JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

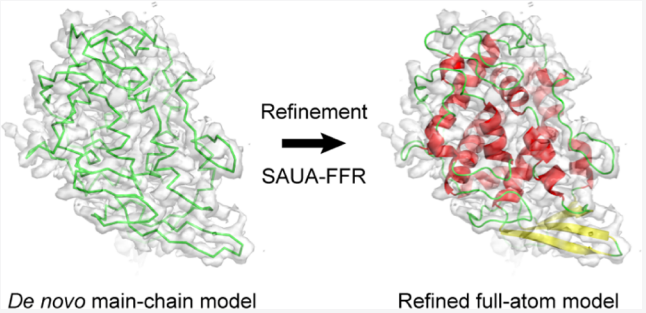
Efficient Flexible Fitting Refinement with Automatic Error Fixing for De Novo Structure Modeling from Cryo-EM Density Maps

Takaharu Mori,* Genki Terashi, Daisuke Matsuoka, Daisuke Kihara, and Yuji Sugita

Cite This: *J. Chem. Inf. Model.* 2021, 61, 3516−3528 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Structural modeling of proteins from cryo-electron microscopy (cryo-EM) density maps is one of the challenging issues in structural biology. De novo modeling combined with flexible fitting refinement (FFR) has been widely used to build a structure of new proteins. In de novo prediction, artificial conformations containing local structural errors such as chirality errors, cis peptide bonds, and ring penetrations are frequently generated and cannot be easily removed in the subsequent FFR. Moreover, refinement can be significantly suppressed due to the low mobility of atoms inside the protein. To overcome these problems, we propose an efficient scheme for FFR, in which the local structural errors are fixed first, followed by FFR using an iterative simulated annealing (SA) molecular dynamics protocol with the united atom (UA) model in an implicit solvent model; we call this scheme "SAUA-FFR". The best model is selected from multiple flexible fitting runs with various biasing force constants to reduce overfitting. We apply our scheme to the decoys obtained from MAINMAST and demonstrate an improvement of the best model of eight selected proteins in terms of the root-mean-square deviation, MolProbity score, and RWplus score compared to the original scheme of MAINMAST. Fixing the local structural errors can enhance the formation of secondary structures, and the UA model enables progressive refinement compared to the all-atom model owing to its high mobility in the implicit solvent. The SAUA-FFR scheme realizes efficient and accurate protein structure modeling from medium-resolution maps with less overfitting.



INTRODUCTION

Single-particle cryo-electron microscopy (cryo-EM) is a powerful tool to determine the three-dimensional (3D) structures of biomolecules at near-atomic resolution.[1] In the method, a 3D density map of the target molecule is reconstructed from a large number of 2D images of the molecule. Owing to the development of various technologies, such as efficient sample preparation, direct electron detection, and software for image processing,[2] high-resolution analyses have been realized for large protein complexes and membrane proteins.[3−5] The method also enables us to understand protein dynamics by capturing snapshots of the structures in their biological processes, such as gene transcription[6] and substrate transport.[7] Although atomic resolution has been recently achieved in some cases,[8,9] typical resolution is still 3−5 Å due to the intrinsic flexibility of proteins in solution. Thus, reliable structure modeling from low- or medium-resolution maps is one of the essential issues in structural molecular biology.

Structure modeling from cryo-EM density maps is usually conducted with computational techniques such as rigid-body docking, flexible fitting, and de novo modeling.[10−13] In rigid-body docking, the entire protein structure is treated as an assembly of component segments, and the positions and orientations of each component are optimized with rigid-body translations and rotations (6D search) to fit the density map.[14−16] Flexible fitting uses a complete model of the target biomolecule. The initial structure, which is typically determined with other methods such as X-ray crystallography, nuclear magnetic resonance (NMR), or homology modeling, is deformed using normal mode analysis (NMA),[17,18] molecular dynamics (MD) simulations,[19−25] or their hybrid approach.[26] De novo modeling predicts the structure from the density map and amino acid sequence information. To date, various methods, including Rosetta,[27] EM-Fold,[28] Pathwalking,[29] and MAINMAST,[30] have been developed. Rosetta constructs a full-atom model based on the fragment assembly algorithm, in which the predicted short fragments are assembled to fit the density map using a 6D search. EM-Fold builds α-helical proteins by placing α-helices on rod-shaped densities in the map based on a Monte Carlo search. Pathwalking traces the Cα atoms in the density map using a traveling salesman problem solver, while

Received: February 27, 2021
Published: June 18, 2021

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

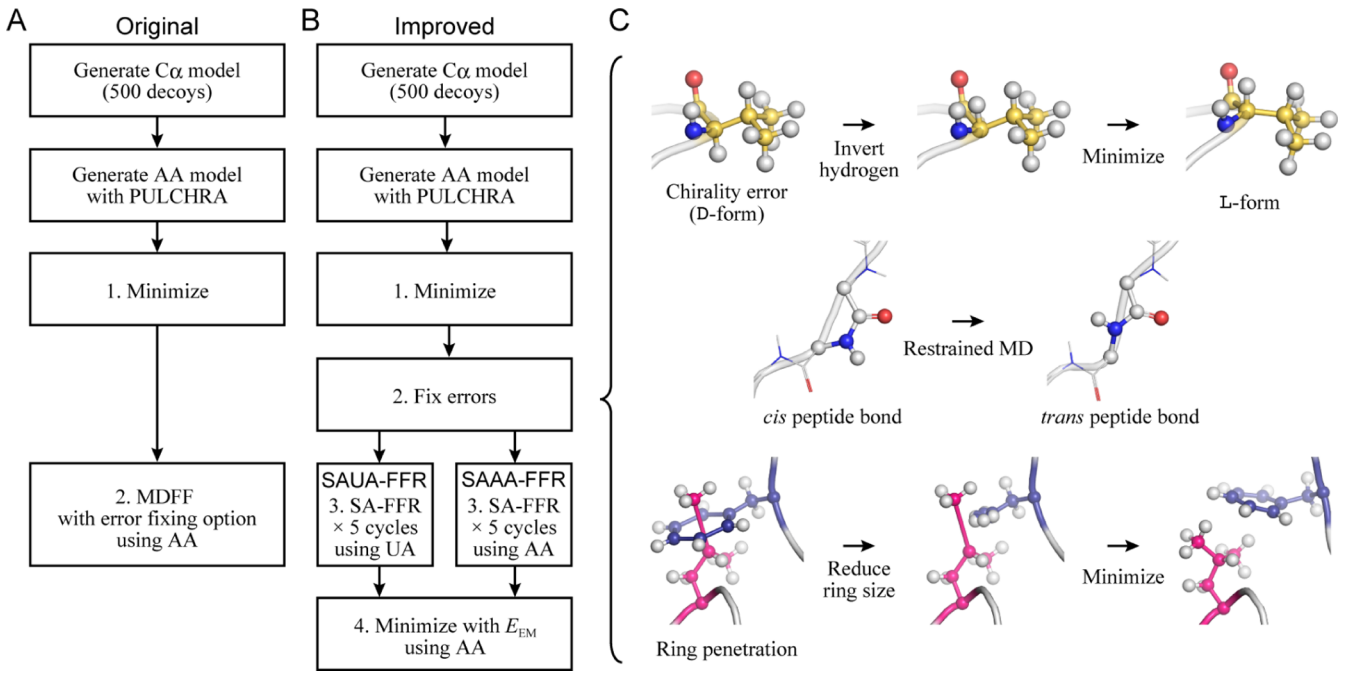


Figure 1. Flowchart of the FFR. (A) Original scheme in MAINMAST, (B) improved scheme proposed in this study (SAUA-FFR and SAAA-FFR), and (C) protocol to fix chirality errors, cis peptide bonds, and ring penetrations in step 2 of the SAUA-FFR or SAAA-FFR.

MAINMAST employs a minimum spanning tree algorithm and tabu search.

The model obtained from de novo modeling is usually refined with MD-based flexible fitting (flexible fitting refinement; FFR) to remove steric clashes or repack the side chains. In the method, biasing potential that guides the structure toward the target density is added to the molecular mechanics force fields. One of the popular methods is cross-correlation coefficient (c.c.)-based flexible fitting, which uses the c.c. between the experimental and simulated density maps in the biasing potential.[19,21] On the other hand, Molecular Dynamics Flexible Fitting (MDFF) introduces a biasing potential that is proportional to the Coulomb potential derived from the experimental density map and also the secondary structure (SS) restraint potential.[20] MAINMAST predicts the Cα model, which is further converted to an all-atom (AA) model with PULCHRA,[31] followed by the refinement with MDFF.[30] The iterative MD-Rosetta protocol, which iteratively performs Rosetta loop prediction and MDFF, has been proposed to improve the quality of the model predicted from EM-Fold.[32,33] Enhanced sampling algorithms such as the replica-exchange method[34] have been widely employed to search the global-energy-minimum structure in flexible fitting.[16,35,36] The CryoFold algorithm[37] first performs the model building using targeted MD[38] combined with Bayesian-inference-based restraints (MELD),[39] which utilizes the Cα atom positions and Cα−Cα distance information obtained from MAINMAST, and then refines the model with resolution-exchange MDFF (ReMDFF).[35] Although replica-exchange schemes with the AA model could provide an accurate structure model compared to the other conventional schemes, they must employ multiple replicas of the target system, resulting in large computational cost. Simple protocols with low computational cost that maintain high accuracy should be needed for efficient structure modeling.

In structure refinement, another important task besides the global-energy-minimum search is the removal of local structural errors such as chirality errors and cis peptide bonds. These errors are frequently generated in de novo structure modeling, especially when a full-atom model is constructed after the main-chain modeling such as MAINMAST and Pathwalking. The errors can be removed using energy minimization or MD simulation with error-fixing restraints such as the dihedral angle restraint at ω = 180° for the cis peptide bond. Many model building tools provide some functions for automatic or manual modifications of the errors, including geometry restraint or inversion of the corresponding atoms.[40−42] In particular, ISOLUDE performs on-the-fly flexible fitting, where the errors visualized in the monitor can be manually removed with a mouse or haptics tool.[43] Another troublesome error is ring penetration, where a covalent bond penetrates an aromatic ring. This situation can accidentally occur, especially when a coarse-grained (CG) model is converted to an AA model.[44] In fact, PULCHRA can generate penetrated rings, even though the algorithm tries to minimize the possibility of occurrence of ring penetration.[31] Because ring penetration is difficult to solve, an effective algorithm that automatically detects or fixes such errors should be developed.

