) The iterative traveling salesman based clustering algorithm, ICON[22] is used to select less than 100 models representing the entire conformational space, and the lowest energy structures based on the Rosetta[23] and dDFIRE[24] energy functions are added to that set. (D) These models are refined using a knowledge-based potential followed by stereo-chemical correction implemented in the KoBaMIN[25-27] server. (E) Finally, GOAP[28], Rosetta[23], dDFIRE[24], and APOLLO,[29] are used to rank the models, leading to a consensus.

The wfFUGT branch. Like wfFUIK, this branch starts with a set of models produced by Foldit then deviates from wfFUIK from step (C) on. It was applied to the prediction of 13 out of the 15 human targets attempted by Foldit and 17 out of the 21 refinement targets attempted by Foldit in CASP10[5]. The replica exchange Monte Carlo simulations that are part of this branch's pipeline (described below) were computationally expensive and for some targets the generation of Foldit models, followed by the filtering step did not allow for enough time for this pipeline to complete. The wfFUGT branch tests combining sampling by Foldit players with filtering algorithms, model selection by the knowledge-based potential GOAP, and the TASSER[30,31] refinement protocol.

Starting from the Unique/Filtered structures, GOAP[28] selects the top 30 models from the enriched set. TASSER[32] next refines the selected models. TASSER is primarily developed for refining template models built upon PDB structures found by threading methods. Here, it is applied to Foldit-generated structures. First, it extracts distance and contact restraints based on consensus conformations of the 30 selected structures. Then, it starts from the 30 structures and moves them to satisfy the distance and contact restraints using replica exchange Monte Carlo simulation[33] in a Cα representation. Low energy trajectories are outputted at fixed step intervals. At the end of the simulation, these trajectories are clustered using SPICKER[34]. Submitted models are the top cluster centroids with rebuilt main-chain and side-chain atoms.

The wfCPUNK branch. This branch is an *ab initio*/free modeling branch that combines secondary structure, beta sheet topology, and contact predictions with the sampling capabilities of coarse-grained replica-exchange molecular dynamics, when templates are unavailable. It was applied to the prediction of 21 small to moderately-sized targets due to the extreme computational cost involved. Of those 21 targets, only 4 belonged in the free modeling category. First, coarse-grained simulations with the UNRES force field[35-39] (www.unres.pl) are employed to carry out Multiplexed Replica Exchange Molecular Dynamics (MREMD)[40]. Dihedral-angle and distance restraints are imposed on the virtual-bond dihedral angles between the consecutive α-carbon (Cα) atoms and virtual side-chain distances. The restraints are obtained using a consensus-based method, CONCORD[41] for secondary-structure prediction, a novel optimization-based approach, BeST[42], for beta-sheet topology prediction, and a physics-based method of inter-residue contact prediction[43,44].

For each protein, 64 MREMD trajectories are run at 32 different temperatures (2 trajectories per temperature). The last 12,800 snapshots (200 snapshots per trajectory), where each snapshot is

saved every 20,000 conformations, are taken for further analysis, which is carried out by using the weighted-histogram-analysis method (WHAM)[45]. This method is used to calculate the relative probability of each conformation from the last portion of the MREMD[38] simulation and to calculate the heat-capacity curve and other thermodynamic and ensemble-averaged properties. Then, the conformations are clustered at the selected temperature, which is equal to T_m-10K, where T_m is the position of the major heat-capacity peak. Five clusters with lowest free energies are chosen as prediction candidates. The conformations closest to the respective average structures corresponding to the found clusters are converted to all-atom structures[46,47] and their energy is minimized using the KoBaMIN server[27].

The WeFold branch. This branch was applied to the prediction of 43 human CASP10[5] targets. It starts with all models from all CASP servers and WeFold methods and assesses them using the APOLLO model quality assessment prediction method. APOLLO[29] first filters out illegal characters and chain-break characters in the models predicted for a target. Next, it performs a full pairwise comparison between these models by calculating GDT_TS scores between a model and all other models using the TM-Score[48] program. The mean pairwise GDT_TS between a model and all other models is used as the predicted GDT_TS of the model. Subsequently, TASSER[32] is employed to refine the top 30 selected models. First, TASSER extracts distance and contact restraints based on consensus conformations of the 30 selected structures. Then, it starts from the 30 structures and moves them to satisfy the distance and contact restraints using replica-exchange Monte Carlo simulations[33] in a Cα representation. Low energy trajectories are output at fixed step intervals. At the end of simulation, these trajectories are clustered using SPICKER[34]. Models selected for submission were the top cluster centroids with rebuilt main-chain and side-chain atoms.

The WeFoldMix branch. This branch was created by a new group that did not participate in CASP10[5] by itself, and was applied only to the prediction of 5 human and 1 refinement CASP10[5] targets due to the extreme cost of performing replica-exchange molecular dynamics simulations and parallelization inefficiencies due to low atom/processor ratios when using implicit solvent. It starts with a small set of high-quality models collaboratively generated and ranked. Each model is energy minimized using the steepest descent method[49]. Initially, no constraints are applied to the protein; in the second step all covalent bonds are constrained with the LINCS algorithm[50]. Simulations are performed using GROMACS 4.5.5[49] with the AMBER99SB-ILDN[51] forcefield and the GBSA[52] implicit solvent model. Replica-exchange molecular dynamics (REMD) is employed to overcome the conformational trapping of the structures in local potential energy minima by diffusion in temperature space. A total of 8 simultaneous simulations (replicas), are performed in the temperature range of 298-473K and are allowed to exchange each 5 ps according to the Metropolis criterion[53]. The observed average exchange probability was 0.2.

After 1-3 ns of REMD, the 298K-trajectory portion reaches convergence and is used for cluster analysis using a single linkage algorithm. Each cluster centroid is submitted to the previously described two-step energy minimization process and each minimized cluster centroid is ranked based on several structural and energetic metrics. These metrics include potential energy, number of intra-protein hydrogen bonds, and SASA. The structures with the best consensus metrics are submitted.

Selection strategy employed by the four Foldit-based teams during CASP10. Here we describe the selection process used by the FOLDIT team, as well as the three teams associated

with it. This serves to explain the different performance of the wfFUIK and wfFUGT teams compared to FOLDIT.

Quality and ranking of Foldit models by the FOLDIT team is determined by the Rosetta full-atom energy[23]. For each CASP target, the lowest Rosetta energy Foldit prediction for each individual Foldit player is kept, in an attempt to select a conformationally diverse set of FOLDIT submissions out of the top-ranked Foldit predictions. Since Foldit allows players to form teams for cooperative gameplay—and share solutions with teammates—the top-ranked predictions were often very similar to one another for players on the same team. This was generally not the case when comparing the top prediction across different teams (or players who are not part of any team), therefore the selection strategy during CASP10[5] for the FOLDIT team was to examine the lowest Rosetta energy Foldit prediction generated by each individual team (players without a team were considered their own team). The five CASP submissions for the FOLDIT team were selected by manually inspecting these representative solutions from each team, and selecting a conformationally diverse set of predictions by visual inspection. This was the same selection strategy used for FOLDIT submissions during CASP9[54].

Before the start of CASP10[5], three Foldit-based teams (Anthropic Dreams, Contenders, and Void Crushers) requested the ability to select and submit their own CASP submissions from a pool of their own team's solutions. Each of these three teams was provided with two top-ranked predictions for each of the players on their Foldit team: the lowest Rosetta energy solution each player generated on their own, and the lowest energy solution that player worked on by sharing with the rest of their team. As Foldit does not allow different teams to share solutions with one another, these three CASP10[5] teams were completely independent from one another, and also independent of the submissions by FOLDIT.

Metrics Used in Analysis. The global distance test total score (GDT_TS[55]) is approximately the percentage of residues that are located in the correct position[12]. It has become a standard evaluation measure in CASP[56] for determining the accuracy of a structure, preferred over the common root-mean squared deviation (RMSD) metric. GDT_HA[57] is a finer metric, which uses tighter Cα distance cutoffs.

GDT_TS and GDT_HA are calculated using the TM-Score[48] program. Both of these metrics can be presented on a zero to one basis or alternatively as a percentage. The higher the value is, the more similar the prediction is to the true structure. GDT_TS is used throughout this paper except for when we refer to the refinement assessors as they used GDT_HA.

Results

The thirteen labs participating in the WeFold initiative were arranged into five branches, each representing five independent protein structure prediction methods that combine different components from their contributing group. Three of the branches produced one remarkable result each and two of these results were featured by the assessors in the refinement[6] and free modeling[58] categories. However, none of them produced consistently good results. In this section, we discuss the strengths and potential of these branches, as well as their weaknesses. We also discuss the strengths and weaknesses of the WeFold experiment as a whole. Our assessments are based on the CASP official results and assessments available at http://predictioncenter.org/CASP10.

What Went Right in the Collaborative Protein Folding Pipelines

The wfCPUNK Branch. The wfCPUNK branch aimed to address free modeling targets. These targets are among the hardest with poor to no sequence identity in the "twilight zone"[59] and thus lack a determinable structural template. Of the four free modeling targets attempted by

wfCPUNK, it achieved its best performance for target T0740_D1, yielding a high-scoring model according to the cumulative plot of α-carbon accuracy as shown in Figure 2. This result is attributed to the *ab initio* contacts predicted as part of the pipeline, which were used in wfCPUNK and not used by the other methods undergoing UNRES sampling (i.e., Cornell-Gdansk and KIAS-Gdansk). In fact, the helix-helix contact predictions[43,44] that were contributed by the FLOUDAS group and used as restraints in the UNRES simulations made a difference in the sampled space. Figure 3 shows the predicted contacts superposed on the experimental structure, the best model (Model 4) from the wfCPUNK group, and the best model from the Cornell-Gdansk group (Model 3). It can be seen that the restraints made the C-terminal α-helix bent and packed against one of the middle α-helices. Unlike the experimental structure, this α-helix is straight in the Cornell-Gdansk Model 3.

The most accurate prediction for this target according to GDT_TS was 38.87 and it was produced by RaptorX-Roll. The GDT_TS of the wfCPUNK prediction was 32.10. However, from Figure 2 it can be seen that the percent of residues within a distance cutoff line for the wfCPUNK Model 4 clearly extends to the right beyond that for any other model, albeit this happens only after the 5Å distance threshold. This feature arises from the middle resolution of the UNRES force field, which reproduces well the overall topology of protein folds and supersecondary structure/domain packing but does not reproduce finer details of protein folds. It should be noted that the same feature of GDT_TS plots was observed for UNRES-predicted structures of T0668 and T0684-D2[60]. For another target, T0663, the UNRES prediction was not among those top ones as far as the GDT_TS plots were concerned; however, UNRES was one of the only two approaches that predicted the correct topology of domain packing and this prediction was, therefore, featured by the CASP assessors.

Table I presents the GDT_TS values for the predictions by wfCPUNK and its component methods, as well as other groups using the component methods. wfCPUNK was able to outperform both the Cornell-Gdansk and KIAS-Gdansk teams in three of four targets, and FLOUDAS in two out of four targets. These results, although not statistically significant, highlight the potential benefits of combining methods for the prediction of free modeling targets.

