BIOCHEMISTRY

Blind protein structure prediction using accelerated free-energy simulations

Alberto Perez,[1] Joseph A. Morrone,[1]* Emiliano Brini,[1] Justin L. MacCallum,[2] Ken A. Dill[1,3,4]†

We report a key proof of principle of a new acceleration method [Modeling Employing Limited Data (MELD)] for predicting protein structures by molecular dynamics simulation. It shows that such Boltzmann-satisfying techniques are now sufficiently fast and accurate to predict native protein structures in a limited test within the Critical Assessment of Structure Prediction (CASP) community-wide blind competition.

INTRODUCTION

Increasingly, our understanding of the structures and properties of protein molecules is based on computer modeling of two main types (*1*). In comparative modeling, protein structures are inferred from a database of other already-known protein structures. In free-energy–based modeling, protein structural populations and dynamics are modeled by computer simulations that satisfy thermodynamic principles, such as detailed balance, on the basis of known interatomic energies.

Although methods that are principally comparative have been the only practical way of inferring protein structures from their amino acid sequences thus far (*2*), free-energy methods have an important future because they do not require structural databases (thus, they could apply to membrane proteins, for example, where structures are few) or alignments to template proteins, and they go beyond native structures, to capture dynamical motions, folding routes, binding affinities, and conformational changes, all of which require a knowledge of the system's free-energy surface. The power of free-energy–based methods derives from their transferrable physical potentials and foundation in the Boltzmann Law.

A key test of any method for predicting protein structures is CASP (Critical Assessment of Structure Prediction), a blind assessment event of a community involving close to 200 research groups (*3*). Groups are given a protein sequence, with the structure blinded, asked to predict the three-dimensional structure in a fixed time frame (typically 3 weeks), and then evaluated when the structure is known. Free-energy–based methods at an atomistic level of detail have not been tested before in this venue because they have been computationally much too slow.

Here, we report the first successful test in CASP of an atomistic free-energy–based method for predicting native structures. We use a recent highly accelerated molecular simulation method called MELD (Modeling Employing Limited Data) (*4*). MELD is a Bayesian method that harnesses generic physical insights ("instructives") (*5*) within atomistic molecular dynamics (MD) force-field–based simulations. Here, we show that these free-energy–based simulations are sufficiently fast and accurate to solve some small simple structures within the competitive venue of CASP. We did not use templates or alignments. All our cluster predictions were generated using a laboratory-sized graphics processing unit (GPU) cluster (~100 GPUs).

MELD is unique in its ability to harness ambiguous instructions for accelerating MD while preserving Boltzmann statistics. For example, MELD can be directed to "make a good hydrophobic core" or "make secondary structures consistent with Web server predictions" (*5*). This information has previously been too vague, combinatory, and misdirective to aid free-energy simulations. Figure 1 (A and B) shows the challenge of constructing a hydrophobic core for protein G. Figure 1A shows the few true native contacts that the method must find (also shown in Fig. 1B as green lines). The red lines in Fig. 1B show the much larger number of possible hydrophobic contacts (the "haystack" that must be searched to find the native-state "needle").

MELD is a Bayesian way to search conformation-dependent subsets of the instructives during the MD simulation in a fast deterministic way (*4*, *5*). Instructives are encoded in flat-bottomed restraining potentials. These restraints direct the simulation toward regions that are consistent with the given information without biasing the sampling inside these regions.

MELD samples conformations using Hamiltonian and temperature replica-exchange molecular dynamics (HT-REMD) (*6*, *7*). HT-REMD ensures the satisfaction of detailed balance (that is, of Boltzmann statistics), and conformations are sampled from the equilibrium canonical ensemble. The stable states are not biased by the Bayesian springs because those springs contribute zero energy at their flat well bottoms. We used the Amber 12SB force field (*8*) with gb-neck2 implicit solvation (*9*) and cMAP corrections (*10*).

