Title
MELD × MD Folds Nonthreadables, Giving Native Structures and Populations.

Permalink
https://escholarship.org/uc/item/22r2b6pt

Journal
Journal of Chemical Theory and Computation, 14(12)

ISSN
1549-9618

Authors
Robertson, James C
Perez, Alberto
Dill, Ken A

Publication Date
2018-12-11

DOI
10.1021/acs.jctc.8b00886

Peer reviewed

Published in final edited form as:
J Chem Theory Comput. 2018 December 11; 14(12): 6734–6740. doi:10.1021/acs.jctc.8b00886.

MELD × MD Folds Nonthreadables, Giving Native Structures and Populations

James C. Robertson[†], **Alberto Perez**[†], **Ken A. Dill**[*,†,‡,§]

[†]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States

[‡]Department of Chemistry, Stony Brook University, Stony Brook, New York 11794, United States

[§]Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, United States

Abstract

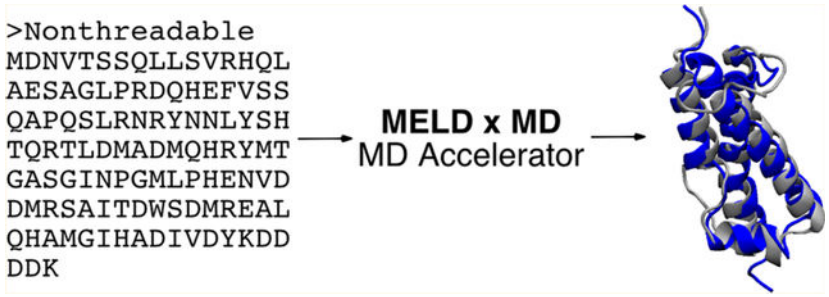
A current challenge is to compute the native structures of proteins from their amino acid sequences. A main approach of bioinformatics is threading, in which a protein to be predicted is computationally threaded onto protein fragments of similar sequence having an already known structure. However, ~15% of proteins cannot be folded in this way; this has been called the glass ceiling, and the proteins are called nonthreadables. For these, physical molecular dynamics (MD) modeling is promising because it does not require templates. We find that MD, when used with an accelerator called MELD, can fold many nonthreadables. For 41 nonthreadable proteins with fewer than 125 residues, MELD-accelerated MD (MELD × MD) folds 20 of them to better than 4 Å error. In 10 cases, MELD × MD succeeds even when the force field does not properly encode the native state. In 11 cases, MELD × MD foretells its own success; seeing large Boltzmann populations in the simulations predicts it has converged to the correct native state. MELD × MD acceleration can be applied to a broad physical protein modeling range.

Graphical Abstract

[*]Corresponding Author: dill@laufercenter.org.

Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00886.
Figures 1–6, Tables 1–3 (PDF)

The authors declare no competing financial interest.



1. INTRODUCTION

Computer modeling of proteins is valuable for understanding biological mechanisms of action, dynamical motions, biological function, and protein folding and binding and for designing ligands as drugs. An important challenge for computational methods, and a useful test bed, is to predict the native folded structure of a protein from its amino acid sequence. There are many methods, but they tend to range between two limits. (i) In bioinformatics, an early step in predicting an unknown protein structure is to find another protein, the template, that has a similar sequence and a known structure in the Protein Data Bank (PDB).[1] In threading, a particular sequence is scored for suitability with known structural fragments from a database of different folds. These database-dependent methods are often relatively successful when the level of homology is high. (ii) In physical molecular dynamics (MD) simulations, no such template is required because computations are fully self-contained within the physics of the model, but physical modeling is limited by some imperfections in force fields and the need for extensive computing resources. MD has not yet been a practical way to compute folded structures of proteins.

However, we recently developed an accelerator for MD simulations, called MELD (modeling employing limited data).[2,3] MELD-accelerated MD (MELD × MD) accelerates the search for important states when some limited (and often vague) information is available. For example, MELD × MD has been able to fold small proteins,[4,5] including in the blind competitive event critical assessment of structure prediction (CASP),[6–15] given only the knowledge that proteins have hydrophobic cores, are compact, and have secondary structure.[16–18] Here, we test MELD × MD in another situation that requires a physics-based method. In particular, ~15% of proteins cannot be threaded onto known templates and cannot be predicted using bioinformatics-based threading methods.[19] Skolnick has called this limitation a glass ceiling and the proteins nonthreadable. Here, we ask if MELD × MD is capable of predicting the native structures of nonthreadable proteins. We describe here a number of successes, but we also comment on the challenges and current limitations.

2. METHODS

2.1. Modeling Employing Limited Data.

MELD accelerates MD simulations, obeys detailed balance, and satisfies Boltzmann’s law. MELD × MD is also a Bayesian inference method, where the force field-generated structural ensemble is the prior, heuristics from general knowledge about proteins provide the likelihood, and the resulting structural ensemble is the posterior distribution. Data that are sparse, ambiguous, or unreliable can be effectively used in MELD × MD to limit conformational searching and accelerate simulations. Besides protein folding, MELD × MD also has applications in identifying pathways, determining protein–protein interactions, and ligand binding in drug discovery.[20]

