Title

Accelerating Protein Folding Molecular Dynamics Using Inter-Residue Distances from Machine Learning Servers

Permalink

https://escholarship.org/uc/item/6378f4w3

Journal

Journal of Chemical Theory and Computation, 18(3)

ISSN

1549-9618

Authors

Nassar, Roy
Brini, Emiliano
Parui, Sridip
et al.

Publication Date

2022-03-08

DOI

10.1021/acs.jctc.1c00916

Peer reviewed

JCTC Journal of Chemical Theory and Computation

Accelerating Protein Folding Molecular Dynamics Using Inter-Residue Distances from Machine Learning Servers

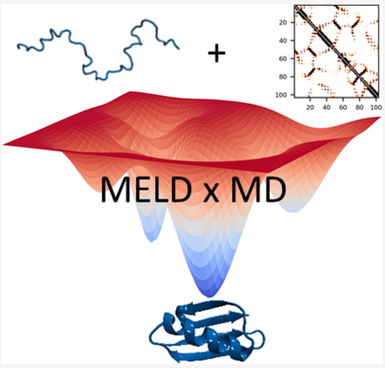
Roy Nassar, Emiliano Brini, Sridip Parui, Cong Liu, Gregory L. Dignon, and Ken A. Dill*

Cite This: J. Chem. Theory Comput. 2022, 18, 1929−1935

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Recently, predicting the native structures of proteins has become possible using computational molecular physics (CMP)—physics-based force fields sampled with proper statistics—but only for small proteins. Algorithms with better scaling are needed. We describe ML x MELD x MD, a molecular dynamics (MD) method that inputs residue contacts derived from machine learning (ML) servers into MELD, a Bayesian accelerator that preserves detailed-balance statistics. Contacts are derived from trRosetta-predicted distance histograms (distograms) and are integrated into MELD's atomistic MD as spatial restraints through parametrized potential functions. In the CASP14 blind prediction event, ML x MELD x MD predicted 13 native structures to better than 4.5 Å error, including for 10 proteins in the range of 115−250 amino acids long. Also, the scaling of simulation time vs protein length is much better than unguided MD: $t_{\mathrm{sim}} \sim e^{0.023N}$ for ML x MELD x MD vs $t_{\mathrm{sim}} \sim e^{0.168N}$ for MD alone. This shows how machine learning information can be leveraged to advance physics-based modeling of proteins.



INTRODUCTION

A key approach for studying the physical equilibria and dynamical actions of protein molecules is computational molecular physics (CMP). CMP captures the physics and dynamics through semiclassical force fields that are sampled by molecular dynamics (MD) or Monte Carlo (MC) methods.[1,2] It is a major tool for protein storytelling. Over the years, CMP modeling has advanced at roughly Moore's Law rates, keeping pace with advances in computer hardware.[3] For example, within the past decade, CMP methods have achieved excellent results in folding proteins from their unfolded states[4−6] and computing binding affinities to small-molecule ligands.[7]

But, huge challenges remain.[8] Protein conformational searching and sampling times increase exponentially with the size of the molecule.[9] To date, the proteins folded by CMP from unfolded states are mostly shorter than 100-mers.[5,6,10−14] Hence, new computer hardware alone, even at Moore's Law rates, gives only a few amino acid length gains each generation. To date, CMP folding simulations using atomic level resolution have been limited to relatively small proteins and short time scales. Reaching the sizes of biologically relevant proteins (at least 100−300 amino acids long) and their larger actions will require new search strategies that have much better scaling.

In principle, conformational search efficiencies can be increased by using some targeted knowledge of the state(s) of interest. However, to ensure correct physical free energies, CMP modeling must give correctly detailed balanced (DB) conformational populations. Incorporating informational restraints, such as springs, in any simple way, breaks DB and will not give proper physical populations. This is because sampling a physical force field that has added nonphysical potentials (informational springs) will not correctly sample the *underlying physical force field*. The MELD acceleration method that we use here,[10,15] "melds in" external information into replica-exchange MD[16] using a Bayesian scheme. The MELD method does preserve proper DB because at convergence all the informational springs become unstretched and contribute approximately zero energy to the physical force field.