RESULTS AND DISCUSSION

As with any prediction method, MELD had successes and failures in CASP, but the following metrics give reason for optimism (see the Supplementary Materials for details). First, 4 of the 26 MELD predictions were the top-ranked by the CASP automatic server page. Second, for 12 of the 26 targets predicted, the root mean square deviation (RMSD) error relative to the native structures was less than 4 Å. These proteins range in size from 67 to 212 residues. Third, as discussed below, cluster populations help detect when MELD has converged.

Figure 2A shows the successfully predicted, lowest free-energy structures for three target proteins for which no information beyond sequence is given (denoted as T0xxx in CASP). For protein T0769 (97 residues), MELD's prediction was 2.8 Å from native (top half of the predictions). For protein T0773 (67 residues), the prediction was 1.4 Å from native (top quarter of all predictions). For protein T0816 (68 residues), MELD was 1.5 Å from native, which was the best prediction across the 121 different entries that attempted this structure. For six other targets in the T0 category, which were all longer than 100 residues, the MELD predictions were worse (see fig. S1). However, this was due to insufficient sampling time, inaccuracies of the force field or inappropriate instructives (some targets were not globular or monomeric, as we had assumed), and not flaws in the MELD framework per se (see the Supplementary Materials).

[1]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA. [2]Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada. [3]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, USA. [4]Department of Physics Astronomy, Stony Brook University, Stony Brook, NY 11794, USA.
*Present address: IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA.
†Corresponding author. Email: dill@laufercenter.org

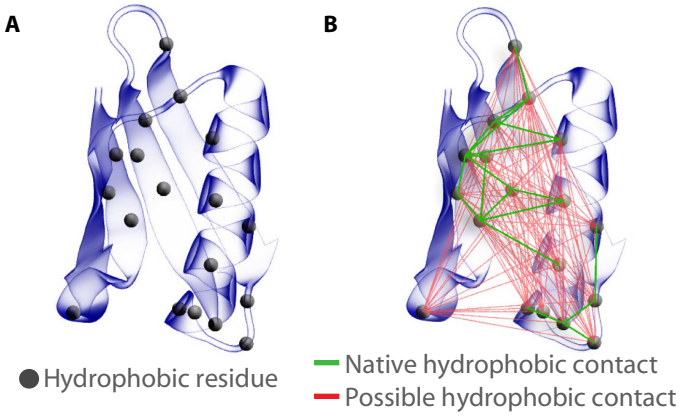

Fig. 1. Hydrophobic core instructive. (**A**) Hydrophobic residues in the native structure of protein G. (**B**) True native contacts that we seek (green lines) and the many possible hydrophobic contacts that must be considered in searching for the native structure (red lines).

CASP has other test categories (targets designated Tcxxx and Tsxxx) in which additional sparse or ambiguous information is provided (see the Supplementary Materials). MELD was tested on 17 protein targets in the Tc (sparse set of contacts present in the native state) and Ts [data simulating an unassigned nuclear magnetic resonance (NMR) spectra] categories. In this framework, MELD gives accurate structures for proteins up to 260 residues long (including three structures that were the best in CASP11; see figs. S2 and S3). MELD predictions resulted in a fifth overall ranking in these categories.

Figure 2B shows a CASP test that simulates having unassigned NMR data. As an example, Ts800 is a 212-residue protein provided with an unassigned NMR spectrum having 2251 NMR peaks, which lead to 29,210 possible assignments—approximately 27,000 of which are wrong. From these data, MELD predicts a structure that is within 3.1 Å Cα RMSD of the experimental data (fig. S3 shows all MELD results in this category). It shows that MELD is not confounded by such low signal-to-noise input data and may be useful for NMR refinements (*11*, *12*) and other sources of noisy data (*13*, *14*).

