Exploring Protocols to build reservoirs to accelerate Replica Exchange MD simulations

Koushik Kasavajhala,[†‡] Kenneth Lam,[⊥‡] and Carlos Simmerling[†,‡,]*

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

[⊥]Molecular and Cellular Biology, Stony Brook University, Stony Brook, New York 11794, United States

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

Keywords: Conformational sampling, converged ensembles, REMD, reservoir REMD, Monte Carlo, protein folding, molecular dynamics

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 around 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 AMBER RREMD code onto GPUs making it 20x faster than the CPU code. Then, we explored protocols for building Boltzmann-weighted reservoirs as well as non-Boltzmann reservoirs, and tested how each choice affects the accuracy of the resulting RREMD simulations. We show that, using the recommended protocols outlined here, RREMD simulations can accurately reproduce Boltzmann-weighted ensembles obtained by much more expensive conventional REMD simulations, with at least 15x faster convergence rates even for larger proteins (>50 amino acids) compared to conventional REMD.

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 even for small 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.

Several enhanced sampling techniques have been developed to accelerate the convergence of MD simulations, a summary of which can be found in recent reviews.[1] 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[2], metadynamics[3], steered MD[4], 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).[5] The efficiency of the first category of methods is dependent on how accurately the collective variables 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.[6]

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 the conformational landscape even at low temperatures. Overall this results in faster convergence of conformational ensembles at all temperatures compared to standard MD. **Figure S1** illustrates the faster convergence of conformational ensembles using REMD for the CLN025[7] hairpin using the same number of MD runs and thermostats as for standard MD, 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 using REMD, compared to MD which requires greater than 400 ns to converge at lower temperatures. This reinforces that the coupling of temperatures allows the efficiently-sampling high temperature simulations to accelerate the low temperature conformational sampling. 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.

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[8]. 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.[6a-d]

In RREMD, a set of structures (reservoir) representing different energy minima 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 similar to J-walking[9], S-walking[10], smart darting[11], annealed swapping[12], and cool-walking[13] 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).

The 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 as in standard REMD. In practice, a crucial assumption is 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).[6a, 6c]

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[6a, 6b] and leucine dipeptide[6a] in gas phase, leucine tripeptide in implicit solvent[6b], Trpzip2 and DPDP in implicit solvent[6c], and $A\beta_{21-30}$ peptide[6e] 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[14]. 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 (relative to forces) 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[15]) and non-physics-based sampling methods (such as homology modeling[16]) 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)[6d] 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[6d], 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 refinement and local exploration of the basin, which can be crucial if the reservoir structure was not generated using the same physics 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[6g], alanine undecapeptide[6g], Trp-cage miniprotein[6g], and Trpzip2 in implicit solvent[6d], and for alanine dipeptide[6f] and RNA (rGACC) tetramer[6f] 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.[6h]

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.[6f] 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[14a] to all the way up to 1.4 μs[6f].

It was not 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[6a, 6c, 6e-g]. If the number of structures is too few, then the populations of different minima will have too low precision and might not reflect the true Boltzmann distribution; using such a reservoir in B-RREMD simulations could 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[6d], and (b) Conformational space annealing (CSA)[17] was used[6g] which, in theory, generates structures with low potential energy that are separated by some distance in conformational space. While either method can result in flat reservoir distributions, 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, and the impact of over-sampling basins is unclear. For example, alanine tetrapeptide simulations using 64 and 256 conformations obtained from CSA produced similar results.[6g] 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.[18] It was not clear which clustering method was the most appropriate, 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?

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.

RREMD code was implemented only on CPUs, which prohibited exploration of alternate approaches to these questions. 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. It is not known if the faster convergence rates of RREMD will translate well to 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, rigorous testing of 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).

In this work, we ported AMBER's RREMD code onto the GPUs which enabled 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 sensitive is a Boltzmann-weighted reservoir simulation to the number of structures selected from a high temperature MD simulation, 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, how sensitive are the results to the potential energy of the reservoir structures, using the energy of the representative structure vs. average energy of all structures in the cluster? (5) Are there ways to identify an 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 (1) both variants of RREMD are at least 2x faster for two small-sized proteins (CLN025 and Trp-cage) and at least 15x faster for a medium-sized protein (Homeodomain). (2) A correlation of cluster populations > 0.7, number of folding/unfolding event pairs > 100 are good indicators of adequate simulation lengths to generate Boltzmann-weighted structure reservoirs. (3) The number of structures in a Boltzmann-weighted reservoir can be as low as 1000 to 5000. (4) KMeans and WL are the best clustering methods to build a non-Boltzmann reservoir, and non-Boltzmann RREMD is only slightly sensitive to the energy assigned to reservoir structures. (5) The reservoir temperature should not be too low, and it should not be too high either, though further work needs to be carried out to understand the effect of temperature on the efficiency of RREMD. Finally, we also observed that the RREMD simulations using the same reservoir will converge to the same, and sometimes, wrong answer indicating that it is critical to ensure that the reservoirs themselves are well converged before running RREMD simulations.

Methods

Model systems

Three proteins – (1) CLN025 (PDB ID: 2RVD, 10 residues)[7], (2) Trp-cage (PDB ID: 1L2Y, 20 residues)[21], and (3) Homeodomain (PDB ID: 2P6J, 52 residues)[22] were considered for this study since these proteins were previously shown[23] 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.

General details

All structures were built via the LEaP module of AmberTools in the AMBER 18 package[24]. 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[25] was used for all simulations. The GB-Neck2[26] (igb=8 in AMBER) implicit solvent model with mbondi3[26] 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[23b]. The 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.

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.

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.

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. 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 was 2 fs. Coordinates were saved every 20 ps. Exchanges between replicas were attempted every 1 ps for all simulations.

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.

System specific details

CLN025 and Trp-cage miniprotein

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.

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.

For all three proteins, the temperatures at which the MD, REMD, and RREMD simulations were carried out are provided in the Supporting Information.

Building Reservoirs

Building Boltzmann-weighted reservoirs

We tested 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 precisely 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*[27] program in AMBER using a seed of 1.

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 in the Supporting Information.

