JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

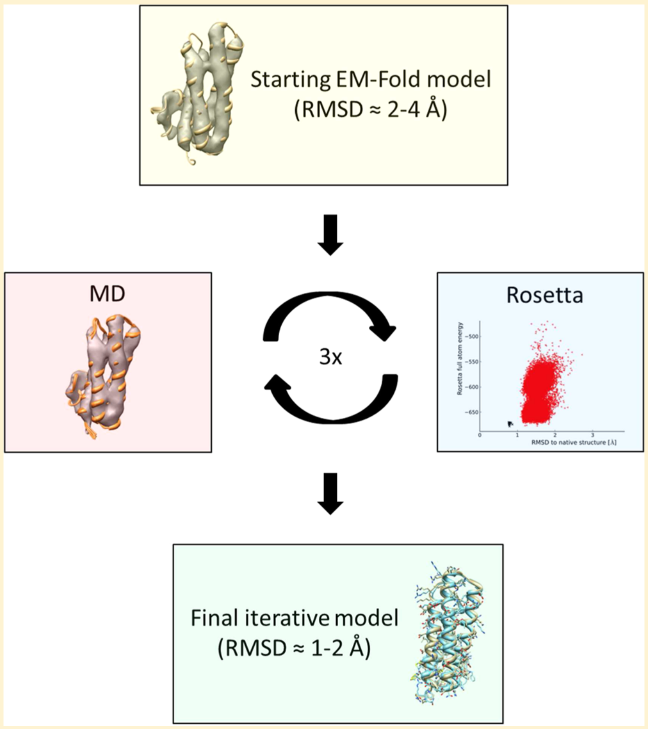
Improved cryoEM-Guided Iterative Molecular Dynamics−Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction

Steffen Lindert*[†,‡] and J. Andrew McCammon[†,‡,§,||]

[†]Department of Pharmacology, [‡]Center for Theoretical Biological Physics, [§]Howard Hughes Medical Institute, and [||]Department of Chemistry & Biochemistry, NSF Center for Theoretical Biological Physics, National Biomedical Computation Resource, University of California San Diego, La Jolla, California 92093, United States

S *Supporting Information*

ABSTRACT: Many excellent methods exist that incorporate cryo-electron microscopy (cryoEM) data to constrain computational protein structure prediction and refinement. Previously, it was shown that iteration of two such orthogonal sampling and scoring methods – Rosetta and molecular dynamics (MD) simulations – facilitated exploration of conformational space in principle. Here, we go beyond a proof-of-concept study and address significant remaining limitations of the iterative MD−Rosetta protein structure refinement protocol. Specifically, all parts of the iterative refinement protocol are now guided by medium-resolution cryoEM density maps, and previous knowledge about the native structure of the protein is no longer necessary. Models are identified solely based on score or simulation time. All four benchmark proteins showed substantial improvement through three rounds of the iterative refinement protocol. The best-scoring final models of two proteins had sub-Ångstrom RMSD to the native structure over residues in secondary structure elements. Molecular dynamics was most efficient in refining secondary structure elements and was thus highly complementary to the Rosetta refinement which is most powerful in refining side chains and loop regions.



■ INTRODUCTION

Computationally predicting a protein's three-dimensional structure from its amino acid sequence is one of the great challenges in biochemistry. This task is especially challenging if no proteins of known structure with sequence or structural homology exist—a field known as *de novo* or *ab initio* protein structure prediction. A plethora of methods have been developed over the last decades which have shown extraordinary promise in predicting the structure of ever larger proteins. Protein energy landscape theory (based on the concept that the energy landscape of a foldable protein looks like a rugged funnel) has been applied to the development of simple folding kinetics models and to obtain optimal energy functions for protein structure prediction.[1] Molecular Dynamics (MD) with physics-based energy functions is the most physically stringent method for realistically predicting protein structure.[2] Traditionally, MD has been confined to folding the smallest of peptides only.[3] However, massively distributed computing has helped MD protein folding simulations access both size and accuracy levels not seen before.[4] Recently, the advent of the Anton supercomputer[5] has allowed researchers to run conventional MD simulations to probe folding pathways of very small proteins within simulation times upward of 100 μs.[6] Arguably the biggest success in the field of *de novo* protein folding has however been reserved to methods using knowledge-based energy functions derived from statistics of the solved structures in deposited PDB.[2] Rosetta[7] and TASSER[8] are probably the two most notable and also the most consistently successful algorithms in the free modeling category of CASP (Critical Assessment of protein Structure Prediction), a biennial blind community-wide experiment for computational

Received: November 5, 2014
Published: February 5, 2015

DOI: 10.1021/ct500995d

protein structure prediction.[9] All the above-mentioned methods have also been applied successfully to the related problem of protein structure refinement where an approximate backbone conformation has to be improved so that side chain coordinates can be assigned properly.

Despite impressive success over the past decade, *de novo* protein modeling is still limited by protein size with the actual limit depending heavily on the methodology used. Most algorithms cannot predict protein structures *de novo* for proteins larger than 150 residues,[10] even though noticeable exceptions exist.[11] To help either predict or refine protein structure for larger proteins or increase modeling accuracy, sparse experimental restraints are routinely used. Sparse experimental data—in itself not sufficient to derive a high-resolution protein structure—can be used in combination with computational methods to either decrease the conformational search space or serve as an additional term in the energy function that rewards compliance with experimental data. The most commonly used experimental restraints are obtained from nuclear magnetic resonance (NMR),[12] electron paramagnetic resonance (EPR) spectroscopy,[13] and low resolution X-ray crystallography[14] as well as electron microscopy (EM).[15]

Particularly recent rapid improvements in the field of cryo-electron microscopy (cryoEM) led to a growing number of medium-resolution cryoEM density maps.[16] In those density maps – at a resolution of 6–9 Å – strong density rods are visible in protein regions with secondary structure, but loop and side chain density generally remains elusive. A multitude of methods has been developed to build secondary structure elements (SSEs) into those density maps[15c,17] as well as to refine those initial models.[15a,e] One obstacle in achieving near-atomic resolution protein models based on Rosetta refinement in medium-resolution cryoEM data is insufficient sampling. When benchmarking the refinement of 27 proteins in medium resolution cryoEM density maps, it was observed that the Rosetta energy function can distinguish native-like from non-native-like conformations by score.[15e] However, generally no sufficiently native-like models were built to take advantage of the deterministic scoring function. Speculating that the conformational search algorithm (based on fragment replacement and side chain repacking) became stuck in "conformational traps" from which no easy escape was possible using the sampling provided by Rosetta, we developed an iterative MD-Rosetta protocol.[18] The idea that an orthogonal sampling and scoring strategy might facilitate exploration of conformational space proved useful in some of the benchmark proteins. For those, steady improvement in model quality was observed in every round of the iterative protocol. While this work was a powerful proof of principle, there were two significant shortcomings that rendered it insufficient to be used outside of benchmark settings. First, while clearly meant to be a hybrid cryoEM-computational method, the MD section of the protocol was not using the cryoEM density map at all. Second, the models generated by the MD protocol had to be picked based on RMSD to the native model. Outside of benchmark settings, a more neutral way of identifying good models had to be used.