2.2. Folding Simulations with MELD × MD.

MELD × MD uses Hamiltonian, temperature-replica exchange molecular dynamics (H,T-REMD)[21,22] with the AMBER ff14SBside[23] force field in the GBneck2 implicit solvent (igb = 8),[24] powered by OpenMM[25] to run on graphical processor units (GPUs). Each MELD × MD run was performed with 30 replicas ranging in temperature from 300 to 450 K, using Langevin dynamics, and a 4.5 fs time step with the hydrogen mass adjusted to 4.0 Da, but keeping the heavy atom and hydrogen pair mass the same. MELD × MD builds an initial extended structure from the sequence using tleap from AmberTools17.[26] Sets of restraints that impose secondary structure, hydrophobic contacts, and hydrogen bonding between β-strand pairs are tabulated from sequence information. Of all possible restraints generated, only a fraction are enforced at each time step; the energies of the restraints are calculated at each exchange step, and the lowest-energy restraints in each replica are activated. Secondary structure restraints are predicted with PSIPRED;[27] all predictions for helices and sheets are accepted, and 70% of the lowest-energy PSIPRED-generated restraints are active at each time step. Sets of hydrophobic and hydrogen bond restraints are generated; hydrophobic pairing is enforced so that there are 1.2 contacts per residue, and secondary structure strand pairing is enforced at 45%. The H,T-REMD is implemented as follows. At high temperatures, the restraints have low force constants and are zero at the highest temperature, while at lower temperatures, the force constants are increased. Exchange between replicas happens by the metropolis Monte Carlo method. Detailed explanations of MELD can be found in ref 3 or 2.

We simulated 41 nonthreadable proteins starting from an extended conformation using MELD × MD. Each system was run for at least 1 μs; 1GYZ, 1HYW, 1KAF, 1PC0, 1RQ6, 1A6S, 1EO0, and 1ND9 were run for 1.5 μs. The computational cost varied with protein size, but for every 1 μs of sampling, the nonthreadables used ~3000 XK node hours on the Blue Waters sustained petascale computing resource at the National Center for Supercomputing Allocations.

2.3. Selecting Nonthreadable Candidates for MELD × MD.

We selected nonthreadable proteins from the three databases at http://cssb2.biology.gatech.edu/threading/download.html.19 These databases contain proteins identified by Skonlick and Zhou as having a template modeling score (TM score)[28] of <0.4, below the value used to determine whether two proteins have the same fold.[29] The number of unique nonthreadable sequences across the three lists was found to be 898:676 from HHpred,[30,31] 637 from SP[3],[32] and 719 from PROSPECTOR_4.[33] Nonthreadable proteins come in a variety of sizes, from 30 to 3440 amino acids in length (though only five are longer than 215 amino acids), cover more than 700 Pfam families, contain a range of secondary structure features, and fold to low- and high-contact order structures (Figure 1 of the Supporting Information). Running MELD × MD simulations on all 898 nonthreadables was not computationally feasible, so a smaller set was selected by filtering out proteins that MELD × MD is not currently optimized to fold. Protein sequences selected for MELD × MD included fewer than 125 residues, were single-chain monomers, had net charges of ≤±5, had an at least 50% secondary structure composition (as predicted by PSIPRED), had no missing residues, and were not known to be membrane proteins. After the pre-MELD filter had been applied, the list was reduced to 41 MELD × MD candidate proteins (Figure 1). The filter eliminated 318 nonthreadables on the basis of size and 370 on the basis of low secondary structure content. The remaining sequences were filtered out because they formed multimers or complexes with other molecules.

2.4. Force Field Stability Tests.

The quality of our modeling results depends on the quality of the force field. Therefore, to establish whether any successes or failures of our modeling were due to a lack of sampling by MELD × MD or flaws in the force field, we first ran control experiments. We ran non-MELD × MD single-trajectory MD simulations (hereafter termed MD runs or stability tests to differentiate them from MELD × MD runs) of each protein starting from its known native conformation to determine whether the native state of the protein was stable in the force field and solvent model that we used in MELD × MD. Stability tests were run for 41 MELD × MD candidates with AMBER pmemd.cuda.[26,34] For these MD runs, we used the native structure downloaded from the PDB as the starting conformation. The ff14SBside protein force field was used with the GBneck2 implicit solvent, the same as in MELD × MD folding simulations. Systems were minimized with 5000 steps of steepest descent followed by 5000 steps of conjugate gradient. The MD systems were each run for 500 ns of production. Temperature REMD (T-REMD) was performed for systems 1PC0 and 1OQK, starting from the native conformation, with ff14SBside and GBneck2 using AMBER pmemd.cuda, from 300 to 450 K (12 replicas for 1PC0 and 14 for 1OQK).

2.5. Seeded MELD × MD Simulations.

For proteins that never sample native conformations in any of the MELD × MD ensembles, we seeded new MELD × MD simulations with the native structure to test whether the problem was insufficient sampling or the force field. The only difference between these simulations and the MELD × MD folding simulations described above is that the lowest-temperature replica started from the native conformation rather than from the extended conformation.

2.6. Ensemble Processing.

We postprocessed trajectories with a combination of scripts included with MELD × MD and CPPTRAJ[35] from AmberTools17. For MELD × MD simulations, trajectories from the five lowest-temperature replicas were clustered using the average-linkage hierarchical agglomerative algorithm with ϵ 2 Å. The conformational clustering was based on the root-mean-square deviation (RMSD) of Cα and Cβ atoms of secondary structure residues, as predicted by PSIPRED. The first 250 ns of trajectory frames was omitted for clustering. Representatives from the top five clusters were assessed in terms of their similarity to the native state by calculating the RMSD of Cα and Cβ atoms of residues in predicted secondary structure elements from the experimental PDB structure. Cluster representatives with RMSDs of <4.0 Å were considered folded to native. MD and T-REMD stability tests were analyzed with CPPTRAJ.