The wfFUIK Branch Applied to Refinement Predictions. According to the refinement category assessors[6], the large majority of the fifty groups that competed in CASP10[5] failed to improve the quality of the starting models and even the successful groups were able to make only modest improvements. Only very few methods could consistently refine the targets. Noteworthy examples are FEIG[20,61] _ENREF_1 (positive ΔGDT_HA for 24 targets), Seok[62] (positive ΔGDT_HA for 16 targets) and KnowMIN (positive ΔGDT_HA for 15 targets)[6]. As the assessors pointed out[6], wfFUIK improved GDT-HA significantly less frequently than the FEIG and Seok groups (i.e. wfFUIK improved GDT-HA for 5 targets), but its models improved GDT_HA by the largest amount, with the same being true for the MolProbity (MP) scores[63] (MP is the MolProbity score that combines the log-scaled counts of all-atom steric clashes, atypical rotamer conformations and unfavorable backbone torsion angles in each prediction). Also, FEIG, Mufold, and wfFUIK are the top groups at side-chain positioning. These conclusions reached by the assessors are reflected in Table II, which uses GDT_TS (the comparative measurement used throughout this paper).

To illustrate the potential additive benefits of the wfFUIK pipeline, Figure 4 shows a walkthrough of the contributions of each step of the wfFUIK method to the refinement of target TR722. First, (A) the starting structure was given to Foldit players. There were two Foldit runs; one run treating the structure as a monomer and another run treating it as a symmetric dimer. In

total, the players produced 256,776 structures. A filtering step was performed on both the monomeric and symmetric set, leaving a Unique/Filtered set of 20,488 monomers and 30,855 symmetric structures, which comprised 20% of the total number of structures generated. They are represented by grey dots in each plot in Figure 4. The starting structure is shown as a red color "X" and had a GDT_TS of 58.0. The structures that would be selected by naïvely taking the one with lowest Rosetta energy for both the monomer and symmetric dimer are shown as pink squares. (B) From the Unique/Filtered set, those structures that are the cluster medoids selected by ICON[22], as well as the lowest energy structures from Rosetta[23] and dDFIRE[24] are highlighted in blue. Several structures from this population are already more accurate than the structures generated in the previous step. Next, (C) the structures that resulted from step B were further refined using KoBaMIN[27] and are shown as cyan dots. After this step, a higher fraction of structures has improved GDT_TS's relative to the step B structures. The cyan dots, generally located down and to the right relative to the dark blue pre-KoBaMIN population, indicate that KoBaMIN structures are refined with lower energies and improved GDT_TS's. (D) The structures submitted in blind prediction are shown in pink stars. Those structures were selected according to a number of energy and quality-assessment metrics and to be diverse from each other.

Figure 5A shows the best model out of the five submitted, which garnered a GDT_TS of 65.95 (the top Model 1 prediction for this target came from one contributing group, FLOUDAS, yielding a GDT_TS of 63.19). The best wfFUIK model achieved the "peak-performance" in terms of ΔGDT_TS for TR722; that is, it was the #1 most refined structure according to ΔGDT_TS considering all the models submitted for this target by all groups. It was the #3 most refined structure according to ΔGDT_TS among all refinement targets considering all the models

submitted by all groups (see Table II) and it was featured by the CASP10[5] assessors in the refinement category for being one of three models where large increases in GDT_HA were observed[6].

The average GDT_TS of blind predictions for this target from all CASP10[5] participants was 52.9±7.5, indicating that wfFUIK's best blind prediction outperformed the rest of the predictions by more than one standard deviation. This model is ranked 93 out of 51,343 models contained in the Unique/Filtered set, which is comparable to the top 2% of structures produced by the Foldit players. There are even better structures in the Unique/Filtered set that were not chosen (please refer to section What Went Wrong in the Collaborative Pipelines for a detailed analysis). Significantly, the best structure contained in the filtered set had a GDT_TS of 71.26, which if selected would have outperformed the average prediction by over 2 standard deviations. The strategy employed by Foldit to use both monomeric and symmetric dimer prediction runs increased the chance of a successful prediction. This example shows that coupling the human players' abilities to refine the proteins with the subsequent clustering, refinement, and scoring methods in the wfFUIK protocol can make it possible to successfully select models among the very best from the remarkably large population of structures produced by Foldit. More importantly, TR722 is not the only target for which wfFUIK produced models that were more accurate than the starting one as shown in Table II.

However, the wfFUIK branch did not achieve consistently good results. In the section "What Went Wrong in the Collaborative Pipelines," we investigate the step-by-step results of wfFUIK applied to other CASP10[5] targets and show that although the structural accuracy remains, it was the last step that consistently failed to select the best models that had been produced by the previous steps.

The wfFUGT Branch Applied to Refinement Predictions. This branch, which is also based on Foldit, did not do as well as wfFUIK. Nevertheless, it produced a noteworthy model for refinement target TR705. In fact, the wfFUGT branch improved the starting GDT_TS of TR705 from 64.84 to 70.05[2]. This blind prediction, which was the best submitted for this target considering only Model 1 and ranked 5th when considering all models, is shown in Figure 5B (green), along with the native (black) and starting (red) structure. Next, we compare the performances of wfFUGT and wfFUIK to their base method Foldit.

Overall Comparison of wfFUIK and wfFUGT to Base Method FOLDIT for Refinement Targets. There were six independent teams that began with structures produced by Foldit players: FOLDIT, Anthropic Dreams, Contenders, Void Crushers, wfFUIK and wfFUGT. Anthropic Dreams, Contenders, and Void Crushers were created and run by the Foldit players themselves (see Methods). Table II shows the top 15 best per-target improvements in GDT_TS during CASP10[5] considering the 5 models predicted by all methods. This table shows that Anthropic Dreams, Void Crushers, and the WeFold wfFUIK branch submitted the three best per-target improvements in GDT_TS over the starting models for the refinement category at CASP10[5]. Also noteworthy is that wfFUIK is the FOLDIT-based group with the largest number of positive ΔGDT_TS.

Figure 6 shows a head-to-head comparison of the refinement models submitted to CASP10[5] using the wfFUIK and wfFUGT methods to those submitted by the base method FOLDIT considering the best of 5 models. This figure is based on the data shown in Table III which provides a comparison of the best refinement structures submitted in CASP10[5] from Foldit-derived branches versus all predictions submitted by all groups. Specifically, the Foldit-derived models are compared against the mean of the GDT_TS values for all models submitted for each

target. This table also provides the standard deviation for each target. wfFUIK submitted more accurate predictions than FOLDIT by GDT_TS in a large majority (74%) of all of the refinement targets attempted by both teams. On the other hand, the wfFUGT method outperformed FOLDIT in 53% of those cases. This indicates that wfFUIK is a better refinement strategy than wfFUGT, and its improved performance is due to the multi-step selection process used in the method. Furthermore, we performed a one-sided t-test comparing the best predictions by FOLDIT to those by wfFUIK, and wfFUGT. The P-value between FOLDIT and wfFUIK is 0.031, indicating a statistically significant improvement. Conversely, the P-value between FOLDIT and wfFUGT is 0.317.

Although the wfFUIK branch amplified the refinement relative to the base method FOLDIT, many of the submitted models did not refine the structures relative to the start. Thus, although it achieved ΔGDT_HA > 0 in 5 of the 23 targets attempted[6] and placed among the top 10 ranked groups where Model 1 ΔGDT_HA was positive[6], overall it ranked below the naïve method of doing nothing to the input structure according to the score used by the CASP10[5] refinement assessors, which includes deviations of GDT_HA, RMSD, and other metrics[6]. This result may not be surprising as the method components have not been optimized to maximize performance given the output from each stage of the prediction pipeline. Optimization of the stages in each WeFold branch to maximize performance relative to the input from the previous stage(s) may lead to improved performance of the branches in the future. Nevertheless, these results show that the wfFUIK pipeline did consistently outperform its base method FOLDIT even without prior optimization.

The WeFold Branch. Figure 7 shows the performance of the WeFold branch in absolute comparison to all Model predictions, and in relative comparison to the top Model predictions by

all groups and all methods. The absolute performance is assessed based on the Z-score of the GDT_TS of the best model submitted by WeFold relative to all predictions by all groups and methods. We calculated the Z-score of each target as $Z-score = \frac{(\text{BestWeFoldPrediction}_{GDT\_TS} - \mu_{GDT\_TS})}{\sigma_{GDT\_TS}}$, where μ_{GDT_TS} and σ_{GDT_TS} denote the mean and standard deviations of the GDT_TS values for all Models submitted to CASP10[5] for that particular target. The relative comparison is based on the ratio between the GDT_TS score of the best WeFold prediction and the best GDT_TS achieved by all groups for each target, calculated as $\%Best = \frac{\text{BestWeFoldPrediction}_{GDT\_TS}}{\text{BestCASPModel}_{GDT\_TS}} x100$. The WeFold branch performed comparably (11 targets) or better (12 targets) than TASSER (one of its base methods) in 53% of the attempted targets as shown in Table IV. This table provides a comparison of best structures submitted in CASP10[5] for tertiary structure prediction by WeFold branches and their component methods to the best and average predictions submitted by all groups. The best prediction for each target among these methods is bolded. In some cases such as T0676, T0700, T0735, and T0744, the WeFold branch did substantially better than TASSER in terms of both GDT_TS and overall ranking (Table IV). The reason for the difference between WeFold and TASSER (human group) is due to the difference between the model selection methods, i.e., APOLLO (used by WeFold) and GOAP (used by TASSER). When consensus information was useful, APOLLO performed better than GOAP and, consequently, WeFold performed better than TASSER. Overall though, TASSER significantly outperformed the WeFold branch "winning" 17 targets in the cross-comparison, with a one-sided P-value of 0.032 (Table IV). This result indicates that WeFold branch has substantial room for improvement.

What Went Wrong in the Collaborative Pipelines

In this section we discuss the main reasons why the collaborative effort did not do as well as expected given the combination of methods. In some cases such as wfCPUNK and WeFoldMix, the branches did not attempt enough targets to make any statistically significant conclusion. In other cases such as wfFUIK and wfFUGT we present a detailed analysis that shows the reason why the collaborative branch failed to produce more positive results.

Problems Identified in the wfCPUNK Branch Applied to Human Free Modeling Targets. This branch submitted only 4 free-modeling targets. This low number of submissions is due to both the uncertainty with which targets are deemed as "Free Modeling*" prior* to their prediction and the high computational cost of performing MREMD calculations which is aggravated by the limited computational resources available. Going forward, the branch plans to explore the use of a consensus from different contact prediction methods rather than the results from a single method to increase its chances for success.