Here, we describe ML x MELD x MD, a novel extension to MELD, formulated to utilize residue distance information predicted by a machine learning (ML) web server[17] previously trained on native structure databases. Specifically, we first derive spatial contacts from the ML distograms and use a set of known structures to parametrize restraining potentials to augment the force field during MD simulations. We observe a significant acceleration in the conformational search for single domain proteins. We show that the designed algorithm populates native states in the lowest free energy basin of its folding landscape. We finally validate all the components together—the MD potential function, MELD accelerator, and ML data focusing—by predicting protein native structures of novel sequences in the CASP14 blind prediction competition, which is currently the highest community-wide standard of

Received: September 13, 2021
Published: February 8, 2022

https://doi.org/10.1021/acs.jctc.1c00916

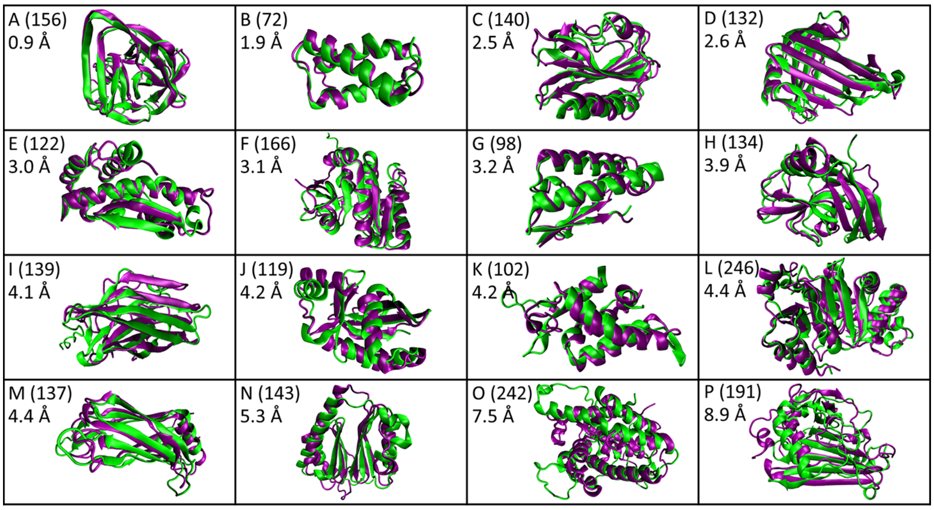


Figure 1. ML x MELD x MD gives good native structures on some targets in CASP14. Panels A−P show the Top1 ensemble representative (purple) superimposed onto the true native structure (green). Target length N (in amino acids) is given in parentheses. The backbone rmsd value from the native structure is also given for each protein. Target IDs assigned by CASP are A (T1034-D1), B (T1046s1-D1), C (T1046s2-D1), D (T1074-D1), E (T1055-D1), F (T1045s2-D1), G (T1065s2-D1), H (T1049-D1), I (T1078-D1), J (T1065s1-D1), K (T1035-D1), L (T1057-D1), M (T1060s3-D1), N (T1054-D1), O (T1041-D1), and P (T1090-D1).

validation for native structure modeling, and has largely been beyond the capabilities of CMP methods in the past.

