Using Structure Reservoirs to Accelerate Biomolecular Simulations

A Dissertation Presented

by

Koushik Kasavajhala

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2020

Stony Brook University

The Graduate School

Koushik Kasavajhala

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Carlos L. Simmerling – Dissertation Advisor
Professor, Department of Chemistry, Stony Brook University

Dr. Daniel P. Raleigh - Chairperson of Defense
Professor, Department of Chemistry, Stony Brook University

Dr. David Green – Third Member of Defense
Associate Professor, Department of Chemistry, Stony Brook University

Dr. Evangelos Coutsias – Outside Member of Defense
Professor, Department of Applied Mathematics and Statistics, Stony Brook University

This dissertation is accepted by the Graduate School

Eric Wertheimer
Dean of the Graduate School

Abstract of the Dissertation

Using Structure Reservoirs to Accelerate Biomolecular Simulations

by

Koushik Kasavajhala

Doctor of Philosophy

in

Chemistry

Stony Brook University

2020

Development of accurate potential energy functions and advances in computing hardware have made molecular dynamics simulations an indispensable tool for studying biomolecular processes. To model biomolecular processes accurately via simulations, at the very least, the end states (such as native folds of proteins) have to be modeled accurately, since these can be directly compared to experiments. Currently, simulations can reliably find native structures of medium sized biomolecules. However, identifying if the native structure is the preferred structure is still a challenging task since it requires extensive sampling of both, native and non-native structures. To address this, Reservoir Replica Exchange Molecular Dynamics (RREMD) enhanced sampling method was developed, in which a set of pre-generated structure snapshots representing native and non-native structures are used as a reservoir. These reservoir structures are then coupled to the replicas using Monte Carlo moves to allow fast structural transitions between native and non-native states followed by thermal reweighting to identify the preferred structure at low temperatures. While the method can, in theory, quickly predict the most preferred structure, the

availability of the code on only the CPUs prohibited testing the method on a wide variety of systems.

In this work, we ported the RREMD code onto GPUs making it 20x faster than the CPU code. Then, we explored protocols for building reservoirs and tested how each choice affects the accuracy of RREMD. Our protocols show that, with careful selection of structure snapshots, the method can accurately reproduce Boltzmann-weighted ensembles obtained by much more expensive conventional REMD, with at least 25x faster convergence rates even for larger proteins (>50 amino acids) compared to conventional REMD. We also demonstrate that these structure reservoirs can be used to predict the accuracy of new force fields and also the effects of mutations, thereby facilitating rapid testing of new force field parameters and also design of new peptides.

Finally, since it is not necessary for the reservoir to be Boltzmann-weighted, structures obtained from non-MD methods such as homology-based models or Rosetta-based models can be integrated into this method making it a powerful tool that combines physics-based approaches with non-physics-based approaches which can lead to better structure predictions of biomolecules.

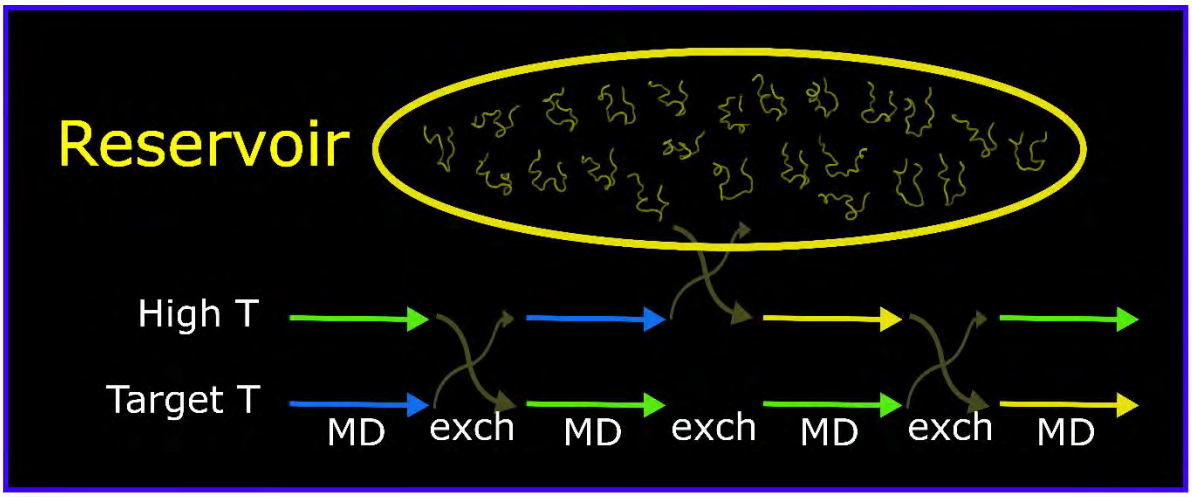
Dedication Page

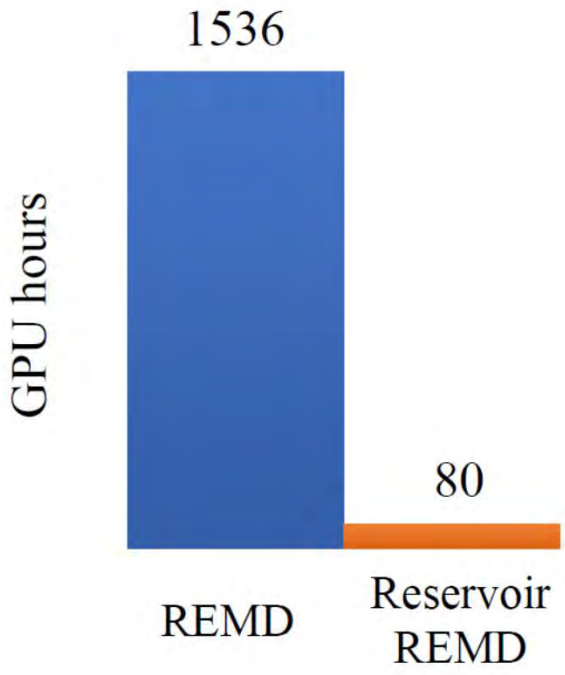
This work is dedicated to Alysse-Danielle Kavanagh, Dr. James Allen Maier, Dr. Alberto Perez, and Dr. Carlos L. Simmerling – without the efforts, advise, and guidance of whom I wouldn't be alive today.

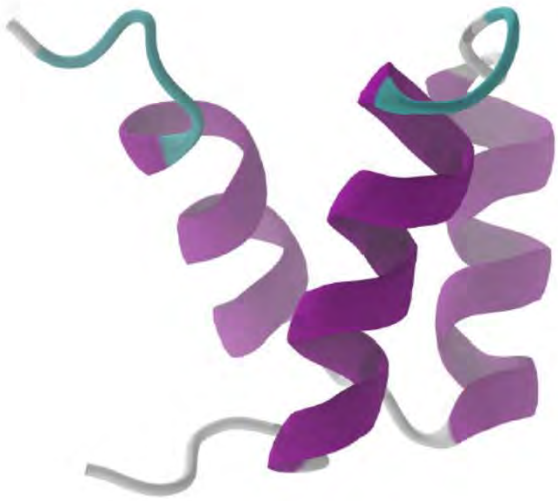
"Because time is all that we have got to live, time is the currency of life."

– Dr. Daniel Kahneman, Artificial Intelligence Podcast, 2020.

Frontispiece







Homeodomain
(52 residues)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GB	Generalized Born
MD	Molecular Dynamics
REMD	Replica Exchange MD
T-REMD	Temperature Replica Exchange MD
RREMD	Reservoir REMD
B-RREMD	Boltzmann-weighted RREMD
nB-RREMD	non-Boltzmann RREMD
AL	Average-Linkage
WL	Ward-Linkage
CRE	Cluster Representative Energy
CAE	Cluster Average Energy
AMBER	Assisted Model Building with Energy Refinement
RMSD	Root Mean Square Deviation
PDB	Protein Data Bank
Trp-cage	Trp-cage miniprotein
MSM	Markov State Model
REST	Replica Exchange with Solute Tempering
V-REMD	Viscosity REMD
MMREMD	Mass-manipulating REMD
Hybrid-REMD	Hybrid-solvent REMD
Hybrid-nB-RREMD	Hybrid-solvent non-Boltzmann RREMD

Acknowledgments

I thank my advisor Prof. Carlos L. Simmerling for guiding and supporting my research during my Ph.D. candidacy. Thank you for your patience and guidance and also for improving my scientific writing, presentation, and communication skills.

I thank Prof. Daniel Raleigh, Prof. David Green, and Prof. Evangelos Coutsias for serving as my dissertation committee members. I thank them for supporting me during the darkest days of my life, and for their help on my research projects.

I thank Prof. Ken A. Dill, Prof. Carlos Simmerling, Prof. Robert Rizzo, Prof. Dima Kozakov, Prof. Evagelos Coutsias, and Prof. Gabor Balaszi, for creating a great environment at the Laufer Center.

I thank the Laufer Center for Physical and Quantitative Biology for access to computational resources and support, and Dr. Feng Zhang for managing computing resources.

I thank Dr. U. Deva Priyakumar, my advisor during my MS by Research program at IIIT-H, without the efforts of whom I wouldn't have made it to Stony Brook University, and also for helping me secure a job as Senior Scientist I, at Schrödinger, India.

I thank all my friends that made my stay at Stony Brook University memorable: Dr. Alberto Perez, Dr. Emiliano Brini, Dzmitry Padhorny, Catherine Tang, Roy Nassar, Dr. Adam De Graff, Dr. Jason A. Wagoner, Dr. Michael Hazoglou, Dr. Lane Votapka, Lauren Prentis, Stephen Telehany, Dr. Tiffany Victor, Dr. Upasana Roy, Dr. Coray McBean, Dr. Luisa Le Donne, Dr. Pratik Kumar, Dr. David Hewitt, Dr. Anusha Shankar, Amogh Akshintala, Alwin James, Mihir Umarani, Yashas Masurekar, Mayank Seth, Dr. Joseph Underwood, Gaurav Guleria, Vamiq Mohammed Mustahsan, and Zubin Darbari.

I thank Katherine Hughes, Nancy Rohring, Diane Vigliotta, and Eileen Dowd, for helping me out at various times during my Ph.D.

I thank all my family members, especially my uncle, Jammi Venkat Rao, and my brother, Vamsee for supporting and helping me during this time.

I thank Baruch Spinoza, Immanuel Kant, Jean-Jacques Rousseau, Will Durant, Ariel Durant, Dr. Werner Heisenberg, Dr. Lex Fridman (Alexei Fedotov), the Joe Rogan Experience, and the Artificial Intelligence podcast, for helping me find solace through their thoughts during the darkest days of my life.

Finally, I thank past and present members of the Simmerling Lab: Dr. James Allen Maier, Dr. Hai Nguyen, Dr. Kevin Hauser, Dr. He Huang (Agnes), Kenneth Lam, Kellon Belfon, Lauren Raguette, Lauren's Mom, Lucy Fallon, Dr. Chuan Tian, Dr. Junjie Zou, Yuzhang Wang, Darya Stepanenko, Jorge Pincay, Kelley Chiu, Abbigayle Cuomo, and Jose Guerra, for their support, help, and all the great moments that I experienced during my stay in the Simmerling Lab. I wouldn't have been able to finish my Ph.D. without torturing all of you with the 100 slides-2 hour-weekly-sampling meetings for the past two years (:P).

List of Publications

1) **K. Kasavajhala**; K. Lam; C. Simmerling, Exploring Protocols to Build Reservoirs to Accelerate Replica Exchange MD simulations. (*to be submitted*)
2) **K. Kasavajhala**; C. Simmerling, Using Structure Reservoirs to Predict the Accuracy of Force Fields and the Effects of Mutations via Thermal Reweighting. (*in preparation*)
3) K. Lam; **K. Kasavajhala**; C. Simmerling, Improving Convergence for RNA stem-loop Simulations using Reservoir Replica Exchange MD. (*in preparation*)
4) J. A. Maier; C. Martinez, **K. Kasavajhala**; L. Wickstrom, K. E. Hauser, C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* 2015, 11(8), 3696-3713.
5) C. Tian; **K. Kasavajhala**; K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q, Wu, C. Simmerling, ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* 2020, 16 (1), 528-552.

1 Introduction

1.1 Modeling biological processes accurately via simulations

Understanding biomolecular processes such as protein folding, drug binding, and protein-protein interactions, besides many others, is crucial to design efficient drugs/treatment procedures against various diseases. It may also be essential to understand the origins of life.[1] Experimental techniques can be used to capture some of these events, however, the resolution at which these events are captured is usually low.

Molecular Dynamics (MD) simulations, on the other hand, provide a higher resolution picture of these biomolecular processes though the "quality" of these simulations is influenced by the accuracy of the underlying MD potential energy function (also known as force fields) used and the computing power available. During the past two decades, significant improvements have been made to the overall quality of MD simulations via improving both, the accuracy of force fields[2-13] and the computing hardware[14-15], resulting in wide-scale applicability of MD simulations to study various biomolecular processes. They have also been used to accurately predict protein-ligand binding affinity[16], study effects of mutations on protein stability[17-18], detect cryptic binding pockets[19-20], design peptides, etc.

Despite the above successes, significant improvements still have to be made to improve the quality of MD simulations. For example, while MD simulations can find the native folds of

some small to medium sized proteins, they still fail to find the native folds of other medium sized proteins[21]. **Figure 1.1** shows the accuracy of simulations compared to experiment in finding (within the simulation time limit) the native folds of 17 different proteins, when the simulation is started from an extended conformation. For 16 out of the 17 proteins, simulations can find structures which have a root mean square deviation (RMSD) of <3 Å to the experimental structure indicating that MD simulations can model small to medium sized proteins reasonably well. However, simulations cannot find the native structure for NuG2 variant – the closest structure is still 4.8 Å away from the experimental structure indicating that simulations need further refinement.

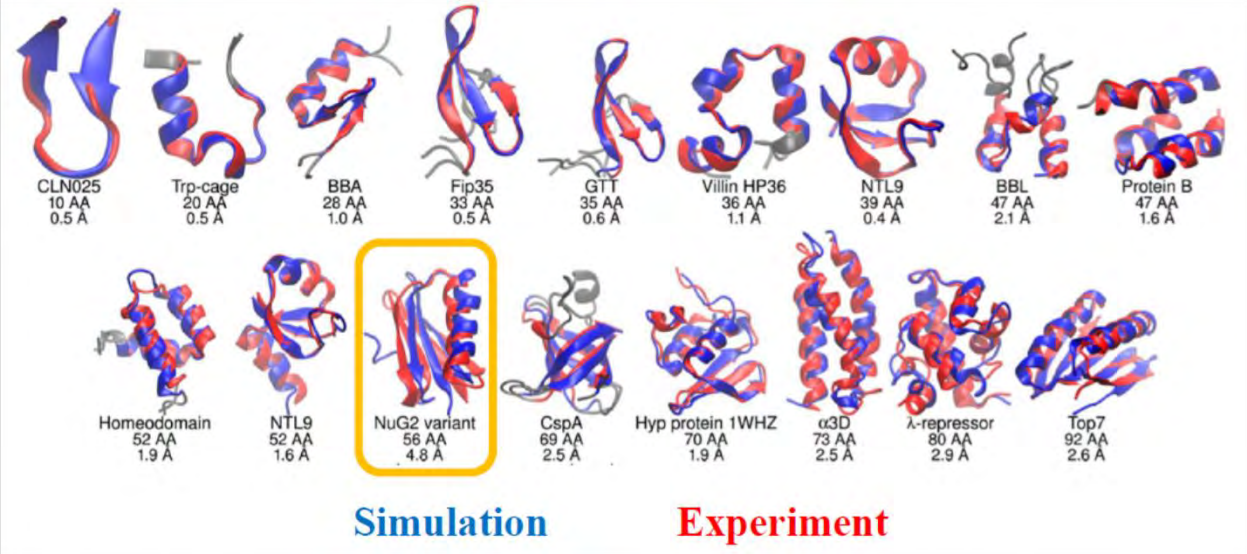


Figure 1-1 Comparing native structures of 17 different proteins obtained from simulations (in blue) vs experiments (in red). Below each protein, the name and length of the protein are shown. The RMSD of the closest native-like structure obtained from simulation is shown below the length of the protein. Grey regions indicate parts of the protein that are poorly characterized in experiments and are excluded from the RMSD analysis. NuG2 variant is enclosed in a yellow box to indicate that the simulation does not find native fold for this protein. Figure adapted from "Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent" by H. Nguyen; J. A. Maier; H. Huang; V. Perrone; C. Simmerling, *J. Am. Chem. Soc.* 2014, 136 (40), 13959-13962. Copyright 2014 by the American Chemical Society.

Moreover, even when MD simulations starting from an extended conformation find the native structure, they do not always favor the native structure[21]. For example, among the 17 proteins shown in **Figure 1.1**, MD simulations favor the native structure for only 9 out of the 17 structures (see **Figure 1.2**).

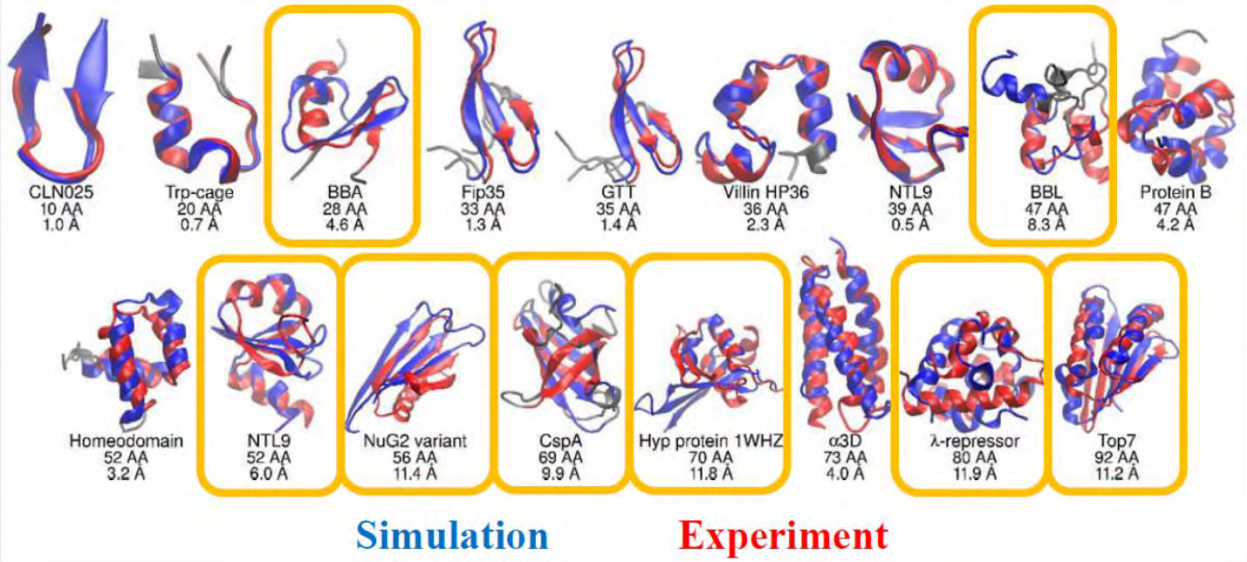


Figure 1-2 Comparing most populated structures obtained from simulations (in blue) vs experiments (in red) for the 17 different proteins shown in **Figure 1.1**, when the simulation is started from extended conformation. Below each protein, the name and length of the protein are shown. The RMSD of the most populated structure obtained from simulation is shown below the length of the protein. Grey regions indicate parts of the protein that are poorly characterized in experiments and are excluded from the RMSD analysis. The proteins for which the most populated structure is not the native structure are enclosed in yellow boxes. Figure adapted from "Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent" by H. Nguyen; J. A. Maier; H. Huang; V. Perrone; C. Simmerling, *J. Am. Chem. Soc.* 2014, 136 (40), 13959-13962. Copyright 2014 by the American Chemical Society.

Since modeling the end states accurately is necessary to determine the thermodynamic equilibrium of a biological process[22], the inability of simulations to find and favor the native protein structures under experimental conditions is concerning. Accurate representations of native states of proteins is essential not only to understand protein stability but also to predict protein-

ligand binding affinity, design novel peptides, etc. Therefore, it is imperative to understand the sources of errors in simulations and to fix them.

The two main possible sources of errors in simulations are shown in **Figure 1.3**. The first source of error stems from the force fields and solvent models used to represent the molecular system. The second source of error stems from sampling i.e. the extent to which the simulation has been carried out. Fixing the errors in force fields and solvent models requires significant manual effort and computational resources and usually takes around 4-5 years currently[10, 13]. On the other hand, addressing errors arising from lack of sampling is much easier and can be solved through either brute force simulations using vast computing resources[14-15, 23] or through various enhanced sampling techniques[24-42] or both.

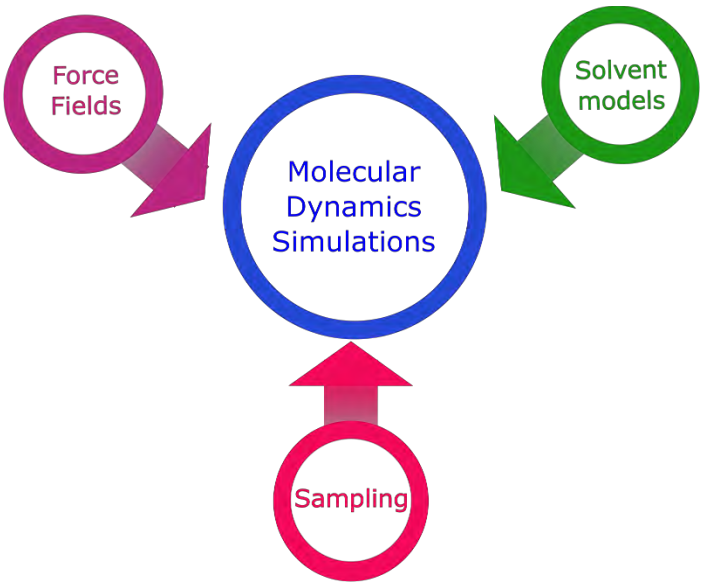


Figure 1-3 Sources of error in Molecular Dynamics Simulations.

More importantly, if the source of error is diagnosed incorrectly, significant manual effort and computational resources will be wasted in fixing areas that need not be fixed. For example, considering the previous example of 17 different proteins, the inability of simulations to favor the

native structure of 8 proteins suggests that the force field might be inaccurate. However, when the same proteins were simulated starting from native structure obtained from experiments instead of an extended structure, 12 (not 9) out of the 17 proteins favored the native structure[21] (see **Figure 1.4**). This indicates that for 3 of the proteins that were modeled incorrectly in **Figure 1.2**, the source of inaccuracy was not from the force fields or solvent model used but from insufficient sampling.

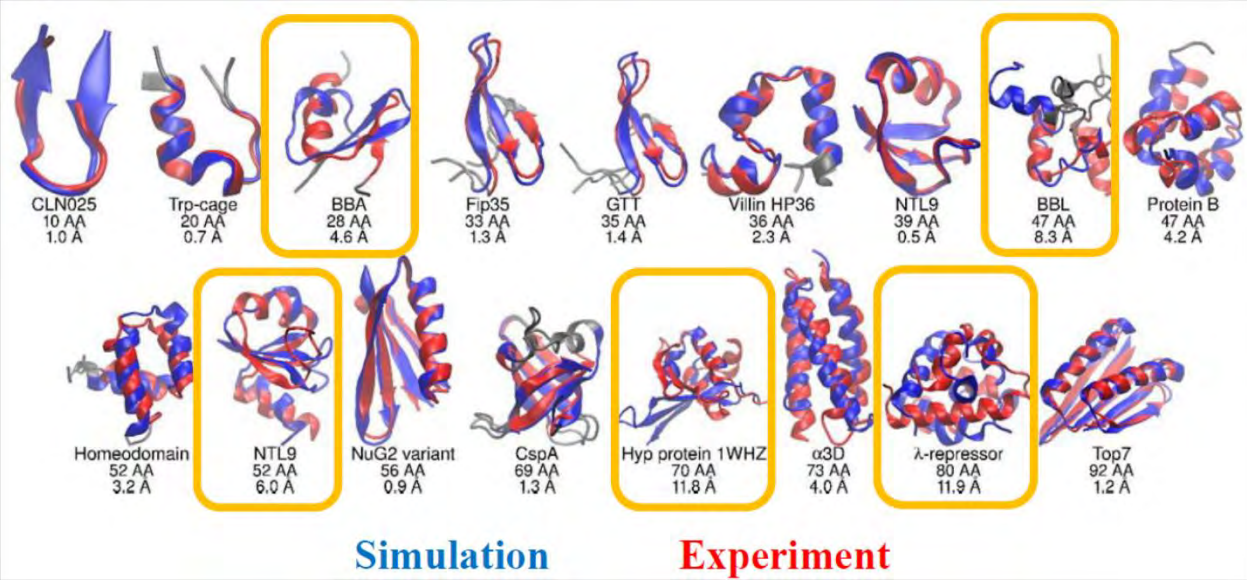


Figure 1-4 Comparing most populated structures obtained from simulations (in blue) vs experiments (in red) for the 17 different proteins shown in **Figure 1.1**, when the simulation is started from native conformation. Below each protein, the name and length of the protein are shown. The RMSD of the most populated structure obtained from simulation is shown below the length of the protein. Grey regions indicate parts of the protein that are poorly characterized in experiments and are excluded from the RMSD analysis. The proteins for which the most populated structure is not the native structure are enclosed in yellow boxes. Figure adapted from "Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent" by H. Nguyen; J. A. Maier; H. Huang; V. Perrone; C. Simmerling, *J. Am. Chem. Soc.* 2014, 136 (40), 13959-13962. Copyright 2014 by the American Chemical Society.

Therefore, it is vital to isolate the different sources of errors to improve the "quality" of simulations. Since sampling errors are easier to isolate than errors arising from force field, the overall aim of this thesis is to reduce the time required to diagnose sampling errors. The following

sections address the source of sampling bottleneck in MD simulations and describe the enhanced sampling methods that have been developed and used in this work, to address this issue.

1.2 Source of sampling bottleneck in MD simulations

In atomistic MD simulations, every atom in the simulation environment is propagated by integrating Newton's second law:

$$F = ma \tag{1-1}$$

where F is the force acting on the particle and is dictated by the force field used, m is the mass of the atom, and a is the acceleration. The integration is typically carried out numerically using algorithms such as Velocity-Verlet[43] which use the above equation to calculate the coordinates of each atom after a certain time step, usually in the order of femtoseconds.

While a longer time step can be taken without significantly compromising the inherent accuracy of the numerical integrators, the vibrational motions of fast-moving hydrogen atoms prohibits using a time step longer than 1 fs. Therefore, attempts have been made to slow or restrict the motion of hydrogen atoms to facilitate the use of a longer time step. To this extent, SHAKE[44] and SETTLE[45] algorithms have been developed to increase the allowed time step to 2 fs. Recently, hydrogen mass repartitioning has further increased the allowed time step to 4 fs[46-47].

Nonetheless, since most biomolecular processes occur on the order of microseconds to milliseconds, studying these processes using MD simulations requires integrating Newton's second law at least 10^9 times. Because of this, the current time scales routinely accessible by simulations is usually in the order of a few microseconds only, significantly limiting the phase space that can be sampled via simulations.

1.3 Increasing temperature to improve sampling in MD simulations

Due to the bottleneck mentioned above, MD simulations carried out at biological (low) temperatures take a long time to sample the relevant phase space of the biological process. To illustrate this, the fraction of unique structures (clusters) observed during a 2 μs simulation of Trpcage (a 20-residue-long protein)[48], using an implicit solvent model[8] is shown in **Figure 1.5**. Even after 2 μs, only 60% of structures are sampled indicating that the simulation time length is not sufficient and needs to be extended further. Moreover, the search process is inefficient at this temperature which is indicated by plateau regions in between 0.3 μs – 0.7 μs and also in between 1.2 μs and 1.4 μs, where no new structures are found.

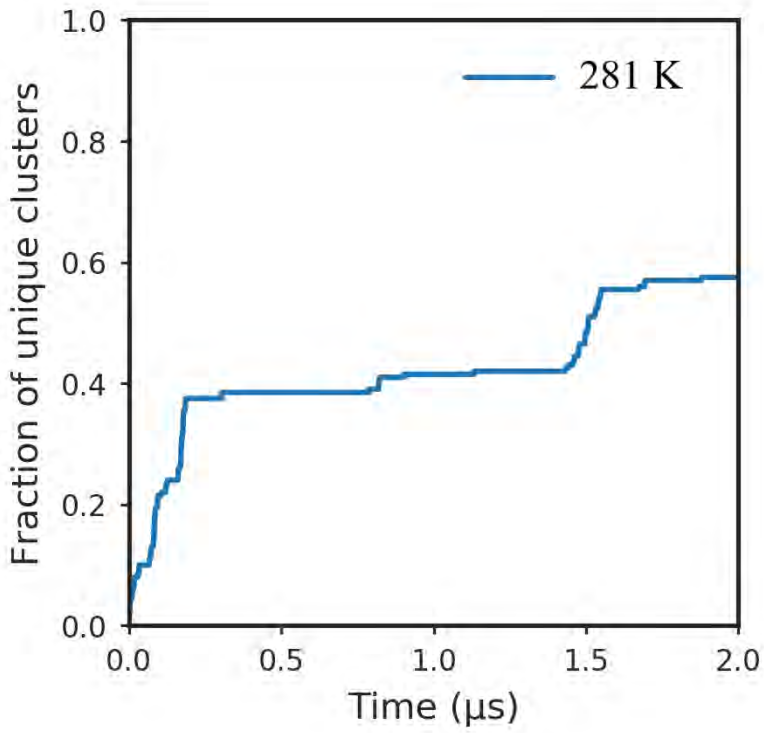


Figure 1-5 Fraction of unique clusters observed during a 2 μs Trpcage MD simulation at 281 K.

If it takes more than 2 μs to sample the phase space of a small protein like Trpcage, it will take even longer to sample the phase space of most biological processes which involve bigger

and/or multiple biomolecules. Therefore, it is necessary to improve the rate of sampling in MD simulations. Since increasing the time step is not a viable solution, alternative methods have to be used to address this problem.

Several enhanced sampling techniques[24-30, 32, 37-42, 49] have been developed to accelerate the convergence of MD simulations, a summary of which can be found in these reviews[34-36]. These techniques can be broadly categorized into two categories: (1) The techniques that require the use of collective variables to enhance sampling of some (not all) degrees of freedom of the biomolecule[24, 27-28], and (2) The techniques that do not require collective variables and enhance sampling of all degrees of freedom of the biomolecule[25-26, 30, 32, 37-38, 49]. This work focuses on enhancing sampling of all degrees of freedom of the biomolecule.

The simplest way to increase the rate of sampling of all degrees of freedom is to run the simulation at a higher temperature so that the added thermal energy can facilitate faster diffusion across phase space. For example, **Figure 1.6** shows the fraction of unique structures (clusters) that are observed at 320 K for Trpcage using the same force field and implicit solvent model as above. At 320 K, >95% of the structures are sampled within the first 300 ns resulting in at least a 6-fold increase in sampling phase space.

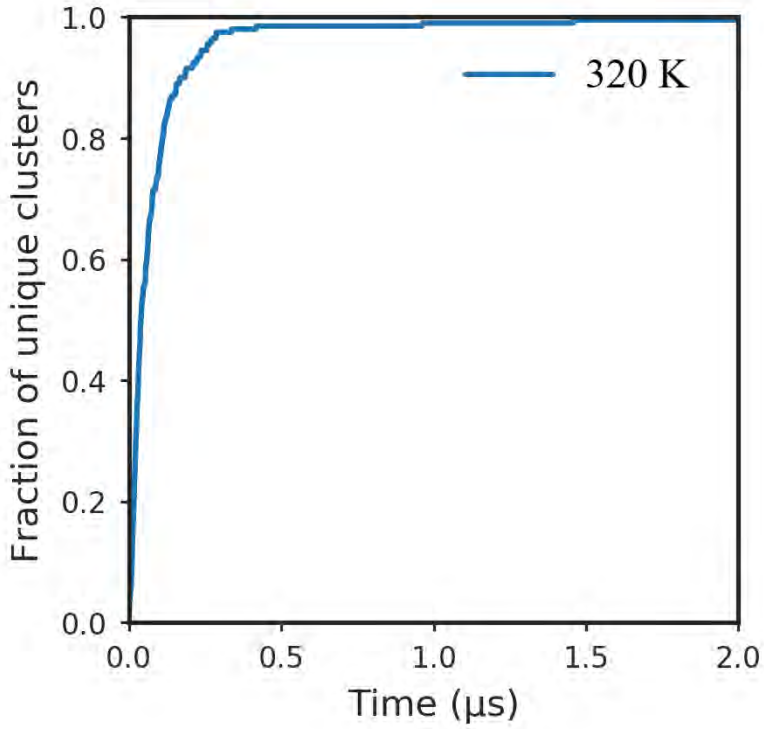


Figure 1-6 Fraction of unique clusters observed during a 2 μs Trpcage MD simulation at 320 K.

The above data shows that it is much more efficient to sample the phase space at high temperatures. However, the equilibrium populations observed at the high temperature might be different from those observed at the biologically relevant temperature.

Nevertheless, since the primary goal of simulations is to sample the phase space at the biologically relevant temperature, which is usually the low temperature, coupling the simulation at high temperature to the simulation at low temperature will allow us to improve sampling at low temperatures. This is the idea behind Temperature Replica Exchange MD[25-26, 29].

1.4 Temperature Replica Exchange MD (scoring and searching)

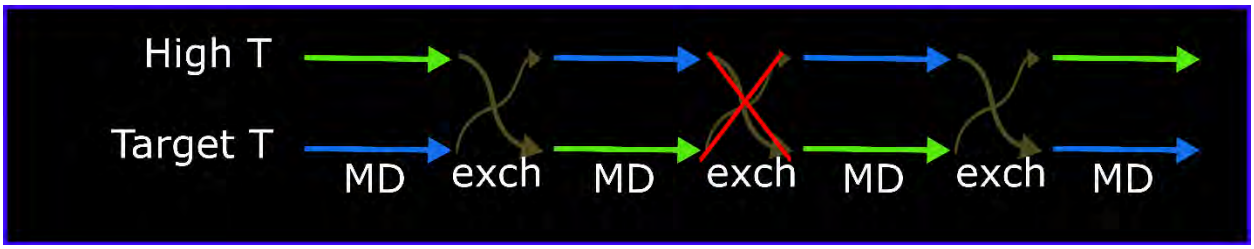


Figure 1-7 T-REMD schema. Blue and green colored arrows represent replicas initiated at different temperatures, Target T and High T (see text). In the schema, the first exchange is successful, therefore, the thermostats of the replicas are switched – blue replica is at high T and green replica is at Target T. The second exchange is not successful (indicated by the red cross), therefore, the thermostats of the replicas are not switched – blue replica is still at high T and green replica is at low T. The third exchange is successful and is indicated by the switch in thermostats. These cycles of MD steps and exchanges are repeated multiple times during a T-REMD simulation.

The basic process of Temperature Replica Exchange MD (T-REMD) is illustrated in **Figure 1.7**. In T-REMD, replicas of the system are simulated simultaneously, at different temperatures. At fixed number of MD steps, exchanges are attempted between the replicas using a Metropolis criterion[50] (see below) that considers the probability of sampling each conformation at a different temperature. If the exchanges are successful, the thermostats of the individual replicas are switched by scaling the velocities. The cycles of MD steps followed by exchanges are repeated multiple times during the course of a typical T-REMD simulation.

Due to the exchanges, each replica can access multiple temperatures in a T-REMD simulation. At high temperatures, the replicas can search the phase space faster, and at low temperatures, the replicas can score the structures that were sampled at the high temperatures. For example, considering the above example of Trp-cage, the fraction of unique clusters sampled by an individual T-REMD replica as a function of time is shown in **Figure 1.8**. As expected, the sampling efficiency is in between that of low temperature MD simulation and a high temperature MD simulation – >90% of structures are sampled within the first 700 ns. More importantly, the

sampled structures are also simultaneously scored at different temperatures including the lowest temperature.

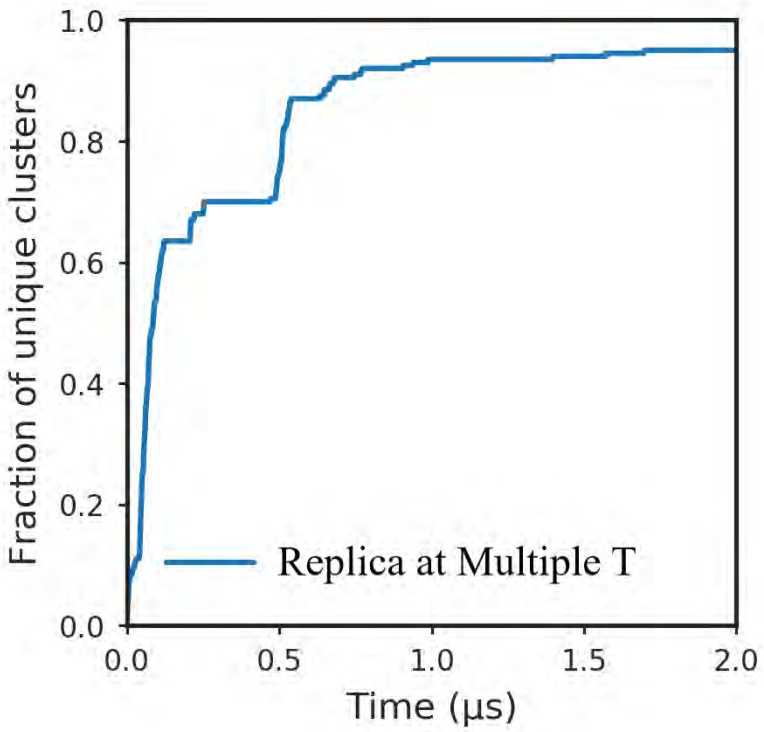


Figure 1-8 Fraction of unique clusters sampled by an individual replica during a 2 μs T-REMD simulation of Trpcage.

Moreover, since there is always a replica at each temperature, ensemble distributions at a particular distribution can be easily obtained by collecting the coordinates of the replicas at that temperature during the course of the simulation. These temperature ensembles allow easy comparison to experimental data such as melting curves. This ability to search and score structures in tandem at different temperatures makes T-REMD a powerful tool to enhance sampling. The key theoretical aspects of T-REMD are discussed below.

1.5 Theoretical details of T-REMD

1.5.1 Derivation of Metropolis criterion

The derivation of Metropolis criterion[50] for a two-replica system is described below in detail, followed by a general extension to an N-replica system.

The positions, momenta, and temperature for replica 1 are denoted by $q^{[1]}$, $p^{[1]}$, and T_m, respectively. Similarly, the positions, momenta, and temperature for replica 2 are denoted by $q^{[2]}$, $p^{[2]}$, and T_n, respectively. The equilibrium probability for this generalized ensemble is

$$W\left(p^{[i]}, q^{[i]}, T_m\right) = \exp\left\{-\sum_{i=1}^{2} \frac{1}{k_B T_m} H\left(p^{[i]}, q^{[i]}\right)\right\} \tag{1-2}$$

where k_B is the Boltzmann constant, Hamiltonian $H(p^{[i]}, q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and the potential energy $E(q^{[i]})$. For simplicity, $(p^{[i]}, q^{[i]})$ at temperature T_m is denoted as $x_m^{[i]}$. With this notation, one state (X) of the generalized ensemble can be written as:

$$X = \left\{x_m^{[1]}, x_n^{[2]}\right\} \tag{1-3}$$

Similarly, the state (X') of the generalized ensemble after exchange can be written as

$$X' = \left\{x_m^{[2]}, x_n^{[1]}\right\} \tag{1-4}$$

To maintain detailed balance of the generalized system, microscopic reversibility has to be satisfied, i.e.,

$$W(X)\rho(X \to X') = W(X')\rho(X' \to X) \tag{1-5}$$

where $\rho(X \to X')$ indicates the probability of going from state X to state X'. Substituting the Boltzmann factors from **Equation 1.2** for the weight of each conformation into **Equation 1.5** yields

$$\exp\left\{-\frac{1}{k_B T_m}H\left(p^{[1]},q^{[1]}\right)-\frac{1}{k_B T_n}H\left(p^{[2]},q^{[2]}\right)\right\}\rho(X\rightarrow X') = \exp\left\{-\frac{1}{k_B T_m}H\left(p^{[2]},q^{[2]}\right)-\frac{1}{k_B T_n}H\left(p^{[1]},q^{[1]}\right)\right\}\rho(X'\rightarrow X) \quad (1\text{-}6)$$

In the canonical ensemble, the momentum can be integrated out. Therefore, the Hamiltonian H reduces to the potential energy E. Rearranging **Equation 1.6** and substituting *1* and *2* by *i* and *j* results in the familiar Metropolis equation generalized for an N-replica system.

$$\rho = min\left(1, exp\left\{\left(\frac{1}{k_B T_m}-\frac{1}{k_B T_n}\right)\left(E_q^{[i]}-E_q^{[j]}\right)\right\}\right) \quad (1\text{-}7)$$

The above equation is valid only for equilibrated ensembles that follow Boltzmann distributions. However, since the replicas are not equilibrated at the beginning of the T-REMD simulation, the above equation drives each replica toward correct equilibrium ensembles. Therefore, until all the replicas are equilibrated, T-REMD simulations cannot be considered converged.

Finally, if the exchange is successful, the thermostats of the individual replicas are adjusted by scaling the velocities as shown below:

$$p^{[i]\prime} = \sqrt{\frac{T_n}{T_m}}p^{[i]} \;;\; p^{[j]\prime} = \sqrt{\frac{T_m}{T_n}}p^{[j]} \quad (1\text{-}8)$$

where $p^{[i]\prime}$ and $p^{[j]\prime}$ are the new momenta of replicas *i* and *j*, respectively.

1.5.2 Factors influencing efficiency of T-REMD simulations

The factors influencing the efficiency of T-REMD simulations are summarized below.

(1) *Choice of lowest temperature*: The lowest temperature should be chosen such that low energy structures such as native structures of proteins can be sampled and should be the temperature at which thermodynamic data is desired.

(2) *Choice of highest temperature*: The highest temperature should be chosen such that the replicas can traverse barriers at this temperature. A good choice is around 350 K – 400 K.

(3) *Number of replicas between the extreme temperatures*: In theory, T-REMD can be used with only two replicas – one at low temperature and the other at high temperature. However, in practice, for systems with large number of degrees of freedom, the two replicas have to be "connected" by multiple replicas in between. To illustrate this, the potential energy distributions at different temperatures are shown for Trpcage in **Figure 1.9**. The overlap between the potential energy distributions of the two extreme temperatures (281 K and 340 K) is very small.

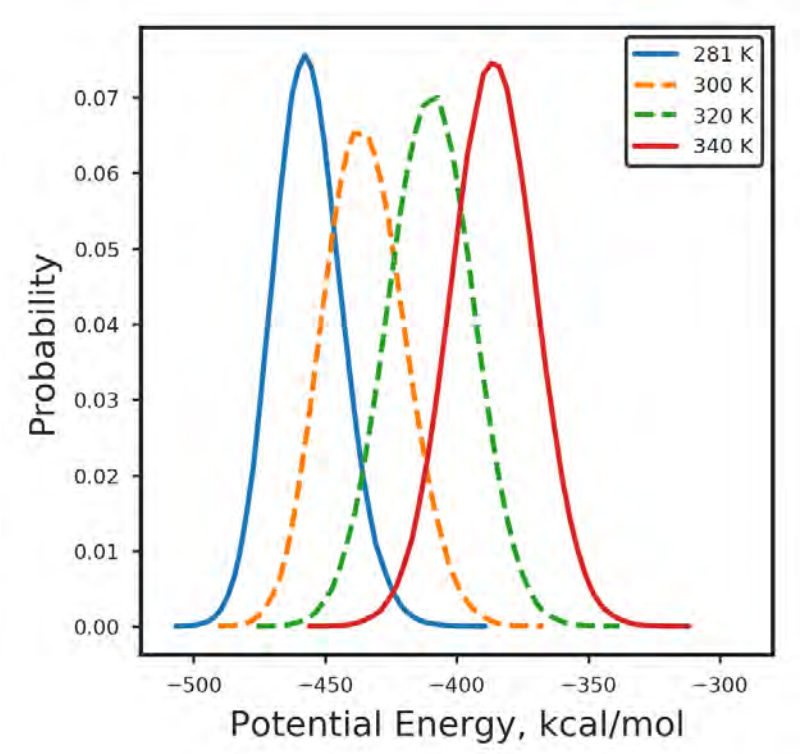


Figure 1-9 Histograms of potential energy at different temperatures for Trpcage simulations in implicit solvent. The extreme cold temperature and the extreme hot temperature are shown as blue and red solid lines, respectively. Additional replicas at intermediate temperatures are shown as orange and green dashed lines.

Since the Metropolis criterion in a canonical ensemble depends on the potential energy only, if the potential energy overlap is too small then most of the exchanges between the two replicas will fail. Therefore, additional replicas at intermediate temperatures have to be used to facilitate fast diffusion across temperature space. Moreover, since the temperature spacing between replicas is proportional to the inverse square root of number of degrees of freedom, for proteins bigger than Trp-cage, many more replicas (increased computational cost) have to be used between the extreme temperatures[51-53]. Nonetheless, estimating the optimal number of replicas between the extreme temperatures for a given system is trivial and tools are available online[54] to calculate the optimal temperature spacing. Typically, replicas are spaced so that 30% of exchanges are successful.

(4) *Frequency of exchanges*: The frequency of exchanges also affects the efficiency of REMD simulations. Typically, exchanges between replicas are attempted every 1 ps though exchanging more frequently might result in faster convergence of T-REMD simulations[55].

(5) *Number of exchanges per exchange cycle*: In AMBER software package[56], one exchange is attempted between every pair of replicas. For example, if 6 replicas are used, then 3 exchange attempts are made every exchange cycle. However, other software packages such as OpenMM[57] support multiple exchanges per exchange cycle so that a replica at low temperature can jump across multiple temperatures instead of jumping to only the adjacent temperature. However, since the potential energy overlap between the replicas decreases as we move farther from the adjacent replica, the acceptance probability is also low for these subsequent exchanges. Therefore, performing multiple exchanges per cycle might not lead to faster convergence since most exchanges might fail.

1.6 Primary issue of T-REMD simulations: changes in temperature occur faster than structural changes

Optimizing the above factors can facilitate faster diffusion across different temperatures, thereby, resulting in improved efficiency of T-REMD simulations. However, faster diffusion across different temperatures does not necessarily correlate with efficient sampling of phase space. To illustrate this, the RMSD of an individual replica to the native structure of Trp-cage is shown as a function of time in **Figure 1.10** as it traverses through different temperatures. The replica (center right image in **Figure 1.10**) appears to be trapped in either low or high RMSD conformations with very few transitions between low/high RMSD conformations. This is probably because the exploration of different structures during the MD part of REMD is still a slow process relative to structure swaps, even at high temperatures.

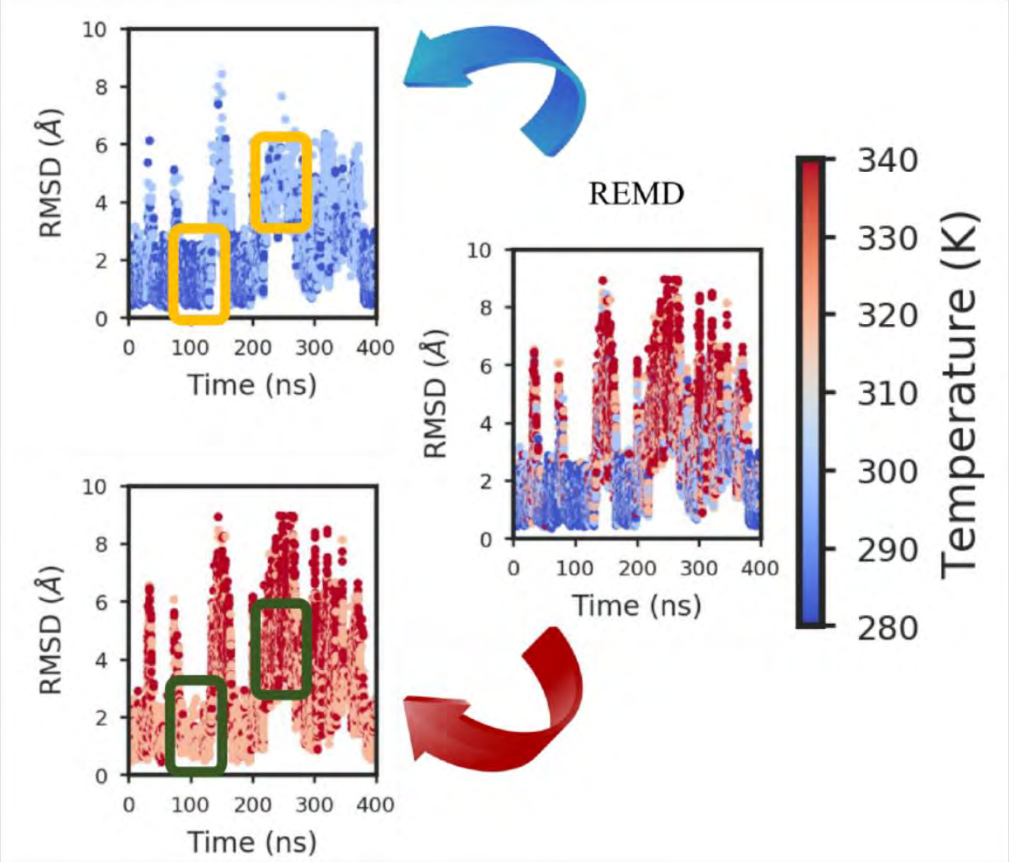


Figure 1-10 The distribution of temperature and RMSD of one replica during the first 400 ns of standard REMD (center right) simulations for Trp-cage. The color of each point represents the temperature while the position of each point on the Y-axis represents the RMSD to native NMR structure. Blue colored points indicate temperatures less than 310 K and red colored points indicate temperatures greater than 310 K. For clarity, the central right image is split into two images. The top left and bottom left images represent the REMD replica data when the temperature of the replica is less than 310 K and greater than 310 K, respectively.

To further clarify this, we split the REMD replica data into two parts, in one part we only show the RMSD and temperature distributions when the temperature of the replica is less than 310 K (top left image in **Figure 1.10**) and in the other part, we only show the RMSD and temperature distributions when the temperature of the replica is greater than 310 K (bottom left image in **Figure 1.10**).