Problems Identified in the wfFUIK and wfFUGT Branches Applied to Human Tertiary Structure Prediction Targets. We observed a stark difference between the performance of the WeFold and FOLDIT methods for tertiary structure prediction targets and refinement targets, and these differences can be explained by discrepancies between the input data. In CASP10[5], Foldit had two runs, called puzzles, for tertiary structure prediction targets. In the first run (Run 1) players are only given the sequences of the target and several alignments to proteins determined by fold recognition methods. In this run, the players can modify the alignments and must determine how to connect the regions that are not well aligned, and have full responsibility for the folding pathway of the proteins. These were the Foldit predictions that were shared with the various WeFold branches. In the second run (Run 2), players are given a number of starting models generated by servers that have performed well in previous CASP experiments. Thus, in

Run 2, folding was somewhat akin to a refinement problem. Often, these initially provided structures are well predicted and are trapped in deep local minima, so subsequent refinement was unable to substantially change the initial structures. It was observed that the structures generated by Run 2 more often yielded lower energy (higher in-game scores) than the structures generated by Run 1, and thus they were usually the final models submitted by the FOLDIT team. Because the server predictions used for Run 2 were not publicly released until six days after the server deadline for each CASP10[5] target, there was not enough time to send these Run 2 Foldit predictions through the WeFold pipelines (often these Run 2 Foldit puzzles would close the day before the CASP10[5] target deadline). As a result, it is unfortunately not possible to draw any fair meaningful comparisons between wfFUIK, wfFUGT, and FOLDIT in tertiary structure prediction. Reflecting on the design of the experiment, this is one area that should be improved upon for the next CASP so that the methods could be directly compared.

Problems Identified in the wfFUIK Branch Applied to Refinement Predictions. Although the wfFUIK method net improved upon its base method FOLDIT, like most refinement methods, wfFUIK suffered from two problems: (1) degrading of the starting model and (2) final model selection. This is not surprising given that the starting structures which are already accurate predictions have been driven into deep local minima. In order to analyze the effects of each component method in the pipeline and show where it failed, we performed a step-by-step analysis of the data starting from the Unique/Filtered set (the U step in wfFUIK) for a subset of 13 randomly selected refinement targets. We use this subset to demonstrate the difficulties in selection as a proof by contradiction.

It is noteworthy to mention that the step-by-step analysis of the performance is very time consuming as it requires evaluating tens to hundreds of thousands of protein models created by

the Foldit players for each target. Therefore, we chose a random subset consisting of half of the total number of refinement targets attempted by this branch.

Figure 8A demonstrates critical weak points of the method and the results of this analysis. First, the Foldit players were able to refine the initial structure in 12/13 of these targets. In 6/13 targets, the most refined structure in the Unique/Filtered set (pink) has a GDT_TS value that is greater than or equal to the best submitted by any group in CASP10[5] (light blue) by an appreciable margin. Table V shows the best GDT_TS value at each step of the wfFUIK branch (columns) for each of the 13 targets (rows) considered in this test. The last 2 rows show the accumulated GDT_TS values (∑GDT_TS) over the 13 targets, as well as the difference (Δ∑GDT_TS) between the ∑GDT_TS for any column and the ∑GDT_TS for the column corresponding to the starting structures. If the best structure in the Foldit+Unique column could be selected then the total improvement by GDT_TS would be +52.38, which is in line with the very best from each single method submitted to CASP10[5] of +53.17 (last column). Unfortunately, none of the selection procedures used by FOLDIT (described in the Methods section) or the WeFold branches were able to select these models.

Despite the ranking problems, the wfFUIK method overall did much better than the naïve approach of taking the lowest energy structure from a single energy function such as dDFIRE[24] or Rosetta[23]. Choosing dDFIRE[24] or Rosetta[23] as a selection strategy would yield a Δ∑GDT_TS of -133.69 and -120.24 for these 13 targets (see Table V), which represent a GDT_TS degradation of 14% and 12% respectively. The ICON step in wfFUIK selected on average a subset of 23 structures. In 11/13 targets, this subset included structures with a higher GDT_TS value than the lowest energy dDFIRE or Rosetta structure. Therefore, the ICON step achieves an

improvement of 104.45 GDT_TS points over the lowest energy dDFIRE and 91 GDT_TS points over the lowest energy Rosetta conformers, a 10% improvement over the naïve method.

KoBaMIN was able to refine the best structures contained in the ICON set in 10/13 targets, thus contributing small, but consistent gains in GDT_TS in the pipeline. Each step (Foldit, Unique, ICON, KoBaMIN) worked together to identify candidate structures that were often better than the lowest energy of a single metric alone. Unfortunately, the last step of the wfFUIK method, which ranked the subset of KoBaMIN-refined structures and selected the 5 models for submission failed to pick the best structures in 7/13 targets (dark blue).

Besides the weakness identified at the final ranking step of the pipeline, another problem in wfFUIK was associated to the small size of the Unique/Filtered set. An ideal selection procedure would be able to enrich the probability of selecting among the most refined structures from the massive number of conformers available. In the Unique/Filtered set, there were only 6/13 targets where there were more than 10 structures that were refined relative to the starting GDT_TS. In 6 of the remaining 7 targets, there were 3 or less better structures than the starting structure. To our knowledge, there is no current method capable of picking those structures among the thousands in the Unique/Filtered sets (average of 15,836 conformers) and hundreds of thousands in the unfiltered set (Supplementary Table S2). The method as it currently stands was unable to consistently select from the small pools of refined models among the significantly larger set of total models contained in the Unique/Filtered sets, as demonstrated by this subset of targets. These results suggest the filtering step (the "U" step in wfFUIK) may have been too stringent for these sets that contain many very similar structures. Based on the challenges observed, we believe a refinement strategy capable of addressing the selection of the best model among a large pool of candidate models would be of utility.

What Went Right and What Went Wrong Beyond the Pipelines

The WeFold experiment showed that part of the protein structure prediction community is ready to collaborate at a larger-scale. More importantly, the WeFold experiment showed that such collaboration can produce the following results, which include:

1. A cyber-infrastructure that facilitates frequent, open discussions among researchers and allows the creation of hybrid pipelines composed of state-of-the-art methods thus leveraging their strengths at a scale that had not been tried before.
2. Resources to share and process extremely large data files and databases and to execute computationally expensive codes.
3. The creation of pipelines by the contributing labs themselves. Having the experts contributing their developed methods in their best way possible presumably is better than the alternative approach of a single lab/person attempting to utilize all of the methods in isolation without any guidance of how to best use them. The collaborative approach also allowed for the ability to adapt sub-methods to work in the setting of WeFold within a reasonable amount of time, since the individual groups are the experts to modify the source code of their own methods. Adaptation of the source codes developed by diverse groups, so that they can work together, may have proven difficult and time-consuming for a single group.
4. The generation of a vast number of decoys. Almost 9 million decoys were contributed by the different groups as shown in Supplemental Table S2. This is over 170 times the number of models submitted to CASP10[5]. If curated, these structures could be very useful for designing, training, and improving energy and scoring functions, which as we

have just shown, have difficulty in selecting top models among an ensemble of very similar decoys.

5. A unique opportunity for students and young researchers to interact on a daily basis with researchers from other labs, to share their models, and to learn new methods and uses that they can then share with their lab members.
6. Data and discussions which are all publicly available and searchable.

Although the WeFold experiment produced enough evidence to warrant its continuation, it did not produce enough good results to categorically claim success. Some of the issues that stood in the way to success include:

1. Only 13% of the manual groups registered for CASP10[5] participated in the experiment. The project needs to scale up to increase the chances for success. The higher the number of components contributed, the higher the chances to create the ideal combination of methods that perfectly complement each other.
2. The gateway is overly restrictive and is based on an approach that does not scale up. All gateway users have to apply for and get NERSC accounts even if they do not need to use the NERSC computers. NERSC can only issue accounts to those users that are affiliated to a trusted institution.
3. The gateway lacks a workflow feature that permits users to quickly assemble pipelines. Members of a lab contributing a component to any of the pipelines needed to run that component and then pass the output to the next lab contributing the next component in the pipeline. This method of operation made it impossible to quickly optimize and benchmark the new pipelines. It also made the execution of the pipelines very time consuming. Thus, the hybrid methods were created during the first days of CASP10[5] and

there was no time to optimize their components to work within the new pipeline. For example, the ICON component of the wfFUIK branch was originally designed by the Floudas lab to deal with the *hundreds to thousands* of structures generated by their own pipeline. Therefore, it needed to be adapted to work with *hundreds of thousands* of structures in the wfFUIK pipeline. Because there was no time for adapting the algorithm or parallelizing the code, a preprocessing step had to be included in wfFUIK in order to filter out structures before using ICON. This filtering solution step may have been too stringent for the refinement targets as the models generated are usually very similar but there was no time to change the procedure.

4. The gateway lacks a feature that allows users to run their individual codes and collaboratively create pipelines directly via the gateway. Therefore, users could not use the gateway as a launching platform to run their codes and their collaborators could not see the status of those runs.

Discussion

Successful blind prediction of a protein's structure requires the correct identification of its secondary structure, the best template and sequence alignment, the accurate prediction of contacts, the accurate prediction of the β-sheet topology, broad sampling, and selection among the populations of structures generated[13]. Each one of these problems is extremely complex and no single lab, no matter how big its resources, has found the optimal solution to all of them. Therefore, we strongly believe that success will come by combining methods and expertise from the different labs, organizations, and individuals that have a stake at solving this problem.

The WeFold experiment was created to realize this potential. WeFold is a social-network-based experiment that comprises both a platform that provides a cyber-infrastructure (high performance

computing, science gateway, and advanced networking) and a community of committed individuals that strongly believe in collaboration to advance science. WeFold brought together junior and seasoned scientists from thirteen labs around the world. Prior to WeFold these groups used their own prediction methods to compete against each other during CASP. WeFold enabled them to compete and collaborate within the same venue; therefore, coopetition accurately describes this interaction. A science coopetition of this magnitude that takes advantage of both expert and citizen scientists from around the world is unprecedented. The execution of such an ambitious project is not straightforward and it is important to determine what needs to be done to warrant its continuity and success.

The WeFold pipelines are a combination of components, which are part of state-of-the-art methods that have been optimized and benchmarked and so, outperforming them was a difficult task. Nevertheless, the WeFold experiment shows that it is possible for a collaborative method to outperform its base method. For example, the combination of UNRES with the contact predictions of Floudas group and the KobaMIN from the Levitt group for the prediction of Target T0740 led to a model that was closer to the experimental one that those produced by UNRES alone; and the combination of Foldit with a multi-step process consisting of filtering, clustering, refinement, and selection could identify improved models among the hundreds of thousands of models created by players in more cases than Foldit alone. However, the improvements were not consistent across base methods and prediction categories and in certain cases it was not possible to perform a fair comparison due to discrepancies between the input data used by the WeFold branches and the base methods. Hence, it is too early in this collaborative approach to science to yield any larger conclusions. Nevertheless, the results shown here warrant its continuity and highlight the areas in need of improvement. It is also evident that

new and different collaborative methods must be tried in future WeFold experiments to maximize the chances of finding the ideal combination of methods that optimally complement each other. For example, it may be interesting for scientific value and for the community in general to create pipelines that combine the very top structure prediction method(s) and the very top refinement methods. These methods themselves are automated pipelines and may be amenable to combination.

Conclusion

The implementation of a coopetitive effort is not a trivial task. Therefore, we decided to tackle it step-wise. First, we had to gauge the community to determine if there was enough interest for a collaborative project of this scope. Fortunately, we realized that at least part of it was ready. Second, we had to develop a prototype infrastructure with the essential features to support such collaboration. We determined that the basic features needed to start were password-protected online discussions and sharing of files and we implemented such prototype. We also obtained the resources that were essential to run the expensive codes and store the files. Using these tools, the community created new hybrid pipelines, contributed an overwhelming number of models and even shared their final models which were ranked by different techniques. The project has recently initiated its next phase and efforts are currently underway on the following fronts:

Science gateway: We are collaborating with members of the Science Gateway Institute (http://sciencegateways.org) to develop a new science gateway using cutting-edge technologies[64]. This gateway will provide users with the tools to collaborate at a larger-scale, as well as to assemble pipelines, run codes, manage large data, and selectively share information with other groups or with the public.