RESULTS

Here, we show tests of ML x MELD x MD on some CASP 14 targets. On the one hand, protein structure prediction *per se* is not a principal objective of MD modeling. MD is a very general tool for broad aspects of protein storytelling—conformational populations, free energies and fluctuations, interpreting experiments, determining dynamical sequences of actions, computing binding affinities, rates and allostery, and more. So, why test the present method in CASP? We do so because CASP is currently the most challenging, detailed, comprehensive, blind, and comparative venue for testing one of the most important properties of proteins, namely, its average static native structure. If an MD model does not correctly encode a native structure, how can we trust it to accurately represent other states? But, finding native states in the short time frame of CASP events has previously been outside the computational reach of most MD simulations. Here, we use the CASP event as a way to test whether MD, accelerated by MELD and ML, can come close to finding native states by physical force field potentials. The larger aspiration, for which the present work is a first step, is to see if advanced MD tools can harness the vast and granular structural knowledge in the PDB to accelerate physics-based protein modeling. We show below that this method often finds fairly accurate native structures, and it does so with much better scaling as a function of chain length than MD alone.

MELD x MD Often Gives Good Native Structures. The critical assessment of protein structure prediction[18] (CASP) event is a time-limited communal blind comparative test of protein structure prediction methods. Figures 1 and 2 show that ML x MELD x MD was capable of determining native states of targets presented in the CASP14 event. The results show that when MELD's assumptions are met (hydrophobic core, good secondary structures, accurate enough ML data) the most populated macrostate in the simulation indeed corresponds to the native basin (RMSD < 4.5 Å) for most targets (13 out of 16). These include targets with α, β, and mixed $\alpha\beta$ topologies. The three largest proteins are longer than 190 amino acids (aa) (panels O, L, and P in Figures 1 and 2). Our simulations populated the native basin for T1057 (246 aa), but the most favored (Top1) models for the other two targets (T1090 191 aa and T1041 242 aa) were less accurate according to the backbone RMSD metric. Nonetheless, the overall topology of the Top1 model for all three largest targets was in fact correct with TM-score > 0.5 as calculated by the CASP14 organizers (Figure S6). The template modeling score[19] (TM-score) is a global metric that compares two structures and is less sensitive to local deviations and protein size than the widely used RMSD value. A TM-score ranges from 0 denoting very different folds to 1 denoting identical structures, with a score greater than 0.5 indicating that the two structures have a similar fold.[20] The size range of proteins sought after here is important because functionally relevant protein domains in prokaryotes and eukaryotes are typically

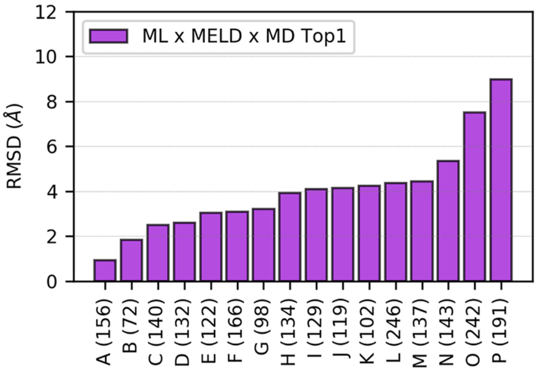


Figure 2. MELD x MD obtains accurate structures for 13 out of 16 targets in CASP14. Most targets are much longer than 100 amino acids. Protein lengths (in number of amino acids) are shown in parentheses. Targets are A (T1034-D1), B (T1046s1-D1), C (T1046s2-D1), D (T1074-D1), E (T1055-D1), F (T1045s2-D1), G (T1065s2-D1), H (T1049-D1), I (T1078-D1), J (T1065s1-D1), K (T1035-D1), L (T1057-D1), M (T1060s3-D1), N (T1054-D1), O (T1041-D1), and P (T1090-D1).

100−150 amino acids long.[21,22] Overall, the results here show that MELD x MD can leverage high quality data from elaborate ML servers to model atomistic native states of proteins.