If the higher temperatures result in faster structural transitions, then the bottom left image in **Figure 1.10** (hot replicas) should have a greater spread of RMSD values compared to the top left image (colder replicas), however, the RMSD distributions in both images are similar indicating that temperature transitions can improve sampling only limitedly. Moreover, the changes in temperature occur much faster than structural transitions – this can be clearly seen in between 80 – 120 ns and also in between 160 – 200 ns where even though the replica visits both high and low temperatures, it only samples structures having an RMSD of <2.0 Å during these phases. It is not clear that switching to higher temperature helps this replica to escape the basin in which it is trapped.

Therefore, in order to improve the efficiency of T-REMD simulations, we need to optimize exploration of phase space not just via diffusion across temperatures but via other means too.

To explore other means of improving T-REMD, let us revisit the T-REMD protocol. Each replica in T-REMD performs two tasks: (1) Search phase space, and (2) Score the structures that are sampled at high temperatures. However, as shown in **Figure 1.6** and **Figure 1.8**, plain MD simulations at high temperatures are still more efficient than T-REMD simulations in sampling phase space. Therefore, if we can use structural data (search) from MD simulations at high temperature (without exchanges) and then couple it to T-REMD simulations to reweight the structures at the temperature at which thermodynamic data is desired, we can, in principle, save significant computational resources since the T-REMD replicas will only to have score the structures and not search for them. This is the idea behind Reservoir REMD methods[30-32, 49].

1.7 Reservoir REMD

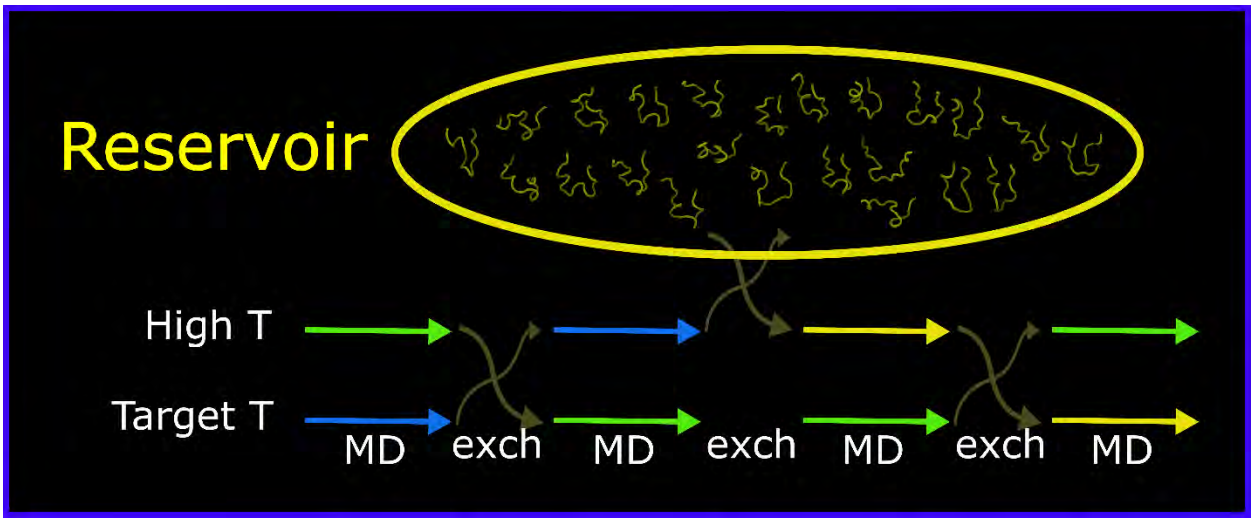


Figure 1-11 Reservoir REMD schema. Blue and green colored arrows represent replicas initiated at different temperatures, Target T and High T (see text). At regular intervals, the replica at high T attempts an exchange with the reservoir. If exchange is successful, the structure in the high T replica is replaced by the reservoir structure. The successful exchange is indicated by the change in color from blue to yellow for the replica at high T. Exchanges with the reservoir are performed in addition to standard T-REMD exchanges.

In Reservoir REMD (RREMD), a set of structures (reservoir) representing different energy minimum are generated by performing extensive MD simulations at a high temperature. Then, these structures are coupled to REMD via allowing exchanges based on the Metropolis criterion between the replica at highest temperature and a randomly chosen structure from the reservoir. If an exchange is successful between the highest temperature replica and the reservoir structure, in addition to rescaling the velocities, the highest temperature replica structure takes on the reservoir structure. After reservoir structures are accepted into the replica at highest temperature, the normal REMD process of simulation across the temperature ladder provides local exploration/refinement of the basins sampled in the reservoir, and also reweights the probability of observing these structures at different temperatures. The exchanges with the reservoir are repeated multiple times, concurrently with the REMD exchanges so that every reservoir structure eventually is accepted

and thermally reweighted many times during the simulation, resulting in converged Boltzmann-weighted ensembles at all REMD temperatures.

The advantage of using RREMD over T-REMD is illustrated further in **Figure 1.12**. In T-REMD, structural transitions can only occur through the MD part of the REMD process. Due to this, the convergence rate of T-REMD is limited by the rate at which different structures are sampled. In RREMD, since the structures are pre-sampled, Monte Carlo (MC) type moves can be used to instantaneously "jump" between structures, thereby eliminating the need to wait for structural transitions during the MD part of T-REMD.

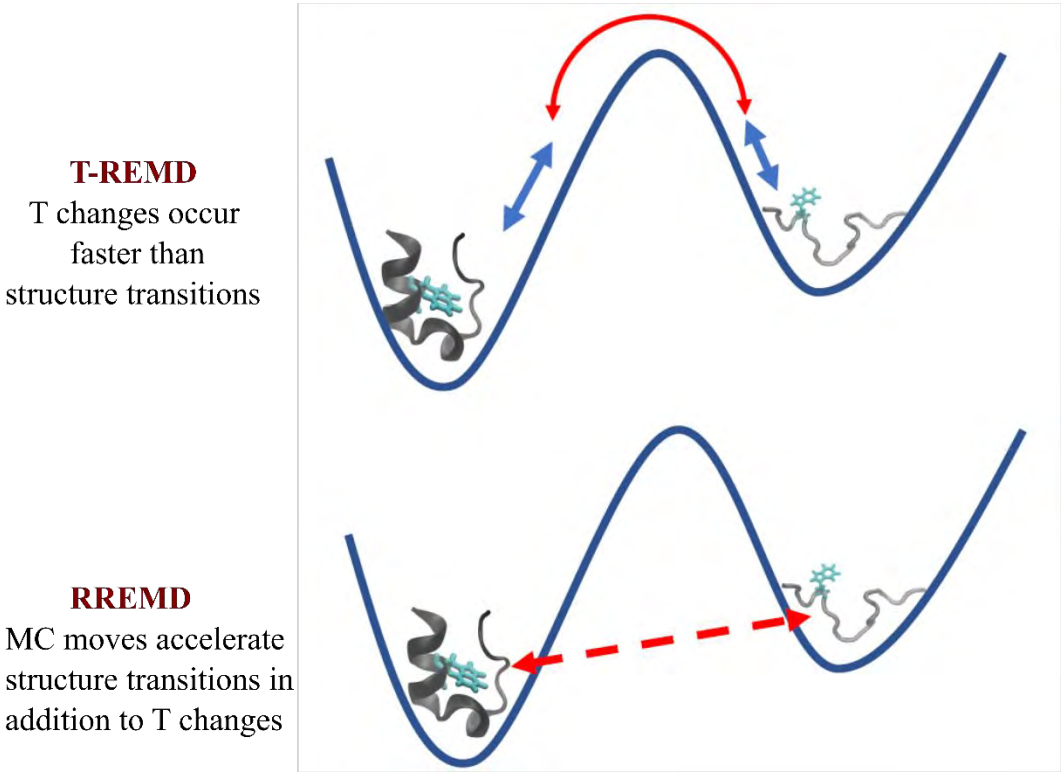


Figure 1-12 Diagram showing the key sampling difference between RREMD and T-REMD. Two arbitrary protein conformations are shown on the energy landscape (dark solid blue line). Structural changes (shown as solid red and blue double-headed arrows) in T-REMD are slow since structures have to evolve via barrier crossing during the MD part of T-REMD. Structural changes in RREMD are performed through MC moves (shown as dashed red double-headed arrow) and are instantaneous since they "skip" barriers.

1.7.1 Types of Reservoir REMD

The two types of exchanges with the reservoir depend on the way the reservoirs are built. These are described in detail in **Chapter 2**. The key differences are highlighted below in brief.

1.7.1.1 Boltzmann-weighted RREMD

In Boltzmann-weighted RREMD (B-RREMD)[30, 32], the reservoir is assumed to be Boltzmann-weighted i.e., the structures in the reservoir occur with correct relative populations of all the relevant minima. Such a reservoir can be generated by extracting equidistant structures from the high temperature MD simulation.

If a Boltzmann-weighted reservoir is used, then exchanges with the reservoir are attempted using the Metropolis criterion derived above and reproduced here again with two small differences.

$$\rho = min\left(1, exp\left\{\left(\frac{1}{k_B\,T_m} - \frac{1}{k_B\,T_R}\right)\left(E_q^{[i]} - E_R\right)\right\}\right) \tag{1.9}$$

where T_R represents the temperature at which the reservoir structures were generated, and E_R represents the energy of structure with which the exchange is attempted.

1.7.1.2 Non-Boltzmann RREMD

In Non-Boltzmann RREMD (nB-RREMD)[49], instead of having multiple structures (reflecting correct ensemble populations) per minima as in B-RREMD, only one structure per minima is used in the reservoir. Such a reservoir can be generated via clustering MD trajectories and extracting cluster representatives, or doing a grid scan, or using non-physics-based models such as homology modeling, or through a combination of these methods. The greater flexibility in generating structures to build a non-Boltzmann reservoir, makes nB-RREMD more attractive than B-RREMD.

If a non-Boltzmann reservoir is used, the following modified exchange criterion has to be used

$$\rho = min\left(1, exp\left\{-\left(\frac{1}{k_B T_m}\right)\left(E_q^{[i]} - E_R\right)\right\}\right) \tag{1.10}$$

where E_R represents the energy of structure with which the exchange is attempted. This is essentially the same as B-RREMD equation except for the weighting factor of ΔE, which looks like a regular Metropolis MC step. An alternative way to think about this is to assume that the reservoir temperature is infinite in **Equation 1.9**, which results in a flat distribution (without populations for each minimum).

However, running simulations at infinite temperature would likely add significant thermal component to the potential energy. Therefore, to obtain structures with reasonable potential energy, non-Boltzmann reservoir generation simulations are run at only slightly high temperatures (<400 K) and then clustered. After clustering, only the cluster representatives are used to build the reservoir. Since the populations of each cluster are ignored, the underlying distribution in the reservoir can be considered flat. Further details on building non-Boltzmann reservoirs and how they work are discussed in **Chapter 2**.

Alternatively, grid-based search algorithms can also be used to generate structures for a non-Boltzmann reservoir. However, such mechanisms are not explored in this thesis.

1.8 Force Fields and Solvent Models

As mentioned before, force fields and solvent models play an important role in determining the accuracy of MD simulations.

1.8.1 Force Fields

The force fields determine the potential energy of the solute. They typically have the form

$$U_{potential\ energy} = U_{bonds} + U_{angles} + U_{dihedral} + U_{improper} + U_{vdw} + U_{electrostatic} \tag{1.11}$$

where the individual components are given by

$$U_{bonds}(\vec{r}) = \sum_{bonds} K_b\,(\vec{r} - r_0)^2 \tag{1.12}$$

$$U_{angles}(\vec{\theta}) = \sum_{angles} K_\theta\,(\vec{\theta} - \theta_0)^2 \tag{1.13}$$

$$U_{dihedral}(\gamma) = \sum_{dihedral} K_\gamma\,(1 + \cos(n\gamma - \delta) \tag{1.14}$$

$$U_{improper}(\vec{\varphi}) = \sum_{improper} K_\varphi\,(\vec{\varphi} - \varphi_0)^2 \tag{1.15}$$

$$U_{vdw}(\vec{r}) = \sum_{vdw} \varepsilon_{ij}\left[\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.16}$$

$$U_{electrostatic}(\vec{r}) = \sum_{electrostatic} \frac{q_i q_j}{\varepsilon r_{ij}} \tag{1.17}$$

The various constants in the above equations determine the accuracy of force fields.

1.8.2 Solvent Models – implicit vs explicit solvent

In MD simulations, solvation effects can be modeled via either explicitly including all the water molecules (referred to as explicit solvent)[2-3, 9] or implicitly by using a continuum dielectric (referred to as implicit solvent)[4, 6, 8, 11, 58-59]. Both methods have their advantages and disadvantages which are discussed in more detail in **Chapter 5**.

During the past two decades, due to the advances in computing power and availability of more experimental data to validate against, significant improvements have been made to the accuracy of force fields and solvent models[2-6, 8, 11, 13, 58, 60-61]. Some of the force fields and solvent models supported by AMBER software package[56] are shown in **Figure 1-13.**

Figure 1-13 Force fields (left) and solvent models (right) supported by AMBER.

1.9 Hybrid-solvent T-REMD

In hybrid-solvent T-REMD[37-38], independent replicas of the system are simulated in explicit solvent, at different temperatures, as in standard T-REMD. However, during exchanges, instead of using the potential energy of the whole system (solute + all explicit water molecules and **Equation 1.18**), the potential energy of the solute solvated in implicit solvent is used (**Figure 1-14** and **Equation 1.19**).

$$U_{potential\ energy\_explicit} = U_{protein-protein} + U_{protein-water} + U_{water-water} \quad (1.18)$$

$$U_{potential\ energy\_hybridsolvent\_REMD} = U_{protein-protein} + U_{protein-water} + U_{implicit} \quad (1.19)$$

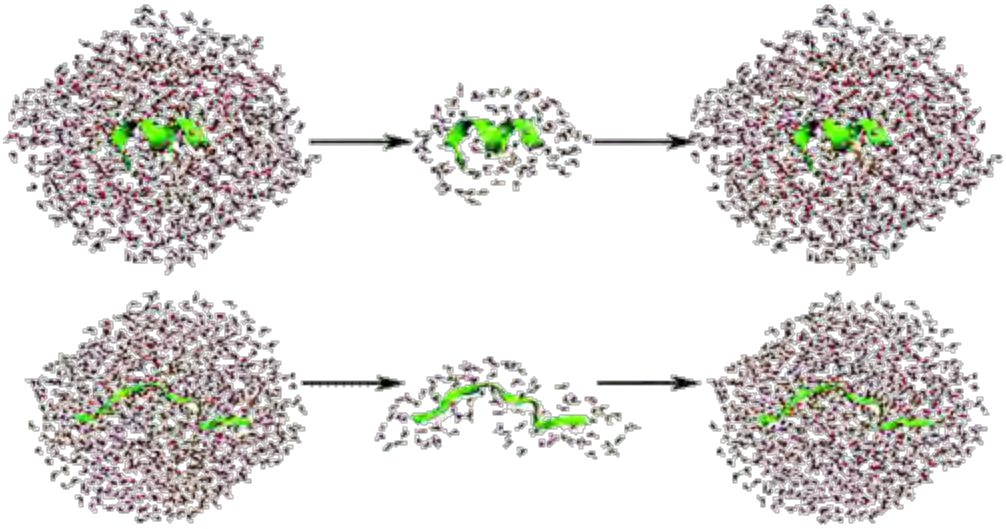


Figure 1-14 Schematic representation of hybrid-solvent T-REMD. Replicas are simulated in fully explicit solvent during the MD part of REMD. During exchanges, only the potential energy of the solute with/without few water molecules solvated in an implicit solvent model is used. After exchange, the water molecules are restored, and the simulation proceeds as usual. Figure taken from "Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvent Model" by A. Okur; L. Wickstrom; M. Layten; R. Geney; K. Song; V. Hornak; C. Simmerling, *J. Chem. Theory Comput.* 2006, 2 (2), 420-433. Copyright 2006 by the American Chemical Society.

The switch to an implicit solvent during exchange allows fewer replicas to be used since the number of degrees of freedom during the exchange are significantly reduced. Water molecules corresponding to the first and second solvation shells ($U_{protein\text{-}water}$) can also be included in the exchange scheme so that close-range interactions between the solute and solvent are modeled accurately. The strengths and weaknesses of a hybrid-solvent exchange scheme are described in more detail in **Chapter 5**.

1.10 Thesis overview

The overall aim of this thesis is to reduce the computational resources and time required to eliminate errors due to insufficient sampling. To achieve this, Reservoir REMD method, which accelerates sampling by coupling T-REMD to a set of pre-generated reservoir structures, is optimized and used.

In **Chapter 2**, the choices involved in building "good" Boltzmann-weighted reservoirs and non-Boltzmann reservoirs are explored. The protocols show that, with careful selection of structures, the method can accurately reproduce Boltzmann-weighted ensembles obtained by much more computationally expensive T-REMD simulations, with 25x faster convergence rates compared to T-REMD.

In **Chapter 3**, the protocols presented in **Chapter 2** for building non-Boltzmann reservoirs are further explored and optimized, making the construction of non-Boltzmann reservoirs fairly straightforward.

In **Chapter 4**, non-Boltzmann reservoirs generated from one Hamiltonian are used to predict the preferences of a different Hamiltonian. Non-Boltzmann reservoirs are also used to predict the effects of mutations. Both these applications show that structure reservoirs can

potentially be used to test new force fields and design new peptides, further increasing the scope of applicability of structure reservoirs in addition to reducing computational time.

In **Chapter 5**, structure reservoirs are used to accelerate conformational sampling in explicit solvent simulations by around 1000x. While this is amazing, the reservoirs used in these simulations were generated from another extensive (100 μs) explicit solvent simulations. Performing explicit solvent simulations for 100 μs still takes at least 6 months on the latest computing hardware.

Therefore, in **Chapter 6**, structures generated from implicit solvent simulations are used as reservoirs to predict the accuracy of explicit solvent simulations, significantly reducing the time required to generate reservoirs to accelerate explicit solvent simulations. The results indicate that structure reservoirs obtained from implicit solvent can be used to predict the accuracy of implicit solvent simulations. However, further refinements to the protocols have to be made before the method can be widely adopted.

Finally, in **Chapter 7**, outstanding issues of using structure reservoirs, and future applications are addressed.

2 Exploring protocols to build reservoirs to accelerate Replica Exchange MD simulations

2.1 Abstract

Replica Exchange Molecular Dynamics (REMD) is a widely used enhanced sampling method for accelerating biomolecular simulations. During the past two decades, several variants of REMD have been developed to further improve the rate of conformational sampling of REMD. One such variant, Reservoir REMD (RREMD), was shown to improve the rate of conformational sampling by at 5-20x. Despite the significant increase in sampling speed, RREMD methods have not been widely used due to the difficulties in building the reservoir and also due to the code not being available on the GPUs.

In this work, we ported the RREMD code onto GPUs making it 20x faster than the CPU code. Then, we explored protocols for building reservoirs and tested how each choice affects the accuracy of RREMD. Our protocols show that, with careful selection of structure snapshots, the method can accurately reproduce Boltzmann-weighted ensembles obtained by much more expensive conventional REMD, with at least 15x faster convergence rates even for larger proteins (>50 amino acids) compared to conventional REMD.

2.2 Acknowledgements

I gratefully acknowledge Kenneth Lam for the numerous helpful discussions. This chapter contains direct excerpts from the manuscript titled “Exploring Protocols to Build Reservoirs to Accelerate Replica Exchange MD simulations” with a few modifications.

2.3 Introduction

Conformational ensembles are essential to understand biological processes such as protein folding, drug binding, and protein-protein interactions, besides many others. They are also needed to evaluate force fields against experimental data, to study intrinsically disordered proteins (IDPs) which have multiple possible conformations under native conditions, and also to study unfolded state of non-IDPs under native conditions. Obtaining converged Boltzmann-weighted ensembles for biomolecules using conventional Molecular Dynamics (MD), however, still takes a long time (weeks to months) even on the latest computational hardware due to the rugged nature of the energy landscape.

To illustrate this, the fraction of native structures as a function of time are shown in **Figure 2-1A** for conventional MD simulations starting from native and fully extended conformations, for the small hairpin peptide CLN025, in implicit solvent, at four different temperatures. Even for a fast-folding system like CLN025, which is only 10-residues long, in implicit solvent, it takes up to 80 ns for the fraction of native structures from the two simulations to converge at 300 K. The simulations at 275 K converge around 200 ns but diverge after that whereas the fraction of native structures from simulations at 252 K and 327 K haven't converged even after 400 ns. Furthermore, the simulation starting from native conformation has stayed native for the entire 400 ns at 252 K indicating that it is trapped in the local minima. If the fraction of native structures takes longer than 400 ns to converge, it will take even longer for the overall ensemble to converge. If the overall ensembles are not converged, any calculated thermodynamic data such as free energy of peptide/protein folding will be unreliable resulting in inaccurate comparisons to experimental observables. This problem will only get worse for bigger biomolecules which might have a more

complex energy landscape than CLN025, or when explicit solvent is used because the added viscosity reduces the rate of conformational changes compared to implicit solvent[62].

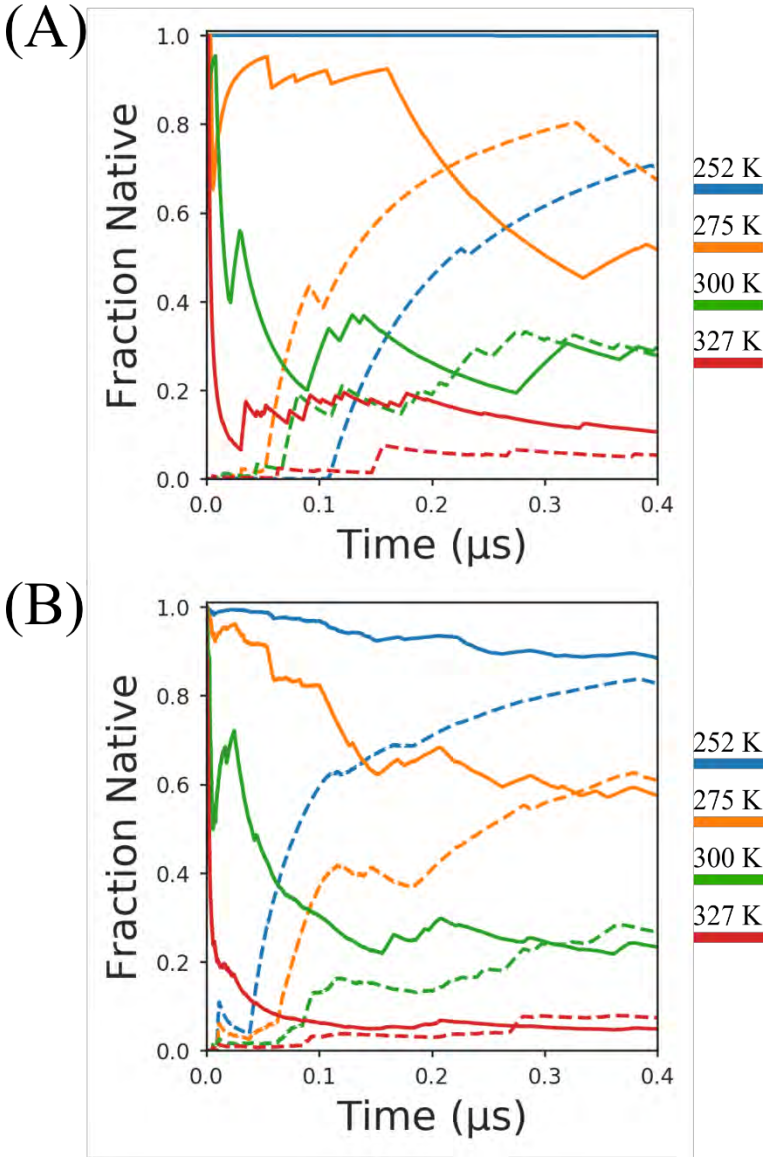


Figure 2-1 Fraction Native vs Time using (A) standard MD, and (B) REMD at four different temperatures. The solid lines are for runs starting from native conformation and the dashed lines are for runs starting from extended conformations.

Several enhanced sampling techniques have been developed to accelerate the convergence of MD simulations, a summary of which can be found in recent reviews.[34-36] These techniques can be broadly categorized into two categories: (1) techniques that require user-defined collective

variables to enhance sampling along these (but not all) degrees of freedom of the biomolecule, such as umbrella sampling[24], metadynamics[27], accelerated MD[28], besides others; and (2) techniques that do not require collective variables, and enhance sampling of all degrees of freedom of the biomolecule, such as temperature Replica Exchange MD (REMD).[25-26, 29] The efficiency of the first category of methods is dependent on whether the collective variables used correspond to the slowest motions of interest. Since identifying appropriate collective variables for biological processes is a non-trivial task, the applicability of these methods is limited, especially when full ensembles and not specific conformational changes need to be modeled. REMD, on the other hand, is a generic method that requires only the knowledge of the number of atoms in the simulation system. However, the applicability of REMD to large biomolecular systems is prohibitive as it requires significant computational resources (see below). Several variants of REMD have been developed to improve the efficiency of REMD for biomolecular systems with large number of degrees of freedom. Here, we explore the Reservoir REMD variants since these were shown to improve convergence of REMD simulations by at least 5x.[30-32, 49, 63-66]

We briefly review the REMD and Reservoir REMD methods for context. In REMD, replicas of the simulation system are simulated simultaneously at different temperatures. At regular intervals, exchanges are attempted between replica pairs based on the Metropolis criterion. If an exchange is successful, the thermostats of the exchanging replica pairs are switched by rescaling the velocities. These exchanges are repeated such that each replica traverses through different temperatures multiple times during the simulation. At high temperatures, the replicas can more easily overcome energy barriers and escape local minima with the added thermal energy. At low temperatures, the replicas explore the local minima that they visited at the high temperatures but with corrected weighting, thereby enhancing exploration of conformational landscape even at low

temperatures. Overall this results in faster convergence of conformational ensembles at all temperatures compared to standard MD. Furthermore, the Metropolis criterion ensures that the canonical ensemble properties are maintained at each temperature, making it easier to compare the conformational probabilities at different temperatures to experimental data such as melting curves. **Figure 2-1B** illustrates the faster convergence of conformational ensembles using REMD for the CLN025[67] hairpin using the same number of MD runs and thermostats as **Figure 2-1A**, but with the addition of exchanges between the MD runs at different temperatures. The fraction of native structures converges within 400 ns for all temperatures, indicating that the coupling of temperatures allows the efficiently-sampling high temperature simulations to accelerate the low temperature conformational sampling.

The faster convergence of REMD sometimes comes at significant computational cost compared to the conventional MD runs due to the following reasons: (1) Even though REMD provides ensemble distributions at all temperatures, we are often interested in the ensemble distributions at only the lowest temperature. Even to obtain the ensemble distributions at only the lowest temperature, multiple replicas have to be simulated in REMD resulting in a *n-fold* increase in computational cost, where *n* is the number of replicas, compared to conventional MD. (2) Since the enhanced exploration of conformations is driven by temperature changes, it is imperative for all the replicas to traverse to high and low temperatures multiple times during the course of the simulation[53, 55, 68]. To ensure this process occurs, the temperature spacing between the replicas must be optimal so that exchanges can be successful. However, since the optimal temperature spacing between the replicas is inversely proportional to the square root of the number of degrees of freedom, a greater number of replicas (increased computational cost) will be required to span the same temperature range between ones where sampling is efficient and ones where

thermodynamic data are desired. (3) Since at low temperatures, replicas only sample the local minima that they visited at the high temperatures, the rate of convergence of REMD is limited by the rate at which different minima are explored at the high temperatures, and until all the minima are properly sampled, none of the replicas represent a Boltzmann-weighted distribution and hence, cannot be considered converged. This means that the data collected from the low temperature replicas should be discarded until the exploration process is complete, resulting in a significant wastage of computational resources at the beginning of REMD simulations.

Since most of the exploration of conformational landscape is done at high temperatures, it seems reasonable to generate a set of structures (reservoir) representing different minima using extensive simulations at a high temperature, and only then to reweight and locally refine them using REMD exchanges along the temperature ladder. In this way the exploration process of REMD at high temperatures can be decoupled from the thermal reweighting process of REMD at low temperatures, with potential for significant reduction in computational cost by avoiding the simulation of all temperatures during the time when only the high temperature runs are exploring new minima. This is the idea behind the Reservoir REMD (RREMD) methods.[30-32, 49]

In RREMD, a set of structures (reservoir) representing different energy minimum are generated by performing extensive MD simulations at a high temperature. Then, these structures are coupled to REMD via allowing exchanges based on the Metropolis criterion between the replica at highest temperature and a randomly chosen structure from the reservoir. If an exchange is successful between the highest temperature replica and the reservoir structure, in addition to rescaling the velocities, the highest temperature replica structure takes on the reservoir structure. After reservoir structures are accepted into the replica at highest temperature, the normal REMD process of simulation across the temperature ladder provides local exploration/refinement of the

basins sampled in the reservoir, and also reweights the probability of observing these structures at different temperatures. The exchanges with the reservoir are repeated multiple times, concurrently with the REMD exchanges so that every reservoir structure eventually is accepted and thermally reweighted many times during the simulation, resulting in converged Boltzmann-weighted ensembles at all REMD temperatures.

RREMD is analogous to J-walking[69], S-walking[70], smart darting[71], annealed swapping[72], and cool-walking[73] methods. The primary difference between RREMD and these methods is that in the former, the highest temperature replica is coupled to the lowest temperature replica through multiple replicas in between, whereas, in the latter methods, the highest temperature replica is directly coupled to the lowest temperature replica through either quenching and heating (annealed swapping), or direct quenching (S-walking and smart-darting), or partial quenching (cool-walking), or no quenching (J-walking).

RREMD method is formally exact and satisfies detailed balance. In principle, the only change from standard REMD is that the highest temperature simulation is not run concurrently with the rest of the REMD ladder, and exchanges can take place to any time point sampled by the reservoir generation MD, rather than only the current time in standard REMD. However, it assumes that the set of structures obtained from the high temperature MD simulation represents a converged Boltzmann-weighted ensemble, i.e., the structures in the reservoir occur with correct relative populations of all the relevant minima. Otherwise, the use of the Boltzmann limiting distribution in the Metropolis criterion is incorrect. In practice, after a successful exchange with the reservoir, the reservoir typically is not updated with the structure in the highest temperature replica since it does not add any new information to what is assumed to be an already-complete reservoir, and bookkeeping is facilitated by simply discarding the structure being exchanged into

the reservoir. Likewise, when the highest temperature accepts a structure from the reservoir, the original copy is not removed from the reservoir. Since the reservoir is finite in practice, adding or removing structures would change the relative weights of the minima and disrupt the ensemble. One assumes that each structure in the reservoir represents an infinite number of copies, and therefore adding or removing a single snapshot would not change the reservoir composition. Since the reservoir is assumed to be Boltzmann-weighted, this method is referred to as Boltzmann-weighted RREMD (B-RREMD).[30, 32]

Prior application of B-RREMD resulted in ensemble distributions that were in excellent agreement with conventional REMD simulations for a wide variety of small molecules and peptides such as butane[30-31] and leucine dipeptide[30] in gas phase, leucine tripeptide in implicit solvent[31], Trpzip2 and DPDP in implicit solvent[32], and $A\beta_{21-30}$ peptide[63] in explicit solvent. B-RREMD also has been used to calculate binding affinities for different host-guest complexes using a combination of Hamiltonian REMD and B-RREMD[74-75]. In all cases except the study involving calculation of binding affinities (where conventional REMD data is not available), B-RREMD simulations converged at least 5x faster (ignoring the time required to generate the reservoir structures) than conventional REMD simulations. Moreover, the infrequent calculations involving exchanges with the reservoir had a negligible effect on the wall clock time of the simulations.

Despite the significant increase in convergence rate and reduced computational cost, the use of B-RREMD method has been limited to only small biomolecules (< 310 atoms) and has not been used for studying larger biomolecules with more complex folding landscapes. In part this is due to the challenge of needing a reliably Boltzmann-weighted ensemble for a reservoir, which may be easier to generate at elevated temperature but remains forbidding for systems with large number of degrees of freedom. On the other hand, simply generating a set of structures

corresponding to the important local minima should be a much easier task (see **Results and Discussions**) than is sampling these same basins with the correct relative populations (a Boltzmann-weighted reservoir). Furthermore, many different physics-based sampling methods (such as metadynamics or accelerated MD) and non-physics-based sampling methods (such as homology modeling[76]) could be used to generate sets of structures corresponding to different important minima, thereby significantly increasing the scope of applicability of structure reservoirs in accelerating biomolecular simulations.

Recognizing this, non-Boltzmann RREMD (nB-RREMD)[49] was developed in which the Metropolis exchange criterion is modified to reflect the altered non-Boltzmann distribution of structures in the reservoir. This criterion is used to exchange between the reservoir and the highest replica temperature. In its simplest form[49], a flat distribution is assumed (i.e., one structure is present for each minimum) and the exchange criterion uses only the energy of the reservoir structure and not its temperature. The exchange is formally equivalent to periodically allowing a Monte Carlo (MC) jump for the highest temperature replica during REMD, with the possible target basin for the jump chosen randomly from those represented in the reservoir. Since the reservoir structures are chosen to correspond to different minima on the energy landscape, the MC exchange moves in nB-RREMD result in rapid exploration of conformational space unhampered by energy barriers, with reasonable acceptance ratios. Thus, the MC swaps in nB-RREMD result in more efficient conformational sampling compared to traditional biomolecular MC simulation, where candidate structures are generated on the fly by adjusting only a few degrees of freedom in order to achieve tolerable acceptance probabilities. Moreover, the decoupling of time between the REMD run and the reservoir allows the REMD ladder to resample basins that may only have a single instance in the reservoir. This can be important to facilitate buildup of population across a

range of low temperature replicas without the need to sample multiple independent (and potentially slow) folding events as would be needed in standard REMD. After the structure is accepted into the highest temperature replica through a non-Boltzmann exchange move, similar to B-RREMD, the REMD process reweights the probability of observing the accepted structure at different temperatures, and also carries out local exploration and refinement, which can be crucial if the reservoir structure was not generated using the same model as used for the REMD run.

nB-RREMD was studied in the past for small systems, and resulted in conformational ensembles that were in good agreement with conventional REMD for alanine tetrapeptide[65], alanine undecapeptide[65], Trp-cage miniprotein[65], and Trpzip2 in implicit solvent[49], and for alanine dipeptide[64] and RNA (rGACC) tetramer[64] in explicit solvent. In all cases, similar to B-RREMD, nB-RREMD was found to be at least 5x faster than conventional REMD simulations. The nB-RREMD method has also been recently adapted to pH replica exchange method (pH-REM), in which, in addition to the pH-REM scheme, the non-Boltzmann exchange move was used to integrate structures from reservoir into the pH-REM replicas.[66]

2.3.1 Current issues with RREMD

Despite the significant promise, both RREMD variants have not been widely used due to the following reasons:

It was observed that the distribution of structures in the reservoir significantly affected the overall ensembles that were sampled by RREMD. If structures were missing from the reservoir, the likelihood of them being observed during RREMD was low.[64] Therefore, it is imperative that the reservoir used in RREMD simulations contains all of the relevant structures.

Running very long simulations can ensure that the reservoir does not have any missing structures. However, it is difficult to estimate how long the simulations have to be run for a given

biomolecule since different biomolecules will require different simulation time lengths to sample important basins, depending on the size and folding/unfolding rates of the biomolecule. For example, the high temperature MD simulations that have been used so far to generate reservoirs ranged from 20 ns[74] to all the way up to 1.4 μs[64]. Nonetheless, the underlying assumption that the convergence rate of REMD simulations is limited by the rate of conformational sampling at the high temperatures means that if a high temperature MD simulation doesn't sample all the conformations in a given amount of time, then, REMD might not also sample all the conformations in the same amount of time per replica. Therefore, it is still advantageous to pre-sample a set of structures and do REMD thermal reweighting and local optimization later.

It wasn't clear how many structures must be extracted from the high temperature MD simulation, to build a Boltzmann-weighted reservoir. The number of structures used so far in Boltzmann-weighted reservoirs ranged from 5000 to 150000[30, 32, 63-65]. If the number of structures is too few, then the populations of different minima will have too low precision and might not reflect true Boltzmann distribution and using such a reservoir in B-RREMD simulations will result in erroneous ensemble distributions at all temperatures.

The nB-RREMD in its simplest form overcomes the issue of selecting structures with correct relative populations for each minimum by using only one structure instead of many structures for each minimum. However, the accuracy of the nB-RREMD depends critically upon the assumption that the underlying reservoir distribution is flat, i.e., there is only one structure corresponding to each local minimum. To obtain a flat reservoir distribution, two methods have been used so far: (a) the high temperature MD trajectories were clustered and cluster representatives were used to build the non-Boltzmann reservoirs[49], and (b) Conformational space annealing (CSA)[77] was used[65] which, in theory, generates structures with low potential energy that

are spaced at least D_{cut} distance apart. While either method can result in flat reservoir distributions, however, both methods have the limitation that the ideal number of clusters (in the case of clustering) or the ideal number of low energy structures (in the case of CSA) is not known in advance. For example, alanine tetrapeptide simulations using 64 and 256 conformations obtained from CSA produced similar results.[65] Also, clustering MD trajectories is a non-trivial task since the clustering results are influenced by the choice of the clustering method, and the clustering metric.[78-82] It wasn't clear which clustering method was the best and which clustering metric should be used to build a non-Boltzmann reservoir. Furthermore, what should the energies of the cluster representatives corresponding to different minima be – Is the energy of the cluster representative an accurate representation of the energy basin or is the less-noisy average energy of all the structures in the cluster a more accurate representation of the energy basin?

The temperature at which the high temperature MD simulation is run also plays a crucial role in generating reservoir structures. The reservoir generation temperature should not only be hot enough to traverse barriers and sample conformational space quickly, but also sample the basins important at low temperatures. Such an ideal reservoir generation temperature for a given biomolecule is difficult to know in advance.

The code was available only on the CPUs which prohibited extensive testing of parameters (How long should the MD simulation be run for different biomolecules? How many structures should be selected? What should the energy of the structures be? What is the ideal temperature?) involved in building reservoirs. Also, because the code was available only on the CPUs, the method has been tested only on very small biomolecules (<310 atoms) with trivial structures so far. It is not known if the faster convergence rates of RREMD might be affected for bigger biomolecules which might require both, (a) a greater number of replicas which can slow down the convergence

speed since the structures from the reservoir have to traverse through more replicas, and (b) more reservoir structures representing different minima on the energy landscape which means more structures have to be evaluated before the simulations can be considered converged.

Finally, testing the method has been difficult due to the cost of obtaining highly precise reference data using the same force field and other simulation conditions, in order to critically compare and evaluate the various approximations made during reservoir construction and isolate these issues from confounding factors in comparison to experiment (such as force field accuracy).

2.3.2 Current work

In this work, we ported AMBER's RREMD code onto the GPUs which facilitated testing various protocols to build reservoirs. We explored protocols for building both Boltzmann-weighted reservoirs and non-Boltzmann reservoirs, and tested how each choice (see below) affects the accuracy of RREMD compared to extensive conventional REMD simulations for CLN025 (10 residues), Trp-cage (20 residues), and Homeodomain (52 residues) proteins.

Specifically, we explore these key questions about the reservoir approach (1) How long should the high temperature MD simulation be run such that all the relevant minima are sampled at all (non-Boltzmann), and in the correct relative weights (Boltzmann-weighted reservoir)? (2) How many structures should be selected from a high temperature MD simulation to build a Boltzmann-weighted reservoir so that the relative populations of different minima are precisely reproduced? (3) Which clustering methods are best suited to select representative structures corresponding to each minimum to build a non-Boltzmann reservoir, and how should one choose the ideal parameters for clustering such as how many clusters, which clustering metric, ideal cluster representative, etc.? (4) For a non-Boltzmann reservoir, what should be used for the potential energy of the reservoir structures? Is the energy of the representative structure for each

cluster a good measure or is the less-noisy average energy of all structures in a cluster a better measure of the cluster's energy basin? (5) What is the ideal temperature to generate structures for building a reservoir for a given protein? (6) For larger proteins, what is the expected accuracy of the ensemble generated at low T with a reservoir as compared to standard REMD? Are any inaccuracies offset by significant performance gains?

Our results show that both variants of RREMD are at least 5x faster for CLN025 and Trp-cage and at least 20x faster for Homeodomain. Moreover, the ensembles obtained by RREMD are in close agreement with standard REMD ensembles indicating that the method is reliable and scalable if around 1000-5000 structures are used in Boltzmann-weighted reservoirs, and KMeans[83] or Ward-linkage[84] clustering methods are used to select structures for building a non-Boltzmann reservoir.

2.4 Methods

2.4.1 Model systems

Three proteins – (1) CLN025 (PDB ID: 2RVD, 10 residues)[67], (2) Trp-cage (PDB ID: 1L2Y, 20 residues)[48], and (3) Homeodomain (PDB ID: 2P6J, 52 residues)[85] were considered for this study since these proteins were previously shown[12, 86] to fold accurately using the force field and solvent models used in this study. This provides us with the ability to generate precise reference data using standard methods, and under conditions that are relevant to experiments.

2.4.2 General details

All structures were built via the LEaP module of AmberTools in the AMBER 18 package[56]. For each protein, two initial conformations were built – (1) Native conformation, for which the first NMR model was used, and (2) Extended conformation, in which φ, ψ angles for all residues

except Proline were set to 180°. Proline residues were set to φ=-61.5°, ψ=-176.6°. The force field ff14SBonlysc[87] was used for all simulations. The GB-Neck2[8] (igb=8 in AMBER) implicit solvent model with mbondi3[8] radii set was used for all simulations. No cutoff was used for calculation of non-bonded interactions. For Homeodomain simulations, in addition to the polar solvation energy term calculated using the GB-Neck2 implicit solvent model, a non-polar solvation energy term was also used by calculating the solvent accessible surface area using a fast pairwise approximation (gbsa=3 in AMBER) with a surface tension of 7 cal.mol^{-1}.Å^{-2} consistent with our previous study[12]. Langevin thermostat with a collision frequency of 1 ps^{-1} was used for all simulations. SHAKE was performed on all bonds including hydrogen with the AMBER default tolerance of 0.00001 Å.

2.4.2.1 Minimization and Equilibration

A time step of 1 fs was used for all MD simulations during equilibration. With 10 kcal.mol^{-1}.Å^{-2} positional restraints on all heavy atoms, the structures built using LEaP were minimized for 1000 cycles using steepest descent and then heated from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. Then, with 10 kcal.mol^{-1}.Å^{-2} positional restraints on only backbone heavy atoms, the structures were again minimized for 1000 cycles using steepest descent and then heated again from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. This was followed by 500 ps of MD at 300 K with 1 kcal.mol^{-1}.Å^{-2} positional restraints on backbone heavy atoms and then another 500 ps of MD at 300 K with 0.1 kcal.mol^{-1}.Å^{-2} positional restraints on backbone heavy atoms. Finally, 5 ns of unrestrained MD was performed at 300 K.

2.4.2.2 Molecular Dynamics simulations

For each protein, MD simulations were performed starting from both native and extended conformations at different temperatures (see System specific details). Chirality constraints and *trans*-peptide ω constraints obtained using *makeCHIR_RST* program in AMBER were used at all

temperatures to prevent chirality inversions and peptide bond flips. A time step of 2 fs was used for all simulations. Coordinates were saved every 20 ps.

2.4.2.3 Replica Exchange Molecular Dynamics

For each protein, REMD simulations were performed starting from both native and extended conformations. Chirality constraints and *trans*-peptide ω constraints were used at all temperatures to prevent chirality inversions and peptide bond flips. A short 50 ps MD simulation using a time step of 1 fs was performed at each target temperature to briefly equilibrate each replica; thereafter the time step as 2 fs. Coordinates were saved every 20 ps. Exchanges between replicas were attempted every 1 ps for all simulations.

2.4.2.4 Reservoir Replica Exchange Molecular Dynamics

B-RREMD and nB-RREMD simulations also used the same procedure as REMD simulations, however, in addition to the exchanges between replicas, exchanges with the reservoir were also attempted every 2 ps, unless otherwise noted. Since the velocities were not saved during the high temperature MD simulations used to build the reservoir, velocities for structures obtained through exchange with the reservoir were assigned by evaluating the forces during the subsequent MD step.

2.4.3 System specific details

2.4.3.1 CLN025

Starting from each initial conformation, MD simulations were carried out for 2 μs to build reservoirs. For REMD and RREMD simulations, 4 replicas were used. Each replica was simulated for 2 μs and 400 ns for REMD and RREMD, respectively. MD simulations were run at six different temperatures – 252.3 K, 275.1 K, 300.0 K, 327.2 K, 356.8 K, and 389.1 K. The MD simulations

at 356.8 K and 327.2 K were used to select structures for a Boltzmann-weighted reservoir and a non-Boltzmann reservoir, respectively.

The replica temperatures for REMD and RREMD simulations were 252.3 K, 275.1 K, 300.0 K, and 327.2 K, and were chosen such that the probability of exchange between replicas is close to 30%.

2.4.3.2 Trp-cage

Starting from each initial conformation, MD simulations were carried out for 2 μs to build reservoirs. For REMD and RREMD simulations, 4 replicas were used. Each replica was simulated for 2 μs and 400 ns for REMD and RREMD, respectively. MD simulations were run at six different temperatures – 281.4 K, 300.0 K, 319.8 K, 340.9 K, 363.3 K, and 387.3 K. The MD simulations at 363.3 K and 340.9 K were used to select structures for a Boltzmann-weighted reservoir and a non-Boltzmann reservoir, respectively.

The replica temperatures for REMD and RREMD simulations were 281.4 K, 300.0 K, 319.8 K, and 340.9 K, and were chosen such that the probability of exchange between replicas was close to 30%.

2.4.3.3 Homeodomain

Starting from each initial conformation, MD simulations were carried out for 4 μs to build reservoirs. For REMD and RREMD simulations, 8 replicas were used. Each replica was simulated for 4 μs and 400 ns for REMD and RREMD, respectively.

MD simulations were run at nine different temperatures – 288.7 K, 300.0 K, 311.7 K, 323.9 K, 336.6 K, 349.8 K, 363.5 K, 377.7 K, and 392.4 K. The MD simulations at 392.4 K and 377.8 K were used to select structures for a Boltzmann-weighted reservoir and a non-Boltzmann reservoir, respectively.

The replica temperatures for REMD and RREMD simulations were 288.7 K, 300.0 K, 311.7 K, 323.9 K, 336.6 K, 349.8 K, 363.5 K, 377.7 K, and were chosen such that the probability of exchange between replicas is close to 30%.

2.4.4 Building Reservoirs

2.4.4.1 Building Boltzmann-weighted reservoirs

To test how long the high temperature MD simulation should be run so that structures in the reservoir occur with correct relative populations of all the relevant minima, for each protein, 6 Boltzmann-weighted reservoirs – BT1_(ext/nat), BT2_(ext/nat), and BEnd_(ext/nat) – were built using structures obtained from different time lengths of the high temperature MD simulations. The "ext" indicates that the reservoir structures were obtained from MD simulations starting from extended conformations, and "nat" indicates that the reservoir structures were obtained from MD simulations starting from native conformations. "T1" and "T2" correspond to simulation time lengths when the correlation of cluster populations between the two independent simulations starting from different initial conformations is 0.40 and 0.65, respectively (see **Results and Discussions**). "End" corresponds to the entire reservoir generation simulation time length. For CLN025, T1, T2, and End correspond to simulation time lengths of 100 ns, 200 ns, and 2 μs, respectively. For Trp-cage, T1, T2, and End correspond to simulation time lengths of 600 ns, 1.3 μs, and 2 μs, respectively. For Homeodomain, T1, T2, and End correspond to simulation time lengths of 500 ns, 1.2 μs, and 4 μs, respectively. For each of these reservoirs, 5000 structures were used to build the reservoir.

To test how many structures should be selected from the high temperature MD simulation to build a Boltzmann-weighted reservoir, for each protein, 8 Boltzmann-weighted reservoirs – B100_(ext/nat), B1000_(ext/nat), B5000_(ext/nat), and B10000_(ext/nat) – were built

representing 100, 1000, 5000, and 10000 structures, respectively. The "ext" and "nat" definitions are the same as above. For CLN025 and Trp-cage, these 100, 1000, 5000, and 10000 structures were extracted from the 2 μs MD runs with equal time spacing of 20 ns, 2 ns, 400 ps, and 200 ps, respectively, between the structures. For Homeodomain, 100, 1000, 5000, and 10000 structures were extracted from the 4 μs MD runs with equal time spacing of 40 ns, 4 ns, 800 ps, and 400 ps, respectively, between the structures. Note that B5000_(ext/nat) and BEnd_(ext/nat) are the same set of reservoir structures since both used the same length of MD simulations and the same number of structures.

The energy for each structure was calculated using the imin=5 flag in *sander* program in AMBER using the same topology file and energy parameters (igb=8, gbsa=0/3) as used in MD and REMD, for each protein. Finally, reservoirs were built using the *createreservoir* command in *cpptraj*[88] program in AMBER using a seed of 1.

2.4.4.2 Building non-Boltzmann reservoirs

For each clustering method, 40000 structures were extracted from the combined MD trajectories starting from two different initial conformations, totaling a time of 4 μs, 4 μs, and 8 μs for CLN025, Trp-cage, and Homeodomain, respectively. For CLN025 and Trp-cage, an equal time spacing of 100 ps was used to extract the structures. For Homeodomain, an equal time spacing of 200 ps was used to extract the structures. Clustering was performed using Average-Linkage (AL), KMeans, and Ward-Linkage (WL) algorithms. For each protein, for each clustering method, the entire backbone RMSD was used as the clustering metric. For each protein, details on choosing the appropriate target number of clusters for each clustering method are provided in **Results and Discussions** section. The clustering algorithm specific details are provided below.

2.4.4.2.1 **Clustering algorithms specific details**

2.4.4.2.1.1 *Average-Linkage (AL) and KMeans*

For AL and KMeans clustering algorithms, clustering was performed on the 40000 structures using *cpptraj* program in AMBER. For KMeans, a seed of 23 was used to randomize initial set of points used. The trajectories for each cluster were saved to be later used to calculate the average energy of each cluster. For each cluster, the structure with the lowest cumulative distance to all other structures within that cluster was chosen as the cluster representative.

2.4.4.2.1.2 Ward-Linkage (WL)

We used a combination of *cpptraj* and python modules to implement WL since neither package had all the capabilities required to do ward-linkage clustering on AMBER MD trajectories. The pairwise RMSDs between all 40000 structures were obtained using *cpptraj* and saved externally. Then, these pairwise RMSDs were used to perform WL clustering using the *scipy.cluster.hierarchy*[89-90] module in python. After clustering, clusters were sorted in descending order of cluster size and the structure numbers corresponding to each cluster were saved externally, separated by comma. Then, the structures corresponding to each cluster were extracted using the *onlyframes* keyword in *cpptraj* and these cluster trajectories were saved externally. Finally, cluster representatives were extracted from these saved cluster trajectories using AL clustering method by setting the target number of clusters to 1 in *cpptraj*. These representatives will, by default, correspond to the structure that has the lowest cumulative distance to every other structure in that cluster.