Pipelines: The community has identified some critical areas that need improvement as described in this work. Efforts are underway to optimize and fine-tune these methods tested during CASP10[5] with the goal to further enhance their combined predictive ability. These efforts include a hierarchical clustering method to efficiently and effectively select representatives from the large set of protein structures and a scoring function that takes advantage of machine learning techniques and the massive amount of structural data generated during the WeFold experiment that is now publically available.

Reach out to the community beyond CASP: Our goal to engage the large community is very ambitious but again, we plan to accomplish this gradually. We will create a database of WeFold decoys that will be available for downloading and querying, as well as its associated metadata that include protein-like features such as those proposed and successfully used for selecting models (Chen Keasar, Personal Communication). We have initiated discussions with members of the machine learning community and will present the scoring function challenge to them during an upcoming workshop.

The barriers to access the gateway must be lowered to allow the expansion of the WeFold community. In this new approach to doing science, ideas can come from anywhere, talents can come from everywhere, and we need to ensure that WeFold provides an environment where people have the opportunity to express their ideas, to bring their best contribution, and to merge those ideas and contributions into methodologies that accelerate innovation and push the protein folding field forward.

Acknowledgements. This research used significant resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract number DE-AC02-05CH11231. CAF acknowledges

support from the National Institutes of Health grant number R01GM052032 and the National Science Foundation. GAK is grateful for support by a National Science Foundation Graduate Research Fellowship under grant number DGE-1148900. GAK, JS, & CAF gratefully acknowledge computing time from Princeton Institute for Computational Science and Engineering (PICSciE). RAF and LOB are grateful for support from the Brazilian funding agencies Coordination of the Enhancement of Higher Education (CAPES) and São Paulo Research Foundation (FAPESP). This research was also supported by grant GM-14312 from the U.S. National Institutes of Health (to HAS), by grant MCB10-19767 from the U.S. National Science Foundation for financial support (to HAS), by grant UMO-2012/06/A/ST4/00376 from the Polish National Science Center (to AL), and by grant the Polish National Science Center UMO-2011/01/N/ST4/01772 (to AKS). PK, MAM, and TW were supported by grant MPD/2010/5 from the Foundation for Polish Science. YH, AL, AKS, PK, MAM, and TW gratefully acknowledge the use of computing resources at the Informatics Center of the Academic Computer Network in Gdansk (TASK), Department of Chemistry and Chemical Biology, Cornell University, and Faculty of Chemistry, University of Gdansk. The work of JS and HZ was supported by grant numbers GM-48835, GM-37408 and GM-08422 of the Division of General Medical Sciences of the National Institutes of Health. JC acknowledges the support from the National Institutes of Health grant (grant number: R01GM093123). SC, JF, KX, and ZP acknowledge support from Office of Naval Research grant N00014-12-C-0158. CK acknowledges support from BSF grant 2009432. The authors thank Johannes Soeding and Armin Meier for access to HHPred predictions during the CASP10 experiment. GC and ML acknowledge the National Institutes of Health award GM063817. ML is the Robert W. and Vivian K. Cahill Professor of Cancer Research. SNC and the WeFold community are very

grateful to Steven Chan, Shreyas Cholia, David Skinner, and especially to Francesca Verdier of NERSC for their HPC support. SNC is very grateful to Cristina Siegerist (http://www.cristinasiegerist.com/ComputingVisualization/index.html) for her artistic design of the WeFold gateway banner.

Author Contributions. SNC conceived the project. GAK, AL, FK, HZ, GC, LOB, SNC, RAF, XD, JF, YH, PK, JL, MAM, AS, JS, TW, KX, CK, and FP performed the research. GAK, CAF, AL, FK, HZ, GC, and SNC analyzed the data. AL, DB, JB, JC, ACBD, CAF, CK, ML, ZP, JS, HS, and SNC supervised the research. GAK, SNC, AL, FK, HZ, LOB, RAF, MAM, and GC wrote the paper. The corresponding lab members participated in the branches by contributing their published methods and CPU time to the collaboration. All authors have read and approved the paper.

References

1. Khatib F, DiMaio F, Foldit Contenders Group, Foldit Void Crushers Group, Cooper S, Kazmierczyk M, Gilski M, Krzywda S, Zabranska H, Pichova I, Thompson J, Popović Z, Jaskolski M, Baker D. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol 2011;18(10):1175-1177.
2. Protein Structure Prediction Center. 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction Davis: University of California, Davis; 2013.
3. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins: Struct, Funct, Bioinf 1995;23(3):ii-iv.
4. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15(3):285-289.
5. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) — round X. Proteins: Struct, Funct, Bioinf 2014;82:1-6.
6. Nugent T, Cozzetto D, Jones DT. Evaluation of predictions in the CASP10 model refinement category. Proteins: Struct, Funct, Bioinf 2014;82(S2):98-111.
7. MacCallum JL, Pérez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. Proteins: Struct, Funct, Bioinf 2011;79(S10):74-90.
8. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Assessment of the protein-structure refinement category in CASP8. Proteins: Struct, Funct, Bioinf 2009;77(S9):66-80.
9. Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. Proteins: Struct, Funct, Bioinf 2014;82(S2):164-174.
10. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;Suppl 5:163-170.
11. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18(3):342-348.
12. Dill KA, MacCallum JL. The Protein-Folding Problem, 50 Years On. Science 2012;338(6110):1042-1046.
13. Khoury GA, Smadbeck J, Kieslich CA, Floudas CA. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol 2014;32(2):99-109.
14. Floudas CA. Computational methods in protein structure prediction. Biotechnol Bioeng 2007;97(2):207-213.
15. Bourne PE. CASP and CAFASP Experiments and Their Findings. Structural Bioinformatics: John Wiley & Sons, Inc.; 2005. p 499-507.
16. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kassner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Honigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Bjorne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJ, Skunca N, Supek F, Bosnjak M, Panov P, Dzeroski S, Smuc T, Kourmpetis YA, van Dijk AD, ter Braak CJ, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney

SD, Friedberg I. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10(3):221-227.
17. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. Protein Sci 2009;14(2):278-283.
18. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ. The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des 2010;24(4):259-279.
19. Callaway E. Mutation-prediction software rewarded. Nature 2010.
20. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Foldit Players. Predicting protein structures with a multiplayer online game. Nature 2010;466(7307):756-760.
21. Hubbard SJ, Thornton JM. 'NACCESS', computer program. 1993.
22. Subramani A, DiMaggio PA, Floudas CA. Selecting High Quality Protein Structures from Diverse Conformational Ensembles. Biophys J 2009;97(6):1728-1736.
23. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2011;487:545-574.
24. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins: Struct, Funct, Bioinf 2008;72(2):793-803.
25. Chopra G, Kalisman N, Levitt M. Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins: Struct, Funct, Bioinf 2010;78(12):2668-2678.
26. Chopra G, Summa CM, Levitt M. Solvent dramatically affects protein structure refinement. Proceedings of the National Academy of Sciences 2008;105(51):20239–20244.
27. Rodrigues JP, Levitt M, Chopra G. KoBaMIN: a knowledge-based minimization web server for protein structure refinement. Nucleic Acids Res 2012;40(W1):W323-W328.
28. Zhou H, Skolnick J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophys J 2011;101(8):2043-2052.
29. Wang Z, Eickholt J, Cheng J. APOLLO: A Quality Assessment Service for Single and Multiple Protein Models. Bioinformatics 2011;27:1715-1716.
30. Zhou H, Skolnick J. Protein Structure Prediction by Pro-Sp3-TASSER. Biophys J 2009;96(6):2119-2127.
31. Zhou H, Skolnick J. Ab Initio Protein Structure Prediction Using Chunk-TASSER. Biophys J 2007;93(5):1510-1518.
32. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 2004;101(20):7594-7599.
33. Gront D, Kolinski A, Skolnick J. A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. J Comput Phys 2001;115(3):1569-1574.
34. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 2004;25(6):865-871.
35. Liwo A, Ołdziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem 1997;18(7):849-873.

36. Liwo A, Czaplewski C, Pillardy J, Scheraga HA. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. The Journal of Chemical Physics 2001;115(5):2323-2347.
37. Liwo A, Czaplewski C, Ołdziej S, Rojas AV, Kazmierkiewicz R, Makowski M, Murarka RK, Scheraga HA, Voth G. Simulation of protein structure and dynamics with the coarse-grained UNRES force field. Coarse-Graining of Condensed Phase and Biomolecular Systems 2008;1:1391-1411.
38. Liwo A, Khalili M, Czaplewski C, Kalinowski S, Ołdziej S, Wachucik K, Scheraga HA. Modification and Optimization of the United-Residue (UNRES) Potential Energy Function for Canonical Simulations. I. Temperature Dependence of the Effective Energy Function and Tests of the Optimization Method with Single Training Proteins. The Journal of Physical Chemistry B 2006;111(1):260-285.
39. He Y, Xiao Y, Liwo A, Scheraga HA. Exploring the parameter space of the coarse-grained UNRES force field by random search: Selecting a transferable medium-resolution force field. J Comput Chem 2009;30(13):2127-2135.
40. Czaplewski C, Kalinowski S, Liwo A, Scheraga HA. Application of Multiplexed Replica Exchange Molecular Dynamics to the UNRES Force Field: Tests with α and α+β Proteins. J Chem Theory Comput 2009;5(3):627-640.
41. Wei Y, Thompson J, Floudas CA. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 2012;468(2139):831-850.
42. Subramani A, Floudas CA. β-sheet Topology Prediction with High Precision and Recall for β and Mixed α/β Proteins. PLoS One 2012;7(3):e32461.
43. Rajgaria R, McAllister SR, Floudas CA. Towards accurate residue–residue hydrophobic contact prediction for α helical proteins via integer linear optimization. Proteins: Struct, Funct, Bioinf 2009;74(4):929-947.
44. Rajgaria R, Wei Y, Floudas CA. Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. Proteins: Struct, Funct, Bioinf 2010;78(8):1825-1846.
45. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 1992;13(8):1011-1021.
46. Kaźmierkiewicz R, Liwo A, Scheraga HA. Addition of side chains to a known backbone with defined side-chain centroids. Biophys Chem 2003;100(1–3):261-280.
47. Kaźmierkiewicz R, Liwo A, Scheraga HA. Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method. J Comput Chem 2002;23(7):715-723.
48. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Struct, Funct, Bioinf 2004;57(4):702-710.
49. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 2008;4(3):435-447.
50. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18(12):1463-1472.

51. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Struct, Funct, Bioinf 2010;78(8):1950-1958.
52. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Struct, Funct, Bioinf 2004;55(2):383-394.
53. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314(1):141-151.
54. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins: Struct, Funct, Bioinf 2011;79(S10):1-5.
55. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370-3374.
56. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins: Struct, Funct, Bioinf 2014;82(S2):7-13.
57. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38-56.
58. Tai C-H, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. Proteins: Struct, Funct, Bioinf 2014;82(S2):57-83.
59. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12(2):85-94.
60. He Y, Mozolewska MA, Krupa P, Sieradzan AK, Wirecki TK, Liwo A, Kachlishvili K, Rackovsky S, Jagieła D, Ślusarz R, Czaplewski CR, Ołdziej S, Scheraga HA. Lessons from application of the UNRES force field to predictions of structures of CASP10 targets. Proceedings of the National Academy of Sciences 2013;110 (37):14936-14941.
61. Mirjalili V, Noyes K, Feig M. Physics based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins: Struct, Funct, Bioinf 2014;82(S2):196-207.
62. Heo L, Park H, Seok C. GalaxyRefine: protein structure refinement driven by side-chain repacking. Nucleic Acids Res 2013:W384–W388.
63. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66(Pt 1):12-21.
64. Crivelli SN, Dooley R, Holmes R, Mock S. The WeFold gateway: Enabling large-scale science coopetition. 2013. 2013 IEEE International Conference on Cluster Computing (CLUSTER).
65. DeLano WL. The PyMOL Molecular Graphics System. Palo Alto, CA: Delano Scientific LLC; 2008.

Figure Legends

Figure 1: Visual depiction of the five WeFold branches collaboratively formed and tested during CASP10. wfFUIK and wfFUGT both began with structures generated by human players in the online multiplayer game Foldit. Foldit allows players to fold proteins independently on their home computers, as well as share their predictions with other players around the world. Foldit players are also able to fold structures by hand using the XBOX Kinect (http://fold.it/portal/node/993534) as well as with the Leap Motion (http://fold.it/portal/node/995117). These generated structures were subsequently clustered, scored, and refined. wfCPUNK is an *ab initio* branch combining secondary structure prediction, beta-sheet topology prediction, contact prediction, coarse-grained replica-exchange molecular dynamics, and clustering. WeFold Branch and WeFold Mix began with the structures generated by the other branches and/or from servers participating in CASP10**.**

Figure 2: wfCPUNK (black lines) outperformed all individual components and all other groups and methods for Free Modeling target T0740_D1. The model produced by wfCPUNK is shown in the inset with a rainbow color, aligned with the native shown in black. The average prediction among all groups for this target had a GDT_TS of 21.68±4.55. Interestingly, the individual groups contributing to the method also outperform the average in a statistically significant fashion, with the combined method outperforming the individual methods. This figure was adapted from a GDT_TS plot generated on the CASP10 website, with permission[2].

Figure 3: Illustration of the *ab initio* predicted helix-helix contacts[43,44] implemented as restraints in UNRES sampling superposed on the experimental structure of T0740. The restraints are superposed on the experimental structure (A), the best model (model 4) from the wfCPUNK group (B), and the best model (model 3) from the Cornell-Gdansk group (C). The restrained parts of the molecule that belong to the same sets of restraints are marked with the same color, from blue to red from the N- to the C-terminus. It can be seen that the restraints marked with orange color made the C-terminal α-helix, which is straight in the Cornell-Gdansk model, bend. The restraints marked with red color made the C-terminal α-helix pack with the long middle α-helix, similar to the packing in the experimental structure. This figure was created using PyMOL[65].

Figure 4: Illustrative walkthrough of wfFUIK pipeline for the blind prediction of refinement target TR722 in CASP10. (A) The starting structure was given to Foldit players, who produced 256,776 structures. A filtering procedure is applied to remove unlikely candidate structures. 20,488 monomers and 30,855 symmetric structures remained after filtering, shown as grey dots in each plot. The starting structure, shown as a red cross and highlighted as a red dotted line, had a GDT_TS of 0.58. The structures that would be chosen by naïvely selecting the ones with the lowest Rosetta energy for both the monomer and symmetric dimer are shown as pink squares. (B) From the Unique/Filtered set, structures that were the cluster medoids selected by ICON, as well as lowest energy Rosetta and dDFIRE structures are highlighted in blue. (C) Structures produced after running the previous set through KoBaMIN are shown as cyan dots. (D) Structures submitted in blind prediction are shown as pink stars. The GDT_TS is shown here normalized on a zero to one basis.

Figure 5: (A) Comparison of native structure (black) and refined structure (green) produced by wfFUIK branch for TR722. The refined structure using this protocol had a GDT_TS of 65.95, whereas the starting model (red) had a GDT_TS of 58.0. This structure is a dimer and adopts a coiled-coil fold. (B) Illustration of best Model 1 prediction produced by any method in CASP10 to improve the metric GDT_TS for target TR705. The WeFold method wfFUGT achieved this improvement, increasing the starting GDT_TS from 64.84 to 70.05. The loops in the upper-right region of the Figure, as well as in the bottom left were the regions where the most refinement occurred. TR705 adopts a β-sandwich fold. These examples highlight significant improvements in refinement for proteins containing only α-helices and β-sheets. This figure was created using PyMOL[65].

Figure 6: Comparison of improvements in Foldit models by WeFold methods. (A) In wfFUIK, 74% of structures were better refinements than the best structure submitted by FOLDIT. (B) Using wfFUGT, 53% of structures were better refinements than the best structure submitted by FOLDIT. The improvements are indicated by the differential bars in white from the base bars in red.

Figure 7: (A) Absolute and (B) relative performance of the WeFold branch on the 43 human targets attempted. The absolute performance is assessed based on the Z-score of the GDT_TS of the best model submitted by WeFold relative to all other predictions by all groups and methods for each target. The relative comparison is based on the ratio between the GDT_TS score of the best WeFold prediction and the best GDT_TS achieved by all groups for each target. In both cases, longer positive bars in the y-direction represent better performance.

Figure 8: (A) Breakdown of the effect of each step in the wfFUIK pipeline in order to identify individual contributions to the pipeline, as well as areas needing further attention. The y-axis is normalized to show the ratio of the GDT_TS of the corresponding model to the GDT_TS of the starting model so that it can be compared across targets. The legend shows the lowest energy dDFIRE structure (black) in the Unique/Filtered set, lowest energy Rosetta structure from all Foldit conformations (red), best GDT_TS contained in the ICON+Lowest E Rosetta+Lowest E dDFIRE step of the pipeline (green), best GDT_TS contained after those structures are refined by KoBaMIN (yellow), best blind prediction of 5 submitted in CASP10 (dark blue), highest GDT_TS structure contained in the Foldit Unique/Filtered set (pink), and the best GDT_TS structure submitted to CASP10 by any team (light blue). (B) Enrichment of candidate conformers by wfFUIK compared to candidate conformers in the Unique/Filtered set. Shown are the probabilities of selecting a better structure than the start in the Unique/Filtered set compared to the enriched probability when choosing from the final set of wfFUIK models.

Target	wfCPUNK	Cornell-Gdansk	KIAS-Gdansk	FLOUDAS
T0740	32.1	25.48	24.03	30.81
T0734	15.57	14.74	12.97	12.97
T0741	12.8	14.8	14.8	13
T0666*	23.47	19.72	18.61	23.89

Table I. GDT_TS values for the top prediction by wfCPUNK compared to other groups contributing to wfCPUNK for Free Modeling targets. *Structure was submitted late in CASP competition and was not evaluated by the assessors. CASP10 discussions and results pertaining this target can be found at http://www.wefold.org.

Group	Target	Best GDT_TS Improvement per target	Number of targets for which ΔGDT_TS > 0
Anthropic_Dreams	TR671	10.51	3
Void_Crushers	TR663	8.06	5
wfFUIK	TR722	7.87	6
Schroderlab	TR705	6.51	7
FEIG	TR723	6.30	23
Schroderlab	TR704	5.85	
FOLDIT	TR710	5.15	3
Seok	TR681	4.84	16
BAKER	TR696	4.50	8
FEIG	TR738	4.42	
FEIG	TR662	4.33	
FEIG	TR750	3.71	
FRESS_server	TR754	3.67	6
Mufold-R	TR720	3.03	6
Pcons-net	TR644	2.84	1

Table II. Top 15 best per-target improvements in GDT_TS made by any team considering all 5 models submitted and the number of targets for which each team improved the original model. Foldit-based methods are highlighted in yellow.

Prediction Target	FOLDIT (068)	Void Crushers (165)	Anthropic Dreams (085)	Contenders (341)	wfFUIK (149)	wfFUGT (260)	Average of All CASP Predictions (± 1 Standard Deviation)
TR644	78.55	71.99	79.61	81.92	82.45	**83.33**	76.57±11.47
TR655	65.86	67.57	64.00	**69.29**	69.00	64.57	65.08±3.74
TR661	64.32	68.65	65.68	NA	65.81	NA	**74.11±7.21**
TR662	81.00	83.67	82.00	83.33	**84.00**	83.33	79.29±9.40
TR663	69.08	**77.30**	76.48	NA	54.93	54.93	60.22±12.10
TR671	61.08	64.20	**66.19**	51.70	59.66	59.66	49.31±11.54
TR674	77.65	77.84	78.98	78.60	**82.01**	NA	81.76±5.08
TR679	67.08	69.72	69.97	**72.24**	69.85	NA	68.42±6.51
TR681	76.57	74.22	74.48	73.30	**77.22**	NA	70.32±14.63
TR688	69.19	**74.32**	72.57	NA	71.76	NA	74.18±5.37
TR689	81.08	**85.05**	84.93	81.42	83.88	NA	83.01±7.29
TR696	59.75	64.75	63.75	61.00	65.25	**67.75**	63.01±9.90
TR698	63.23	64.58	64.92	61.34	64.71	**65.13**	63.12±3.44
TR705	64.84	64.06	69.27	63.80	64.84	**70.05**	60.39±9.66
TR708	78.57	82.65	81.50	78.32	**85.71**	81.00	82.44±3.54
TR710	**80.28**	77.83	74.87	74.23	78.09	76.16	72.33±3.74
TR712	81.18	86.83	85.22	82.26	**88.04**	86.42	86.65±5.90
TR722	47.24	59.05	54.33	60.83	**65.94**	43.11	52.92±7.47
TR723	83.40	79.01	83.02	86.45	**86.83**	82.82	82.57±4.10
TR747	**86.11**	83.89	81.11	84.72	83.89	79.72	79.99±5.09
TR750	67.44	66.35	67.17	68.68	72.67	71.57	**72.88±5.76**
TR752	86.49	88.85	**89.02**	88.34	88.51	88.51	86.75±3.12
TR754	73.90	73.90	**75.73**	60.66	71.69	56.25	70.26±7.54
# Wins	2	3	3	2	7	4	2

Table III: Comparison of best refinement structures submitted in CASP10 by Foldit-derived branches to the average GDT_TS value calculated over all predictions submitted by all groups. Besides the average GDT_TS value, the corresponding standard deviations are provided for each target. GDT_TS values were tabulated using the prediction center's official results on the CASP10 webpage[2]. NA indicates that a structure was not submitted and evaluated for this target. Group numbers are denoted in parentheses. The best prediction for each target among these methods is bolded. Foldit-based human groups Void Crushers and Anthropic Dreams were able to refine and select the best structure among the cohort three times each, whereas FOLDIT was able to select it for two targets. In these two cases though, Foldit's selection produced the top improvement in GDT_TS of any method in CASP10. These teams were a collaboration of citizen scientists. The WeFold branches wfFUIK selected the best structure seven times and wfFUGT four times indicating the enhanced performance of using these synergistic branches utilizing the structures refined by the citizen scientists followed by subsequent clustering, further refinement, and selection.