To solve these problems, we propose an efficient flexible fitting scheme for refining the decoys obtained from de novo modeling, where MD-based flexible fitting with an iterative simulated annealing (SA) protocol is conducted, during which the united atom (UA) model and implicit solvent model are employed; this scheme is, therefore, called "SAUA-FFR". The UA model, which incorporates hydrogen atoms of CH_3, CH_2, and CH groups into the carbon atoms, can maintain the atomic resolution, and the implicit solvent model considers solvent effects with low computational cost. We use the decoys obtained from MAINMAST.[30] The local structural errors in the decoys are automatically fixed using new functions implemented in MD software GENESIS,[45,46] which can address ring penetrations, cis peptide bonds, and chirality errors. Our refinement scheme is compared with the original scheme of MAINMAST using eight

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

Table 1. Summary of the Target Systems

protein	EMD ID	res. (Å)	PDB ID	α-helix/β-sheet	chain/residues	rmsd (Å)[a]
FrhA	2513	3.36	4CI0	181/58	A/2−386	3.80
PCS	3231	3.6	5FMG	56/41	K/2−195	15.00
SV	5495	3.5	3J26	29/151	A/1−508	9.46
BPP-1	5764	3.5	3J4U	65/49	A/5−331	33.10
TRPV1	5778	3.275	3J5P	220/0	A/381−719	6.04
MAVS	5925	3.64	3J6J	68/0	A/1−97	3.34
BmCPV-1	6374	2.90	3JB0	103/37	D/1−292	1.67
PCV2	6555	2.90	3JCI	0/86	A/42−231	2.32

[a]The Cα rmsd with respect to the native structure calculated with the MMTSB toolset (*rms.pl*).[58]

selected proteins: F420-reducing hydrogenase α subunit (FrhA), 20S proteasome core subunit (PCS), Sputnik virophage (SV), Bordetella phage (BPP-1), transient receptor potential vanilloid 1 (TRPV1), CARD domain of mitochondria antiviral signaling protein (MAVS), bombyx mori cypovirus 1 (BmCPV-1), and porcine circovirus 2 (PCV2). The models refined with our scheme are also compared with those refined with the Phenix *real_space_refine* tool. The results demonstrate that our scheme can achieve progressive formation of SSs due to the high mobility of atoms in the UA model, realizing efficient FFR.

METHODS

MAINMAST. MAINMAST is a powerful method for de novo modeling.[30] In the method, a tree structure is first constructed by connecting local dense points in the density map (minimum spanning tree). The structure is further refined with a tabu search to find the longest pathway, which corresponds to the main chain of the protein. The amino acid sequence is aligned to the pathway by evaluating the matching between local densities in the experimental map and those predicted from each amino acid on the path, and then, the Cα model is constructed. Finally, the AA model is generated from the Cα model, and it is subjected to FFR. Previous work demonstrated good performance [average root-mean-square deviation (rmsd) = 2.68 Å)] for the selected 30 proteins with 2.6−4.8 Å resolution maps.[30]

Figure 1A illustrates the flowchart of the original scheme in MAINMAST, focusing on the protocol after generating the Cα model. First, 500 possible models (decoys) are generated, and then, the Cα model is converted to the AA model using PULCHRA.[31] Each model is subjected to energy minimization, followed by refinement with MDFF.[10,11] In the refinement, only a single run is carried out at 300 K with *g*-scale 0.5. The restraints that maintain the trans peptide bond and proper chirality are employed, while the restraints of the SS are not applied in the system. The best model is selected from the 500 models according to the MDFF energy, which is composed of the EM biasing potential and restraint energy for fixing chirality errors and cis peptide bonds.

The SAUA-FFR Scheme. In this study, we propose an efficient scheme for this FFR method (namely, SAUA-FFR) using the decoys obtained from MAINMAST, which is mainly composed of four steps after generating the AA model (Figure 1B, left scheme). In step 1, energy minimization is carried out to remove steric clashes in the initial decoy. In step 2, energy minimization and restrained MD simulation are further carried out to fix local structural errors such as chirality errors, cis peptide bonds, and ring penetrations, where the specific treatment and restraints are applied to the errors (Figure 1C; for details, see the next paragraph). In step 3, FFR is carried out using the UA model in an implicit solvent model, where the SA MD is iterated five times. In step 4, the UA model is converted to the AA model by generating hydrogen atoms, and the FFR with the same EM biasing potential is carried out once in the implicit solvent. In steps 3 and 4, c.c.-based flexible fitting is employed, which introduces the biasing potential $E_{EM} = k\ (1 - \text{c.c.})$ with the force constant k. Here, various force constants ranging from low to high values (N force constants) are examined to generate a "pool" containing strongly or weakly fitted structures because the optimal value for the force constant is unknown. Thus, the pool contains $500 \times N$ decoys in total. Any restraints except for the EM biasing potential are not applied in the system. The best model is selected from the $500 \times N$ decoys based on three validation scores: the c.c. between the experimental and simulated density maps, the RWplus score,[47] and the MolProbity score[48] (for details, see the Results section). For comparison, we also examine the AA model in step 3 (SAAA-FFR; Figure 1B, right scheme).

Local structural errors are frequently observed in the energy-minimized structure at step 1. The chirality error can occur in the Cα atoms of the amino acids except for Gly or Cγ atoms of Thr and Ile. To fix the error, the corresponding hydrogen atom attached to the chiral center is inverted, and energy minimization is carried out (Figure 1C top).[49] The errors can also be easily fixed through the MD simulation using the UA model (step 3 in the SAUA-FFR), because the geometry around the chiral center, which involves the hydrogen atom, is regulated with the improper torsion angle potential. In the cis peptide bond, the backbone dihedral angle ω is close to 0°, which is energetically unstable. To invert the cis peptide bond to a trans peptide bond, the dihedral angle restraint at $\omega = 180°$ is applied to the corresponding part during the MD simulation (Figure 1C, middle). Note that cis peptide bond is not always an error, and it can often be found even in high-resolution X-ray crystal structures.[49] In this study, we applied restraints to all backbone peptide bonds except for those in proline. Another typical error is ring penetration, in which one covalent bond accidentally penetrates the ring of Phe, Tyr, Trp, His, or Pro (Figure 1C, bottom). This error can be detected based on the bond length of the ring because the penetration makes a ring larger. To fix the error, we first geometrically reduce the ring size and then carry out energy minimization, which allows the penetrating bond to escape from the ring owing to the quick recovery of the natural ring size (see Video S1). The functions for automatically detecting and fixing errors are available in MD software GENESIS ver 1.6 or later.[45,46]

Test Systems. To examine the efficiency and reliability of our scheme, we selected eight proteins: FrhA, PCS, SV, BPP-1, TRPV1, CARD domain of MAVS, BmCPV-1, and PCV2 (see Table 1). In this study, we used the same initial decoys and target density maps as those used in the previous work.[30] We consider

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

Table 2. Average Number of Local Structural Errors (Chirality Errors, cis Peptide Bonds, and Ring Penetrations) in One Decoy Obtained at Step 1 or Step 2 of the Original and Improved Schemes[a]

system	error	original scheme step 2	improved scheme step 1	improved scheme step 2
FrhA	chirality error	1.216 (1.703)	0.006 (0.077)	0.000 (0.000)
	cis peptide bond	10.708 (3.698)	11.546 (3.578)	3.716 (1.733)
	ring penetration	0.226 (0.489)	0.180 (0.428)	0.002 (0.045)
PCS	chirality error	1.210 (2.011)	0.000 (0.000)	0.000 (0.000)
	cis peptide bond	9.348 (3.393)	9.870 (3.816)	0.192 (0.437)
	ring penetration	0.280 (0.605)	0.244 (0.541)	0.000 (0.000)
SV	chirality error	2.204 (2.467)	0.002 (0.045)	0.000 (0.000)
	cis peptide bond	16.698 (6.353)	19.138 (6.485)	4.080 (2.051)
	ring penetration	0.542 (0.867)	0.400 (0.724)	0.006 (0.077)
BPP-1	chirality error	1.260 (2.205)	0.000 (0.000)	0.000 (0.000)
	cis peptide bond	12.166 (4.110)	13.504 (4.200)	1.700 (1.302)
	ring penetration	0.218 (0.463)	0.206 (0.442)	0.002 (0.045)
TRPV1	chirality error	2.064 (2.467)	0.006 (0.077)	0.000 (0.000)
	cis peptide bond	17.014 (5.317)	19.378 (6.296)	1.654 (1.248)
	ring penetration	0.668 (1.100)	0.488 (0.700)	0.000 (0.000)
MAVS	chirality error	0.574 (1.098)	0.000 (0.000)	0.000 (0.000)
	cis peptide bond	4.338 (2.427)	4.648 (2.591)	1.198 (0.942)
	ring penetration	0.074 (0.277)	0.058 (0.234)	0.002 (0.045)
BmCPV-1	chirality error	0.234 (0.632)	0.000 (0.000)	0.000 (0.000)
	cis peptide bond	4.282 (1.958)	4.990 (2.072)	1.582 (1.050)
	ring penetration	0.170 (0.503)	0.158 (0.448)	0.004 (0.063)
PCV2	chirality error	0.788 (1.805)	0.000 (0.000)	0.000 (0.000)
	cis peptide bond	3.426 (2.372)	4.406 (2.355)	1.310 (1.118)
	ring penetration	0.228 (0.522)	0.216 (0.491)	0.002 (0.045)

[a]The values in parentheses represent the standard deviation.

the PDB coordinates as the answer of the prediction, which have been determined by fitting the reference X-ray crystal structure or homology model (MAVS and PCS),[50,51] manual building (SV, BPP-1, BmCPV-1, and PCV2),[52−55] or their combinations (FrhA and TRPV1).[56,57] The proteins are composed of α-helices, β-sheets, or their mixtures (5th column in Table 1). Note that each system is a part of a large complex (6th column in Table 1). Thus, to make the target density map, the corresponding region was clipped out of the experimental map. The resolution of the original map is approximately 3 Å. Previous work demonstrated that these test sets show different qualities in the predicted best model in terms of rmsd with respect to the native conformation.[30] Specifically, the prediction was successful for BmCPV-1 (rmsd = 1.67 Å) but not for PCS (15.00 Å), SV (9.46 Å), or BPP-1 (33.10 Å).