Figure 2C shows how added information can help MELD to recover from wrong instructives. All CASP entrants mispredicted target T0824 (Fig. 2C, left). Subsequently, added experimental data were given (Fig. 2C, right) for the Ts824 target. MELD yielded the lowest RMSD struc-

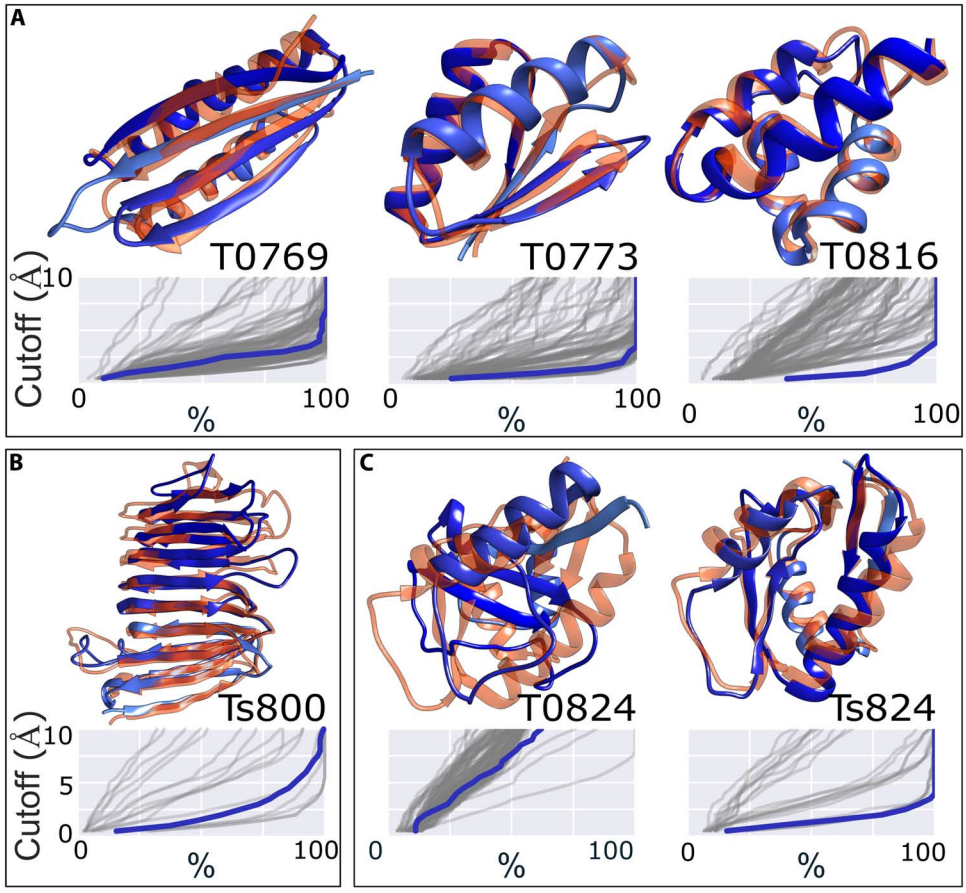

Fig. 2. MELD performance in CASP11. MELD predictions submitted prospectively are shown in blue. Experimental structures are shown in red. Below the structures are Hubbard plots (cumulative α-carbon accuracy), where each line assesses the quality of one submitted model of the target, and where the blue line is the MELD result. The best predictions are given by "elbows" that pass most closely through the bottom-right corner. (**A**) MELD predictions of three proteins given only sequence data. (**B**) MELD prediction of 212-residue target with simulated unassigned NMR data. (**C**) MELD prediction in the presence and absence of data.

ture according to CASP assessment. In retrospect, this protein was found to have a positively charged hole through its center to accommodate single-stranded DNA (see fig. S4 and the Supplementary Materials) (*15*). In this case, the missing DNA was vital to correctly sample the native structure—a limitation that was overcome with the provided data.