In the present work, we have addressed those shortcomings in the iterative Rosetta–MD protocol and are presenting significant improvements to computational high-resolution refinement of protein structures guided by medium-resolution cryoEM data. Starting from low-RMSD protein conformations (RMSDs of 2–5 Å to the native structure, generated with EM-Fold[15e]) we are using a combination of Rosetta and MD, guided by medium-resolution cryoEM density maps, to gradually improve model quality through iterative refinement. All models are now picked based on criteria such as score or simulation time.

MATERIAL AND METHODS

System Preparation and Molecular Dynamics (MDFF) Simulations. Four different proteins were chosen for which low-RMSD models had been built *de novo* with EM-Fold and Rosetta:[15e] 1X91, 1ICX, 1DVO, and 2FD5. Three of those were α-helical proteins, while one was an α–β-protein (1ICX). The size of the proteins was between 152 and 180 residues. The best scoring model after the third round of Rosetta-only refinement in ref 15e for each protein was extracted and used as input model for the first round of the iterative cryoEM-guided MD–Rosetta protocol. For the MD section of the protocol, the four systems were prepared for molecular dynamics simulations in NAMD using Molecular Dynamics Flexible Fitting (MDFF[15h,19]). VMD was used to add a TIP3P water box with a 14 Å padding. K^+ and Cl^- ions were added to neutralize the system and obtain a physiologically relevant ionic strength (150 mM). 34 K^+ and 33 Cl^- ions were added to 1X91, 32 K^+ and 25 Cl^- ions were added to 1ICX, 53 K^+ and 60 Cl^- ions were added to 1DVO, and 36 K^+ and 34 Cl^- ions were added to 2FD5. The fully solvated systems contained approximately 38000 (1X91), 28000 (1ICX), 59000 (1DVO), and 39000 (2FD5) atoms, respectively. The actual number varied slightly during the three rounds of MDFF. The mdff package was used for the density specific preparations. The griddx command was used to convert the cryoEM density maps from mrc format to dx format (the MDFF potential file). Subsequently, a PDB file containing the per-atom scaling factors was generated with gridpdb and dihedral angle restraints in SSEs were applied using ssrestraints. Peptide bond configurations and chiral centers were also restrained (using cispeptide and chirality restrain). NAMD configuration files were generated using the mdff setup command. In every round of the iterative protocol, two MDFF simulations were performed in two consecutive steps as outlined in the MDFF tutorial.[15h,19] The CHARMM27 force field[20] was used for the simulations. In the first step, a short 200 time step minimization using NAMD 2.9[21] was followed by 1 ns of MD using a low density scaling factor of 0.3 (gscale = 0.3). In a second step, a 200 ps minimization using a higher density scaling factor of 5 to 10 followed the simulation in the first step. Periodic boundary conditions were used, along with a nonbonded interaction cutoff of 10 Å for Particle Mesh Ewald (PME) long-range electrostatic interaction calculations. Bonds involving hydrogen atoms were constrained using the SHAKE algorithm,[22] allowing for a time step of 2 fs. Structures were saved every 2 ps.

Rosetta All Atom Refinement in Density Map. In each round of the iterative protocol, the final models from both steps of the MDFF simulations were subjected to loop rebuilding and refinement within Rosetta[7a,15a] guided by the cryoEM density map. The Rosetta refinement protocol is identical to the protocol described in ref 18. In summary, regions of the models that agree least with the density map of the protein are identified (loops_from_density.linuxgccrelease) and rebuilt guided by the density map (loopmodel.linuxgccrelease). Each round performs a full atom relaxation of the entire structure.

Iterative MD–Rosetta Protocol. Three rounds of iterative MDFF and Rosetta were run for all four benchmark proteins.

