Improving low-accuracy protein structures using enhanced sampling techniques

Tianwu Zang,¹ Tianqi Ma,¹ Qinghua Wang,² and Jianpeng Ma^{1,2,a)}

¹*Applied Physics Program and Department of Bioengineering, Rice University, Houston, Texas 77005, USA*

²*Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, BCM-125, Houston, Texas 77030, USA*

(Received 28 February 2018; accepted 23 May 2018; published online 11 June 2018)

In this paper, we report results of using enhanced sampling and blind selection techniques for high-accuracy protein structural refinement. By combining a parallel continuous simulated tempering (PCST) method, previously developed by Zang *et al.* [J. Chem. Phys. **141**, 044113 (2014)], and the structure based model (SBM) as restraints, we refined 23 targets (18 from the refinement category of the CASP10 and 5 from that of CASP12). We also designed a novel model selection method to blindly select high-quality models from very long simulation trajectories. The combined use of PCST-SBM with the blind selection method yielded final models that are better than initial models. For Top-1 group, 7 out of 23 targets had better models (greater global distance test total scores) than the critical assessment of structure prediction participants. For Top-5 group, 10 out of 23 were better. Our results justify the crucial position of enhanced sampling in protein structure prediction and refinement and demonstrate that a considerable improvement of low-accuracy structures is achievable with current force fields. *Published by AIP Publishing.* <https://doi.org/10.1063/1.5027243>

I. INTRODUCTION

To determine the three-dimensional structure of a protein solely from its primary sequence remains an elusive goal in modern computational biology. At the current stage, computational prediction methods^{1–7} can produce models that are usually of 3–5 Å main chain root-mean-square deviation (RMSD) from the native structure.^{8–10} New methods are needed to refine these low-accuracy models to high-accuracy ones (with 1–2 Å RMSD).

Since 1994, the biennial Critical Assessment of Structure Prediction (CASP)^{11,12} provides a benchmark for all prediction techniques in an outcome-oriented way. Particularly, the refinement category of CASP (also known as CASPR) benchmarks all the refinement methodologies. While the past two decades have witnessed significant improvements of low-accuracy structure prediction,^{8–10} the progress in CASPR is relatively slow.^{13,14} The main reason is that most of the methods used in low-accuracy modeling lost their effectiveness when dealing with high-accuracy refinement. In this regard, molecular dynamics (MD) simulation at the atomic resolution, i.e., without any coarse-graining, has a high promise for high-accuracy refinement within limited computing time. However, even today's fastest supercomputers especially built for MD simulation can only reach several milliseconds,^{15,16} which are far shorter than the time scale of folding a large protein in real life. Therefore, powerful enhanced sampling methods are needed to overcome this difficulty.

In this work, we focus on developing computational protocols that are especially designed for high-accuracy refinement. By combining a temperature-based method that we previously introduced, parallel continuous simulated tempering (PCST),¹⁷ with structure-based model (SBM)^{18,19} restraints, we were able to perform a more thorough sampling in the configurational space. Furthermore, the introduction of a novel blind model selection method was able to select final models from very long MD trajectories.

The effectiveness of our refinement protocol is demonstrated on 23 targets selected from the refinement category of CASP10 and CASP12. The final models are significantly improved over the initial models as judged by the global distance test total score, or GDT.TS.^{20–22} In Top-1 group, 7 out of 23 cases, and in Top-5 group, 10 out of 23 cases, our final models are better than the CASP participants. The results suggest that, despite the inaccuracy of current force fields, the combination of effective sampling and model selection techniques can significantly improve low-accuracy models.

II. METHODS

A. Temperature-based enhanced sampling

One of the major approaches in enhanced sampling is to cross the energy barriers by accelerating dynamics using a generalized ensemble.²³ As its name implies, generalized ensemble switches the Boltzmann distribution in the canonical ensemble to other mathematical forms $W(X, \beta)$ to encourage sampling of rare events such as protein folding. For example, multicanonical ensembles^{24–30} generate a flat potential energy histogram with $W(X) \propto \Omega^{-1}(E(X)) = e^{-\beta TS(E(X))}$,

^{a)}Author to whom correspondence should be addressed: jpm@bcm.tmc.edu.
 Telephone: 713-798-8187. Fax: 713-796-9438.

while tempering methods^{31–35} generate a flat or biased temperature energy histogram with $W(T) = \text{const.}$ or $W(\beta) = \text{const.}$ Their efficiency has been shown to be similar in protein folding simulations.³⁶