Prediction Target	TASSER (079)	FLOUDAS (077)	Cornell-Gdansk (152)	FOLDIT (068)	wfFUIK[1] (149)	wfFUGT[1] (260)	WeFold Branch (101)	wfCPUNK (287)	WeFoldMix (441)	Average of All CASP10 Predictions (± 1 Standard Deviation)	Best CASP10 Prediction by All Groups, All Methods
T0644-D1	83.51	28.72	17.38	**84.93**	75.89	60.99	84.04	18.97		61.78±23.62	85.28
T0649-D1	**33.42**	21.74	16.44				28.40			20.21±8.10	36.82
T0651	**33.47**	11.42								28.30±7.28	37.70
T0655-D1	76.67	25.67	21.33				**76.83**	23.67		60.62±17.10	79.83
T0663	**40.62**	19.08	23.19				**40.46**			32.80±8.90	42.93
T0666-D1	**32.64**	23.89	19.72				31.11			22.66±4.50	33.75
T0668-D1	**40.06**	33.01	35.90	36.54	32.05	30.45	36.54	30.77		30.34±4.73	44.23
T0673-D1	**65.73**	34.68	34.27	50.81	40.73	46.37	46.37	27.02		35.97±10.27	66.94
T0676-D1	26.45	20.38	22.11				**38.87**	18.50		21.75±7.35	43.21
T0678-D1	37.99	28.41	25.65	35.55	34.09		**39.29**	25.00		25.81±6.71	42.53
T0680-D1	**75.52**	51.82	31.51	50.26	49.22	50	67.19			45.35±14.74	77.60
T0684	14.52	13.07	**15.14**				14.42			12.12±1.87	18.67
T0687-D1	**76.50**	58.75					**76.50**			64.91±15.10	78.25
T0691-D1	52.59	46.46		48.35	53.30	**53.77**	30.90	23.82		33.57±12.32	57.31
T0700-D1	85.00	62.86	55.00	65.71	55.71	55	**94.29**	55.00	52.86	61.46±14.53	96.43
T0704-D1	**69.91**	53.68					69.81			62.28±6.97	71.00
T0707-D1	**54.59**	38.55					52.65			38.24±13.75	54.59
T0709-D1	**96.88**	96.88	53.12	95.83	93.75	94.79	**96.88**	50.00	93.75	86.43±16.78	98.96
T0711-D1	**89.84**	87.50	42.19	83.59	82.81	85.94	89.06	40.62	45.31	73.36±17.35	90.62
T0713	27.41	**27.81**					26.94			21.69±6.62	30.28
T0717	33.14	29.24	9.55				**33.39**			23.73±8.28	38.54
T0719	11.59	**12.73**					11.21			8.82±2.86	16.15
T0720-D1	56.31	53.66	14.39				**56.44**			37.25±13.50	65.78
T0724	29.70	**31.62**					26.92			21.86±6.25	31.62
T0732	**42.38**	36.71	12.14				38.15			31.09±11.07	42.96
T0734-D1	**18.51**	12.97	14.74				14.51	15.57		13.92±2.41	23.82
T0735	12.38	10.05	9.35				**22.20**			10.54±4.65	28.66
T0739	9.89	7.36					**10.26**			6.24±2.78	14.38
T0740-D1	30.16	30.81	25.48	24.52	29.52	22.58	24.84	**32.10**		21.68±4.55	38.87
T0741-D1	13.20	13	**14.80**				11.40	12.80		12.09±1.23	17.20
T0742-D1	**52.45**	22.34	13.51				20.00			23.32±15.96	57.13
T0743-D1	**70.17**	69.74	25.88	63.60	61.40	64.03	49.78	25.66		42.67±15.12	73.25
T0744-D1	38.69	35.54	9.54				**44.00**	15.00		31.49±13.17	61.62
T0746-D1	**52.82**	49.16	7.62				48.70	7.85		37.82±15.26	52.82
# Wins	**17**	3	2	1	0	1	12	1			

Table IV: Comparison of best structures submitted in CASP10 for tertiary structure prediction by WeFold branches and their component methods to the best and average predictions submitted by all groups. The average GDT_TS value and corresponding standard deviations for all CASP Model submissions are provided for each target. GDT_TS values were taken from the prediction center's official results available on the CASP10 website[2]. An empty cell indicates that a structure was not submitted and evaluated for this target. Group numbers are denoted in parentheses. The best prediction for each target among these methods is bolded. [1]Due to a discrepancy in the input data used for FOLDIT, wfFUIK, and wfFUGT tertiary structure predictions as described in the main text, the results cannot be directly compared to those obtained by other methods. The data is tabulated strictly for reporting and it not meant for comparison since they are not fairly comparable.

TARGET	Start GDT_TS	Best in Unique/ Filtered Set	Lowest E dDFIRE	Lowest E Rosetta	Best in ICON	Best after ICON+ KoBaMIN	Best Submitted Model	Best Model 1 in CASP10	Best MODEL in CASP
TR661	80.68	78.51	59.46	55.27	69.69	71.49	65.81	81.35	81.35
TR663	69.24	69.74	41.94	53.62	54.44	54.77	54.93	74.84	77.3
TR679	71.73	74.25	68.22	67.71	71.48	71.61	69.85	73.74	74.37
TR688	78.24	78.24	68.78	67.43	72.7	72.43	71.76	79.73	80.14
TR696	71.5	73.5	61.75	51.25	63.5	63.5	63.5	75.5	76
TR698	65.55	72.27	64.71	59.66	64.71	65.13	64.71	67.02	67.65
TR705	64.84	73.7	65.36	63.02	70.57	70.83	64.84	70.05	71.35
TR710	75.13	83.76	71.78	70.62	80.41	79.51	78.09	77.83	80.28
TR722	57.09	69.69	42.32	58.27	61.02	65.75	65.75	65.75	65.75
TR723	85.11	89.31	72.52	86.45	86.45	86.83	86.83	88.17	91.41
TR747	83.61	89.44	86.39	78.33	86.39	87.78	83.89	85.28	86.11
TR752	90.37	91.22	86.32	83.28	88.01	88.51	88.51	90.71	90.88
TR754	78.31	80.15	48.16	56.25	72.79	71.32	71.32	79.04	81.98
GDT_TS	**971.4**	**1023.78**	**837.71**	**851.16**	**942.16**	**949.46**	**929.79**	**1009.01**	**1024.57**
GDT_TS from Start	**0**	**52.38**	**-133.69**	**-120.24**	**-29.24**	**-21.94**	**-41.61**	**37.61**	**53.17**

Table V: Illustration of the effect of each step in the wfFUIK method on 13 CASP10 refinement targets. The first column indicates the CASP10 target. The second column indicates the starting structure's GDT_TS value as calculated using the TMScore program[48]. The third and fourth columns indicate the GDT_TS of the lowest energy dDFIRE[24] and Rosetta[23] structures, respectively. The fifth column indicates the best structure contained in the set of candidates after ICON clustering, in addition to the lowest energy dDFIRE and Rosetta structures. The seventh column indicates the GDT_TS of the best structure after refinement by KoBaMIN[27]. The eighth column is the best structure submitted in blind predictions during CASP10 by the wfFUIK method. The ninth and tenth columns indicate the best Model 1 and best of the five Models submitted in CASP10 by any method. Green text indicates that the structure was a refinement of the start. Blue text indicates that the best contained model is more accurate than any submitted during CASP10. This demonstrates the selection challenge the structure prediction community faces. That is, even though there are sampling methods that can generate high-accuracy structures, the best forcefields and selection methods still cannot pick these "best" structures among the ensembles. If one were able to select the best produced Foldit structure, generated by the game players, this would be comparable to the best structure submitted in CASP by all groups.

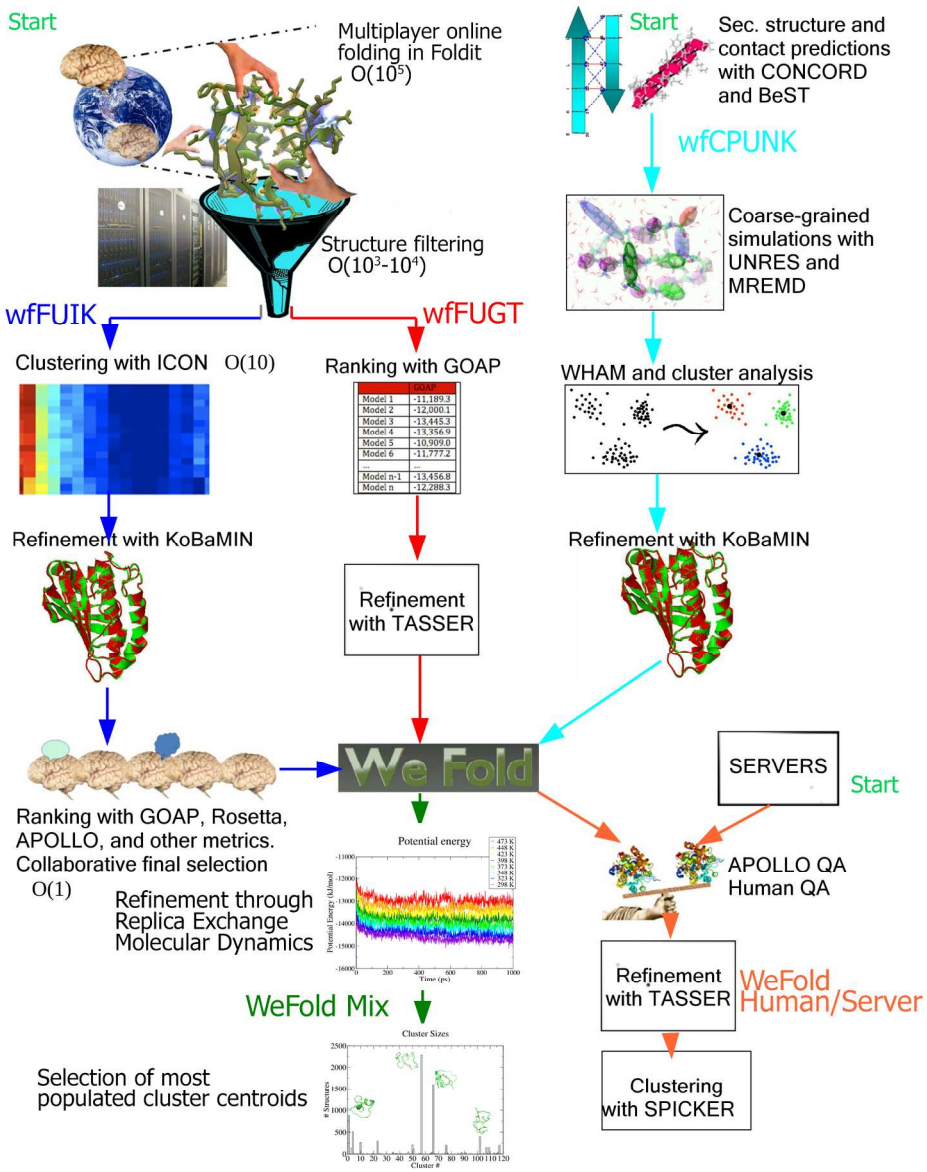

Figure 1: Visual depiction of the five WeFold branches collaboratively formed and tested during CASP10. wfFUIK and wfFUGT both began with structures generated by human players in the online multiplayer game Foldit. Foldit allows players to fold proteins independently on their home computers, as well as share their predictions with other players around the world. Foldit players are also able to fold structures by hand using the XBOX Kinect (http://fold.it/portal/node/993534) as well as with the Leap Motion (http://fold.it/portal/node/995117). These generated structures were subsequently clustered, scored, and refined. wfCPUNK is an ab initio branch combining secondary structure prediction, beta-sheet topology prediction, contact prediction, coarse-grained replica-exchange molecular dynamics, and clustering. WeFold Branch and WeFold Mix began with the structures generated by the other branches and/or from servers participating in CASP10.
187x239mm (300 x 300 DPI)