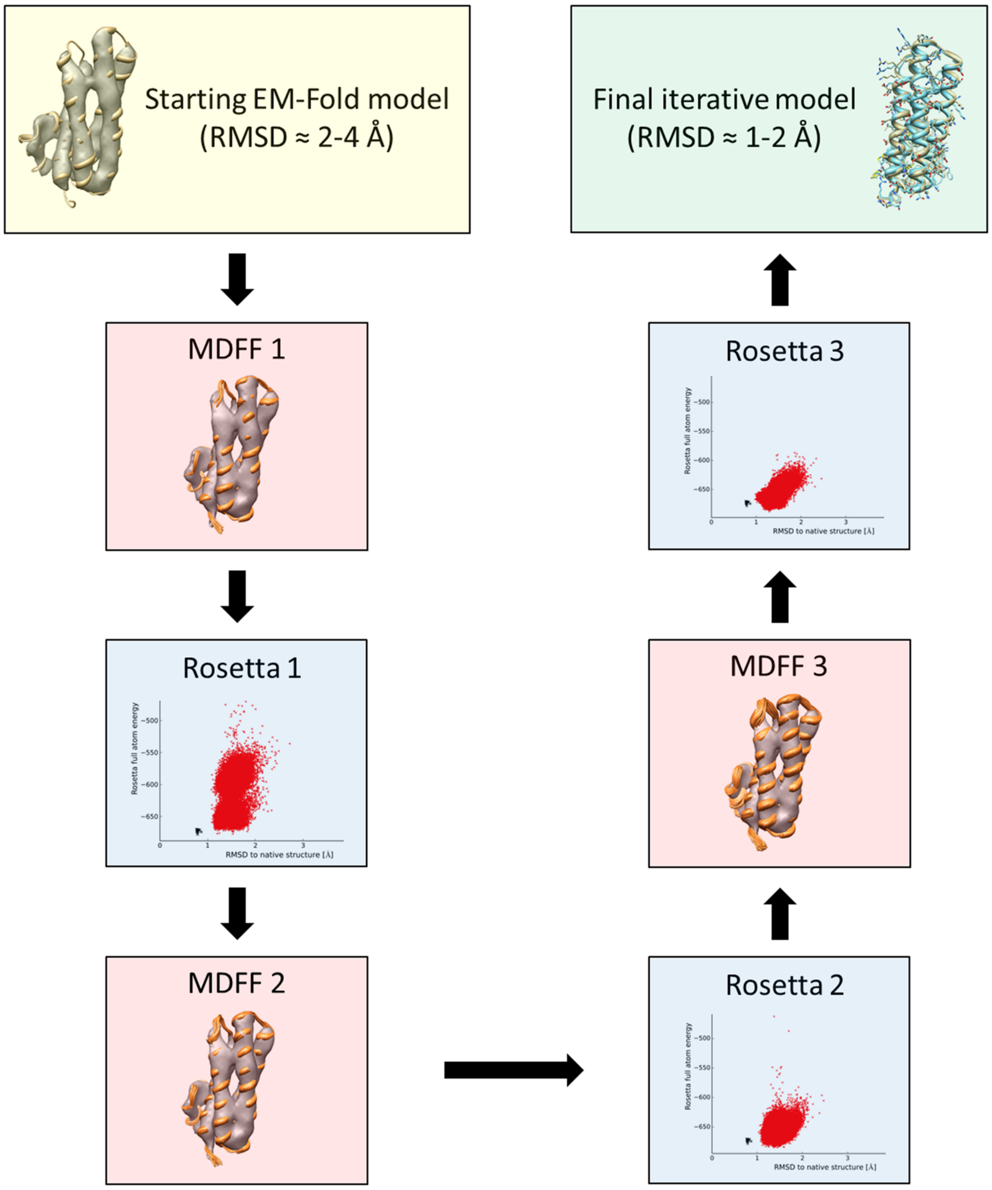


Figure 1. Flowchart of the iterative cryoEM-guided Rosetta−MD protocol. For each of the proteins, the starting model for the first round of the protocol was the best scoring model from a previous benchmark. The protocol started with an MDFF run, followed by a Rosetta run, repeated twice: MDFF1−Rosetta1−MDFF2−Rosetta2−MDFF3−Rosetta3. The best scoring models after the third round of Rosetta was picked as the final model.

Figure 1 shows a flowchart of the iterative Rosetta−MDFF protocol. For each of the proteins, the starting model for the first round of the protocol was the best scoring model after the third round of a Rosetta-only refinement from ref 15e. The protocol started with an MDFF run, followed by a Rosetta run, repeated twice: MDFF1−Rosetta1−MDFF2−Rosetta2−MDFF3−Rosetta3. All MDFF runs were short (1 ns (step 1) followed by 0.2 ns of minimization into the density map (step 2)). After each MDFF run two models were picked to transition into the following Rosetta round. For this work we picked the final models after each of the two steps of the MDFF protocol: (a) the model after 1 ns of MDFF simulation with modest density map scaling factor and (b) the model after the terminal MDFF minimization into the density map using a high density map scaling factor. For both of these models the regions that agreed least with the density map were identified and rebuilt using Rosetta, followed by an all atom refinement of the models. The Rosetta models were then sorted by score. The five best-scoring models were picked as input into the next MDFF round. Thus, for MDFF2 and MDFF3, a total of five models underwent system preparation and MDFF simulations. Hence, Rosetta2 and Rosetta3 rounds were based on a total of 10 starting models each. Final results are reported after the third round of Rosetta. All reported RMSDs were calculated over the protein backbone atoms N, C_{α}, C, and O. The BCL::Quality application[11b] was used for all RMSD calculations. The SCit Web server was used for the rotamer analysis (http://bioserv.rpbs.jussieu.fr//cgi-bin/SCitCompare[23]) with an angular deviation limit of 40°.

■ RESULTS AND DISCUSSION

Iterative MD−Rosetta Protocol Successful in Refining All Four Benchmark Proteins without User Intervention. An improved iterative cryoEM-guided MD−Rosetta protocol aimed at refining proteins to near-native resolution is presented here. The rationale for iteratively using two orthogonal

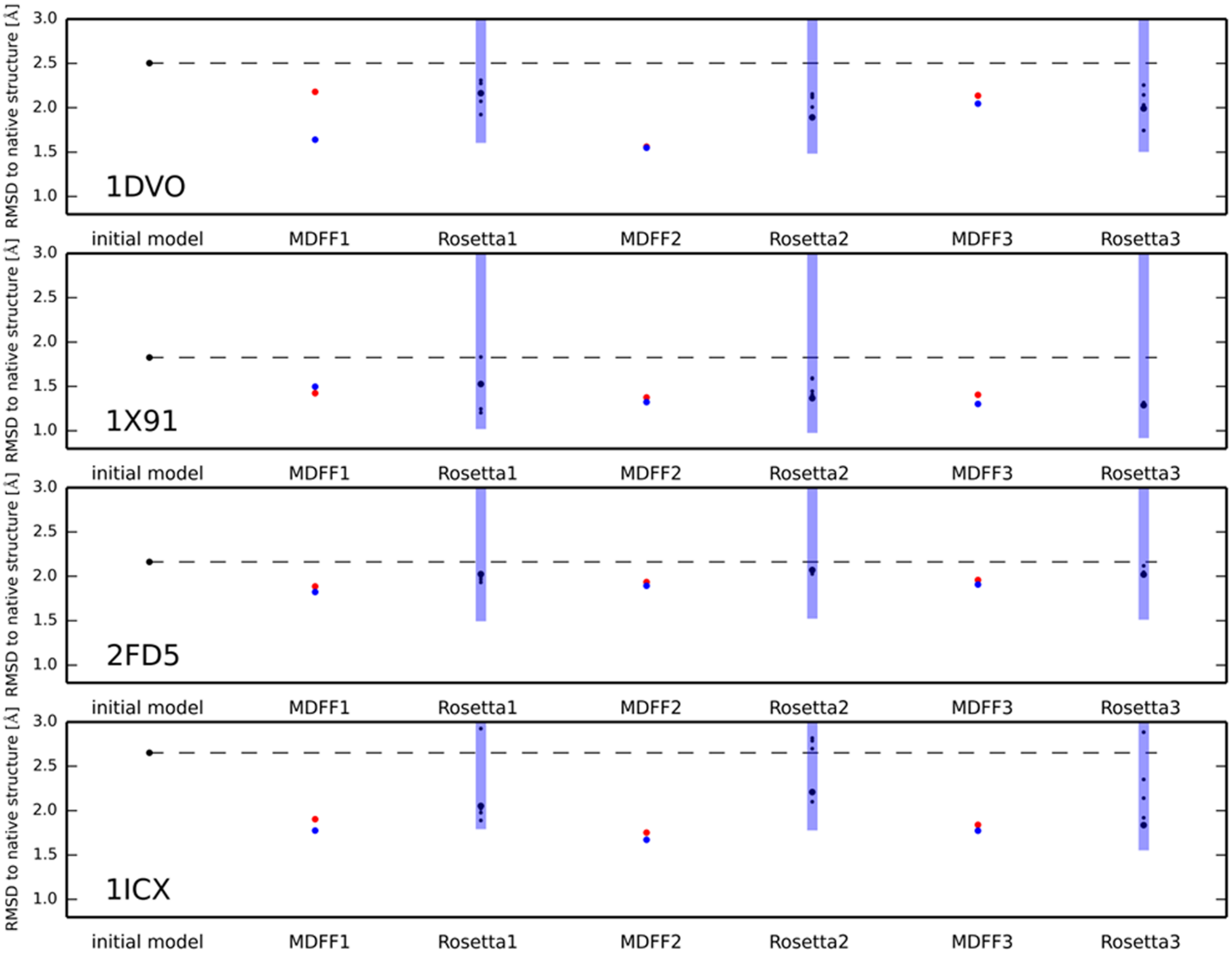


Figure 2. Summary of the results of all rounds of the iterative MDFF—Rosetta refinement. The RMSD values over all protein residues with respect to the native structure in all three rounds of iterative cryoEM-guided protein structure refinement are shown. The first panel (black dot, labeled initial model) shows the top-scoring model of the last round of Rosetta-only refinement in 15e. A trend line has been added to facilitate comparison to the initial model. For the MDFF simulations, final models after each of the two separate steps (MD simulation with low density forces (red) and subsequent minimization with high forces (blue)) are shown. For each of the three rounds of Rosetta refinement, the RMSD values of the top scoring model (thick black dot) and the subsequent four best-scoring models (four thin black dots) are shown. A blue bar indicates the range of the RMSDs of all models built during that round of Rosetta refinement (irrespective of their individual scores), with the low-RMSD end of the bar corresponding to the lowest-RMSD model built. The high-RMSD end of the bar is beyond the plotting limit of 3 Å for all proteins and rounds.