2.4.4.2.2 Combining the cluster representatives and cluster energies to build non-Boltzmann reservoirs

After clustering, the energy of each cluster representative (CRE) was calculated using the imin=5 flag in *sander* program as described above. The average energy of each cluster (CAE) was obtained by repeating the above step for all structures within a cluster and taking the average of the energies thus obtained. Finally, reservoirs were built using the *createreservoir* command in *cpptraj* program in AMBER using a seed of 1.

2.4.5 Analyses:

2.4.5.1 Melting curves and convergence plots

Temperature-based trajectories were extracted from REMD and RREMD simulations using the *cpptraj* program in AMBER. For each protein, the fraction of native structures in each temperature-based trajectory was calculated using the following criteria consistent with our previous published studies[8, 12, 86] for these proteins: for CLN025, a structure was characterized as native if the backbone RMSD of residues 1 to 10 was <2.0 Å to the first NMR structure, for Trp-cage, a structure was characterized as native if the backbone RMSD of residues 3 to 18 was <2.0 Å to the first NMR structure, and for Homeodomain, a structure was characterized as native if the backbone RMSD of residues 5 to 48 was <5.0 Å to the first NMR structure. For REMD simulations, since the simulations converged slowly, only for the calculation of melting curves, the last 1 μs, 1 μs, and 3 μs of data was used for CLN025, Trp-cage, and Homeodomain, respectively. Since the RREMD simulations converged quickly, all 400 ns of data was used to calculate the melting curves without excluding any data from the beginning of the simulations. The error bars indicate the half difference of the melting curves obtained from the two runs – one starting from native conformation and the other starting from extended conformation.

2.4.5.2 Calculating fraction of unique clusters observed vs time

For each protein, for each temperature, 25000 structures were extracted from the MD trajectories starting from each initial conformation. For CLN025 and Trp-cage, an equal time-spacing of 80 ps was used. For Homeodomain, an equal time-spacing of 160 ps was used. These 25000 structures from each MD run were clustered using KMeans, using *cpptraj* program, by setting the target number of clusters to 500, 1000, and 2000, for CLN025, Trp-cage, and Homeodomain, respectively. These target number of clusters were chosen based on the analysis used for identifying the ideal number of clusters for each protein (see **Results and Discussions** section). The entire backbone RMSD was used as the clustering metric. A seed of 23 was used to randomize initial set of points used. After clustering, the fraction of unique clusters observed over time were calculated by dividing the number of observed unique clusters with the target number of clusters used for each protein. A cluster was identified as unique if it was not seen in the simulation up to that point.

2.4.5.3 Calculating Correlation of cluster populations vs time

For each protein, for each temperature, the MD trajectories starting from two different initial conformations were combined and clustered together so that the resulting clusters will be the same for the two independent simulations, thereby, allowing direct comparison of the populations of each cluster obtained from each of the two independent MD simulations.

For each protein, for each temperature, 25000 structures were extracted and clustered from the combined MD trajectories starting from both initial conformations. For CLN025 and Trp-cage, an equal time-spacing of 160 ps was used to extract the structures. For Homeodomain, an equal time-spacing of 320 ps was used. These 25000 structures at each temperature were clustered using KMeans, using *cpptraj* program, by setting the target number of clusters to 500, 1000, and 2000,

for CLN025, Trp-cage, and Homeodomain, respectively. These target number of clusters were chosen based on the analysis used for identifying the ideal number of clusters for each protein (see **Results and Discussions** section). The entire backbone RMSD was used as the clustering metric. A seed of 23 was used to randomize initial set of points used.

After clustering, the population of each cluster from each MD run was calculated in cumulative time intervals of 4 ns, 4 ns, and 8 ns, for CLN025, Trp-cage, and Homeodomain, respectively. For example, for CLN025, the population of each cluster from each MD run was calculated for the first 4 ns, then for the first 8 ns, then for the first 12 ns, and so on until 2 μs of simulation time. Finally, at each of these cumulative time intervals, the Pearson correlation coefficient was calculated between the current cluster populations obtained from each MD run.

2.4.5.4 Calculation of number of folding/unfolding pair events for each protein

A folding/unfolding pair event was defined as a round trip going from a native state to a non-native state and back to the native state again. The number of folding/unfolding pair events during each high temperature MD simulation for each protein were recorded and the average and the half difference between the two runs were calculated. The same mask that was used for the calculation of melting curves was used to define the native state. A structure was characterized as non-native if the RMSD was >5.0 Å, >5.0 Å, and >10.0 Å for CLN025, Trp-cage, and Homeodomain, respectively. A larger cutoff than that used to define native was used to ensure that the non-native structures had completely unfolded and were not just fluctuations of native-like structures.

2.4.5.5 Comparing cluster populations between REMD, B-RREMD, and nB-RREMD simulations

For each protein, the trajectories close to the calculated melting temperature (275.1 K – CLN025, 300.0 K – Trp-cage, 349.8 K – Homeodomain) for each sampling method were combined (see below) and clustered together so that direct comparisons can be made between the cluster populations obtained using REMD, B-RREMD, and nB-RREMD simulations.

For REMD simulations, for each protein, 20000 structures were extracted from the combined REMD trajectories starting from both initial conformations, with an equal time-spacing of 200 ps for CLN025 and Trp-cage, and with an equal time-spacing of 400 ps for Homeodomain.

For B-RREMD simulations, for each protein, 10000 structures with an equal time-spacing of 80 ps were extracted from the combined B-RREMD trajectories starting from both initial conformations using the B5000_ext reservoirs, and another 10000 structures with an equal time-spacing of 80 ps were extracted from the combined B-RREMD trajectories starting from both initial conformations using the B5000_nat reservoirs, for all three proteins.

For nB-RREMD simulations the following protocol was used for each protein: for each of the three clustering methods used to build the non-Boltzmann reservoirs, 10000 structures with an equal time-spacing of 80 ps were extracted from the combined nB-RREMD trajectories starting from both initial conformations using the reservoir with CREs, and another 10000 structures with an equal time-spacing of 80 ps were extracted from the combined nB-RREMD trajectories starting from both initial conformations using the reservoir with CAEs. This provided 60000 structures for each protein.

Then, for each protein, the 100000 structures (20000 from REMD + 20000 from B-RREMD + 60000 from nB-RREMD) were clustered together using KMeans, using *cpptraj*

program, by setting the target number of clusters to 100. The entire backbone RMSD was used as the clustering metric for all three proteins. A random seed of 23 was used to randomize initial set of points used.

Finally, for each cluster thus obtained for each protein, the relative population of that cluster from each REMD, B-RREMD, and nB-RREMD simulation was calculated. The Pearson correlation coefficient was calculated between the cluster populations obtained from each method. The slope reported in the figures was obtained by doing a linear regression fit of the cluster populations. While the entire backbone was used for clustering, however, the RMSD values reported in the figures used the same mask as the melting curves calculations so that direct comparison can be made between the fraction of native structures shown in the melting curves and native-like clusters obtained from clustering analysis.

2.5 Results and Discussion

The following questions are addressed here: (1) How long should the high temperature MD simulation be run to ensure all relevant structures are sampled at all (non-Boltzmann), and in the correct relative weights (Boltzmann-weighted reservoir)? (2) How many structures should be selected from the high temperature MD simulation to build a Boltzmann-weighted reservoir so that the relative populations of different minima are precisely reproduced? (3) Which clustering methods are best suited to select representative structures corresponding to each minimum to build a non-Boltzmann reservoir, and how should one choose the ideal parameters for clustering such as how many clusters, which clustering metric, ideal cluster representative, etc.? (4) For a non-Boltzmann reservoir, what should the energy of each cluster representative structure be? Is the energy of the representative structure for each cluster a good measure or is the less-noisy average energy of all structures in a cluster a better measure of the cluster's energy basin? (5) What is the

ideal temperature to generate structures for building a reservoir? Testing the above choices will require generating accurate and precise reference data for each system, since the experimental data will not represent the "correct" answer using a given force field and solvent model.

To obtain the reference data, we performed extensive independent REMD simulations (see **Methods** for details) starting from native and extended conformations for each protein. For CLN025, Trp-cage, and Homeodomain, each independent REMD simulation used 4, 4, and 8 replicas, spanning a temperature range of 252.3 K to 327.2 K, 281.4 K to 340.9 K, and 288.7 K to 377.7 K, respectively. Furthermore, each replica was simulated for 2 μs, 2 μs, and 4 μs, for CLN025, Trp-cage, and Homeodomain, respectively.

To check for the convergence of REMD simulations (though we apply stricter metrics below), we calculated the fraction of native structures (shown in column 1 in **Figure 2-2**) as a function of time at each temperature, for each independent simulation, for each protein. Since the two independent REMD simulations started from very distinct initial conformations (native and fully extended) for each protein, sampling a reproducible fraction of native structures at each temperature suggests reasonable convergence.

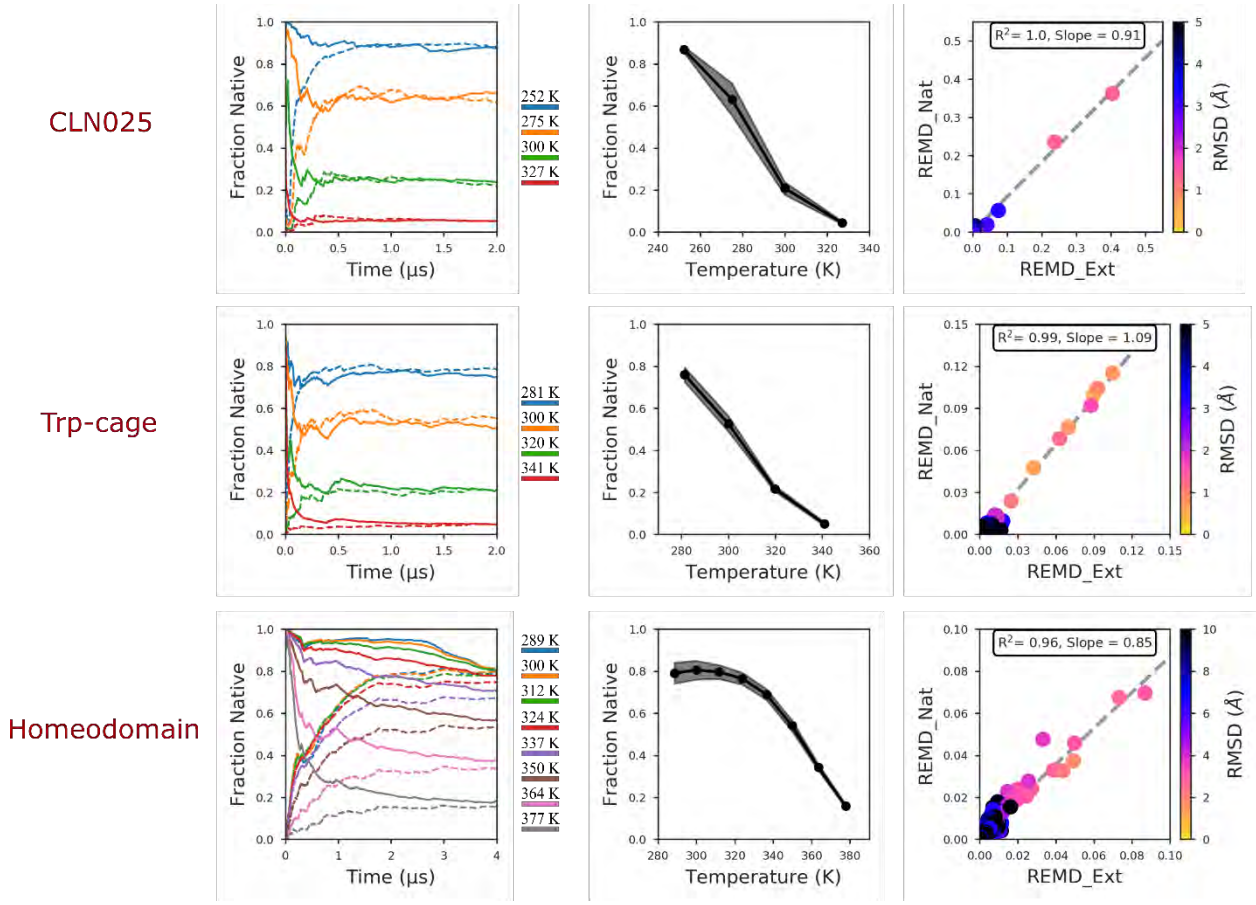


Figure 2-2 Reference data obtained from extensive standard REMD simulations. The fraction of native structures vs time at each temperature (column 1), the melting curves (column 2) and the cluster populations at the calculated melting temperature (column 3) are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). The solid and dashed lines in column 1 indicate the fraction of native structures obtained from REMD simulations starting from native and extended conformations, respectively. The error bars in column 2 (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations. "REMD_Nat" and "REMD_Ext" in column 3 indicate that the clusters were obtained from REMD simulations starting from native conformation and extended conformation, respectively. The color of each point in column 3 indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from the two independent REMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each figure in column 3 for each protein.

For CLN025, both independent simulations converge to the same amount of fraction of native structures after 700 ns of simulation per replica. Interestingly, Trp-cage simulations converge faster than CLN025 even though Trp-cage is a bigger protein than CLN025 – the two

simulations converge after 500 ns of simulation per replica. For Homeodomain, the two independent simulations converge after 3 μs of simulation per replica.

As an alternate illustration of the convergence of the native population for each protein at all temperatures, we calculated the average melting curves (shown in column 2 in **Figure 2-2**) across the two independent REMD simulations, with error bars reflecting the difference. For each protein, even though the two independent REMD simulations were started from distinct initial conformations, they result in similar melting curves (error bars are <5% at all temperatures).

Good agreement between the melting curves obtained from the two independent REMD simulations indicates that similar population of native structure is observed at all temperatures. It is important, however, to analyze the convergence of the entire ensemble, including non-native as well as native structures.[33, 91-92] To test this, we combined the trajectories from the two independent runs at the calculated melting temperature, which is 275.1 K, 300.0 K, and 349.8 K for CLN025, Trp-cage, and Homeodomain, respectively, and performed cluster analysis (see **Methods** for details). The cluster populations obtained from the two independent REMD simulations at the calculated melting temperature are shown in column 3 in **Figure 2-2**. For all three proteins, a high correlation (R^2 >0.96) with slope close to 1 is observed between the cluster populations at the calculated melting temperature, indicating that the REMD simulations not only sample similar population of native structures but also sample similar populations of non-native structures.

Overall, the melting curves and the cluster populations indicate that these standard REMD simulations are very well converged and can be used as reference data to test the various choices involved in building reservoirs. In the following sections, we explore the various choices involved in building Boltzmann-weighted reservoirs and non-Boltzmann reservoirs and test how they affect the accuracy of RREMD simulations compared to extensive conventional REMD simulations.

2.5.1 Protocols for building Boltzmann-weighted reservoirs

2.5.1.1 How long should the high temperature MD simulations be run to build a reservoir with accurate Boltzmann weighting?

In order to represent the ensemble accurately, a Boltzmann-weighted reservoir must have an appropriate number of structures populating each important local minimum. To ensure that the simulations are populating different local minima, as an initial check, we calculated the number of folding/unfolding pair events (see **Methods** section) for each protein. **Table 2-1** shows the average number of folding/unfolding pair events during the 2 μs, 2 μs, and 4 μs high temperature MD simulations for CLN025, Trp-cage, and Homeodomain, respectively. The average number of folding/unfolding pair events for CLN025, Trp-cage, and Homeodomain, are 338 ± 4, 109 ± 5, and 294 ± 21, respectively, indicating that the simulations are not trapped in native-like clusters. Surprisingly, even though Homeodomain is a bigger protein than Trp-cage, it has more folding/unfolding pair events than Trp-cage, perhaps due to the use of different temperatures used to generate the reservoirs for the two proteins (see *Ideal Temperature to generate reservoir section*).

Table 2-1 Average Number of folding/unfolding pair events for each protein.

Protein	Number of Folding/Unfolding pair events[‡]
CLN025	338 ± 4
Trp-cage	109 ± 5
Homeodomain	294 ± 21

[‡]The uncertainties correspond to half the difference between the two independent MD simulations.

Then, to ensure that the reservoir faithfully represents the correct relative populations of the different local minima, we measured the correlation of cluster populations obtained from the two simulations (starting from different initial conformations) as a function of time. This is expected to be a more stringent measure of simulation convergence as compared to only evaluating the time-dependent population of the native-like cluster. The correlation between the cluster populations should improve as the simulation length is extended.

Figure 2-3 (top) shows the correlation between cluster populations obtained from 2 high temperature MD simulations initiated from native and extended conformations. For all three proteins, the correlation of cluster populations starts at 0 and increases over time resulting in a final correlation coefficient of 0.93 (after 2 μs), 0.72 (after 2 μs), and 0.87 (after 4 μs), for CLN025, Trp-cage, and Homeodomain, respectively.

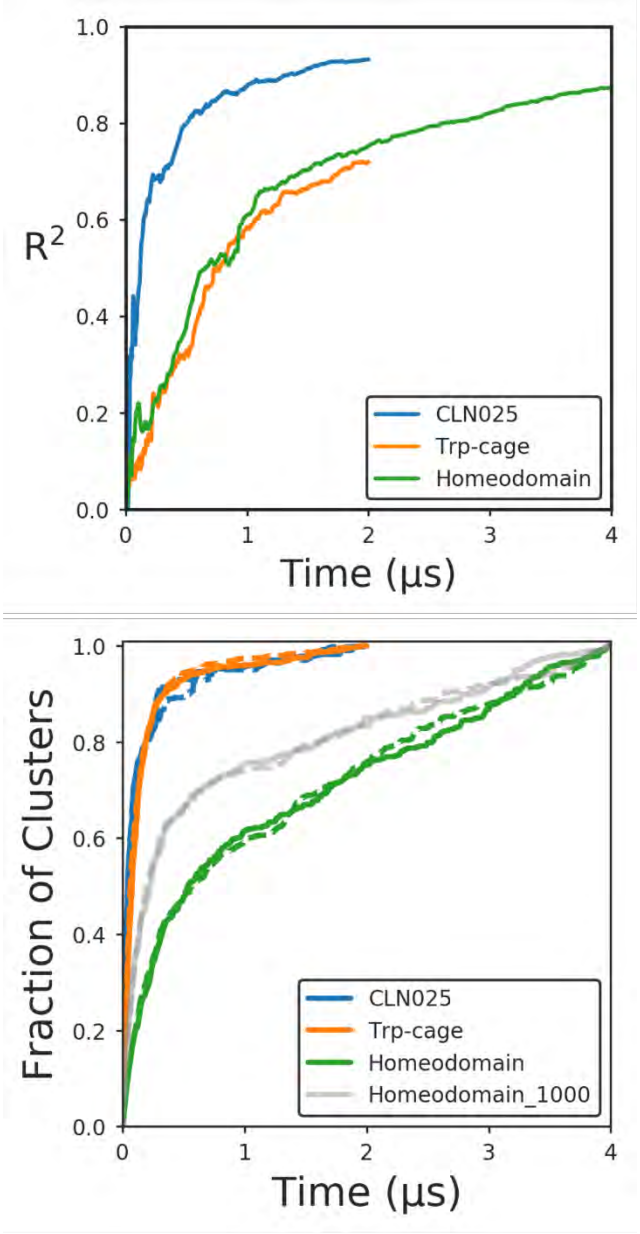


Figure 2-3 The correlation of cluster populations (top) between the two independent simulations starting from different initial conformations for each protein is shown as a function of time. The fraction of unique clusters (bottom) observed during each independent simulation for each protein is shown as a function of time. The solid lines indicate simulations starting from native conformation and the dashed lines indicate simulations starting from extended conformation. The Homeodomain_1000 indicates that the clustering was done with the target number of clusters set to 1000 instead of 2000.

As expected, due to its relatively small size compared to the other two proteins, CLN025 has the fastest increase in correlation of cluster populations (R^2 = 0.8 after 0.5 μs). Surprisingly,

even though Trp-cage is a much smaller protein than Homeodomain, the correlation of cluster populations increases at a similar rate for Trp-cage (R^2 = 0.72 after 2 μs) and Homeodomain (R^2 = 0.75 after 2 μs), suggesting that Trp-cage, like Homeodomain, might need longer simulations for complete convergence.

In contrast to the population correlation, **Figure 2-3** (bottom) shows that more than 0.9 fraction of unique clusters are observed within the first 0.5 μs of simulations starting from both initial conformations for CLN025 and Trp-cage. In contrast to CLN025 and Trp-cage, the Homeodomain simulations sample only 0.75 fraction of unique clusters within the first 2 μs, suggesting that these simulations of the larger protein still may not have sampled all relevant minima and perhaps should be extended. The discrepancy between the rate of sampling of unique clusters for Trp-cage and Homeodomain simulations could also be due to the higher target number of clusters for Homeodomain (2000 clusters compared to only 1000 for Trp-cage) which means that to achieve a similar fraction as Trp-cage simulations, more clusters must be sampled for Homeodomain as compared to Trp-cage. However, even after setting the target number of clusters to 1000, the Homeodomain simulations sample only 0.85 fraction of unique clusters after 2 μs indicating that Homeodomain reservoir generation simulations must be extended. Therefore, Homeodomain reservoir generation simulations were extended to 4 μs to ensure that the two reservoir generation simulations are well converged.

Nevertheless, **Figure 2-3** also shows that, especially for Trp-cage, both independent simulations sample the majority of the minima significantly earlier than they sample these minima with the correct relative populations. This is because, in the latter scenario, each minimum has to be revisited multiple times to obtain precise population estimates and thus, is a time-consuming process compared to simply sampling each minimum at least once. This illustrates the potential

advantage of nB-RREMD compared to B-RREMD since nB-RREMD reservoirs do not require relative populations of different minima, and thus could be generated more quickly. This will be explored in more detail in a later section.

Overall, multiple folding/unfolding pair events, the correlation of cluster populations between the two independent simulations, and the rate of observing unique clusters in each of the two independent simulations indicate that the simulation times of 2 μs, 2 μs, and 4 μs might be sufficient to build precisely Boltzmann-weighted reservoirs for each independent run for CLN025, Trp-cage, and Homeodomain, respectively. Moreover, if the two independent simulations are reasonably converged, then B-RREMD simulations employing reservoirs built from these independent simulations should generate ensembles at low T that are similar to the reference data without reservoir use.

To test this, for each B-RREMD simulation used below, we built 6 (3*2) reservoirs using structures obtained from three different time lengths (T1, T2, and End) from each independent high temperature MD simulation to determine the length of reservoir generation time lengths required to build a Boltzmann-weighted reservoir for each protein. "T1" and "T2" correspond to simulation time lengths when the correlation of cluster populations (shown in **Figure 2-3** (top)**)** between the two independent reservoir generation simulations starting from different initial conformations is 0.40 and 0.65, respectively. "End" corresponds to the entire reservoir generation simulation time length. Consequently, T1, T2, and End correspond to different reservoir generation time points for the three proteins since the correlation values are different at different time points for each protein (see **Methods** for details).

To distinguish between the six reservoirs for each protein, the following reservoir naming convention was used – B(T)_(c) reservoir indicates a Boltzmann-weighted reservoir with 5000

structures obtained from the high temperature MD simulation starting from "c" conformation up to the reservoir generation time length "T". For example, BT1_ext and BT1_nat indicate that the reservoir has 5000 structures that were obtained from the high temperature MD simulations up to time T1 starting from extended conformation and native conformation, respectively. In summary, these reservoirs represent 2 different ensembles, each sampled at 3 different precisions.

Figure 2-4 shows the melting curves obtained from B-RREMD simulations using each of the six B(T)_(c) reservoirs compared to standard REMD simulations, for all three proteins. When the correlation of cluster populations between the two independent reservoir generation simulations starting from different initial conformations is 0.40 (T1 time point), the melting curves obtained from B-RREMD do not match with standard REMD melting curves – the fraction of native structures are often overestimated or underestimated at all temperatures. For CLN025, the B-RREMD simulations using BT1_nat and BT1_ext result in 0.44 and 0.39 fraction of native structures at 300 K, respectively, compared to 0.21 from standard REMD simulations. For Trp-cage, the B-RREMD simulations using BT1_nat and BT1_ext result in 0.63 and 0.30 fraction of native structures at 300 K, respectively, compared to 0.53 from standard REMD simulations. For Homeodomain, the B-RREMD simulations using BT1_nat and BT1_ext result in 0.43 and 0.24 fraction of native structures at 300 K, respectively, compared to 0.80 from standard REMD simulations.

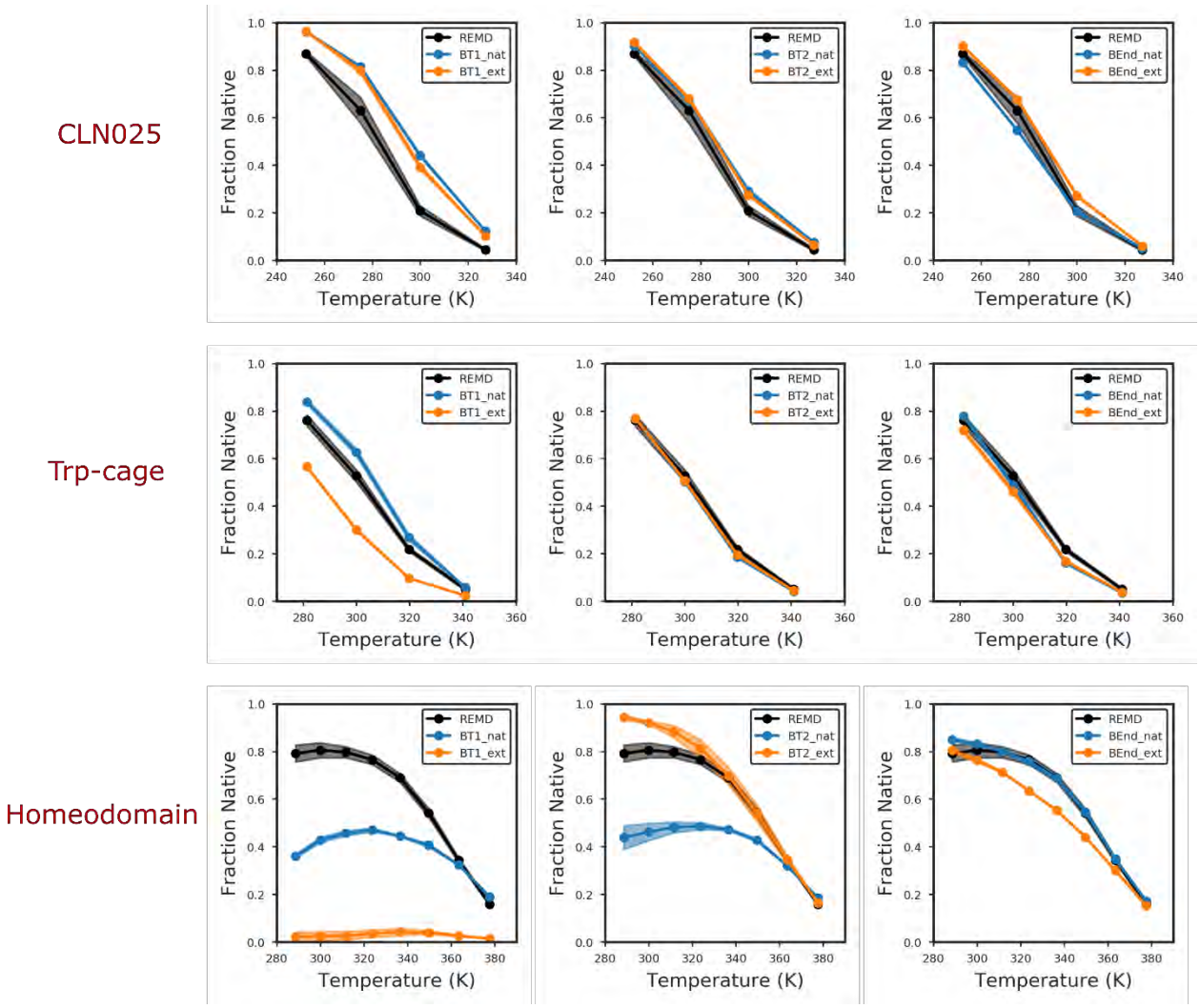


Figure 2-4 The melting curves obtained using standard REMD (black) and B-RREMD simulations using structures obtained from three different time lengths are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). T1, T2, and End, indicate different time lengths for which the reservoir generation simulations were run (see text for details). The "nat" (blue) and "ext" (orange) indicate that the B-RREMD simulations were carried out with reservoir structures obtained from high temperature MD simulations starting from native and extended conformations, respectively. The error bars indicate the half difference of the melting curves obtained from two B-RREMD simulations – one starting from native conformation and the other starting from extended conformation, using the same set of reservoir structures. For some B-RREMD simulations, the error bars are negligible and hence are not visible on the graphs.

Increasing the time length to T2 ($R^2 = 0.65$) significantly improves the agreement between B-RREMD and standard REMD simulations for CLN025 and Trp-cage but not for Homeodomain. For CLN025, the B-RREMD simulations using BT2_nat and BT2_ext result in 0.29 and 0.29

fraction of native structures at 300 K, respectively. For Trp-cage, the B-RREMD simulations using BT2_nat and BT2_ext result in 0.50 and 0.51 fraction of native structures at 300 K, respectively. For Homeodomain, the B-RREMD simulations using BT2_nat and BT2_ext result in 0.46 and 0.92 fraction of native structures at 300 K, respectively.

Further increasing the reservoir generation time length to End (R^2 >0.70) results in good agreement between the melting curves obtained using B-RREMD and standard REMD simulations for all three proteins including Homeodomain. For CLN025, Trp-cage, and Homeodomain, the BEnd_nat and BEnd_ext reservoirs result in 0.21 and 0.27, 0.49 and 0.46, and 0.83 and 0.76, fraction of native structures, respectively.

In summary, the above results indicate that, as expected, there is no one size fits all solution to estimate length of reservoir generation simulations for building a precisely Boltzmann-weighted reservoir for a given protein – CLN025, Trp-cage, and Homeodomain require 200 ns, 1.3 μs, and 4 μs, of reservoir generation simulations, respectively. More importantly, while standard REMD simulations can be improved by extending them, B-RREMD simulations using imprecise reservoirs will possibly converge to wrong answer and running the B-RREMD simulations longer won't help. Instead, it is critical to ensure that the reservoirs are well converged by measuring observables such as the correlation of cluster populations between the two reservoir generation simulations as a function of time before running B-RREMD simulations.

2.5.1.2 How many reservoir structures are needed to represent a Boltzmann-weighted ensemble?

If the number of structures in the reservoir are too few, the reservoir might not reflect the relative populations of the different minima precisely enough and using such a reservoir for B-

RREMD simulations could result in erroneous ensemble distributions at all temperatures. To determine the minimum number of structures required for building a precisely Boltzmann-weighted reservoir, for each protein, we built 8 (4*2) reservoirs using 100, 1000, 5000, and 10000 equidistant structures selected from each independent high temperature MD simulation. Note that these structures were obtained from the full 2 μs, 2 μs, and 4μs, reservoir generation simulations for CLN025, Trp-cage, and Homeodomain, respectively (see **Methods** for details).

To distinguish between the eight reservoirs for each protein, the following reservoir naming convention was used – B(N)_(c) reservoir indicates a Boltzmann-weighted reservoir with "N" structures obtained from the high temperature MD simulation starting from "c" conformation. For example, B100_ext and B100_nat indicate that the reservoir has 100 structures that were obtained from the high temperature MD simulations starting from extended conformation and native conformation, respectively. In summary, these reservoirs represent 2 different ensembles, each sampled at 4 different precisions. We expect that the B-RREMD simulations using these reservoirs will match the standard REMD results, unless the reduction in number of structures reduces the precision such that the reservoir is no longer properly Boltzmann weighted.

Figure 2-5 shows the melting curves obtained from B-RREMD simulations using six (B10000_(ext/nat), B1000_(ext/nat), and B100_(ext/nat)) of the eight reservoirs compared to standard REMD simulations, for all three proteins. B-RREMD simulations using B5000_(ext/nat) are the same as BEnd_(ext/nat) shown in **Figure 2-4** and hence, are not shown in **Figure 2-5**. Irrespective of the number of structures used in the reservoir, the error bars on the melting curves for RREMD simulations are significantly smaller (sometimes negligible) than the error bars for standard REMD simulations indicating that, given a set of structures, RREMD simulations starting from different initial replica conformations converge to the same answer.

Furthermore, when only 10000 structures are used in the reservoir, the melting curves obtained using B-RREMD simulations are in close agreement with standard REMD melting curves. For CLN025, the B-RREMD simulations using B10000_nat and B10000_ext result in 0.22 and 0.27 fraction of native structures at 300 K, respectively, compared to 0.21 from standard REMD simulations. For Trp-cage, the B-RREMD simulations using B10000_nat and B10000_ext result in 0.50 and 0.48 fraction of native structures at 300 K, respectively, compared to 0.53 from standard REMD simulations. For Homeodomain, the B-RREMD simulations using B10000_nat and B10000_ext result in 0.82 and 0.79 fraction of native structures at 300 K, respectively, compared to 0.80 from standard REMD simulations.

Reducing the number of structures to 5000 does not affect the melting curves significantly (see BEnd_(ext/nat) data in **Figure 2-4**). The melting curves obtained using B-RREMD simulations are still in close agreement with standard REMD melting curves. For CLN025, Trp-cage, and Homeodomain, the B5000_nat and B5000_ext reservoirs result in 0.21 and 0.27, 0.49 and 0.46, and 0.83 and 0.76, fraction of native structures, respectively.

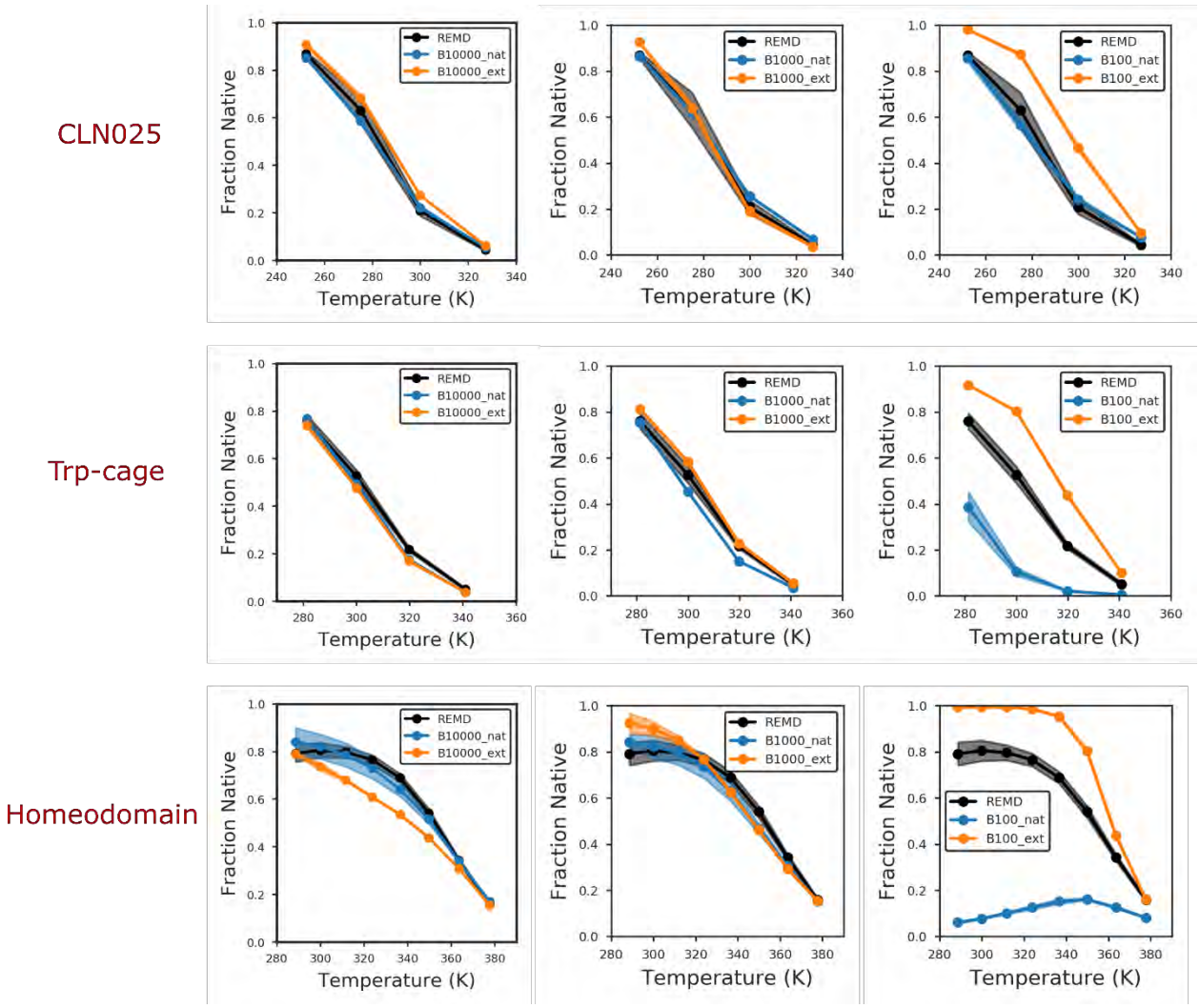


Figure 2-5 The melting curves obtained using standard REMD (black) and B-RREMD simulations using different number of reservoir structures are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). B10000, B1000, and B100 reservoirs indicate reservoirs having 10000, 1000, and 100 structures, respectively. The "nat" (blue) and "ext" (orange) indicate that the B-RREMD simulations were carried out with reservoir structures obtained from high temperature MD simulations starting from native and extended conformations, respectively. The error bars indicate the half difference of the melting curves obtained from two B-RREMD simulations – one starting from native conformation and the other starting from extended conformation, using the same set of reservoir structures. For some B-RREMD simulations, the error bars are negligible and hence are not visible on the graphs.

Further reducing the number of structures to 1000 results in slightly but not significantly worse agreement between B-RREMD and standard REMD melting curves. For CLN025, Trp-

cage, and Homeodomain, the B1000_nat and B1000_ext reservoirs result in 0.26 and 0.19, 0.45 and 0.58, and 0.82 and 0.90, fraction of native structures, respectively.

In contrast, when only 100 structures are used, the melting curves obtained using B-RREMD simulations do not match with standard REMD melting curves – the fraction of native structures are significantly overestimated or underestimated at all temperatures. For CLN025, the B-RREMD simulations using B100_nat and B100_ext result in 0.24 and 0.46 fraction of native structures at 300 K, respectively. For Trp-cage, the B-RREMD simulations using B100_nat and B100_ext result in 0.10 and 0.80 fraction of native structures at 300 K, respectively. For Homeodomain, the B-RREMD simulations using B100_nat and B100_ext result in 0.08 and 0.99 fraction of native structures at 300 K, respectively.

The above results indicate that the optimal number of structures for building a Boltzmann-weighted reservoir at high T can be as low as 1000 to 5000. This number is significantly less than the 10000 to 150000 number of structures that have been previously used to build Boltzmann-weighted reservoirs for similar sized biomolecules.[32, 49, 63-64] It may be that a few thousand structures is sufficient at high T because the important basins have similar populations when sampled well above the melting temperature. At lower T, the differences in population are likely amplified, and thus more structures would be needed in the reservoir to be able reproduce these population differences.

2.5.1.3 Can B-RREMD simulations reproduce the <u>overall ensemble</u> obtained from standard REMD simulations?

Good agreement between the melting curves obtained from B-RREMD and standard REMD simulations suggests that both methods result in similar population of native structure at

all temperatures. However, B-RREMD simulations should reproduce the full ensemble from the reference data, including populations of non-native as well as native structures.

To test if B-RREMD simulations can accurately reproduce the ensemble obtained from standard REMD simulations, we performed clustering on the combined trajectories of B-RREMD simulations and standard REMD simulations at the temperature close to the calculated melting temperature, where multiple conformations should contribute (see **Methods**). This combined clustering ensures a consistent set of clusters for the reference and B-RREMD ensembles. The cluster populations at 275.1 K, 300.0 K, and 349.8 K, for CLN025, Trp-cage, and Homeodomain, respectively, from B-RREMD and standard REMD simulations are shown in **Figure 2-6**.

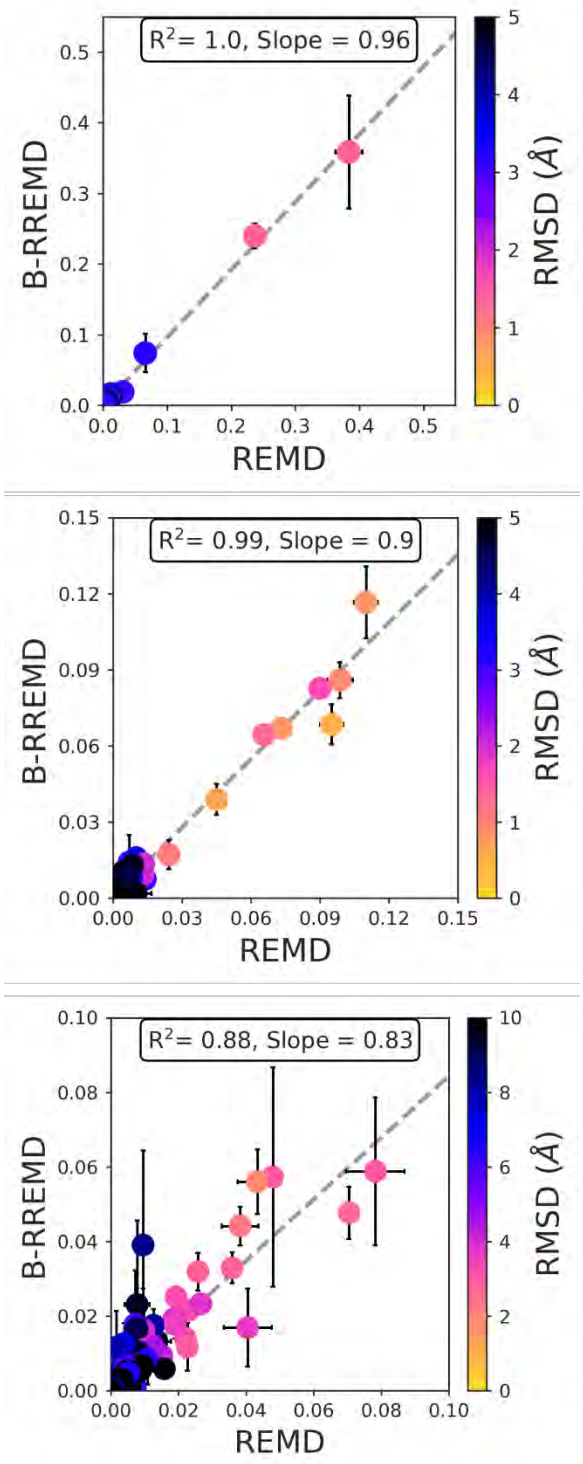


Figure 2-6 Cluster populations obtained using standard REMD (X-axis) and B-RREMD (Y-axis) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom) at 275.1 K, 300.0 K, and 349.8 K, respectively. The color of each point indicates the RMSD of the cluster representative to the native structure. The error bars on the X-axis indicate the half difference of cluster populations obtained from the two REMD runs – one starting from native conformation and the other starting

from extended conformation. The error bars on the Y-axis indicate the standard deviation of cluster populations obtained from the two sets of B-RREMD runs (4 simulations in total) – one starting from native conformation and the other starting from extended conformations, for both B5000_ext and B5000_nat reservoirs. Error bars for some clusters are not visible since they are smaller than the point size. The Pearson correlation coefficient between the cluster populations obtained from standard REMD and B-RREMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each protein.

For all three proteins, the ensembles sampled by B-RREMD simulations are in good agreement with ensembles sampled by standard REMD simulations. The most populated cluster is the same between standard REMD and B-RREMD. In addition to accurately reproducing the most populated cluster from standard REMD, B-RREMD also reproduces the populations of other clusters reasonably well. For CLN025, the correlation of cluster populations between the two methods is 1.00 and the slope is 0.96. For Trp-cage, the correlation of cluster populations is 0.99 and the slope is 0.90. For Homeodomain, the correlation of cluster populations is 0.88 and the slope is 0.83.

For all three proteins, the error bars on the X-axis (standard REMD) in **Figure 2-6** are small. This is expected since the two standard REMD simulations starting from different initial conformations were extended until they were well converged (see **Figure 2-2**). The error bars on the Y-axis, on the other hand, are non-negligible and reflect the difference in populations of four different B-RREMD simulations – two simulations (starting from different initial conformations) each using B5000_ext and B5000_nat reservoirs.

There are two possible sources for these non-negligible differences in the cluster populations of B-RREMD simulations: (1) Given the same reservoir, B-RREMD simulations starting from two different initial conformations may result in different ensembles, with significant differences in cluster populations between the two simulations, or (2) the differences in the populations could stem from the use of two different reservoirs and not from sensitivity to the

initial structures. However, the small error bars in the B-RREMD melting curves in **Figure 2-4** and **Figure 2-5** suggest minimal uncertainty from varying the initial structures.

To further explore this, the cluster populations between B-RREMD simulations, using the same reservoir but starting from two different initial conformations, are shown in **Figure 2-7**. When the same set of reservoir structures are used (either B5000_ext reservoir or B5000_nat reservoir), the clusters obtained by the two B-RREMD simulations starting from different initial conformations are in excellent agreement – the correlation of cluster populations is >0.96 for all three proteins. This supports the earlier conclusion that independent B-RREMD simulations using the same reservoir converge to the same ensembles, irrespective of initial structure. Therefore, it is critical to ensure that the reservoirs are well converged before running B-RREMD simulations.

Nonetheless, performing long MD simulations at a single high temperature followed by short B-RREMD simulations at all temperatures saves considerable computing resources compared to performing long REMD simulations at all temperatures.

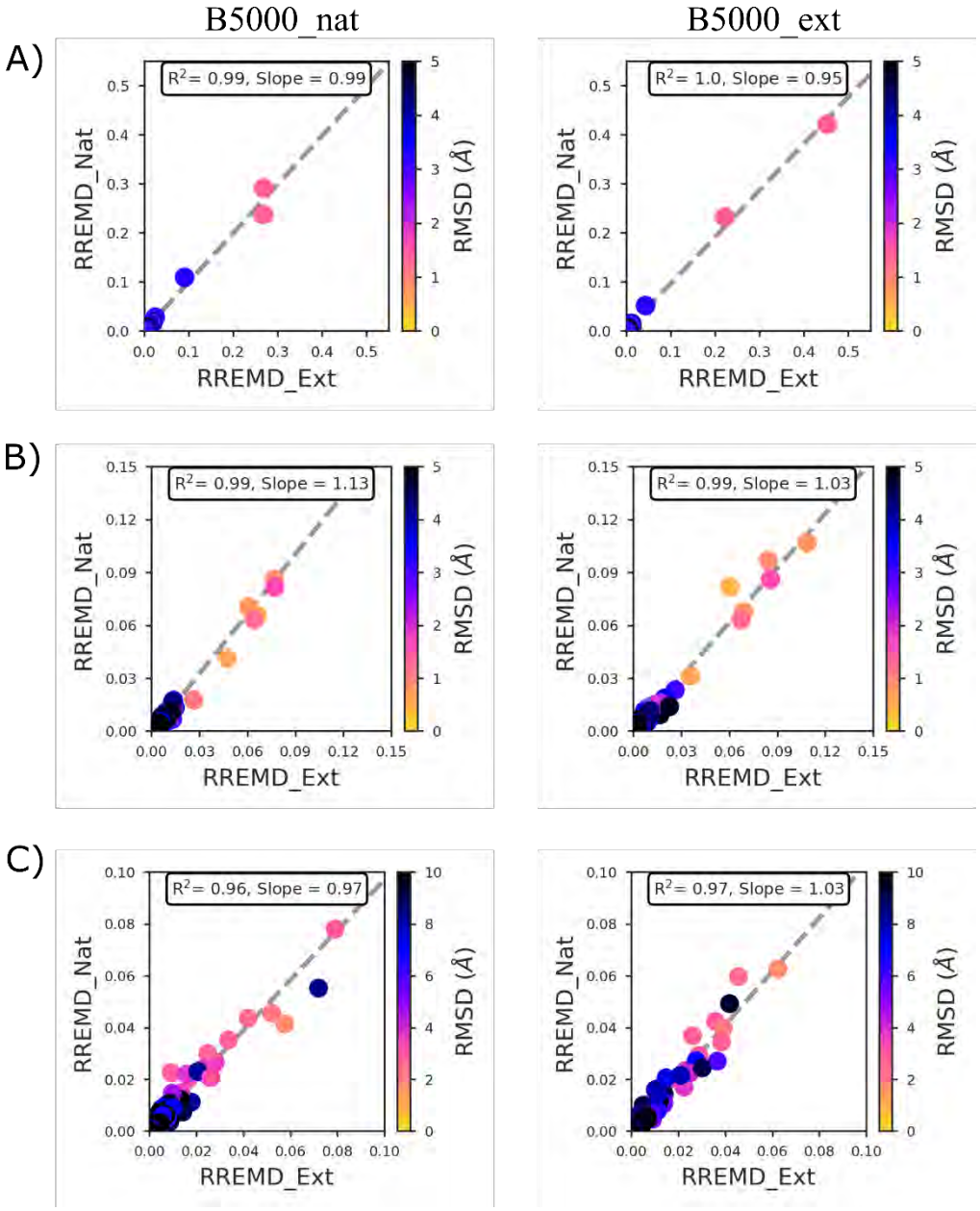


Figure 2-7 Cluster populations obtained using B-RREMD simulations starting from extended conformation (X-axis) and B-RREMD simulations starting from native conformation using the same set of reservoir structures (Y-axis) for A) CLN025, B) Trp-cage, and C) Homeodomain at 275.1 K, 300.0 K, and 349.8 K, respectively, using B5000_nat reservoir (left column) and B5000_ext reservoir (right column). The color of each point indicates the RMSD of the cluster representative to the native structure. The Pearson correlation coefficient between the cluster populations obtained from standard REMD and B-RREMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each protein.

2.5.2 Protocols for building non-Boltzmann reservoirs

2.5.2.1 Motivation for use of non-Boltzmann reservoirs.

Generating a properly Boltzmann-weighted ensemble of structures at high temperatures requires considerable computing resources for a Homeodomain-sized biomolecule (4 μs of simulation), and this will only get more challenging for larger biomolecules (particularly in explicit solvent). Moreover, only those sampling methods that generate a canonical ensemble of structures can be used, further limiting the range of efficient sampling methods that could be harnessed to rapidly construct a reservoir.