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. Note that the non-Boltzmann reservoir structures were not minimized since minimization would alter the thermal energy and affect the MC exchanges with the reservoir.

Analyses:

The various analysis methods used to generate the data presented in **Results and Discussion** are described in detail in the Supporting Information.

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 energy should be assigned to each representative structure? (5) What is the ideal temperature to generate structures for building a reservoir?

Testing the above choices requires 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 standard 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.

As an initial 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 1**) 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.

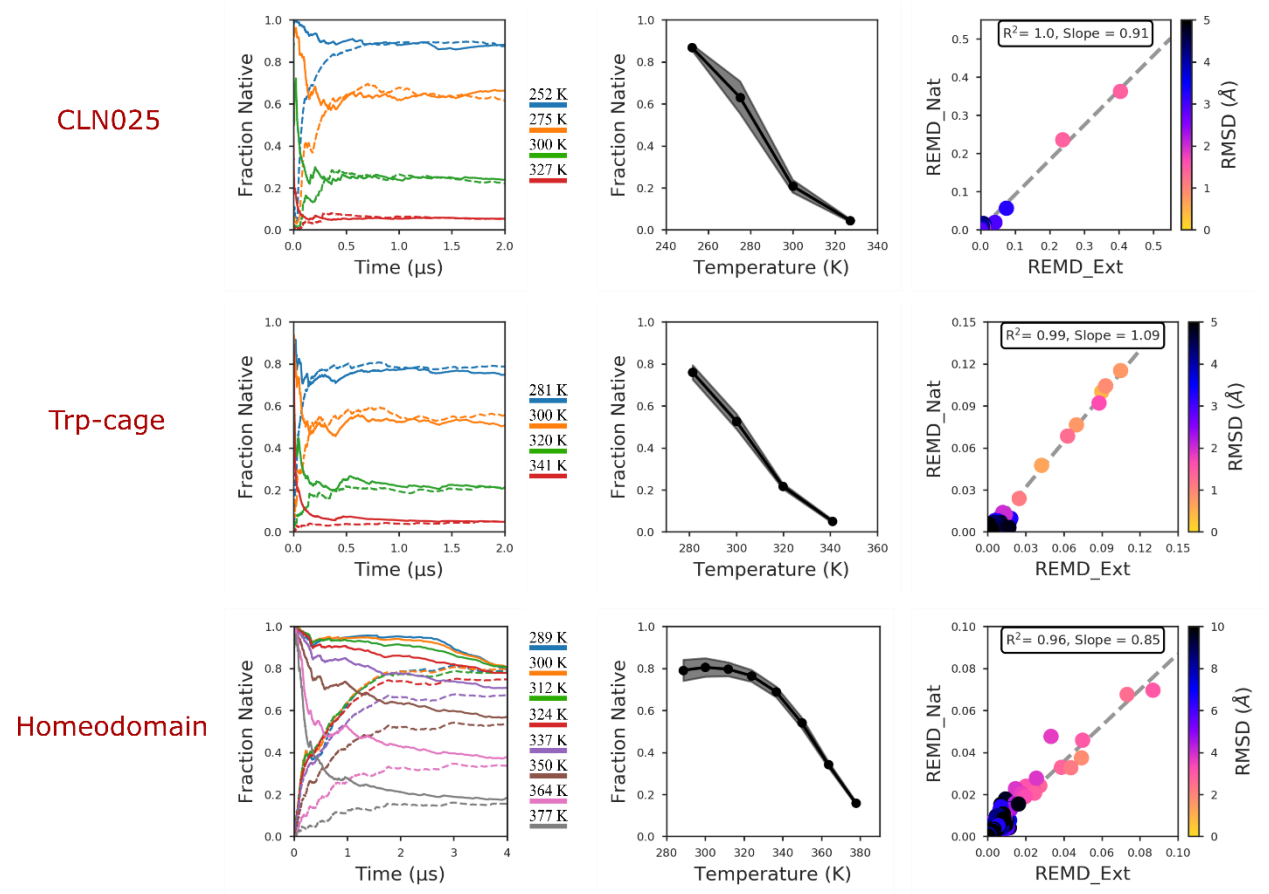


Figure 1. Reference data obtained from extensive standard REMD simulations. The fraction of native structures vs time at each temperature (column 1), the melting curves (column 2) and the cluster populations at the calculated melting temperature (column 3) are shown for CLN025 (top), Trp-cage (middle), 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 melting curves 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 of the cluster representative structure 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 1**) across the two independent REMD simulations, with error bars reflecting the half-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.[28] 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 1**. 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 have generated highly-converged ensembles, and can be used as reference data to test the various choices involved in building Boltzmann-weighted reservoirs and non-Boltzmann reservoirs. In the following sections, we explore the impact of these choices, and how they affect the accuracy and efficiency of RREMD simulations as compared to the reference data from conventional REMD.

Protocols for building Boltzmann-weighted reservoirs

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 appropriate structures populating each important local minimum. To ensure that the simulations are sampling different local minima, as an initial check, we calculated the number of folding/unfolding pair events (see **Methods** section) for each protein. **Table 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 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. It also requires no prior knowledge of the native structure, as is needed for folding event analyses.

Figure 2A 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.