As a temperature-based method, PCST inherits the idea of simulated tempering (ST)^{31,37,38} while it adopts the spirit of parallel tempering (PT),^{32–35} or the replica-exchange method, by employing multiple copies with different temperature distributions and performing random walk in temperature space according to each copy's own distribution. In PCST, the temperature in the i th copy is no longer a constant, but is updated via a stochastic process as follows:

$$d\left(\frac{1}{\beta}\right) = \left(E - \bar{E}(\beta) - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2}\right)dt + \frac{\sqrt{2}}{\beta}dW_t, \quad (1)$$

where the reciprocal temperature β is defined in a large range ($\beta_{\min}, \beta_{\max}$) and dW_t is the Wiener process. Evaluated during the simulation, $\bar{E}(\beta)$ denotes the average potential energy in the canonical ensemble at the current temperature, i.e., $\bar{E}(\beta) \equiv Z^{-1}(\beta) \int E(X)e^{-\beta E(X)}dX$ where $Z(\beta)$ is the canonical partition function at temperature β . This random walk in temperature space will generate a probability distribution determined by three parameters $\{\gamma, \beta_i^0, \sigma_i\}$, with the form of $P(X, \beta) \propto Z^{-1}(\beta)e^{-\beta E(X)}\beta^{-\gamma}e^{-(\beta - \beta_i^0)^2/2\sigma_i^2}$. The probability distribution can be regarded as a sum of Boltzmann distributions at different temperatures with a polynomial-like weight function $\beta^{-\gamma}$ as well as a Gaussian-like weight function with the mean value of β_i^0 and width of σ_i .

During the tempering simulation, many small bins (β_i, β_{i+1}) are created in a large temperature range ($\beta_{\min}, \beta_{\max}$). As the bin size is very small, we assume that the average potential energy $\bar{E}(\beta)$ within this bin is constant, which is already shown to be valid even for finite-size phase transition problems. Thus, when the temperature falls into the i th bin at the current simulation step, the current potential energy is collected to calculate the average energy in this bin $\bar{E}_i$, which serves as an estimation of $\bar{E}(\beta)$. Since the energy data are not reliable in the early stage of simulation when the system has not equilibrated properly, various advanced statistical methods³⁸ are employed to improve the accuracy of $\bar{E}_i$. First, an adaptive averaging scheme is used via the assignment of larger weights toward newer collected statistics. Suppose the k th sample of potential energy in the i th small bin is $E_i^{(k)}$ and its weight is $w_i^{(k)}$, then the average energy of samples is $\bar{E}_i = \sum E_i^{(k)} w_i^{(k)}$. In our scheme, the weight is $w_i^{(k)} = (1 - C/k)^{-1}$, where C is a constant smaller than 1 and gradually reduces $w_i^{(k)}$ from $(1 - C)^{-1}$ to 1 as $k \rightarrow \infty$. Second, the statistics in neighboring bins are borrowed to provide a much less biased estimation of $\bar{E}_i$ since they are strongly correlated with each other under an integral identity. The integral identity assures that the estimated potential energy converges to the exact potential energy asymptotically with long time simulation. These two schemes work together to overcome the low accuracy of $\bar{E}_i$ in short simulation due to the limited number of available samples in each small bin.

In addition, a specific parameter exchange protocol¹⁷ between copies is introduced to facilitate the barrier crossing of low temperature copies. The acceptance ratio for copy

exchange between the i th copy and the j th copy is defined as $\min\{1, e^{-(\Delta\beta\Delta q_0 + \Delta\beta^2\Delta q_1)}\}$ to satisfy the detailed balance, where $\Delta\beta \equiv \beta_i - \beta_j$, $\Delta\beta^2 \equiv \beta_i^2 - \beta_j^2$, $\Delta q_0 \equiv \beta_i^0/\sigma_i^2 - \beta_j^0/\sigma_j^2$, and $\Delta q_1 \equiv -1/(2\sigma_i^2) + 1/(2\sigma_j^2)$. An important feature of this protocol is that the acceptance ratio is only related to the parameters in the temperature space, whereas it does not have an explicit dependence on the extensive quantities, i.e., the potential energy E . Therefore, compared with parallel tempering, which usually employs more than 20 copies,^{39–44} this method does not suffer from a decrease of exchange rate with the increase of the system size. In practice, only 2–3 copies are employed, yet it is still capable of maintaining a high rate of exchange between neighboring copies.

There are three important time intervals in PCST: dt_{CE} , the time step in the canonical ensemble at a fixed temperature, i.e., the simulation time step; dt_{walk} , the time step for integrating the Langevin equation guiding the temperature random walk; and dt_{ex} , the time interval to attempt an exchange of parameters between different copies. The flowchart of PCST is shown in Fig. S1 of the [supplementary material](#).

B. Coordinate-based enhanced sampling

Different from generalized ensembles, which do not change the shape of the energy landscape, the coordinate-based (or energy-biasing) approaches can cross the energy barriers by incorporating additional energy terms to the original Hamiltonian. The biased energy term can be sophisticatedly designed to make the system favor or disfavor conformations with certain reaction coordinates.^{45–48} In our application, a restraint term serving as the biased energy makes the system perform random walk around a “reference state” where the restraint reaches its minima. Note that all the energy-biasing methods can be used together with tempering methods to generate enhanced targeted sampling in the configurational space.