Computational Details of SAUA-FFR and SAAA-FFR. We employed the CHARMM C19[59] and C36m force fields[60] for the UA and AA models, respectively. In step 1 of the improved scheme (Figure 1B), a 1000-step energy minimization was carried out in vacuum. In step 2, energy minimization and restrained MD simulations were performed. To fix the cis peptide bond, we conducted MD simulations at 300 K in vacuum using dihedral angle restraints with a force constant of 10 kcal/mol/rad^2, where the positional restraint was also applied to all Cα atoms (k = 0.5 kcal/mol/Å^2). In step 3 of the SAUA-FFR, c.c.-based FFR[19] was carried out using the UA model with the SA MD protocol (100 ps × 5 cycles = 500 ps in total), where the temperature was decreased from 600 to 300 K in each cycle. We used the effective energy function (EEF1) model for the implicit solvent model.[61] In step 3 of the SAAA-FFR, the AA model was employed with the generalized Born/solvent-accessible surface area (GB/SA) implicit solvent model (OBC2 model) using the same simulation conditions.[62] Here, we examined five different force constants (k = 2000, 4000, 6000, 8000, and 10,000 kcal/mol) in the EM biasing potential for each system. In the MD simulations, we used a cutoff distance of 18 Å. The Langevin thermostat was employed for temperature control, and the equations of motion were integrated with the leapfrog algorithm. All MD simulations were performed using GENESIS.[45,46]

RESULTS

Removal of Local Structural Errors from Initial Decoys. First, we analyzed the number of local structural errors in all decoys to investigate the efficiency of our error-fixing algorithms. In Table 2, we list the average number of chirality errors, cis peptide bonds, and ring penetrations in the 500 decoys obtained at step 2 of the original scheme, step 1 of the improved scheme, and step 2 of the improved scheme. Here, the cis peptide bond was counted using the VMD *cispeptide* plugin,[40] and the chirality errors and ring penetrations were counted with the GENESIS *check_structure* function. Note that the numbers in the table are based on a "warning" message for the suspicious moiety in the molecule. We can see that in the original scheme, there are still some errors even after the FFR with error-fixing restraints. On the other hand, in the improved scheme, chirality errors, cis peptide bonds, and ring penetrations are resolved or significantly reduced. Specifically, in TRPV1, ring penetrations were found in 309 of 500 decoys at step 1 of the improved scheme, but they were completely removed at step 2. These results suggest that the simple protocol used in the original scheme cannot fully solve local structural errors, and careful removal of the errors is necessary before the FFR.

Structural Change during the FFR. To monitor the progress of the refinement in each scheme, we analyzed the c.c. between the experimental and simulated density maps for all 500

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

models. Here, we focus on the five highest c.c. models obtained at 500 ps of step 3. Figure 2A,B illustrates the time evolution of

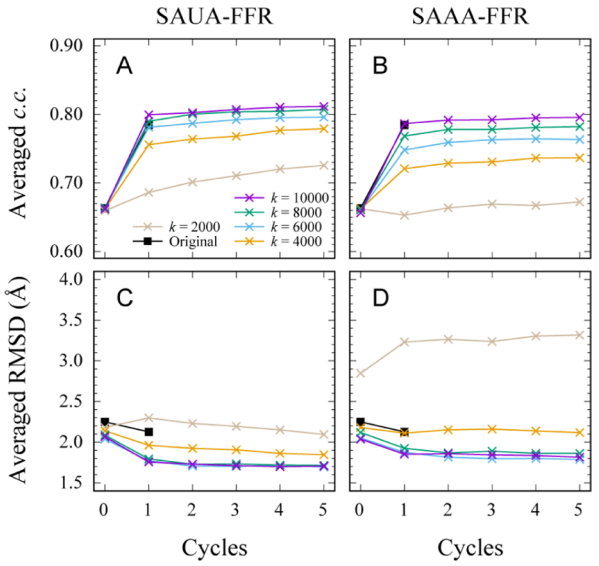


Figure 2. Time evolution of the averaged c.c. and Cα rmsd values for the five highest c.c. models of PCV2. (A) c.c. in SAUA-FFR, (B) c.c. in SAAA-FFR, (C) Cα rmsd in SAUA-FFR, and (D) Cα rmsd in SAAA-FFR. Brown, orange, blue, green, and purple lines are the results obtained from the FFR using k = 2000, 4000, 6000, 8000, and 10,000 kcal/mol, respectively, and the black line corresponds to the original scheme (previous work).[30]

the averaged c.c. of the five highest c.c. models for PCV2 in the SAUA-FFR and SAAA-FFR, respectively. The average was calculated at the last step of each SA cycle. Note that cycle = 0 corresponds to the initial model generated from PULCHRA. For comparison, we also plot the results of the original scheme (black). In all cases in the SAUA-FFR and SAAA-FFR using different biasing force constant k values, the averaged c.c. value increased from the initial value, demonstrating that most models were successfully fitted to the target density map during the iterative FFR. The original scheme also showed an increase in the averaged c.c. value, and the result was similar to the averaged c.c. value of SAUA-FFR$_{k=8000}$ and SAAA-FFR$_{k=10,000}$.

The averaged c.c. value also increased as the force constant increased (from brown to purple lines in Figure 2A,B). If the same force constant was used in SAUA-FFR and SAAA-FFR, a higher c.c. value was obtained in SAUA-FFR than in SAAA-FFR. This is mainly because the structural energy [(i.e., molecular-mechanics potential energy (E_{MM}) + solvation free energy (ΔG_{solv})] in the UA model is lower than that in the AA model, and thus, the EM biasing energy (E_{EM}) in the SAUA-FFR had a relatively larger contribution to the fitting than that in the SAAA-FFR. Similar tendencies were observed in the other seven systems (see Figure S1). We also found that the c.c. value could decrease from the initial value, especially when weak force constants (e.g., k = 2000 kcal/mol) are used for large systems such as BmCPV-1. In such cases, E_{EM} is still inferior to the structural energy, resulting in less fitting to the density map. These results indicate that the optimal force constant for the FFR depends on the system size as well as force fields or molecular models, although it is unknown a priori.

We analyzed the averaged rmsd of the Cα atoms with respect to the native structure for the five highest c.c. models [Figure 2C,D]. In most cases, except for SAAA-FFR$_{k=2000}$, the averaged rmsd value decreased from the initial value by 0.2−0.4 Å, and it almost converged at cycle = 2 or 3. In SAAA-FFR$_{k=2000}$, the biasing force might be too weak to guide the initial model toward the native structure. We can see that a smaller rmsd was mostly obtained with a strong force constant (e.g., k = 6000, 8000, and 10,000 kcal/mol). However, the smallest rmsd was not always obtained with the strongest force constant, as in SAAA-FFR$_{k=6000}$ [blue line in Figure 2D]. One of the reasons might be that a moderate force constant is required in some cases to prevent the structure from becoming trapped in local energy minima. Another reason might be overfitting, where the obtained structure is distorted due to fitting to the noisy density map. This issue is further discussed in the next subsection. A comparison between the three schemes suggests that the SAUA-FFR and SAAA-FFR schemes seem to be more effective than the original scheme. In fact, SAUA-FFR$_{k=8000}$ and SAAA-FFR$_{k=10,000}$ yielded a smaller rmsd compared to the rmsd of the original scheme, even though these three schemes showed almost identical c.c. values [compare the black, green, and purple lines in Figure 2A,B].

For the other systems, we observed similar results, where the stronger force constant yielded a smaller rmsd (see Figure S1). In MAVS and BmCPV-1, the averaged rmsd successfully decreased by 0.2−0.6 Å using k = 6000−10,000 kcal/mol in both SAUA-FFR and SAAA-FFR. On the other hand, in FrhA, PCS, SV, BPP-1, and TRPV1, the rmsd did not change even with the strongest force constant. This is presumably because the structure of the initial model deviated largely from the native structure (e.g., averaged initial rmsd = 6.5−8.8 and 3.5−4.6 Å in TRPV1 and FrhA, respectively), and the conformational search was not conducted sufficiently. These results suggest that the FFR seems to work effectively if the initial rmsd is less than 3.5 Å.

Evaluation of the Decoys. One of the difficult issues in protein structure prediction is the selection of the best model from a large number of decoys. In typical flexible fitting approaches using an X-ray crystal structure as the initial model, we may simply choose a model according to a score that represents goodness of fitting, such as c.c., because the model would already have a protein-like structure. In the FFR for de novo models, however, many structures that show a high c.c. value but include a nonprotein-like conformation might be contained in the decoys and should be discriminated from near-native structures. In addition, we should address overfitting. Thus, the best model must be carefully selected based on not only the goodness of fitting but also any scores that validate the protein structure.