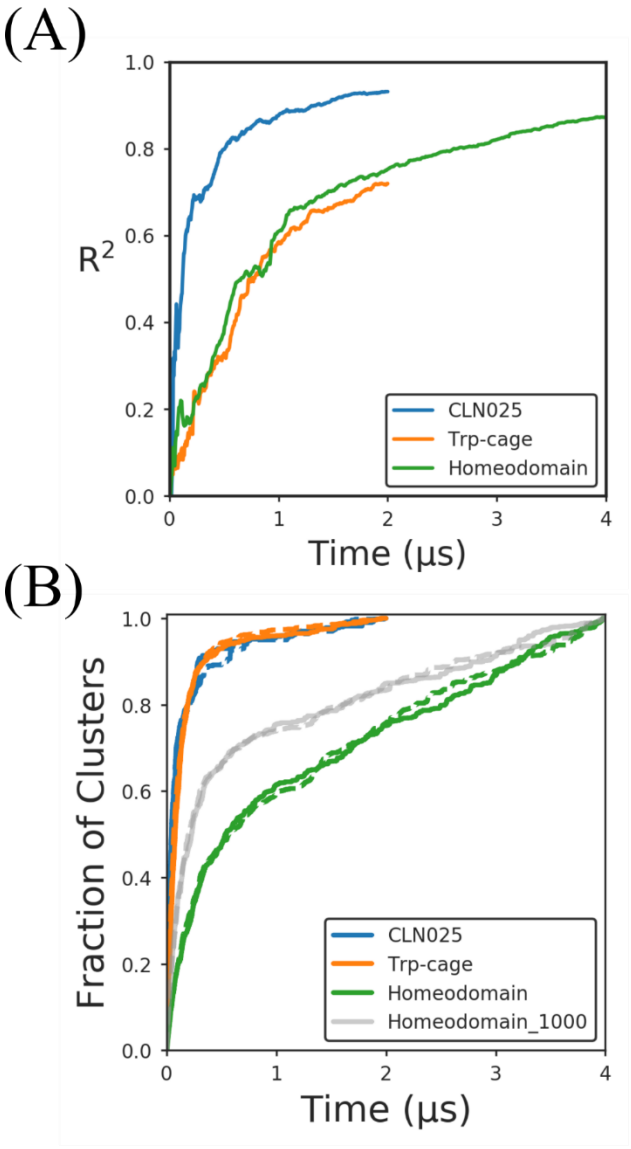


Figure 2. The correlation of cluster populations (A) between the two independent high-temperature simulations starting from different initial conformations for each protein, shown as a function of time. The fraction of unique clusters (B) observed during each independent simulation for each protein is 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 2B** 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 were well converged.

Nevertheless, **Figure 2A** and **Figure 2B** also show 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 that were obtained without using reservoirs.

To test these hypotheses, for each B-RREMD simulation used below, we built 6 (3 different time lengths * 2 high temperature MD simulations) 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 required to build a reliable Boltzmann-weighted reservoir for each protein. “T1” and “T2” correspond to simulation time lengths when the correlation of cluster populations (shown in **Figure 2A)** between the two independent reservoir generation simulations starting from different initial conformations (native

and extended) are 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 cluster population 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: reservoir B(T)_(c) 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 accuracies. **Figure 3** shows the melting curves obtained from B-RREMD simulations using each of these six 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.

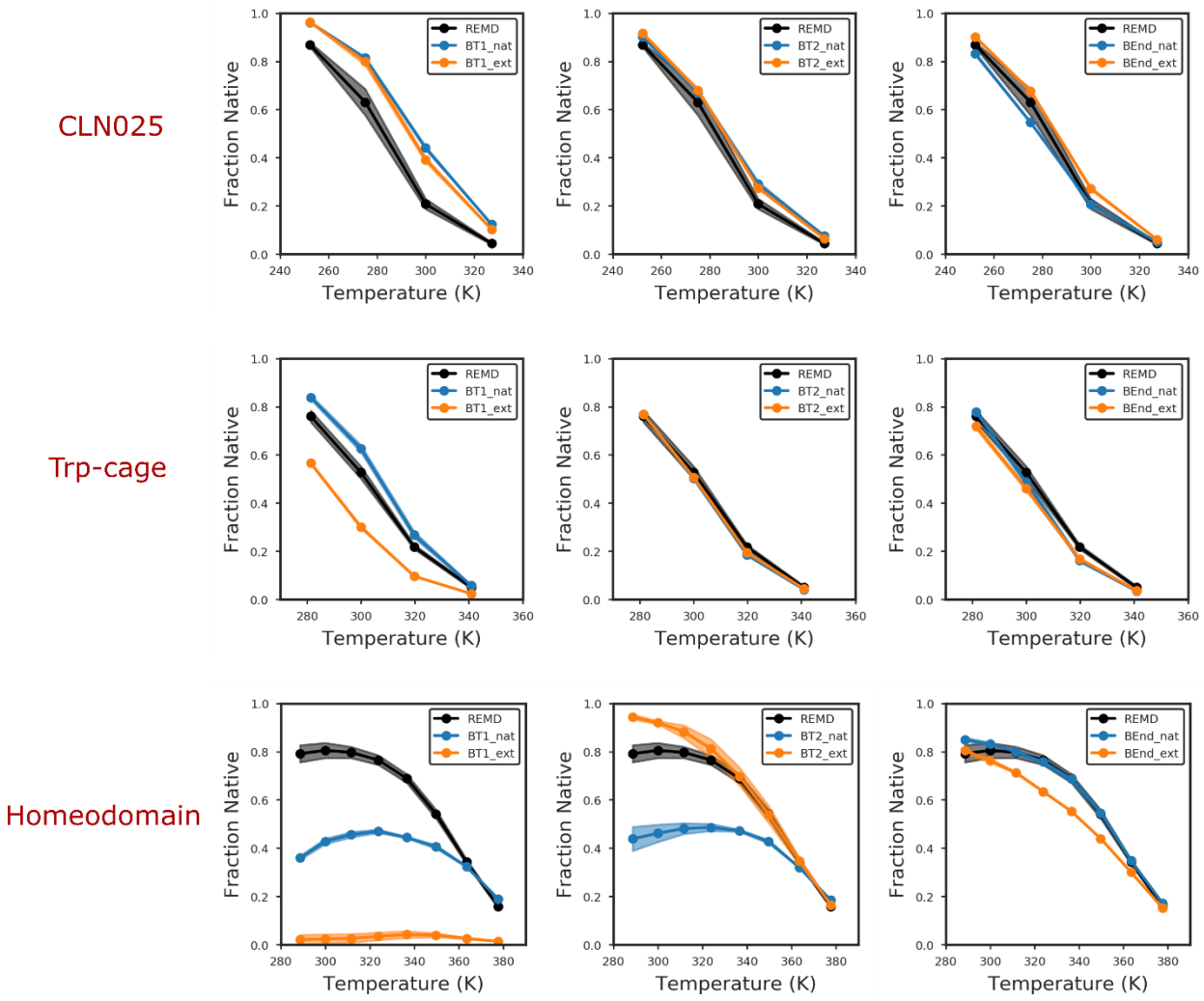


Figure 3. The melting curves obtained using standard REMD (black) and B-RREMD simulations using structures obtained from three different time lengths are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). T1, T2, and End, indicate different time lengths for which the reservoir generation 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 (compared to 0.21 from standard REMD), 0.49 and 0.46 (compared to 0.53 from standard REMD), and 0.83 and 0.76 (compared to 0.80 from standard REMD), fraction of native structures, respectively.