1. Structure-based model (SBM) restraints

We introduce the structure-based model (SBM)^{18,19} restraints to increase the sampling efficiency around an important conformation. While initially proposed as a solvent-free model to simulate protein conformational changes, SBM has many advantages over coarse-grained models. For example, the SBM potential can generate a funnel-like energy landscape by itself, which can be compared to the protein energy landscape in water. The mathematical form of the SBM bias potentials is

$$\begin{aligned} V = & \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{improper dihedrals}} \epsilon_\chi (\chi - \chi_0)^2 + \sum_{\text{backbone}} \epsilon_{BB} F_D(\phi) \\ & + \sum_{\text{sidechain}} \epsilon_{SC} F_D(\phi) + \sum_{\text{contacts}} \epsilon_C \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^6 \right] \\ & + \sum_{\text{non-contacts}} \epsilon_{NC} \left(\frac{\sigma_{ij}}{r} \right)^{12}, \end{aligned} \quad (2)$$

where $F_D(\phi) \equiv [1 - \cos(\phi - \phi_0)] + [1 - \cos(3(\phi - \phi_0))]/2$, σ_{ij} and all the ϵ s are pre-set parameters of SBM; r , θ , χ , and ϕ are calculated from the current coordinates of the protein and r_0 , θ_0 , χ_0 , and ϕ_0 are calculated from the coordinate of the reference state which corresponds to the important conformation we want to sample around. The selection procedure of “contacts” follows the “shadow map” scheme,⁴⁹ which only counts the atom pairs that are in each other’s “first coordination shell.” Specifically, atom pairs are considered as contacts only when the distance between these two atoms is in a certain range (e.g., 4–8 Å) and no other atoms between them are “blocking the way” (Fig. S2 of the [supplementary material](#)).

The funnel-like landscape of SBM inspired us to use it as structural restraints in the configuration space. Compared with the elastic network approaches that use harmonic functions, SBM has a much flatter energy landscape and less energy penalties for states that are far away from the reference state. This encourages the sampling of states with large deviations from the reference state but lower free energy. SBM is particularly useful in the protein structure refinement when the reference state is accurate enough to reflect most information about the native state. As an energy-biasing method, SBM is orthogonal to the tempering methods such as PCST.

2. Temperature-dependent restraining protocol

While the restraining strength is usually kept as constant in conventional schemes, the amplitude of a restraint is temperature-dependent in our restraining protocol.⁵⁰ At room temperature, the restraining strength goes to zero such that the original energy landscape can be recovered. The restraint

strength reaches its maximum at the highest temperature, preventing the system from drifting too far away from the reference state during the energy barrier crossing. Specifically, a biased potential $V(X)$, expressed as a function of molecular coordinates X , serves as restraints and can be added to the generalized ensemble without disturbing room-temperature properties. This is achieved via a temperature-dependent Hamiltonian $H(X) = H_0(X) + \lambda(T)V(X)$ where $H_0(X)$ is the original Hamiltonian and $\lambda(T)$ is a linear function of temperature T . Two parameters $\{\lambda_{min}, \lambda_{max}\}$ control the shape of $\lambda(T)$, representing the value of λ at T_{min} (set as zero in our case) and the one at T_{max} , respectively. The reason of choosing the linear form of $\lambda(T)$ is that linear functions can be combined with PCST seamlessly by rewriting the stochastic equation above as

$$d\left(\frac{1}{\beta}\right) = \left(E' - \overline{E'}(\beta) - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2}\right)dt + \frac{\sqrt{2}}{\beta}dW_t, \quad (3)$$

where $E' \equiv E(X) - V(X)$ denotes the difference between the original potential energy and the biased energy. It can be shown that the new equation generates the desired probability distribution as

$$P(X, \beta) \propto Z^{-1}(\beta)e^{-\beta[E(X)+\lambda(T)V(X)]}\beta^{-\gamma}e^{-(\beta-\beta_i^0)^2/2\sigma_i^2}. \quad (4)$$

As mentioned above, this temperature-dependent restraining scheme adds biases on the high temperature range while it de-biases the energy landscape at room temperature. This brings some significant improvement over conventional restraints, especially when the latter contain too much inaccurate information about the system and fail to guide the simulation toward the correct direction.

FIG. 1. Time series of GDT_TS of the MD trajectories corresponding to 18 CASP10 targets. In each trajectory, the red line and the black line represent the high temperature copy and the low temperature copy in PCST, respectively. The blue line indicates GDT_TS of the initial model.