sampling and scoring methods is to symbiotically leverage both methods' strengths and increase sampling efficiency in high-resolution protein structure refinement. The original implementation of the iterative protocol was described in ref 18, and several main improvements are presented in this work. The starting point for the iterative protocol was the best-scoring Rosetta structure from a previous benchmark.[15e] The protocol then performs three rounds of iterative cryoEM-guided molecular dynamics (MDFF) followed by cryoEM-guided Rosetta refinement. The MDFF simulations contained two separate steps (one with low density forces and a second one with high forces), and the final models after each of these steps were chosen as input into a Rosetta loop rebuilding and all-atom refinement run guided by the cryoEM density map. After each round of Rosetta refinement, the five best-scoring models were extracted and were used as starting models in the subsequent MDFF simulations. Thus, during the first round of MDFF, simulations were run on one model, while simulations were run on five models during MDFF2 and MDFF3. Similarly, Rosetta runs were started from two different models for Rosetta1, while ten starting models were used during Rosetta2 and Rosetta3. Success of the protocol is quantified by improvement in RMSD of the models built with respect to the native structure. Importantly, all starting structures resulted from several rounds of cryoEM-guided Rosetta-only refinement in which model improvement had completely seized.[15e] Thus, any improvement in RMSD in the current benchmark can be attributed to the iterative combination of MDFF and Rosetta.

Figure 2 summarizes the results of the entire refinement. It shows the RMSD over all protein residues with respect to the native structure after each of the rounds of the iterative cryoEM-guided protein structure refinement. For the MDFF simulations, final models after each of the two separate steps (MD simulation with low density forces (red) and subsequent minimization with high forces (blue)) are shown. In virtually all the cases, the subsequent minimization did at least slightly improve the RMSD of the final model (blue points generally have slightly lower RMSD than red points). For each of the three rounds of Rosetta refinement, the RMSD of the top scoring model (thick black dot) and the subsequent four best-

Table 1. RMSDs of the Models Built with Respect to Native Structure over All Residues and over All Residues in Secondary Structure Elements (in Parentheses)[I]

protein	start[a]	MDFF1[b]	Rosetta1[c]	MDFF2[d]	Rosetta2[e]	MDFF3[f]	Rosetta3[g]	best[h]
1X91	1.82(1.19)	1.42(0.88)	1.53(0.94)	1.38(0.89)	1.37(0.85)	1.41(0.92)	1.29(0.79)	0.98(0.74)
1DVO	2.50(1.65)	2.18(1.28)	2.16(1.19)	1.56(0.99)	1.89(1.01)	2.14(1.17)	2.14(0.98)	1.54(1.01)
1ICX	2.65(2.14)	1.90(1.45)	2.05(1.72)	1.75(1.38)	2.21(1.87)	1.84(1.39)	1.80(1.37)	1.80(1.37)
2FD5	2.16(1.46)	1.89(1.17)	2.03(1.26)	1.94(1.19)	2.07(1.31)	1.96(1.27)	2.01(1.38)	1.55(1.29)

[a] RMSDs of the starting models. [b] RMSDs of the models after 1 ns of the first round of MDFF (step 1, density map scaling factor 0.3). [c] RMSDs of the top scoring model after the first round of Rosetta refinement. [d] RMSDs of the models after 1 ns of the second round of MDFF (step 1, density map scaling factor 0.3). [e] RMSDs of the top scoring model after the second round of Rosetta refinement. [f] RMSDs of the models after 1 ns of the third round of MDFF (step 1, density map scaling factor 0.3). [g] RMSDs of the top scoring model after the third round of Rosetta refinement. [h] RMSDs of the best Rosetta models ever built during the iterative MD−Rosetta refinement protocol. [I] All RMSDs shown are in Å.

scoring models (thin black dots) are shown. Those are the five models that were used as input in the following MDFF round. A blue bar indicates the range of the RMSDs of all models built. The low-RMSD end of that bar corresponds to the lowest-RMSD model built during that round of Rosetta. In summary, for all four proteins, models significantly lower in RMSD than the initial model were built over the course of the protocol. For three of the four proteins (1X91, 2FD5, 1ICX) there is a gradual improvement in model quality over the three rounds of the iterative refinement. For 1DVO, the third round of the protocol did undo some of the improvements from the previous rounds. But even for 1DVO, the second round MDFF model has its RMSD improved by more than 1 Å compared to the initial starting structure. Also, the quality of the SSE parts of the 1DVO model improved through all three rounds. For all proteins, the biggest improvement is seen in the first round of MDFF. However, the subsequent rounds of MDFF generally still improve model quality. An in-depth analysis showed that none of the five top-scoring models in any Rosetta run (for any of the proteins) did originate from a model that underwent strong MDFF density minimization (blue dots). This is important for two reasons. First, it suggests that density map scaling factors of 10 (and even 5; we did lower the scaling factor over the course of the benchmark) perturb the protein structure too much for Rosetta to be able to still build well-scoring models. Second, in the future it may not be necessary to even transition the MDFF step 2 models into Rosetta. This will cut the computational cost of the protocol in half without affecting the performance at all.

Supplemental Figure 1 shows an equivalent plot of the results of the entire refinement but with the RMSD values calculated just over SSEs in the proteins. The model progression is less steady when only secondary structure elements are considered. Particularly, the Rosetta refinement and subsequent model selection more frequently seemed to increase the model RMSD. This can be understood when considering that the Rosetta protocol is aimed at improving model quality in loop regions. Sections of the models that agree least with the density map of the protein are identified and rebuilt. These sections tend to be in loop regions. The final model selection is based on the Rosetta energy function scoring all protein residues. Thus, less favorable conformations within SSEs may be selected at the expense of better agreement in loop regions. Nonetheless, for two of the four benchmark proteins (1DVO and 1X91), a more or less steady quality improvement is observed even when the RMSD values are calculated just over SSEs in the proteins.