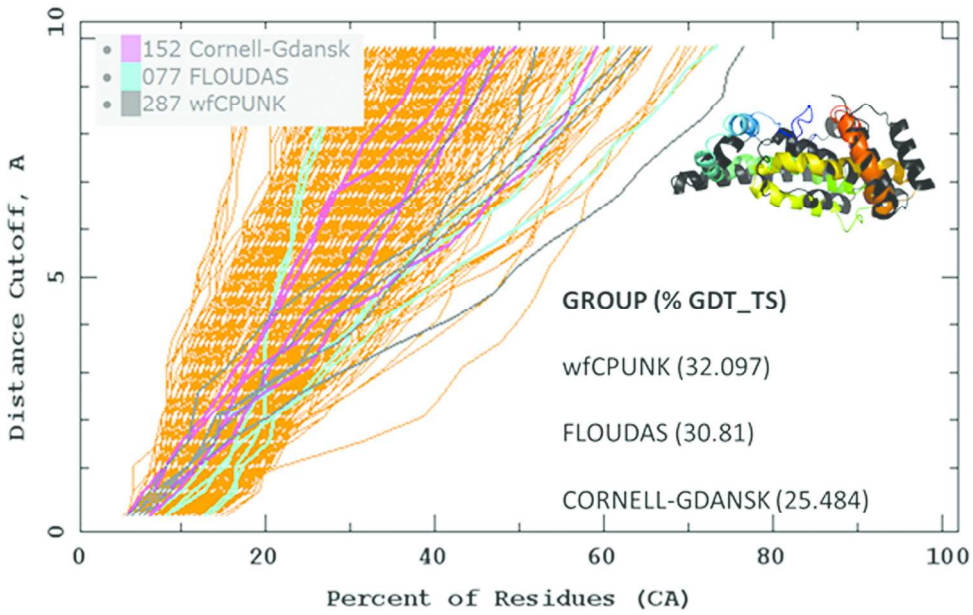

Figure 2: wfCPUNK (black lines) outperformed all individual components and all other groups and methods for Free Modeling target T0740_D1. The model produced by wfCPUNK is shown in the inset with a rainbow color, aligned with the native shown in black. The average prediction among all groups for this target had a GDT_TS of 21.68±4.55. Interestingly, the individual groups contributing to the method also outperform the average in a statistically significant fashion, with the combined method outperforming the individual methods. This figure was adapted from a GDT_TS plot generated on the CASP10 website, with permission.
203x124mm (300 x 300 DPI)

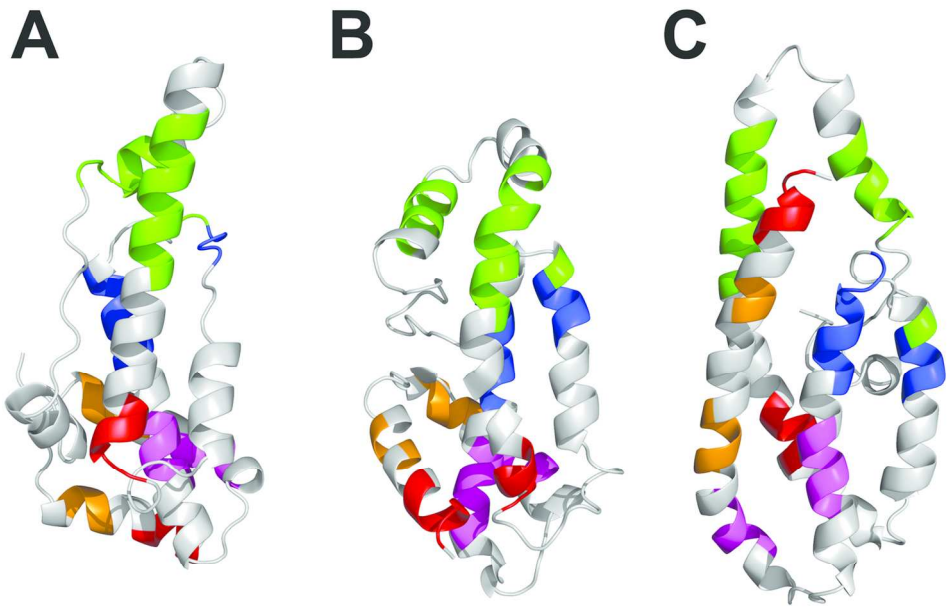

Figure 3: Illustration of the ab initio predicted helix-helix contacts implemented as restraints in UNRES sampling superposed on the experimental structure of T0740. The restraints are superposed on the experimental structure (A), the best model (model 4) from the wfCPUNK group (B), and the best model (model 3) from the Cornell-Gdansk group (C). The restrained parts of the molecule that belong to the same sets of restraints are marked with the same color, from blue to red from the N- to the C-terminus. It can be seen that the restraints marked with orange color made the C-terminal α-helix, which is straight in the Cornell-Gdansk model, bend. The restraints marked with red color made the C-terminal α-helix pack with the long middle α-helix, similar to the packing in the experimental structure. This figure was created using PyMOL.
143x95mm (300 x 300 DPI)

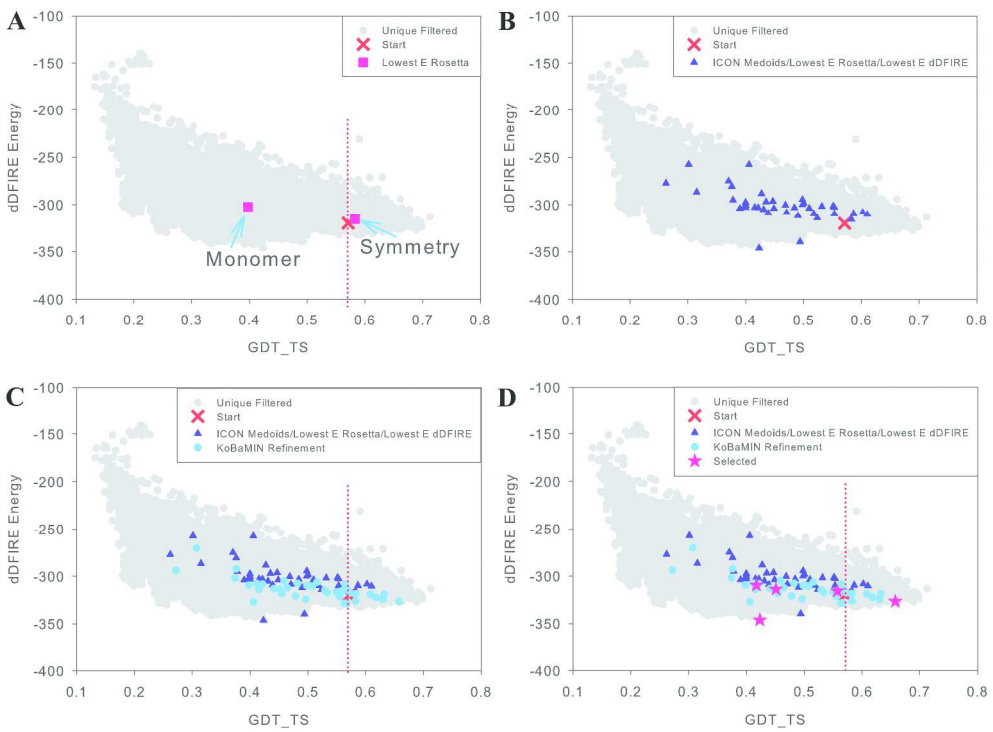

Figure 4: Illustrative walkthrough of wfFUIK pipeline for the blind prediction of refinement target TR722 in CASP10. (A) The starting structure was given to Foldit players, who produced 256,776 structures. A filtering procedure is applied to remove unlikely candidate structures. 20,488 monomers and 30,855 symmetric structures remained after filtering, shown as grey dots in each plot. The starting structure, shown as a red cross and highlighted as a red dotted line, had a GDT_TS of 0.58. The structures that would be chosen by naïvely selecting the ones with the lowest Rosetta energy for both the monomer and symmetric dimer are shown as pink squares. (B) From the Unique/Filtered set, structures that were the cluster medoids selected by ICON, as well as lowest energy Rosetta and dDFIRE structures are highlighted in blue. (C) Structures produced after running the previous set through KoBaMIN are shown as cyan dots. (D) Structures submitted in blind prediction are shown as pink stars. The GDT_TS is shown here normalized on a zero to one basis. 378x279mm (300 x 300 DPI)

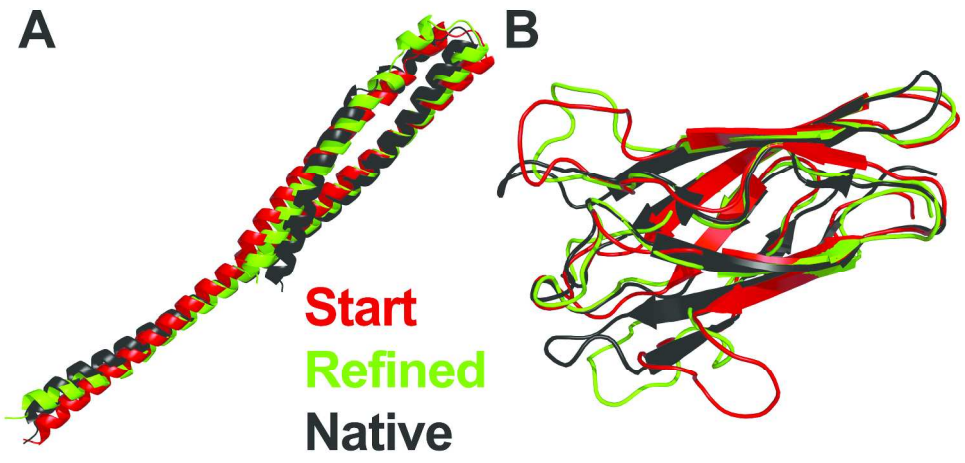

Figure 5: (A) Comparison of native structure (black) and refined structure (green) produced by wfFUIK branch for TR722. The refined structure using this protocol had a GDT_TS of 65.95, whereas the starting model (red) had a GDT_TS of 58.0. This structure is a dimer and adopts a coiled-coil fold. (B) Illustration of best Model 1 prediction produced by any method in CASP10 to improve the metric GDT_TS for target TR705. The WeFold method wfFUGT achieved this improvement, increasing the starting GDT_TS from 64.84 to 70.05. The loops in the upper-right region of the Figure, as well as in the bottom left were the regions where the most refinement occurred. TR705 adopts a β-sandwich fold. These examples highlight significant improvements in refinement for proteins containing only α-helices and β-sheets. This figure was created using PyMOL54.
259x127mm (300 x 300 DPI)