C. Model selection methods

While MD simulation can generate sufficient number of structures, it is essential to select proper candidates for practical use. Currently, two mainstream methods with totally different philosophies are used for model selection. Widely used in MD simulations, the clustering method^{50–52} attempts to divide the whole trajectory into “clusters” such that each cluster can represent a conformation. In ideal cases when the system reaches the equilibrium, the largest cluster corresponds to the native state since it is the most thermodynamically stable. The second approach, the statistical potential method,^{53–56} uses knowledge-based information to score each structure and select ones with higher scores. Each method has its advantages and disadvantages. Our novel model selection method combines these two methods to effectively select only tens of structures, with higher model accuracy, from more than 10^5 candidates along the whole trajectory. The method is especially designed for tempering simulations, in which any snapshot of the simulation contains both the coordinates of the current state and its corresponding temperature.

We start the model selection process with model elimination according to physics principles (Step I). In the generalized ensemble, the probability for a state $\{X, \beta\}$ to transit to another state $\{X, \beta_0\}$ is

$$\frac{P(X, \beta_0)}{P(X, \beta)} = \frac{Z(\beta)w(\beta)}{Z(\beta_0)w(\beta_0)} e^{-(\beta_0 - \beta)E(X)}, \quad (5)$$

where $Z(\beta)$ is the canonical partition function, $E(X)$ is the potential energy, and $w(\beta)$ denotes the weight function of the canonical ensemble at temperature β . The transition probability is also proportional to the overlapping area between energy histograms of these two canonical ensembles and usually decreases as $|\beta_0 - \beta|$ increases. Thus, if a state is sampled at a high temperature, it is unlikely to be sampled at a low temperature. This inspires us to add a temperature threshold to the trajectory such that the states corresponding to temperatures above the threshold (e.g., 300 K) are not considered as a candidate, given that the native structure of a protein is experimentally measured at room temperature. On the other hand, as the temperature along the simulation changes continuously, the setup of the temperature threshold naturally divides the whole trajectory into clusters. By eliminating the states in the high temperature range, the number of candidates can be reduced to 10^3 – 10^4 , which is 5%–10% of the total number of states.

TABLE I. Improvements in C α -RMSD and GDT_TS of the best PCST-SBM models (Best) compared to the initial models provided by CASP (Initial). TR644 to TR750 are CASP10 targets and TR862 to TR870 are CASP12 targets.

Target	Size	Initial		Best	
		RMSD (Å)	GDT_TS	Δ RMSD (Å)	Δ GDT_TS
TR644	141	2.71	84.75	−1.49	8.51
TR655	175	4.65	68.86	−0.99	4.43
TR661	185	2.74	80.68	−0.9	4.59
TR662	75	2.03	84.00	−0.65	6.33
TR663	152	3.37	69.24	−1.69	20.56
TR671	88	7.72	55.68	−5.61	31.25
TR674	132	3.44	85.80	−1.34	6.06
TR688	185	2.52	78.24	−0.53	6.81
TR696	100	3.52	71.50	−0.82	9.75
TR698	119	4.65	65.55	−0.8	3.57
TR704	235	2.78	70.13	−1.2	15.07
TR705	96	4.71	64.84	−1.81	11.72
TR708	196	4.63	86.48	−0.79	2.04
TR710	194	2.44	75.13	−0.27	7.60
TR720	202	8.52	58.46	−2.47	4.80
TR723	132	2.23	85.11	−0.59	7.07
TR747	98	1.96	83.61	−0.21	5.00
TR750	182	2.12	77.75	−0.55	5.49
TR862	93	5.55	58.60	−1.19	5.65
TR866	104	3.27	79.57	−1.4	12.02
TR868	105	1.97	80.95	−0.8	11.19
TR869	104	12.36	38.94	0.14	0.96
TR870	109	7.64	42.43	−1.38	10.32
Average	139	4.24	71.58	−1.19	8.73

Step II is to score the candidates using a statistical potential. The one we chose is generalized orientation-dependent, all-atom statistical potential (GOAP),⁵⁵ one of the most effective statistical potentials for recognizing native states of proteins. Instead of ranking all the single structures using GOAP, we rank all the clusters by calculating the average GOAP score of each cluster. In practice, the size of a cluster might be too large and the variance of the GOAP score within the cluster is huge. In this case, we divide the cluster into multiple sub-clusters such that the variance of the GOAP score within each sub-cluster is controlled. The ranking process lets us select clusters with higher scores and eliminate clusters with lower scores. For instance, leaving only 10–20 clusters in the candidate pool would reduce the number of candidates to 10^2 – 10^3 . The final step involves generation of our final models using the clusters provided in the previous step. This is achieved by first

FIG. 2. Time series of C α -RMSD (left) and GDT_TS (right) of TR671 where distinct conformations are observed. The red line and the black line represent the high temperature copy and the low temperature copy in PCST, respectively. The blue line indicates the C α -RMSD (left) and GDT_TS (right) of the initial model.