Also, for all three proteins, irrespective of the length of simulation used to generate reservoir structures, the error bars on the individual melting curves for B-RREMD simulations using the same reservoir are significantly smaller (sometimes negligible) than the difference between simulations using different reservoirs, indicating that RREMD simulations starting from different initial replica conformations but using the same reservoir will converge to the same (perhaps incorrect) answer.

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 an accurate 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 inaccurate reservoirs will possibly converge to the wrong answer and running 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 before running B-RREMD simulations.

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

For a given reservoir generation MD length, one next needs to determine how many snapshots to place in a reservoir. 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 (the "End" data sets in the previous section).

To distinguish between the eight reservoirs for each protein, the following reservoir naming convention was used: reservoir B(N)_(c) 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. Since all of these reservoirs were built using the entire reservoir generation simulations, we expect that the B-RREMD simulations using these reservoirs will match well to 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 4 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 the BEnd_(ext/nat) shown in **Figure 3** and hence, are not included in **Figure 4**. Similar to melting curves in **Figure 3**, the error bars on the individual melting curves for RREMD simulations are significantly smaller (sometimes negligible) than the difference between simulations using different reservoirs, irrespective of the number of structures used in the reservoir.

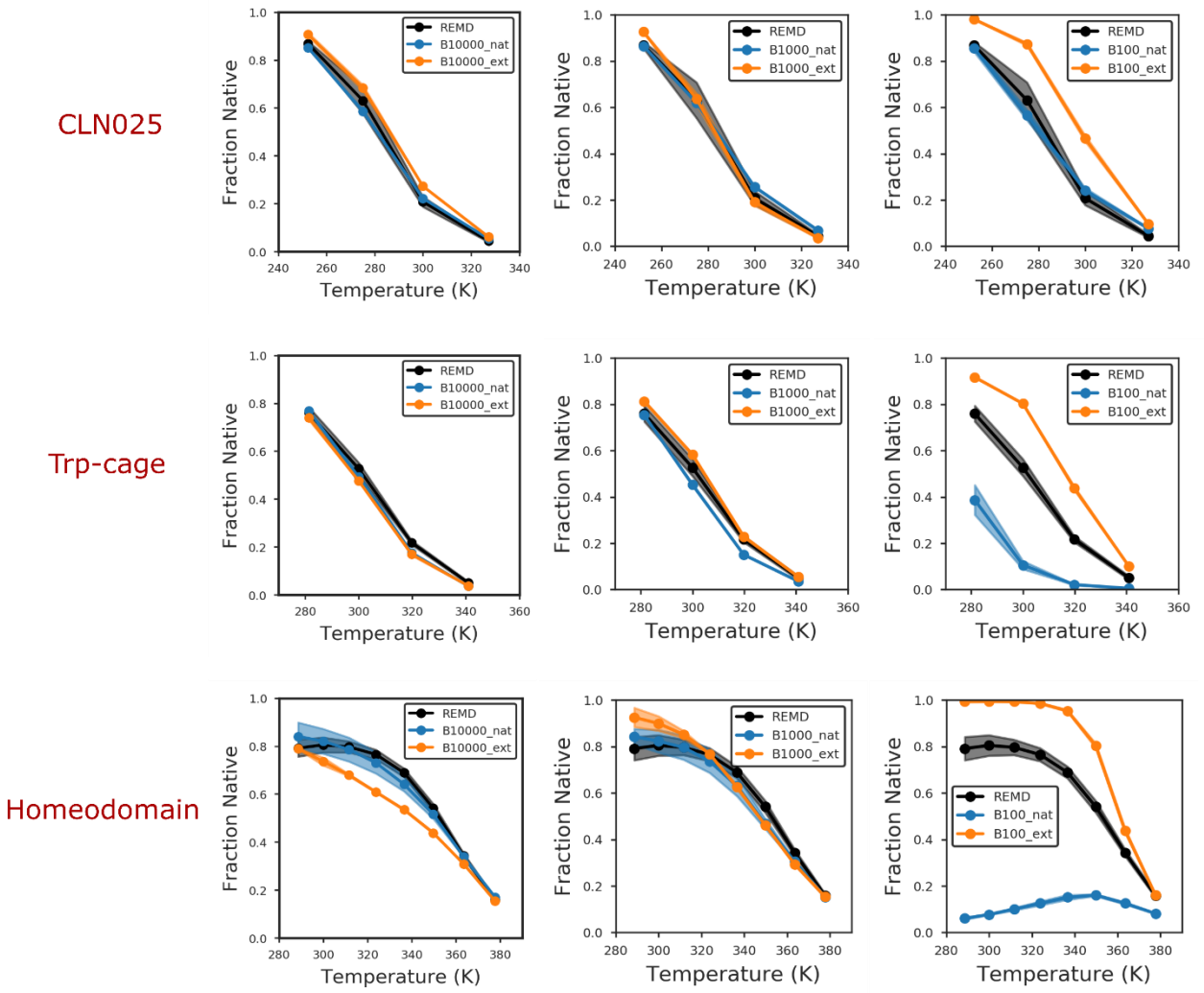


Figure 4. The melting curves obtained using standard REMD (black) and B-RREMD simulations using different number of reservoir structures are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). B10000, B1000, and B100 reservoirs indicate reservoirs having 10000, 1000, and 100 structures, 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 (shaded area) 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.

When 10000 structures are used in the reservoir, the melting curves obtained using B-RREMD simulations are in close agreement with standard REMD melting curves. Simulations using 10000 structure reservoirs obtained from two different MD runs are also in good agreement. 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 3** and related discussion). 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. Furthermore, the results from 2 different sets of 100 structures also do not match, even though the data above show that the reservoir generation simulations are themselves well converged. When only 100 snapshots

are used, the reservoirs become unreliable at representing the long MD run from which they were obtained. 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.[6c-f] 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.