Historically, a major challenge for all prediction groups in CASP has been in determining which of their own five allowed submissions is the best one (*16*). That is, how can we predictively rank-order submissions? In principle, free-energy–based methods can compute the relative populations, or free energies, which provide a rational means to rank targets. If the underlying force field and sampling are adequate, then the highest population will predict the native structure. Figure 3 shows that for the MELD simulations that converge (that is, for chains shorter than ~100-mers), the most populated states correctly predict the native structures. In these cases, MELD can prospectively tell whether it has found a native structure.

CONCLUSIONS

Most proteins remain not yet foldable by any computational method. Still challenging are proteins that are large or multidomain or have prosthetic groups or reside in membranes, or are found in bound or complexed states. For those small proteins that are computationally tractable, the methods of choice are still the best comparative methods, because they are fast to compute and have been tested and proven extensively (*17*–*20*). Our tests here are on only a very limited number of the smallest and simplest proteins.

However, the main result here is the demonstration that, with MELD acceleration, MD force-field simulations with implicit solvent have now reached a speed and accuracy sufficient for evaluation within the accepted community-wide blind test venue of CASP. In many of the cases we tested, MELD performed very well. The promise of MELD is that it does not require templates or alignments, it predicts populations (which indicate convergence), it scales better to larger proteins than pure MD does (*5*), and it is applicable beyond native structure prediction to treat in principle conformational equilibria, kinetics, binding, and mechanisms, because it satisfies the Boltzmann Law.

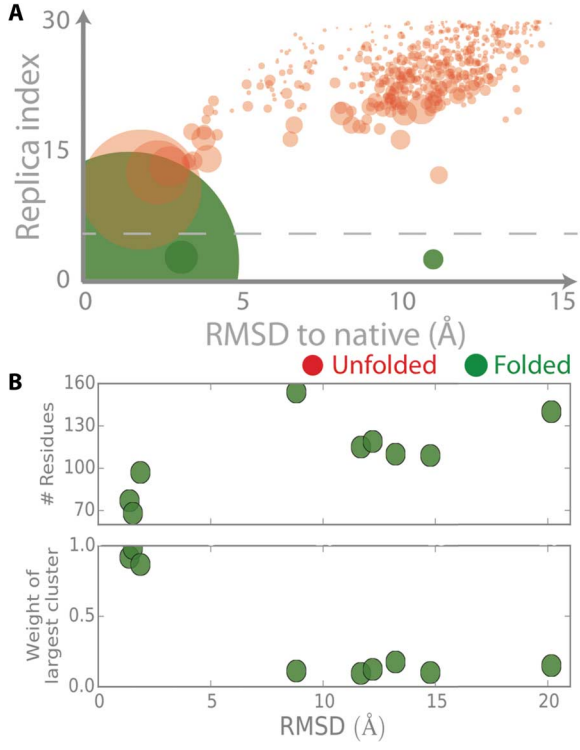

Fig. 3. Cluster populations are indicative of success. (**A**) Clustering of all conformations for protein T0816. (A) Conformational clustering occurs most strongly at low replica index (low temperature) and in near-native structures. Circle sizes represent cluster populations. The circle center coordinates denote the average RMSD and average replica index of the structures belonging to that cluster. Unfolded clusters tend to have low population. (**B**) Summary of top cluster populations for tested T0 targets. The populations converge to a single dominant native-like structure for the three protein targets we tried that were less than 100 residues long. Modeling longer chains did not converge to a single population.

MATERIALS AND METHODS

An introduction to MELD

MELD provides a readily extendable framework for physics-based molecular simulation guided by information (*4*). The method can be understood in terms of Bayes' theorem (*21*, *22*) as a way to find the posterior probability, $p(x|D)$, of finding a configuration, x, given the data, D

$$p(x|D) = \frac{p(D|x)p(x)}{p(D)}$$

The prior probability, $p(x)$, is the Boltzmann probability distribution yielded by sampling the system under the given force field. The likelihood of the data given the structure, $p(D|x)$, is proportional to $e^{-E_c(x)/kT}$, where T is the temperature and $E_c(x)$ is the overall constraint energy as formulated by MELD. The normalization factor $p(D)$ cancels out when considering ratios of the posterior probability of sampling different configurations. Information can be taken from sources such as experiments or coarse physical insights.