To examine this, we first analyzed the distribution of rmsd as a function of c.c. Figure 3A shows the c.c.−rmsd plot obtained at the last step of SAUA-FFR (step 4 in Figure 1B) for PCV2, where the brown, orange, blue, green, and purple points were obtained with k = 2000, 4000, 6000, 8000, and 10,000 kcal/mol, respectively. Hereafter, we define the 500 models obtained with each force constant as a "decoy set". We see that the rmsd decreases as the c.c. increases in each decoy set, and the distribution shifts toward a higher c.c. as the force constant increases. Each decoy set exhibits a funnel-like distribution, where the bottom decoy is close to the native structure (black point). Similar distributions were observed in the other systems except for PCS and BPP-1 (see Figure S2). In these two cases, improvement of the initial main-chain modeling might be required to reproduce the funnel-like distribution. We suggest that if higher c.c. models are selected, smaller rmsd models can be obtained.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

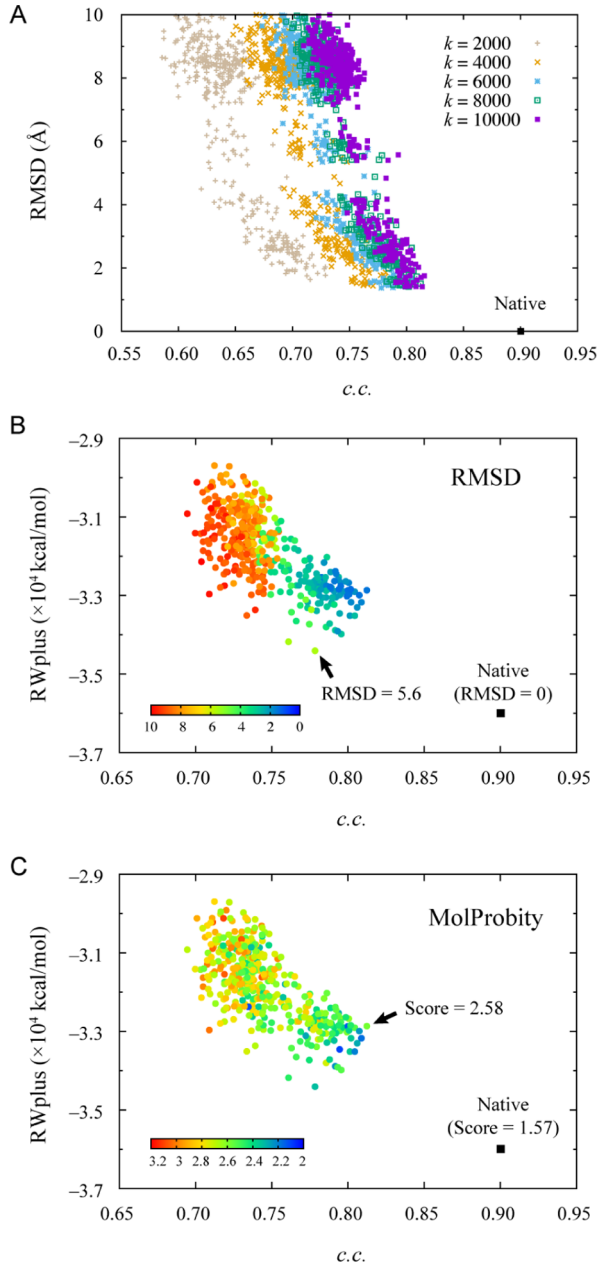


Figure 3. Distribution of the decoys obtained from SAUA-FFR for PCV2. (A) Distribution of the Cα rmsd with respect to the native structure. Note that decoys with a large rmsd (>10 Å) were excluded. (B) Heat map of the rmsd projected onto the c.c.−RWplus plot obtained from SAUA-FFR$_{k=8000}$. (C) Heat map of the MolProbity score projected onto the c.c.−RWplus plot obtained from SAUA-FFR$_{k=8000}$.

Here, the model with the highest c.c. was usually obtained with the strongest force constant, and it corresponded to the model with the smallest rmsd in some cases (e.g., SAUA-FFR for FrhA and SAAA-FFR for TRPV1). However, in most cases, the smallest rmsd was often obtained with a moderate force constant (see Table S1). Particularly, in SAUA-FFR for PCV2 [Figure 3A], the model with the highest c.c. was found in the decoy set with k = 10,000 kcal/mol (c.c. = 0.816 and rmsd = 1.75 Å), while the model with the smallest rmsd was in k = 6000 kcal/mol (c.c. = 0.778 and rmsd = 1.37 Å). These results suggest that to select a model with a smaller rmsd, we should search all decoys generated with various force constants ranging from low to high values. Then, we can eliminate the dependency of the force constant or simultaneously determine the optimal force constant that gives the best model.

The best model should have a protein-like conformation. To investigate the accuracy of the protein structure in the decoys, we first examined the RWplus score, which is a statistical energy function that evaluates the protein structure based on the orientation of the side chains.[47] Figure 3B shows a heat map of the rmsd projected onto the c.c.−RWplus plot obtained from SAUA-FFR$_{k=8000}$ for PCV2. The decoy set exhibits a funnel-like distribution, where the RWplus decreases as the c.c. increases. The rmsd also decreases as the c.c. increases and RWplus decreases, and thus, the bottom of the funnel is close to the native structure (black point). Similar tendencies were observed in the other decoy sets (data not shown). We also found that the model with the best RWplus does not always correspond to the model with the smallest rmsd (rmsd = 5.6 Å) if it has a low c.c., as indicated by the arrow in Figure 3B. Therefore, to find near-native and protein-like structures in the decoys, we should choose decoys that have good scores for both c.c. and RWplus.

Another useful method to validate the protein structure is MolProbity, which assesses protein geometry using clash-score and conformational outliers in the main chain and side chains.[48] Figure 3C shows a heat map of the MolProbity score projected onto the c.c.−RWplus plot. We can see that the MolProbity score decreases as the c.c. increases and the RWplus decreases, suggesting that the decoys at the bottom of the funnel are again likely to have a protein-like structure. Interestingly, as indicated by the arrow, the model with a high c.c. showed worse MolProbity and RWplus scores than the other nearby decoys. For such decoys, we should suspect overfitting, in which the structure is distorted to some extent due to fitting to the noisy density. Thus, we exclude such decoys from the candidates of the best model.

Selection of the Best Model. Based on the above observations, we propose a scheme for the best model selection, which consists of three steps (Figure 4). After obtaining N decoy sets ($N \times 500$ decoys in total) using various force constants, we first decide the Top5N models, where the 5 highest c.c. models are selected from each decoy set. This step can filter out large rmsd models or less-fitted models. Then, we select Top5 models from the Top5N models based on the RWplus score to filter out nonprotein-like structures and eliminate the dependency of the force constant. Finally, we select the best model from the Top5 models based on the MolProbity score to further filter out

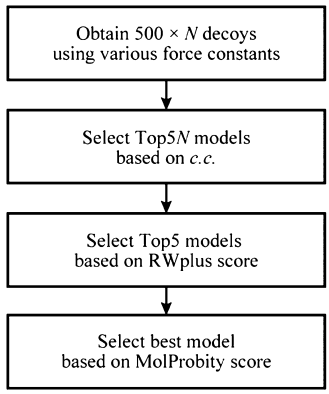


Figure 4. Proposed scheme for the selection of the best model from the decoys obtained from either SAUA-FFR or SAAA-FFR.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

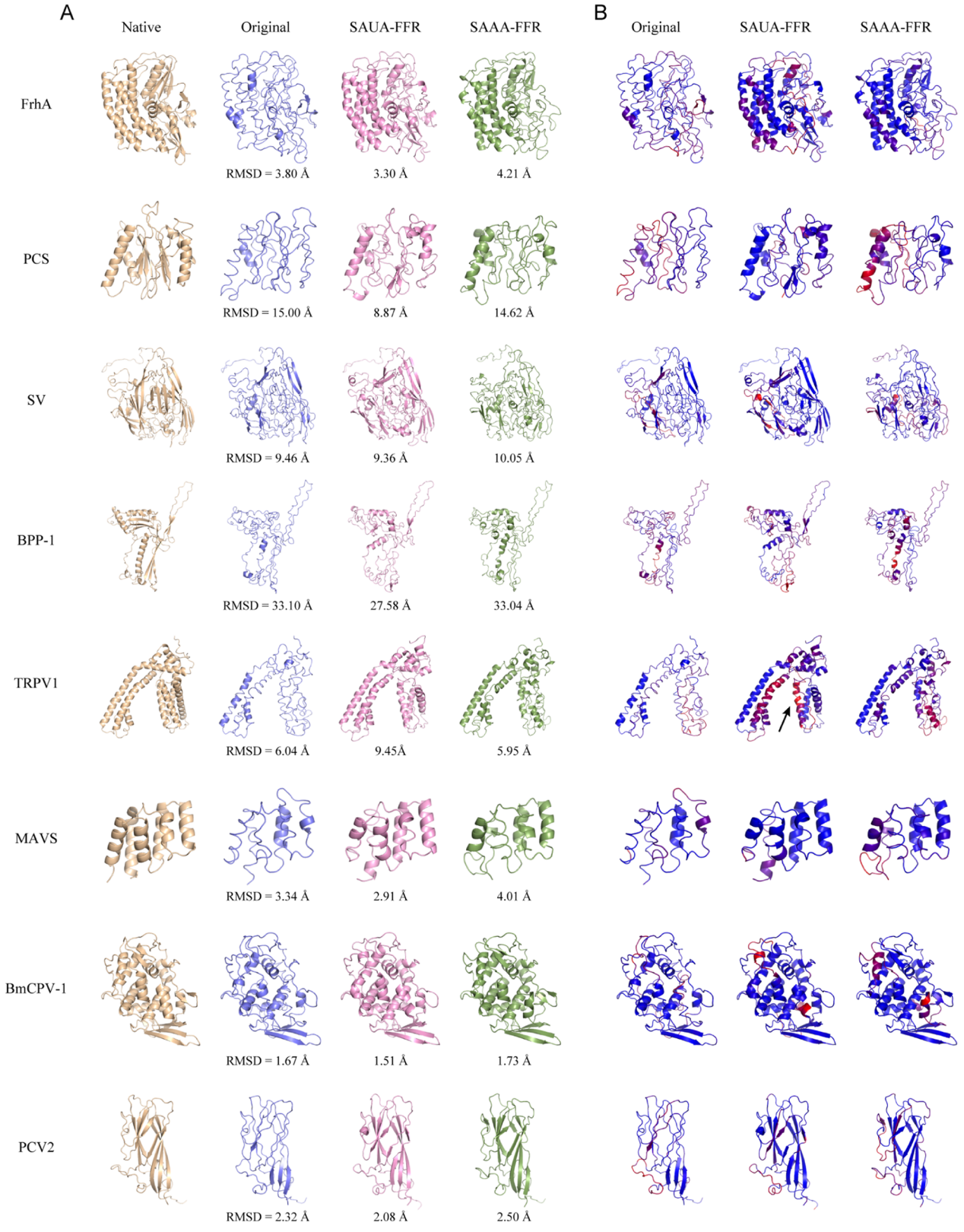


Figure 5. Best models obtained from the original, SAUA-FFR, and SAAA-FFR schemes. (A) Comparison of the native structure and the predicted best models. (B) Deviations of the predicted best model from the native structure (blue: small deviation and red: large deviation). The PyMOL *colorbyrmsd.py* tool was used to make a color map.[63]

nonprotein-like structures and eliminate the possibility of overfitting.