3. RESULTS AND DISCUSSION

3.1. MELD × MD Folds Nonthreadable Proteins.

We find that MELD × MD successfully folded 20 of our 41 nonthreadable targets (Figure 2 and Table 1 of the Supporting Information). We refer to proteins that MELD × MD successfully folded as “folders” and the rest as “nonfolders”. Of the 20 folders, 14 folded to structures having the single lowest free energy. For the other six, the true native was among the three lowest-free energy conformations.

3.2. MELD × MD Often Foretells When It Succeeds with Large Populations.

An important challenge is to know in advance when to trust that a computer simulation may have found the native state. The power of physical modeling, such as force field-based MD, is that it gives free energies and, hence, populations. Therefore, when MELD × MD converges on a state with a large population, it is evidence that the force field “thinks” it has found the state with the lowest free energy among all the states it has sampled. Indeed, we found this to be a good sign of success. When MELD × MD cluster populations exceeded 40%, the structure it found was within4.0 Å of native in all cases but one (Figure 3). Therefore, for blind predictions, this criterion is a good measure of confidence that the simulation has found the native state. When we see smaller conformational populations, it is inconclusive (Figure 2 of the Supporting Information).

3.3. MELD × MD Rescues 10 Predictions for Which the Native Protein Is Not Stable in the Force Field.

MELD × MD is just a search strategy, in principle always limited by the quality of the force field on which it relies. If a simulated protein is put into the true experimental native structure and if that structure is not stable in the force field, we should not expect a sampling strategy like MELD × MD to fix it. However, remarkably, we find that MELD × MD correctly identifies the native states of 10 proteins that are not stable in the force field (Figure 4 and Figure 3 of the Supporting Information). For example, initiating 1AA3 in its true native state in stability tests leads to its complete unfolding to structures 10 Å from native, but MELD × MD starting from unfolded found the correct native state and populated it. There were also nine other examples. The reason, apparently, is that external knowledge of secondary structures and a hydrophobic core were sufficient to help the force field find the correct native state.

However, not surprisingly, MELD × MD cannot always rescue force field failures. For example, 1W09 has a proline in the middle of the third helix. Typically, this predicts helix breaking. MELD × MD generated a kinked third helix. However, the true native structure has three straight helices. The force field problems shown with 29 nonthreadables that sample non-native ensembles provide additional data for benchmarking new protein force fields. Especially interesting for force field development might be the 10 proteins that

MELD × MD folded despite the force field favoring other conformations. This indicates that secondary structure propensities are likely at fault and the restraints used in MELD × MD for secondary structure help push those conformations to higher energies. In addition, two all-β proteins, 1PC0 and 1OQK, were failures of insufficient MELD × MD sampling, not the force field. Both proteins were mostly in low-RMSD conformations at 300 K in the T-REMD stability tests, indicating that the force field was not the reason MELD × MD did not populate these native conformations (Figure 4 of the Supporting Information).

3.4. MELD × MD Sometimes Cannot Rescue a Prediction from Poor Secondary Structure Predictions.

We found only one example in which predicting a large-population state did not correctly predict the native state (Figure 3 and Figure 5 of the Supporting Information). Upon further inspection, we found that the PSIPRED secondary structure prediction failed to predict the β-sheets present in the native conformation. Instead of PSIPRED predicting $\beta\alpha\alpha\beta\alpha$ for 1ND9, PSIPRED predicted $\alpha\alpha\alpha$. The result was that MELD × MD folded 1ND9 to a structure 5.2 Å from native, with secondary structures that agreed with the PSIPRED prediction. Therefore as a test, we reran 1ND9 in MELD × MD, giving it only the correct native secondary structures this time. The best prediction was still non-native, now 5.1 Å from native. It was somewhat improved but with helices that were longer than those of the native form. The force field is known to overstabilize helices.[23,36,37] Even so, by a different measure, the global distance test (GDT),[38] the structure was found to be closer to native when given the correct secondary structures (Figure 5 of the Supporting Information). In short, while we know that MELD × MD can rescue structures from wrong input knowledge sometimes, it cannot always.

3.5. MELD × MD Found and Sampled Most Native Structures Well.

MELD × MD is an efficient search strategy that was previously shown to decrease folding time on 20 fast folding proteins by up to 5 orders of magnitude compared to those seen with single-trajectory, “traditional” molecular dynamics.[3] Here, we show that MELD × MD finds native states of nonthreadables within 1 μs per replica simulation time. In fact, many fold to native within 250 ns per replica sampling time (see Figure 6 of the Supporting Information), although sampling was extended to see whether others would eventually find native or move away from native with an increased level of sampling.

For 15 proteins that never sampled native in the original MELD × MD runs, only one protein folded to native in seeded MELD × MD simulations (Table 2 of the Supporting Information). This suggests that the force field was responsible for 14 of these nonfolders, while sampling was an issue for 1LN4. The PSIPRED secondary structure predictions fed into the original MELD × MD simulation of 1LN4 were quite accurate, but the lowest-free energy structure was 10.5 Å from native with a helix in place of β2. In addition, the three other β-sheets were not properly paired. This suggests a combination of problems in 1LN4: the force field is stabilizing α over β, and MELD × MD is not properly pairing the other β-sheets. A possible improvement for MELD × MD is a better β-strand pairing scheme. Ultimately, however, we found that by seeding 15 new MELD × MD simulations with their true native structures, only one folded to native, indicating problems with the force field rather than sampling for those proteins.