Combining ML with MELD Gives Good Acceleration and Scaling with Protein Size. Conformational searching and sampling of large proteins starting from the completely unfolded state is challenging and grows exponentially with protein size. Figure 3 shows that MELD x MD uses the ML data effectively to generate native states from the extended linear chain of proteins up to 150 aa. This indicates that ML x MELD x MD can converge quickly onto native basins that are compatible with the ML input and with the force field. The efficiency in folding with ML x MELD x MD stems from its rapid assembly of helices and strands, followed by the protein core (Figures S10 and S11). The simulation time as a function of protein size N for conventional MD is estimated from Figure 3 to scale as $t \approx 0.5e^{0.168N}$, whereas the MELD-accelerated simulation time goes as $t \approx 8e^{0.023N}$ (see the Supporting Information for details). This indicates that the present method accelerates native finding for 100-mers by about 5 orders of magnitude and 200-mers by about 11 orders of magnitude. Most importantly, because native finding with CMP requires exponential searching, it means that an approximately 100-mer protein was a fairly hard limit for brute-force atomistic MD, but that for the present method the limit on chain length is no longer so hard. Our folding simulations only span the few microsecond time range (Figure 3 and Table S1), which is quite attainable on current lab-scale GPU clusters.

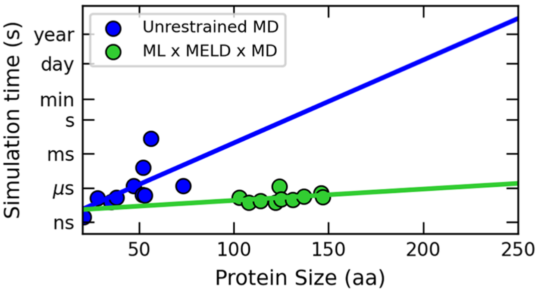


Figure 3. ML x MELD x MD efficiently scales up physics-based folding simulations. Pre-CASP14 simulations (green points) handle 10 large single-domain proteins of sizes 100−150 aa while maintaining practical μs simulation time. These simulations start from the completely unfolded linear chain and give orders of magnitude acceleration over unguided MD (blue points).

Figure 3 shows a comparison of MELD sampling vs vanilla MD with the same atomic resolution and solvent model. But, another form of speedup is coarse-grained modeling.[23]

ML x MELD x MD Predetermines Its Success by Lowest Free Energies and Highest Populations. The CASP event allows a team to submit five predicted structures for a protein target. A perpetual challenge has been to know in advance which of a team's five submissions is the best one, i.e., which predicted model is most likely to be the true native structure. For prediction methods that do give confidence estimates, they are statistical, based on estimates of past successes. However, much better, in principle, would be a physical potential that predicts the native structure as being the conformation of lowest free energy. Here, we show that the conformations of lowest free energy in the force field correctly predict the experimental native structure. We compare these native (experimental) snapshots with a centroid representative of the Top1 macrostate (lowest free-energy ensemble) from our MELD simulation.

The RMSD values in Figure 2 and Figure 4B show that in such a blind setting (as in CASP), where no experimental data are known about the protein structure, the Top1 (most populated) macrostate of ML x MELD x MD was indeed a native state for 13 out of 16 targets. A similar result was observed in proofs of concept simulations (Pre-CASP14) for nine out of 10 proteins (Figure 4A) where RMSD < 4 Å for the Top1 state was obtained. When the population of its Top1 macrostate is high (>50%), ML x MELD x MD is more certain that its conformational sampling using the input data has converged onto a deep energy minimum corresponding to the native basin. Interestingly, the simulation can still find the native basin even if the observed populations are less dominant (Figure 4B), but MELD x MD cannot be as assertive about its Top1 success in these cases. The present method also gives conformational distributions. The RMSD distribution of the microstates constituting the Top1 macrostate for each protein is shown in Figure S4 of the Supporting Information.