Figure 5A shows the best model of each system obtained from the original scheme (blue), SAUA-FFR (purple), and SAAA-FFR (green) compared to the native structure (gold). Note that the best model in the original scheme was selected using the previous protocol (see the Methods section).[30] Obviously, all best models obtained from the SAUA-FFR and SAAA-FFR have

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

Table 3. Summary of the Structural Properties of the Best Model Predicted from Each Scheme[a]

system	scheme	c.c.	rmsd (Å)	rank	chiral errors/cis peptide bonds/ring penetrations	SS (%)	MolProbity	RWplus (kcal/mol)
FrhA	original	0.795	3.80	9	0/6/0	3.8	2.55	−67,852
	SAUA-FFR	0.793	3.30	3	0/1/0	53.1	2.42	−74,796
	SAAA-FFR	0.739	4.21	100	0/0/0	48.5	2.37	−72,563
	Phenix	0.781	3.82	4	0/4/0	16.7	2.46	−67,886
PCS	original	0.725	15.00	102	1/7/0	0.0	2.43	−33,168
	SAUA-FFR	0.737	8.87	4	0/1/0	26.8	2.57	−35,001
	SAAA-FFR	0.640	14.62	273	0/0/0	9.3	2.49	−33,713
	Phenix	0.698	14.50	14	0/5/0	4.1	2.64	−33,728
SV	original	0.758	9.46	1	0/8/0	15.6	2.52	−89,093
	SAUA-FFR	0.732	9.36	3	0/0/0	44.4	2.15	−88,184
	SAAA-FFR	0.528	10.05	10	0/0/0	21.7	2.19	−90,578
	Phenix	0.757	9.43	1	0/8/0	12.8	2.49	−88,216
BPP-1	original	0.799	33.10	417	0/4/0	3.5	2.94	−42,912
	SAUA-FFR	0.667	27.58	274	0/0/0	7.0	2.25	−44,807
	SAAA-FFR	0.709	33.04	2201	0/0/0	7.9	2.35	−45,125
	Phenix	0.781	33.18	411	0/1/0	0.0	2.93	−42,729
TRPV1	original	0.755	6.04	6	0/17/0	9.1	2.60	−55,334
	SAUA-FFR	0.705	9.45	396	0/0/0	54.1	2.03	−62,981
	SAAA-FFR	0.710	5.95	25	0/1/0	42.7	2.18	−61,773
	Phenix	0.732	7.99	40	0/6/0	9.5	2.42	−50,946
MAVS	original	0.830	3.34	31	1/4/0	14.7	2.69	−14,178
	SAUA-FFR	0.803	2.91	110	0/0/0	64.7	1.62	−18,265
	SAAA-FFR	0.820	4.01	1406	0/0/0	57.4	2.23	−17,147
	Phenix	0.805	3.43	22	0/4/0	17.6	2.47	−15,059
BmCPV-1	original	0.837	1.67	2	0/4/0	43.6	2.22	−58,214
	SAUA-FFR	0.856	1.51	3	0/0/0	85.7	2.15	−62,604
	SAAA-FFR	0.834	1.73	16	0/1/0	70.7	1.90	−61,319
	Phenix	0.802	1.86	2	0/2/0	50.7	2.18	−58,035
PCV2	original	0.777	2.32	30	0/0/0	23.3	2.21	−29,185
	SAUA-FFR	0.798	2.08	124	0/1/0	69.8	2.09	−33,403
	SAAA-FFR	0.798	2.50	168	0/0/0	69.8	1.82	−33,314
	Phenix	0.772	1.72	1	0/1/0	38.4	2.30	−31,378

[a]rmsd: Cα rmsd with respect to the native structure using the MMTSB toolset (*rms.pl*).[58] c.c.: cross-correlation coefficient between the experimental and simulated density maps using the VMD *mdffi* tool.[20,40] Rank: rank of the rmsd over all decoys (500 decoys in the original scheme and 500 × 5 decoys in the improved schemes). Chirality errors: number of chirality errors using the GENESIS *check_structure* function. Cis peptide bonds: number of cis peptide bonds using the VMD *cispeptide* plugin. Ring penetrations: number of ring penetrations using the GENESIS *check_structure* function. SS: reproducibility of the α-helix and β-strand residues using the DSSP program (symbols H and E were counted).[64]

much more SS than those from the original scheme. In the case of FrhA, both α-helices and β-strands are successfully yielded in the SAUA-FFR and SAAA-FFR but not in the original scheme. The rmsd in the original, SAUA-FFR, and SAAA-FFR schemes was 3.80, 3.30, and 4.21 Å, respectively, demonstrating the good performance of SAUA-FFR. Similar tendencies were observed in PCS, SV, BPP-1, MAVS, BmCPV-1, and PCV2. In the case of TRPV1, SAUA-FFR showed a larger rmsd (9.45 Å) than that of the other two schemes. However, the latter two schemes also showed a large rmsd (6.04 or 5.95 Å). These large rmsds mainly originate from a specific region of the protein. Figure 5B illustrates the deviations of the predicted model from the native structure, where the red and blue regions have large and small deviations in the Cα atom position, respectively. In the case of TRPV1, the prediction for most regions was successful, but one α-helix (indicated by the arrow) shifted by three turns, resulting in a large rmsd. This shift in the α-helix might be difficult to discriminate from the native conformation using the current protocol. This point will be discussed later.

In Table 3, we summarize the structural properties of each of the best models. In most cases, the rmsd in SAUA-FFR is smaller than that in the original scheme or SAAA-FFR. Some structural errors were generated in the original scheme, while they were significantly reduced in SAUA-FFR and SAAA-FFR. The reproducibility of the SS in SAUA-FFR is better than that in the original scheme or SAAA-FFR. One of the most remarkable results was 85.7% in SAUA-FFR for BmCPV-1, in which both α-helices (93.2%) and β-strands (64.9%) were well formed in the predicted model. SAUA-FFR also showed better MolProbity and RWplus scores. Similar tendencies were observed in the averaged structural properties of the Top5 models (see Table S2). In Tables 3 and S2, we also show the results of the refinement using Phenix *real_space_refine*.[42] We carried out a basic scheme consisting of minimization global, local grid search, morphing, and SA (simulated_annealing = every_macro_cycle and the other options are default) for the AA model generated from PULCHRA. The best model was selected based on the c.c. value. We see that the results are similar to those in the original scheme, and the reproducibility of the SS is still low. If the initial model deviates largely from the ideal structure, the refinement might not work well in the basic protocol of Phenix. Overall, the SAUA-FFR showed better performance for most cases than the other schemes.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

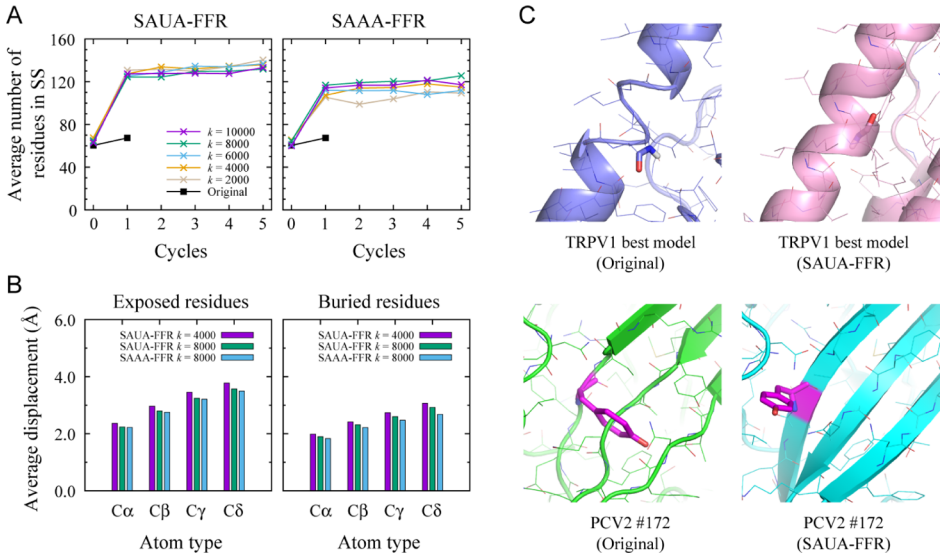


Figure 6. Formation of SS during the FFR. (A) Time evolution of the average number of SS (α-helix + β-strand) in the five highest c.c. models of BmCPV-1 during SAUA-FFR (left) and SAAA-FFR (right). In the analysis, symbols H (α-helix) and E (β-strand) obtained from DSSP were counted. (B) Average displacement of the Cα, Cβ, Cγ, and Cδ atoms in the surface-exposed residues (left) and buried residues (right) before and after the FFR. Here, exposed and buried residues were defined according to the relative solvent accessibility of each residue (criteria = 50%), which was computed with the Naccess program.[65] (C) Examples of the local structural errors that prevented the formation of α-helices (top) and β-sheets (bottom). Right panels show the correct formation of SS after fixing the errors using our algorithms.

In the 5th column of Table 3, we show the rank of the best models in terms of the rmsd. Some best models have a high rank. In particular, the rank of the best model obtained from the SAUA-FFR for FrhA, SV, and BmCPV-1 was 3/2500. On the other hand, the rank was 124/2500 in the case of PCV2. In this case, the smallest rmsd decoy (1.37 Å) was filtered out because it had worse validation scores (MolProbity = 2.56 and RWplus = −32,867 kcal/mol) and a lower reproducibility of SS (66.3%) than those in the best model. Basically, our scheme aims to reduce overfitting as much as possible in the candidate decoys. Here, we also emphasize that the Top5 models actually contained the smallest rmsd model in some cases (e.g., SAUA-FFR for TRPV1, SAAA-FFR for MAVS, and SAAA-FFR for BmCPV-1; see Table S2), demonstrating the good performance of our scheme.