3.6. What Protein Properties Determine Whether MELD × MD Can Fold Them or Not?

Here, we describe which proteins are foldable, and which are not, by MELD × MD (Figure 5). First, we looked at protein size. Folding succeeded for proteins ranging in size from 46 to 108 amino acids and failed for proteins ranging in size from 49 to 110 amino acids. Therefore, size, at least in this range, is not a critical determinant. However, not surprisingly, successes were greater for sequences in the range of 50–75 amino acids. Previous studies with MELD × MD have folded up to 97 amino acids with heuristics-informed restraints (same input as in this study) and up to 212 with experimental data-informed restraints.[5] Folding 1R5E (105 amino acids) and 1KAF (108 amino acids) demonstrates that MELD × MD goes beyond 100-mers without experimental data or co-evolutionary information. Importantly, MELD × MD folded 1R5E to within 2.5 Å, with a cluster population of 85%.

We looked at net charge. Proteins having a small net charge, ranging from −5 to +5, were equally likely to fold or not fold, indicating that this range of charges was tolerable. Proteins with higher net charges were prefiltered out to avoid known problems with implicit solvent models such as the one we use here.[37]

We also looked at the protein contact order, a measure of how nonlocal the average contacts are. Larger contact order proteins tend to fold more slowly,[39] indicating that it is physically more difficult for the protein to find its native state in test tubes. However, MELD × MD folded proteins with relatively high contact orders (Table 3 of the Supporting Information). The relative contact order of the native state PDB structure was determined using Plaxco's[39] perl script and the default 6 Å heavy atom cutoff. MELD × MD folded proteins 1HDN, 1J27, and 1KN6, which had relative contact orders of 0.18, 0.19, and 0.21, respectively, which are all higher than 0.17, the highest contact order for a nonfolder. This shows that MELD × MD is not limited by proteins with both high contact order and sequence lengths approaching 100 amino acids, because 1HDN, 1J27, and 1KN6 had 85, 98, and 73 residues, respectively, though both MELD × MD folders with more than 100 amino acids (1KAF and 1R5E) had relative contact orders close to 0.10.

We also looked at whether the quality of the secondary structure predictions that were input into MELD × MD was a predictor of folding success or failure. We used PSIPRED-predicted secondary structure to enforce secondary structure restraints for α-helices, β-sheets, and β-sheet strand pairing. The PSIPRED restraints for the set of MELD × MD candidates matched quite well with the native secondary structure content of these nonthreadables. The distributions of secondary structures were similar for folders and nonfolders, although folders had more α-helical content in the range of 50–75% and more β-sheets compared to nonfolders. Poor PSIPRED predictions were overridden by MELD × MD in some cases but not in others. For example, PSIPRED was 70% accurate in predicting secondary structure for 1AA3, a protein that MELD × MD was able to fold. In contrast, PSIPRED was 96% accurate for 2EZK; however, the lowest-RMSD structure that MELD × MD sampled was only 4.2 Å, and the lowest-RMSD cluster representative was 6.6 Å, because of force field deficiencies.

4. CONCLUSIONS

We have shown that molecular dynamics force field simulations, accelerated by a Bayesian method called MELD × MD, predict well the native structures of 20 nonthreadable proteins that are smaller than 125-mers. These are proteins that cannot currently be folded by bioinformatics-based threading methods. A virtue of such physics-based simulations is that they give free energies and state populations, which gives a confidence measure in advance that the method is finding the right structure. Proteins may have features that make them more or less likely to fold with our method, but none were identified in this study. MELD × MD may be useful for leveraging physics-based modeling for molecules or actions that are larger than can otherwise be handled by normal MD alone.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors are grateful to Jeff Skolnick for suggesting this problem to us. The authors thank the National Science Foundation (NSF) Blue Waters team for computing resources and their considerable help. Blue Waters is supported by NSF Grants ACI-0725070 and ACI-1238993, the State of Illinois, and the University of Illinois. This work is also part of the “Petascale integrative approaches to protein structure prediction” PRAC allocation support from the National Science Foundation (Grant ACI1514873). The authors also thank the Laufer Center for support.

Funding

The authors are thankful for IRACDA NY-CAPS Grant K12-GM-102778 to J.C.R. and National Institutes of Health Grant GM125813 to A.P. and K.A.D.

REFERENCES

(1). Burley SK; Berman HM; Christie C; Duarte JM; Feng Z; Westbrook J; Young J; Zardecki C RCSB Protein Data Bank: Sustaining a Living Digital Data Resource that Enables Breakthroughs in Scientific Research and Biomedical Education. Protein Sci 2018, 27, 316–330. [PubMed: 29067736]

(2). MacCallum JL; Perez A; Dill K 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. [PubMed: 26038552]

(3). Perez A; MacCallum JL; Dill K Accelerating Molecular Simulations of Proteins using Bayesian Inference on Weak Information. Proc. Natl. Acad. Sci. U. S. A 2015, 112, 11846–11851. [PubMed: 26351667]

(4). Perez A; Morrone JA; Simmerling C; Dill K Advances in Free-Energy-Based Simulations of Protein Folding and Ligand Binding. Curr. Opin. Struct. Biol 2016, 36, 25–31. [PubMed: 26773233]

(5). Perez A; Morrone JA; Brini E; MacCallum JL; Dill K Blind Protein Structure Prediction Using Accelerated Free-Energy Simulations. Sci. Adv 2016, 2, e1601274–e1601274. [PubMed: 27847872]