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

The good agreement seen above 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. Ideally, B-RREMD simulations also 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 ensembles 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 5**.

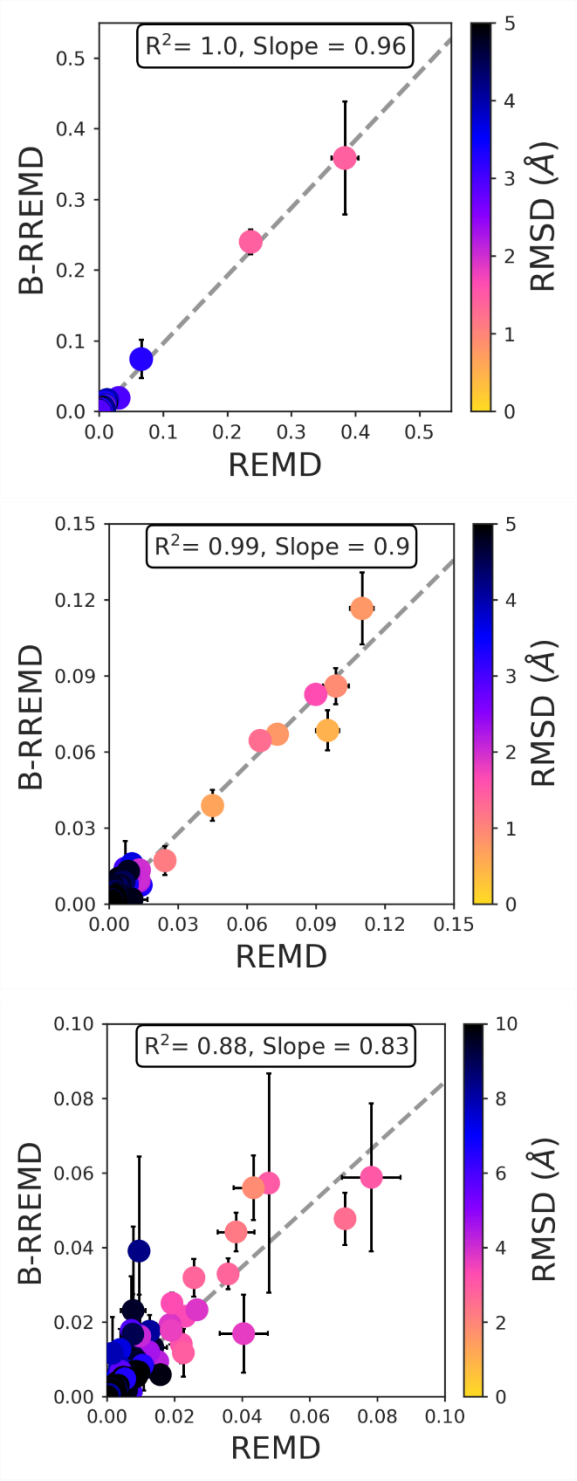


Figure 5. Cluster populations obtained using standard REMD (X-axis) and B-RREMD (Y-axis) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom) at 275.1 K, 300.0 K, and

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 5** 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 1**). 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 3** and **Figure 4** 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 S2**. 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 our conclusion above that independent B-RREMD simulations using the same reservoir converge to the same ensembles, irrespective of initial structure for the REMD step. Therefore, it is critical to ensure that the reservoirs are well converged before running B-RREMD simulations, perhaps by comparing cluster populations as shown above. Nonetheless, performing long MD simulations at a single high temperature followed by short B-RREMD simulations at all temperatures may save considerable computing resources compared to performing long REMD simulations at all temperatures.

Protocols for building non-Boltzmann reservoirs

Motivation for use of non-Boltzmann reservoirs. As seen above, 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 costly 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 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.[6d] 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.

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.[6d] 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 are often 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 suitable[18c]. However, to identify metastable states with very low density, a partitioning algorithm such as KMeans might be better[18e]. 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 intuitive, 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.

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 reduce 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 directly corresponds to a standard MC jump to a randomly selected basin on the energy landscape, which happens in this case 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. Importantly, since the B-RREMD simulations gave similar results to standard REMD (**Figures 3, 4, and 5**), if the NB-REMD gives different results it would indicate issues with the protocols for selecting representative structures and energies, rather than problems with the source MD data for the reservoir.

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 i.e., the weights for each cluster are absent/ignored. While removing the cluster weights could result in reduced accuracy vs. a converged Boltzmann reservoir, using a method that does not employ weights may improve the results as compared to using a poorly converged Boltzmann reservoir with inaccurate weights.

Identifying a good clustering protocol for building non-Boltzmann reservoirs

A good clustering protocol should result in homogeneous clusters, i.e. all the structures in each cluster should look alike. Therefore, a metric for validating 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 variants of each key step.

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[18d], 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 which resulted in too few (<15) and mostly native-like clusters (data not shown).

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 ensures that the reservoir accounts for alternate arrangements of termini or loop regions even if they are not well ordered in the native fold; furthermore, these regions of the protein also contribute to the energy of the structures (see below) that is assigned to the cluster in the

reservoir. If not included in the metric, it is possible that the single representative structure for a basin could exhibit an atypical conformation for these regions.

Identifying the ideal number of clusters

To approximately identify the ideal number of clusters for each clustering method and for each protein, we performed cluster analysis by setting the target number of clusters to different numbers. We then looked for consistent trends across the model systems. 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 S3**.

For all the clusters, the difference between average and median intra-cluster RMSDs is always 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 S3. Based on these results, we chose the clusters by setting the target number of clusters to 2000 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 1**.

Table 1. 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.

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 3, 4,** and **5**) 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. This also allows comparison of clusters between the reservoir generation runs.

Alternatively, since most of the clusters were observed in the first 500 ns for CLN025 and Trp-cage (see **Figure 2**), using the trajectories up to 500 ns for these two proteins and clustering on them might be sufficient. Here, we initially focus on quantifying the accuracy of the nB-RREMD approach given complete reservoir sampling, rather than identifying the most inexpensive approach to generating the reservoir; that will be explored in the future.

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 also can be explored in more detail in future work.