MELD went beyond other methods that encoded information using restraints in MD simulation by using a flexible algorithm in which a fraction of the number of possible constraints that was most compatible with the present point in configuration space could be enforced. This was achieved in a deterministic fashion by ranking possible constraints according to their energy and enforcing a given number of the lowest energy options. In the hierarchy of MELD, groups of constraints could be housed in a collection, and a given number of these groups could be enforced. Each restraint had a flat-bottomed functional form, so a sizable number of conformations were consistent with the information. This formalism yielded a well-defined Hamiltonian. Because the constraint energy was zero in regions of space consistent with information, the ratio of populations in different regions consistent with the MELD constraints was unbiased and reflected free-energy differences.

The constraints imposed high-energy barriers between regions of configuration space consistent with the information. Sampling was therefore facilitated by the use of REMD. In REMD, multiple copies of the system were simulated under different conditions, and exchanges between conditions were attempted at fixed time intervals. Temperature and the strength of the MELD restraints were altered along the replica ladder, increasing and decreasing from low to high replica index, respectively. Thus, moving down the temperature ladder led from relatively flat energy surfaces at high temperatures to relatively funneled regions at lower temperatures, to energy wells that were flat-bottomed in regions consistent with the informational constraints as discussed above.

Computational details

The driving engine in MELD was MD simulations running on graphics card technology (GPU), powered by OpenMM (*23*). The protein systems were represented in full atomic detail using the Amber ff12SB force field (*8*) with cMAP (*10*) correction and using gb-neck2 (*9*) as solvent model. All systems were started from extended conformations as produced by the AmberTools (*24*) leap sequence command. We used the MELD plug-in for OpenMM to create an HT-REMD. Swaps between conditions were attempted every 50 ps in our simulations. Each individual replica as it moved up and down the replica exchange ladder was called a "walker." The Hamiltonian was changed by adding information into the system—either given to us or from coarse physical insights (CPIs) (*5*)—and this perturbation to the force field Hamiltonian was scaled down as individual walkers moved to higher replicas.

We used hydrogen mass repartitioning (*25*) and rigid bonds to allow a time step of 3.5 fs. Langevin dynamics with a 1-ps^{-1} coupling constant was used. We set simulations to be 500 ns long. Time constraints during CASP precluded us from achieving this in many cases.

T0 targets.

We set up 30 replicas for each system, ran for 500 ns each, and clustered at the end of the simulations. In the case of the first target, T0759, we attempted to use contacts from homology modeling as additional heuristic. We were not satisfied with our knowledge of homology tools, so we did not use this for the remaining targets. Target T0759 was repeated after CASP with the same protocol as the rest of the proteins for comparison purposes. Here are the CPI constraints we used:

(1) Secondary structures. We ran a local psipred on the sequence and enforced secondary structure at an accuracy level between 70 and 85%. We started CASP taking 85% as a good estimate (for targets T0759, T0769, and T0773) and later used 70 to 75% for the rest of the targets, relying more heavily on the force field. This restraint type was not scaled down in higher replicas.

(2) Hydrophobic pairing. We created a restraint between each pair of C_β in hydrophobic residues (alanine, isoleucine, leucine, methionine, phenylalanine, proline, tryptophan, and valine), excluding those that are less than seven residues apart in sequence. We then counted how many hydrophobic residues were in the chain (N_h). During CASP, we enforced 2.7 N_h hydrophobic contacts. Our post-CASP analysis showed that 1.3 would be a better option.