(6). Moult J; Pedersen JT; Judson R; Fidelis K A Large-Scale Experiment to Assess Protein Structure Prediction Methods. Proteins: Struct., Funct., Genet 1995, 23, ii–iv. [PubMed: 8710822]

(7). Moult J; Fidelis K; Zemla A; Hubbard T Critical Assessment of Methods of Protein Structure Prediction (CASP): Round IV. Proteins: Struct., Funct., Genet 2001, 45, 2–7. [PubMed: 11536354]

(8). Moult J; Fidelis K; Zemla A; Hubbard T Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round V. Proteins: Struct., Funct., Genet 2003, 53, 334–339. [PubMed: 14579322]

(9). Moult J; Fidelis K; Rost B; Hubbard T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round 6. Proteins: Struct., Funct., Genet 2005, 61, 3–7. [PubMed: 16187341]

(10). Moult J; Fidelis K; Kryshtafovych A; Rost B; Hubbard T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction-Round VII. Proteins: Struct., Funct., Genet 2007, 69, 3–9. [PubMed: 17918729]

(11). Moult J; Fidelis K; Kryshtafovych A; Rost B; Tramontano A Critical Assessment of Methods of Protein Structure Prediction-Round VIII. Proteins: Struct., Funct., Genet 2009, 77, 1–4.

(12). Moult J; Fidelis K; Kryshtafovych A; Tramontano A Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round IX. Proteins: Struct., Funct., Genet 2011, 79, 1–5.

(13). Moult J; Fidelis K; Kryshtafovych A; Schwede T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round x. Proteins: Struct., Funct., Genet 2014, 82, 1–6.

(14). Moult J; Fidelis K; Kryshtafovych A; Schwede T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction: Progress and New Directions in Round XI. Proteins: Struct., Funct., Genet 2016, 84, 4–14. [PubMed: 27171127]

(15). Moult J; Fidelis K; Kryshtafovych A; Schwede T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round XII. Proteins: Struct., Funct., Genet 2018, 86, 7–15. [PubMed: 29082672]

(16). Dill KA Dominant Forces in Protein Folding. Biochemistry 1990, 29, 7133–7155. [PubMed: 2207096]

(17). Pauling L; Corey RB; Branson HR The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. Proc. Natl. Acad. Sci. U. S. A 1951, 37, 205–211. [PubMed: 14816373]

(18). Dill KA Theory for the Folding and Stability of Globular Proteins. Biochemistry 1985, 24, 1501–1509. [PubMed: 3986190]

(19). Skolnick J; Zhou H Why is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? J. Phys. Chem. B 2017, 121, 3546–3554. [PubMed: 27748116]

(20). Morrone JA; Perez A; Deng Q; Ha SN; Holloway MK; Sawyer TK; Sherborne BS; Brown FK; Dill KA Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. J. Chem. Theory Comput 2017, 13, 863–869. [PubMed: 28042965]

(21). Sugita Y; Okamoto Y Replica-Exchange Molecular Dynamics Method for Protein Folding. Chem. Phys. Lett 1999, 314, 141–151.

(22). Fukunishi H; Watanabe O; Takada S On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. J. Chem. Phys 2002, 116, 9058–9067.

(23). Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput 2015, 11, 3696–3713. [PubMed: 26574453]

(24). Nguyen H; Roe DR; Simmerling C Improved Generalized Born Solvent Model Parameters for Protein Simulations. J. Chem. Theory Comput 2013, 9, 2020–2034. [PubMed: 25788871]

(25). Eastman P; Friedrichs MS; Chodera JD; Radmer RJ; Bruns CM; Ku JP; Beauchamp KA; Lane TJ; Wang L-P; Shukla D; Tye T; Houston M; Stich T; Klein C; Shirts MR; Pande VS OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. J. Chem. Theory Comput 2013, 9, 461–469. [PubMed: 23316124]

(26). Case DA; Betz RM; Cerutti DS; Cheatham TE; Darden TA; Duke RE; Gohlke H; Goetz AW; Homeyer N; Izadi S; Janowski P; Kaus J; Kovalenko A; Lee TS; LeGrand S; Li P; Lin C; Luchko T; Luo R; Madej B; Mermelstein D; Merz KM; Monard G; Nguyen H; Nguyen H; Omelyan I; Onufriev A; Roe DR; Roitberg A; Sagui C; Simmerling CL; Botello-Smith W; Swails J; Walker R; Wang J; Wolf RM; Wu X; Kollman PA Amber 16; University of California, San Francisco: San Francisco, 2016.

(27). Jones DT Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. J. Mol. Biol 1999, 292, 195–202. [PubMed: 10493868]

(28). Zhang Y; Skolnick J Scoring Function for Automated Assessment of Protein Structure Template Quality. Proteins: Struct., Funct., Genet 2004, 57, 702–710. [PubMed: 15476259]

(29). Xu J; Zhang Y How Significant is a Protein Structure Similarity with TM-score = 0.5? Bioinformatics 2010, 26, 889–895. [PubMed: 20164152]

(30). Söding J Protein Homology Detection by HMM-HMM Comparison. Bioinformatics 2005, 21, 951–960. [PubMed: 15531603]

(31). Meier A; Söding J Automatic Prediction of Protein 3D Structures by Probabilistic Multi-Template Homology Modeling. PLoS Comput. Biol 2015, 11, e1004343. [PubMed: 26496371]

(32). Zhou H; Zhou Y SPARKS 2 and SP3 Servers in CASP6. Proteins: Struct., Funct., Genet 2005, 61, 152–156.