Why the Formation of SS Was Enhanced? One of the remarkable features of our improved schemes is the progressive formation of SS (7th column in Table 3). To investigate how SS were formed during the refinement, we analyzed the time evolution of the average number of residues that formed α-helices or β-strands. Figure 6A shows the results of the DSSP analysis[64] for the five highest c.c. models of BmCPV-1 in SAUA-FFR (left panel) and SAAA-FFR (right panel). In the native structure, there are 140 residues that form SS (103 α-helix and 37 β-strand residues). We found that in the SAUA-FFR, SS were quickly generated at cycle 1 and gradually increased up to ∼135 at cycles 2−5. On the other hand, fewer SS were generated in the SAAA-FFR (∼120) and the original scheme (∼70). Similar results were obtained in the other seven systems (see Figure S3).

One of the reasons for the progressive formation of SS is presumably the high mobility of atoms in the UA model. Figure 6B shows the averaged displacement of the Cα, Cβ, Cγ, and Cδ atoms in the surface-exposed residues (left panel) and buried residues (right panel) of PCV2. Here, the purple, green, and blue bars were obtained from SAUA-FFR$_{k=4000}$, SAUA-FFR$_{k=8000}$, and SAAA-FFR$_{k=8000}$, respectively. We see that the displacement is suppressed if the stronger force constant is used (compare purple and green bars) or if the residues are buried inside the protein (compare left and right panels). If we compare SAUA-FFR$_{k=4000}$ (purple) and SAAA-FFR$_{k=8000}$ (blue), which showed identical results in terms of c.c. and rmsd (see Figure 2), the displacement of the UA model is larger than that of the AA model. In the UA model, hydrogen atoms of CH_3, CH_2, and CH groups are incorporated into the carbon atoms. Therefore, the molecule represented with the UA model is less dense than that with the AA model, resulting in a higher mobility of atoms in the UA model. The implicit solvent model can also contribute to the high mobility of atoms or quick relaxation of the system because there are no explicit waters around the protein.

Treatment of local structural errors such as chirality errors, cis peptide bonds, and ring penetrations is important for the formation of SS because these errors can prevent polypeptides from forming a proper hydrogen bond network. In Figure 6C, we show two examples, where SS were correctly formed by fixing cis peptide bonds (top panels) and ring penetrations (bottom panels). Particularly, fixing ring penetrations is important to stabilize the MD simulation because these errors can easily cause the SHAKE error around the aromatic ring or penetrating covalent bonds. Chirality errors in the Cα atom can also disrupt the α-helix due to steric clashes between side chains.[66] Overall, we suggest fixing these errors before the FFR, and using the combination of the UA model and implicit solvent model are useful for the efficient structure refinement of the decoys obtained from de novo modeling.

DISCUSSION

Computational de novo modeling is usually useful for the density maps at 3.5−5 Å. In fact, if the resolution of the density map is high enough to recognize the type of side chains (e.g., higher than 3.0 Å), manual de novo modeling should be possible by tracing the high dense points in the map. If the resolution is not so high (e.g., lower than 4.0 Å), computational de novo modeling can give us a good hint for reliable structure modeling. MAINMAST is useful for 4−5 Å or higher resolution maps

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

because the backbone is recognized in such maps.[30] In this study, we employed the PDB coordinates determined from the density maps at a slightly higher resolution (2.9−3.6 Å) because the native structure, which is needed to evaluate the protocols, is more reliable.

Although the obtained model can be refined with the flexible fitting, local structural errors such as chirality errors, cis peptide bonds, and ring penetrations should be fixed in advance to obtain a more realistic model. Fixing of the errors can enhance the formation of SS during the refinement. We note that the errors can be mainly generated when the full-atom model is constructed from the main-chain model. In fact, we found that the full-atom model constructed from the Cα-model predicted with Pathwalking contained some local structural errors in the tested systems as in MAINMAST (see Table S3). On the other hand, there were no chirality errors, no ring penetrations, and a small number of cis peptide bonds in the Rosetta model, which directly constructs a full-atom model.

Because more than 99.5% of peptide bonds in the native proteins have cis conformation,[67] we applied the dihedral angle restraint (ω = 180°) to all peptide bonds for simplicity. Spontaneous transition from cis to trans is difficult due to a high energy barrier. To examine the possibility of the transition in the decoys, we performed SAUA-FFR without the dihedral angle restraints starting from the structure in which almost all peptide bonds have cis conformation. We observed that 50% of the peptide bonds had still cis conformation after the refinement in the case of BmCPV-1 with SAUA-FFR$_{k=10,000}$, suggesting that the dihedral angle restraints are essential to fix the cis peptide bonds.

In the proposed SAUA-FFR scheme, iterative SA MD simulation is carried out using the UA model (CHARMM19 force field) in combination with the implicit solvent model (EEF1). One of the advantages of the UA model is its low computational cost compared to that of the AA model. The EEF1 model is also faster than the GB/SA model.[61] In fact, the benchmark performance of SAUA-FFR was 46.5 ns/day for FrhA using a typical Linux machine (Intel Xeon Gold 6130 2.10 GHz; 32 CPU cores), while it was 6.0 ns/day in SAAA-FFR. In the scheme, a short MD simulation (typically 10−100 ps) seems to be enough to obtain a converged structure because the implicit solvent model enables quick equilibration of the system. These features allow us to perform many parallel runs using a supercomputer.

Another advantage of the UA model is that the model is already close to the atomic resolution, enabling an easy conversion to the AA model by just adding hydrogen to the heavy atoms. To date, various CG models, such as Go-model,[68] PRIMO,[69] and SICHO,[70] have been utilized for structure modeling, including not only cryo-EM flexible fitting[24,36,71] but also general de novo protein structure prediction.[72] Although low-resolution molecular models are usually used to reduce computational cost or to enhance conformational sampling, they should eventually be converted to the AA model, which in turn may cause structural errors such as ring penetration. A multiscale protocol combining the CG and AA models can avoid such issues. For example, targeted MD simulation is carried out starting from a certain conformation with the AA model using the Cα atom positions as the reference, as in the CryoFold algorithm[37] or our multiscale flexible fitting protocol proposed recently.[73] However, such a conversion scheme requires additional computation in addition to structure refinement.

In this study, we also proposed a new scheme for the best model selection, where the decoys obtained from multiple FFR runs with various force constants are screened using the combination of the c.c., RWplus score, and MolProbity score. This idea is based on the fact that we do not know the optimal force constant that can minimize overfitting. In the first screening, we use the c.c. to select the models that are fitted to the density map. In the second and third screenings, we try to find a model that has a protein-like conformation and minimal overfitting using the RWplus and MolProbity scores without considering the density map. For validation of the map-to-model quality, various algorithms have been proposed.[74−76] EMRinger is useful to evaluate the side chain modeling based on the consistency between the dihedral angle in the rotamer and the local density of the map.[77] The solvation free energy is also useful for scoring because it can evaluate the exposure of hydrophobic and hydrophilic residues on the protein surface.[78] We suggest that screening decoys through multiple steps and scores with and without the density map is essential in de novo modeling from cryo-EM density maps.

Finally, we discuss further improvements of our scheme. Among the eight proteins employed in this study, TRPV1 is the only membrane protein. For such cases, using the implicit membrane model[79] is more reasonable than the implicit water model. We applied the implicit micelle model (IMIC)[80] to TRPV1 in SAUA-FFR but found that the obtained results were similar to those in the implicit water model. The structural properties of the best model were rmsd = 6.79 Å, c.c. = 0.716, reproducibility of SS = 55.5%, MolProbity score = 2.29, and RWplus score = −59,940 kcal/mol. This is presumably because the membrane environment does not significantly affect the movement of atoms in the flexible fitting. The fitting force still seems to be superior to the effect from the membrane or solvent. However, we expect that the solvation free energy calculated in the membrane environment is useful to select the best model or to filter out the decoys that have abnormal conformations, such as the shift of the transmembrane α-helices, as observed in our calculations. Yuzlenko and Lazaridis[81] and Dutagaci et al.[82] suggested using a scoring function that includes the solvation free energy calculated in the implicit membrane model for the discrimination of the native conformation from decoys of membrane proteins. The early stage of MAINMAST should also be improved by considering the effect of the solvent and/or membrane environment.

CONCLUSIONS

In this study, we propose the SAUA-FFR scheme for efficient FFR, in which c.c.-based flexible fitting with the iterative SA MD protocol is carried out using the UA model in combination with the implicit solvent model. To obtain a model with less overfitting, we carried out multiple FFR runs with various force constants ranging from weak to strong values and screened the decoys using a combination of the c.c., RWplus score, and MolProbity score. Our scheme showed progressive formation of SS owing to the high mobility of atoms in the UA model. Our new algorithm for fixing local structure errors also contributed to the correct formation of the hydrogen bond network in SS. We expect that our scheme is useful for reliable de novo structure modeling from cryo-EM density maps with low computational cost.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00230.

Summary of the smallest rmsd among all decoys, force constant k that gave the smallest rmsd, and c.c. of the smallest rmsd model; summary of the average structural properties of the Top5 models; number of structural errors found in the decoys obtained from Pathwalking and Rosetta; time evolution of the averaged c.c. and Cα rmsd for the five highest c.c. models; distribution of the decoys obtained from SAUA-FFR; and formation of the SSs during SAUA-FFR and SAAA-FFR (PDF)

Fixing ring penetration (MP4)

■ AUTHOR INFORMATION

Corresponding Author

Takaharu Mori − *RIKEN Cluster for Pioneering Research, Wako-shi, Saitama 351-0198, Japan;* orcid.org/0000-0002-8717-2926; Phone: +81-48-462-1407; Email: t.mori@riken.jp; Fax: +81-48-467-4532

Authors

Genki Terashi − *Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, United States;* orcid.org/0000-0002-5339-909X

Daisuke Matsuoka − *RIKEN Cluster for Pioneering Research, Wako-shi, Saitama 351-0198, Japan*

Daisuke Kihara − *Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, United States; Department of Computer Science, Purdue University, West Lafayette, Indiana 47907, United States;* orcid.org/0000-0003-4091-6614

Yuji Sugita − *RIKEN Cluster for Pioneering Research, Wako-shi, Saitama 351-0198, Japan; RIKEN Center for Computational Science, Kobe, Hyogo 650-0047, Japan; RIKEN Center for Biosystems Dynamics Research, Kobe, Hyogo 650-0047, Japan;* orcid.org/0000-0001-9738-9216

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00230

Notes

The authors declare no competing financial interest.