(3) Strand pairing. We based this on secondary structure predictions. We never applied restraints between residues that belong to the same strand. We added restraints among all other pairings of residues belonging to different strands. For each pair of residues, we created two restraints in one group (see Materials and Methods); one corresponded to the N_i–O_j pairing, and the other was the N_j–O_i hydrogen bond pairing between residues i and j. We allowed that only one of those two hydrogen bonds needed to be satisfied. All of the groups were added onto a collection. The number of active restraints was set by counting the number of residues predicted (psipred) to be extended (N_E) and multiplying times 0.65.

(4) Confinement restraint. This constraint enforced that the protein be relatively compact. It increased the probability of forming high–contact order contacts. The radius of the protein's occupancy sphere was set to be $r(\text{nm}) = 16.9 \times \log(N_{res} - 15.8)/28$. This restraint type was not scaled down at higher replica index.

At the end of the simulations, we clustered the last 250 ns of the lowest five temperature replicas. We used a hierarchical agglomerative clustering (*26*, *27*) on the basis of C_α/C_β RMSD over residues with predicted secondary structure (psipred). We did this using a linkage algorithm with an ε of 2 for the agglomeration, as defined in cpptraj (*28*). For the clustering, we sieved every 10 frames, resulting in 2500 frames. We then assigned the 25,000 frames to the corresponding clusters. We looked at the structure closer to the centroid of each of the 10 top clusters, submitting the top 5 to CASP (except in cases in which the RMSD between two centroids was too small; and then, to favor diversity, we replaced one of them with the cluster having the next lowest population or a structure minimized toward the cluster average).

Ts targets.

Ts targets constituted a CASP experiment in which constraints were provided beyond the amino acid sequence to represent what would be obtainable from an unassigned NMR spectrum. Participants were given the sequence data and about 1 week (sometimes less) to solve the structures. It was known how many NMR peaks there were, and several possible atomic contacts were given for each peak (see Table 1 for NMR

Table 1. Summary of NMR-like data provided for Ts targets.

Target	Length	Oligomeric state	NMR peaks	Possible contacts	Reduced peaks	Reduced restraints	Days	RMSD
Ts761	237	2	3867	29,210	828	5183	6	8.6
Ts763	131	2	2537	11,516	693	2619	5	3.0
Ts785	112	3	1072	4,009	210	510	4	5.1
Ts800	212	1	2251	19,759	489	3313	6	3.2
Ts802	118	3	900	2,014	223	475	7	2.0
Ts810	113	1	1174	3,627	129	317	4	5.3
Ts818	134	1	873	2,228	149	426	14	4.5
Ts824	110	?	867	1,600	182	365	12	1.8
Ts826	201	1	2531	23,959	152	816	8	11.9
Ts832	209	1	2146	17,630	274	2346	5	4.0

peaks and possible contacts and http://predictioncenter.org/download_area/CASP11/extra_experiments/README_Ts for more details about the origin of the data).

We first reduced the number of possible contacts by enforcing distances between heavy atoms rather than hydrogens in the cases of degenerate hydrogens (for example, methyl groups in alanine, valine, leucine, or isoleucine). We also excluded any peak that could be explained by a "contact order" of 4 or less. Contact order refers to the number of residues along the sequence between the two residues of interest. This information significantly reduced the number of restraints in the simulation (see Table 1 for reduced peaks and reduced restraints). Along with the possible contacts, the data provided the NMR distance between atoms—when tracing the hydrogen back to the heavy atom to which it was attached. We added 1 Å to this distance. The table also shows the time allowed for solution and the resulting RMSD of our predictions compared to the experimental structure.