FIG. 3. Demonstration of the model selection method using a 500 ns PCST-SBM simulation of TR662. In each panel, the red and the black line represent the high temperature copy and the low temperature copy, respectively. (a) Time series of $C\alpha$ -RMSD. (b) Time series of the system temperature. (c) $C\alpha$ -RMSD of the rest of the trajectories after a temperature threshold of 300 K is applied. (d) Corresponding GOAP score (in thousands) of the states in (c).

aligning all the structures within each cluster via least square fitting and obtaining only one averaged structure representing each cluster.

III. RESULTS

A. MD simulation of CASP10 and CASP12 targets using PCST-SBM method

We used the PCST-SBM method to refine low-accuracy models, 18 selected from the refinement category of CASP10 and 5 from that of CASP12. The targets have a great variety of sizes, ranging from 75 to 235 residues, and contain different secondary structure components. The $C\alpha$ -RMSD and the GDT_TS of the initial models, compared to the native structures, are in the range of 1.96–12.36 Å and 38.94–86.48, respectively.

All-atom MD simulation in explicit solvent was performed using GROMACS 4.5.^{57–60} The TIP3P model⁶¹ was used for explicit solvent, AMBER99SB-ILDN⁶² was used as the force field, and explicit ions were used for neutralizing net charges. For CASP10 targets, each was simulated for 500 ns, and for CASP12 targets, each was simulated for 1000 ns.

In each simulation, we set the initial model provided by the CASP organizers as both the “reference state” of SBM and the “starting point” of simulation. The SMOG server at Rice University⁶³ was used to extract the SBM parameters from the reference state. The temperature range for our simulation was 293–500 K. Two copies were employed in PCST, with the parameters $\{\beta^0, \sigma\}$ set at $\{0.38, 0.05\}$ for the low temperature copy and $\{0.27, 0.13\}$ for the high temperature copy (the peaks of the Gaussian temperature distribution were at 316 K and 445 K, respectively). The time step for MD integration was 0.002 ps. The time step for integrating the Langevin equation was 0.04 ps, which was the same as the neighbor-list refreshing interval. For the PCST method, the time interval of the exchange attempt was 10 ns. For each system, we calculated time series of GDT_TS to quantify the model quality (Fig. 1, only CASP10 targets are shown).

TABLE II. GDT_TSs in the refinement. TR644 to TR750 are CASP10 targets and TR862 to TR870 are CASP12 targets. Numbers in Δ CSPR5 are the GDT_TS improvements of the best models among all CASP participants evaluated by the organizers using the native structures (each participating group contributed 5 models). Numbers in Top5 are the GDT_TS improvements of the best models evaluated by the native structures among the Top5 choices provided by our blind selection scheme. Numbers in Rank5 are the ranking of the models in Top5 among the five blindly selected models; therefore, 1 in Rank5 means the best model indicated by the blind selection scheme is the same as the one evaluated by the native structures (there are totally 7 targets with 1 in Rank5). Numbers in Δ CSPR1 are the GDT_TS improvements for the best models declared by the CASP participants among all participating groups. Numbers in Top1 are the GDT_TS improvements for the best models provided by our blind selection scheme (11 of them are better than the initial models and 7 of them are better than those in Δ CSPR1).

Target	Initial	Δ CSPR5	Top5	Rank5	Δ CSPR1	Top1
TR644	84.75	2.84	4.97	2	2.13	−5.14
TR655	68.86	1.71	2.29	2	0.85	−0.71
TR661	80.68	0.67	1.21	5	0.67	−0.82
TR662	84.00	4.33	4.33	4	2.33	−2.67
TR663	69.24	8.06	15.63	2	5.60	13.16
TR671	55.68	10.51	19.60	4	4.55	18.75
TR674	85.80	1.7	2.46	2	1.70	−0.95
TR688	78.24	1.9	1.90	5	1.49	−2.16
TR696	71.50	4.5	3.75	3	4.00	1.25
TR698	65.55	2.1	0.21	2	1.47	−4.42
TR704	70.13	5.85	12.77	1	3.09	12.77
TR705	64.84	6.51	7.82	1	5.21	7.82
TR708	86.48	1.02	−0.64	1	0.89	−0.64
TR710	75.13	5.15	3.35	3	2.70	1.80
TR720	58.46	3.03	0.13	1	1.89	0.13
TR723	85.11	6.3	6.49	4	3.06	4.39
TR747	83.61	2.5	0.83	4	1.67	−0.83
TR750	77.75	3.71	0.96	4	3.71	−0.28
TR862	58.60	4.57	1.08	1	4.57	1.08
TR866	79.57	9.61	10.81	1	9.61	10.81
TR868	80.95	8.34	−2.57	4	8.34	−7.09
TR869	38.94	8.18	−2.58	2	8.18	−2.88
TR870	42.43	9.41	5.96	1	5.73	5.96
Average	71.58	4.89	4.38		3.63	2.14