The methods developed here are freely available in MD software GENESIS ver 1.6 or later from the website (https://www.r-ccs.riken.jp/labs/cbrt/). MAINMAST is also freely available from the website (https://kiharalab.org/emsuites/mainmast.php). The initial structures of each protein, input cryo-EM density maps, refined Top5 models, and sample control files of GENESIS and MDFF are available from GitHub (https://github.com/RikenSugitaLab/). The tutorial of the SAUA-FFR is available in the GENESIS website.

■ ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI (numbers 19K06532 and 19H05645), a grant from Innovative Drug Discovery Infrastructure through Functional Control of Biomolecular Systems, Priority Issue 1 in Post-K Supercomputer Development (hp170254), and the RIKEN Pioneering Projects, Dynamic Structural Biology, and Glycolipidologue. MD simulations were partially carried out on HOKUSAI GreatWave and BigWaterFall at RIKEN. DK acknowledges support from the National Institutes of Health (R01GM123055 and R01GM133840) and the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, and DBI2003635).

■ REFERENCES

(1) Danev, R.; Yanagisawa, H.; Kikkawa, M. Cryo-electron Microscopy Methodology: Current Aspects and Future Directions. *Trends Biochem. Sci.* **2019**, *44*, 837−848.

(2) Cheng, Y. Single-particle Cryo-EM−How Did It Get Here and Where Will It Go. *Science* **2018**, *361*, 876−880.

(3) Chen, S.; Zhao, Y.; Wang, Y.; Shekhar, M.; Tajkhorshid, E.; Gouaux, E. Activation and Desensitization Mechanism of AMPA Receptor-TARP Complex by Cryo-EM. *Cell* **2017**, *170*, 1234−1246.

(4) Kishikawa, J.; Nakanishi, A.; Furuta, A.; Kato, T.; Namba, K.; Tamakoshi, M.; Mitsuoka, K.; Yokoyama, K. Mechanical inhibition of isolated Vo from V/A-ATPase for proton conductance. *eLife* **2020**, *9*, No. e56862.

(5) Matoba, K.; Kotani, T.; Tsutsumi, A.; Tsuji, T.; Mori, T.; Noshiro, D.; Sugita, Y.; Nomura, N.; Iwata, S.; Ohsumi, Y.; Fujimoto, T.; Nakatogawa, H.; Kikkawa, M.; Noda, N. N. Atg9 is a Lipid Scramblase That Mediates Autophagosomal Membrane Expansion. *Nat. Struct. Mol. Biol.* **2020**, *27*, 1185−1193.

(6) Ehara, H.; Kujirai, T.; Fujino, Y.; Shirouzu, M.; Kurumizaka, H.; Sekine, S.-i. Structural Insight into Nucleosome Transcription by RNA Polymerase II with Elongation Factors. *Science* **2019**, *363*, 744−747.

(7) Hiraizumi, M.; Yamashita, K.; Nishizawa, T.; Nureki, O. Cryo-EM Structures Capture the Transport Cycle of the P4-ATPase Flippase. *Science* **2019**, *365*, 1149−1155.

(8) Nakane, T.; Kotecha, A.; Sente, A.; McMullan, G.; Masiulis, S.; Brown, P.; Grigoras, I. T.; Malinauskaite, L.; Malinauskas, T.; Miehling, J.; Uchanski, T.; Yu, L. B.; Karia, D.; Pechnikova, E. V.; de Jong, E.; Keizer, J.; Bischoff, M.; McCormack, J.; Tiemeijer, P.; Hardwick, S. W.; Chirgadze, D. Y.; Murshudov, G.; Aricescu, A. R.; Scheres, S. H. W. Single-particle Cryo-EM at Atomic Resolution. *Nature* **2020**, *587*, 152−156.

(9) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587*, 157−161.

(10) Kim, D. N.; Sanbonmatsu, K. Tools for the Cryo-EM Gold Rush: Going from the Cryo-EM Map to the Atomistic Model. *Biosci. Rep.* **2017**, *37*, BSR20170072.

(11) Malhotra, S.; Träger, S.; Dal Peraro, M.; Topf, M. Modelling Structures in Cryo-EM Maps. *Curr. Opin. Struct. Biol.* **2019**, *58*, 105−114.

(12) Alnabati, E.; Kihara, D. Advances in Structure Modeling Methods for Cryo-electron Microscopy Maps. *Molecules* **2019**, *25*, 82.

(13) Dodd, T.; Yan, C.; Ivanov, I. Simulation-based Methods for Model Building and Refinement in Cryoelectron Microscopy. *J. Chem. Inf. Model.* **2020**, *60*, 2470−2483.

(14) Roseman, A. M. Docking Structures of Domains into Maps from Cryo-electron Microscopy Using Local Correlation. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56*, 1332−1340.

(15) Wriggers, W.; Birmanns, S. Using Situs for Flexible and Rigid-body Fitting of Multiresolution Single-molecule Data. *J. Struct. Biol.* **2001**, *133*, 193−202.

(16) Zhang, B.; Zhang, X.; Pearce, R.; Shen, H.-B.; Zhang, Y. A New Protocol for Atomic-level Protein Structure Modeling and Refinement Using Low-to-medium Resolution Cryo-EM Density Maps. *J. Mol. Biol.* **2020**, *432*, 5365−5377.

(17) Tama, F.; Miyashita, O.; Brooks, C. L. Flexible Multi-scale Fitting of Atomic Structures into Low-resolution Electron Density Maps with Elastic Network Normal Mode Analysis. *J. Mol. Biol.* **2004**, *337*, 985−999.

(18) Lopéz-Blanco, J. R.; Chacón, P. iMODFIT: Efficient and Robust Flexible Fitting Based on Vibrational Analysis in Internal Coordinates. *J. Struct. Biol.* **2013**, *184*, 261−270.