(33). Lee SY; Skolnick J TASSER_WT: A Protein Structure Prediction Algorithm with Accurate Predicted Contact Restraints for Difficult Protein Targets. Biophys. J 2010, 99, 3066–3075. [PubMed: 21044605]

(34). Götz AW; Williamson MJ; Xu D; Poole D; Le Grand S; Walker RC Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput 2012, 8, 1542–1555. [PubMed: 22582031]

(35). Roe DR; Cheatham TE III PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput 2013, 9, 3084–3095. [PubMed: 26583988]

(36). Hornak V; Abel R; Okur A; Strockbine B; Roitberg A; Simmerling C Comparison of Multiple AMBER Force Fields and Development of Improved Protein Backbone Parameters. Proteins: Struct., Funct., Genet 2006, 65, 712–725. [PubMed: 16981200]

(37). Roe DR; Okur A; Wickstrom L; Hornak V; Simmerling C Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. J. Phys. Chem. B 2007, 111, 1846–1857. [PubMed: 17256983]

(38). Zemla A; Venclovas C; Moult J; Fidelis K Processing and Analysis of CASP3 Protein Structure Predictions. Proteins: Struct., Funct., Genet 1999, 37, 22–29.

(39). Plaxco KW; Simons KT; Baker D Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins. J. Mol. Biol 1998, 277, 985–994. [PubMed: 9545386]

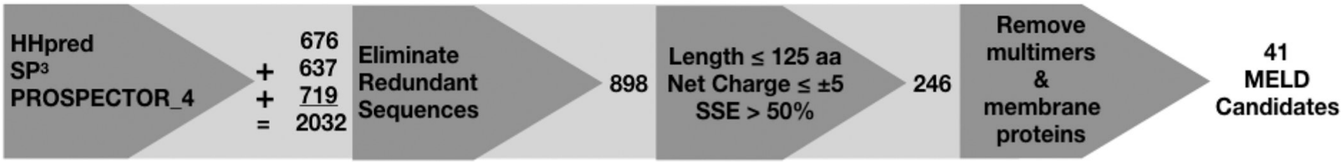


Figure 1.

Subset of nonthreadables that were selected for MELD × MD. Nonthreadables were filtered out to eliminate sequences least likely to be folded by MELD × MD. A set of 41 protein monomers with fewer than 125 residues, a low net charge, and a high secondary structure content were selected for MELD × MD simulations.

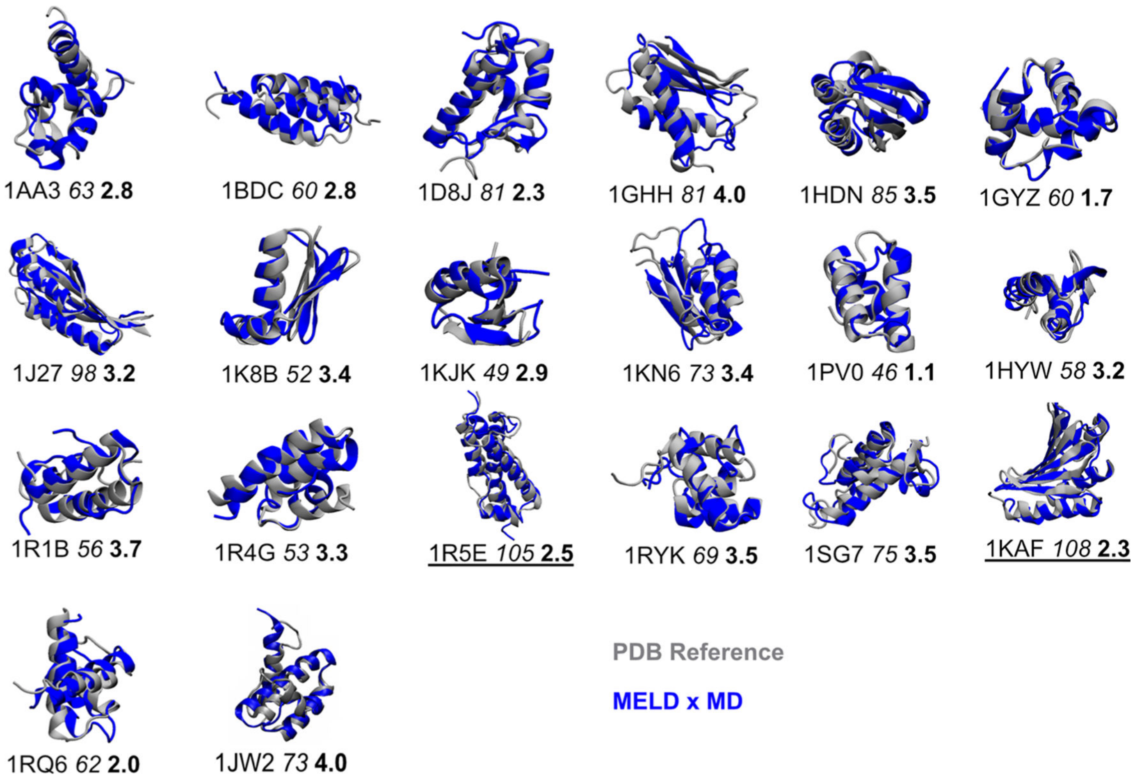


Figure 2.

MELD × MD predictions vs true experimental natives. (Blue) MELD × MD predicted structures, folded from fully extended. (Gray) True natives from the PDB. Also given are the PDB identification numbers, sequence lengths in italics (>100-mers are underlined), and the root-mean-square deviations (RMSD) in angstroms of the MELD × MD structure from the PDB reference. The RMSD was calculated for residues in secondary structure elements.