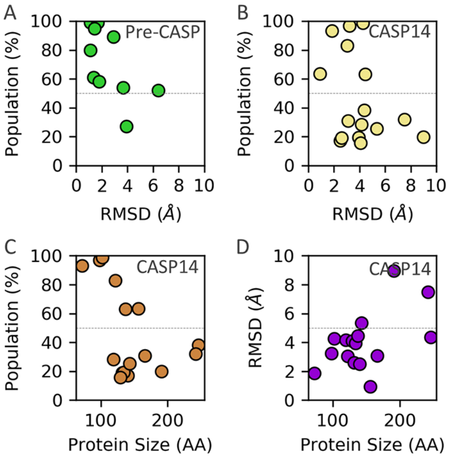


Figure 4. MELD x MD orders macrostates of proteins by free energy as inferred through their Boltzmann populations. (A, B) Boltzmann populations of the most favored macrostate for each protein in the training (Pre-CASP14) and the CASP14 simulations. The most populated macrostate is the lowest free energy minimum on the landscape and accurately resembles the experimental structure (RMSD ≤ 4.0 Å) in most cases. (C, D) Boltzmann populations and RMSD as a function of protein size for the CASP14 targets.

■ COMMENTS, LIMITATIONS, AND CAVEATS

First, in the Supporting Information, we look at the source of the computational efficiency of ML x MELD x MD folding. Summarizing, we find that it quickly achieves concerted assembly of helices and strands, followed by beta sheets and protein cores (Figure S10). The native state is often reached early and remains stable until the end of the simulation (Figure S11). This indicates that when secondary structures pack correctly into a protein core, the high quality ML restraints can stabilize the thermodynamically favorable native at the physiological temperatures of the ensemble. Figure S12 shows misfolded proteins trapped in alternative conformations, mostly due to incorrect tertiary packing even when most strands and helices individually assume a native-like arrangement, suggesting that when the conformations derange from the native state it is likely due to misfolded protein cores that get stabilized by a certain subset of the data.

Second, we note that ML x MELD x MD is not the best way to predict the single optimal native structure of a protein. It is computationally expensive, and several deep-learning methods are generally more accurate across a larger set of proteins.[24,25] Also, there remain inaccuracies in implicit solvent and force field models.[3,26] Convergence onto the native state becomes more difficult for proteins in the range of 200 aa or longer (Figure 4C, D). Further work is required for targeting multidomain proteins.

MELD x MD relies on distance predictions from the ML algorithm, so successful predictions usually depend on the amount and precision of the input (contacts) information. Intuitively, the number of contact hubs (clusters) is a better measure of the quantity of information provided. We investigate a possible correlation between the folding accuracy and the number of medium- and long-range contact hubs per amino acid. This analysis indeed reveals that protein domains with more available nonlocal data from machine learning are folded to better structures than those with less such data (Figure 5). Interestingly, some proteins with few contact hubs are still folded to good native structures, indicating that MELD x MD can sometimes compensate for insufficient ML information through reliance on its force field, secondary structure predictions, and hydrophobic interactions.

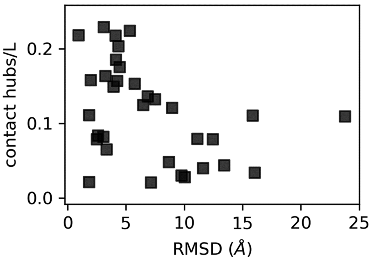


Figure 5. Better models are obtained for proteins with a larger set of predicted nonlocal contact hubs. The number shown on the plot is that of medium- and long-range contact hubs for the 33 CASP14 targets attempted by MELD x MD. The number of hubs was normalized by the domain sequence length for comparison.

Different algorithms used as source of ML input restraints require different setpoints of confidence intervals for usage in MELD; we study these differences in Figure S13 and Table S3 and show their effect on folding performance.

CONCLUSIONS AND PERSPECTIVE

Ultimately, the telling of protein stories requires physics—the distribution of conformational populations, dynamics, and motions and binding affinities of a protein's various states. Computational molecular physical modeling is increasingly successful at this but is limited today to small proteins. New methods are needed that scale better for searching and sampling physical force fields to larger proteins and larger dynamics. Here, we use machine-learning (ML) native-structure contact predictions to accelerate replica-exchange molecular dynamics through a Bayesian method called MELD that preserves the proper sampling physics.