Within the MELD structure, each peak was expressed as a "group" of restraints, where we required that only one of the possible interpretations of the restraints needed to be satisfied. All groups were part of the same collection, which had an accuracy of 100% (meaning that every peak had to be satisfied by one restraint). This kind of data coupled with large protein sizes produced large bottlenecks in the REMD ladder when using a small number of replicas. Thus, we used more replicas, as many as 106. Because of the time constraints, we were never able to sample sufficiently for convergence. To choose structures, we selected the subsets of restraints from the original CASP data that had no ambiguity and checked how many were violated in each structure in our ensemble. We submitted to CASP structures with the smallest number of restraint violations. We later improved on this by looking at restraint energies rather than restraints violated. Target Ts826 turned out to be a membrane protein and hence was poorly predicted here because it violated our assumption that it was water-soluble.

Tc targets.

In this category, we were given the sequence and the top $L/5$ correctly predicted contacts from contact prediction groups—where L is the protein length in amino acids. We took the different predictions from different groups and used all nonrepetitive instances in our simulations. In this case, there was no ambiguity. As in the previous case, we used agreement with the original data to select structures from our ensemble.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/11/e1601274/DC1

Supplementary Text

table S1. MELD energies of CASP structures.

fig. S1. Ab initio folding.

fig. S2. Ab initio folding guided by contact predictions.

fig. S3. Ab initio folding guided by an unassigned NMR-like data set.

fig. S4. Native structure for target T0824.

fig. S5. Comparison of the average performances of different groups over the targets we tackled.

fig. S6. Simple two-state folding mechanism.

fig. S7. Complex folding mechanism.

fig. S8. Identification of multiple misfolded intermediates for T0769.

fig. S9. Prediction of a mirror topology.

fig. S10. When the instructives are sufficiently wrong, MELD will not find the correct native structure.

References (*29*–*41*)

REFERENCES AND NOTES

1. D. Baker, A. Sali, Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
2. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
3. J. Moult, A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **15**, 285–289 (2005).
4. J. L. MacCallum, A. Perez, K. A. Dill, Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6985–6990 (2015).
5. A. Perez, J. L. MacCallum, K. A. Dill, Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11846–11851 (2015).
6. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
7. H. Fukunishi, O. Watanabe, S. Takada, On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **116**, 9058–9067 (2002).
8. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
9. H. Nguyen, D. R. Roe, C. Simmerling, Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9**, 2020–2034 (2013).
10. A. Perez, J. L. MacCallum, E. Brini, C. Simmerling, K. A. Dill, Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations. *J. Chem. Theory Comput.* **11**, 4770–4779 (2015).
11. J. Meiler, D. Baker, Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15404–15409 (2003).
12. S. Raman, Y. J. Huang, B. Mao, P. Rossi, J. M. Aramini, G. Liu, G. T. Montelione, D. Baker, Accurate automated protein NMR structure determination using unassigned NOESY data. *J. Am. Chem. Soc.* **132**, 202–207 (2010).
13. D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE* **6**, e28766 (2011).
14. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
15. A. Kryshtafovych, J. Moult, A. Baslé, A. Burgin, T. K. Craig, R. A. Edwards, D. Fass, M. D. Hartmann, M. Korycinski, R. J. Lewis, D. Lorimer, A. N. Lupas, J. Newman, T. S. Peat, K. H. Piepenbrink, J. Prahlad, M. J. van Raaij, F. Rohwer, A. M. Segall, V. Seguritan, E. J. Sundberg, A. K. Singh, M. A. Wilson, T. Schwede, Some of the most interesting CASP11 targets through the eyes of their authors. *Proteins*, **84**, 34–50 (2016).
16. J. L. MacCallum, A. Pérez, M. J. Schnieders, L. Hua, M. P. Jacobson, K. A. Dill, Assessment of protein structure refinement in CASP9. *Proteins* **79** (suppl. 10), 74–90 (2011).
17. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **37** (suppl. 3), 171–176 (1999).
18. A. Roy, A. Kucukural, Y. Zhang, I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
19. J. Lee, H. A. Scheraga, S. Rackovsky, New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J. Comput. Chem.* **18**, 1222–1232 (1997).
20. A. Šali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
21. W. Rieping, M. Habeck, M. Nilges, Inferential structure determination. *Science* **309**, 303–306 (2005).
22. V. A. Voelz, G. Zhou, Bayesian inference of conformational state populations from computational models and sparse experimental observables. *J. Comput. Chem.* **35**, 2215–2224 (2014).
23. P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, V. S. Pande, OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* **9**, 461–469 (2013).
24. D. A. Case, V. Babin, J. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, P. A. Kollman, *Amber 14* (University of California, 2014).
25. C. W. Hopkins, S. Le Grand, R. C. Walker, A. E. Roitberg, Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
26. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.* **38**, 236–240 (1999).
27. J. Shao, S. W. Tanner, N. Thompson, T. E. Cheatham III, Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **3**, 2312–2334 (2007).
28. D. R. Roe, T. E. Cheatham III, PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