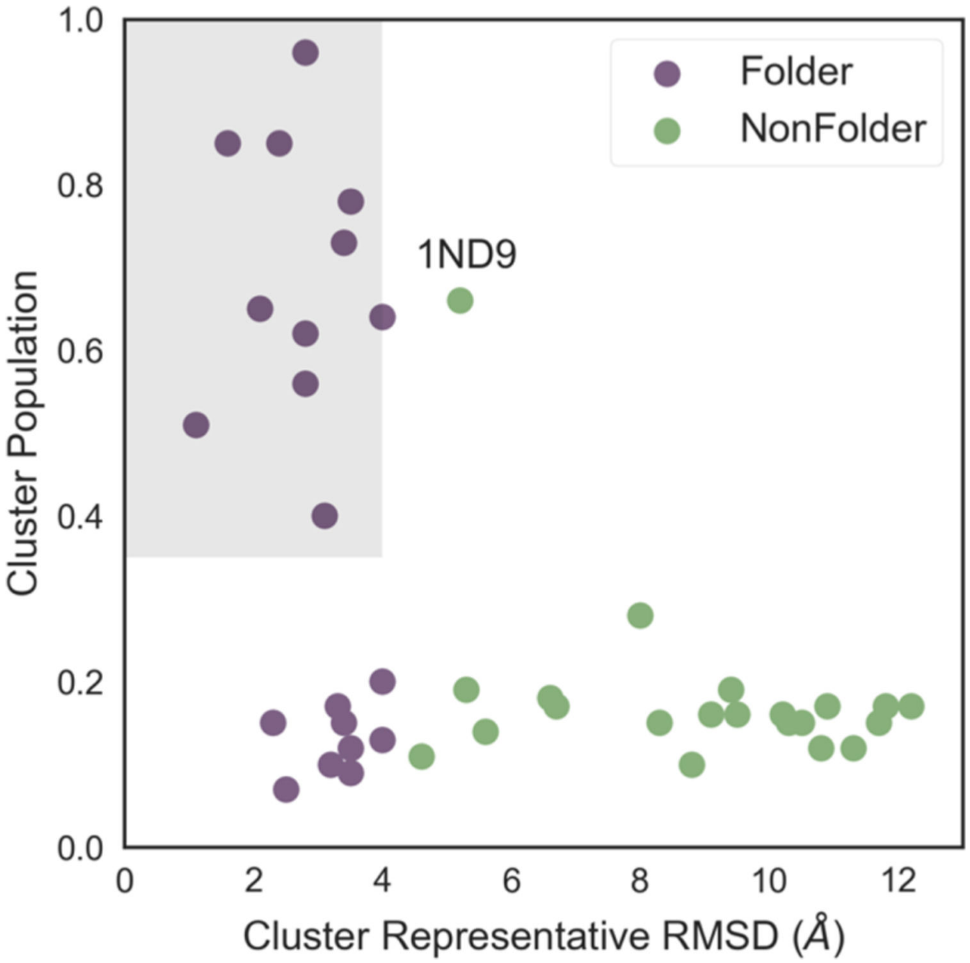


Figure 3.

Large populations from MELD × MD foretell its success. Proteins that have large MELD × MD cluster populations fold to the native structure (low RMSD). Protein 1ND9 is an exception (see the text). MELD × MD also folded some proteins to native that had small cluster populations, but usually small populations imply non-native folds or a lack of convergence.