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 in the cluster was chosen as the representative. This was a straightforward choice, since the Amber *cpptraj*[27] program outputs this representative structure by default. (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 structures in the cluster, denoted as 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 a non-Boltzmann reservoir, built using representative structures obtained from Average-linkage clustering method, and using the cluster representative energy (CRE) or cluster average energy (CAE), respectively.

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. All simulations were repeated with different initial replica conformations. **Figure 6** 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.

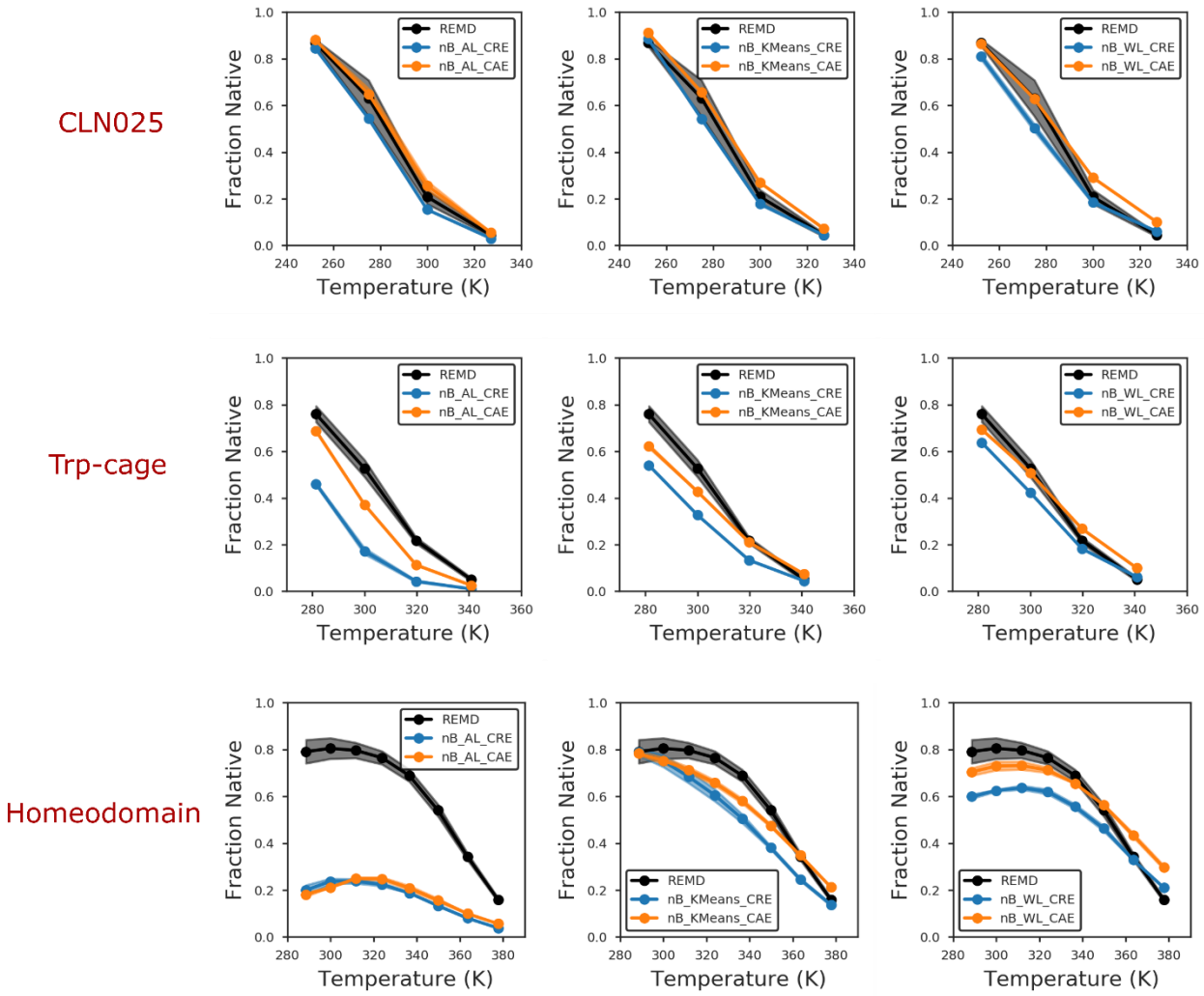


Figure 6. The melting curves obtained using standard REMD (black) and nB-RREMD simulations using different clustering methods are shown for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). AL, KMeans, and WL indicate that the cluster representatives used to build the reservoir were obtained from Average-Linkage, KMeans, and Ward-Linkage

clustering algorithms, respectively. CRE (blue) and CAE (orange) 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. Overall, CLN025 appears insensitive to the clustering details.

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 increases 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 produce melting curves that are in reasonable agreement with the standard REMD melting curves, whereas AL performs most poorly and 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 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.

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 simulated 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 7**.