The models from PCST-SBM were compared with the initial models (Table I). The best model in each target (identified by comparing with the known native structure) is consistently better than the initial model. The average GDT_TS improvement is 8.73 over the initial models while the greatest GDT_TS enhancement (TR671) reaches 31.25. Furthermore, these best models from PCST-SBM exhibit much smaller C α -RMSD from the native structures. The average improvement in C α -RMSD is 1.19 Å for all 23 systems, with the highest being TR671 (at 5.61 Å). Importantly, out of the 23 systems, the best models are within 2 Å C α -RMSD from the native structures for 11 of them (TR644, TR661, TR662, TR663, TR688, TR704, TR723, TR747, TR750, TR866, and TR868), within 2–3 Å C α -RMSD for five systems (TR671, TR674, TR696, TR705, and TR710), and within 3–4 Å C α -RMSD for three systems (TR655, TR698, and TR708). Clearly, PCST-SBM can sample the configurational space around the initial model thoroughly and the best models of the majority of these tested systems are accurate enough to be search models for molecular replacement (which generally requires C α -RMSD smaller than 2–3 Å from native structures) in phasing X-ray diffraction data. Particularly, for several targets, distinct conformations can be observed in the time series of C α -RMSD and GDT_TS. Results for the most striking example TR671 are shown in Fig. 2.

Note that the results in Table I only reflect the quality of the best models in the refinement and the results of the blind model selection procedure are in Sec. III B.

B. Blind selection of high-quality models from MD trajectories

We also applied our model selection method to blindly select high-quality models from simulation trajectories (Fig. 3). As indicated in the Methods section, there are three steps in this process. We use TR662 as an example to illustrate the procedure. Step I is to apply a temperature threshold

to eliminate the states in the high temperature range that may have a good GOAP score but physically forbidden. This step eliminates more than 90% structures in an MD trajectory [Figs. 3(b) and 3(c)]. Step II is to score the candidates as clusters using the statistical potential GOAP instead of directly ranking all the single structures, which enhances the accuracy by canceling out systematic errors of the statistical potential on each single structure [Fig. 3(d)]. The final step (Step III) involves the generation of only one model representing one cluster by structural averaging, which also serves to refine side-chain packing and the hydrogen network within a cluster of structures.

Similar to CASP competition, we selected refinement results in terms of Top-1 group and Top-5 group. To evaluate the quality of the models we selected, we computed the GDT_TS with respect to the known native structures. The results are shown in Table II.

In Top5 group (Top5), 20 out of 23 targets have their GDT_TSs better than the initial models (Initial), out of which 10 targets are better than the best results in CASP (Δ CSPR5). For the GDT_TS improvements in the best models provided by our blind selection scheme (Top1), 11 of them are better than the initial models (Initial) and 7 of them are better than CASP (Δ CSPR1), and these 7 models are TR663, TR671, TR704, TR705, TR723, TR866, and TR870. The improvements in some targets are particularly significant such as TR671, TR704, TR663, and TR705. The final GDT_TSs of these targets are significantly better than those of the CASP results and their differences (Top1 over Δ CSPR1) are 14.20, 9.68, 7.56, and 2.61, respectively. The superpositions of these targets are shown in Fig. 4.

It is very important to point out that the results of our method listed in Table II are the results by our single group, while the results of CASP (Δ CSPR5 and Δ CSPR1) are the best results among ALL participants. This may suggest that, in terms of the performance of the single group, our results are more significant than what the numbers indicate.

FIG. 4. The refined structures of four targets, TR704 (a), TR663 (b), TR671 (c), and TR705 (d) are compared to the native structure. In each case, the native structure is shown in green, the initial model is in pink, and the final model generated by our refinement protocol is shown in red. The regions with large structural improvements in our final model are highlighted by blue dashed lines.

IV. CONCLUDING DISCUSSION

In this paper, we combined a temperature-based enhanced sampling method, parallel continuous simulated tempering (PCST), and a temperature-dependent coordinate-based restraint, the structure-based model (SBM), to perform an effective search of configurational space during structural refinement. A novel model selection method was also developed to blindly select high-quality models from very long simulation trajectories. The simulation protocol yielded final models that are better than initial models. For Top-1 group, 7 out of 23 targets had better models (greater GDT_TS) than the CASP participants. For Top-5 group, 10 out of 23 were better.

The temperature-based sampling method, the PCST method, permits faster rate of energy barrier crossing. It has two copies of simulation, each with a Gaussian weight function added in the temperature space. These two copies play different roles during the simulation. The low-temperature copy facilitates better sampling within energy basins without large changes in the energy landscape, while the high-temperature copy is responsible for crossing the large energy barriers. We also tried, for example, 3 copies, but the improvement did not outweigh the increased computational burden.

The temperature-dependent coordinate-based restraint, SBM, was used to prevent the system from drifting too far from the initial structure especially at higher temperatures. A novel feature is that the strength of the restraint is linearly decayed from high to low temperatures so that the low temperature sampling is less influenced by the restraints. Such a scheme can be generalized to almost any kind of restraints. With this kind of restraining scheme, the main sampling efficiency is achieved at lower temperatures where weaker restraints are imposed. At higher temperatures, the system could be biased toward the restraining model.