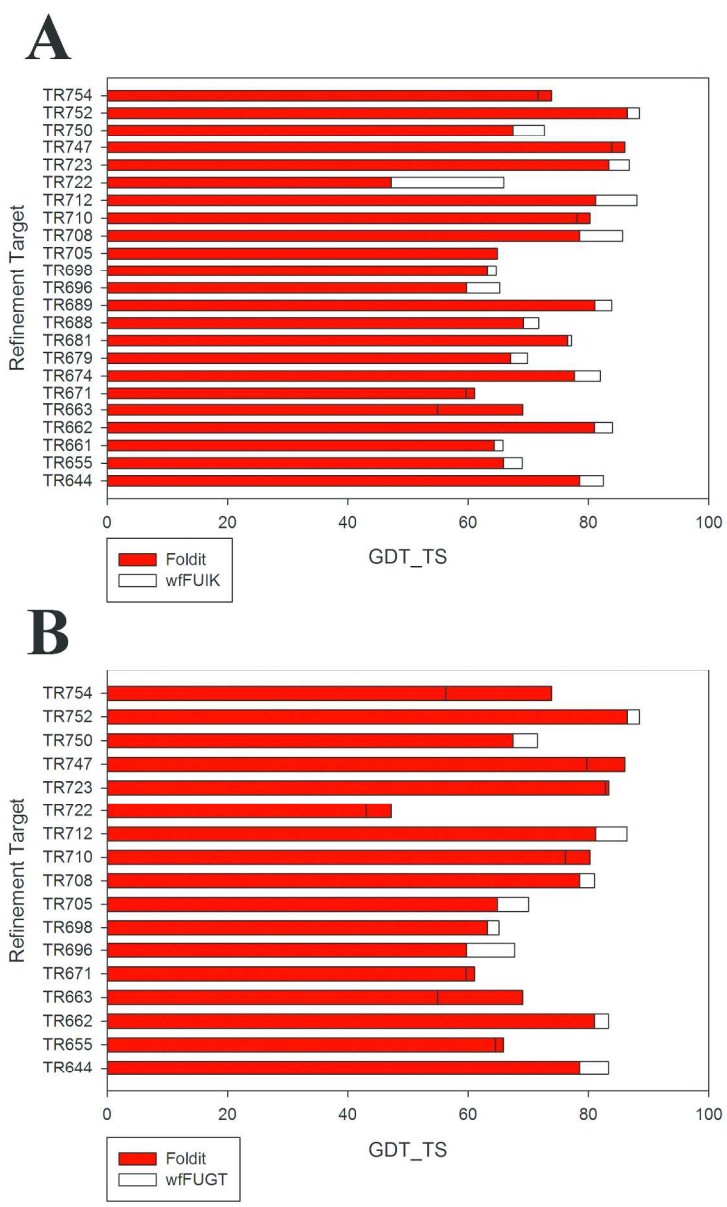

Figure 6: Comparison of improvements in Foldit models by WeFold methods. (A) In wfFUIK, 74% of structures were better refinements than the best structure submitted by FOLDIT. (B) Using wfFUGT, 53% of structures were better refinements than the best structure submitted by FOLDIT. The improvements are indicated by the differential bars in white from the base bars in red.
172x279mm (300 x 300 DPI)

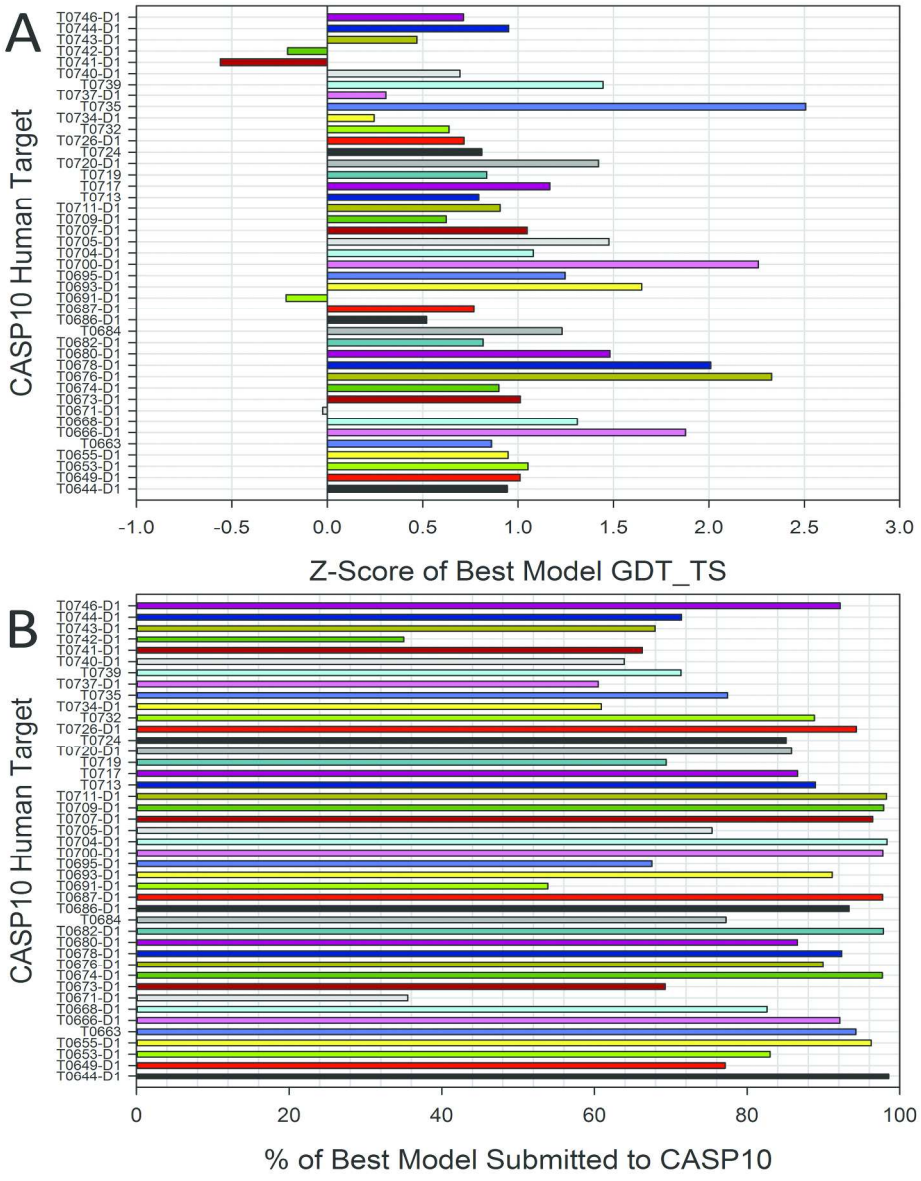

Figure 7: (A) Absolute and (B) relative performance of the WeFold branch on the 43 human targets attempted. The absolute performance is assessed based on the Z-score of the GDT_TS of the best model submitted by WeFold relative to all other predictions by all groups and methods for each target. The relative comparison is based on the ratio between the GDT_TS score of the best WeFold prediction and the best GDT_TS achieved by all groups for each target. In both cases, longer positive bars in the y-direction represent better performance.
228x295mm (300 x 300 DPI)

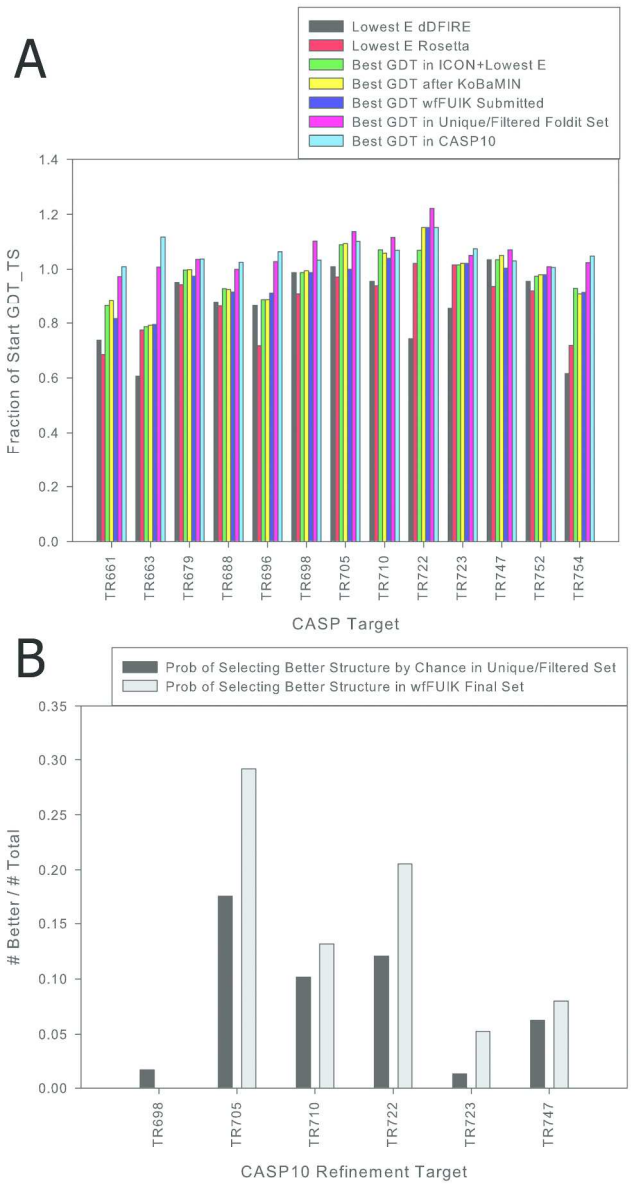

Figure 8: (A) Breakdown of the effect of each step in the wfFUIK pipeline in order to identify individual contributions to the pipeline, as well as areas needing further attention. The y-axis is normalized to show the ratio of the GDT_TS of the corresponding model to the GDT_TS of the starting model so that it can be compared across targets. The legend shows the lowest energy dDFIRE structure (black) in the Unique/Filtered set, lowest energy Rosetta structure from all Foldit conformations (red), best GDT_TS contained in the ICON+Lowest E Rosetta+Lowest E dDFIRE step of the pipeline (green), best GDT_TS contained after those structures are refined by KoBaMIN (yellow), best wfFUIK blind prediction of 5 submitted in CASP10 (dark blue), highest GDT_TS structure contained in the Foldit Unique/Filtered set (pink), and the best GDT_TS structure submitted to CASP10 by any team (light blue). (B) Enrichment of candidate conformers by wfFUIK compared to candidate conformers in the Unique/Filtered set. Shown are the probabilities of selecting a better structure than the start in the Unique/Filtered set compared to the enriched probability when choosing from the final set of wfFUIK models.
164x304mm (300 x 300 DPI)