Here are our main results: (1) ML x MELD x MD correctly predicted 13 native structures to better than 4.5 Å accuracy in CASP14; ten of those (115−250 amino acids long) are considerably larger than what atomistic MD folding has done before. (2) The simulation time scaling exponent in going from unfolded to folded states with protein size ($t \sim e^{0.023N}$) is much better (smaller) than MD alone ($t \sim e^{0.168N}$), giving an 8 order of magnitude speed advantage for 150 amino acid proteins, for example. (3) In 13 of the CASP14 proteins, convergence of the force field predicts the correct basin of the lowest free energy state in this blind test. Also, while our tests here focus only on native structure predictions—since that is where the most granular and definitive data are—nevertheless, the real potential is in having a CMP modeling method with better scaling for the ultimate purpose of protein storytelling.

METHODS

MELD (modeling employing limited data) is an enhanced sampling technique for reducing the vast conformational spaces of proteins.[10,15] It aims to accelerate the MD sampling to quickly produce low energy states on the folding landscape. Here, low energy states are those favored by some input data and by the MD force field. MELD uses a Bayesian framework where it combines two probability distributions: a prior, corresponding to the unbiased distribution of states given by the MD force field energy, and a likelihood, corresponding to states compatible with the provided external information. The external data are converted into restraints, which penalize the energy of states that do not fit the guiding external information. During the MELD x MD simulation, many possible mixes of the restraints are thoroughly searched, and the states are ultimately ranked based on their relative populations which, in the limit of converged populations, is a good proxy for their relative free energies.[27]

I. Conversion of Data into MELD Restraints. MELD can utilize various different sources of information as input to its Bayesian inference engine, including that which is noisy, imperfect, ambiguous, or combinatorically challenging (details in the Supporting Information and in MacCallum et al.[15] and Perez et al.[10]). In the present work, we utilize the output from the public ML server trRosetta,[17] which predicts distances between the amino acids, as input to MELD x MD. Such information has given good insights about static native structures[28,29] but not yet applied to guide detailed-balance preserving physical simulations, as far as we know. Here, we calculate the (cumulative) probability of each pair of noncovalently interacting residues to be within an interaction distance $d_{ij} \leq 8.0$ Å directly from the trRosetta predicted distogram matrix. Pairs with probability $p \geq 0.5$ (high probability) are kept as possible contacts to be converted into distance restraints between C_β−C_β atoms. We configure this new class of ML distance restraints in MELD by parametrizing (flat-bottom) potentials using a training set of known structures from the PDB. The potentials consist of a spatial region where the contacts are satisfied (zero restraint energy) and other regions with positive energy penalty to be added to the force field when the contacts are not satisfied.

A distinguishing feature of MELD, compared to other methods that use restraints to model proteins, is its ability to intelligently enforce only a subset of the entire input data in order to account for imperfections and noise from the data source. Briefly, at each cycle during the MD, MELD computes the energy of each restraint based on the current structure, and ranks all the input restraints according to their energies. It then automatically enforces the lowest-energy restraints up until a specified number to be set *a priori* by the user, which corresponds to the user's trust level of the data source. Here, the activation fraction of the ML-derived restraints was tuned to 80% during the training stage by empirically checking the accuracy of the derived restraints on a set of 24 globular proteins with known structures in the PDB. More details on how the MELD potentials were parametrized for this type of data are provided in the Supporting Information methods.

Independently, restraints corresponding to secondary structure elements predicted by PSIPRED[30] and to hydrophobic associations obtained from the sequence[10] were also built and used in the simulations. We emphasize, however, that data supplied by the machine learning are the primary driver of search-focusing here, as the other sources are derived from general protein properties[9,10] which are too combinatorially numerous (and nonspecific) to scale to larger systems despite their previous success on small domains (see the Supporting Information methods for full details of the restraint types and their implementation).