The blind selection scheme uses the empirical potential GOAP⁵⁵ in combination with a clustering method. An advantageous feature of the method is the use of cluster scoring to select high-quality clusters instead of single structures, which avoids many limitations of statistical potentials. Empirical potentials applied alone are usually not accurate enough for model selection; however, the cluster averaging reduces the intrinsic bias and noise existing in the empirical potentials. Although there are many ways of clustering the structures along a simulation trajectory, the one in our blind selection scheme, which utilizes temperature criteria for initial screening, seems to be effective. This temperature based initial screening eliminates all structures sampled at higher temperatures, only the ones in lower temperature range are used for further analysis. The model-averaging step (the final step) generates only one model to represent one cluster.

We want to emphasize that the capability of blindly selecting low-free-energy conformers from simulation trajectories is one of the most important and most challenging problems in structural prediction. There has been little progress in this aspect in the field. Although the blind selection scheme reported in this paper seems to be effective, we believe that

there is a long way to go for optimizing the scheme. Another problem in protein structure refinement is that, aside from the possible errors in force fields and limitations in sampling, the outcome of refinement depends on many other factors that are often beyond one's control. For example, some native structures may be stable due to the existence of partner proteins, co-factors, crystal packing forces, etc. None of these factors are usually taken into account in refinement simulation. This is surely an issue that many real applications would face, not just CASP participants.

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for figures of the PCST-SBM method, including the flowchart of the PCST method and SBM potential function.

ACKNOWLEDGMENTS

J.M. thanks support from the National Institutes of Health (Nos. R01-GM067801 and R01-GM116280) and the Welch Foundation (Q-1512). Q.W. thanks support from the National Institutes of Health (Nos. R01-AI067839 and R01-GM116280), the Gillson-Longenbaugh Foundation, and The Welch Foundation (Q-1826). The authors thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Stampede2 at the TACC through Allocation No. MCB150013.

- ¹Y. Zhang, *Curr. Opin. Struct. Biol.* **18**(3), 342 (2008).
- ²E. Faraggi, Y. Yang, S. Zhang, and Y. Zhou, *Structure* **17**(11), 1515 (2009).
- ³Y. Zhang, *BMC Bioinf.* **9**(1), 40 (2008).
- ⁴S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker, *Proteins: Struct., Funct., Bioinf.* **77**(S9), 89 (2009).
- ⁵M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291 (2000).
- ⁶J. Peng and J. Xu, *Proteins* **79**(Suppl. 10), 161 (2011).
- ⁷S. T. Wu, J. Skolnick, and Y. Zhang, *BMC Biol.* **5**, 17 (2007).
- ⁸Y. J. Huang, B. Mao, J. M. Aramini, and G. T. Montelione, *Proteins* **82**(Suppl. 2), 43 (2014).
- ⁹V. Mariani, F. Kiefer, T. Schmidt, J. Haas, and T. Schwede, *Proteins: Struct., Funct., Bioinf.* **79**(S10), 37 (2011).
- ¹⁰J. Moult, K. Fidelis, A. Kryshchuk, T. Schwede, and A. Tramontano, *Proteins: Struct., Funct., Bioinf.* **82**, 1 (2014).
- ¹¹J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis, *Proteins: Struct., Funct., Genet.* **23**(3), ii (1995).
- ¹²J. Moult, *Curr. Opin. Struct. Biol.* **15**(3), 285 (2005).
- ¹³T. Nugent, D. Cozzetto, and D. T. Jones, *Proteins: Struct., Funct., Bioinf.* **82**, 98 (2014).
- ¹⁴J. L. MacCallum, A. Perez, M. J. Schnieders, L. Hua, M. P. Jacobson, and K. A. Dill, *Proteins* **79**(Suppl. 10), 74 (2011).
- ¹⁵K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**(6055), 517 (2011).
- ¹⁶S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **109**(44), 17845 (2012).
- ¹⁷T. Zang, L. Yu, C. Zhang, and J. Ma, *J. Chem. Phys.* **141**(4), 044113 (2014).
- ¹⁸P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, *Proteins: Struct., Funct., Bioinf.* **75**(2), 430 (2009).
- ¹⁹C. Clementi, H. Nymeyer, and J. N. Onuchic, *J. Mol. Biol.* **298**(5), 937 (2000).
- ²⁰A. Zemla, C. Venclovas, J. Moult, and K. Fidelis, *Proteins: Struct., Funct., Genet.* **37**(Suppl. 3), 22–29 (1999).