Table 1 quantizes the RMSD values of the generated models for all four proteins throughout all three rounds of the iterative protocol. Three of the four proteins exhibited excellent overall improvement. The RMSD of 1ICX improved by 0.85 Å (0.77 Å measured over residues in secondary structure elements), and the RMSD of 1X91 improved by 0.53 Å (0.40 Å measured over residues in secondary structure elements). While the RMSD of 1DVO improved by a respectable 0.36 Å over all residues of the protein, it did improve by 0.77 Å measured over residues in secondary structure elements. The native structures relaxed in the Rosetta force field exhibit RMSDs of around 0.5−0.8 Å, thus the observed improvement in the protocol corresponds to around 30−60% of the maximally possible improvement. Two of the final best-scoring Rosetta models (1X91 and 1DVO) have sub-Ångstrom RMSD values when measured over residues in secondary structure elements − arguably the most important core part of the protein. During the course of the protocol, models were built that had an even higher quality than the best-scoring models. Those models showed improvement of almost 1 Å over only three rounds of iteration. We were even able to build a sub-Ångstrom RMSD model for 1X91 (as measured over all 153 residues). This demonstrates the power of the iterative MDFF−Rosetta protocol to build models of excellent quality that would not have been generated by any of the individual methods.

To visualize the quality of the generated models, Figure 3 and Supplemental Figure 2 show the best models for all four benchmark proteins overlaid with their native structure. Both the backbone (ribbon-only representation) and side chain agreement are shown. For proteins of 150 to 180 residues, RMSD-to-native values of around 1 Å over SSE residues represent a virtually correct backbone prediction. Some deviations in loop regions still exist, but given the more flexible nature of those regions, this is not unexpected. The side chain coordinate prediction is excellent particularly in secondary structure elements. To quantify the side chain prediction accuracy, a rotamer analysis was performed. Between 57% (1DVO) and 69% (1X91) of the χ_1 rotamers were predicted correctly. Within the core of the proteins these values were as high as 77%.

The successful model refinement for 1ICX is of particular interest since it is the α−β-protein in the benchmark. The previous implementation of the method was only successful in iteratively refining α-helical proteins.[18] This suggests that cryoEM-guided sampling in all steps of the protocol can be particularly useful for accurately refining β-sheets. In summary, the novel implementation of the iterative Rosetta−MD protein structure refinement protocol has been more successful than the previous version. This is particularly notable since the RMSD-based model selection in the MD stage of the protocol

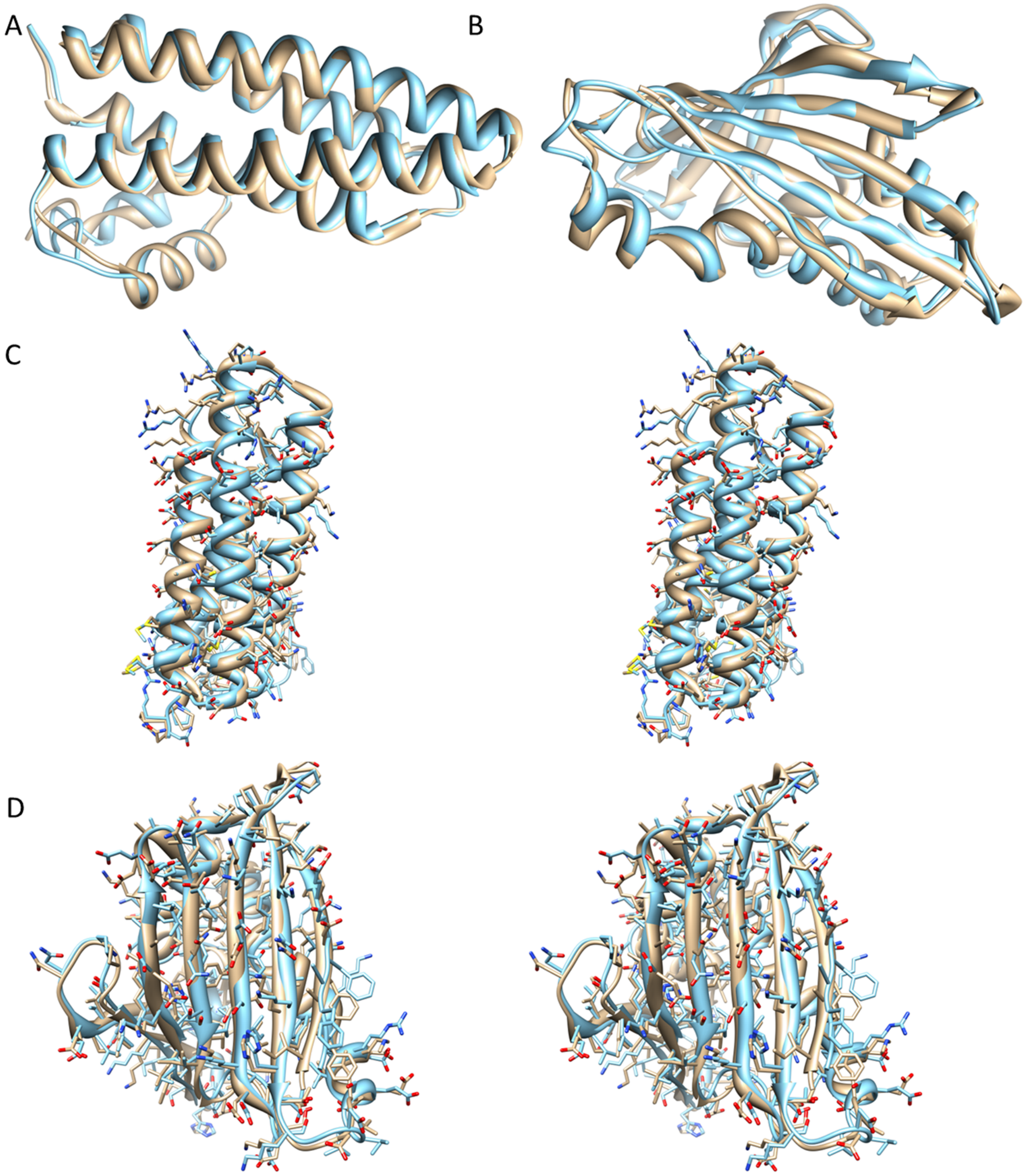


Figure 3. Lowest RMSD models after three rounds of iterative MDFF−Rosetta refinement for 1X91 (panels A and C) and 1ICX (panels B and D). The native structure is shown in gold, while the model is shown in turquois. A, B) Ribbon backbone representation of the proteins. The overall structure within secondary structure elements has been recovered in the models. C, D) Stereoviews of side chain non-hydrogen coordinates are shown in addition to ribbon backbone representation. Most side chain conformations within the interface of secondary structure elements have been built correctly.

has been abolished. The current protocol can be used without any previous knowledge of the native protein structure.