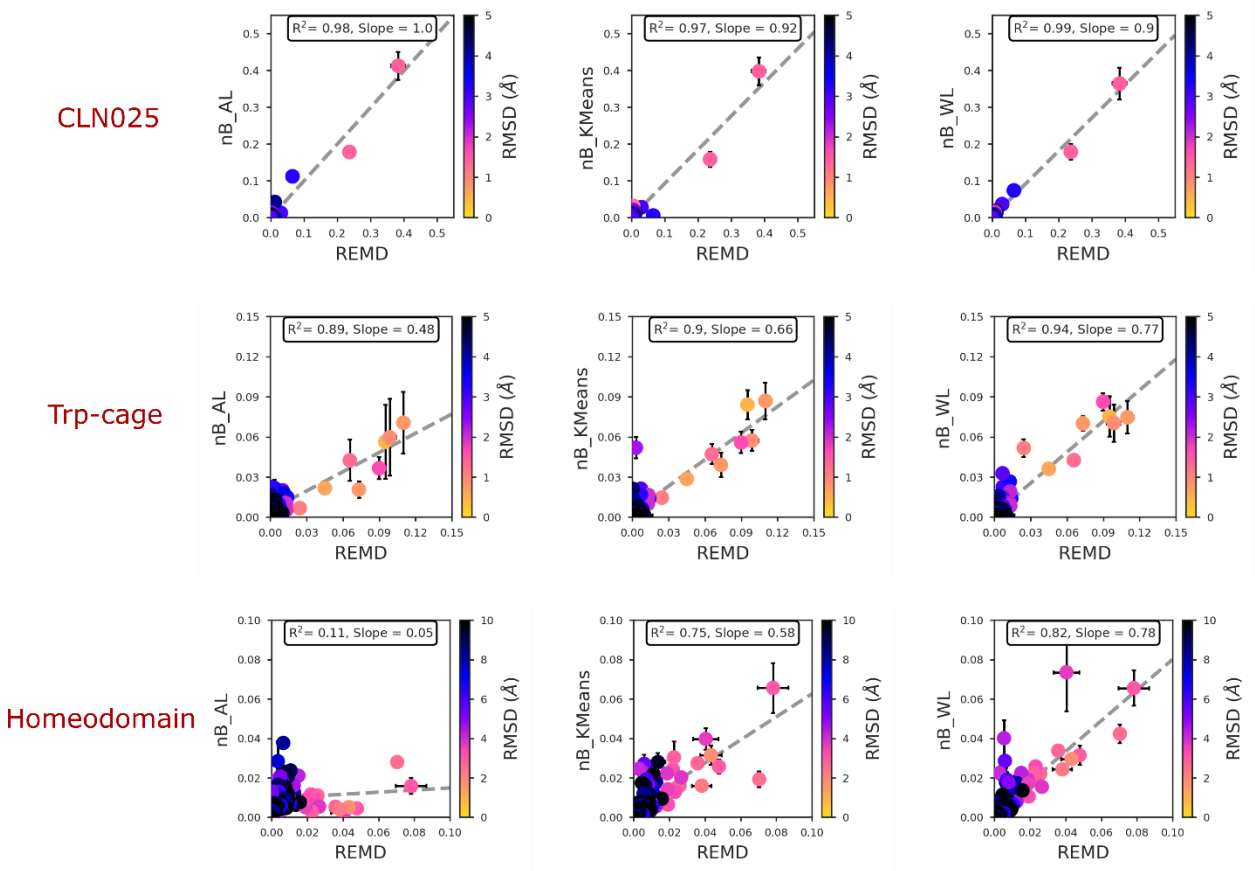


Figure 7. Cluster populations obtained using standard REMD (X-axis) and nB-RREMD (Y-axis) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom) at 275.1 K, 300.0 K, and

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.

CLN025 once again shows very little sensitivity to reservoir clustering method and energies, and 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 reinforces that simple peptides may have limitations when used to validate protocols that will be applied to more complex systems.

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 inferior, 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 given the major simplification of the reservoirs.

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 6** that using CAEs results in slightly more stable melting curves compared to using CREs. However, the small Y-error bars in **Figure 7** 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 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 S4**) 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.

Ideal temperature to generate the reservoir

To explore the sensitivity to the temperature at which the reservoir is generated, we performed MD simulations at different temperatures for each protein (see **Methods**). Then, similar to the analyses in **Figure 2** and **Table 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 8**, and the number of folding/unfolding event pairs are shown in **Table 2**. The corresponding data for CLN025 and Homeodomain are shown in **Figures S5** and **S6**, **Tables S1** and **S2**, respectively.

Low temperature simulations would not be expected to be useful for reservoir generation, since this abandons the temperature-accelerated barrier crossing 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 8**) indicating that the simulation is stuck in a local minimum. The correlation of cluster populations increases after 0.7 μs but

remains below 0.5. Moreover, the average number of folding/unfolding event pairs observed at this temperature is only 13 ± 5, indicating that 281.4 K is too low a temperature to build the reservoir (as expected).

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 simulations would be needed.

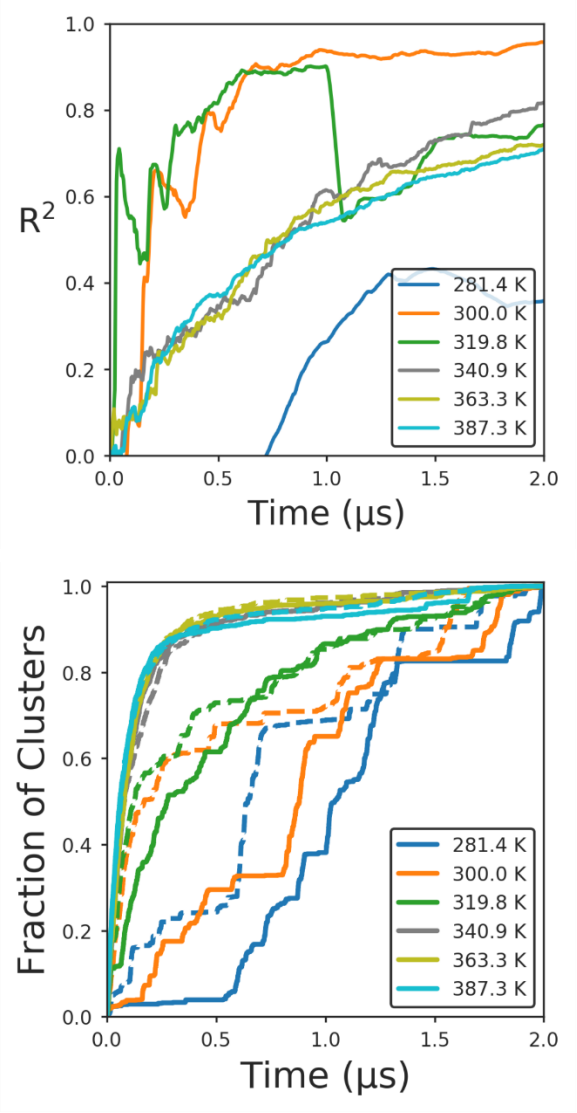


Figure 8. Identifying optimal temperatures for Trp-cage reservoir generation. The correlation of cluster populations (top) between the two independent simulations at different temperatures starting from different initial conformations is shown as a function of time. The fraction of unique clusters (bottom) 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. 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 319.8 K, around 0.85 fraction of unique clusters are observed within the first 1 μs, and the average number of folding/unfolding event pairs is 175 ± 10, more than double that from 300 K. 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. Thus 319.8 K improves sampling but is still susceptible to kinetic trapping.