- ²¹A. Zemla, Č. Venclovas, J. Moulton, and K. Fidelis, *Proteins: Struct., Funct., Genet.* **45**(Suppl. 5), 13 (2001).
- ²²A. Zemla, *Nucleic Acids Res.* **31**(13), 3370 (2003).
- ²³A. Mitsutake, Y. Sugita, and Y. Okamoto, *Biopolymers* **60**(2), 96 (2001).
- ²⁴B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68**(1), 9 (1992).
- ²⁵B. Baumann, *Nucl. Phys. B* **285**(3), 391 (1987).
- ²⁶F. G. Wang and D. P. Landau, *Phys. Rev. Lett.* **86**(10), 2050 (2001).
- ²⁷P. Dayal, S. Trebst, S. Wessel, D. Wurtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith, *Phys. Rev. Lett.* **92**(9), 097201 (2004).
- ²⁸Q. L. Yan, R. Faller, and J. J. de Pablo, *J. Chem. Phys.* **116**(20), 8745 (2002).
- ²⁹Q. L. Yan and J. J. de Pablo, *Phys. Rev. Lett.* **90**(3), 035701 (2003).
- ³⁰M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, *Phys. Rev. E* **66**(5), 056703 (2002).
- ³¹E. Marinari and G. Parisi, *Europhys. Lett.* **19**(6), 451 (1992).
- ³²R. H. Swendsen and J. S. Wang, *Phys. Rev. Lett.* **57**(21), 2607 (1986).
- ³³M. Falcioni and M. W. Deem, *J. Chem. Phys.* **110**(3), 1754 (1999).
- ³⁴Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.* **314**(1-2), 141 (1999).
- ³⁵D. J. Earl and M. W. Deem, *Phys. Chem. Chem. Phys.* **7**(23), 3910 (2005).
- ³⁶P. Jiang, F. Yasar, and U. H. E. Hansmann, *J. Chem. Theory Comput.* **9**(8), 3816 (2013).
- ³⁷C. Zhang and J. P. Ma, *J. Chem. Phys.* **130**(19), 194112 (2009).
- ³⁸C. Zhang and J. P. Ma, *J. Chem. Phys.* **132**(24), 244101 (2010).
- ³⁹G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, *J. Am. Chem. Soc.* **128**(41), 13435 (2006).
- ⁴⁰C. Camilloni, D. Provati, G. Tian, and R. A. Broglia, *Proteins* **71**(4), 1647 (2008).
- ⁴¹A. Barducci, M. Bonomi, M. K. Prakash, and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **110**(49), E4708 (2013).
- ⁴²R. Zhou, B. J. Berne, and R. Germain, *Proc. Natl. Acad. Sci. U. S. A.* **98**(26), 14931 (2001).
- ⁴³A. Plazinska, W. Plazinski, and K. Jozwiak, *J. Comput. Chem.* **35**(11), 876 (2014).
- ⁴⁴A. E. García and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **100**(24), 13898 (2003).
- ⁴⁵G. M. Torrie and J. P. Valleau, *J. Comput. Phys.* **23**(2), 187 (1977).
- ⁴⁶C. Bartels and M. Karplus, *J. Phys. Chem. B* **102**(5), 865 (1998).
- ⁴⁷A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **99**(20), 12562 (2002).
- ⁴⁸A. Barducci, G. Bussi, and M. Parrinello, *Phys. Rev. Lett.* **100**(2), 020603 (2008).
- ⁴⁹J. K. Noel, P. C. Whitford, and J. N. Onuchic, *J. Phys. Chem. B* **116**(29), 8692 (2012).
- ⁵⁰C. Zhang and J. Ma, *Proc. Natl. Acad. Sci. U. S. A.* **109**(21), 8139 (2012).
- ⁵¹J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, *J. Chem. Theory Comput.* **3**(6), 2312 (2007).
- ⁵²A. Raval, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.* **80**(8), 2071 (2012).
- ⁵³J. Zhang and Y. Zhang, *PLoS One* **5**(10), e15386 (2010).
- ⁵⁴C. Zhang, S. Liu, Q. Zhu, and Y. Zhou, *J. Med. Chem.* **48**(7), 2325 (2005).
- ⁵⁵H. Zhou and J. Skolnick, *Biophys. J.* **101**(8), 2043 (2011).
- ⁵⁶M. Lu, A. D. Dousis, and J. Ma, *J. Mol. Biol.* **376**(1), 288 (2008).
- ⁵⁷H. J. C. Berendsen, D. Vanderspoel, and R. Vandrunen, *Comput. Phys. Commun.* **91**(1-3), 43 (1995).
- ⁵⁸E. Lindahl, B. Hess, and D. van der Spoel, *J. Mol. Model.* **7**(8), 306 (2001).
- ⁵⁹D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comput. Chem.* **26**(16), 1701 (2005).
- ⁶⁰B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, *J. Chem. Theory Comput.* **4**(3), 435 (2008).
- ⁶¹W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**(2), 926 (1983).
- ⁶²K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, *Proteins* **78**(8), 1950 (2010).
- ⁶³J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford, *PLoS Comput. Biol.* **12**(3), e1004794 (2016).