On the other hand, since it is significantly faster to sample the unique clusters at least once than to sample enough transitions to obtain the correct relative populations (see **Figure 2-3**), generating a reservoir with a non-Boltzmann population should be a much more tractable task in MD. Moreover, a non-Boltzmann reservoir offers greater flexibility in generating structures since many sampling methods (physics-based and non-physics-based) could, in principle, be used to obtain the structures. However, a non-Boltzmann reservoir still requires a well-defined distribution of structures; ensuring such distributions is also challenging, and may require careful consideration.[49] In the following section, we explore protocols for building a non-Boltzmann reservoir, and the impact of these choices on the resulting nB-RREMD ensembles generated using these reservoirs.

2.5.2.2 Critical components for building a non-Boltzmann reservoir

nB-RREMD in its simplest form requires an unweighted ("flat") distribution of reservoir structures, i.e., structures corresponding to each relevant local minimum should be selected and represented equally.[49] One way to obtain a flat distribution is to cluster the structures and select only one structure per cluster. However, there is not a one-size-fits-all clustering algorithm, and

different clustering protocols have to be tailored to the specific problem for which clustering is used. For example, to identify the most populated cluster, density-based clustering algorithms such as DBSCAN might be better[80]. However, to identify metastable states with very low density, a partitioning algorithm such as KMeans might be better[82]. Besides, the parameters of both DBSCAN and KMeans can be fine-tuned and used for the same clustering problem.

So, which clustering methods are ideal for building a non-Boltzmann reservoir? Which metric should be used for clustering? How should the ideal number of clusters (or minima) be identified for a given clustering method (or protein)? We explored above how many structures are needed to represent basin weights in a Boltzmann reservoir, but the ideal number of structures for a non-Boltzmann reservoir likely differs. Also, since only a single structure is used to represent each local minimum, the accuracy of nB-RREMD might be sensitive to the energy used for the structure during the exchange. While the potential energy of the representative structure seems reasonable, is the average energy of all structures in the cluster more indicative of the energy of the basin it represents? In this section, we explore different protocols that can help in making the appropriate choices necessary to build a "good" non-Boltzmann reservoir.

2.5.2.2.1 Trajectories used for building non-Boltzmann reservoirs

Since the exchange criterion for the non-Boltzmann RREMD (nB-RREMD) uses only the energy of the reservoir structure and not its temperature, for building non-Boltzmann reservoirs, we used the MD trajectories at the same temperature as the highest replica temperature for each protein. This will minimize the thermal energy differences between the reservoir structures and the structures in the highest temperature replica, ensuring efficient MC exchanges with the reservoir. In this manner, exchange with the non-Boltzmann reservoir simply corresponds to a

standard MC jump to a randomly selected basin on the energy landscape, which happens to have been sampled in advance during reservoir generation.

We begin construction of the non-Boltzmann reservoirs by combining the trajectories from the same high temperature MD simulations that were used to build Boltzmann reservoirs. Since the B-RREMD simulations gave similar results to standard REMD (**Figures 2-4, 2-5, and 2-6**), if the NB-REMD gives different results it will indicate issues with the protocols for selecting representative structures and energies, rather than problems with the source MD data.

2.5.2.2.2 Identifying a good clustering protocol for building non-Boltzmann reservoirs

Next, we clustered the combined trajectories using different clustering methods (see **Methods**) to extract representative structures corresponding to each minimum. A reservoir built using only these representative structures should, in principle, result in a flat distribution. While removing the cluster weights may result in reduced accuracy vs. a converged Boltzmann reservoir, it may improve the results as compared to using a poorly converged Boltzmann reservoir with inaccurate weights.

A good clustering protocol should result in homogeneous clusters, i.e. all the structures in each cluster should look alike. Therefore, a metric for obtaining good clusters is a low intra-cluster RMSD variance. Also, if a given cluster is homogeneous (no outliers), the difference between average and median intra-cluster RMSD should be close to zero.

In practice, obtaining homogeneous clusters from MD trajectories is a non-trivial task due to the following reasons: (1) the choice of the clustering method influences the final clusters that are obtained, (2) the metric (RMSD or backbone dihedrals, which region, besides others) used for clustering also changes the clustering results, (3) it is difficult to know beforehand the ideal number of clusters for a given trajectory, and (4) as stated before, the choice of the clustering method, the

choice of clustering metric, and identifying the ideal number of clusters are in turn dependent on the clustering problem. Since it is prohibitive to test all possible combinations, we outline here a simple clustering protocol to extract structures for building a non-Boltzmann reservoir and compare results with a few common variants.

2.5.2.2.3 Clustering methods used in this study

Selecting structures for building non-Boltzmann reservoirs is akin to identifying the macro-states to build a Markov-state model. We should be able to identify the native states, intermediate states, and also the unfolded ensemble. Since hierarchical clustering using Ward-Linkage (WL) and partitioning clustering using KMeans were found to result in the best Markov-state models[81], we used these clustering algorithms in this study. We also used hierarchical clustering using Average-Linkage (AL) since it is a commonly used clustering algorithm for clustering MD trajectories. In addition to these three clustering methods, we also tried hierarchical clustering using Complete-Linkage which resulted in clusters with very high variance (data not shown) and DBSCAN clustering algorithm which resulted in too few (<15) and mostly native-like clusters (data not shown).

2.5.2.2.4 Clustering metric used in this study

For all three proteins, for all three clustering algorithms, we used the entire backbone heavy-atom RMSD of the protein as the clustering metric. Using the entire backbone for clustering will ensure that the variation in flexible termini or loop regions of the protein are also accounted for because these regions of the protein will also contribute to the energy of the structures (see below) that is assigned to the cluster in the reservoir.

2.5.2.2.5 Identifying the ideal number of clusters

To approximately identify the ideal number of clusters for each clustering method for each protein, we performed cluster analysis by setting the target number of clusters to different numbers for each clustering method for each protein. For example, for Homeodomain, we performed six different cluster analyses using KMeans by setting the target number of clusters to 100, 500, 1000, 2000, 3000, and 4000. Then, for each target number of clusters using KMeans, we calculated the average, median, and variance of intra-cluster RMSDs for all the clusters that were obtained using that target number of clusters. These RMSD values corresponding to each cluster obtained by setting the target number of clusters to 100, 500, 1000, 2000, 3000, and 4000, using KMeans for Homeodomain, are shown in **Figure 2-8**.

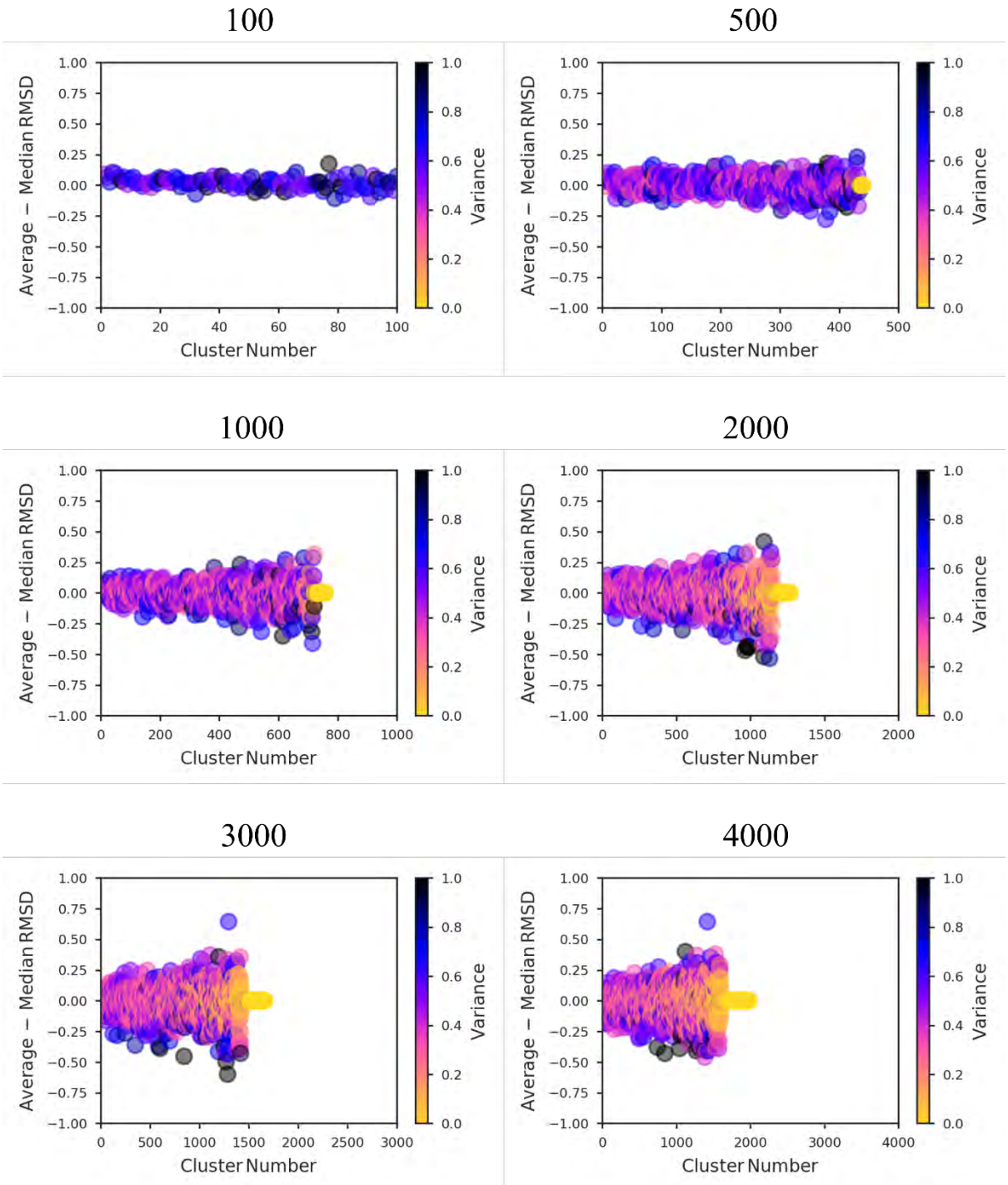


Figure 2-8 The difference between the intra-cluster average RMSD and the intra-cluster median RMSD for each cluster obtained by setting the target number of clusters to 100, 500, 1000, 2000, 3000, and 4000, using KMeans for Homeodomain, is shown. The color of each point (shown with a transparency of 50% to make overlapping points visible) indicates the intra-cluster RMSD variance and is used to identify best clusters.

For all the clusters, the difference between average and median intra-cluster RMSDs is close to zero indicating that the clusters might be homogeneous, however, setting the target number of clusters to only 100 results in many (>85%) clusters having a high intra-cluster RMSD variance (>0.5Å^2) indicating that a higher target number of clusters should be used for clustering to obtain homogeneous clusters. Increasing the target number of clusters to 500 or 1000 results in around 51% and 70% of non-singleton clusters with an intra-cluster RMSD variance of <0.5 Å^2, respectively. Further increasing the target number of clusters to 2000 results in >84% of the clusters with an intra-cluster RMSD variance of <0.5Å^2. While setting the target number of clusters to 3000 or 4000 also results in many clusters (>85%) with an intra-cluster RMSD variance of <0.5Å^2, they also result in many (>45%) singleton clusters. These singleton clusters (with only one structure in them) have a variance of zero and hence, are invisible in **Figure 2-8**. Based on the above results, we chose the clusters obtained by setting the target number of clusters to 2000 as the best clusters for Homeodomain using KMeans, since it resulted in >80% clusters with a low intra-cluster RMSD variance and only 35% singleton clusters.

After clustering, we sorted the best clusters in descending order based on the number of structures in them and selected the representative structures of clusters that contain at least 4 structures, rounded off to the nearest upper multiple of 50. For example, if the first 1145 clusters have at least 4 structures in them and the clusters after that have 3 or fewer structures, we picked representative structures from the first 1150 clusters to build the reservoir.

The choice of using intra-cluster RMSD variance <0.5 Å^2, using <35% singleton clusters as cutoff, picking clusters that have at least 4 structures, and rounding off to the nearest multiple of 50 are arbitrary and the possible impact of these choices could be explored further in future work.

The above process was repeated for all three proteins using all three clustering methods and the resulting number of clusters that were selected for each clustering method for each protein are shown in **Table 2-2**.

Table 2-2 Number of clusters used for each protein for each clustering method.

	CLN025	Trp-cage	Homeodomain
Average-Linkage (AL)	400 (500)	1000 (1000)	1400 (2000)
KMeans	500 (500)	1000 (1000)	1150 (2000)
Ward-Linkage (WL)	500 (500)	1000 (1000)	2000 (2000)

Numbers in parenthesis indicate the target number of clusters that was used to obtain these clusters.

2.5.2.2.6 Limitations in the clustering protocol

Using trajectories from both independent simulations and clustering them together might seem very costly since two simulations were used to generate structures for the reservoir. However, since the B-RREMD simulations indicated (see **Figures 2-4, 2-5,** and **2-6**) that the high temperature MD simulations have sampled all the relevant minima, using the combined trajectories and clustering them can inform us which clustering methods are the best at selecting structures for each minimum given that all the relevant minima are present.

Alternatively, since most of the clusters were observed in the first 500 ns for CLN025 and Trp-cage (see **Figure 2-3**), using the trajectories up to 500 ns for these two proteins and clustering on them might be sufficient. This is something that we plan to explore in the future, but for now, we cluster the combined trajectories.

Using only the backbone heavy-atoms for clustering might also be a limitation since each backbone conformation may have multiple side chain rotamers in the original ensemble. For simplicity here we assume that the best backbone representative will also have the best side chain rotamers, and that side chain transitions can be sampled readily during the REMD phase. In principle, however, this approach neglects possible side chain entropy differences between the backbone clusters, since side chain variants on the same backbone correspond to unique clusters on the multidimensional landscape. Inclusion of side chains in the cluster analysis should be a straightforward extension. Nevertheless, influence of the side chain rotamer variance on reservoirs can be explored in more detail in future work.

2.5.2.2.7 Creating the non-Boltzmann reservoirs

After following the above clustering protocol, we built six (3*2) non-Boltzmann reservoirs for each protein as follows: (1) For each cluster, the structure with the lowest cumulative backbone heavy-atom RMSD (which is the same mask used for clustering) to all other structures was chosen as the cluster representative. This was a reasonable and easy choice, since, *cpptraj*, by default, outputs the structure with the lowest cumulative RMSD as the cluster representative. (2) Then, for each set of representative structures from each clustering method, we built two sets of reservoirs using (a) the energy of only the representative structure for each cluster, denoted as cluster representative energy (CRE), and (b) the average of the energies of all the structures in each cluster, denoted cluster average energy (CAE). Since only the backbone was used for clustering, using the average energy of all structures in a cluster could account for possible multiple side chain orientations of the cluster representative and thus, may be a better representative of the overall relative energy of that cluster (minimum).

To distinguish between the six non-Boltzmann reservoirs for each protein, the following reservoir naming convention was used – nB_(CM)_(CE) reservoir indicates that the "CM" clustering method and the "CE" cluster energy was used for each structure in the non-Boltzmann reservoir. For example, nB_AL_CRE and nB_AL_CAE indicate that the non-Boltzmann reservoir was built using representative structures obtained from Average-linkage clustering method and using the cluster representative energy (CRE) and the cluster average energy (CAE), respectively.

2.5.2.3 Sensitivity of nB-RREMD to clustering method and the energy (CRE or CAE) used in building the non-Boltzmann reservoirs.

For each non-Boltzmann reservoir built using the above protocols, we performed nB-RREMD simulations for each protein to test the influence of the chosen clustering method (Average-Linkage, KMeans, or Ward-Linkage) and the chosen energy (CRE or CAE) of each cluster for building non-Boltzmann reservoirs. **Figure 2-9** shows the melting curves obtained from nB-RREMD simulations using each of the six non-Boltzmann reservoirs compared to standard REMD simulations, for all three proteins.

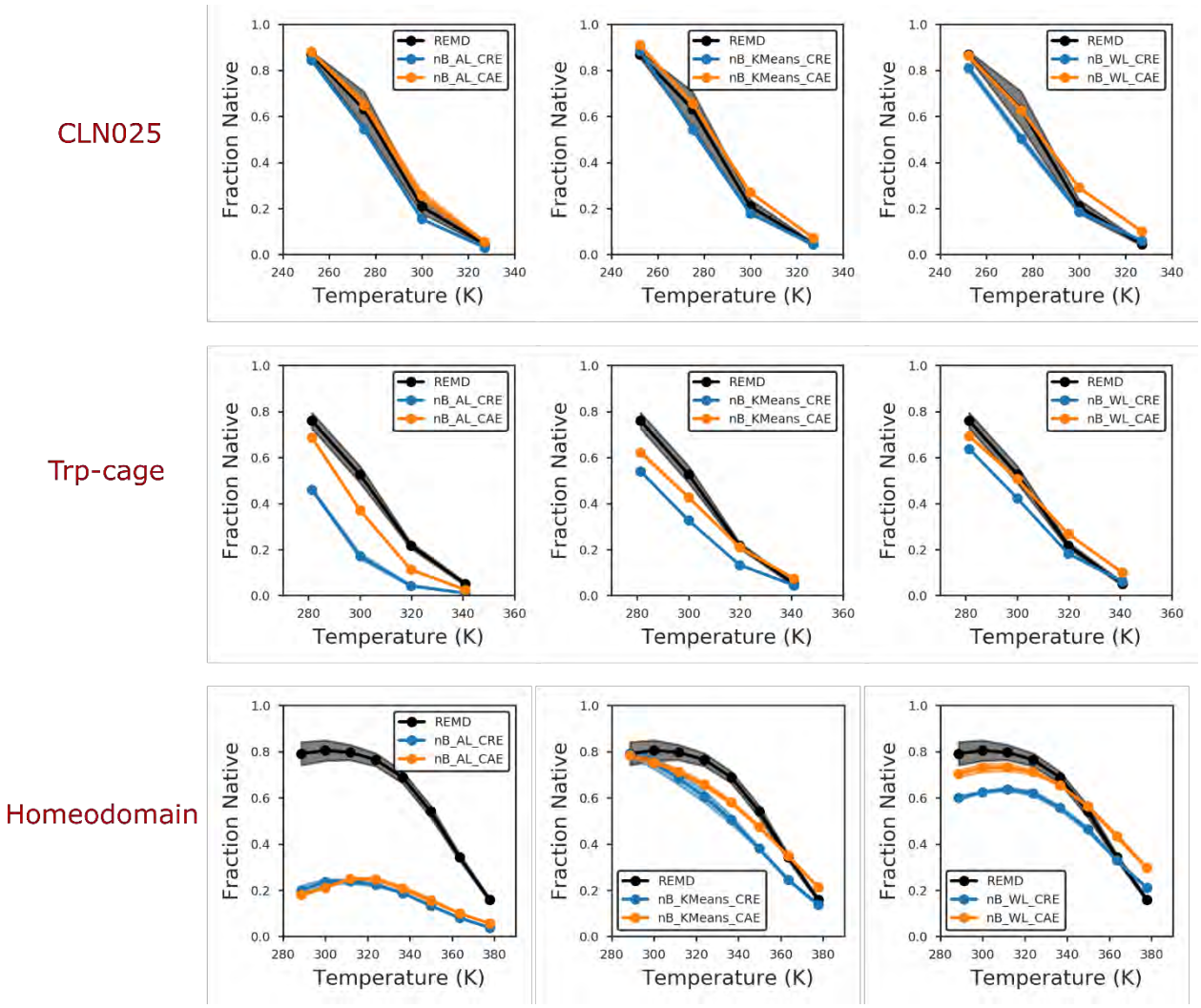


Figure 2-9 The melting curves obtained using standard REMD (black) and nB-RREMD simulations using different clustering methods are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). AL, KMeans, and WL indicate that the cluster representatives used to build the reservoir were obtained from Average-Linkage, KMeans, and Ward-Linkage clustering algorithms, respectively. CRE and CAE indicate that the cluster representative energy and the cluster average energy were used to build the reservoir (see text), respectively. The error bars indicate the half difference of the melting curves obtained from two nB-RREMD simulations – one starting from native conformation and the other starting from extended conformation, using the same set of reservoir structures. For some nB-RREMD simulations, the error bars are negligible and hence are not visible on the graphs.

For CLN025, for all three clustering methods, the melting curves obtained from nB-RREMD simulations are in good agreement with the standard REMD melting curve – the fraction of native structures at 300 K obtained from reservoirs built using CREs are 0.15, 0.18, and 0.18,

for AL, KMeans, and WL clustering methods, respectively. The corresponding reference value from standard REMD simulations is 0.21. Using CAEs instead of CREs results in only slightly more stable melting curves – the fraction of native structures at 300 K are 0.26, 0.27, and 0.29, for AL, KMeans, and WL clustering methods, respectively.

For Trp-cage, the nB-RREMD simulations using clusters obtained from WL perform the best followed by KMeans and AL clustering methods – the fraction of native structures at 300 K using CREs are 0.17, 0.33, and 0.42, for AL, KMeans, and WL clustering methods, respectively. These values are lower than the 0.53 fraction of native structures obtained from reference standard REMD simulations. Similar to CLN025, using CAEs improves the stability of the melting curves for nB-RREMD Trp-cage simulations, thereby, resulting in a better match between nB-RREMD simulations and standard REMD simulations – the fraction of native structures at 300 K using CAEs are 0.37, 0.43, and 0.51 for AL, KMeans, and WL, respectively, compared to the reference 0.53 fraction.

For Homeodomain, only KMeans and WL clustering methods reproduce melting curves that are in reasonable agreement with the standard REMD melting curves whereas AL results in mostly non-native ensembles even at 300 K – the fraction of native structures at 300 K using CREs are 0.24, 0.75, and 0.63, for AL, KMeans, and WL, respectively, compared to the reference 0.80 from standard REMD. For AL clustering method, the agreement between nB-RREMD data and standard REMD data does not improve even when CAEs are used – the fraction of native structures at 300 K is only 0.21 which is similar to the fraction of native structures obtained from reservoir using CREs. For KMeans, the fraction of native structures at 300 K is 0.75 with CAEs, while WL using CAEs results in a fraction of 0.73; both are in good agreement with the reference fraction of 0.80.

In all cases except for Homeodomain nB-RREMD simulations using AL, using CAEs results in slightly more stable melting curves compared to using CREs. Overall, irrespective of whether CREs or CAEs were used to build the reservoir, WL and KMeans clustering methods result in melting curves that are in good agreement with standard REMD melting curves for all three proteins while AL clustering method results in melting curves that are in reasonable agreement for only CLN025 and Trp-cage but not for Homeodomain.

2.5.2.4 Can nB-RREMD simulations reproduce the overall ensemble obtained from standard REMD simulations?

As stated above, good agreement between the melting curves obtained from nB-RREMD and reference standard REMD simulations indicates that both methods result in similar population of native structure at all temperatures. As discussed above, it is important to verify that the reservoir approach also reproduces non-native structures in the reference ensemble. Since the non-Boltzmann reservoir includes only one structure to represent each minimum, there is a greater risk that an important structure may be missed in the reservoir. Therefore, evaluating the populations of the entire ensemble can further validate the clustering methods used to select the structures, as well as any potential impact of the choice of representative energy on the final populations.

We performed clustering on the combined trajectories of nB-RREMD simulations and reference standard REMD simulations at a temperature close to the melting temperature (see **Methods**). The cluster populations at 275.1 K, 300.0 K, and 349.8 K, for CLN025, Trp-cage, and Homeodomain, respectively, from nB-RREMD and standard REMD simulations are shown in **Figure 2-10**.

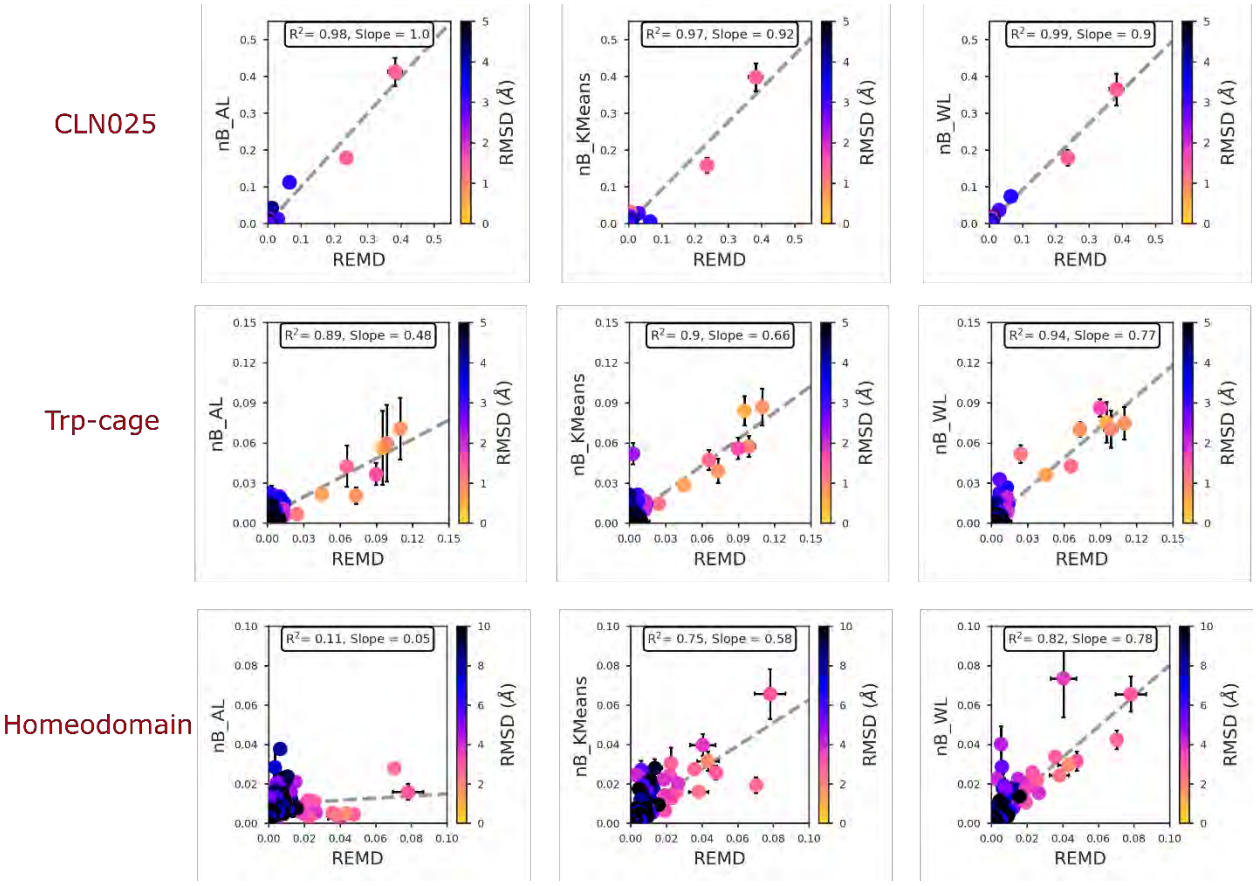


Figure 2-10 Cluster populations obtained using standard REMD (X-axis) and nB-RREMD (Y-axis) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom) at 275.1 K, 300.0 K, and 349.8 K, respectively. AL, KMeans, and WL indicate that the cluster representatives used to build the reservoir were obtained from Average-Linkage, KMeans, and Ward-Linkage clustering algorithms, respectively. The color of each point indicates the RMSD of the cluster representative to the native structure. The error bars on the X-axis indicate the half difference of cluster populations obtained from the two REMD runs – one starting from native conformation and the other starting from extended conformation. The error bars on the Y-axis indicate the standard deviation of cluster populations obtained from the two sets of nB-RREMD runs (4 simulations in total) – one starting from native conformation and the other starting from extended conformations, for both CRE and CAE reservoirs. Error bars for some clusters are not visible since they are smaller than the point size. The Pearson correlation coefficient between the cluster populations obtained from standard REMD and B-RREMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each protein.

For CLN025, similar to the results observed with melting curves, all three clustering methods result in ensembles that are in close agreement with standard REMD ensembles – the correlation of cluster populations is 0.98, 0.97, and 0.99, and the slope is 1.00, 0.92, and 0.90, for

AL, KMeans, and WL, respectively. This suggests that simple peptides may be relatively insensitive to the details of reservoir clustering method and energies.

For Trp-cage, all three clustering methods result in similar correlation of cluster populations – the correlation coefficient is 0.89, 0.90, and 0.94, for AL, KMeans, and WL, respectively. However, the three clustering methods result in different slopes – AL, Kmeans, and WL, result in a slope of 0.48, 0.66, and 0.77, respectively. The higher slopes using KMeans and WL indicates that using KMeans and WL result in not only the correct relative populations of different clusters but also in better absolute cluster populations compared to reference data. This is also reflected in the melting curves where KMeans and WL result in melting curves with a better match to the reference standard REMD melting curves than AL.

For Homeodomain, the correlation of cluster populations is 0.75 and 0.82, and the slope is 0.58 and 0.78, for KMeans and WL, respectively, indicating that the ensembles sampled by nB-RREMD simulations using clusters obtained from these two clustering methods are in reasonable agreement with reference ensembles. Once again, the performance with AL is much weaker, and the correlation of cluster populations is only 0.11 with a slope of 0.05.

KMeans results in the same most populated cluster as reference ensembles for all three proteins. WL results in the same most populated cluster as reference simulations for CLN025 but not for Trp-cage and Homeodomain. However, the difference in the populations sampled by reference REMD simulations and nB-RREMD using WL amounts to a free energy difference of <0.2 kcal.mol^{-1} for the most populated cluster from REMD simulations and vice-versa. AL results in the same most populated cluster as standard REMD simulations for CLN025 and Trp-cage, however, it favors non-native clusters for Homeodomain. Furthermore, for all three proteins, similar to B-RREMD simulations, multiple native-like clusters are observed. Overall, KMeans and

WL clustering algorithms, but not AL, tend to result in ensembles that are in good agreement with reference REMD data.

2.5.2.5 Effect of CREs and CAEs on nB-RREMD ensembles

Since the exchange criterion for nB-RREMD uses only the energy of the reservoir and not its temperature, the nB-RREMD results might be sensitive to the energies assigned to the reservoir structures. We know from **Figure 2-9** that using CAEs results in slightly more stable melting curves compared to using CREs. However, the small Y-error bars in **Figure 2-10** indicate that the nB-RREMD simulations are mostly insensitive to the energies used to build the reservoir. To confirm this, we combined the trajectories from the simulations nB-RREMD simulations using CREs and CAEs and clustered them together (see **Methods**). For all three proteins, for all three clustering methods, the correlation of cluster populations obtained from nB-RREMD using CREs vs. CAEs is >0.84 (**Figure 2-11**) indicating that nB-RREMD ensembles obtained with reservoirs using CREs are in close agreement with ensembles obtained using CAEs. Nevertheless, using CAEs tends to result in more population of native-like clusters which is indicated by a slope >1.04 for all proteins, for all clustering methods.

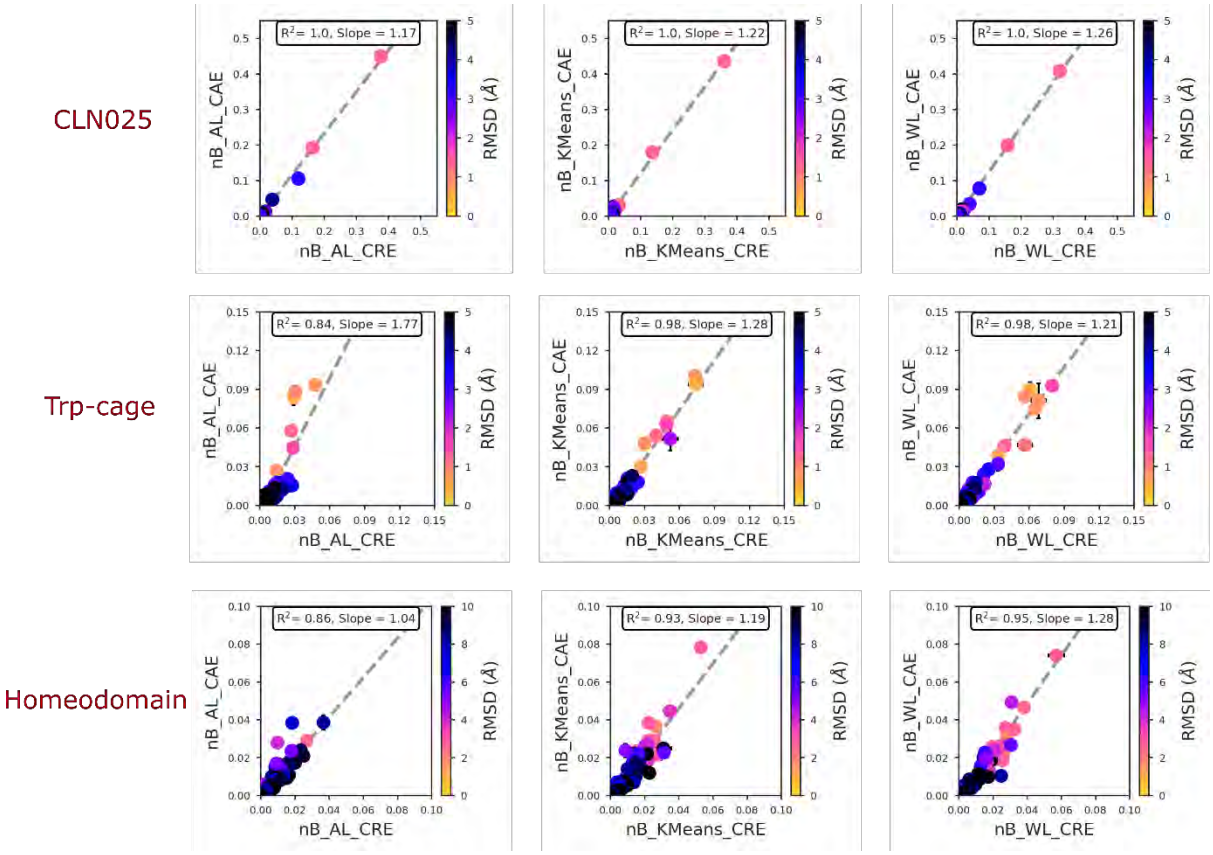


Figure 2-11 Cluster populations obtained from nB-RREMD simulations using reservoirs built using CREs (X-axis) and CAEs (Y-axis) with the three different clustering methods. AL, KMeans, and WL, represent Average-Linkage, KMeans, and Ward-Linkage clustering methods, respectively. The color of each point indicates the RMSD of the cluster representative to the native structure. The error bars on the X-axis and Y-axis indicate the standard deviation of cluster populations obtained from two independent nB-RREMD using CREs and two independent nB-RREMD runs using CAEs, respectively. Error bars for some clusters are not visible since they are smaller than the point size. The Pearson correlation coefficient between the cluster populations obtained from nB-RREMD simulations using CREs and CAEs and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each protein.

The above clustering protocols used to build the non-Boltzmann reservoirs may not be optimal, i.e., identifying the ideal number of clusters using intra-cluster RMSD variance might not result in the perfect number of clusters for a given protein for a given clustering method. However, we do note that the difference between the nB-RREMD simulation results using KMeans and WL clustering methods, and those using AL clustering method, are too big for these differences to stem only from the clustering protocols and not from the underlying clustering algorithm itself. Also,

our results are consistent with previous studies that have identified WL and KMeans clustering methods as the best for clustering MD trajectories.[78-79, 81] Moreover, the clustering methods and the protocols we outlined here are easy to implement on modern computers making construction of non-Boltzmann reservoirs fairly straightforward. More importantly, structures used to build non-Boltzmann reservoirs can be selected from many different sampling methods (physics-based and non-physics-based), thereby significantly increasing the scope of applicability of structure reservoirs in accelerating biomolecular simulations.

In the following section, we explore the ideal temperature at which the reservoir must be generated.

2.5.3 Ideal temperature to generate the reservoir

To explore the sensitivity to the temperature at which the reservoir is generated, we did MD simulations at different temperatures for each protein (see **Methods**). Then, similar to the analyses in **Figure 2-3** and **Table 2-1**, we calculated the correlation of cluster populations, fraction of unique clusters, and number of folding/unfolding event pairs at each temperature. For Trp-cage, the correlation of cluster populations and fraction of unique clusters are shown in **Figure 2-12**, and the number of folding/unfolding event pairs are shown in **Table 2-3**. The corresponding data for CLN025 and Homeodomain is shown in **Figure 2-13** and **Figure 2-14**, **Table 2-4** and **Table 2-5**, respectively.

Low temperature simulations would not be expected to be useful for reservoir generation, since this abandons the enhanced sampling at higher temperatures that underpins REMD. At 281.4 K, the correlation of cluster populations between runs is <0.0 for the first 0.7 μs. The fraction of unique clusters observed during the MD simulation starting from the native structure plateaus at this temperature (solid blue line in bottom panel of **Figure 2-12**) indicating that the simulation is

stuck in a local minimum. The correlation of cluster populations increases after 0.7 μs but is never >0.5. Moreover, the average number of folding/unfolding event pairs observed at this temperature are only 13 ± 5, indicating that 281.4 K is too low a temperature to build the reservoir.

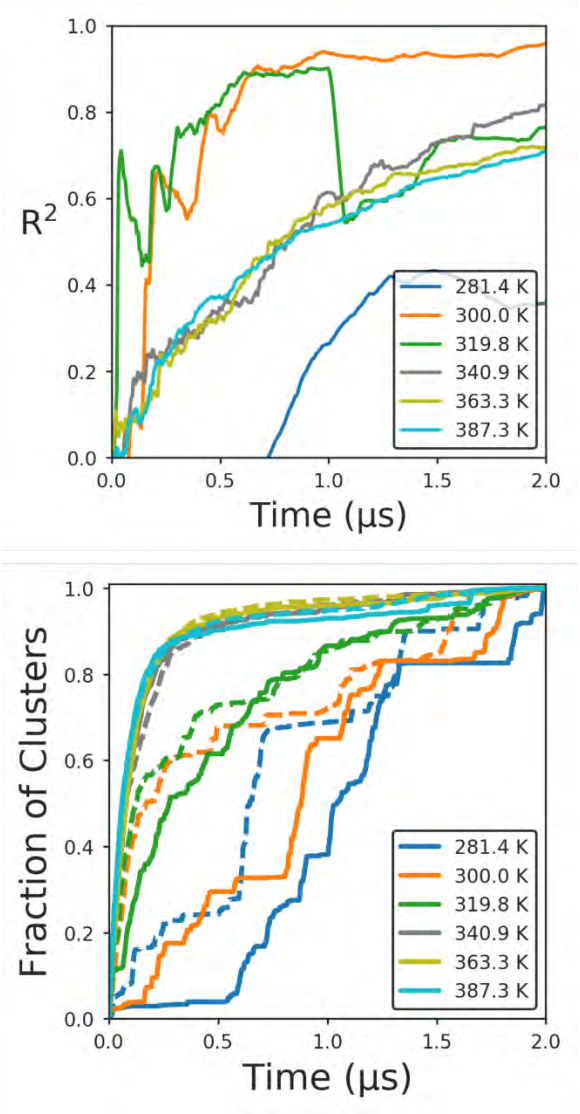


Figure 2-12 Identifying optimal temperatures for Trp-cage reservoir generation. The correlation of cluster populations (top) between the two independent simulations at different temperatures starting from different initial conformations is shown as a function of time. The fraction of unique clusters (bottom) observed during each independent simulation at different temperatures is shown as a function of time. The solid lines indicate simulations starting from native conformation and the dashed lines indicate simulations starting from extended conformation.

Table 2-3 Average Number of folding/unfolding pair events for Trp-cage at different temperatures.

Temperature	Number of Folding/Unfolding pair events
281.4 K	13 ± 5
300.0 K	80 ± 13
319.8 K	175 ± 10
340.9 K	175 ± 8
363.3 K	109 ± 5
387.3 K	54 ± 3

The error indicates the half difference between the two independent MD simulations at the same temperature.

At 300.0 K, the correlation of cluster populations is >0.9 after 1 μs of simulation, and the average number of folding/unfolding pair events is 80 ± 13, indicating that 300.0 K might be a suitable temperature to build the reservoir. However, the two independent simulations have only sampled 0.85 fraction of unique clusters after 1.5 μs indicating that long simulation would be needed.

At 319.8 K, around 0.85 fraction of unique clusters are observed within the first 1 μs, the average number of folding/unfolding event pairs are 175 ± 10. Surprisingly, the correlation of cluster populations is close to 0.9 within the first 0.5 μs but drops to 0.55 around 1 μs (possibly because one of the simulations becomes trapped in a local minimum), and then gradually rises again to 0.77.

At temperatures higher than 319.8 K, the rate at which unique clusters are explored, and the correlation of cluster populations are relatively insensitive to temperature. Around 0.9 fraction of unique clusters are observed within the first 0.5 μs and the final correlation of cluster populations is in the range of 0.7-0.8 suggesting that any of these temperatures should be suitable for building

a reservoir. However, the average number of folding/unfolding event pairs drops significantly as the temperature increases – at 340.9 K, 175 ± 8 number of folding/unfolding event pairs are observed followed by 109 ± 5 and 54 ± 3 at 363.3 K and 387.3 K, respectively. This likely reflects the non-Arrhenius behavior of folding at high temperatures.

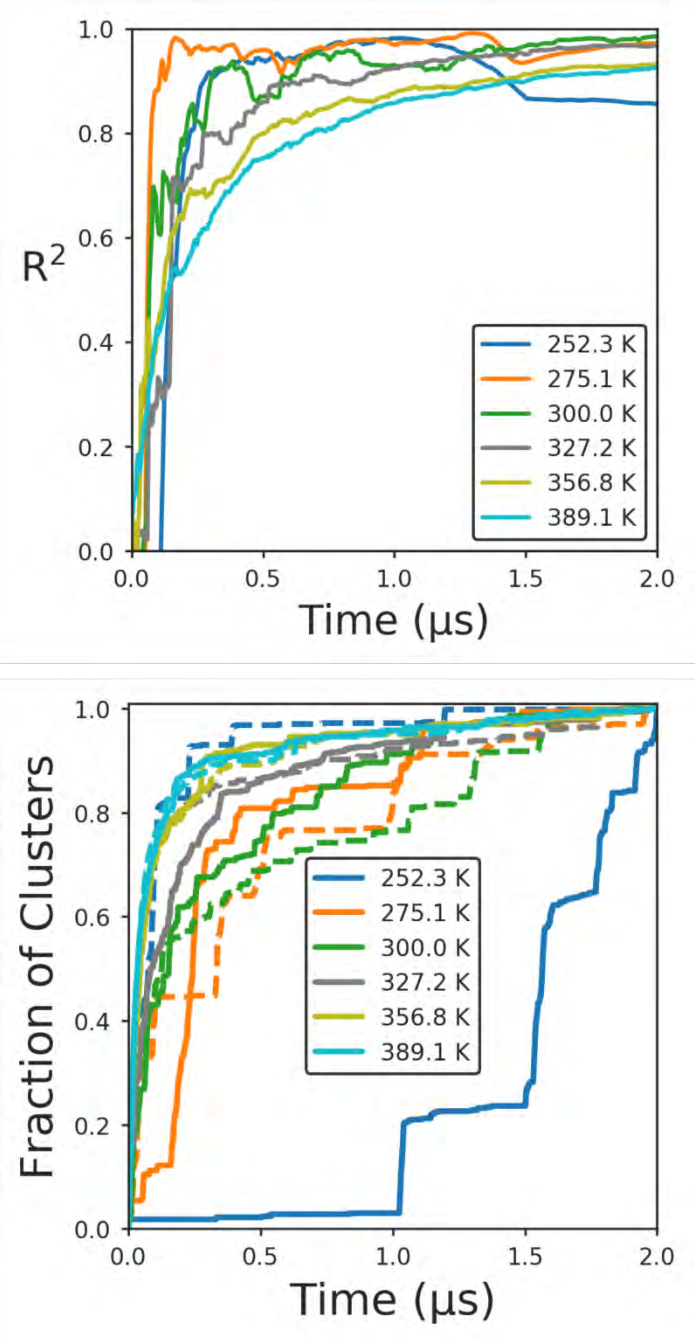


Figure 2-13 Identifying optimal temperatures for CLN025 reservoir generation. The correlation of cluster populations (top) between the two independent simulations at different temperatures starting from different initial conformations is shown as a function of time. The fraction of unique clusters (bottom) observed during each independent simulation at different temperatures is shown as a function of time. The solid lines indicate simulations starting from native conformation and the dashed lines indicate simulations starting from extended conformation.

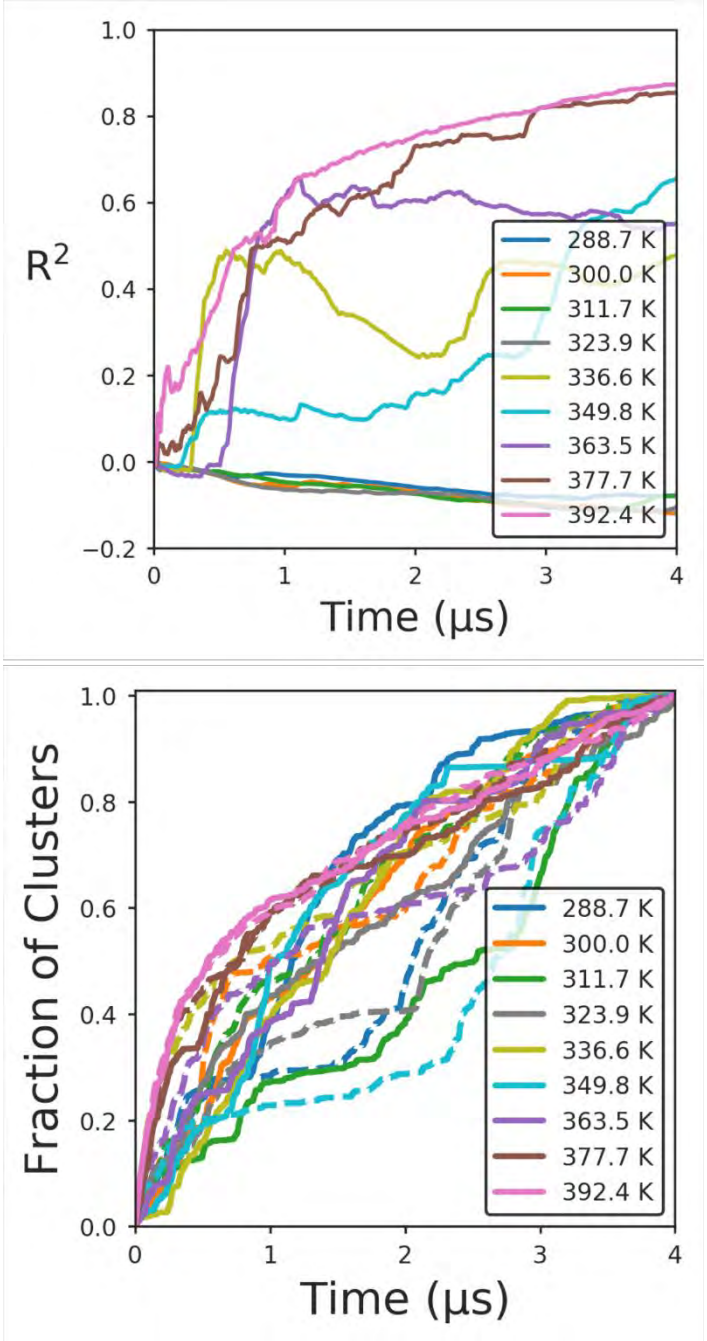


Figure 2-14 Identifying optimal temperatures for Homeodomain reservoir generation. The correlation of cluster populations (top) between the two independent simulations at different temperatures starting from different initial conformations is shown as a function of time. The fraction of unique clusters (bottom) observed during each independent simulation at different temperatures is shown as a function of time. The solid lines indicate simulations starting from native conformation and the dashed lines indicate simulations starting from extended conformation.

Table 2-4 Average Number of folding/unfolding pair events for CLN025 at different temperatures.

Temperature	Number of Folding/Unfolding pair events
252.3 K	5 ± 1
275.1 K	57 ± 1
300.0 K	223 ± 4
327.2 K	405 ± 2
356.8 K	338 ± 4
389.1 K	255 ± 15

The error indicates the half difference between the two independent MD simulations at the same temperature.

Table 2-5 Average Number of folding/unfolding pair events for Homeodomain at different temperatures.

Temperature	Number of Folding/Unfolding pair events
288.7 K	0 ± 0
300.0 K	0 ± 0
311.7 K	0 ± 0
323.9 K	2 ± 2
336.6 K	23 ± 17
349.8 K	53 ± 7
363.5 K	156 ± 8
377.8 K	278 ± 6
392.4 K	294 ± 22

The error indicates the half difference between the two independent MD simulations at the same temperature.

The above results indicate that using low temperatures is not ideal to generate structures for building the reservoir since the simulations tend to get trapped in local minima. On the other hand, using very high temperatures to generate reservoir structures is also detrimental since the native state is less accessible at these high temperatures. Therefore, the ideal temperature to build a reservoir should be somewhere in between. While we have tested many different temperatures to identify the ideal temperature for reservoir structure generation for each protein, in practice, and in our experience, it is sufficient to generate the structures at two temperatures and pick the one that has around 10-25% fraction of native structures (which is the range used in this current work).

If the native structure is not known, radius of gyration can be used to identify the range of sampled conformations. The ideal temperature in that case will be the one that samples conformations with low radius of gyration and also conformations with high radius of gyration multiple times, and quickly. Alternatively, REMD simulations with two high temperature replicas can be used to identify the ideal temperature to generate the reservoir structures for a given protein. It is beyond the scope of this work to generate Boltzmann-weighted reservoirs and non-Boltzmann reservoirs at every temperature and predict the effect of reservoir generation at different temperatures on the overall ensembles that can be obtained from RREMD.

2.5.4 How efficient are RREMD simulations compared to standard REMD simulations?

The sections above focused largely on the accuracy of the reservoir methods, and how the final results depend on choices made in reservoir generations. Next, we compare the convergence speeds of RREMD and standard REMD simulations and explore why RREMD simulations are much more efficient than standard REMD simulations.

In RREMD, extensive MD simulations are performed only at one high temperature, whereas standard REMD simulations require extensive simulations at all temperatures. Therefore, at least in theory, RREMD should be more efficient than standard REMD simulations.

To test this, we calculated the fraction of native structure as a function of time for standard REMD simulations, B-RREMD simulations, and nB-RREMD simulations. Since we performed two different simulations starting from very distinct initial conformations (native and fully extended) for each method, the rate at which the two different simulations for each method converge to the same amount of fraction of native structures at all temperatures serves as a good indicator for measuring the convergence rate of the different methods. The fraction of native structures observed in the simulations as a function of time for standard REMD simulations, B-RREMD simulations using B5000_nat reservoirs, nB-RREMD simulations built using KMeans clustering method and CAEs, are shown in **Figure 2-15** for all three proteins.