29. J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—Round x. *Proteins* **82**, 1–6 (2013).
30. J. Moult, J. T. Pedersen, R. Judson, K. Fidelis A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–iv (1995).
31. T. J. P. Hubbard, RMS/Coverage graphs: A qualitative method for comparing three-dimensional protein structure predictions. *Proteins* **37**, 15–21 (1999).
32. A. Zemla, LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
33. A. Perez, Z. Yang, I. Bahar, K. A. Dill, J. L. MacCallum, FlexE: Using elastic network models to compare models of protein structure. *J. Chem. Theory Comput.* **8**, 3985–3991 (2012).
34. I. W. Davis, L. W. Murray, J. S. Richardson, D. C. Richardson, MOLPROBITY: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32** (suppl. 2), W615–W619 (2004).
35. D. T. Capraro, M. Roy, J. N. Onuchic, P. A. Jennings, Backtracking on the folding landscape of the β-trefoil protein interleukin-1β? *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14844–14848 (2008).
36. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
37. M. Korycinski, R. Albrecht, A. Ursinus, M. D. Hartmann, M. Coles, J. Martin, S. Dunin-Horkawicz, A. N. Lupas, STAC—A new domain associated with transmembrane solute transport and two-component signal transduction systems. *J. Mol. Biol.* **427**, 3327–3339 (2015).
38. A. N. Naganathan, V. Muñoz, Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127**, 480–481 (2005).
39. A. Perez, J. A. Morrone, C. Simmerling, K. A. Dill, Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol.* **36**, 25–31 (2016).
40. V. Mirjalili, M. Feig, Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. *J. Chem. Theory Comput.* **9**, 1294–1303 (2013).
41. A. Perez, A. Roy, K. Kasavajhala, A. Wagaman, K. A. Dill, J. L. MacCallum, Extracting representative structures from protein conformational ensembles. *Proteins* **82**, 2671–2680 (2014).

Acknowledgments: We want to thank J. Wagoner for his help during the CASP event. We are grateful for support from the Laufer Center and NIH grants GM107104 and GM090205. We thank Blue Waters and NSF through ACI1514873 for computational time used to complete this manuscript after the CASP11 event. **Author contributions:** A.P., J.L.M., and K.A.D. designed the research. A.P., E.B., and J.A.M. carried out the experiments and collected and analyzed the data. A.P., E.B., J.A.M., J.L.M., and K.A.D wrote and revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. CASP results are available through the CASP website www.predictioncenter.org/casp11/index.cgi; our group is denoted as Laufer, group 428. Additional data related to this paper may be requested from the authors.

Submitted 6 June 2016
Accepted 7 October 2016
Published 11 November 2016
10.1126/sciadv.1601274

Citation: A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum, K. A. Dill, Blind protein structure prediction using accelerated free-energy simulations. *Sci. Adv.* **2**, e1601274 (2016).

Blind protein structure prediction using accelerated free-energy simulations

Alberto Perez, Joseph A. Morrone, Emiliano Brini, Justin L. MacCallum and Ken A. Dill

Sci Adv **2** (11), e1601274.
DOI: 10.1126/sciadv.1601274