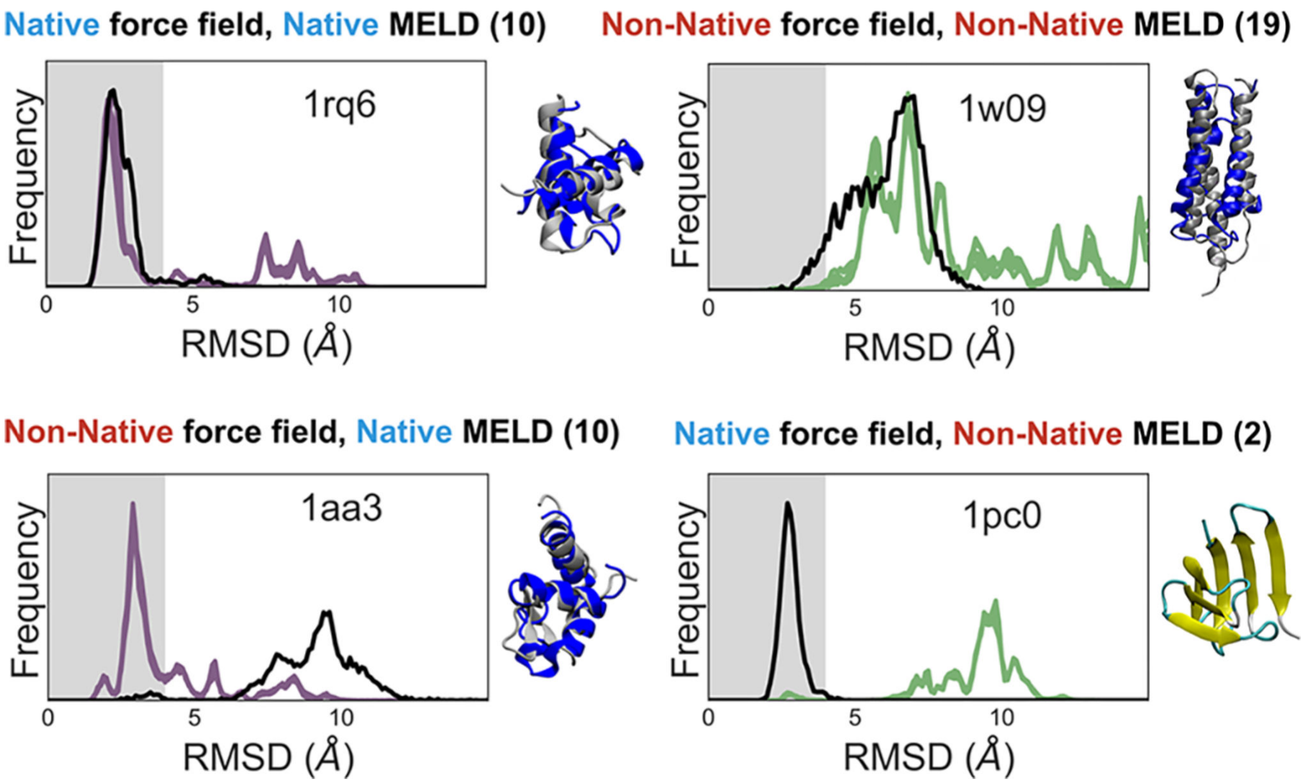


Figure 4.

MELD × MD distributions of folding compared to MD of native structures. (Top left) Example in which the native structure is stable in the force field and MELD × MD samples it well. (Top right) The force field gives the wrong structure, and MELD × MD finds the wrong structure. (Bottom left) The force field gives the wrong structure, but MELD × MD rescues it and finds the right structure. (Bottom right) The native structure is stable in the force field, but MELD × MD does not sample it. The number of occurrences of each type is given in parentheses. (Silver) True natives. (Blue) MELD × MD prediction. (Yellow) β-Sheets of true native. In short, in half of the cases, MELD × MD finds good native structures, and in the other half, force field errors cannot be rescued by MELD × MD.

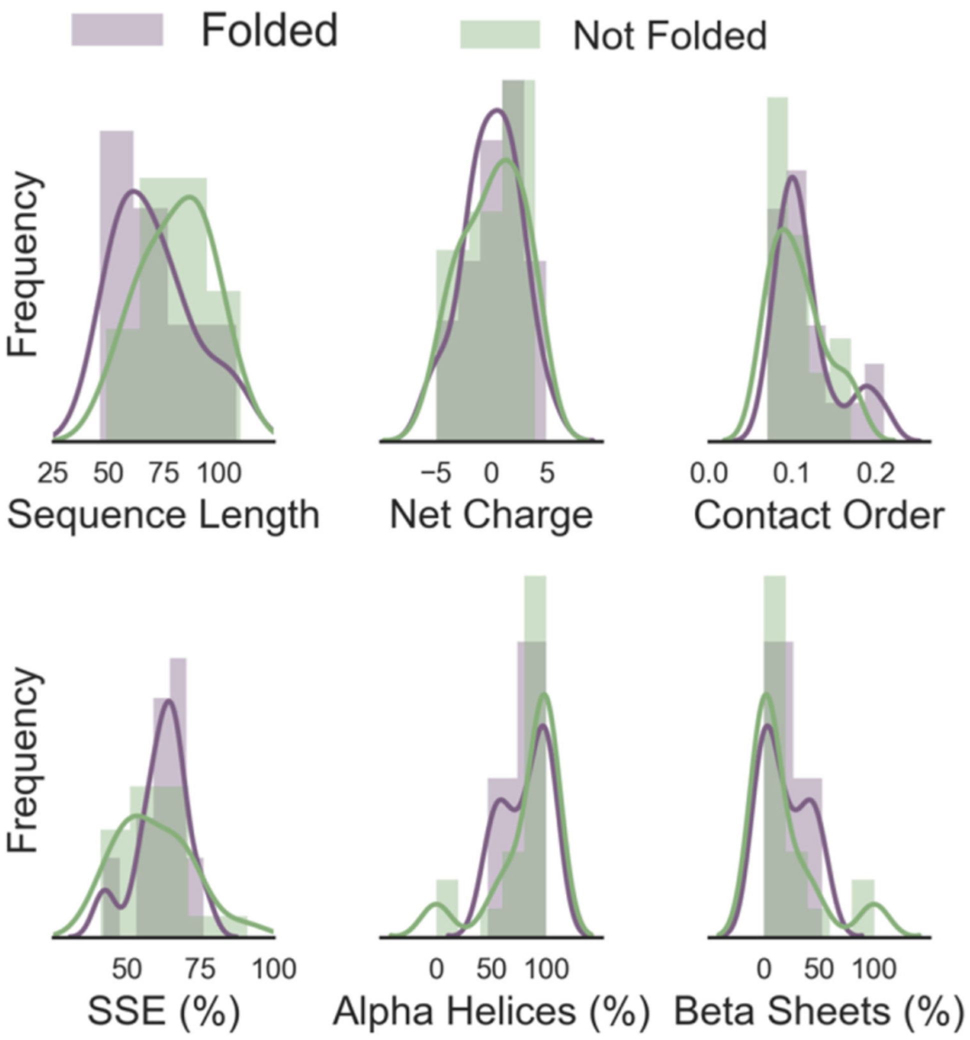


Figure 5.

Features that do not determine MELD × MD success. Histograms of sequence and structural features for nonthreadable proteins (purple) folded by MELD × MD compared to those (green) not folded by MELD × MD with (solid line) a smoothed estimate of the distributions from kernel density estimation. For the features we examined, none could be used to predict MELD × MD success a priori.