For CLN025, standard REMD requires more than 400 ns of simulation per replica for each independent simulation to converge to the same amount of fraction of native structures. In contrast, B-RREMD and nB-RREMD simulations converge in 150 ns and 50 ns, respectively, resulting in at least 3-8-fold increase in convergence speed, excluding the time required to generate reservoirs.

Similar to CLN025, Trp-cage simulations with standard REMD also take longer to converge compared to the RREMD methods – each replica has to be simulated for 400 ns for each independent simulation using standard REMD, whereas B-RREMD and nB-RREMD simulations converge in 200 ns and 40 ns, respectively, resulting in 2-10-fold increase in convergence speed, excluding the time required to generate reservoirs. These results are consistent with the 5-20-fold increase in convergence speed observed in previous studies for similar sized molecules.[32, 49, 63-66]

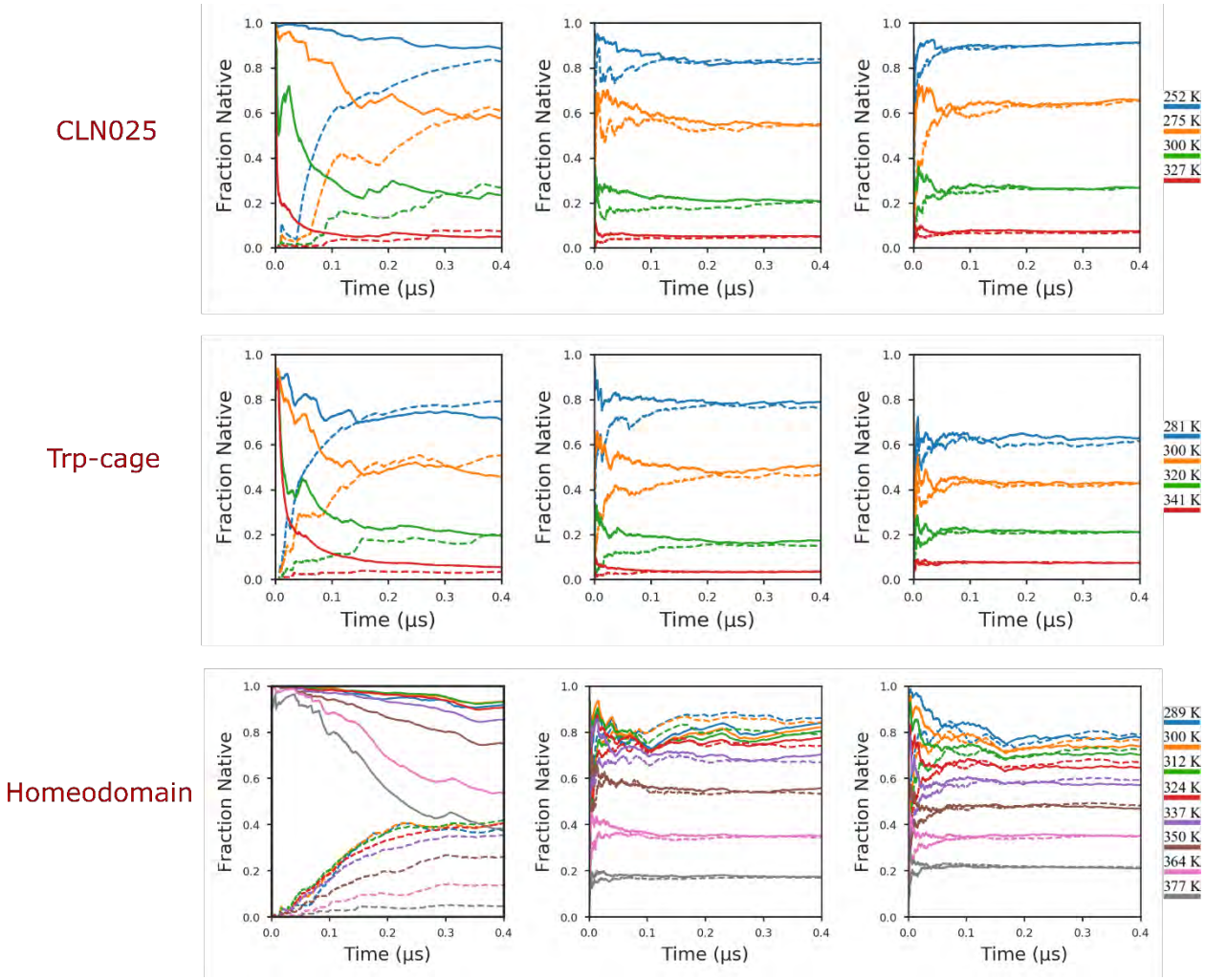


Figure 2-15 Fraction Native vs Time using standard REMD (left), B-RREMD (center), and nB-RREMD (right) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). The solid lines indicate the fraction of native structures for simulations starting from native conformation while the dashed lines indicate the fraction of native structures for simulations starting from extended conformations. B-RREMD data is from simulations using B5000_nat reservoirs. nB-RREMD data is from simulations using nB-KMeans_CAE reservoir. The standard REMD data is the same as the data shown in **Figure 2-2** except that only the first 0.4 μs are shown here for all three proteins.

However, as mentioned before, RREMD methods have only been used so far for small sized biomolecules (<310 atoms) and it is not known if the faster convergence rates will also be observed for medium to large sized biomolecules such as Homeodomain. The bottom row of **Figure 2-15** indicates that RREMD methods can significantly improve the convergence rates of

even medium sized proteins like Homeodomain. None of the replicas converge within the first 400 ns for Homeodomain. In fact, each replica has to be simulated for around 3-4 μs (see **Figure 2-2**) for each independent standard REMD simulation to converge to the same amount of fraction of native structure. On the other hand, both the B-RREMD and the nB-RREMD simulations converge within the first 200 ns of simulations for each replica at each temperature resulting in at least a 15-fold increase in convergence speed, excluding the time required to generate reservoirs. Moreover, even though the Homeodomain simulation required twice the number of replicas as the other two proteins, RREMD simulations still converge on a time scale similar to the other two proteins indicating that having more replicas does not slow down the convergence speed of RREMD simulations.

The above analysis indicates that RREMD simulations are significantly more efficient than standard REMD simulations. However, to effectively characterize the efficiency of RREMD compared to standard REMD, one must also include the time required to generate the reservoir. Our B-RREMD simulations indicate that reservoir generation simulations should be run for 200 ns, 1.3 μs, and 4 μs, to generate precise Boltzmann-weighted ensembles for CLN025, Trp-cage, and Homeodomain, respectively. Taking these times into account, B-RREMD simulations are 2-fold, 0.8-fold, and 6-fold faster than standard REMD simulations for CLN025, Trp-cage, and Homeodomain, respectively.

Nonetheless, the reservoir generation simulation lengths mentioned above are upper bounds. As shown in **Figure 2-3**, non-Boltzmann reservoirs will require significantly shorter reservoir generation simulations. Moreover, structures obtained from different enhanced sampling techniques such as accelerated MD, metadynamics, umbrella sampling, or even standard REMD

can be used in combination with structures from experiments or homology modeling, further reducing the time required to generate reservoir structures.

2.5.5 Why are Reservoir REMD simulations more efficient than standard REMD?

In RREMD simulations, in addition to scaling velocities between replicas, structures are also swapped between the highest replica temperature and the pre-sampled reservoir structures. If these MC steps to pre-sampled reservoir structures result in faster structural transitions compared to REMD, then these structural transitions should be reflected in the trajectories of each replica. To check this, we calculated the temperature and also the RMSD of the structure for each replica during standard REMD and RREMD simulations for each protein. In **Figure 2-16,** the temperature and RMSD of one replica (similar results are obtained for other replicas as well) during the first 400 ns of standard REMD and B-RREMD simulations for Trp-cage are shown.

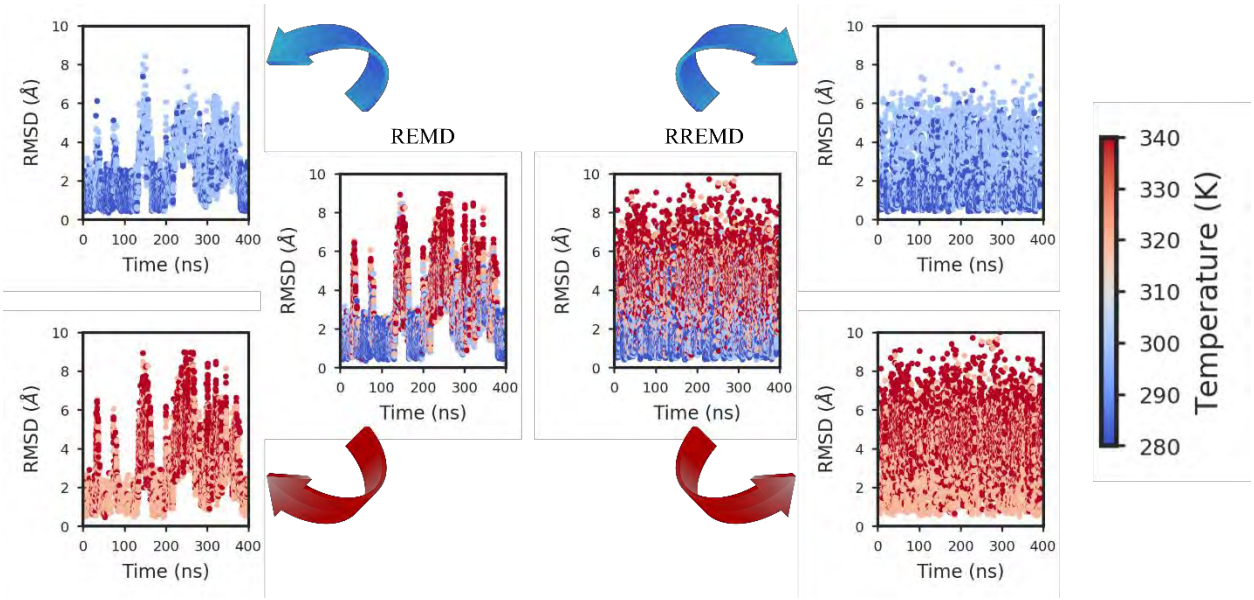


Figure 2-16 The distribution of temperature and RMSD of one replica during the first 400 ns of standard REMD (center left) and B-RREMD (center right) simulations for Trp-cage. The color of each point represents the temperature while the position of each point on the Y-axis represents the RMSD to native NMR structure. Blue colored points indicate temperatures less than 310 K and red colored points indicate temperatures greater than 310 K. For clarity, the central two images are

split into four images. The top left and bottom left images represent the REMD replica data when the temperature of the replica is less than 310 K and greater than 310 K, respectively. The top right and bottom right images represent the B-RREMD replica data when the temperature of the replica is less than 310 K and greater than 310 K, respectively.

The standard REMD replica (center left image in **Figure 2-16**) appears to be trapped in either low or high RMSD conformations with very few transitions between low/high RMSD conformations compared to RREMD. This is probably because the exploration of different structures during the MD part of REMD is still a slow process relative to structure swaps, even at high temperatures.

To illustrate this, we split the REMD replica data into two parts, in one part we only show the RMSD and temperature distributions when the temperature of the replica is less than 310 K (top left image in **Figure 2-16**) and in the other part, we only show the RMSD and temperature distributions when the temperature of the replica is greater than 310 K (bottom left image in **Figure 2-16**).

If the higher temperatures result in faster structural transitions, then the bottom left image in **Figure 2-16** (hot replicas) should have a greater spread of RMSD values compared to the top left image (colder replicas), however, the RMSD distributions in both images are similar indicating that temperature transitions can improve sampling only limitedly. Moreover, the changes in temperature occur much faster than structural transitions – this can be clearly seen in between 80 – 120 ns and also in between 160 – 200 ns where even though the replica visits both high and low temperatures, it only samples structures having an RMSD of <2.0 Å during these phases. It is not clear that swapping to higher temperature helps this replica to escape the basin in which it is trapped.

In contrast, a replica in RREMD simulation (center right image in **Figure 2-16**) samples low and high conformations at a significantly faster rate than standard REMD simulation. This can be seen clearly from the split RREMD replica data where high (>2.0 Å) RMSD structures are routinely sampled even at low temperatures (top right image in **Figure 2-16**) and low (<2.0 Å) RMSD structures are routinely sampled even at high temperatures (bottom right image in **Figure 2-16**). These fast-structural transitions are only possible because of MC steps using structure reservoirs.

Since the exploration of structures is done beforehand in RREMD, when a successful exchange with the reservoir occurs, a different region of the conformational landscape is immediately explored, thus, eliminating the lag time that will otherwise be required to traverse the barrier between the two structures. Since structures are swapped via MC, the MD part of REMD is no longer a limiting step. On the contrary, the MD part of REMD is mostly used to refine (locally explore) the accepted reservoir structure as it passes through different temperatures in the REMD ladder. Therefore, the more the number of successful exchanges with the reservoir the sooner the overall conformational landscape is explored by all replicas, thus, resulting in significantly faster convergence for RREMD simulations.

Likewise, exchanging less often with the reservoir might slow down the convergence rate of RREMD simulations. To illustrate this, we performed nB-RREMD simulations for Trp-cage, in which, instead of exchanging with the reservoir every 2 ps, we exchanged with the reservoir every 50 ps. When exchanges with the reservoir are attempted every 50 ps, the nB-RREMD simulations take 200 ns to converge compared to only 40 ns when exchanges are attempted every 2 ps (**Figure 2-17**).

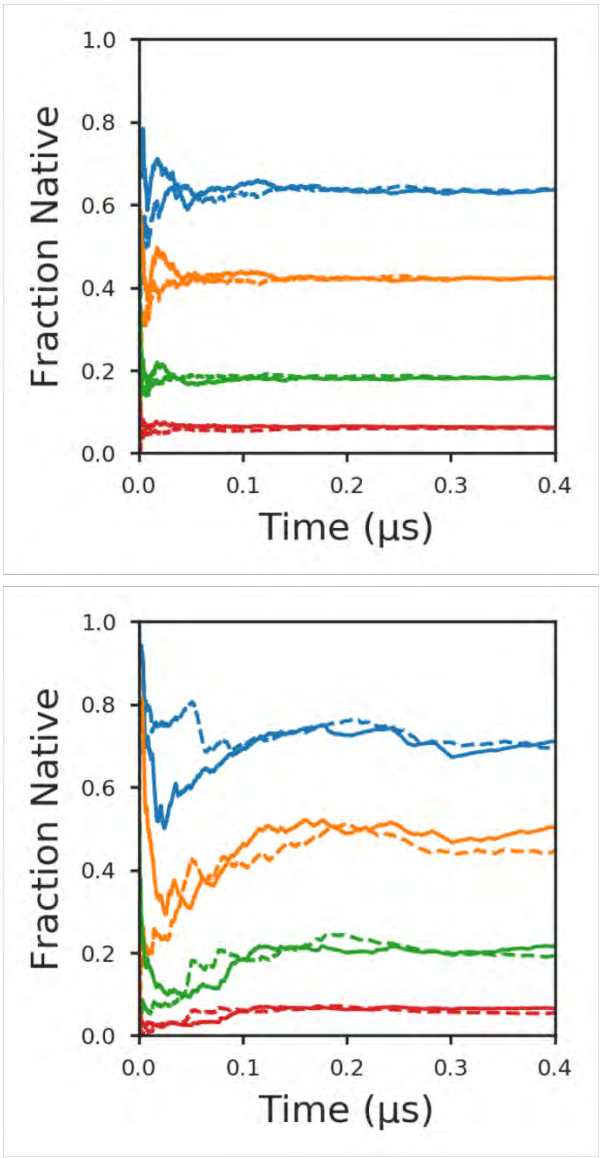


Figure 2-17 Fraction Native vs Time from nB-RREMD simulations with reservoirs built using Ward-Linkage and CREs, for Trp-cage. In top, exchanges with the reservoir are attempted every 2 ps. In bottom, exchanges with the reservoir are attempted every 50 ps.

Nevertheless, since RREMD simulations swap structures with a pre-sampled reservoir, as long as the rate of exchange with the reservoir is faster than the rate of conformational change at high temperatures, RREMD simulations will always be more efficient than standard REMD simulations.

Moreover, exchanging less frequently with the reservoir can, sometimes, also improve the accuracy of nB-RREMD simulations. **Figure 2-18** shows the effect of reservoir exchange frequency on the melting curves obtained using nB-RREMD simulations using structure reservoirs obtained using the three clustering methods.

When the exchanges with the reservoir are attempted every 50 ps, the melting curve obtained from nB-RREMD simulations using WL clustering method matches very closely with the standard REMD melting curve. Likewise, exchanging less frequently with reservoir also improves the melting curve obtained from nB-RREMD simulations using KMeans clustering method compared to exchanging every 2 ps. A similar trend is observed for nB-RREMD simulations using AL, but the improvement is only marginal. These improvements in the melting curves might be due to the possibility that the accepted structures have more time to explore and sample alternate rotamer conformations that might not be sampled if the exchanges are too rapid (since the reservoir structures include only a single rotamer example for each backbone). These results also indicate that exchanging less frequently with the reservoir can potentially fix slightly imperfect reservoirs (WL and KMeans) but not poorly built reservoirs (AL).

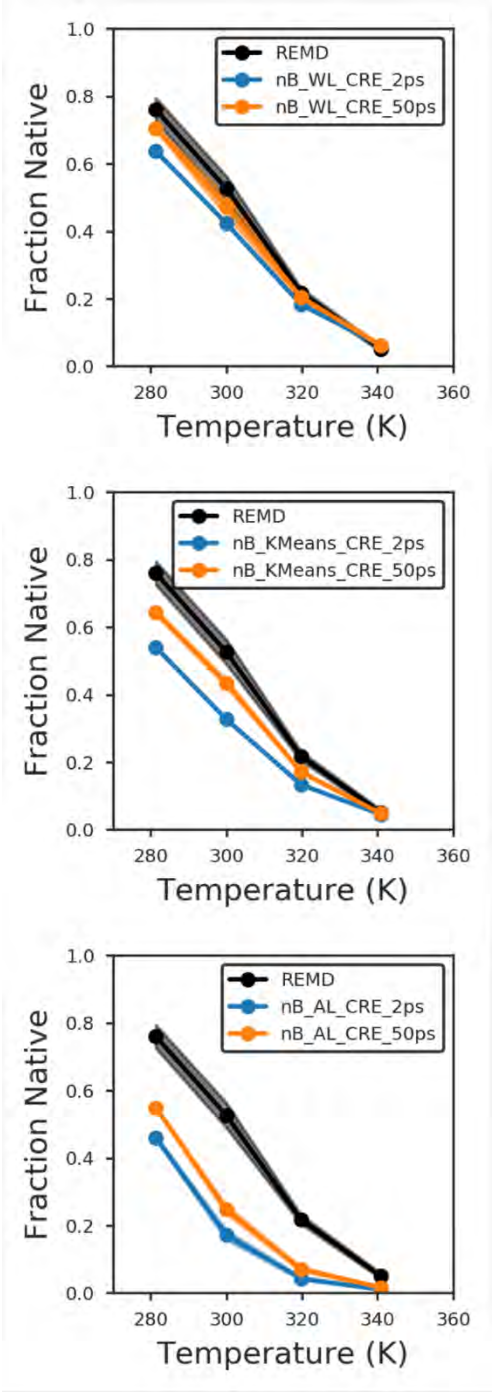


Figure 2-18 Effect of exchange frequency on the accuracy of nB-RREMD simulations of Trp-cage. The black curves indicate the melting curves obtained from standard REMD (same as **Figure 1**). The blue curves are the melting curves obtained when exchanges with the reservoir are attempted every 2 ps (same as **Figure 2-9**). The orange curves are the melting curves obtained when exchanges with the reservoir are attempted every 50 ps. CRE indicates that the energy of the representative structure for each cluster is used to build the reservoir. Error bars, if visible, indicate the half differences between two simulations starting from different initial conformations using the same set of reservoir structures.

2.6 Conclusions

Despite the significant increase in convergence speed, reservoir REMD methods have not been widely used due to the difficulties in building the reservoir and also due to the code not being available on the GPUs.

In this work, we have explored protocols to build reservoirs to accelerate the convergence of REMD simulations. Specifically, we presented protocols that use the correlation of cluster populations and the fraction of unique clusters observed as a function of time, and the number of folding/unfolding event pairs to identify how long the high temperature MD simulations should be run and how many structures should be used to build a Boltzmann-weighted reservoir. Our results indicate that a correlation of cluster populations >0.7, number of folding/unfolding event pairs >100 are good indicators of simulation time lengths to generate the structure reservoirs. Our results also show that the number of structures in a Boltzmann-weighted reservoir can be as low as 1000 to 5000.

We have also explored protocols to build a non-Boltzmann reservoir. Specifically, we explored which clustering methods are most suitable to pick structures for building a non-Boltzmann reservoir and observed that KMeans and WL are the best clustering methods to build a non-Boltzmann reservoir. Our results also show that the non-Boltzmann RREMD is only slightly sensitive to the energy used to build the reservoir. nB-RREMD simulations using CAEs tend to slightly favor more native-like structures compared to the corresponding simulations using CREs. Nonetheless, the overall ensembles obtained from using either CREs or CAEs are similar. We have also shown that the imperfections in a non-Boltzmann reservoir can be slightly fixed by exchanging less frequently with reservoir.

We have also shown protocols for choosing the ideal temperature to run the MD simulations to generate reservoir structures. Our results indicate that the reservoir temperature should not be too low, and it should not be too high either.

We have also shown that MC moves using structure reservoirs significantly accelerate convergence of REMD simulations. Our results indicate that RREMD is 2-15x faster (excluding the time required to generate the reservoir) than standard REMD, and the improvement in convergence speed becomes even more apparent for biomolecules which undergo slow structural transitions in standard MD/REMD simulations.

Finally, since the non-Boltzmann reservoir does not require a canonical ensemble of structures, structures obtained from physics-based enhanced sampling methods such as accelerated MD, metadynamics, umbrella sampling, standard REMD, etc. can be used in conjunction with non-physics-based methods such as homology modeling, further increasing the scope of applicability of RREMD simulations. Such non-Boltzmann reservoirs can be used to quickly test the accuracy of new force fields, and design novel peptides.

3 Optimizing protocols for building non-Boltzmann reservoirs

3.1 Abstract

In this chapter, the protocols presented in the previous chapter for building non-Boltzmann reservoirs are further optimized by (a) simplifying the protocol for finding the ideal number of clusters, and (b) attempting exchanges with the reservoir less frequently so that the accepted structures have more time to explore and sample alternate rotamer conformations. Using the optimized protocol, non-Boltzmann RREMD simulations result in melting curves that are in excellent agreement with standard REMD simulations for all three proteins listed in the previous chapter.

3.2 Introduction

In the simplest form of non-Boltzmann RREMD (nB-RREMD)[49], the reservoir is built by choosing one structure corresponding to different minima on the energy landscape. Building such a reservoir should, in principle, be much easier than generating structures for a Boltzmann-weighted reservoir since the different minima do not have to be precisely populated. Moreover, the greater flexibility in choosing different sampling methods to generate structures to build a non-Boltzmann reservoir makes nB-RREMD an attractive tool to enhance sampling. With this in view, in the previous chapter, a clustering protocol to select structures for building a non-Boltzmann reservoir was presented. The reservoirs built using these cluster representatives resulted in nB-RREMD melting curves that were in good agreement with standard REMD melting curves.

Nonetheless, some arbitrary choices were made during the reservoir building process. Specifically, in the previous chapter, the target number of clusters were adjusted so that the intra-cluster variance for each cluster was <0.5 $Å^2$, and the number of singleton clusters were <35% of the total number of clusters. Furthermore, from the resulting clusters, representative structures were picked from only those clusters that had at least 4 structures in them.

However, in the previous chapter, it was also shown that exchanging less frequently with the reservoir can potentially fix slightly imperfect reservoirs, indicating that the above arbitrary choices can be relaxed.

Therefore, in this chapter, the previously presented clustering protocols for building a non-Boltzmann reservoir are optimized. The results show that, by attempting exchanges less frequently with the reservoir, the ideal number of clusters for a given protein can be significantly reduced. With the new protocol, nB-RREMD simulations result in melting curves that are in excellent agreement with standard REMD melting curves for CLN025, Trp-cage, and Homeodomain.

3.3 Methods

Unless otherwise stated, the **Model systems, General details, System-specific details**, and **Cluster algorithms specific details** are the same as in the previous chapter and are not described here again to avoid redundancy. Only the methods for additional simulations and the protocols for building non-Boltzmann reservoirs are provided below.

In addition to the simulations described in the previous chapter, the following simulations were also performed:

(1) The reason for the shift in temperatures for the below simulations will become more apparent in **Chapter 4**.

(2) For Trp-cage, MD simulations starting from two different initial conformations (native and extended) were also performed at 319.8 K for 4 μs each, to generate structures for the reservoir. These MD simulations will be referred to as MD2_nat and MD2_ext, where "nat" and "ext" indicate that the simulations started from native and extended conformations, respectively.

(3) For Trp-cage, standard REMD simulations starting from two different initial conformations (native and extended) were also run with the replica temperatures set to 264.0 K, 281.4 K, 300.0 K, and 319.8 K, instead of the previously used 281.4 K, 300.0 K, 319.8 K, and 340.9 K temperatures. These standard REMD simulations will be referred to as REMD2_nat and REMD2_ext, where "nat" and "ext" indicate that the simulations started from native and extended conformations, respectively. These standard REMD simulations using a shifted temperature ladder were run for 6 μs per replica.

(4) To allow direct comparison to standard REMD simulations REMD2_nat and REMD2_ext, for Trp-cage reservoirs built using MD2_nat and MD2_ext trajectories (see below), nB-

RREMD simulations starting from two different initial conformations (native and extended) were also run for 400 ns per replica, with the replica temperatures set to 264.0 K, 281.4 K, 300.0 K, and 319.8 K.

3.3.1 Building non-Boltzmann reservoirs for the three proteins

Clustering was performed using Average-Linkage (AL), KMeans[83], and Ward-Linkage[84] (WL) algorithms. For each clustering method, 40000 structures were extracted from the combined MD trajectories at temperatures 327.2 K, 340.9 K, and 388.7 K, starting from two different initial conformations, totaling a time of 4 μs, 4 μs, and 8 μs for CLN025[67], Trp-cage[48], and Homeodomain[85], respectively. For CLN025 and Trp-cage, an equal time spacing of 100 ps was used to extract the structures. For Homeodomain, an equal time spacing of 200 ps was used to extract the structures. For each protein, for each clustering method, the entire backbone RMSD was used as the clustering metric and the target number of clusters was set to 200.

After clustering, the energy of each cluster representative (CRE) was calculated using the imin=5 flag in *sander* program in AMBER[56] using the same topology file and energy parameters (igb=8, gbsa=0/3) as used in MD and REMD, for each protein. The average energy of each cluster (CAE) was obtained by repeating the above step for all structures within a cluster and taking the average of the energies thus obtained. Finally, reservoirs were built using the *createreservoir* command in *cpptraj*[88] program in AMBER using a seed of 1.

For Trp-cage, an additional 30 reservoirs were also built. For these reservoirs, 20000 structures were extracted from the combined MD trajectories (MD2_nat and MD2_ext) at 319.8 K using an equal time spacing of 400 ps. Then, for each clustering method, these 20000 structures were clustered by setting the target number of clusters to values ranging from 100 to 1000 with

increments of 100, resulting in 10 different reservoirs. To distinguish between the 10 different reservoirs for each clustering method, the following naming convention was used – nB_(N)_CRE where "N" indicates the number of structures in the reservoir. After clustering, only CREs were calculated using the imin=5 flag in *sander* program as described above. Finally, reservoirs were built using the *createreservoir* command in *cpptraj* program in AMBER using a seed of 1.

3.3.2 Running nB-RREMD simulations

In all nB-RREMD simulations described in this chapter, except for the nB-RREMD simulations using nB_(N)_CRE reservoirs, exchanges with the reservoir were attempted every 50 ps. For nB_(N)_CRE reservoirs, two sets (each set containing two simulations – one starting from a native conformation and the other from an extended conformation) of nB-RREMD simulations were performed – one in which exchanges with the reservoir were attempted every 2 ps, and the other in which exchanges with the reservoir were attempted every 50 ps. For nB_(N)_CRE reservoirs built using WL, the nB-RREMD simulations (in which exchanges with the reservoir were attempted every 2 ps) were only performed for reservoirs having N<=500 structures.

3.3.3 Analysis

3.3.3.1 Melting curves

Except for REMD2_nat and REMD2_ext simulations, the melting curves were calculated in the same way as mentioned in the previous chapter. For REMD2_nat and REMD2_ext, the last 5 μs of data was used to calculate the fraction of native structures at each temperature.

3.4 Results and Discussions

In the previous chapter, it was shown that exchanging less frequently with the reservoir can potentially fix imperfect non-Boltzmann reservoirs due to the possibility that the accepted structures have more time to explore and sample alternate rotamer conformations that might not be sampled if the exchanges are too rapid. In this chapter, this idea is tested thoroughly.

3.4.1 Can exchanging less frequently with the reservoir fix imperfections in the reservoir?

To test if exchanging less frequently with the reservoir can fix imperfections in the reservoir, for each clustering method, we built 10 reservoirs by varying the target number of clusters (see **Methods**). (Varying the number of clusters in the reservoir might introduce imperfections in the reservoir and if these imperfections can be fixed, the protocol for building non-Boltzmann reservoirs can be simplified significantly.)

Then, for each of the above 10 reservoirs for each clustering method, we performed two sets of nB-RREMD simulations – one in which exchanges with the reservoir were attempted every 2 ps and the other in which exchanges with the reservoir were attempted every 50 ps. The melting curves obtained from standard REMD simulations and nB-RREMD simulations using these 10 reservoirs are shown in **Figure 3.1**, **Figure 3.2**, and **Figure 3.3**, for AL, KMeans, and WL, respectively. Note that these melting curves are different from the melting curves shown in the previous chapter due to the change in replica temperatures (see **Methods**).

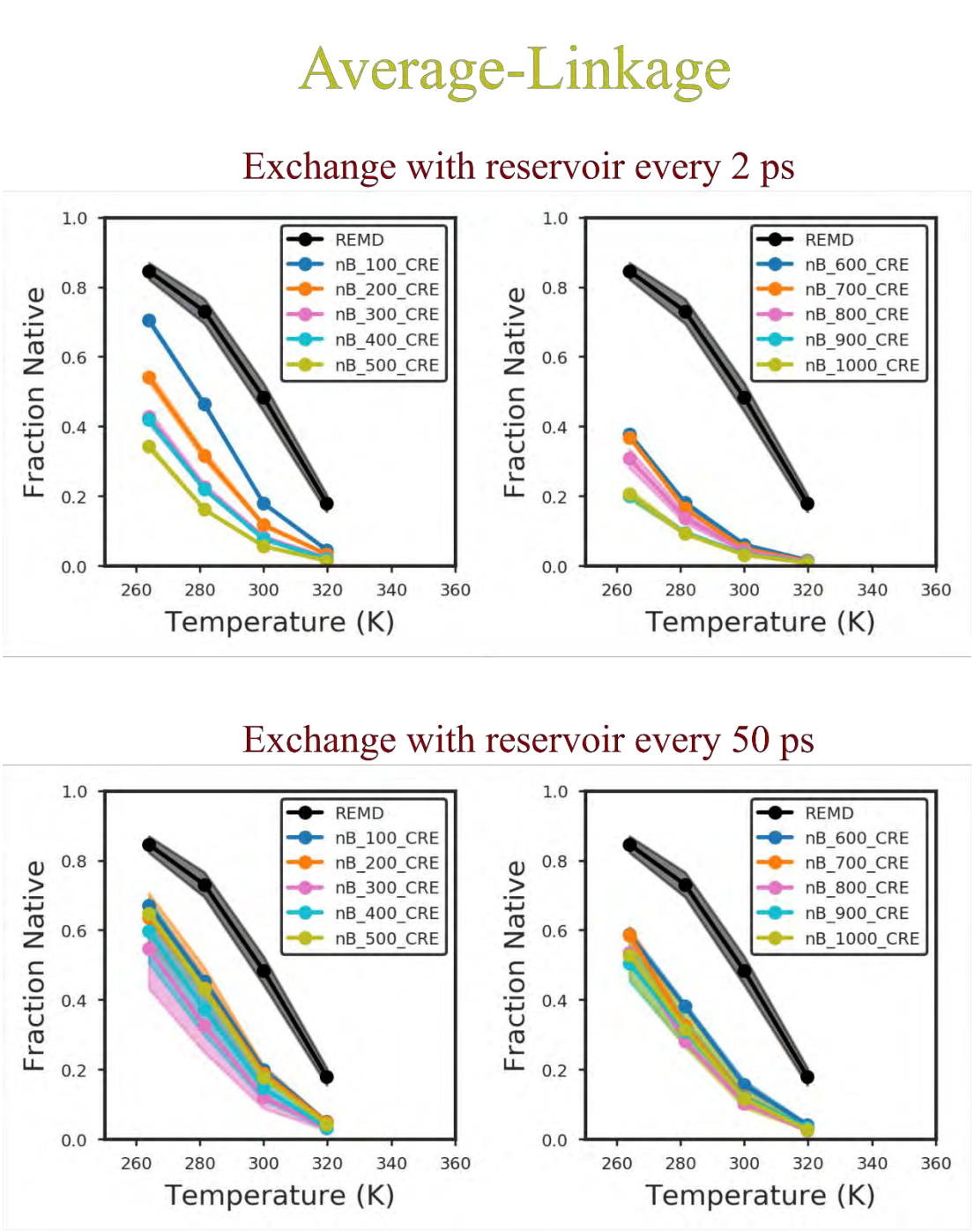


Figure 3-1 Melting curves obtained from standard REMD simulations (black) and nB-RREMD simulations using reservoirs nB_(N)_CRE built using AL. The number in each label (for both left and right plots) indicates the number of structures in the reservoir. The data in the top and bottom rows are from nB-RREMD simulations in which exchanges with the reservoir were attempted every 2 ps, and every 50 ps, respectively. The error bars (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations starting from different conformations.

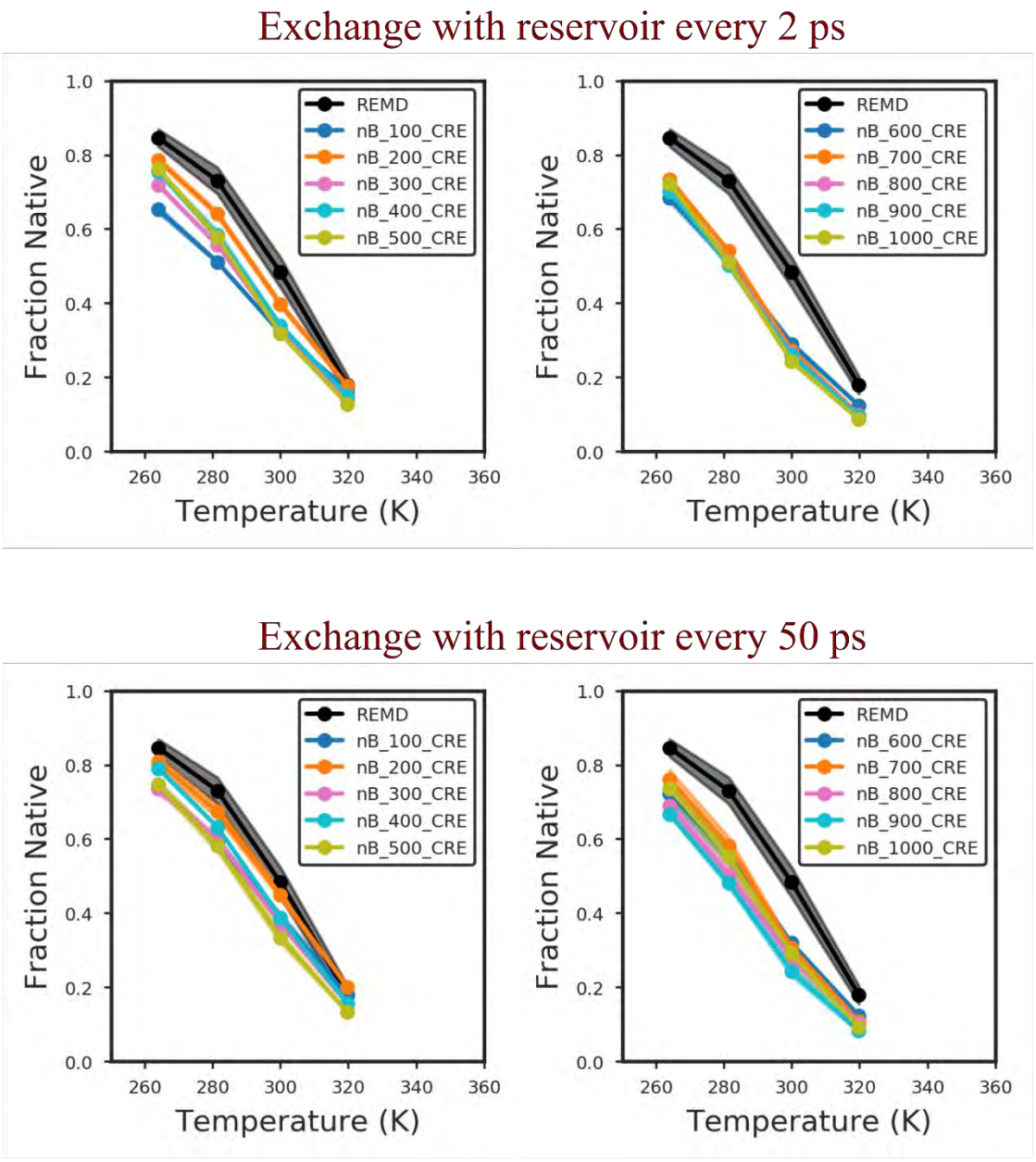


Figure 3-2 Melting curves obtained from standard REMD simulations (black) and nB-RREMD simulations using reservoirs nB_(N)_CRE built using KMeans. The number in each label (for both left and right plots) indicates the number of structures in the reservoir. The data in the top and bottom rows are from nB-RREMD simulations in which exchanges with the reservoir were attempted every 2 ps, and every 50 ps, respectively. The error bars (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations starting from different conformations.

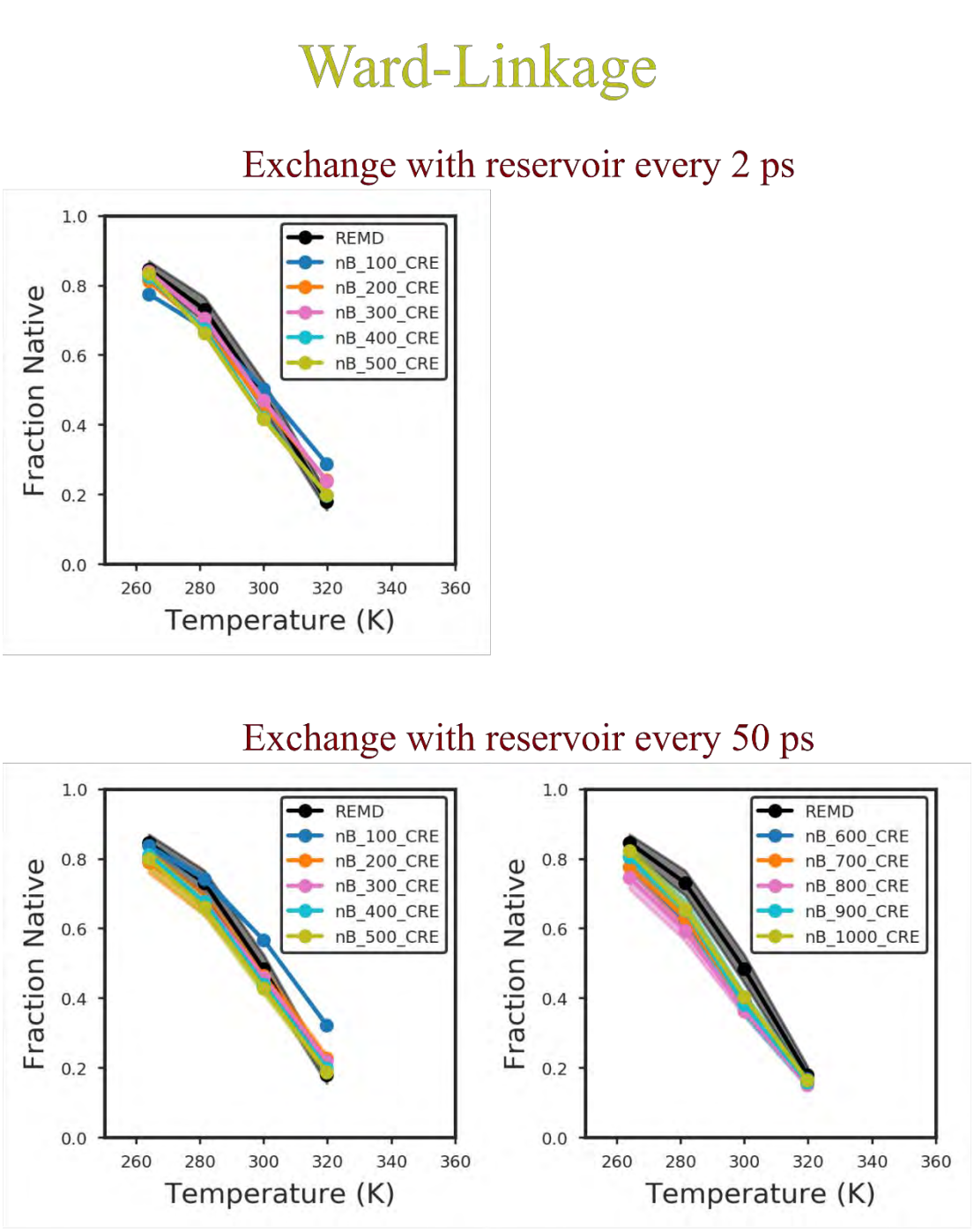


Figure 3-3 Melting curves obtained from standard REMD simulations (black) and nB-RREMD simulations using reservoirs nB_(N)_CRE built using WL. The number in each label (for both left and right plots) indicates the number of structures in the reservoir. The data in the top and bottom rows are from nB-RREMD simulations in which exchanges with the reservoir were attempted every 2 ps, and every 50 ps, respectively. For N>500, nB-RREMD simulations in which exchanges with the reservoir were attempted every 2 ps were not performed, and hence, not shown here. The error bars (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations starting from different conformations.

When AL is used to build the reservoirs, irrespective of the number of clusters used to build the reservoir, when exchanges with the reservoir are attempted every 2 ps, the melting curves obtained from nB-RREMD simulations poorly reproduce the standard REMD melting curves, consistent with the results from the previous chapter. Surprisingly, increasing the number of clusters worsens the match between nB-RREMD melting curves and standard REMD melting curves. This is could possibly be due to the increase in number of non-native clusters when more target number of clusters are chosen.

Exchanging with the reservoir less frequently slightly alleviates the problem but does not fix it completely. The melting temperature is still 35 K lower compared to standard REMD melting temperature.

When KMeans is used to build the reservoirs, when exchanges with the reservoir are attempted every 2 ps, the melting curves obtained from nB-RREMD simulations are close to the standard REMD melting curves, consistent with the results from the previous chapter. Changing the number of structures does not significantly alter the melting curves – the melting curve obtained with 300 structures in the reservoir is similar to the melting curves obtained with 400 to 1000 structures in the reservoir. Exchanging less frequently with the reservoir results in slightly better melting curves.

When WL is used to build the reservoirs, irrespective of how often exchanges are attempted with the reservoir, and irrespective of the number of structures used in the reservoir, the melting curves obtained from nB-RREMD simulations are in excellent agreement with the standard REMD melting curves, consistent with the results in the previous chapter.

More importantly, the above results also indicate that, for all three clustering methods, exchanging less frequently with the reservoir will result in similar melting curves irrespective of

the number of structures used to build the reservoir. This means that the clustering protocol outlined in the previous chapter can be greatly simplified without affecting the overall quality of nB-RREMD simulations as long as exchanges with the reservoir are attempted less frequently. It could also be that the clustering protocol outlined in the previous chapter was subdividing a cluster into multiple clusters. Here, by reducing the number of clusters, these subdivided clusters might be part of the same cluster.

To further test this idea, 6 (3*2) different non-Boltzmann reservoirs were built for CLN025, Trp-cage, and Homeodomain. These 6 reservoirs have the same nomenclature as used in the previous chapter. The only difference in the construction of these non-Boltzmann reservoirs compared to the non-Boltzmann reservoirs used in the previous chapter is that, for building these reservoirs, the target number of clusters was set to 200 for all three clustering methods. After clustering, all 200 clusters were used to build the non-Boltzmann reservoirs without excluding any of the clusters. Due to the significant reduction in number of clusters, the intra-cluster RMSD variance is >1.0 $Å^2$ for quite a few clusters and <10% of the clusters were singleton clusters.

Using these 200 structure reservoirs, nB-RREMD simulations were performed for all three proteins by attempting exchanges with the reservoir every 50 ps. The melting curves obtained from these nB-RREMD simulations, and from nB-RREMD simulations using the clustering protocol outlined in the previous chapter are shown in **Figure 3.4** along with melting curves obtained from standard REMD simulations.

For CLN025, when AL and KMeans clustering methods are used to build the reservoir, nB-RREMD simulations using the new protocol result in melting curves that match well with standard REMD melting curves. Using WL results in slightly but not significantly higher melting temperature.

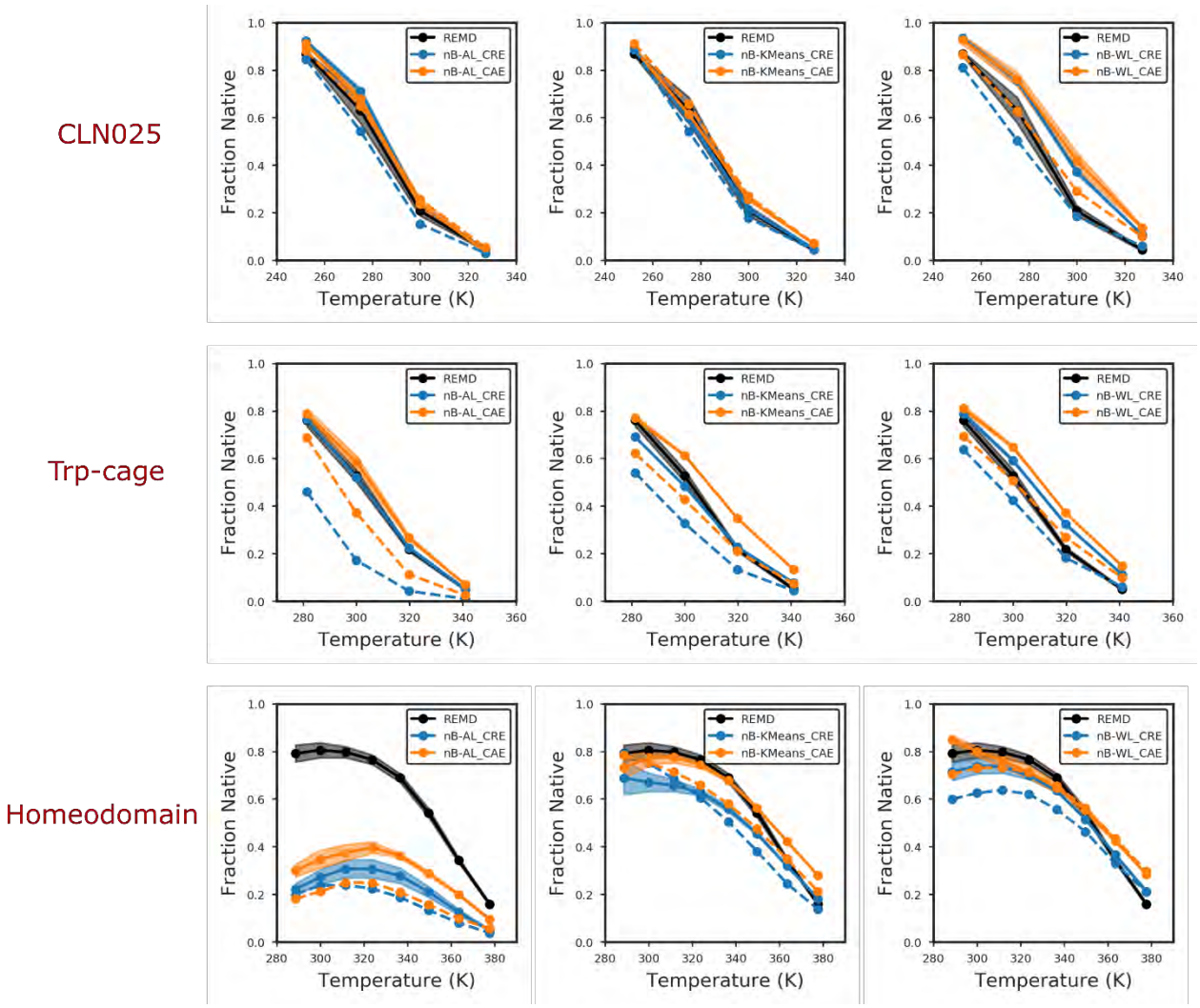


Figure 3-4 The melting curves obtained using standard REMD (black) and nB-RREMD simulations using different clustering methods are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). AL, KMeans, and WL indicate that the cluster representatives used to build the reservoir were obtained from Average-Linkage, KMeans, and Ward-Linkage clustering algorithms, respectively. CRE and CAE indicate that the cluster representative energy and the cluster average energy were used to build the reservoir (see text), respectively. The error bars indicate the half difference of the melting curves obtained from two nB-RREMD simulations – one starting from native conformation and the other starting from extended conformation, using the same set of reservoir structures. For some nB-RREMD simulations, the error bars are negligible and hence are not visible on the graphs. The solid lines represent the melting curves obtained with the new clustering protocol (see text) and the dashed lines indicate the melting curves obtained with the clustering protocol outlined in the previous chapter.

For Trp-cage, for all three clustering methods including AL, nB-RREMD simulations using the new protocol result in melting curves that are in reasonable agreement with the standard REMD melting curves. The improvement for AL nB-RREMD simulations could possibly be due to the shift to a structure with more favorable rotamers that would otherwise not be possible if exchanges with the reservoir are too frequent.

For Homeodomain, KMeans and WL nB-RREMD simulations using the new protocol result in melting curves that are in close agreement with standard REMD melting curves. More importantly, the shape of the melting curve is also better when exchanging less frequently with the reservoir. However, AL nB-RREMD simulations using the new protocol still result in non-native ensembles.