MDFF Routinely Builds Improved Models. As part of the iterative cryoEM-guided Rosetta−MD protein structure refinement protocol, three rounds of MDFF simulations were run for the benchmark proteins. The simulation length was kept to 1 ns (plus an additional 0.2 ns of MDFF minimization with high density map scaling factor) based on testing and previous results[18] which suggested that the proteins exhibit their largest improvements within the first 0.5 ns of MD simulation. As an example, Figure 4 shows the evolution of the model quality of 1X91 during all three rounds of MDFF. At the start of the simulations, the protein quickly improves its RMSDs compared to that of the previous-round Rosetta model (red lines) as the agreement with the density map improves. This effect is most pronounced in the first round of MDFF when an improvement in RMSD to the native structure of about 0.3 Å is seen within only 100 ps of MDFF. While in a previous implementation of the protocol, models had to be picked based on RMSD, it is possible to now simply pick the last model at the end of each step of the MDFF simulation (i.e., one model after 1 ns and another model after 1.2 ns) and still see consistent model quality improvement. Despite the fact that the final models generally have improved RMSDs (with respect

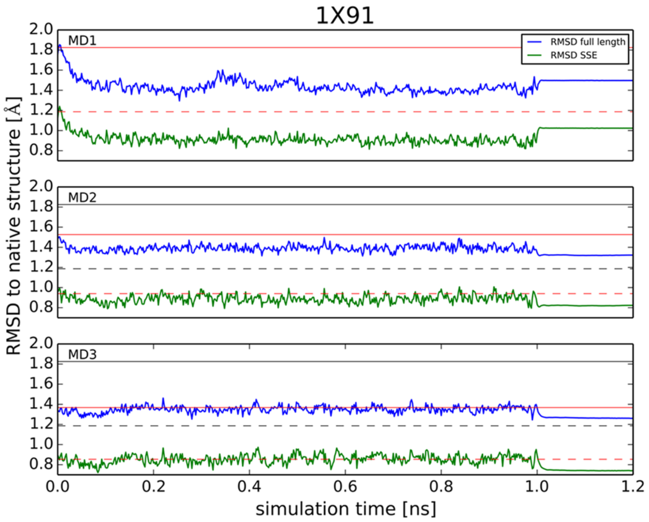


Figure 4. Model quality evolution of 1X91 during the three rounds of MD. The RMSD of the MD structure with respect to the native model is shown for all protein residues (blue) and for residues in secondary structure elements (green). RMSDs of specific reference models are displayed by vertical lines: the full length RMSD of the starting model (black line), the RMSD over SSEs of the starting model (dashed black line), the full length RMSD of the best scoring model from the previous Rosetta round (red line), and the RMSD over SSEs of the best scoring model from the previous Rosetta round (dashed red line). For the first round of MD, the red and black lines coincide. Success in this stage of the protocol is characterized by the blue line breaking through the red line (corresponding to MD sampling lower RMSD models than the best scoring model seen in the last Rosetta round) and the green line breaking through the dashed red line.

to the starting model in that round of MDFF), there are still conformations sampled during the MDFF simulations that have lower RMSDs. As long as no reliable method for identification of those models exists, this constitutes untapped potential. Compared to results without density map guidance, notable progress has been achieved in the improvement of the full length RMSD, something that was previously routinely not achieved during the second and third round of MD refinement. Lastly, as can be seen from Figure 4, the MDFF refinement is still clearly most effective at improving segments of the proteins in secondary structure elements.

Rosetta Refinement Most Powerful at Improving Model Quality in Loop Regions. Figure 2 displayed the RMSDs of the five top-scoring models and the lowest RMSD model during each round of Rosetta refinement. For a more in-depth analysis of the Rosetta refinement performance, it is necessary to consider RMSD vs Rosetta energy score plots for all rounds of refinement. Figure 5 shows the RMSD vs score plots of all four benchmark proteins (panels A through D). For each protein the RMSD vs score distributions are plotted in rounds 1, 2, and 3 of the iterative protocol (the lower three subpanels). Additionally RMSD vs score distributions from the final round of a Rosetta-only refinement of the same proteins[15e] is shown in the top subpanel. Those can be considered the limit of the protocol without the use of iterative MDFF simulations. Additionally, the native structure is plotted for reference. The Rosetta refinement protocol can only be successful if the native structure scores better than models with higher RMSD values. The nonzero RMSD values for the native structures are a result of a relaxation in the Rosetta force field before scoring. Those values (between 0.5 and 0.8 Å for the benchmark proteins) can be considered the positive limit of RMSD refinement in Rosetta. All proteins exhibit a general funnel-shaped RMSD vs score distribution. Models with RMSD values larger than 0.5−1 Å compared to the best models built all have higher (more unfavorable) scores than low-RMSD models. The funnel shape is less pronounced within the 0.5−1 Å RMSD range of the best models built. This is particularly true for 1X91. The final RMSD vs score distributions are significantly shifted to lower RMSD value (and also lower scores) compared to the initial starting distributions (upper panel). This demonstrates overall success of the protocol to improve model quality. While a large fraction of the Rosetta models have higher RMSD than the previous round MDFF models, the RMSD vs score distributions improve with each round, due to iterative orthogonal sampling by molecular dynamics. This suggests that the rounds of iterative Rosetta−MD refinement do indeed gradually improve the quality of the protein models. For 1DVO, 1ICX, and 2FD5, the native models (relaxed and scored in the Rosetta force field) score better than any refined protein models, indicating that there is indeed still room for model improvement. However, for 1X91 there are about 800 Rosetta3 models that score better than any of the realxed native models. This may explain the disappearance of a funnel shape for low RMSD models for this protein since the scoring function cannot distinguish native from native-like models any more. More importantly, this may also indicate a limitation to what model quality can be achieved with the iterative Rosetta−MD protocol. Potentially, the cryoEM-guided Rosetta scoring function is only able to refine models up to about 1 Å RMSD. The protocol converges after three iterative MD−Rosetta rounds. An additional fourth round of MDFF did not improve the proteins considerably.

■ CONCLUSIONS

Here we presented the results of an improved iterative MD−Rosetta protocol to computationally refine protein structures guided by medium resolution cryoEM density maps. The presented protocol is tailored toward high resolution structure refinement if native-like starting models exist. Molecular dynamics flexible fitting and Rosetta were used iteratively to improve the model quality of the benchmark proteins. All four benchmark proteins exhibited improvement of up to 1 Å in RMSD over only three rounds of iteration. This work was based on the original idea of an iterative MD−Rosetta protein structure refinement protocol,[18] where we demonstrated that such a combination of MD and Rosetta could indeed help to overcome some of the "conformational traps" in which cryoEM-guided Rosetta refinement may have been trapped. Several improvements over the original implementation were presented. Particularly, all limitations discussed in ref 18 were addressed here. First, both the MD and Rosetta part of the protocol are guided by medium-resolution cryoEM density maps now, allowing the method to leverage the full potential from the sparse experimental data. Second, and more importantly, the need to cherry-pick MD models based on RMSD has been obviated. Models are now picked from the end of the MDFF simulations regardless of their RMSD with respect to the native structure. Furthermore, not only the top scoring model from each Rosetta round enters the next MDFF stage, but rather the five best-scoring models are chosen. All these method enhancements did also improve the success rate of the protocol, so that now all benchmark proteins showed improvements. Lastly, while the benchmark set only contained