II. Maximizing the Information Content of ML-Derived Restraints. As explained earlier, MELD ranks the restraints at each iteration according to their computed restraint energy, and the ones closest to zero energy get activated until the next reassessment. This design allows the tightening of conformations around subsets of the input data, but it also means that MELD always takes the "easy way out" in restraint energy space by enforcing the least-stretched springs. A data set of predicted contacts derived from ML predictions usually contains local and nonlocal contacts. Nonlocal contacts are more informative because they are most restrictive of conformational space. If all contacts were grouped together and MELD was asked to enforce a desired fraction from the group, then it will activate, on average, more of the local ("easy") restraints, leaving many useful nonlocal restraints unexploited. Therefore, in order to distribute the information during the MD more evenly across the predicted 2D contact map, the contacts were grouped into three categories based on their sequence contact order: short-range, medium-range and long-range contacts. The restraints in each group were enforced separately from the other two groups.

Naturally, a similar scenario would also arise within each of these groups. Therefore, a second layer of restraints was built to maximize data utility without enforcing all of it. This second layer is constructed by clustering the predicted contacts into contact hubs, where each hub includes all nearby contacts on the 2D contacts map (see Figure S2 and the Supporting Information methods for how this was done). In a similar fashion to the direct contacts, we parametrize potentials for the contact hubs, but here, we instruct MELD to enforce one contact as a representative of each hub. Both collections, short/medium/long contacts and their hubs, are enforced simultaneously but in separate groups in order to maximize the true contacts in the simulations.

III. Simulation Details. In all the simulations presented here, MELD x MD sampling creates system conformations (microstates) using the Amber's ff14SBside[31] force field and gbNeck2[32] implicit solvent model. MELD x MD traverses the complex folding landscape by relying on a one-dimensional Hamiltonian and temperature replica exchange molecular dynamics (H,T-REMD) with coupled temperatures and restraint force constants along the replica ladder.[10,15] Low temperature replicas are associated with high restraint force constants in order to refine the search in an energy minimum that is tightly compatible with the input data. The high temperature replicas have vanishing restraints and are able to unfold the protein and allow for fast exploration of chain dynamics. The temperature range is set to 300−500 K, with temperatures in the middle replicas automatically varying to improve exchanges. The REMD ladder contains 28−32 replicas in the simulations presented here. Macrostates, defined as ensembles of similar structures, are then built from the trajectories by clustering microstates in the low temperature replicas. The most populated macrostates correspond to low free-energy basins on the MELD x MD landscape. We focus on the single most populated macrostate (Top1) and compare its representative (centroid) conformation to the true (experimental) native structure.

Since CASP14 targets comprise the biggest challenge for folding methods, we focus on those proteins in the Results section. Because of computational expense and the CASP14 timeline for submitting models, we restricted our attempted targets to lengths less than 250 amino acids. In the main analysis, we focus our assessment on the 16 submitted targets whose native structures have no violations (i.e., zero restraint energy) to any of MELD's assumptions: the protein native state has a hydrophobic core, good secondary structure predictions by PSIPRED, and 80% accuracy in the input ML data (see the Supporting Information for a full analysis of all 33 submitted targets). We also note that the RMSD values reported correspond to deviations of the predicted model from the native one based on ordered secondary structure elements present in the experimental structure.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00916.

> Figures S1−S13, Tables S1−S3, MELD overview and restraints description, simulated protein sets, simulation details and analysis, time scales estimation, other sources of restraints, restraining function parametrization, macrostate time evolution, microstate rmsd distributions, TM-scores, MELD energies, contact maps derived from ML data, native stability tests, and misfolded states (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ken A. Dill − *Laufer Center for Physical and Quantitative Biology, Department of Physics and Astronomy, and Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States;* orcid.org/0000-0002-2390-2002; Email: dill@laufercenter.org

Authors

Roy Nassar − *Laufer Center for Physical and Quantitative Biology and Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States;* orcid.org/0000-0001-9822-6128

Emiliano Brini − *Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States;* orcid.org/0000-0002-1314-8405

Sridip Parui − *Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States;* orcid.org/0000-0003-0906-963X