As observed in the previous chapter, for all three proteins, for all three clustering methods, using CAEs results in melting curves that are slightly but not significantly more stable compared to using CREs.

Overall, the above results indicate that, even with a simplistic clustering protocol, exchanging less frequently with the reservoir can result in melting curves that are close to the standard REMD melting curves when KMeans and WL are used.

Crucially, for all three proteins, the melting curves obtained from nB-RREMD simulations using the new protocol are better than or close to the melting curves obtained from nB-RREMD simulations using the old protocol. The only choices to be made are the number of clusters and the frequency of exchanges with the reservoir. While these choices are still arbitrary, they are fewer and more intuitive than iteratively choosing the number of clusters by varying the target number of clusters.

Moreover, Markov State Models (MSMs)[93-95], which provide information on the time required to transition from one macrostate to another macrostate, can be used to estimate the ideal frequency of exchanges with the reservoir. The frequency of exchanges in such an instance should be less than the fastest macrostate transitions so that transitions from one minimum to another minimum can be avoided during the REMD part of RREMD, since structures corresponding to each minimum are already present in the reservoir. nB-RREMD simulations can then be efficiently used to estimate the probability of observing the reservoir structures at the lowest temperature.

3.5 Conclusions

In this chapter, it was shown that exchanging less frequently with the reservoir can be used to greatly simplify the protocol for building non-Boltzmann reservoirs. The new protocol makes construction of reservoirs for nB-RREMD simulations fairly straightforward.

4 Using Structure Reservoirs to Predict the Accuracy of Force Fields and the Effects of Mutations via Thermal reweighting

4.1 Abstract

During the past two decades, many force fields have been developed to model biological processes. Testing the accuracy of each of these force fields requires precise sampling to validate against experimental data and also to compare against previous force fields. In the previous two chapters, we showed how reservoirs built from extensive high temperature MD simulations can be used to quickly obtain precise ensembles. While this is exciting, running extensive reservoir generation simulations for each force field will be computationally expensive. Therefore, in this chapter, we explore the applicability of nB-RREMD method for predicting the accuracy of new force fields without generating structures using the new force field. Our results show that structures generated from a reference Hamiltonian can be used to predict the accuracy of a different Hamiltonian indicating that nB-RREMD can be used to accelerate the testing of new force field parameters. We also extend this methodology to predict the effects of mutations on peptide stability without generating structures for different mutants. We use conformations obtained from wild-type simulations to predict the stability of mutants.

4.2 Introduction

An accurate description of forces is crucial to understand the energetics of various biological processes using Molecular Dynamics (MD) simulations. It is also essential for computational design of peptides/proteins[96]. Several force fields have been developed in the past to improve the accuracy of biomolecular simulations[5, 10, 13, 60-61]. Due to these improvements, currently, MD simulations are routinely used to fold proteins[21, 23], measure protein-ligand binding affinities[16], study effects of mutations[97], etc.

Despite the above successes, considerable manual and computational efforts are still dedicated to further improve the quality of force fields. These efforts often take around 4-5 years and involve training the force field parameters (usually against quantum mechanical calculations) followed by testing the new force field parameters. While the training part is different for different force fields, the testing part is similar and involves running MD simulations on a wide variety of test molecules and comparing the results to experimental observables.

However, precisely validating the accuracy of a new force field against experimental data requires extensive sampling to ensure that the error bars obtained from a given simulation are smaller than the error within the force field. Such extensive sampling also allows us to compare the accuracy of the new force field against previous force fields. While enhanced sampling methods such as Replica Exchange MD (REMD)[25-26], accelerated MD[28], and others[27, 32, 34-37, 39, 41, 49] can be used to obtain precise ensembles, running these simulations for every new force field requires significant computational resources.

Therefore, techniques such as one-step perturbation[98-103] have been developed to predict the accuracy of new force fields. In one-step perturbation, ensemble averages of the new force field are predicted by using time series data from an old reference force field, via Hamiltonian

reweighting[103]. Since the method uses previously existing data, significant computational resources can be saved, and the accuracy of new force fields can be predicted rapidly. One-step perturbation has been successfully used in development of GROMOS force fields[103], besides other applications[104-107].

While this is promising, since the structures sampled by the old reference force field cannot be adjusted during one-step perturbation, only changes to the soft degrees of freedom can be predicted using one-step perturbation. Moreover, previous studies[100] have shown that the accuracy of one-step perturbation is determined by the degree of overlap between the ensembles sampled by the new force field and the reference force field – the higher the overlap the better the accuracy. Since different force fields sample different conformations, it is difficult to ensure high degree of overlap between the ensembles of the two force fields. Using soft-core potentials[100] can ensure that the reference force field samples all the relevant conformations. However, applying soft-core potentials on all the protein atoms might result in highly unphysical conformations. Due to this limitation, the one-step perturbation technique was used only during the training phase of GROMOS force fields, where the ensembles of dipeptides were reweighted, but not during the testing phase[103]. The testing phase still required MD simulations on a wide variety of proteins. Therefore, methods that can quickly predict the accuracy of force fields are still desired.

In the previous chapters, we have shown that, by using the same force field for both high temperature reservoir generation simulations and non-Boltzmann-Reservoir REMD (nB-RREMD)[49] simulations, precise ensembles can be quickly obtained at all temperatures. However, since the non-Boltzmann reservoirs do not need canonical ensembles, in theory, the reservoir structures can be generated using any sampling method or as an extension, any force field. Since simulations of test molecules using old force fields are readily available, non-Boltzmann reservoirs

can be easily built using the protocols outlined in the previous chapters. By coupling these non-Boltzmann reservoirs built using old force fields to the REMD simulations using the new force fields, the basins sampled in the reservoir can be locally explored/refined using the new force field, thereby allowing rapid prediction of the accuracy of the new force field (see **Figure 4-1**).

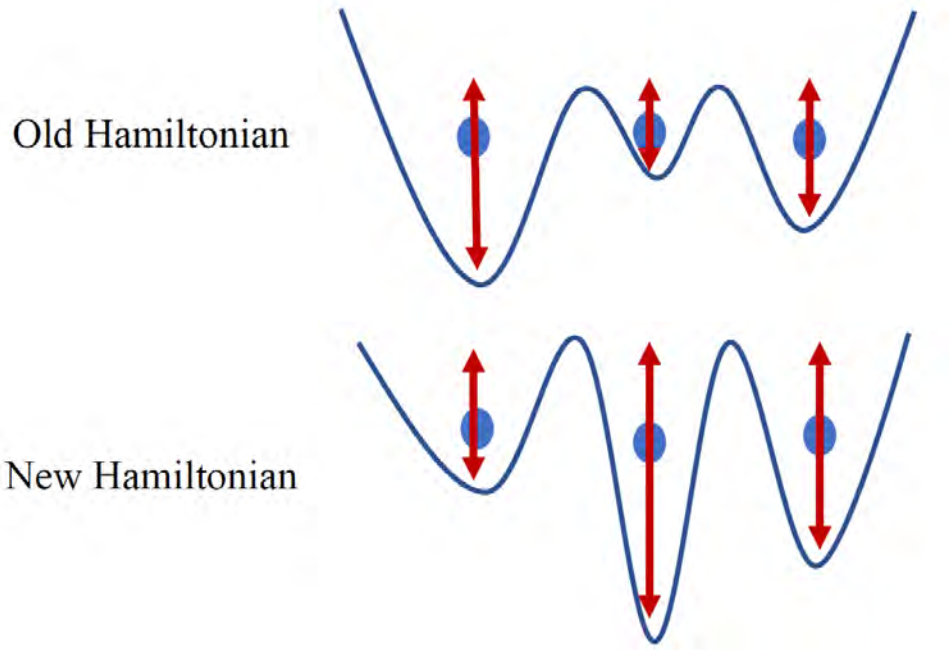


Figure 4-1 Illustration of Hamiltonian switching using nB-RREMD. Hamiltonian switching using nB-RREMD assumes that the old and new Hamiltonians sample the same basins but with different weights. The structure reservoirs corresponding to each basin are shown as blue circles. The red arrows indicate that the weights of the basins are different for different Hamiltonians. Besides changing the Hamiltonian, mutations can also be introduced into the reservoir. Since the relevant conformations are pre-sampled, using nB-RREMD can predict the favorability of a mutation in a given conformation, thereby facilitating rapid exploration of stabilizing/destabilizing mutations.

The same idea can also be extended to study the effects of mutations. Currently, alchemical transformations[97, 108], which require extensive sampling of the mutant and the wild type, are used to quantify the effects of mutations on protein stability. However, if all the relevant conformations are pre-sampled (as it is in RREMD), introducing mutations into these structures followed by thermal reweighting should predict the effect of mutation on each of the pre-sampled

conformations. The conformations in which the mutation is favorable will be sampled more at the low temperatures. If more native-like conformations are sampled with the mutation compared to the wild-type, then the mutation can be considered to be stabilizing. Likewise, if non-native conformations are sampled with a higher probability at the low temperatures, then the mutation can be considered destabilizing. Since introducing mutations into a few hundred structures takes only a few minutes, and since nB-RREMD simulations converge quickly, the effects of mutations on protein stability can be rapidly studied. Moreover, the effects of the mutation can also be quantified using the converged ensembles obtained from nB-RREMD simulations at different temperatures.

To test these ideas, we generated non-Boltzmann reservoirs by running high temperature MD simulations. Then, we perturbed the structures by using a different force field during the nB-RREMD simulation or by introducing mutations into the reservoir structures. Our results show that nB-RREMD simulations can be used to predict the accuracy of new force fields remarkably. Furthermore, we also show that the method can be used to accurately predict the effects of mutations on protein stability.

4.3 Methods

4.3.1 Model systems

The Trp-cage miniprotein (PDB ID: 1L2Y, 20 residues)[48] was used for this study since we can generate precise reference data using standard methods for this protein, and under conditions that are relevant to experiments.

4.3.2 Force fields and solvent models used for switching Hamiltonians

Four different Hamiltonians were considered: (1) ff14SBonlysc[10]/GB-Neck2[8] (referred to as H1 in this study), (2) ff99SB-ILDN[109]/GB-Neck2 (referred to as H2 in this study), (3) fffb15[110]/GB-Neck2 (referred to as H3 in this study), and (4) ff99SB[5]/GB-OBC[4] (referred to as H4 in this study). These four Hamiltonians were chosen to represent different perturbations. H1 and H2 differ in side-chain dihedral parameters for amino acids ILE, LEU, ASP, GLN. H1 and H3 have different bonds, angles, and dihedral parameters. H1 and H4 have different side chain dihedrals and also different implicit solvent model.

4.3.2.1 General details

All structures were built via the LEaP module of AmberTools in the AMBER 18 package[56]. For each Hamiltonian, two initial conformations were built – (1) Native conformation, for which the first NMR model was used, and (2) Extended conformation, in which φ, ψ angles for all residues except proline were set to 180°, proline residues were set to φ=-61.5°, ψ=-176.6°. mbondi2[4], and mbondi3[8] radii sets were used for all simulations using GB-OBC implicit solvent model, and GB-Neck2 implicit solvent model, respectively. No cutoff was used for calculation of non-bonded interactions. Langevin thermostat with a collision frequency of 1 ps^{-1} was used for all simulations. SHAKE[44] was performed on all bonds including hydrogen with the AMBER default tolerance of 0.00001 Å.

4.3.2.2 Minimization and Equilibration

A time step of 1 fs was used for all MD simulations during equilibration. With 10 $kcal.mol^{-1}.Å^{-2}$ positional restraints on all heavy atoms, the structures built using LEaP were minimized for 1000 cycles using steepest descent and then heated from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. Then, with 10 $kcal.mol^{-1}.Å^{-2}$ positional restraints on only backbone heavy

atoms, the structures were again minimized for 1000 cycles using steepest descent and then heated again from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. This was followed by 500 ps of MD at 300 K with 1 kcal.mol^{-1}.Å^{-2} positional restraints on backbone heavy atoms and then another 500 ps of MD at 300 K with 0.1 kcal.mol^{-1}.Å^{-2} positional restraints on backbone heavy atoms. Finally, 5 ns of unrestrained MD was performed at 300 K.

4.3.2.3 Molecular Dynamics simulations

For each Hamiltonian, MD simulations were performed starting from both native and extended conformations at 319.8 K, for 4 μs each, to generate reservoir structures. Chirality constraints and *trans*-peptide ω constraints obtained using *makeCHIR_RST* program in AMBER were used at all temperatures to prevent chirality inversions and peptide bond flips. A time step of 2 fs was used for all simulations. Coordinates were saved every 20 ps.

4.3.2.4 Replica Exchange Molecular Dynamics

For each protein, REMD simulations were performed starting from both native and extended conformations. Chirality constraints and *trans*-peptide ω constraints were used at all temperatures to prevent chirality inversions and peptide bond flips. A short 50 ps MD simulation using a time step of 1 fs was performed at each target temperature to briefly equilibrate each replica; thereafter the time step as 2 fs. Coordinates were saved every 20 ps. Exchanges between replicas were attempted every 1 ps for all simulations. 4 replicas were used, and the simulations were performed for 6 μs per replica. The replica temperatures were 264.0 K, 281.4 K, 300.0 K, 319.8 K, and were chosen such that the probability of exchange between replicas is close to 30%.

4.3.2.5 Non-Boltzmann Reservoir Replica Exchange Molecular Dynamics

nB-RREMD simulations also used the same procedure as REMD simulations, however, in addition to the exchanges between replicas, exchanges with the reservoir were also attempted

every 50 ps, unless otherwise noted. Since the velocities were not saved during the high temperature MD simulations used to build the reservoir, velocities for structures obtained through exchange with the reservoir were assigned by evaluating the forces during the subsequent MD step. The simulations were performed for 400 ns per replica. The replica temperatures were the same as REMD simulations.

4.3.3 Mutations considered in this study

Two stabilizing and two destabilizing mutations were considered for this study. Tc5b-W6F[48] and Tc5b-S14A[111] mutations were found to destabilize wild-type Tc5b variant of Trp-cage, whereas Tc5b-S13A[111] and Tc5b-K8A[111] were found to stabilize wild-type Tc5b variant. For each mutant, two different initial conformations were built. (1) Native conformation, for which the first NMR model was used. The native structure was mutated using the *Chimera*[112] software package and choosing the default options to result in "native" mutant structure. (2) Extended conformation, in which φ, ψ angles for all residues except proline were set to 180°, proline residues were set to φ=-61.5°, ψ=-176.6°. The mutant sequence was given as input to LEaP to build the extended conformation.

For each mutation, MD simulations, REMD, and nB-RREMD simulations were performed in the same way as described above. The only difference is that REMD simulations for the mutants were performed for 4 μs, instead of 6 μs since the simulations converged sooner (see **Results and Discussions**). For all simulations of mutants, ff14SBonlysc[10]/GB-Neck2[8] Hamiltonian was used.

4.3.4 Building non-Boltzmann Reservoirs

4.3.4.1 Extracting structures from the MD simulations of each Hamiltonian and each mutation

For each Hamiltonian, and for each mutant, the following protocol was used to extract structures for building non-Boltzmann reservoirs.

40000 structures were extracted from the combined MD trajectories at 319.8 K starting from two different initial conformations, totaling a time of 8 μs, using an equal time spacing of 200 ps. The entire backbone RMSD was used as the clustering metric. The target number of clusters were set to 200. Ward-Linkage (WL)[84] algorithm was used for clustering. The detailed clustering protocol is given below.

We used a combination of *cpptraj* and python modules to implement WL since neither package had all the capabilities required to do WL clustering on AMBER MD trajectories. The pairwise RMSDs between all 40000 structures were obtained using *cpptraj* and saved externally. Then, these pairwise RMSDs were used to perform WL clustering using the *scipy.cluster.hierarchy*[89-90] module in python. After clustering, clusters were sorted in descending order of cluster size and the structure numbers corresponding to each cluster were saved externally, separated by comma. Then, the structures corresponding to each cluster were extracted using the *onlyframes* keyword in *cpptraj* and these cluster trajectories were saved externally. Finally, cluster representatives were extracted from these saved cluster trajectories using AL clustering method by setting the target number of clusters to 1 in *cpptraj*. These representatives will, by default, correspond to the structure that has the lowest cumulative distance to every other structure in that cluster. Only the representative structure of each cluster was used for further processing.

4.3.4.2 Building reservoirs for Hamiltonian switching

For Hamiltonian switching, 6 (3 in each direction) different reservoirs were built to model the switch from H1 to H2, H1 to H3, H1 to H4, and H2 to H1, H3 to H1, H4 to H1. Since different Hamiltonians have different overall energy, for each switch, the structures obtained from one Hamiltonian were briefly equilibrated using the other Hamiltonian to make the energy of each reservoir structure correspond to that in the new Hamiltonian. This way exchanges with the reservoir with the new Hamiltonian will be more efficient. The equilibration was performed by a very short MD simulation for 5 ps with positional restraints of 10 $kcal.mol^{-1}.Å^{-2}$ on the backbone heavy atoms to prevent the structures from transitioning to a different conformation. The other input parameters for this short equilibration were the same as the high temperature MD simulation parameters. Then, the equilibrated structures were used to build the reservoirs. The energy of the equilibrated structure was calculated using the imin=5 flag in *sander* program using the topology file and input parameters of the new Hamiltonian. Finally, reservoirs (containing 200 structures with energies corresponding to the new Hamiltonian) were built using the *createreservoir* command in *cpptraj*[88] program in AMBER using a seed of 1.

4.3.4.3 Building super-Hamiltonian reservoirs

For building super-Hamiltonian reservoirs, we combined the 200 reservoir structures from each Hamiltonian resulting in 800 reservoir structures. Then, for each of these 800 structures, for each Hamiltonian, equilibration was performed by a very short MD simulation for 5 ps with positional restraints of 10 $kcal.mol^{-1}.Å^{-2}$ on the backbone heavy atoms to prevent the structures from transitioning to a different conformation. The other input parameters for this short equilibration were the same as the high temperature MD simulation parameters. Then, the equilibrated structures were used to build the reservoirs. The energy of the equilibrated structure

was calculated using the imin=5 flag in *sander* program using the topology file and input parameters of each Hamiltonian. Finally, reservoirs (containing 800 structures with energies corresponding to the each Hamiltonian) were built using the *createreservoir* command in *cpptraj* program in AMBER using a seed of 1.

4.3.4.4 Building reservoirs for predicting effects of mutations

For predicting effects of mutations, 4 different reservoirs were built – one for each mutation. To build these mutated reservoirs, every structure in the reservoir generated using high temperature MD simulation of the wild-type Trp-cage Tc5b variant, was mutated using *Chimera*[112] software package. These mutated structures were minimized for 25 cycles using steepest descent to fix any steric clashes/bad contacts that might have been introduced due to the mutation. Then, each minimized mutated structure was equilibrated by a very short MD simulation for 5 ps with positional restraints of 10 $kcal.mol^{-1}.Å^{-2}$ on the backbone heavy atoms to prevent the structures from transitioning to a different conformation. The other input parameters for this short equilibration were the same as the high temperature MD simulation parameters. Then, the equilibrated structures were used to build the reservoirs. The energy of the equilibrated structure was calculated using the imin=5 flag in *sander* program using the topology file and input parameters of the mutant. Finally, reservoirs (containing 200 structures with energies corresponding to the mutant) were built using the *createreservoir* command in *cpptraj*[88] program in AMBER using a seed of 1.

4.3.5 Analyses:

4.3.5.1 Melting curves and convergence plots

Temperature-based trajectories were extracted from REMD and nB-RREMD simulations using the *cpptraj*[88] program in AMBER. The fraction of native structures in each temperature-based trajectory was calculated using the criteria consistent with our previous published studies for Trp-cage[8, 12, 21], a structure was characterized as native if the backbone RMSD of residues 3 to 18 was <2.0 Å to the first NMR structure. For REMD simulations, since the simulations converged slowly, only for the calculation of melting curves, the first 1 μs of discarded. Since the nB-RREMD simulations converged quickly, all 400 ns of data was used to calculate the melting curves without excluding any data from the beginning of the simulations. The error bars indicate the half difference of the melting curves obtained from the two runs – one starting from native conformation and the other starting from extended conformation.

4.3.5.2 Comparing cluster populations between REMD and nB-RREMD simulations

4.3.5.2.1 Hamiltonian switching simulations

For each Hamiltonian switching pair, for example H1 to H2 and H2 to H1, 10000 structures (5000 from each Hamiltonian) were extracted from the combined high temperature MD trajectories at 319.8 K using an equal time spacing of 160 ps. 10000 structures (5000 from each Hamiltonian) were also extracted from the combined temperature-based trajectories at 264.0 K. Finally, another 10000 structures (5000 from H1 to H2 switch and another 5000 from H2 to H1 switch) were extracted from the combined temperature-based nB-RREMD trajectories at 264.0 K. Then, these 30000 structures were clustered together so that direct comparisons can be made between the cluster populations obtained using high temperature MD from each Hamiltonian,

standard REMD from each Hamiltonian, and nB-RREMD from each Hamiltonian switch simulation.

KMeans[83] clustering algorithm was used, using *cpptraj* program, by setting the target number of clusters to 50. The entire backbone RMSD was used as the clustering metric. A random seed of 23 was used to randomize initial set of points used.

Finally, for each cluster thus obtained for each protein, the relative population of that cluster from each MD, standard REMD, and nB-RREMD simulation was calculated. The Pearson correlation coefficient was calculated between the cluster populations obtained from each method. The slope reported in the figures was obtained by doing a linear regression fit of the cluster populations. While the entire backbone was used for clustering, however, the RMSD values reported in the figures used the same mask as the melting curves calculations so that direct comparison can be made between the fraction of native structures shown in the melting curves and native-like clusters obtained from clustering analysis.

4.3.5.3 Calculation of average persistence time

The persistence time for each reservoir structure is defined as the simulation time between successive successful exchanges with the reservoir. For example, if Replica N accepts structure number "X" from the reservoir, and "t" ns of simulation time elapses before Replica N accepts another/same structure from the reservoir, the persistence time for structure number "X" is then "t" ns. Since each structure gets accepted multiple times from the reservoir during the RREMD simulation, there will be multiple persistence times for each structure. The average persistence time is the average of all these multiple persistence times for each structure. Since energetically favorable structures will be annealed to and retained at lower temperatures more often than other

structures, these structures are expected to have longer persistence times compared to structures that are energetically unfavorable.

4.4 Results and Discussions

The following questions are addressed here: (1) Can nB-RREMD be used to predict accuracy of new force fields? and (2) Can nB-RREMD be used to predict effects of mutations?

To answer the first question, we considered 4 different Hamiltonians for this study: (1) ff14SBonlysc/GB-Neck2 (referred to as H1), (2) ff99SB-ILDN/GB-Neck2 (referred to as H2), (3) fffb15/GB-OBC (referred to as H3), and (4) ff99SB/GB-OBC (referred to as H4). H1 and H2 have different parameters for side chain dihedrals of ILE, LEU, ASP, and GLN amino acids. H1 and H3 have different parameters for not just dihedrals but also for bonds and angles. H1 and H4 have different dihedral parameters and different implicit solvent model.

Since H1 and H3, and H1 and H4 have bigger differences than H1 and H2, predicting the accuracy of H1 using structures obtained from H3 or H4 might be harder than predicting the accuracy of H1 using H2, and vice versa. If nB-RREMD can be used to predict the accuracy of these three force field transitions (different dihedrals; different bonds, angles, and dihedrals; different dihedrals, and implicit solvent), then nB-RREMD simulations can be used to predict the accuracy of new force fields.

To answer the second question, we considered 4 different mutants of Trp-cage protein: (1) Tc5b S13A, (2) Tc5b K8A, (3) Tc5b W6F, and (4) Tc5b S14A. The first two mutations are stabilizing mutations while the last two are destabilizing. Capturing the effects of these stabilizing/destabilizing mutations using nB-RREMD can serve as a good initial test.

Nonetheless, answering the above questions will require generating accurate and precise reference data for each Hamiltonian and each mutation, since the experimental data will not represent the "correct" answer using a given force field and solvent model.

4.4.1 Convergence of standard REMD simulations using each Hamiltonian

To obtain the reference data for predicting the accuracy of each Hamiltonian, we performed extensive independent REMD simulations (see **Methods** for details) starting from native and extended conformations of Trp-cage using each Hamiltonian. Each independent REMD simulation used 4 replicas, spanning a temperature range of 264.0 K to 319.8 K. Furthermore, each replica was simulated for 6 μs.

To check for the convergence of REMD simulations, we calculated the average melting curves (shown in column 1 in **Figure 4-2**) across the two independent REMD simulations, with error bars reflecting the half difference. For each Hamiltonian, even though the two independent REMD simulations were started from distinct initial conformations, they result in similar melting curves (error bars are <5% at all temperatures).

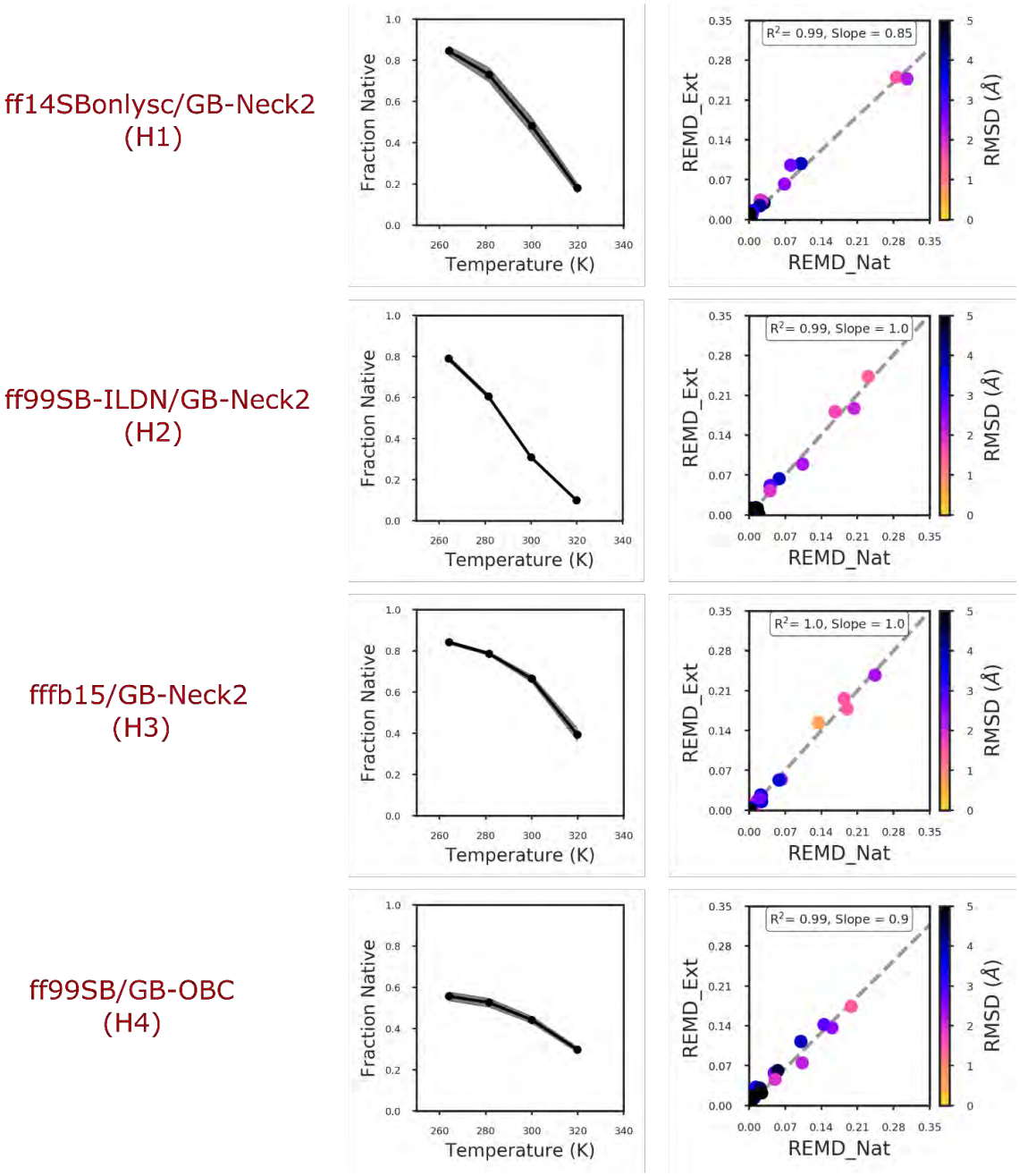


Figure 4-2 Reference data obtained from extensive standard REMD simulations. The melting curves (column 1) and the cluster populations at 264.0 K (column 2) are shown for each Hamiltonian. The error bars in column 1 (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations. "REMD_Nat" and "REMD_Ext" in column 2 indicate that the clusters were obtained from REMD simulations starting from native conformation and extended conformation, respectively. The color of each point in column 2 indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from the two independent REMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each figure in column 2 for each Hamiltonian.

Good agreement between the melting curves obtained from the two independent REMD simulations indicates that similar population of native structure is observed at all temperatures. It is important, however, to analyze the convergence of the entire ensemble, including non-native as well as native structures.[33, 91-92] To test this, we combined the trajectories from the two independent runs at 264.0 K and performed cluster analysis (see **Methods** for details). The cluster populations obtained from the two independent REMD simulations at the calculated melting temperature are shown in column 2 in **Figure 4-2**. For all Hamiltonians, correlation close to 1 with slope >0.85 is observed between the cluster populations at 264.0 K, indicating that the REMD simulations not only sample reproducible population of native structures but also sample reproducible populations of non-native structures.

4.4.2 Convergence of standard REMD simulations of each mutant

Similar to the standard REMD simulations using each Hamiltonian, standard REMD simulations were also performed for each Trp-cage mutant. The melting curves and the cluster populations at 264.0 K, are shown in **Figure 4-3** for each mutant. For each mutant, the melting curves have small error bars (<5% at all temperatures). The correlation of cluster populations is close to 1 for the stabilizing mutants but only 0.74 and 0.75 for the destabilizing mutants Tc5b W6F and Tc5b S14A.

Overall, the melting curves and the cluster populations indicate that these standard REMD simulations are reasonably well converged and can be used as reference data to test the accuracy of force fields and effects of mutations using nB-RREMD.

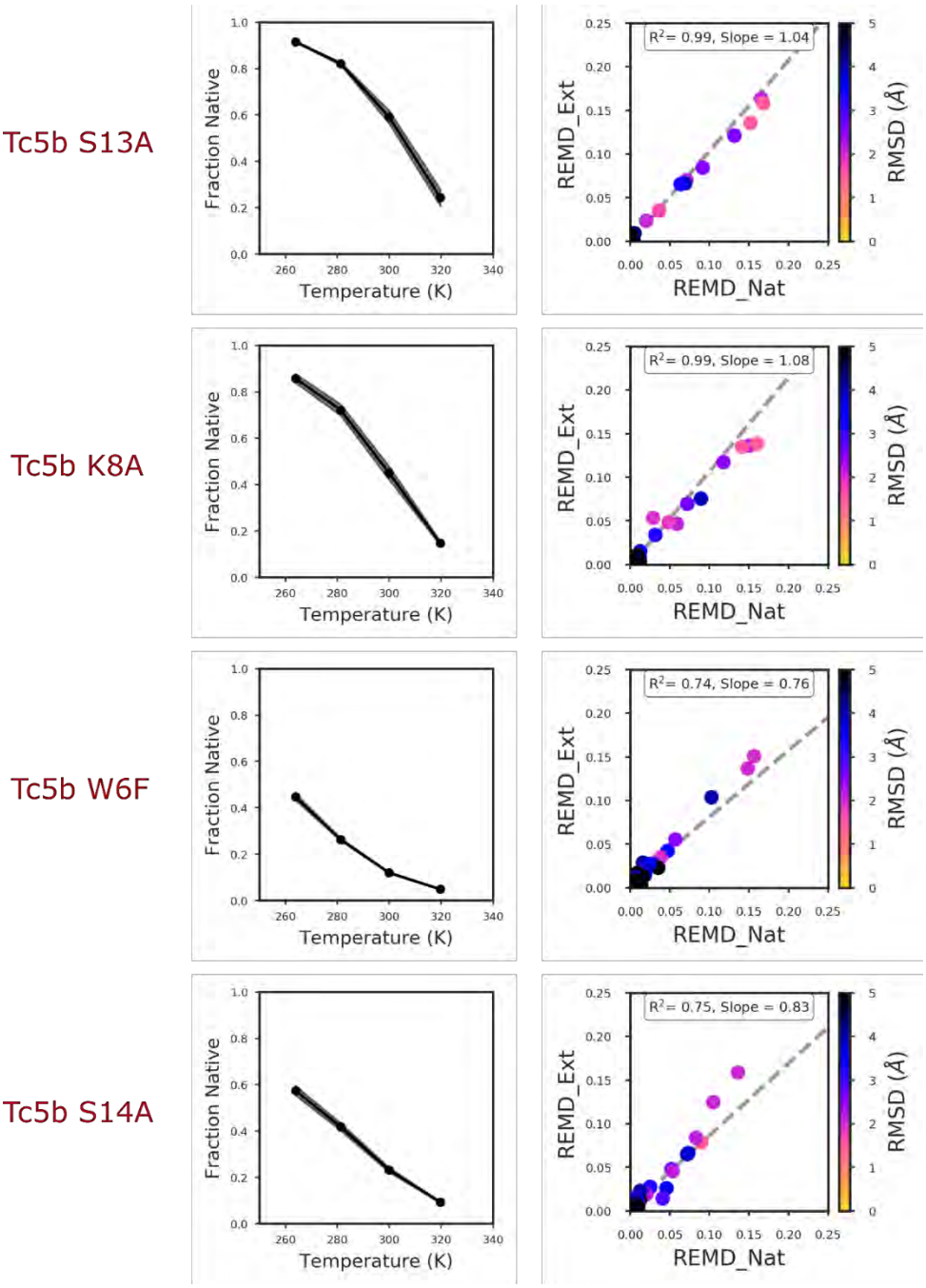


Figure 4-3 Reference data obtained from extensive standard REMD simulations. The melting curves (column 1) and the cluster populations at 264.0 K (column 2) are shown for each mutant. The error bars in column 1 (shown as shaded regions) represent the half difference between the average melting curve and the melting curve obtained from each of the two-independent simulations. "REMD_Nat" and "REMD_Ext" in column 2 indicate that the clusters were obtained from REMD simulations starting from native conformation and extended conformation, respectively. The color of each point in column 2 indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from the two independent REMD simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for each figure in column 2 for each mutant.

4.4.3 Predicting the accuracy of force fields using nB-RREMD

In nB-RREMD, structures generated from one force field can, in theory, be used to predict the results using a different force field. To test this, we built non-Boltzmann reservoirs using the protocols outlined in **Chapter 3**, for each Hamiltonian. Then, we performed nB-RREMD simulations by using a Hamiltonian different from the one used to build the reservoir. If these nB-RREMD simulations can reproduce the ensemble populations of the different Hamiltonian, then nB-RREMD simulations can be more broadly used to predict the accuracy of new force fields.

To rigorously test the applicability of this approach, we performed nB-RREMD simulations reflecting 6 different Hamiltonian perturbations. These 6 different Hamiltonian perturbations can be grouped into 3 pairs: (1) H1 and H2, (2) H1 and H3, and (3) H1 and H4. For each of these pairs, the non-Boltzmann reservoirs built from one force field in the pair was used to predict the accuracy of the other force field in the pair and vice versa. For example, for H1 and H2 pair, structure reservoirs built using H1 were used to perform nB-RREMD simulations using H2 and vice versa. For the method to be broadly applicable, it should predict the effect of perturbations in either directions. In the following sections, we explore the accuracy of our approach in predicting different Hamiltonian perturbations.

4.4.3.1 Can nB-RREMD be used to predict the accuracy of force fields differing in only dihedral parameters?

To test if nB-RREMD can predict the structural preferences of the force fields differing in only dihedral parameters, we performed nB-RREMD simulations with H1 using the structures generated from H2 and vice versa. The melting curves and the cluster populations for each of the force fields and their perturbations are shown in **Figure 4-4.**

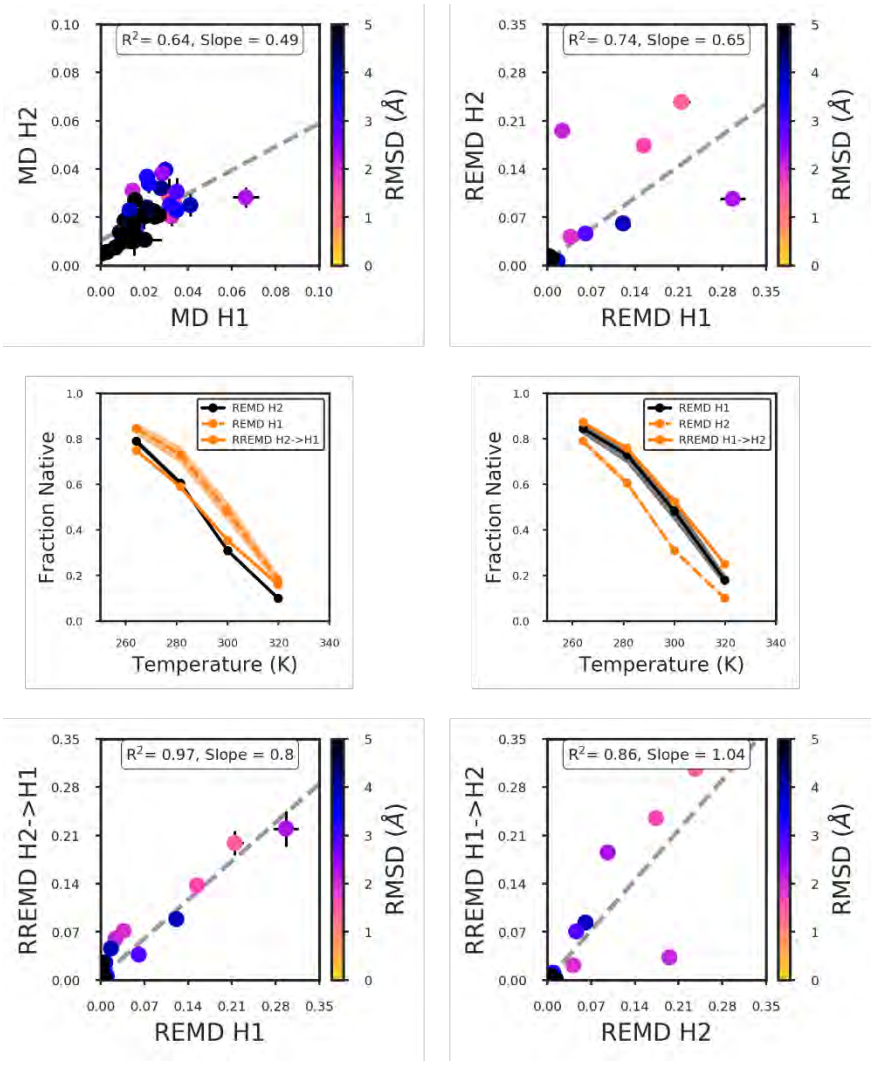


Figure 4-4 The melting curves and the cluster populations obtained from various simulations using Hamiltonians H1 and H2 are shown. "MD H1" and "MD H2" represent the high temperature MD simulations using H1 and H2 Hamiltonians, respectively. "REMD H1" and "REMD H2" represent the standard REMD simulations using H1 and H2 Hamiltonians, respectively. "REMD H1->H2" indicates that the structures from H1 were used to do nB-RREMD simulations using H2 Hamiltonian. Similarly, "REMD H2->H1" indicates that the structures from H2 were used to do nB-RREMD simulations using H1 Hamiltonian. The color of each point in the cluster population plots indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from different simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for in column 2 for each mutant. The melting curve of the reference Hamiltonian is shown as dashed orange lines. The melting curve of the Hamiltonian from which the structures were generated is shown as solid black lines. The melting curve obtained from nB-RREMD simulations is shown as solid orange lines. The solid orange line should overlap with the dashed orange line. The error bars (if present) on the X-axis and Y-axis in the cluster population plots indicate the half difference between the two independent simulations starting from different structures. The error bars on the melting curves indicate the half difference between the melting curves obtained using each independent simulation.

The two Hamiltonians sample slightly different ensembles during the reservoir generation MD simulations at 319.8 K – the correlation of cluster populations is 0.64. They also sample different ensembles at 264.0 K – the correlation of cluster populations is 0.74 between the two REMD simulations. This difference in the ensembles is also reflected in the melting curves obtained from the standard REMD simulations using each Hamiltonian – the melting temperature using H1 is ~5 K greater than the melting temperature using H2.

When structures from H2 are used to do nB-RREMD simulations using H1, the nB-RREMD simulation using H1 results in a melting curve that is still close to H2. Likewise, when structures from H1 are used to do nB-RREMD simulations using H2, the nB-RREMD simulation using H2 results in a melting curve that is closer to H1. This suggests that thermally reweighting the structures using nB-RREMD might not be a viable solution to predict the accuracy of force fields.

However, the bottom row of **Figure 4-4** indicates that nB-RREMD simulations accurately capture the underlying energy landscape of the new Hamiltonian – the correlation of cluster populations between standard REMD simulations using H1 and the nB-RREMD simulations using structures from H2 is 0.97. Similarly, the correlation of cluster populations between standard REMD simulations using H2 and the nB-RREMD simulations using structures from H1 is 0.86. These correlations are significantly higher than the 0.64 and 0.74 correlations observed between the ensembles populated by the two Hamiltonians. The discrepancy between the melting curves and the cluster correlations could be due to the value of cutoffs used to calculate the fraction of native structures.

Overall, the above results indicate that nB-RREMD simulations can be used to predict the accuracy of force fields differing in dihedral parameters.

4.4.3.2 Can nB-RREMD be used to predict the accuracy of force fields differing in bonds, angles, and dihedral parameters?

To test if nB-RREMD can predict the structural preferences of the force fields differing in bonds, angles, and dihedral parameters, we performed nB-RREMD simulations with H1 using the structures generated from H3 and vice versa. The melting curves and the cluster populations for each of the force fields and their perturbations are shown in **Figure 4-5.**

The two Hamiltonians sample significantly different ensembles during the reservoir generation MD simulations at 319.8 K – the correlation of cluster populations is only 0.36. They also sample different ensembles at 264.0 K – the correlation of cluster populations is 0.60 between the two REMD simulations. This difference in the ensembles is also reflected in the melting curves obtained from the standard REMD simulations using each Hamiltonian – the melting temperature using H1 is ~10 K less than the melting temperature using H3.

When structures from H3 are used to do nB-RREMD simulations using H1, the nB-RREMD simulation using H1 results in a melting curve that is still close to H3. However, the correlation of cluster populations between the standard REMD simulations using H1 and nB-RREMD simulations using structures from H3 is 0.87 indicating that the structures can adapt to the different Hamiltonian.

On the other hand, when structures from H1 are used to do nB-RREMD simulations using H3, the nB-RREMD simulations using H3 result in a melting curve that matches very well with the standard REMD melting curves obtained using H3. This is also reflected in the high degree of correlation ($R^2 = 0.99$) of cluster populations between the standard REMD simulations using H3 and the nB-RREMD simulations using structures from H1. The above results indicate that nB-RREMD simulations can be used to predict the effects of perturbing from H1 to H3 and vice versa.

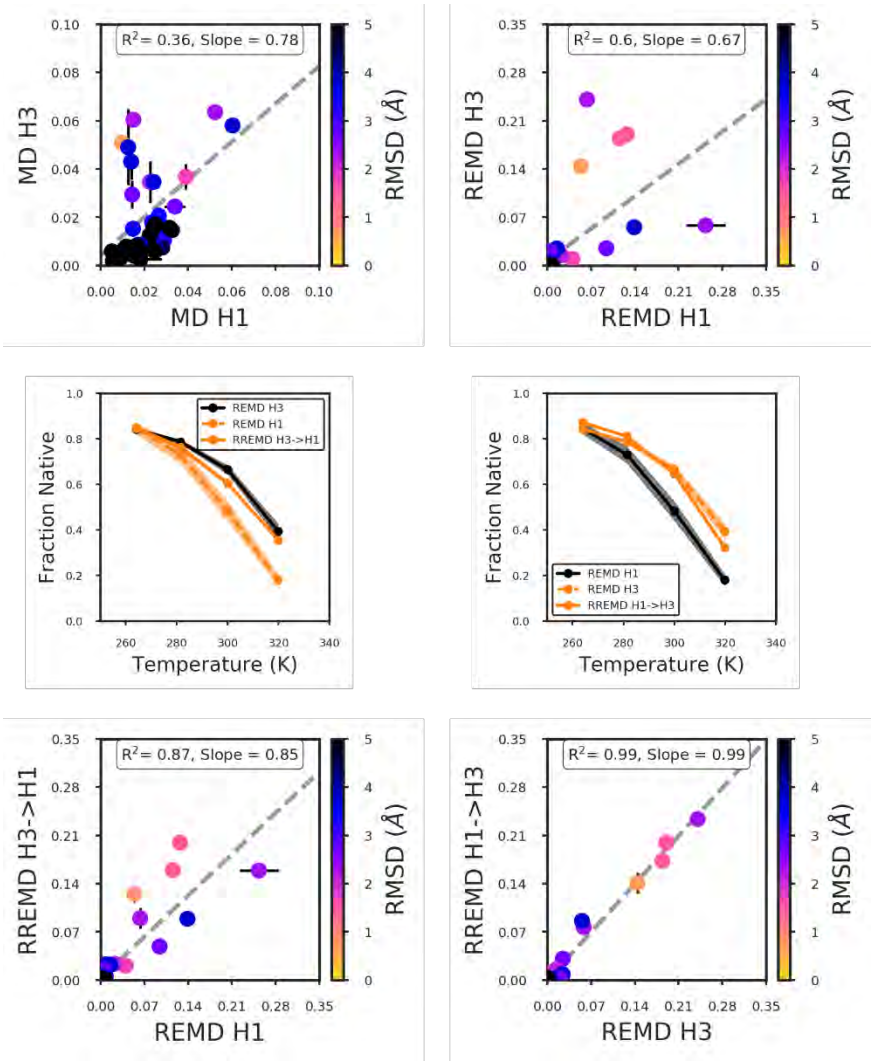


Figure 4-5 The melting curves and the cluster populations obtained from various simulations using Hamiltonians H1 and H3 are shown. "MD H1" and "MD H3" represent the high temperature MD simulations using H1 and H3 Hamiltonians, respectively. "REMD H1" and "REMD H3" represent the standard REMD simulations using H1 and H3 Hamiltonians, respectively. "REMD H1->H3" indicates that the structures from H1 were used to do nB-RREMD simulations using H3 Hamiltonian. Similarly, "REMD H3->H1" indicates that the structures from H3 were used to do nB-RREMD simulations using H1 Hamiltonian. The color of each point in the cluster population plots indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from different simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for in column 2 for each mutant. The melting curve of the reference Hamiltonian is shown as dashed orange lines. The melting curve of the Hamiltonian from which the structures were generated is shown as solid black lines. The melting curve obtained from nB-RREMD simulations is shown as solid orange lines. The solid orange line should overlap with the dashed orange line. The error bars (if present) on the X-axis and Y-axis in the cluster population plots indicate the half difference between the two independent simulations starting from different structures. The error bars on the melting curves indicate the half difference between the melting curves obtained using each independent simulation.

4.4.3.3 Can nB-RREMD be used to predict the accuracy of force fields differing in dihedrals and solvation parameters?

To test if nB-RREMD can predict the structural preferences of the force fields differing in dihedral and solvation parameters, we performed nB-RREMD simulations with H1 using the structures generated from H4 and vice versa. The melting curves and the cluster populations for each of the force fields and their perturbations are shown in **Figure 4-6.**

The two Hamiltonians sample very different ensembles during the reservoir generation MD simulations at 319.8 K – the correlation of cluster populations is <0.30. They also sample different ensembles at 264.0 K – the correlation of cluster populations is 0.60 between the two REMD simulations. This difference in the ensembles is also reflected in the melting curves obtained from the standard REMD simulations using each Hamiltonian – the fraction of native structures at 264.0 K are 0.86 and 0.56 for H1 and H4, respectively.

When structures from H4 are used to do nB-RREMD simulations using H1, the nB-RREMD simulation using H1 results in a melting curve that matches remarkably well with the standard REMD reference melting curve obtained using H1. Even the correlation of cluster populations between the standard REMD simulations using H1 and nB-RREMD simulations using structures from H4 is 0.93.

Crucially, the nB-RREMD simulations also capture the perturbations in the other direction accurately. When structures from H1 are used to do nB-RREMD simulations using H4, the nB-RREMD simulations using H4 result in a melting curve is in excellent agreement with the standard REMD melting curves obtained using H4. This is also reflected in the high degree of correlation ($R^2 = 0.94$) of cluster populations between the standard REMD simulations using H4 and the nB-RREMD simulations using structures from H1.