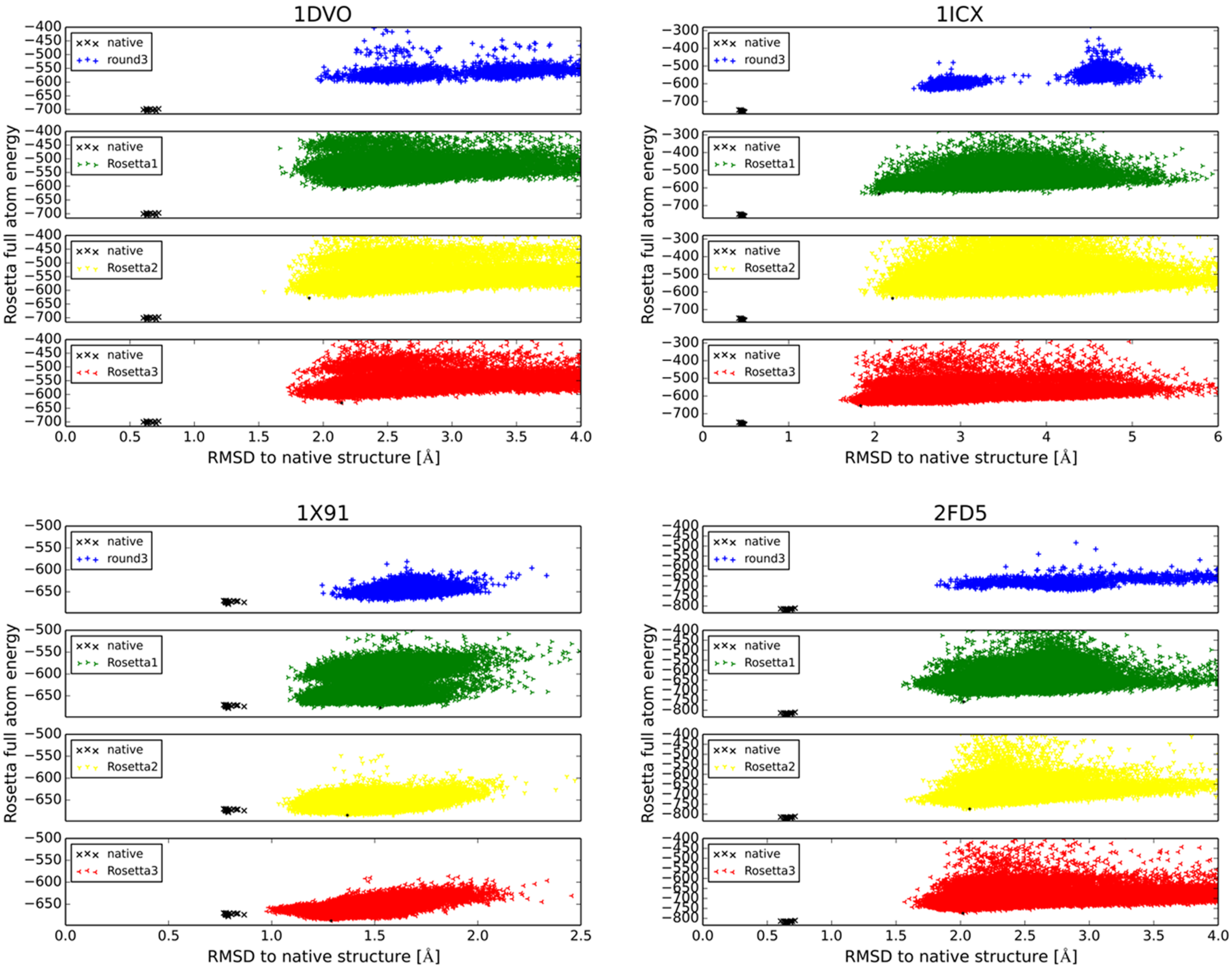


Figure 5. RMSD vs score plots for all four benchmark proteins. The first panel (blue, labeled round3) shows the results of the last round of Rosetta-only refinement in ref 15e. The other three panels show the results for the first (green), second (yellow), and third (red) Rosetta round of the iterative MDFF−Rosetta protocol, respectively. The native structure, relaxed in the Rosetta force field, is shown in all panels (black). The best-scoring structure in each of the three rounds of the iterative MDFF−Rosetta protocol is shown as a solid black dot.

one α−β-protein, this protein was not only successfully refined but did show the most significant improvement of all proteins in the benchmark (the RMSD of 1ICX improved by 0.85 Å when measured over all protein residues and by 0.77 Å when measured over residues in secondary structure elements). This demonstrates a promising extension of the scope of the protocol to β-sheet containing proteins. In summary, the improved protocol is considerably better than the original iterative MD−Rosetta protocol. The protocol converges after three iterative MD−Rosetta rounds. An encouraging symbiosis between MDFF and Rosetta was observed, in that MDFF was best at improving the RMSDs over residues in secondary structure elements, while Rosetta improvements were greatest in loop regions. This suggests that each method has its strength in an area where the other method faces challenges. Used in an iterative fashion, MDFF and Rosetta can thus contribute constructively to overcome sampling limitations of the individual methods. While these results are very encouraging, room for future methodology improvement remains. Future work will focus on optimizing the role of the density map scaling factor in MDFF, as it pertains to the successful transition of models into Rosetta. We will also focus on identifying lower RMSD models built by MDFF.

■ ASSOCIATED CONTENT

Supporting Information

Figure 1 shows the summary of the results of all rounds of the iterative MDFF−Rosetta refinement as measured over SSEs. Figure 2 shows the lowest RMSD models after three rounds of iterative MD/Rosetta refinement for proteins 2FD5 and 1DVO. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 858-534-2913. Fax: 858-534-4974. E-mail: slindert@ucsd.edu. Corresponding author address: Department of Chemistry & Biochemistry, University of California San Diego, 9500 Gilman Drive, Mail Code 0365, La Jolla, CA 92093-0365.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health, the National Science Foundation, the Howard Hughes Medical Institute, the National Biomedical Computation Resource, and the NSF Supercomputer Centers. Computational resources were supported, in part, by the National Science Foundation grant PHY-0822283, the Center for Theoretical Biological Physics. S.L. was supported by the American Heart Association (12POST11570005) and the Center for Theoretical Biological Physics.

■ REFERENCES

(1) (a) Schafer, N. P.; Hoffman, R. M.; Burger, A.; Craig, P. O.; Komives, E. A.; Wolynes, P. G. Discrete kinetic models from funneled energy landscape simulations. *PLoS One* **2012**, *7*, e50635. (b) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 2932−7.