(19) Orzechowski, M.; Tama, F. Flexible Fitting of High-resolution X-ray Structures into Cryoelectron Microscopy Maps Using Biased Molecular Dynamics Simulations. *Biophys. J.* **2008**, *95*, 5692−5705.
(20) 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*, 673−683.
(21) Topf, M.; Lasker, K.; Webb, B.; Wolfson, H.; Chiu, W.; Sali, A. Protein Structure Fitting and Refinement Guided by Cryo-EM Density. *Structure* **2008**, *16*, 295−307.
(22) Ishida, H.; Matsumoto, A. Free-energy Landscape of Reverse tRNA Translocation through the Ribosome Analyzed by Electron Microscopy Density Maps and Molecular Dynamics Simulations. *PLoS One* **2014**, *9*, No. e101951.
(23) Bonomi, M.; Pellarin, R.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-electron Microscopy. *Biophys. J.* **2018**, *114*, 1604−1613.
(24) Mori, T.; Kulik, M.; Miyashita, O.; Jung, J.; Tama, F.; Sugita, Y. Acceleration of Cryo-EM Flexible Fitting for Large Biomolecular Systems by Efficient Space Partitioning. *Structure* **2019**, *27*, 161−174.
(25) Igaev, M.; Kutzner, C.; Bock, L. V.; Vaiana, A. C.; Grubmüller, H. Automated Cryo-EM Structure Refinement Using Correlation-driven Molecular Dynamics. *eLife* **2019**, *8*, No. e43542.
(26) Costa, M. G. S.; Fagnen, C.; Vénien-Bryan, C.; Perahia, D. A New Strategy for Atomic Flexible Fitting in Cryo-EM Maps by Molecular Dynamics with Excited Normal Modes (MDeNM-EMfit). *J. Chem. Inf. Model.* **2020**, *60*, 2419−2423.
(27) Wang, R. Y.-R.; Kudryashev, M.; Li, X.; Egelman, E. H.; Basler, M.; Cheng, Y.; Baker, D.; DiMaio, F. De Novo Protein Structure Determination from Near-atomic-resolution Cryo-EM Maps. *Nat. Methods* **2015**, *12*, 335−338.
(28) Lindert, S.; Staritzbichler, R.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. EM-Fold: De Novo Folding of α-Helical Proteins Guided by Intermediate-Resolution Electron Microscopy Density Maps. *Structure* **2009**, *17*, 990−1003.
(29) Chen, M.; Baldwin, P. R.; Ludtke, S. J.; Baker, M. L. De Novo Modeling in Cryo-EM Density Maps with Pathwalking. *J. Struct. Biol.* **2016**, *196*, 289−298.
(30) Terashi, G.; Kihara, D. De Novo Main-chain Modeling for EM Maps Using MAINMAST. *Nat. Commun.* **2018**, *9*, 1618.
(31) Rotkiewicz, P.; Skolnick, J. Fast Procedure for Reconstruction of Full-atom Protein Models from Reduced Representations. *J. Comput. Chem.* **2008**, *29*, 1460−1465.
(32) Lindert, S.; McCammon, J. A. Improved CryoEM-guided Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction. *J. Chem. Theory Comput.* **2015**, *11*, 1337−1346.
(33) Leelananda, S. P.; Lindert, S. Iterative Molecular Dynamics-Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. *J. Chem. Theory Comput.* **2017**, *13*, 5131−5145.
(34) Sugita, Y.; Okamoto, Y. Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(35) Singharoy, A.; Teo, I.; McGreevy, R.; Stone, J. E.; Zhao, J.; Schulten, K. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* **2016**, *5*, No. e16105.
(36) Miyashita, O.; Kobayashi, C.; Mori, T.; Sugita, Y.; Tama, F. Flexible Fitting to Cryo-EM Density Map Using Ensemble Molecular Dynamics Simulations. *J. Comput. Chem.* **2017**, *38*, 1447−1461.
(37) Shekhar, M.; Terashi, G.; Gupta, C.; Debussche, G.; Sisco, N. J.; Nguyen, J.; Zook, J.; Vant, J.; Sarkar, D.; Fromme, P.; Van Horn, W. D.; Dill, K.; Kihara, D.; Tajkhorshid, E.; Perez, A.; Singharoy, A. CryoFold: Ab-initio Structure Determination from Electron Density Maps Using Molecular Dynamics. **2019**, bioRxiv:687087.
(38) Schlitter, J.; Engels, M.; Krüger, P.; Jacoby, E.; Wollmer, A. Targeted Molecular Dynamics Simulation of Conformational Change-Application to the T ↔ R Transition in Insulin. *Mol. Simul.* **1993**, *10*, 291−308.
(39) 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.
(40) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics Modell.* **1996**, *14*, 33−38.
(41) Emsley, P.; Cowtan, K. Coot: Model-building Tools for Molecular Graphics. *Acta Crystallogr., Sect. D: Struct. Biol.* **2004**, *60*, 2126−2132.
(42) Afonine, P. V.; Poon, B. K.; Read, R. J.; Sobolev, O. V.; Terwilliger, T. C.; Urzhumtsev, A.; Adams, P. D. Real-space refinement inPHENIXfor cryo-EM and crystallography. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018**, *74*, 531−544.
(43) Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018**, *74*, 519−530.
(44) Badaczewska-Dawid, A. E.; Kolinski, A.; Kmiecik, S. Computational Reconstruction of Atomistic Protein Structures from Coarse-grained Models. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 162−176.
(45) Jung, J.; Mori, T.; Kobayashi, C.; Matsunaga, Y.; Yoda, T.; Feig, M.; Sugita, Y. GENESIS: A Hybrid-parallel and Multi-scale Molecular Dynamics Simulator with Enhanced Sampling Algorithms for Biomolecular and Cellular Simulations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 310−323.
(46) Kobayashi, C.; Jung, J.; Matsunaga, Y.; Mori, T.; Ando, T.; Tamura, K.; Kamiya, M.; Sugita, Y. GENESIS 1.1: A Hybrid-parallel Molecular Dynamics Simulator with Enhanced Sampling Algorithms on Multiple Computational Platforms. *J. Comput. Chem.* **2017**, *38*, 2193−2206.
(47) Zhang, J.; Zhang, Y. A Novel Side-chain Orientation Dependent Potential Derived from Random-walk Reference State for Protein Fold Selection and Structure Prediction. *PLoS One* **2010**, *5*, No. e15386.
(48) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: All-atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr., Sect. D: Struct. Biol.* **2010**, *66*, 12−21.
(49) Schreiner, E.; Trabuco, L. G.; Freddolino, P. L.; Schulten, K. Stereochemical Errors and Their Implications for Molecular Dynamics Simulations. *BMC Bioinf.* **2011**, *12*, 190.
(50) Wu, B.; Peisley, A.; Tetrault, D.; Li, Z.; Egelman, E. H.; Magor, K. E.; Walz, T.; Penczek, P. A.; Hur, S. Molecular Imprinting As a Signal-activation Mechanism of the Viral RNA Sensor RIG-I. *Mol. Cell* **2014**, *55*, 511−523.
(51) Li, H.; O'Donoghue, A. J.; van der Linden, W. A.; Xie, S. C.; Yoo, E.; Foe, I. T.; Tilley, L.; Craik, C. S.; da Fonseca, P. C. A.; Bogyo, M. Structure- and Function-based Design of Plasmodium-selective Proteasome Inhibitors. *Nature* **2016**, *530*, 233−236.
(52) Zhang, X.; Sun, S.; Xiang, Y.; Wong, J.; Klose, T.; Raoult, D.; Rossmann, M. G. Structure of Sputnik, a virophage, at 3.5-A resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 18431−18436.
(53) Zhang, X.; Guo, H.; Jin, L.; Czornyj, E.; Hodes, A.; Hui, W. H.; Nieh, A. W.; Miller, J. F.; Zhou, Z. H. A new topology of the HK97-like fold revealed in Bordetella bacteriophage by cryoEM at 3.5 A resolution. *eLife* **2013**, *2*, No. e01299.
(54) Yu, X.; Jiang, J.; Sun, J.; Zhou, Z. H. A Putative ATPase Mediates RNA Transcription and Capping in a dsRNA Virus. *eLife* **2015**, *4*, No. e07901.
(55) Liu, Z.; Guo, F.; Wang, F.; Li, T.-C.; Jiang, W. 2.9 Å Resolution Cryo-EM 3D Reconstruction of Close-Packed Virus Particles. *Structure* **2016**, *24*, 319−328.
(56) Allegretti, M.; Mills, D. J.; McMullan, G.; Kühlbrandt, W.; Vonck, J. Atomic Model of the F420-reducing [NiFe] Hydrogenase by Electron Cryo-microscopy Using a Direct Electron Detector. *eLife* **2014**, *3*, No. e01963.
(57) Liao, M.; Cao, E.; Julius, D.; Cheng, Y. Structure of the TRPV1 Ion Channel Determined by Electron Cryo-microscopy. *Nature* **2013**, *504*, 107−112.
(58) Feig, M.; Karanicolas, J.; Brooks, C. L. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.
(59) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105*, 1902−1921.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528

(60) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71−73.
(61) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins* **1999**, *35*, 133−152.
(62) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383−394.
(63) Schrodinger LLC. *The PyMOL Molecular Graphics System*, version 1.8, 2015.
(64) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(65) Hubbard, S.; Thornton, J. *Naccess*, version 2.1.1, 1993.
(66) Krause, E.; Bienert, M.; Schmieder, P.; Wenschuh, H. The Helix-Destabilizing Propensity Scale of D-Amino Acids: The Influence of Side Chain Steric Effects. *J. Am. Chem. Soc.* **2000**, *122*, 4865−4870.
(67) Weiss, M. S.; Jabs, A.; Hilgenfeld, R. Peptide Bonds Revisited. *Nat. Struct. Biol.* **1998**, *5*, 676.
(68) Taketomi, H.; Ueda, Y.; Gō, N. Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.
(69) Kar, P.; Gopal, S. M.; Cheng, Y.-M.; Predeus, A.; Feig, M. PRIMO: A Transferable Coarse-grained Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 3769−3788.
(70) Kolinski, A.; Jaroszewski, L.; Rotkiewicz, P.; Skolnick, J. An Efficient Monte Carlo Model of Protein Chains. Modeling the Short-range Correlations between Side Group Centers of Mass. *J. Phys. Chem. B* **1998**, *102*, 4628−4637.
(71) Whitford, P. C.; Ahmed, A.; Yu, Y.; Hennelly, S. P.; Tama, F.; Spahn, C. M. T.; Onuchic, J. N.; Sanbonmatsu, K. Y. Excited States of Ribosome Translocation Revealed through Integrative Molecular Modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 18943−18948.
(72) Feig, M.; Gopal, S. M.; Vadivel, K.; Stumpff-Kane, A. Conformational Sampling in Structure Prediction and Refinement with Atomistic and Coarse-grained Models. In *Multiscale Approaches to Protein Modeling: Structure Prediction, Dynamics, Thermodynamics and Macromolecular Assemblies*; Kolinski, A., Ed.; Springer New York: New York, NY, 2011, pp 85−109.
(73) Kulik, M.; Mori, T.; Sugita, Y. Multi-scale Flexible Fitting of Proteins to Cryo-EM Density Maps at Medium Resolution. *Front. Mol. Biosci.* **2021**, *8*, 631854.
(74) Afonine, P. V.; Klaholz, B. P.; Moriarty, N. W.; Poon, B. K.; Sobolev, O. V.; Terwilliger, T. C.; Adams, P. D.; Urzhumtsev, A. New Tools for the Analysis and Validation of Cryo-EM Maps and Atomic Models. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018**, *74*, 814−840.
(75) Sazzed, S.; Scheible, P.; Alshammari, M.; Wriggers, W.; He, J. Cylindrical Similarity Measurement for Helices in Medium-resolution Cryo-electron Microscopy Density Maps. *J. Chem. Inf. Model.* **2020**, *60*, 2644−2650.
(76) Ramírez-Aportela, E.; Maluenda, D.; Fonseca, Y. C.; Conesa, P.; Marabini, R.; Heymann, J. B.; Carazo, J. M.; Sorzano, C. O. S. FSC-Q: a CryoEM Map-to-atomic Model Quality Validation Based on the Local Fourier Shell Correlation. *Nat. Commun.* **2021**, *12*, 42.
(77) Barad, B. A.; Echols, N.; Wang, R. Y.-R.; Cheng, Y.; DiMaio, F.; Adams, P. D.; Fraser, J. S. EMRinger: Side Chain Directed Model and Map Validation for 3D Cryo-electron Microscopy. *Nat. Methods* **2015**, *12*, 943−946.
(78) Li, Z.; Yang, Y.; Zhan, J.; Dai, L.; Zhou, Y. Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. *Annu. Rev. Biophys.* **2013**, *42*, 315−335.
(79) Mori, T.; Miyashita, N.; Im, W.; Feig, M.; Sugita, Y. Molecular Dynamics Simulations of Biological Membranes and Membrane Proteins Using Enhanced Conformational Sampling Algorithms. *Biochim. Biophys. Acta, Biomembr.* **2016**, *1858*, 1635−1651.
(80) Mori, T.; Sugita, Y. Implicit Micelle Model for Membrane Proteins Using Superellipsoid Approximation. *J. Chem. Theory Comput.* **2020**, *16*, 711−724.
(81) Yuzlenko, O.; Lazaridis, T. Membrane Protein Native State Discrimination by Implicit Membrane Models. *J. Comput. Chem.* **2013**, *34*, 731−738.
(82) Dutagaci, B.; Wittayanarakul, K.; Mori, T.; Feig, M. Discrimination of Native-like States of Membrane Proteins with Implicit Membrane-based Scoring Functions. *J. Chem. Theory Comput.* **2017**, *13*, 3049−3059.

https://doi.org/10.1021/acs.jcim.1c00230
J. Chem. Inf. Model. 2021, 61, 3516−3528