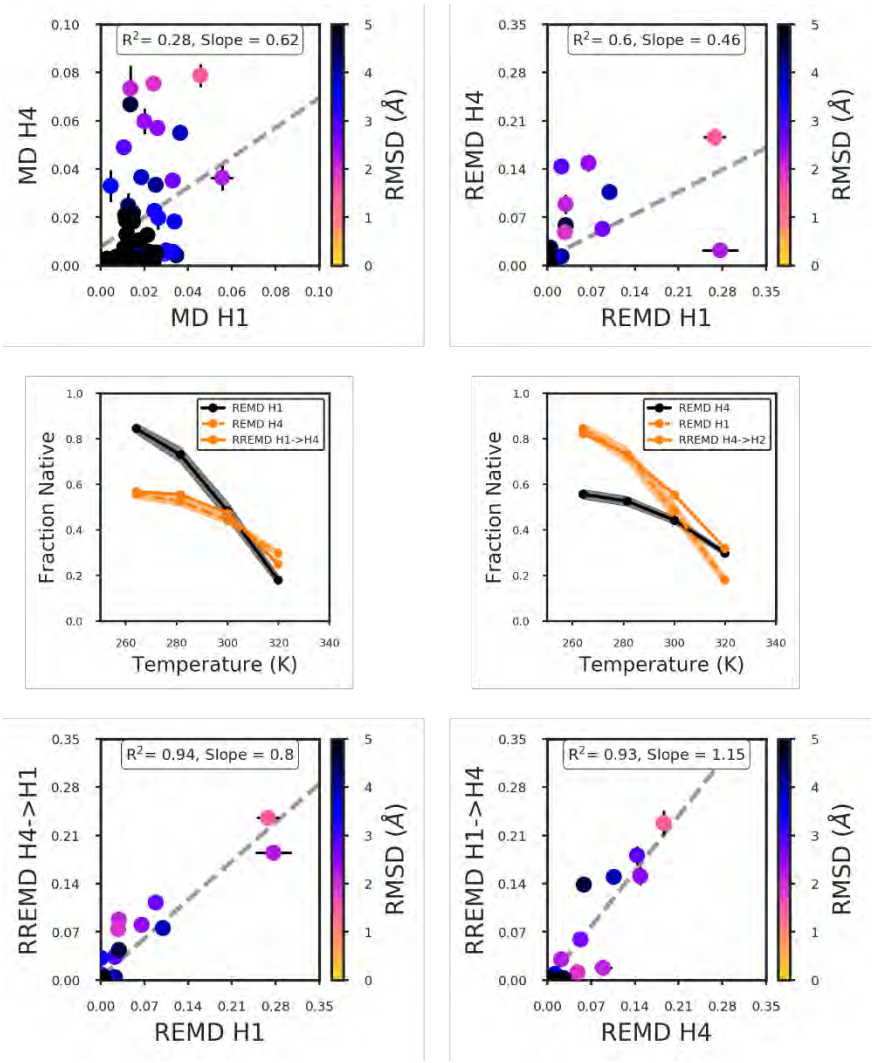


Figure 4-6 The melting curves and the cluster populations obtained from various simulations using Hamiltonians H1 and H4 are shown. "MD H1" and "MD H4" represent the high temperature MD simulations using H1 and H4 Hamiltonians, respectively. "REMD H1" and "REMD H4" represent the standard REMD simulations using H1 and H4 Hamiltonians, respectively. "REMD H1->H4" indicates that the structures from H1 were used to do nB-RREMD simulations using H4 Hamiltonian. Similarly, "REMD H4->H1" indicates that the structures from H4 were used to do nB-RREMD simulations using H1 Hamiltonian. The color of each point in the cluster population plots indicates the RMSD to the native structure. The Pearson correlation coefficient between the cluster populations obtained from different simulations and the slope of the best fit line (represented by the dashed grey line) are shown inside the box for in column 2 for each mutant. The melting curve of the reference Hamiltonian is shown as dashed orange lines. The melting curve of the Hamiltonian from which the structures were generated is shown as solid black lines. The melting curve obtained from nB-RREMD simulations is shown as solid orange lines. The solid orange line should overlap with the dashed orange line. The error bars (if present) on the X-axis and Y-axis in the cluster population plots indicate the half difference between the two independent simulations starting from different structures. The error bars on the melting curves indicate the half difference between the melting curves obtained using each independent simulation.

The above results indicate that nB-RREMD simulations can be used to predict the effects of perturbing from H1 to H4 and vice versa. Overall, **Figures 4-4**, **4-5**, and **4-6** indicate that structures derived from Hamiltonian can be used to predict the accuracy of a different Hamiltonian via nB-RREMD.

4.4.4 Predicting the effects of mutations using nB-RREMD

Since the relevant conformations are already sampled in the reservoir, mutating the structures in the reservoir and then reweighting them through nB-RREMD can, in principle, be used to predict the effects of mutations. To test this, we built non-Boltzmann reservoirs using the protocols outlined in **Chapter 3**, for the wild-type Trp-cage protein. Then, we mutated each of the structures in the reservoir (see **Methods**). Then, we performed nB-RREMD simulations using these mutated reservoir structures. If these nB-RREMD simulations can quantitatively reproduce the ensemble populations of the different mutants, then nB-RREMD simulations can be more broadly used to design proteins with the desired properties.

To test the applicability of this approach, we performed nB-RREMD simulations with 4 different mutations (see **Methods**). The melting curves and the cluster populations for each of the mutants compared to the standard REMD reference simulations are shown in **Figure 4-7.** For all mutants except Tc5b W6F, the melting curves from nB-RREMD simulations match remarkably well with the standard REMD melting curves. However, the correlation of cluster populations is close to 0.8 even for Tc5b W6F mutant indicating that nB-RREMD simulations can, indeed, be used to capture the effects of mutations.

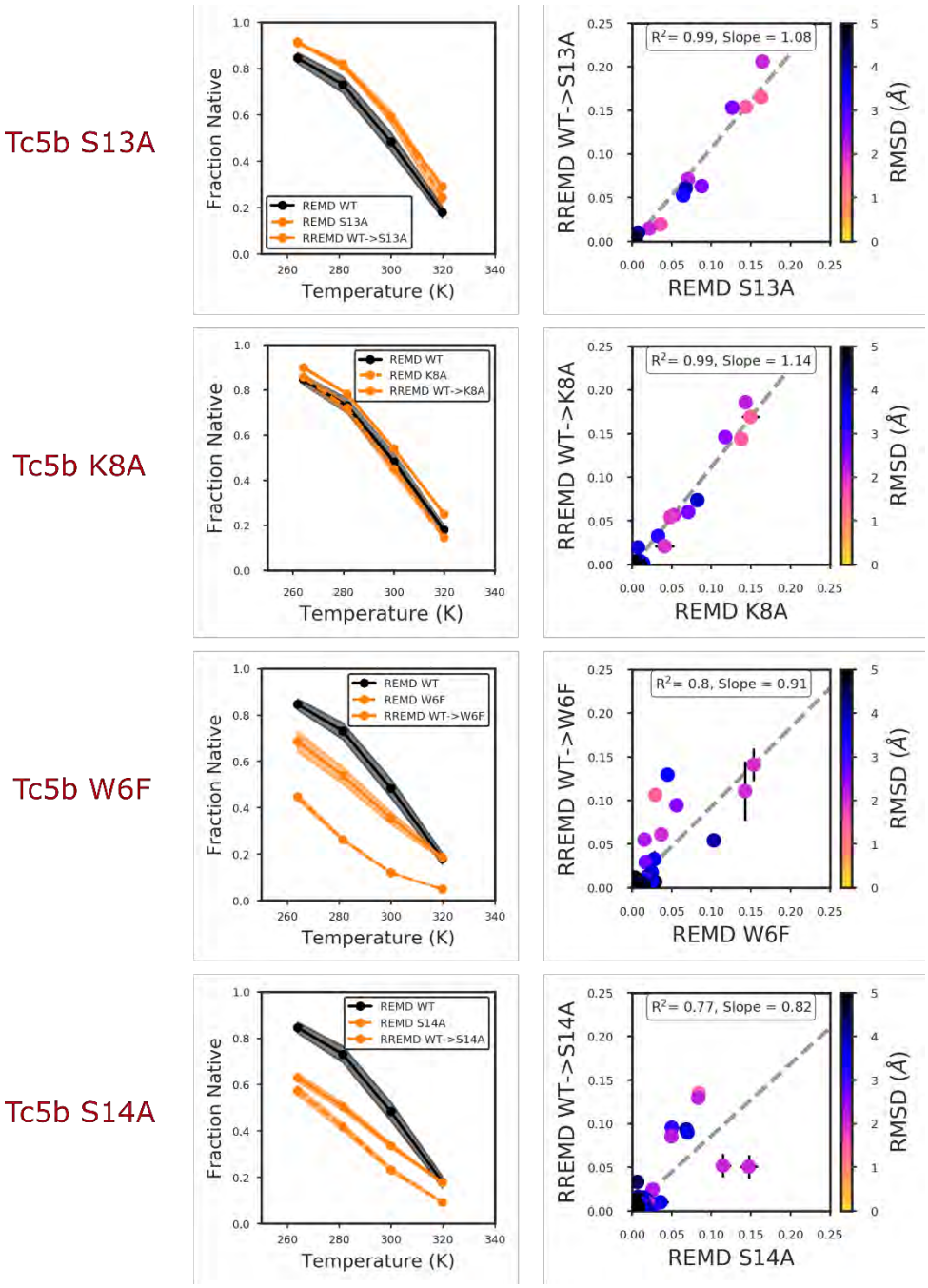


Figure 4-7 The melting curves obtained from standard REMD simulations for the wild-type (solid black lines) and the mutants (dashed orange lines) are shown in column 1. The melting curves obtained from nB-RREMD simulations (solid orange lines) using the mutated reservoirs are also shown in column 1. The cluster populations obtained from the standard REMD reference simulations for each mutant are shown on the X-axis. The cluster populations obtained from the nB-RREMD simulations using the mutated reservoir structures are shown in column 2. The error bars (if visible) on the X-axis and Y-axis in the cluster population plots indicate the half difference between the two independent simulations starting from different structures.

Nonetheless, since it is difficult to sample structures that are absent in the reservoir, to accurately predict the accuracy of force fields or the effects of mutants, the basins that are sampled by the Hamiltonian used in nB-RREMD simulations must also be sampled by the Hamiltonian used to generate the reservoir structures. This may not be true always. However, since different Hamiltonians sample different structures, combining the structures from each of these Hamiltonians might eventually exhaust the conformational search for a given protein. These combined structures (super-reservoirs) for a given protein can be used in a similar manner to predict the accuracy of a new force field. From this super-reservoir of structures, the new force field must be able to selectively sample the structures it favors the most.

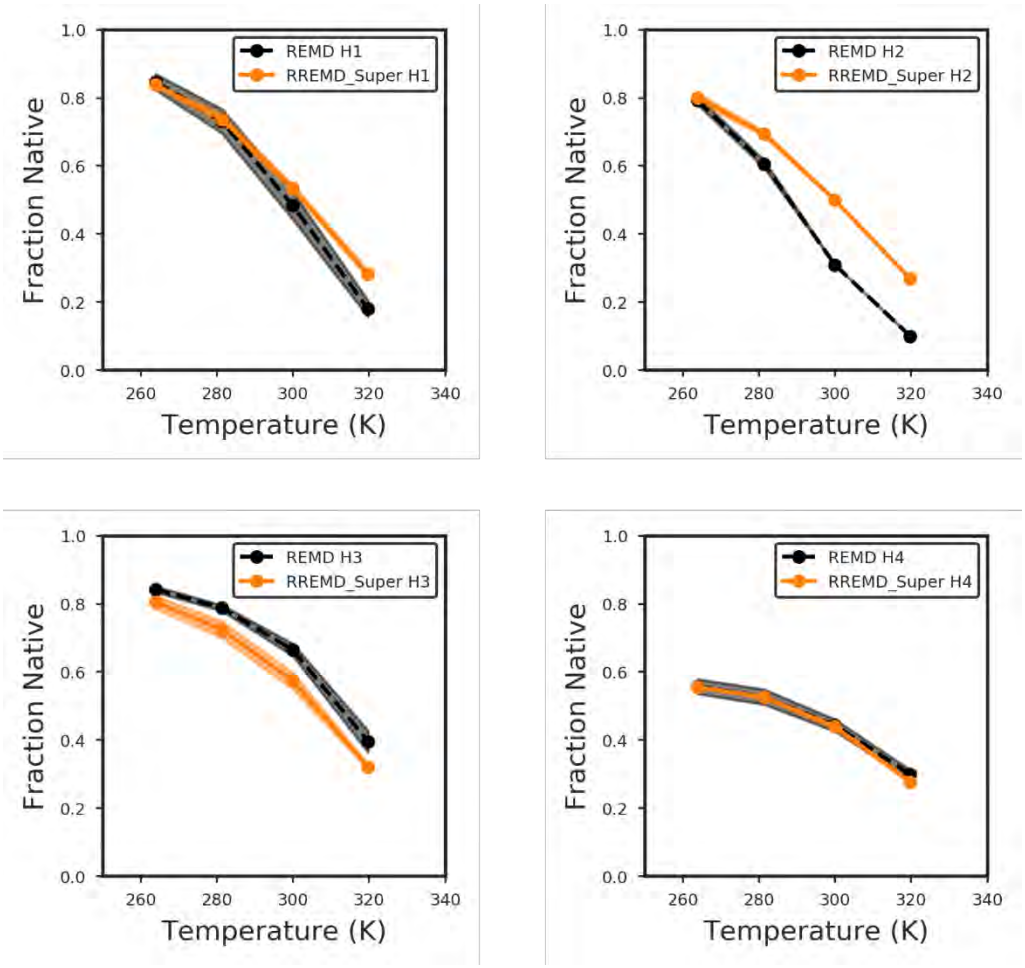


Figure 4-8 Melting curves obtained from standard REMD simulations and nB-RREMD simulations using super-reservoirs are shown. The solid orange lines match remarkably well with the dashed black lines for all Hamiltonians.

To test this, we built super-reservoirs (see **Methods**) using the structures from 4 different Hamiltonians used in this study. The melting curves obtained using these super-reservoirs for each of the Hamiltonians are shown in **Figure 4-7**. As expected, for all Hamiltonians, the melting curves obtained from nB-RREMD simulations using these super-reservoirs are in excellent agreement with the melting curves obtained from standard REMD simulations.

Also, since the structures sampled at the low temperatures can be directly mapped to a structure in the reservoir (because of lack of transitions from one reservoir structure to another reservoir structure during REMD part of RREMD), the structures preferred by each Hamiltonian

can be easily identified by calculating the average persistence time. The average persistence (see **Methods**) for a given structure indicates the time that the structure spends in the REMD part of RREMD before exchanging back with the reservoir. The longer the persistence time, the more favored the structure is. **Figure 4-8** shows the average persistence time of different reservoir structures during the REMD part of nB-RREMD for the various structures in the super-reservoir for H1 and H4 Hamiltonians. When H1 is used to do nB-RREMD simulations using the super-reservoir (top image in **Figure 4-8**), structures that were obtained from H1 have a high average persistence time. Likewise, when H4 is used to the nB-RREMD simulations (bottom image in **Figure 4-8**), the structures obtained from H4 have higher average persistence time.

Broadly, this indicates that the average persistence times can be used to characterize the structural preferences of a given Hamiltonian. More importantly, they can also be used to diagnose issues in force fields. If a non-native structure is preferred by a given Hamiltonian, it is because the structure is favored energetically. Then, these structures can be visualized in a visualization software to help understand the reason for its favorability. Likewise, the average persistence times can also be used to predict the effects of mutations on each structure in the reservoir. Such information can be used to design peptides with the desired properties.

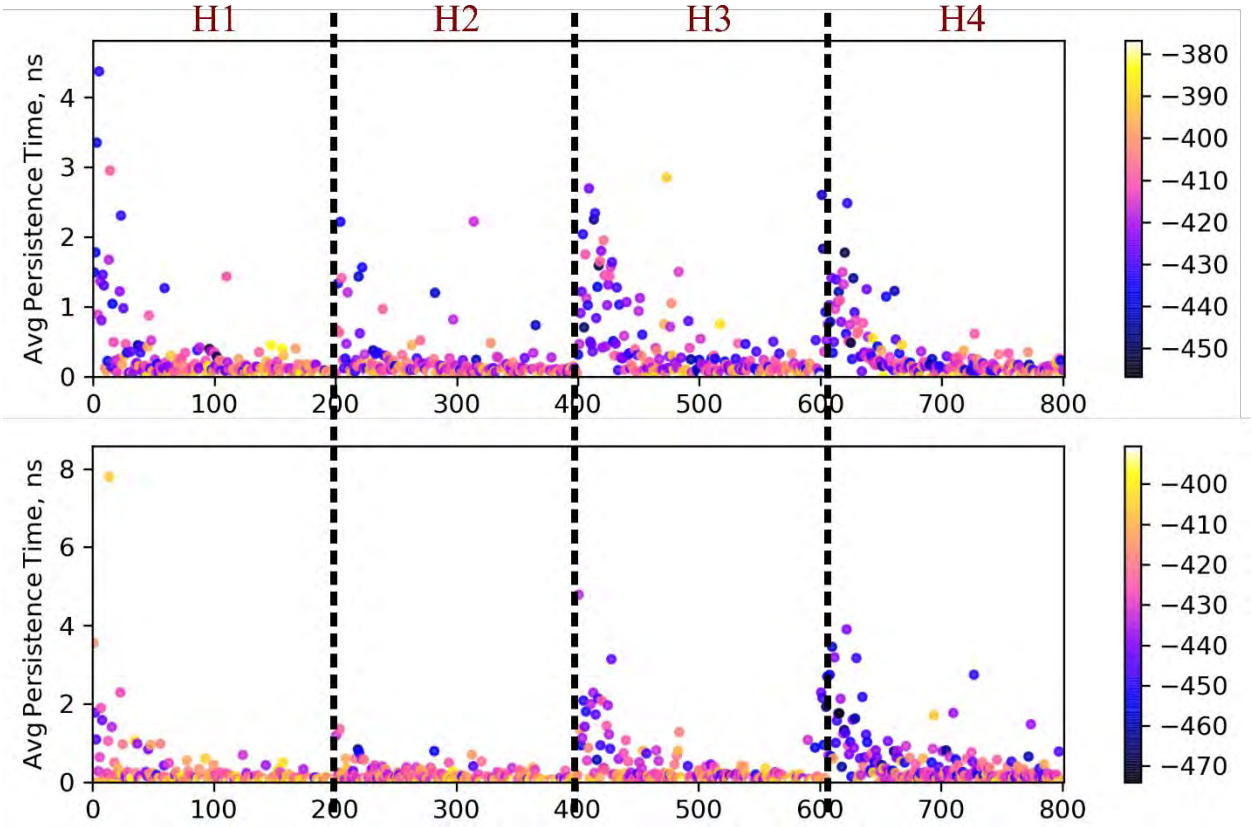


Figure 4-9 The average persistence of each of the 800 structures in the super-reservoir when nB-RREMD simulation used H1 (top), and when nB-RREMD simulation used H4 (bottom). The vertical lines represent the Hamiltonians from which the structures were obtained. The first 200 structures were obtained from H1, the next 200 from H2, the next 200 from H3, and the last 200 from H4. The color of each point indicates the energy of the structure. Low energy structures usually have a high average persistence time.

4.5 Conclusions

In this chapter, reservoir built using structures from one Hamiltonian were used to predict the accuracy of a different Hamiltonian. We also used nB-RREMD to predict the effects of mutations reasonably well. These results indicate that nB-RREMD can be broadly used to reduce the time required to test force fields, and also to design peptides.

5 Using Structure Reservoirs to Accelerate Explicit Solvent Simulations

5.1 Abstract

In explicit solvent simulations, the large number of degrees of freedom and high viscosity reduces the rate of conformational sampling significantly. To address this, several variants of T-REMD have been developed. While these variants reduce the number of replicas required, the enhancement in sampling is still limited by the high viscosity. As an alternative to these sampling methods, in this chapter, B-RREMD simulation was performed to test if pre-sampled structure reservoirs can accelerate sampling in explicit solvent. The results show that RREMD can, indeed, remarkably accelerate conformational sampling in explicit solvent simulations by around1000x.

5.2 Introduction

Accurate treatment of solvation in MD simulations is crucial to understand the function of biomolecules in an aqueous environment. Solvation effects can be modeled either implicitly by using a continuum dielectric[58-59] (usually referred to as implicit solvent) or explicitly by including all the water molecules with atomistic representation[2-3, 9] (usually referred to as explicit solvent).

While representing solvation effects using a continuum dielectric might be too simplistic, the low computational cost of implicit solvent models due to the absence of explicit solvent molecules makes them an attractive choice for simulation of biomolecules, and hence are widely used[21, 113] and improved[4, 6, 8, 11]. In addition to the low computational cost, the low viscosity in implicit solvent compared to explicit solvent can accelerate conformational sampling significantly[62, 114-115]. To illustrate this, the RMSD vs. time for MD simulations of Trp-cage in implicit and explicit solvents, at the same temperature, are shown in **Figure 5-1**. In implicit solvent, multiple (>50) folding and unfolding (RMSD >2.0 Å) are observed whereas only 1 major folding/unfolding event is observed in explicit solvent.

The above advantages notwithstanding, most implementations of implicit solvent models estimate only the polar component of solvation energy, and hence have limited accuracy. The non-polar component, which is usually approximated by the solvent accessible surface area, is often ignored due to the cost of calculating the surface area and its derivatives, and also because the magnitude of the non-polar term is much smaller than the polar term. While an algorithm to speed up the calculation of solvent accessible area has been recently developed and implemented[12], using solvent accessible area to estimate non-polar solvation itself has several limitations[116-118]. Moreover, the effects of structured water molecules close to the surface of the biomolecule are

completely ignored which can lead to incorrect salt-bridge geometries, besides other issues[37, 119-123].

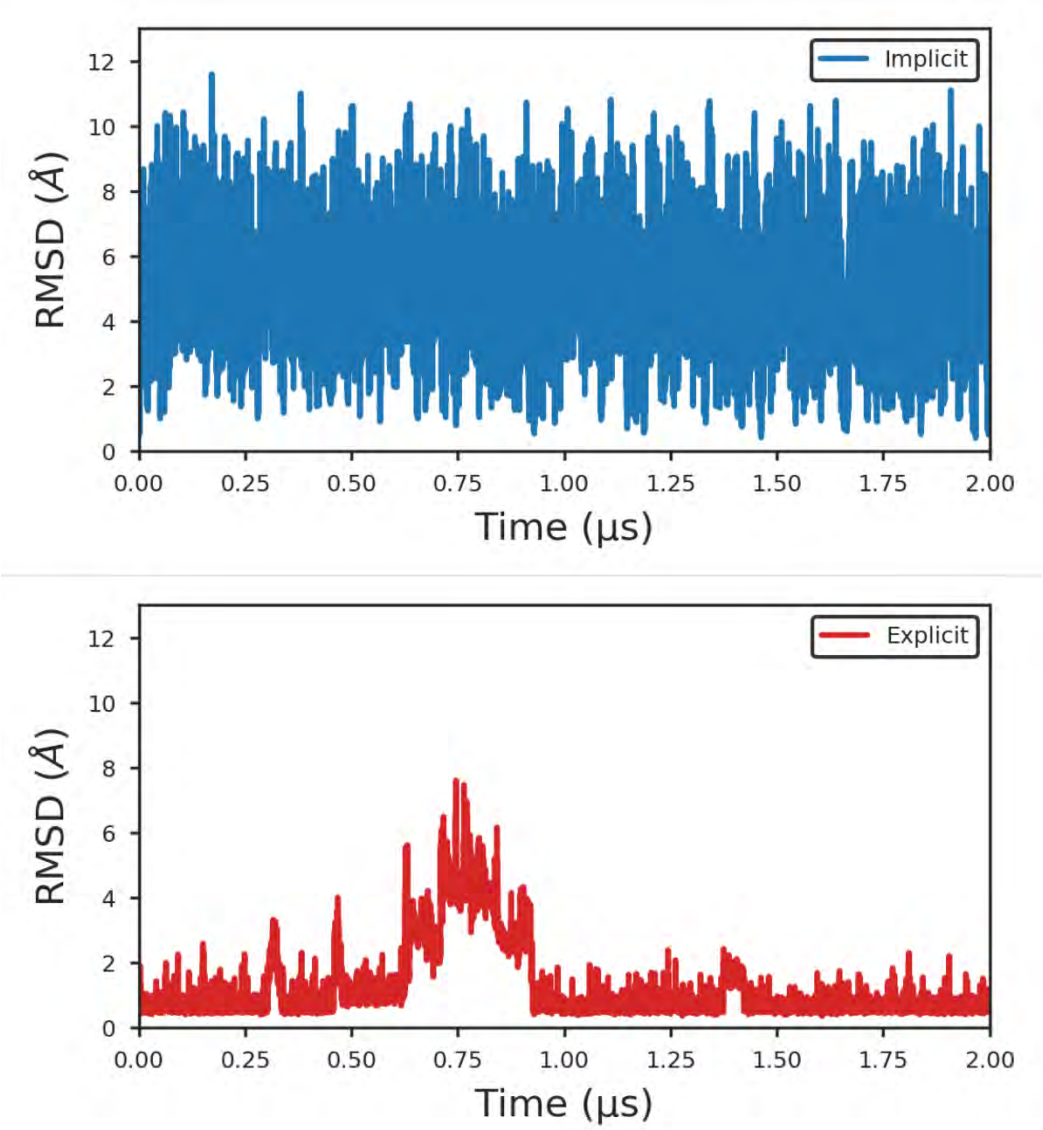


Figure 5-1 The RMSD as a function of time in implicit solvent (top) and explicit solvent (bottom) simulations of Trp-cage variant Tc5b at the same temperature are shown.

Therefore, treating the solvation effects using an explicit solvent is the gold standard of MD simulations. However, as mentioned above, calculating the interactions between all the solvent molecules and the solute atoms is computationally demanding even when Periodic

Boundary Conditions (PBC) and Particle Mesh Ewald (PME) summation algorithm[124] is used. Moreover, as shown in **Figure 5-1**, the presence of water molecules introduces viscosity into the simulation system which also slows down the rate of conformational sampling.

Algorithms such as T-REMD can be used to accelerate conformational sampling in explicit solvent[120]. However, since the temperature spacing between replicas is proportional to the inverse square root of number of degrees of freedom, the huge number of water molecules means that several replicas have to be used to span a small temperature range for the T-REMD exchanges to be efficient. For example, **Figure 5-2** shows that ten replicas are required to span a temperature range of 20 K when explicit solvent is used compared to only 4 replicas required to span a temperature range of 60 K when implicit solvent is used (see **Figure 1-9**).

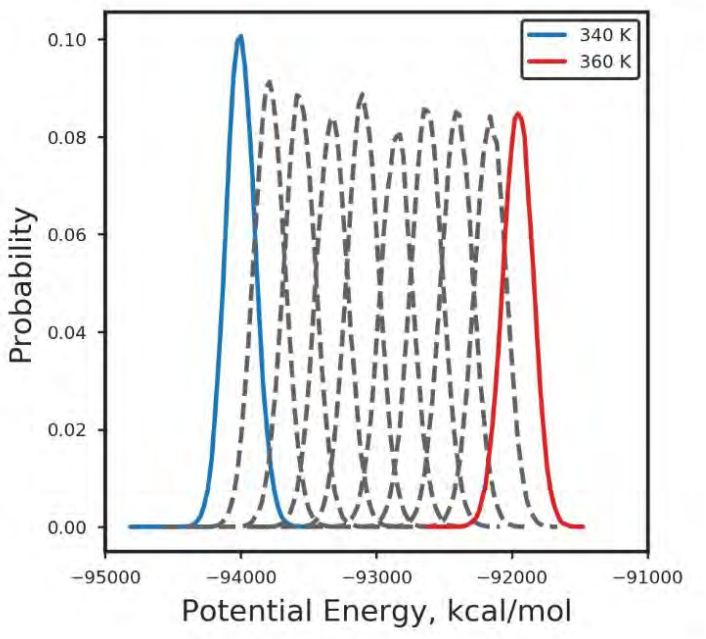


Figure 5-2 Histograms of potential energy at different temperatures for Trpcage simulations in explicit solvent. The cold temperature (340 K) and the hot temperature (360 K) are shown as blue and red solid lines, respectively. Additional replicas at intermediate temperatures are shown as grey dashed lines.

If the conformational sampling in explicit solvent simulations at high temperatures is significantly faster than explicit solvent simulations at low temperature, then the need to use large number of replicas can be addressed by using highly parallelized high-performance computing clusters. However, **Figure 5-3** shows that, when explicit solvent is used, increasing the temperature results in only a modest increase in rate of conformational sampling – the number of folding and unfolding (RMSD >2.0 Å) are slightly but not significantly higher at 360 K compared to 340 K.

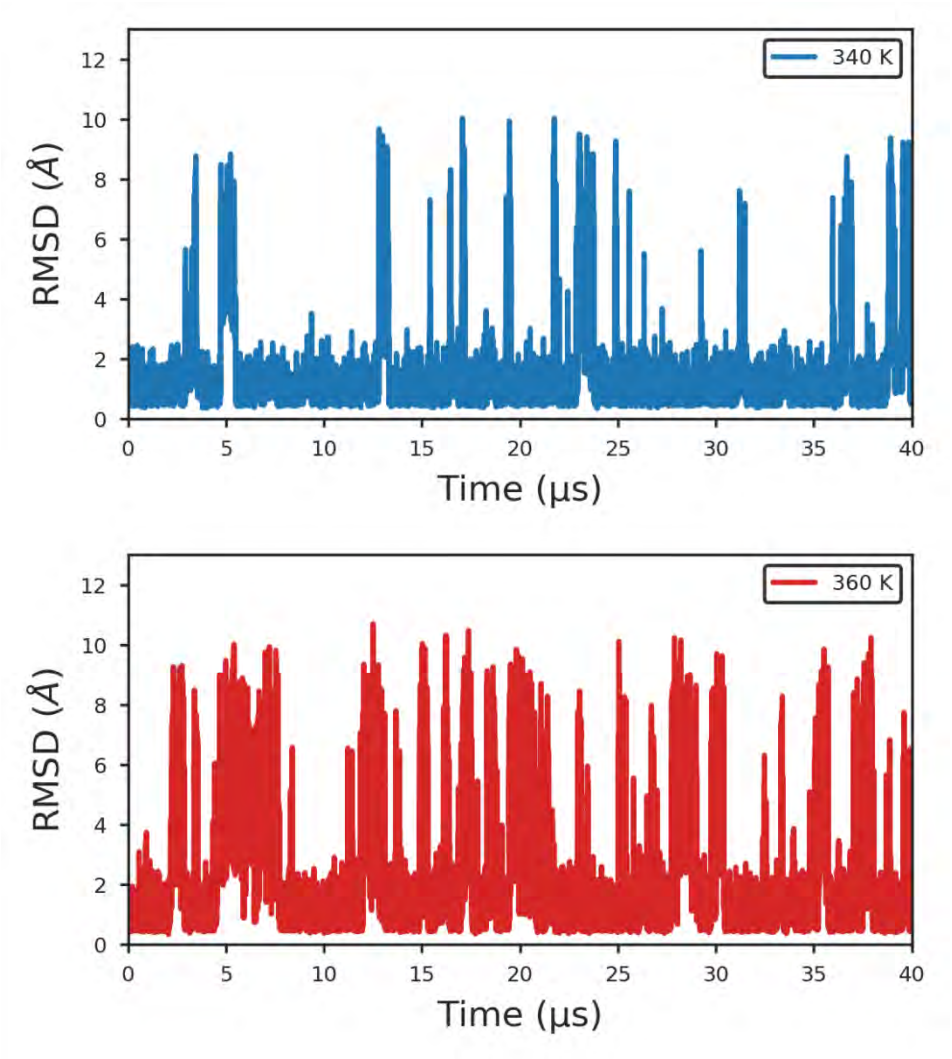


Figure 5-3 The RMSD as a function of time for explicit solvent simulations of Trp-cage variant Tc10b at 340 K (top) and 360 K (bottom) are shown.

Moreover, it still takes around 1-2 μs to fold and unfold even at 360 K. Currently, it takes around 2-3 days to simulate a small protein like Trp-cage in explicit solvent, for 1 μs, on a GTX 1080. Therefore, to simulate at least 50 folding and unfolding events, T-REMD simulations will have to be run for >50 μs, resulting in an overall wall-clock time of at least three months.

Therefore, several variants of T-REMD[37-39, 125-134] have been developed to accelerate conformational sampling in explicit solvent. These variants either (a) reduce the number of replicas required to span the desired temperature range, or (b) reduce the viscosity so that conformational sampling can occur faster.

5.2.1 Reducing the number of replicas in explicit solvent T-REMD simulations

Among this category of T-REMD variants, Hybrid-solvent T-REMD (Hybrid-REMD)[37-38, 127, 129-130] and Replica Exchange with Solute Tempering (REST)[39, 41] algorithms are the most popular ones.

In Hybrid-REMD, independent replicas of the system are simulated at different temperatures. However, during exchanges, the potential energy of a stripped system instead of the entire system is used to do the exchange. The stripped system usually consists of the solute, first or up to second shell of water molecules, and implicit solvent. The shell of water molecules is used to capture the structured water effects accurately, and the implicit solvent is used to model the bulk water. The use of implicit solvent to model the bulk water significantly reduces the number of degrees of freedom of the stripped system, thereby facilitating the use of fewer replicas to span the desired temperature range.

Hybrid-REMD methods have been successfully used to study secondary structure preferences of small peptides, fix salt-bridge geometries, and to fold small proteins, with a reduced number of replicas[37-38, 123]. However, it was also shown that the results were sensitive to the number

of water molecules included and the implicit solvent used during exchange[37]. Since identifying the ideal number of water molecules required to capture structured water effects for a given biomolecule is non-trivial, and since implicit solvent models used are approximate and have their own limitations[8], hybrid-REMD has not been widely used. Moreover, even though the number of replicas is significantly reduced, the conformational changes are still slow due to the high viscosity of fully explicit solvent, which is used during all MD steps.

In REST[39] and its more recent version REST2[41], only the solute (and not solvent) is simulated at different temperatures and only those degrees of freedom are used during exchange. Since only a few degrees of freedom are used, the number of replicas required to simulate the system is also less. However, since the solvent degrees of freedom are essentially "frozen", the bulk water molecules prohibit large scale conformational changes, resulting in only local enhancement of conformational sampling in REST2 simulations compared to global enhanced sampling obtained from corresponding T-REMD simulations[42, 135].

5.2.2 Reducing viscosity in explicit solvent T-REMD simulations

To reduce the viscosity in explicit solvent simulations without introducing artifacts in protein folding dynamics and kinetics, viscosity-REMD (V-REMD)[128] was developed. In V-REMD and in its generalized formalism mass-manipulating REMD (MMREMD)[132], the solvent viscosity is decreased by decreasing the mass of the solvent waters. However, it was observed that decreasing the viscosity did not significantly improve the conformational sampling of systems with large energetic barriers. This is expected since viscosity mainly affects the rate of diffusion through the water molecules but not significantly alter the transition from one minimum to another. In such cases, despite the decrease in viscosity, the improvement in the conformational sampling was mainly due to the high temperature replicas[128].

5.2.3 Current work

Nevertheless, since transitioning from one minimum to another is the primary bottleneck of T-REMD explicit solvent simulations, using pre-sampled reservoirs with structures corresponding to different minimum should, in principle, be able to accelerate conformational sampling. Indeed, the results shown in this chapter indicate that RREMD simulations can accelerate biomolecular simulations in explicit solvent by around 1000x.

5.3 Methods

5.3.1 General Details

The Trp-cage variant Tc10b[111] (PDB ID: 2JOF) was considered for this study. The first NMR model was used as the initial structure. LEaP module of AmberTools in the AMBER 18 package[56] was used to build the structure for simulation. The ff14SB force field[87] and TIP3P water model[2] were used for all simulations. To ensure that even the non-compact unfolded states are adequately solvated, a 25 Å buffer was used to solvate the system resulting in 9471 water molecules. Non-bonded interactions were calculated directly up to 8.0 Å with cubic spline switching and PME approximation with direct sum tolerances of 0.00001. SHAKE[44] was performed on all bonds including hydrogen with the AMBER default tolerance of 0.00001 Å.

5.3.2 Minimization and Equilibration

Minimization and equilibration were performed with a weak-coupling (Berendsen[136]) thermostat and barostat targeted to 1.0 bar with isotropic position scaling as follows. With 100.0 kcal mol^{-1} $Å^{-2}$ positional restraints on peptide heavy atoms, structure was minimized for up to 10000 cycles and then heated at constant volume from 100.0 K to 300.0 K over 100.0 ps, followed by another 100.0 ps at 300.0 K. The pressure was equilibrated for 100.0 ps and then 250.0 ps with

time constants of 100.0 fs and then 500.0 fs on coupling of pressure and temperature to 1.0 bar and 300.0 K, with 100.0 kcal mol^{-1} $Å^{-2}$ and then 10.0 kcal mol^{-1} $Å^{-2}$ Cartesian positional restraints on peptide heavy atoms. The system was again minimized, with 10.0 kcal mol^{-1} $Å^{-2}$ force constant Cartesian restraints on only the protein main chain N, Cα, and C for up to 10000 cycles. Then, three 100.0 ps simulations with temperature and pressure time constants of 500.0 fs were performed, with backbone restraints of 10.0 kcal mol^{-1} $Å^{-2}$, 1.0 kcal mol^{-1} $Å^{-2}$, and then 0.1 kcal mol^{-1} $Å^{-2}$. Finally, the system was simulated unrestrained with pressure and temperature time constants of 1.0 ps for 500.0 ps with a 2.0 fs time step, removing center-of-mass translation every 1.0 ps.

5.3.3 MD simulations

MD simulations were performed at two different temperatures 340 K, and 360 K, for 119.6 μs, and 119.34 μs, respectively. The simulation at 360 K was run for 119.34 μs instead of 119.6 μs due to a cluster shutdown which prevented the simulation from finishing. Chirality constraints and *trans*-peptide *ω* constraints obtained using *makeCHIR_RST* program in AMBER were used at all temperatures to prevent chirality inversions and peptide bond flips. A time step of 2.0 fs was used for all simulations for the first 5 μs. Then, to increase the rate of conformational sampling, a 4.0 fs time step was used for the rest of the simulations by modifying the masses of the hydrogen atoms of solute[46]. Coordinates were saved every 100 ps.

5.3.4 RREMD simulation

The RREMD simulation was run in NVT ensemble using 10 replicas. The replica temperatures were set to 338.0, 340.0, 342.2, 344.3, 346.5, 348.7, 351.0, 353.2, 355.4, and 357.7. The parameters for each replica in the REMD simulations are the same as those in the production MD simulations except that each replica was simulated for 100.0 ns with exchanges attempted

every 1.0 ps. A 4.0 fs time step was used for the entire simulation. The center-of-mass translation was removed every 1.0 ps and the coordinates were also saved every 20 ps resulting in 5000 coordinates. Since velocities were not saved during the MD simulations, velocities for structures obtained through exchange with the reservoir were assigned by evaluating the forces during the subsequent MD step.

5.3.5 Building reservoir for RREMD simulation

The 119.34 μs MD simulation trajectory at 360 K was used to select structures to build the reservoir. A total of 5967 structures were selected with an equal time spacing of 20 ns resulting in a Boltzmann-weighted reservoir.

The energy for each structure was calculated using the imin=5 flag in *sander* program in AMBER using the same topology file and energy parameters as used in MD. Finally, reservoir was built using the *createreservoir* command in *cpptraj* program in AMBER using a seed of 1.

5.3.6 Analysis

5.3.6.1 Calculating Fraction of Native Structures

A structure was characterized as native if the RMSD of residues 3 to 18 was <2.0 Å, consistent with our previous studies on Trp-cage[8, 12, 21].

5.4 Results and Discussions

Since it is hard to obtain reference data using standard REMD for explicit solvent simulations due to the reasons cited in **Section 5.2**, to check if RREMD simulations can accelerate conformational sampling in explicit solvent, extensive MD simulations (for >100 μs) were performed at two different high temperatures (see **Methods**) to generate converged reference data.

5.4.1 Are the high temperature MD simulations in explicit solvent converged?

Even in explicit solvent, running the simulations at high temperatures for a long time must help in overcoming quasi-ergodicity. To check this, RMSD vs. time for each of the high temperature MD simulations are shown in **Figure 5-4**. At both temperatures, multiple transitions between native and non-native structures are observed indicating that the sampling might be sufficient.

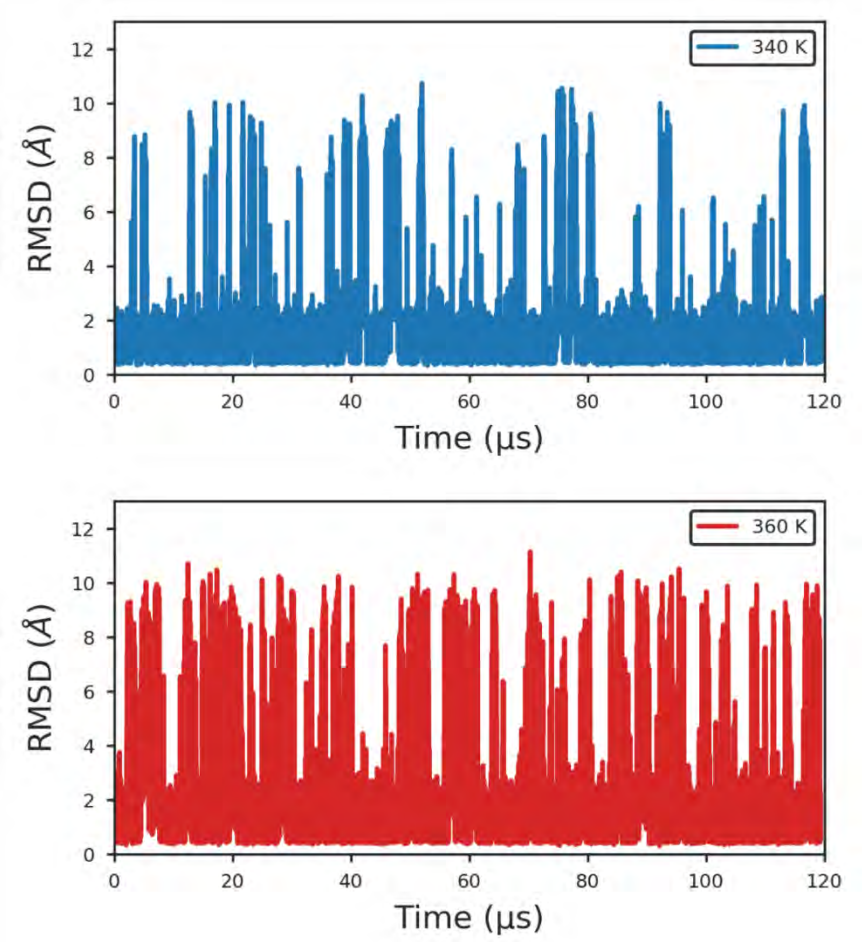


Figure 5-4 The RMSD as a function of time for explicit solvent simulations of Trp-cage variant Tc10b at 340 K (top), and 360 K (bottom) are shown. Multiple folding and unfolding events are observed at each temperature. This is the same data as shown in **Figure 5-3** but including the data up to 119.6 μs and 119.34 μs, at 340 K and 360 K, respectively.

To further check the convergence of the high temperature MD simulations, the fraction of native structures as a function of time was calculated (**Figure 5-5).** At both temperatures, the fraction of native structures plateaus around 50 μs to 60 μs indicating that the simulations are well converged. Surprisingly, the simulations of Trp-cage at 340 K do not take longer to converge compared to simulations at 360 K. Looking at cluster populations might reveal differences in convergence speeds; however, such analysis is not performed here and will be looked at in future.

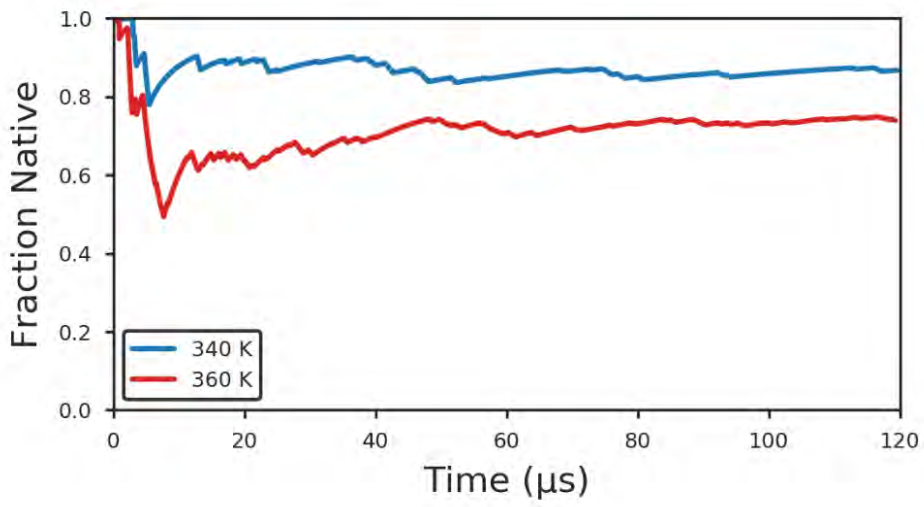


Figure 5-5 Fraction of native structures at the two different temperatures are shown.

5.4.2 Can RREMD accelerate conformational sampling in explicit solvent simulations?

To test this, the 360 K high temperature trajectory was used to select structures with equal time spacing to build the Boltzmann-weighted reservoir. (Ideally, similar to the protocols outlined in **Chapter 2**, the convergence for the reservoir generation simulation should be measured by comparing cluster populations between the first half and second half of the trajectory, however, we haven't done such analysis here.) Then, using this reservoir, B-RREMD simulation was performed for 100 ns per replica. The fraction of native structures as a function of time are shown

in **Figure 5-6**. Strikingly, the fraction of native structures at 340 K from the B-RREMD simulation matches excellently with the fraction of native structures obtained from the long MD simulation, after just 50 ns of simulation per replica resulting in around 1000x speed up (without accounting for reservoir generation time) compared to standard MD simulation. This indicates that RREMD can, indeed, accelerate conformational sampling in explicit solvent significantly.

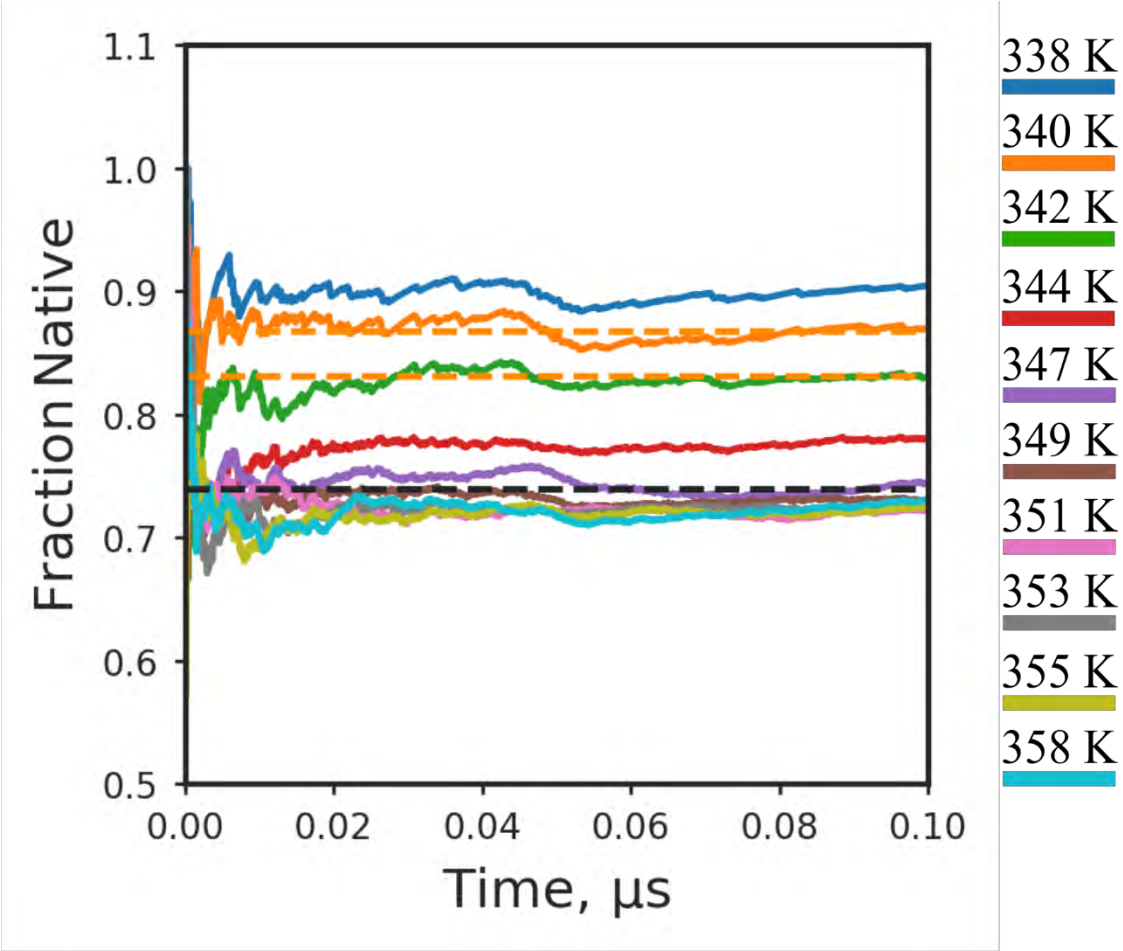


Figure 5-6 Fraction of native structures as a function of time for the RREMD simulation at each temperature are shown. The solid lines indicate the fraction of native structure obtained from B-RREMD simulation. The color codes for the temperature are shown to the right of the graph. The dashed black line and the dashed orange line indicate the fraction of native structures obtained from the 119.34 μs simulation at 360 K, and 119.6 μs simulation at 340 K, respectively. Note that the Y-axis ranges from 0.5 to 1.1 instead of 0 to 1. Also note that the fraction of native structures at temperatures 349 K to 358 K are similar to each other. This is because of the absence of velocities in the reservoir (data not shown).

While this speed up is remarkable, generating structures for building the Boltzmann-weighted reservoir still takes a long time (at least 3 months on a GTX 1080). This problem will only be exacerbated for bigger proteins than Trp-cage. Moreover, if the accuracy of a different force field or explicit solvent model has to be predicted, then MD simulations will have to be performed with the different force field and the explicit solvent model to generate canonical ensembles, which is highly prohibitive. Therefore, in the next chapter, we explore alternative ways of generating structures for building reservoirs to accelerate explicit solvent biomolecular simulations.

5.5 Conclusions

In this chapter, it was shown that B-RREMD simulations can accurately reproduce the data obtained from long MD simulations at baseline temperature. Moreover, B-RREMD simulations resulted in converged ensembles within 50 ns indicating compared to 50 μs using standard MD indicating a speed up of around 1000x (excluding reservoir generation time).

However, it was also shown that it takes a long time to generate a Boltzmann-weighted reservoir in explicit solvent simulations, even at high temperature. Therefore, in the next chapter, we explore protocols to reduce the time required to generate structures for building reservoirs for explicit solvent simulations.

6 Accelerating Explicit Solvent Simulations by using Structure Reservoirs generated from Implicit Solvent Simulations

6.1 Abstract

In the previous chapter, it was shown that Boltzmann-weighted reservoirs can be used to reproduce the fraction of native structures obtained from long MD simulations accurately. However, the sheer cost of generating Boltzmann-weighted reservoirs in explicit solvent means that the method results in only a slight improvement in sampling efficiency compared to standard MD simulations. In this chapter, implicit solvent simulations are used to reduce the time required to generate structures, which are then minimized and equilibrated in explicit solvent to generate explicit solvent reservoirs. These explicit solvent reservoirs were then coupled to REMD via a non-Boltzmann exchange scheme. The results indicate that such reservoirs can be used to reasonably reproduce ensembles obtained from long MD simulations.

6.2 Introduction

In **Chapter 4**, structure reservoirs generated from one force field were used to accurately predict the structural preferences of a different force field. However, this application of nB-RREMD was tested in only implicit solvent simulations. Since rapidly obtaining converged ensembles in explicit solvent is the holy grail of sampling methods, in this chapter, the applicability of Hamiltonian switching via nB-RREMD to predict the structural preferences of explicit solvent simulations is explored.