Cong Liu − *Laufer Center for Physical and Quantitative Biology and Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States;* orcid.org/0000-0002-1687-2220

Gregory L. Dignon − *Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00916

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Alberto Perez for valuable discussions and suggestions. The authors are grateful for the support from NIH for Grants RM1GM135136 and R01GM125813, the University of Virginia GF10946, and the Laufer Center for Physical and Quantitative Biology at Stony Brook University. Roy Nassar also acknowledges the support from the Natural Sciences and Engineering Research Council of Canada (PGSD2-517099-2018). This research is part of the Frontera computing project at the Texas Advanced Computing Center. Frontera is made possible by National Science Foundation Award OAC-1818253. This research used resources of the Oak Ridge Leadership Computing Facility, which is a U.S. DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. This work used resources services and support provided via the COVID-19 HPC Consortium (https://covid19-hpc-consortium.org/), which is a unique private−public effort to bring together government, industry, and academic leaders who are volunteering free compute time and resources in support of COVID-19 research.

■ REFERENCES

(1) Brini, E.; Simmerling, C.; Dill, K. Protein storytelling through physics. *Science* **2020**, *370*, eaaz3041.

(2) Schlick, T.; Portillo-Ledesma, S.; Myers, C. G.; Beljak, L.; Chen, J.; Dakhel, S.; Darling, D.; Ghosh, S.; Hall, J.; Jan, M.; Liang, E.; Saju, S.; Vohr, M.; Wu, C.; Xu, Y.; Xue, E. Biomolecular modeling and simulation: a prospering multidisciplinary field. *Annu. Rev. Biophys.* **2021**, *50*, 267−301.

(3) Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A. Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol.* **2016**, *36*, 25−31.

(4) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.

(5) Lindorff-Larsen, K.; Piana, S.; Dror, R.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.

(6) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **2014**, *136*, 13959−13962.

(7) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.

(8) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **2010**, *6*, 751−758.

(9) Perez, A.; Morrone, J. A.; Dill, K. A. Accelerating physical simulations of proteins by leveraging external knowledge. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, e1309.

(10) Perez, A.; MacCallum, J. L.; Dill, K. A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 11846−11851.

(11) Pan, A. C.; Weinreich, T. M.; Piana, S.; Shaw, D. E. Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems. *J. Chem. Theory Comput.* **2016**, *12*, 1360−1367.

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

(13) Robertson, J. C.; Perez, A.; Dill, K. A. MELD x MD folds nonthreadables, giving native structures and populations. *J. Chem. Theory Comput.* **2018**, *14*, 6734−6740.

(14) Adhikari, U.; Mostofian, B.; Copperman, J.; Subramanian, S. R.; Petersen, A. A.; Zuckerman, D. M. Computational estimation of microsecond to second atomistic folding times. *J. Am. Chem. Soc.* **2019**, *141*, 6519−6526.

(15) MacCallum, J. L.; Perez, A.; Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A* **2015**, *112*, 6985−6990.

(16) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(17) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A* **2020**, *117*, 1496−1503.

(18) CASP14. *Protein Structure Prediction Center*. https://predictioncenter.org/index.cgi (accessed July 1, 2021).

(19) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 702−710.

(20) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889−895.

(21) Berman, A. L.; Kolker, E.; Trifonov, E. N. Underlying order in protein sequence organization. *Proc. Natl. Acad. Sci. U. S. A* **1994**, *91*, 4044−4047.

(22) Xu, D.; Nussinov, R. Favorable domain size in proteins. *Folding Des.* **1998**, *3*, 11−17.

(23) Sanyal, T.; Mittal, J.; Shell, M. S. A hybrid, bottom-up, structurally accurate, Go-like coarse-grained protein model. *J. Chem. Phys.* **2019**, *151*, 044111.

(24) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(25) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millan, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(26) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative

evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98−105.

(27) Zuckerman, D. M. Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.

(28) Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A* **2019**, *116*, 16856−16865.

(29) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.

(30) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.

(31) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(32) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020−2034.