(2) Lee, J.; Wu, S.; Zhang, Y. *Ab initio protein structure prediction. From Protein Structure to Function with Bioinformatics, Chapter 1*; Rigden, D. J., Ed.; Springer: London, 2009; pp 1−26.

(3) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740−4.

(4) (a) Adam, L. B. In *Folding@home: Lessons from eight years of volunteer distributed computing*; Daniel, L. E., Guha, J., Siraj, K., Vijay, S. P., Eds.; 2009; pp 1−8. (b) Shirts, M.; Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−4. (c) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **2002**, *420*, 102−6.

(5) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91−97.

(6) (a) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−20. (b) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17845−50.

(7) (a) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209−25. (b) Das, R.; Qian, B.; Raman, S.; Vernon, R.; Thompson, J.; Bradley, P.; Khare, S.; Tyka, M. D.; Bhat, D.; Chivian, D.; Kim, D. E.; Sheffler, W. H.; Malmstrom, L.; Wollacott, A. M.; Wang, C.; Andre, I.; Baker, D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **2007**, *69* (Suppl 8), 118−28. (c) Bradley, P.; Misura, K. M.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **2005**, *309*, 1868−71.

(8) (a) Zhang, Y.; Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7594−9. (b) Zhang, Y.; Kolinski, A.; Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* **2003**, *85*, 1145−64. (c) Zhang, Y.; Kihara, D.; Skolnick, J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **2002**, *48*, 192−201. (d) Wu, S.; Skolnick, J.; Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **2007**, *5*, 17.

(9) (a) Kinch, L.; Yong Shi, S.; Cong, Q.; Cheng, H.; Liao, Y.; Grishin, N. V. CASP9 assessment of free modeling target predictions. *Proteins* **2011**, *79* (Suppl 10), 59−73. (b) Tai, C. H.; Bai, H.; Taylor, T. J.; Lee, B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins* **2014**, *82* (Suppl 2), 57−83.

(10) Bonneau, R.; Strauss, C. E.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmstrom, L.; Robertson, T.; Baker, D. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **2002**, *322*, 65−78.

(11) (a) Karakas, M.; Woetzel, N.; Staritzbichler, R.; Alexander, N.; Weiner, B. E.; Meiler, J. BCL::Fold−de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One* **2012**, *7*, e49240. (b) Woetzel, N.; Karakas, M.; Staritzbichler, R.; Muller, R.; Weiner, B. E.; Meiler, J. BCL::Score−knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS One* **2012**, *7*, e49242.

(12) (a) Bowers, P. M.; Strauss, C. E.; Baker, D. De novo protein structure determination using sparse NMR data. *J. Biomol. NMR* **2000**, *18*, 311−8. (b) Meiler, J.; Baker, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 15404−9. (c) Meiler, J.; Baker, D. The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J. Magn. Reson.* **2005**, *173*, 310−6. (d) Rohl, C. A.; Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* **2002**, *124*, 2723−9. (e) Boomsma, W.; Tian, P.; Frellsen, J.; Ferkinghoff-Borg, J.; Hamelryck, T.; Lindorff-Larsen, K.; Vendruscolo, M. Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 13852−7.

(13) (a) Alexander, N.; Bortolus, M.; Al-Mestarihi, A.; McHaourab, H.; Meiler, J. De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure* **2008**, *16*, 181−95. (b) Hanson, S. M.; Dawson, E. S.; Francis, D. J.; Van Eps, N.; Klug, C. S.; Hubbell, W. L.; Meiler, J.; Gurevich, V. V. A model for the solution structure of the rod arrestin tetramer. *Structure* **2008**, *16*, 924−34. (c) Hirst, S. J.; Alexander, N.; McHaourab, H. S.; Meiler, J. RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *J. Struct. Biol.* **2011**, *173*, 506−514.

(14) (a) DiMaio, F.; Echols, N.; Headd, J. J.; Terwilliger, T. C.; Adams, P. D.; Baker, D. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **2013**, *10*, 1102−4. (b) Schroder, G. F.; Levitt, M.; Brunger, A. T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **2010**, *464*, 1218−22. (c) Fenn, T. D.; Schnieders, M. J. Polarizable atomic multipole X-ray refinement: weighting schemes for macromolecular diffraction. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 957−65.

(15) (a) DiMaio, F.; Tyka, M. D.; Baker, M. L.; Chiu, W.; Baker, D. Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* **2009**, *392*, 181−90. (b) Lindert, S.; Hofmann, T.; Wotzel, N.; Karakas, M.; Stewart, P. L.; Meiler, J. Ab initio protein modeling into CryoEM density maps using EM-Fold. *Biopolymers* **2012**, *97*, 669−77. (c) Lindert, S.; Staritzbichler, R.; Wotzel, N.; Karakas, M.; Stewart, P. L.; Meiler, J. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure* **2009**, *17*, 990−1003. (d) Lindert, S.; Stewart, P. L.; Meiler, J. Hybrid approaches: applying computational methods in cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **2009**, *19*, 218−25. (e) Lindert, S.; Alexander, N.; Wotzel, N.; Karakas, M.; Stewart, P. L.; Meiler, J. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure* **2012**, *20*, 464−78. (f) Chan, K. Y.; Trabuco, L. G.; Schreiner, E.; Schulten, K. Cryo-electron microscopy modeling by the molecular dynamics flexible fitting method. *Biopolymers* **2012**, *97*, 678−86. (g) Trabuco, L. G.; Schreiner, E.; Gumbart, J.; Hsin, J.; Villa, E.; Schulten, K. Applications of the molecular dynamics flexible fitting method. *J. Struct. Biol.* **2011**, *173*, 420−7. (h) Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **2008**, *16*, 673−83.

(16) (a) Lawson, C. L.; Baker, M. L.; Best, C.; Bi, C.; Dougherty, M.; Feng, P.; van Ginkel, G.; Devkota, B.; Lagerstedt, I.; Ludtke, S. J.; Newman, R. H.; Oldfield, T. J.; Rees, I.; Sahni, G.; Sala, R.; Velankar, S.; Warren, J.; Westbrook, J. D.; Henrick, K.; Kleywegt, G. J.; Berman, H. M.; Chiu, W. EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **2011**, *39*, D456−64. (b) Esquivel-Rodriguez, J.; Kihara, D. Computational methods for constructing protein structure models from 3D electron microscopy maps. *J. Struct Biol.* **2013**, *184*, 93−102.

(17) Baker, M. L.; Ju, T.; Chiu, W. Identification of secondary structure elements in intermediate-resolution density maps. *Structure* **2007**, *15*, 7−19.

(18) Lindert, S.; Meiler, J.; McCammon, J. A. Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol to Improve Model Quality. *J. Chem. Theory Comput.* **2013**, *9*, 3843−3847.

(19) Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **2009**, *49*, 174−80.

(20) MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257−65.

(21) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−802.

(22) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(23) Gautier, R.; Camproux, A. C.; Tuffery, P. SCit: web tools for protein side chain conformation analysis. *Nucleic Acids Res.* **2004**, *32*, W508−11.