To test the applicability of Hamiltonian switching via nB-RREMD for explicit solvent simulations, ideally, structure reservoirs must also be generated using explicit solvent simulations. However, as mentioned in **Chapter 5**, generating structures, even for building a non-Boltzmann reservoir, is extremely slow due to solvent viscosity. On the other hand, implicit solvent models, despite their limited accuracy, sample the conformational space significantly faster than explicit solvent simulations[62, 114]. Since structures from many different sources can be used to build a non-Boltzmann reservoir, using implicit solvent simulations to generate structures and then thermally reweighting them using explicit solvent nB-RREMD simulations might be a viable solution to quickly obtain converged ensembles in explicit solvent.

Therefore, in this chapter, protocols for using structure reservoirs generated from implicit solvent simulations to predict the accuracy of explicit solvent simulations are explored. The results indicate that structure reservoirs from implicit solvent simulations can be used to qualitatively predict the accuracy of explicit solvent simulations, however, predicting the accuracy quantitatively will require further research.

6.2.1 Potential issue with using non-Boltzmann reservoirs for explicit solvent simulations

In the B-RREMD[32] simulation described in the previous chapter, since the reservoir is Boltzmann-weighted, multiple structures are present with similar solute conformations but different water conformations i.e., the effects of the water are "averaged out".

nB-RREMD simulations[49], on the other hand, use only one structure per each minimum. Using such reservoirs is reasonable for implicit solvent simulations since there is a one-to-one mapping between the conformation of the solute and the overall potential energy. Indeed, in **Chapter 2**, we have shown that the accuracy of nB-RREMD is insensitive to the energy used to build the reservoir. However, this conclusion might not hold true for nB-RREMD simulations in explicit solvent.

In explicit solvent simulations, the huge number of water molecules affect this one-to-one mapping between the solute conformation and the overall potential energy of the system. For example, the overall potential energy of the system can be similar even when the solute conformations are drastically different due to similar conformations of water molecules, and vice versa. Due to the above reason, using the potential energy of a single conformation might not reflect the "true" potential energy of a given minimum, which might significantly affect the MC exchanges with the reservoir.

The above potential issue can be fixed in three ways:

(1) Using multiple structures per cluster: Reservoirs were built (see below) with single structure per cluster, and also with multiple structures per cluster. The multiple structures per cluster were obtained by equilibrating the single structure with positional restraints and taking snapshots from this equilibrated run. This way, multiple structures

in the reservoir will correspond to the same solute structure but multiple water conformations, essentially "averaging out" the effects of the water molecules in the exchange criterion.

(2) Using a hybrid-solvent exchange scheme: The effects of the water can also be "averaged out" by using a Hybrid-solvent exchange scheme[37-38, 123], in which exchanges with the reservoir and between the replicas are attempted by solvating the solute in an implicit solvent. The accuracy of such a scheme was also tested here by building reservoirs (see below) using the same conformations as the fully solvated system but with energies corresponding to only the solute solvated in GB.

(3) Using average energy of each cluster: Using the average energy of each cluster might also help.

In this chapter, only the first two ways are addressed. The third way can be looked into in future.

6.3 Methods

6.3.1 General Details

The Trp-cage variant Tc10b[111] (PDB ID: 2JOF) was considered for this study. All structures were built via the LEaP module of AmberTools in the AMBER 18 package[56].

6.3.2 Implicit solvent simulation to generate structures for reservoir

Two initial conformations were built – (1) Native conformation, for which the first NMR model was used, and (2) Extended conformation, in which φ, ψ angles for all residues except Proline were set to 180°, proline residues were set to φ=-61.5°, ψ=-176.6°. The force field ff14SBonlysc[10] was used for all implicit solvent simulations. The GB-Neck2[8] (igb=8 in AMBER)

implicit solvent model with mbondi3 radii set was used for all simulations. No cutoff was used for calculation of non-bonded interactions. Langevin thermostat with a collision frequency of 1 ps^{-1} was used for all simulations. SHAKE[44] was performed on all bonds including hydrogen with the AMBER default tolerance of 0.00001 Å.

6.3.2.1 Minimization and Equilibration

A time step of 1 fs was used for all MD simulations during equilibration. With 10 $kcal.mol^{-1}.Å^{-2}$ positional restraints on all heavy atoms, the structures built using LEaP were minimized for 1000 cycles using steepest descent and then heated from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. Then, with 10 $kcal.mol^{-1}.Å^{-2}$ positional restraints on only backbone heavy atoms, the structures were again minimized for 1000 cycles using steepest descent and then heated again from 100 K to 300 K for 250 ps followed by another 250 ps at 300 K. This was followed by 500 ps of MD at 300 K with 1 $kcal.mol^{-1}.Å^{-2}$ positional restraints on backbone heavy atoms and then another 500 ps of MD at 300 K with 0.1 $kcal.mol^{-1}.Å^{-2}$ positional restraints on backbone heavy atoms. Finally, 5 ns of unrestrained MD was performed at 300 K.

6.3.2.2 Molecular Dynamics simulations

For each conformation, MD simulations were performed for 2 μs at 342.2 K, starting from both native and extended conformations. Chirality constraints and *trans*-peptide ω constraints obtained using *makeCHIR_RST* program in AMBER were used at all temperatures to prevent chirality inversions and peptide bond flips. A time step of 2 fs was used for all simulations. Coordinates were saved every 20 ps.

6.3.3 Building explicit solvent reservoirs using structures obtained from implicit solvent

In brief, clustering was performed to select structure representatives corresponding to each minimum. Then, these structure representatives were solvated, minimized, and equilibrated, in explicit solvent. The equilibrated structures were then used to build reservoirs. The protocols for clustering, minimization, and equilibration, are provided below.

6.3.3.1 Selecting structure representatives from implicit solvent simulations using clustering

From the combined MD trajectories (totaling a time of 4 μs) starting from the two different initial conformations, 10000 structures were extracted using an equal time spacing of 400 ps. The entire backbone RMSD was used as the clustering metric and KMeans[83] clustering algorithm was used. The target number of clusters were set to 500. A seed of 23 was used to randomize initial set of points used. For each cluster, the structure with the lowest cumulative distance to all other structures within that cluster was chosen as the cluster representative. The representative structures of only the top 50 clusters were selected for building reservoirs. These representative structures were then solvated using the same number of water molecules for each explicit water model (see below).

6.3.3.2 Minimization and equilibration

The minimization and equilibration described below is similar to the protocols used in the previous chapter. However, there are **two things to note:** (1) The temperature for equilibration was set to 360.0 K since the final reservoir will be at 360.0 K. This way the equilibrated reservoir structures will have the appropriate thermal energy for efficient exchanges with the highest temperature replica. (2) Since the reservoir structures correspond to different minima, positional

restraints on the backbone heavy atoms are always applied during the minimization and equilibration process to prevent the representative structures from transitioning to a different minimum.

For each of the solvated structures, minimization and equilibration were performed with a weak-coupling (Berendsen[136]) thermostat and barostat targeted to 1.0 bar with isotropic position scaling as follows. With 100.0 kcal mol^{-1} $Å^{-2}$ positional restraints on peptide heavy atoms, structure was minimized for up to 10000 cycles and then heated at constant volume from 100.0 K to 360.0 K over 100.0 ps, followed by another 100.0 ps at 360.0 K. The pressure was equilibrated for 100.0 ps and then 250.0 ps with time constants of 100.0 fs and then 500.0 fs on coupling of pressure and temperature to 1.0 bar and 360.0 K, with 100.0 kcal mol^{-1} $Å^{-2}$ and then 10.0 kcal mol^{-1} $Å^{-2}$ Cartesian positional restraints on peptide heavy atoms. The system was again minimized, with 10.0 kcal mol^{-1} $Å^{-2}$ force constant Cartesian restraints on only the protein main chain N, Cα, and C for up to 10000 cycles. Then, a 100.0 ps simulation with temperature and pressure time constants of 500.0 fs was performed, with backbone restraints of 10.0 kcal mol^{-1} $Å^{-2}$. Then, another 100.0 ps simulation with temperature and pressure time constants of 1.0 ps was performed using a time step of 1 fs, with backbone restraints of 10.0 kcal mol^{-1} $Å^{-2}$. This was followed by another 100.0 ps simulation with the same parameters as previous step but with a 2 fs time step. Finally, the system was simulated with pressure and temperature time constants of 1.0 ps for 1.0 ns with a 4.0 fs time step (using hydrogen-mass repartitioning), removing center-of-mass translation every 2.0 ps.

6.3.3.3 Adjusting the box sizes of the equilibrated structures to be the same

Equilibration in NPT ensemble is essential to remove the gaps that are introduced when the system is built initially. However, equilibration in NPT ensemble also changes the dimensions of the truncated octahedron unit cell leading to different box dimensions for each equilibrated

structure. Since nB-RREMD simulations are performed in NVT ensemble, it is essential to make sure all the equilibrated structures have the same box dimensions to avoid simulation artifacts such as “flying” water molecules.

To ensure that the box dimensions of all the equilibrated structures are the same, the box dimensions of each of the equilibrated structures were read from the NetCDF restart files using the *ncdump* command. Then, the maximum box dimension out of all the box dimensions was saved. Finally, *ncgen* command was used to regenerate the restart files with the previous coordinates but with the same maximum box dimensions used for every equilibrated structure.

6.3.3.4 Final equilibration in NVT ensemble

Since the box dimensions were modified in the previous step, another 10 ns equilibration was performed in NVT ensemble to allow the molecules to adjust to the new dimensions, and also to sample different water conformations for a given protein conformation (see below). The input variables for this equilibration were the same as the last step of NPT equilibration except that the pressure coupling is turned off. Velocities were also saved during this equilibration.

6.3.3.5 Combining the solvated cluster representatives and the cluster energies to build non-Boltzmann reservoirs

For each explicit solvent model (see below), 6 (3*2) reservoirs were built using the NVT equilibrated structures. Each of these reservoirs differ in the number of structures per cluster, and the energy used to build the reservoir. The number of structures per cluster were either 1, 20, or 50. If 1 structure per cluster was selected, this was the restart file obtained at the end of the NVT equilibration. If multiple (20 or 50) structures were selected, these were equally spaced structures selected from the last 5 ns of the NVT equilibration. Finally, either the potential energy of the full system was used to build the reservoir (see **nB-RREMD simulations** below) or the potential

energy of only the solute solvated in GB (see **hybrid-nB-RREMD simulations** below) was used to build the reservoir.

The energy of each structure in each reservoir was calculated using the imin=5 flag in *sander* program with the corresponding input flags used for nB-RREMD and hybrid-nB-RREMD simulations. Finally, reservoirs were built using the *createreservoir* command in *cpptraj*[88] program in AMBER using a seed of 1. These reservoirs also contain velocity information for each structure.

6.3.4 nB-RREMD simulations

6.3.4.1 General Details

The ff14SB[10] force field was used for all simulations using TIP3P[2] water model. The ff19SB[13] force field was used for all simulations using OPC[9] water model. Non-bonded interactions were calculated directly up to 8.0 Å with cubic spline switching and PME approximation with direct sum tolerances of 0.00001. SHAKE[44] was performed on all bonds including hydrogen with the AMBER default tolerance of 0.00001 Å.

To allow for large conformational changes, a large box with 9471 and 10532 water molecules were used for simulations using TIP3P and OPC water models, respectively. Ideally, the number of water molecules between the two solvent models should be the same so that the two simulations systems have the same chemical potential. However, the difference between the melting curves (obtained from nB-RREMD simulations using top 50 clusters with 50 structures per cluster) using 9471 TIP3P water molecules and 10532 TIP3P water molecules was insignificant (data not shown), and hence the simulations for 9471 water molecules were not performed again.

6.3.4.2 Replica Exchange simulations

To ensure that the box dimensions of the REMD simulations are the same as the reservoirs, the restart file of the first cluster after NVT equilibration was used as the starting structure for each replica. This structure was equilibrated for 50 ps at the replica target temperatures using a time step of 1 fs. Then, nB-RREMD simulations were performed for 100 ns per replica with exchanges attempted every 1 ps. The lowest replica temperature was 340 K since we generated individual replica temperatures were 340.0 K, 342.2 K, 344.3 K, 346.5 K, 348.7 K, 351.0 K, 353.2 K, 355.4 K, 357.7 K, and 360.0 K.

6.3.5 Hybrid-nB-RREMD simulations

Hybrid-nB-RREMD simulations used the same setup as above with three differences: (1) All exchanges (including exchanges with reservoir and exchanges between replicas) were attempted using the energy of the protein solvated in GBNeck2[8]. Only the polar component of the solvation energy was included during exchanges. (2) Only 4 replicas were used with the replica temperatures set to 340.0 K, 346.5 K, 353.2 K, and 360.0 K. (3) The number of water molecules for TIP3P simulations were adjusted from 9471 to 10532. Again, this change in number of water molecules was only done to keep the chemical potential the same across different simulations. The results were not significantly affected due to the change in number of water molecules (data not shown).

6.3.6 Analysis

6.3.6.1 Melting curves

Temperature-based trajectories were extracted from REMD and RREMD simulations using the *cpptraj* program in AMBER. The fraction of native structures was calculated for each temperature-based trajectory. A structure was characterized as native if the backbone RMSD of

residues 3 to 18 was <2.0 Å to the first NMR structure. Since the RREMD simulations in explicit solvent converged quickly (see **Chapter 5**), all 100 ns of data was used to calculate the melting curves without excluding any data from the beginning of the simulations. The error bars indicate the half difference of the melting curves obtained from first 50 ns and last 50 ns.

6.4 Results and Discussions

Using implicit solvent simulations can reduce the time required to generate conformations to build a reservoir. However, the MC exchange scheme in nB-RREMD simulations using explicit solvent can be sensitive to the energy used to build the reservoir. To overcome this issue, multiple structures per cluster can be used, or a hybrid-solvent exchange scheme can be used. Both of these approaches are explored below.

6.4.1 Generating reservoir structures using implicit solvent simulations

To reduce the time required to generate structure reservoirs for explicit solvent simulations, high temperature MD simulations were performed in implicit solvent (see **Methods**). Then, the trajectories were clustered using KMeans clustering algorithm using the protocols outlined in **Chapter 2**. Then, these structures were solvated, minimized, and equilibrated in explicit solvent. During equilibration, the backbone heavy atoms of the protein were restrained to prevent transitions of the cluster representatives from one minimum to another. Finally, the energies of these equilibrated structures were calculated and used to build the reservoir.

6.4.2 Sensitivity of nB-RREMD in explicit solvent to the number of structures per cluster?

To test the sensitivity of nB-RREMD simulations in explicit solvent to the number of structures per cluster, for each force field and explicit solvent water model, 3 reservoirs were built with different number of structures per cluster (see **Methods**). To distinguish between the 3 reservoirs, the following reservoir naming convention was used – nB_(N)_(n) indicates a non-Boltzmann reservoir using top "N" clusters, in this instance 50, and "n" structures per cluster. For example, nB_50_1 indicates that the top 50 clusters from GB were used and only 1 structure per cluster was used to build the reservoir. The melting curves obtained from nB-RREMD simulations using the 3 reservoirs are shown in **Figure 6-1**, for each force field and explicit solvent model combination.

If explicit solvent nB-RREMD simulations are sensitive to the number of structures, the melting curves obtained from reservoirs with different number of structures should be different. Surprisingly, when ff14SB force field and TIP3P water model combination is used, irrespective of the number of structures per cluster used to build the reservoir, nB-RREMD simulations of Trp-cage result in the similar melting curves compared to each other. Also, the fraction of native structures (0.8) at 340 K is close to the reference value (0.86) obtained from 119.6 μs long MD simulations.

On the other hand, when ff19SB force field and OPC water model combination is used, the number of structures used per cluster affects the final melting curves significantly. The fraction of native structures (0.25) at 340 K matches with the reference data (0.25) only when 1 structure per cluster is used while the other two simulations are significantly more stable, >0.6 fraction of native structures at 340 K. However, the match with reference data when only 1 structure per cluster is

used could be due to coincidence (maybe “good” potential energies) and needs to be looked into in future.

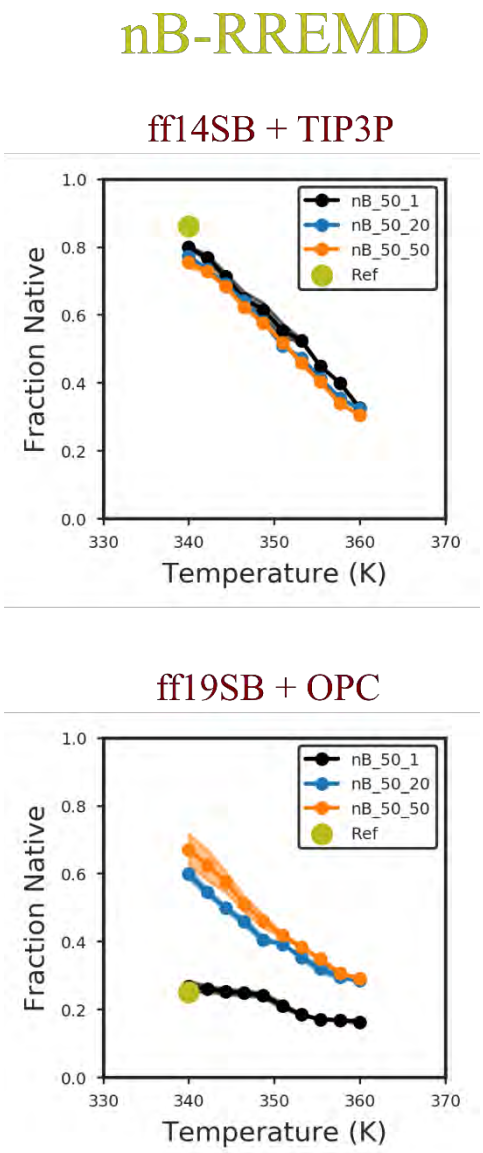


Figure 6-1 Melting curves obtained from nB-RREMD simulations using the 3 reservoirs (see text), for each force field and explicit solvent model. “Ref” indicates the fraction of native obtained from standard MD simulation for each force field. The “Ref” data for ff19SB+OPC is from Dr. Chuan Tian’s unpublished work. The “Ref” data for ff14SB+TIP3P is from **Chapter 5.** The error bars indicate the half difference between the melting curves obtained from the first 50 ns and the last 50 ns of the trajectory.

Nonetheless, since the same set of solute structures were used to build the reservoirs for nB-RREMD simulations, the decrease in protein stability using ff19SB+OPC compared to ff14SB+TIP3P indicates that nB-RREMD can capture the differences between force fields and explicit solvent models qualitatively.

Considering that the implicit solvent simulations used to generate the structures required only a day of simulation on a GTX 1080, and the nB-RREMD simulations required only 6 hours on GTX 1080s, the ability to qualitatively predict the preference of the different force fields and explicit solvent combinations is remarkable. In comparison, obtaining the reference data for Trp-cage using standard MD requires at least 3 months on a GTX 1080.

6.4.3 Can Hybrid-nB-RREMD qualitatively reproduce the ensembles obtained from long MD simulations in explicit solvent?

The differences in the melting curves of nB-RREMD simulations with reservoirs using different number of structures per cluster when the ff19SB+OPC combination is used indicates that the energies of the structures in the reservoir might influence the final ensembles of nB-RREMD simulations. Therefore, Hybrid-nB-RREMD simulations were performed to test if "averaging out" the water with a continuum solvent can result in accurate conformational ensembles. The melting curves obtained from nB-RREMD simulations using the 3 reservoirs are shown in **Figure 6-2**, for each force field and explicit solvent model combination.

Using a hybrid-solvent exchange scheme should nullify the effect of using multiple structures per cluster. The melting curves in **Figure 6-2** show that this is indeed the case – all three reservoirs for each force field and solvent model combination result in the same melting curves.

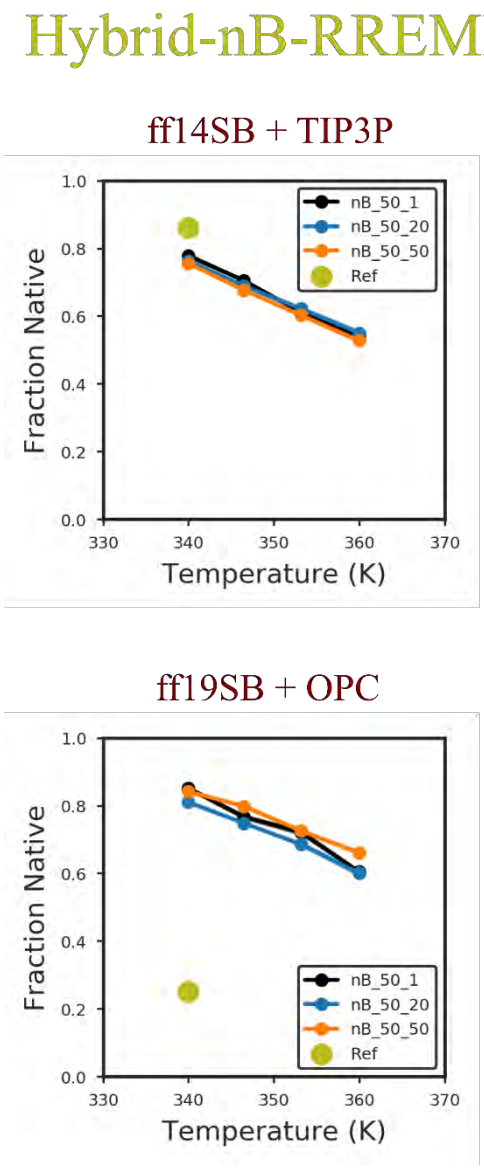


Figure 6-2 Melting curves obtained from Hybrid-nB-RREMD simulations using the 3 reservoirs (see text), for each force field and explicit solvent model. "Ref" indicates the fraction of native obtained from standard MD simulation for each force field. The "Ref" data for ff19SB+OPC is from Dr. Chuan Tian's unpublished work. The "Ref" data for ff14SB+TIP3P is from **Chapter 5.** The error bars indicate the half difference between the melting curves obtained from the first 50 ns and the last 50 ns of the trajectory.

However, they also result in the same melting curves across different force fields and solvent models indicating the implicit solvent affects the final ensembles that will be obtained using Hybrid-nB-RREMD simulations. This could be due to frequent exchanges with the reservoir.

Exchanging less frequently with the reservoir might help structures adapt to the new Hamiltonian and needs to be addressed in future. Alternatively, using solvent molecules in the first shell and second shell might be a better representative of the underlying energetics and can be addressed in the future.

6.5 Conclusions

In this chapter, protocols for, and potential issues of, using implicit solvent simulations to generate structure reservoirs for explicit solvent simulations were outlined. Using these protocols, structural preferences of force fields in explicit solvent can be modeled qualitatively.

Further work needs to be carried out to explore the sensitivity of nB-RREMD simulations to not just the number of structures per cluster used but also to the overall number of clusters used. The role of solvent molecules in the first and second shell during hybrid-nB-RREMD also needs to be explored in the future. Finally, using average energy of each cluster might also help in reproducing the ensembles obtained from long MD explicit solvent simulations, and this needs to be explored in future as well.

For all of these future works, it is crucial to compare not just the fraction of native conformations but also the cluster populations obtained from nB-RREMD and long MD simulations in explicit solvent. Eventually, standard REMD simulations in explicit solvent must be used to generate reference data at a wide (~300 K – ~400 K) range of temperatures for a variety of solutes to thoroughly test the protocols mentioned in this chapter (see next chapter).

7 Future Directions

In this thesis, protocols for building Boltzmann-weighted reservoirs and non-Boltzmann reservoirs were explored were in **Chapters 2** and **3.** Using these protocols, we showed that RREMD simulations can significantly accelerate the rate of conformational sampling in biomolecular simulations using implicit solvent. In **Chapter 4**, we showed how nB-RREMD can be used to predict the accuracy of force fields and also the effects of mutations on peptide stability. In **Chapters 5**, we showed that Boltzmann-weighted RREMD can be used to accelerate the convergence of explicit solvent simulations by at least 1000x, however, generating Boltzmann-weighted reservoirs took a long time. In **Chapter 6**, structures from implicit solvent were used to build the reservoirs for explicit solvent simulations. These reservoirs can be used to qualitatively predict the accuracy of different force fields and explicit solvent models. Overall, in this thesis work, structure reservoirs were used to accelerate sampling in biomolecular simulations significantly.

The above developments notwithstanding, there is still significant scope for both, improving the methodology for building reservoirs, and widening the potential areas of applicability of RREMD methods. These outstanding issues and future directions are addressed below.

7.1 Reducing the time required to generate reservoirs

In this thesis, high temperature MD simulations were used to generate structures for building both Boltzmann-weighted and non-Boltzmann reservoirs. As mentioned in **Chapter 2**, high temperature MD simulations are upper bounds for the time required to generate reservoir structures. It should be possible to generate Boltzmann-weighted reservoirs by using REMD with two to four replicas. Having multiple walkers might further improve the rate of sampling at high temperatures.

While sampling methods that produce canonical ensembles have to be used for building Boltzmann-weighted reservoirs, non-Boltzmann reservoirs do not have this restriction. This means that many different sampling methods such as accelerated MD could be used to generate conformations. Accelerated MD (aMD)[28] and its variants add boost potentials to accelerate sampling. However, removing the bias introduced by the boost potentials is challenging. Nonetheless, since correctly weighted ensembles are not required for non-Boltzmann reservoirs, the boost potentials in aMD can be used to quickly generate the different backbone conformations for a given protein. These different conformations can then be thermally reweighted using nB-RREMD.

As an alternative to enhanced sampling methods, coarse grained simulations[137] could also be used to quickly generate reservoir structures. Due to the absence of hydrogen atoms, time steps as long as 20 fs can be used in coarse grained simulations. This way large scale structural changes can be modeled quickly. These coarse grained structures can then be back mapped[138] to fully atomistic models which can then be used for nB-RREMD simulations.

Homology modeling can also be used to generate structures that can be used in nB-RREMD reservoirs. Current homology modeling tools such as Rosetta[76] take around 10-20

minutes to generate a single snapshot for proteins that are around 80-100 amino acids long. Usually, 5000-10000 structures have to be generated using Rosetta to sample the conformational space adequately. While this indicates it might take weeks to months to generate different conformations, the highly parallelizable nature of modeling using Rosetta means that using 100 processors can result in 10000 structures in less than 2 days. This is faster than the simulation time required (around 2 weeks) to generate non-Boltzmann reservoirs for Homeodomain.

Since Rosetta uses empirical energy functions to score structures, the ranking of minima might not be accurate. However, reweighting using nB-RREMD can re-rank the different minima. Consequently, using Rosetta to generate structures and reweighting the structures using nB-RREMD can accelerate testing of physics-based force fields.

Lastly, simulation programs such as mdgx allow multiple copies (8 for Homeodomain sized protein in implicit solvent) of proteins to be simulated on a single GPU. This way, multiple different trajectories can be used to build reservoir structures instead of a single MD simulation, resulting in significant reduction in reservoir generation times.

7.2 Applications of RREMD simulations

7.2.1 Predicting accuracy of force fields

We have shown that structure reservoirs can be used to predict the accuracy of force fields with significantly different energy landscapes. However, wide-spread applicability of nB-RREMD to predict force field accuracies will require studies using bigger proteins than Trp-cage. Fip35 and HP36 might serve as good test systems for testing this approach further.

7.2.2 Predicting effects of mutations

nB-RREMD simulations were also used to predict the effects of mutations. However, alchemical free energy calculations can also be used for this purpose. Nonetheless, large scale conformational changes (or too large a chemical perturbation) can introduce significant uncertainties into free energy calculations. For this reason, methods such as TI[108] are mostly used to predict the accuracy of only one or two mutations per thermodynamic cycle. Reservoir methods, on the other hand, can be used to study the effects of multiple mutations in a single run as long the mutations do not result in significant steric clashes.

However, the applicability of nB-RREMD to predict the accuracy of force fields and effects of mutations assumes that the relevant basins are already present in the reservoir.

7.3 Super-reservoirs are the way to go

To overcome the lack of sampling from a particular force field, super-reservoirs that combine structures from different sources can be used. While completely exhausting the conformational space for a given protein is impossible, super-reservoirs can at least represent all the relevant basins that the protein might sample. Since these super-reservoirs are usually small and contain all the relevant conformations for a given sequence, these can be easily distributed to labs across the world. This way, different labs from across the world can use the same set of structures to quantify the accuracy of force fields. Moreover, if any new structures are observed, these can be easily integrated into the super-reservoir. This way the conformational space can be gradually exhausted.

To use an analogy from computer science, super-reservoirs (and reservoirs in general) are like compressed folders containing the markers for the accurate description of the biomolecule, and nB-RREMD is the *winzip* (for people old enough to remember this) software used to

uncompress the markers. Improving the accuracy of biomolecular simulations using RREMD is then, similar to reinforcement learning.

7.4 RREMD for explicit solvent simulations

As mentioned before, the holy grail of most sampling methods is to reduce the time required for generating converged ensembles in explicit solvent. In this work, we used two very long MD simulations as reference data to validate RREMD protocol for explicit solvent simulations. This was done because the large number of replicas required to span the desired temperature range prohibited us from using T-REMD. However, eventually, RREMD has to be compared to T-REMD simulations. In my experience, 40 to 60 replicas have to be simulated for around 3-4 months to generate converged ensembles in explicit solvent. Doing such simulations requires access to supercomputers. Nonetheless, this should be addressed eventually. If it can be shown that RREMD simulations can reliably predict ensembles obtained using standard REMD simulations, then it can eventually replace REMD altogether. However, one should always remember that REMD can sample new conformations and fix itself whereas RREMD cannot.

Also, the number of clusters that have to be selected from implicit solvent simulations, the number of structures per cluster, and the energy used to build the reservoir are choices that still have to be addressed through a more detailed study. While there isn't a big difference between using 20 structures per cluster or 50 structures per cluster, the different melting curves for ff19SB+OPC indicate that including multiple structures per cluster does influence the ensembles obtained from nB-RREMD simulations. Moreover, in my experience (data not shown), the number of clusters selected from implicit solvent plays a crucial role in determining the final ensembles.

Lastly, hybrid-nB-RREMD simulations have to be performed by including structured water or by including nonpolar contributions to the solvation free energy. The influence of these choices has to be thoroughly tested.

8 References

1. Schrödinger, E., *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press: 1944.

2. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983,** *79* (2), 926-935.

3. Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T., Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **2004,** *120* (20), 9665-78.

4. 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* (2), 383-94.

5. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.

6. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A., Generalized Born model with a simple, robust molecular volume correction. *J Chem Theory Comput* **2007,** *3* (1), 156-169.

7. Shang, Y.; Nguyen, H.; Wickstrom, L.; Okur, A.; Simmerling, C., Improving the description of salt bridge strength and geometry in a Generalized Born model. *J Mol Graph Model* **2011,** *29* (5), 676-84.

8. Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J Chem Theory Comput* **2013,** *9* (4), 2020-2034.

9. Izadi, S.; Anandakrishnan, R.; Onufriev, A. V., Building Water Models: A Different Approach. *J Phys Chem Lett* **2014,** *5* (21), 3863-3871.

10. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015,** *11* (8), 3696-713.

11. Nguyen, H.; Perez, A.; Bermeo, S.; Simmerling, C., Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. *J Chem Theory Comput* **2015,** *11* (8), 3714-28.

12. Huang, H.; Simmerling, C., Fast Pairwise Approximation of Solvent Accessible Surface Area for Implicit Solvent Simulations of Proteins on CPUs and GPUs. *J Chem Theory Comput* **2018,** *14* (11), 5797-5814.

13. Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C., ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J Chem Theory Comput* **2020,** *16* (1), 528-552.

14. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C., Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* **2008,** *51* (7), 91-97.

15. Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L. S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y. H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Ben Schafer, U.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C., Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *Int Conf High Perfor* **2014**, 41-53.

16. Zou, J.; Tian, C.; Simmerling, C., Blinded prediction of protein-ligand binding affinity using Amber thermodynamic integration for the 2018 D3R grand challenge 4. *J Comput Aided Mol Des* **2019,** *33* (12), 1021-1029.

17. Galiano, L.; Ding, F.; Veloro, A. M.; Blackburn, M. E.; Simmerling, C.; Fanucci, G. E., Drug pressure selected mutations in HIV-1 protease alter flap conformations. *J Am Chem Soc* **2009,** *131* (2), 430-1.

18. Childers, M. C.; Daggett, V., Insights from molecular dynamics simulations for computational protein design. *Mol Syst Des Eng* **2017,** *2* (1), 9-33.

19. Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A., Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol* **2018,** *44*, 1-8.

20. Zhu, J.; Hoop, C. L.; Case, D. A.; Baum, J., Cryptic binding sites become accessible through surface reconstruction of the type I collagen fibril. *Sci Rep* **2018,** *8* (1), 16646.

21. Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C., Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J Am Chem Soc* **2014,** *136* (40), 13959-62.

22. Garcia, H. G.; Kondev, J.; Orme, N.; Theriot, J. A.; Phillips, R., Thermodynamics of biological processes. *Methods Enzymol* **2011,** *492*, 27-59.

23. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* **2011,** *334* (6055), 517-520.

24. Torrie, G. M.; Valleau, J. P., Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *Journal of Computational Physics* **1977,** *23*, 187.

25. Hansmann, U. H. E., Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* **1997,** *281* (1-3), 140-150.

26. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **1999,** *314* (1-2), 141-151.

27. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc Natl Acad Sci U S A* **2002,** *99* (20), 12562-6.

28. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics* **2004,** *120* (24), 11919-11929.

29. Earl, D. J.; Deem, M. W., Parallel tempering: Theory, applications, and new perspectives. *Phys Chem Chem Phys* **2005,** *7* (23), 3910-3916.

30. Li, H. Z.; Li, G. H.; Berg, B. A.; Yang, W., Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces. *Journal of Chemical Physics* **2006,** *125* (14).

31. Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M., Resolution exchange simulation. *Phys Rev Lett* **2006,** *96* (2), 028105.

32. Okur, A.; Roe, D. R.; Cui, G. L.; Hornak, V.; Simmerling, C., Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir. *J Chem Theory Comput* **2007,** *3* (2), 557-568.

33. Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C., Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J Phys Chem B* **2007,** *111* (7), 1846-57.

34. Pietrucci, F., Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics* **2017,** *2*, 32-45.

35. Wang, A.-h.; Zhang, Z.-c.; Li, G.-h., Advances in enhanced sampling molecular dynamics simulations for biomolecules. *Chinese Journal of Chemical Physics* **2019,** *32* (3), 277-286.

36. Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q., Enhanced sampling in molecular dynamics. *J Chem Phys* **2019,** *151* (7), 070902.

37. Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C., Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J Chem Theory Comput* **2006,** *2* (2), 420-433.

38. Chaudhury, S.; Olson, M. A.; Tawa, G.; Wallqvist, A.; Lee, M. S., Efficient Conformational Sampling in Explicit Solvent Using a Hybrid Replica Exchange Molecular Dynamics Method. *J Chem Theory Comput* **2012,** *8* (2), 677-687.

39. Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J., Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proc Natl Acad Sci U S A* **2005,** *102* (39), 13749-54.

40. Huang, X.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R.; Berne, B. J., Replica exchange with solute tempering: efficiency in large scale systems. *J Phys Chem B* **2007,** *111* (19), 5405-10.

41. Wang, L.; Friesner, R. A.; Berne, B. J., Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J Phys Chem B* **2011,** *115* (30), 9431-8.

42. Smith, A. K.; Lockhart, C.; Klimov, D. K., Does Replica Exchange with Solute Tempering Efficiently Sample Aβ Peptide Conformational Ensembles? *J Chem Theory Comput* **2016,** *12* (10), 5201-5214.

43. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R., A Computer-Simulation Method for the Calculation of Equilibrium-Constants for the Formation of Physical Clusters of Molecules - Application to Small Water Clusters. *Journal of Chemical Physics* **1982,** *76* (1), 637-649.

44. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

45. Miyamoto, S.; Kollman, P. A., Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J Comput Chem* **1992,** *13* (8), 952-962.

46. Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E., Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2015,** *11* (4), 1864-1874.

47. Balusek, C.; Hwang, H.; Lau, C. H.; Lundquist, K.; Hazel, A.; Pavlova, A.; Lynch, D. L.; Reggio, P. H.; Wang, Y.; Gumbart, J. C., Accelerating Membrane Simulations with Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2019,** *15* (8), 4673-4686.

48. Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H., Designing a 20-residue protein. *Nat Struct Biol* **2002,** *9* (6), 425-430.

49. Roitberg, A. E.; Okur, A.; Simmerling, C., Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. *J Phys Chem B* **2007,** *111* (10), 2415-2418.

50. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **1953,** *21* (6), 1087-1092.

51. Kofke, D. A., On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics* **2002,** *117* (15), 6911-6914.

52. Rathore, N.; Chopra, M.; de Pablo, J. J., Optimal allocation of replicas in parallel tempering simulations. *Journal of Chemical Physics* **2005,** *122* (2).

53. Trebst, S.; Troyer, M.; Hansmann, U. H. E., Optimized parallel tempering simulations of proteins. *Journal of Chemical Physics* **2006,** *124* (17).

54. Patriksson, A.; van der Spoel, D., A temperature predictor for parallel tempering simulations. *Phys Chem Chem Phys* **2008,** *10* (15), 2073-7.

55. Sindhikara, D.; Meng, Y. L.; Roitberg, A. E., Exchange frequency in replica exchange molecular dynamics. *Journal of Chemical Physics* **2008,** *128* (2).

56. Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; III, T. E. C.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *AMBER 2018*, University of California: San Francisco, CA, 2018.

57. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y. T.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *Plos Comput Biol* **2017,** *13* (7).

58. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **1990,** *112* (16), 6127-6129.

59. Gilson, M. K.; Davis, M. E.; Luty, B. A.; Mccammon, J. A., Computation of Electrostatic Forces on Solvated Molecules Using the Poisson-Boltzmann Equation. *J Phys Chem-Us* **1993,** *97* (14), 3591-3600.

60. Huang, J.; MacKerell, A. D., CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* **2013,** *34* (25), 2135-2145.

61. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J Chem Theory Comput* **2015,** *11* (7), 3499-3509.

62. Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V., Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophys J* **2015,** *108* (5), 1153-1164.

63. Ruscio, J. Z.; Fawzi, N. L.; Head-Gordon, T., How Hot? Systematic Convergence of the Replica Exchange Method Using Multiple Reservoirs. *J Comput Chem* **2010,** *31* (3), 620-627.

64. Henriksen, N. M.; Roe, D. R.; Cheatham, T. E., Reliable Oligonucleotide Conformational Ensemble Generation in Explicit Solvent for Force Field Assessment Using Reservoir Replica Exchange Molecular Dynamics Simulations. *J Phys Chem B* **2013,** *117* (15), 4014-4027.

65. Okur, A.; Miller, B. T.; Joo, K.; Lee, J.; Brooks, B. R., Generating Reservoir Conformations for Replica Exchange through the Use of the Conformational Space Annealing Method. *J Chem Theory Comput* **2013,** *9* (2), 1115-1124.

66. Damjanovic, A.; Miller, B. T.; Okur, A.; Brooks, B. R., Reservoir pH replica exchange. *Journal of Chemical Physics* **2018,** *149* (7).

67. Honda, S.; Akiba, T.; Kato, Y. S.; Sawada, Y.; Sekijima, M.; Ishimura, M.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K., Crystal Structure of a Ten-Amino Acid Protein. *J Am Chem Soc* **2008,** *130* (46), 15327-15331.

68. Nadler, W.; Hansmann, U. H. E., Generalized ensemble and tempering simulations: A unified view. *Phys Rev E* **2007,** *75* (2).

69. Frantz, D. D.; Freeman, D. L.; Doll, J. D., Reducing Quasi-Ergodic Behavior in Monte-Carlo Simulations by J-Walking - Applications to Atomic Clusters. *Journal of Chemical Physics* **1990,** *93* (4), 2769-2784.

70. Zhou, R. H.; Berne, B. J., Smart walking: A new method for Boltzmann sampling of protein conformations. *Journal of Chemical Physics* **1997,** *107* (21), 9185-9196.

71. Andricioaei, I.; Straub, J. E.; Voter, A. F., Smart darting Monte Carlo. *Journal of Chemical Physics* **2001,** *114* (16), 6994-7000.

72. Opps, S. B.; Schofield, J., Extended state-space Monte Carlo methods. *Phys Rev E* **2001,** *63* (5).

73. Brown, S.; Head-Gordon, T., Cool walking: A new Markov chain Monte Carlo sampling method. *J Comput Chem* **2003,** *24* (1), 68-76.

74. Gallicchio, E.; Levy, R. M., Prediction of SAMPL3 host-guest affinities with the binding energy distribution analysis method (BEDAM). *J Comput Aid Mol Des* **2012,** *26* (5), 505-516.

75. Wickstrom, L.; He, P.; Gallicchio, E.; Levy, R. M., Large Scale Affinity Calculations of Cyclodextrin Host-Guest Complexes: Understanding the Role of Reorganization in the Molecular Recognition Process. *J Chem Theory Comput* **2013,** *9* (7), 3136-3150.

76. Song, Y.; DiMaio, F.; Wang, R. Y.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D., High-resolution comparative modeling with RosettaCM. *Structure* **2013,** *21* (10), 1735-42.

77. Lee, J.; Scheraga, H. A.; Rackovsky, S., New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J Comput Chem* **1997,** *18* (9), 1222-1232.

78. Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J Chem Theory Comput* **2007,** *3* (6), 2312-2334.

79. De Paris, R.; Quevedo, C. V.; Ruiz, D. D. A.; de Souza, O. N., An Effective Approach for Clustering InhA Molecular Dynamics Trajectory Using Substrate-Binding Cavity Features. *Plos One* **2015,** *10* (7).

80. Lemke, O.; Keller, B. G., Density-based cluster algorithms for the identification of core sets (vol 145, 164104, 2016). *Journal of Chemical Physics* **2016,** *145* (19).

81. Husic, B. E.; Pande, V. S., Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J Chem Theory Comput* **2017,** *13* (3), 963-967.

82. Peng, J. H.; Wang, W.; Yu, Y. Q.; Gu, H. L.; Huang, X. H., Clustering Algorithms to Analyze Molecular Dynamics Simulation Trajectories for Complex Chemical and Biological Systems. *Chinese Journal of Chemical Physics* **2018,** *31* (4), 404-420.

83. MacQueen, J. B., Some Methods for Classification and Analysis of MultiVariate Observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Cam, L. M. L.; Neyman, J., Eds. University of California Press: 1967; Vol. 1, pp 281-297.

84. Ward, J. H., Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* **1963,** *58* (301), 236-244.

85. Shah, P. S.; Hom, G. K.; Ross, S. A.; Lassila, J. K.; Crowhurst, K. A.; Mayo, S. L., Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol* **2007,** *372* (1), 1-6.

86. Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C., Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *J Am Chem Soc* **2014,** *136* (40), 13959-13962.

87. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015,** *11* (8), 3696-3713.

88. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-3095.

89. Bar-Joseph, Z.; Gifford, D. K.; Jaakkola, T. S., Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **2001,** *17* (suppl_1), S22-S29.

90. Müllner, D., Modern hierarchical, agglomerative clustering algorithms. *arXiv e-prints* **2011**, arXiv:1109.2378.

91. Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D. W.; Zuckerman, D. M., Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J Comput Mol Sci* **2018,** *1* (1).

92. van Gunsteren, W. F.; Daura, X.; Hansen, N.; Mark, A. E.; Oostenbrink, C.; Riniker, S.; Smith, L. J., Validation of Molecular Simulation: An Overview of Issues. *Angew Chem Int Ed Engl* **2018,** *57* (4), 884-902.

93. Pande, V. S.; Beauchamp, K.; Bowman, G. R., Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010,** *52* (1), 99-105.

94. Chodera, J. D.; Noe, F., Markov state models of biomolecular conformational dynamics. *Curr Opin Struc Biol* **2014,** *25*, 135-144.

95. Husic, B. E.; Pande, V. S., Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018,** *140* (7), 2386-2396.

96. Childers, M. C.; Daggett, V., Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design & Engineering* **2017,** *2* (1), 9-33.

97. Zou, J.; Song, B.; Simmerling, C.; Raleigh, D., Experimental and Computational Analysis of Protein Stabilization by Gly-to-D-Ala Substitution: A Convolution of Native State and Unfolded State Effects. *J Am Chem Soc* **2016,** *138* (48), 15682-15689.

98. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases. *Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.

99. Lin, Z. X.; Liu, H. Y.; Van Gunsteren, W. F., Using One-Step Perturbation to Predict the Effect of Changing Force-Field Parameters on the Simulated Folding Equilibrium of a beta-Peptide in Solution. *J Comput Chem* **2010,** *31* (13), 2419-2427.

100. Oostenbrink, C., Free Energy Calculations from One-Step Perturbations. *Methods Mol Biol* **2012,** *819*, 487-499.

101. Bachmann, S. J.; Dolenc, J.; van Gunsteren, W. F., On the use of one-step perturbation to investigate the dependence of different properties of liquid water on a variation of model parameters from a single simulation. *Mol Phys* **2013,** *111* (14-15), 2334-2344.

102. Lin, Z. X.; Oostenbrink, C.; van Gunsteren, W. F., On the use of one-step perturbation to investigate the dependence of NOE-derived atom-atom distance bound violations of peptides upon a variation of force-field parameters. *Eur Biophys J Biophy* **2014,** *43* (2-3), 113-119.

103. Diem, M.; Oostenbrink, C., Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field. *J Chem Inf Model* **2020,** *60* (1), 279-288.

104. Liu, H. Y.; Mark, A. E.; vanGunsteren, W. F., Estimating the relative free energy of different molecular states with respect to a single reference state. *J Phys Chem-Us* **1996,** *100* (22), 9485-9494.

105. Oostenbrink, C.; van Gunsteren, W. F., Efficient calculation of many stacking and pairing free energies in DNA from a few molecular dynamics simulations. *Chem-Eur J* **2005,** *11* (15), 4340-4348.

106. Hritz, J.; Oostenbrink, C., Efficient Free Energy Calculations for Compounds with Multiple Stable Conformations Separated by High Energy Barriers. *J Phys Chem B* **2009,** *113* (38), 12711-12720.

107. Li, D. W.; Bruschweiler, R., NMR-Based Protein Potentials. *Angew Chem Int Edit* **2010,** *49* (38), 6778-6780.

108. Kirkwood, J. G., Statistical mechanics of fluid mixtures. *Journal of Chemical Physics* **1935,** *3* (5), 300-313.

109. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* **2010,** *78* (8), 1950-1958.

110. Wang, L. P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martinez, T. J.; Pande, V. S., Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J Phys Chem B* **2017,** *121* (16), 4023-4039.

111. Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H., The Trp-cage: optimizing the stability of a globular miniprotein. *Protein Eng Des Sel* **2008,** *21* (3), 171-185.

112. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-1612.

113. Feig, M.; Brooks, C. L., Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr Opin Struc Biol* **2004,** *14* (2), 217-224.

114. Zagrovic, B.; Pande, V., Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J Comput Chem* **2003,** *24* (12), 1432-1436.

115. Rhee, Y. M.; Pande, V. S., Solvent viscosity dependence of the protein folding dynamics. *J Phys Chem B* **2008,** *112* (19), 6221-6227.

116. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K., On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. *J Am Chem Soc* **2003,** *125* (31), 9523-9530.

117. Wagoner, J. A.; Baker, N. A., Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *P Natl Acad Sci USA* **2006,** *103* (22), 8331-8336.

118. Chen, J.; Brooks, C. L., Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys* **2008,** *10* (4), 471-481.

119. Zhou, R. H.; Berne, B. J., Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water? *P Natl Acad Sci USA* **2002,** *99* (20), 12777-12782.

120. Pitera, J. W.; Swope, W., Understanding folding and design: Replica-exchange simulations of "Trp-cage" fly miniproteins. *P Natl Acad Sci USA* **2003,** *100* (13), 7587-7592.

121. Zhou, R. H., Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins-Structure Function and Genetics* **2003,** *53* (2), 148-161.

122. Zhou, R. H.; Krilov, G.; Berne, B. J., Comment on "Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?" The Poisson-Boltzmann equation. *J Phys Chem B* **2004,** *108* (22), 7528-7530.

123. Okur, A.; Wickstrom, L.; Simmerling, C., Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models. *J Chem Theory Comput* **2008,** *4* (3), 488-498.

124. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.

125. Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J., Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *P Natl Acad Sci USA* **2005,** *102* (39), 13749-13754.

126. Huang, X. H.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R. H.; Berne, B. J., Replica exchange with solute tempering: Efficiency in large scale systems. *J Phys Chem B* **2007,** *111* (19), 5405-5410.

127. Mu, Y. G.; Yang, Y.; Xu, W. X., Hybrid Hamiltonian replica exchange molecular dynamics simulation method employing the Poisson-Boltzmann model. *Journal of Chemical Physics* **2007,** *127* (8).

128. Nguyen, P. H., Replica exchange simulation method using temperature and solvent viscosity. *Journal of Chemical Physics* **2010,** *132* (14).

129. Kouza, M.; Hansmann, U. H. E., Velocity scaling for optimizing replica exchange molecular dynamics. *Journal of Chemical Physics* **2011,** *134* (4).

130. Wang, J. N.; Zhu, W. L.; Li, G. H.; Hansmann, U. H. E., Velocity-scaling optimized replica exchange molecular dynamics of proteins in a hybrid explicit/implicit solvent. *Journal of Chemical Physics* **2011,** *135* (8).

131. Wang, L. L.; Friesner, R. A.; Berne, B. J., Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J Phys Chem B* **2011,** *115* (30), 9431-9438.

132. Nagai, T., General Formalism of Mass Scaling Approach for Replica-Exchange Molecular Dynamics and its Application. *J Phys Soc Jpn* **2017,** *86* (1).

133. Kulke, M.; Geist, N.; Moller, D.; Langel, W., Replica-Based Protein Structure Sampling Methods: Compromising between Explicit and Implicit Solvents. *J Phys Chem B* **2018,** *122* (29), 7295-7307.

134. Geist, N.; Kulke, M.; Schulig, L.; Link, A.; Langel, W., Replica-Based Protein Structure Sampling Methods II: Advanced Hybrid Solvent TIGER2hs. *J Phys Chem B* **2019,** *123* (28), 5995-6006.

135. Zhang, T.; Nguyen, P. H.; Nasica-Labouze, J.; Mu, Y. G.; Derreumaux, P., Folding Atomistic Proteins in Explicit Solvent Using Simulated Tempering. *J Phys Chem B* **2015,** *119* (23), 6941-6951.

136. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.

137. Machado, M. R.; Barrera, E. E.; Klein, F.; Sonora, M.; Silva, S.; Pantano, S., The SIRAH 2.0 Force Field: Altius, Fortius, Citius. *J Chem Theory Comput* **2019,** *15* (4), 2719-2733.

138. Machado, M. R.; Pantano, S., SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **2016,** *32* (10), 1568-1570.