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.

These results confirm that use of 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 can also be 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 for the system (or for the simulation conditions) is not known, radius of gyration can be used to validate the sampling large transitions across the energy landscape. The ideal temperature in that case would be one that quickly alternates between conformations with low and high radius of gyration, ideally multiple times (see **Figure S7**). 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.

Is reservoir REMD more efficient than standard REMD?

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 each method. 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, and nB-RREMD simulations built using KMeans clustering method and CAEs (the most reliable nB protocol that we explored) are shown in **Figure 9** 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.[6c-h]

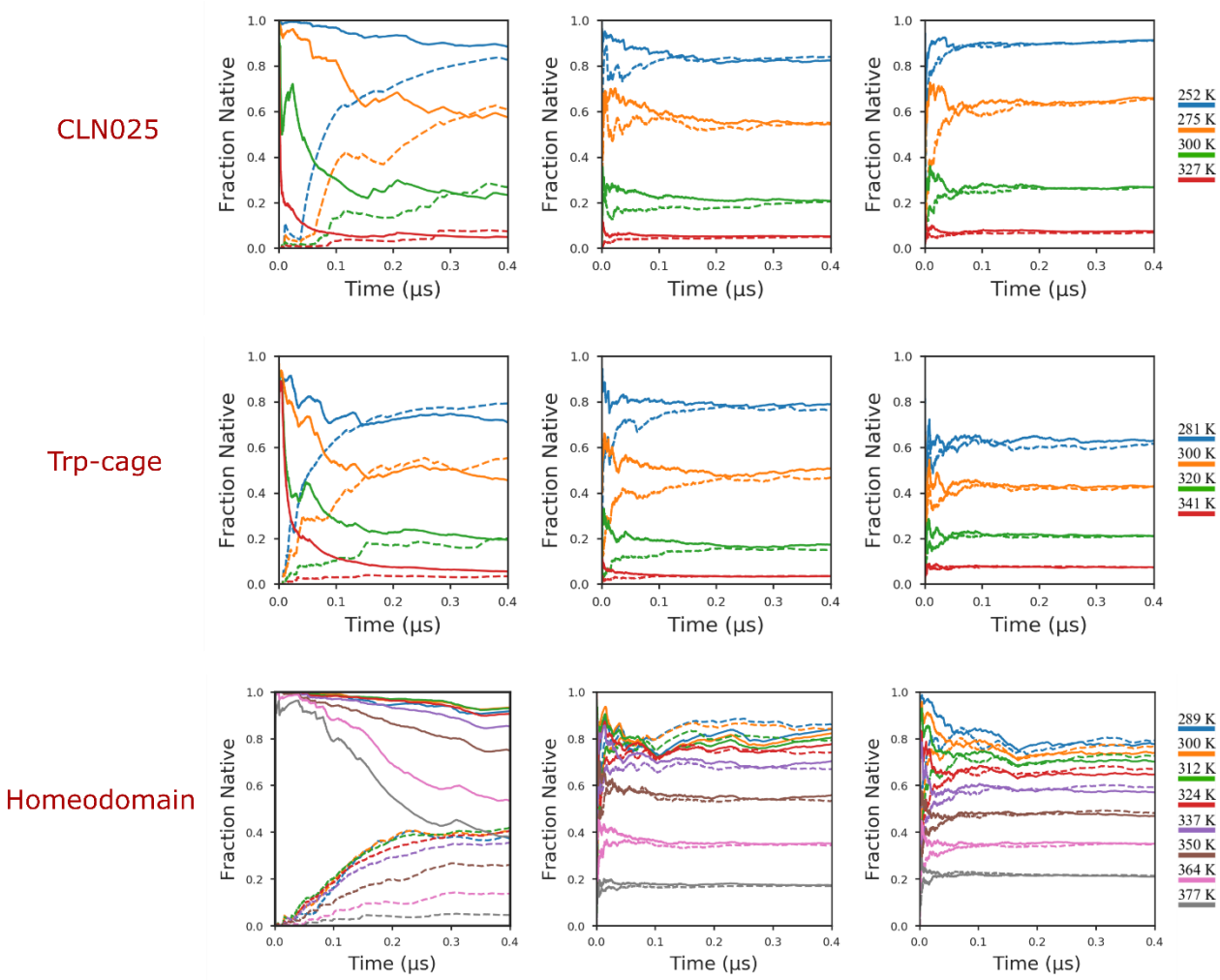


Figure 9. Fraction Native vs Time using standard REMD (left), B-RREMD (center), and nB-RREMD (right) for CLN025 (top), Trp-cage (middle), and Homeodomain (bottom). The solid lines indicate the fraction of native structures for simulations starting from native conformation while the dashed lines indicate 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** except that only the first 0.4 μs are shown here for all three proteins so the methods can be directly compared.

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 9** 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**) 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 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 cost 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. As we showed above, Trp-cage folds rapidly at temperatures above 300 K and may be too "easy" to require advanced methods, especially in implicit solvent. Overall, there can be some gain with using a Boltzmann-weighted reservoir, but the cost of using MD to converge the ensemble even at a single T can be significant. Nonetheless, the reservoir generation simulation lengths mentioned above are upper bounds.

The main advantage to non-Boltzmann reservoirs is that the simulations used to generate reservoirs do not need long MD simulations to converge the populations. As shown in **Figure 2**, non-Boltzmann reservoirs should require significantly shorter reservoir generation simulations. Moreover, slowly converging MD may not be necessary at all for sampling the reservoir basins in the reservoir; structures obtained from different enhanced sampling techniques such as accelerated MD, metadynamics, umbrella sampling, or even standard REMD could be used in combination with structures from experiments or homology modeling, further reducing the time required to generate reservoir structures. This will be explored in future work.

Why are Reservoir REMD simulations more efficient than standard REMD?

In RREMD simulations, in addition to swapping thermostats / scaling velocities between replicas, structures also can be 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 10,** the temperature and RMSD of one replica during the first 400 ns of standard REMD and B-RREMD simulations for Trp-cage are shown (similar results are observed for other replicas, data not shown).

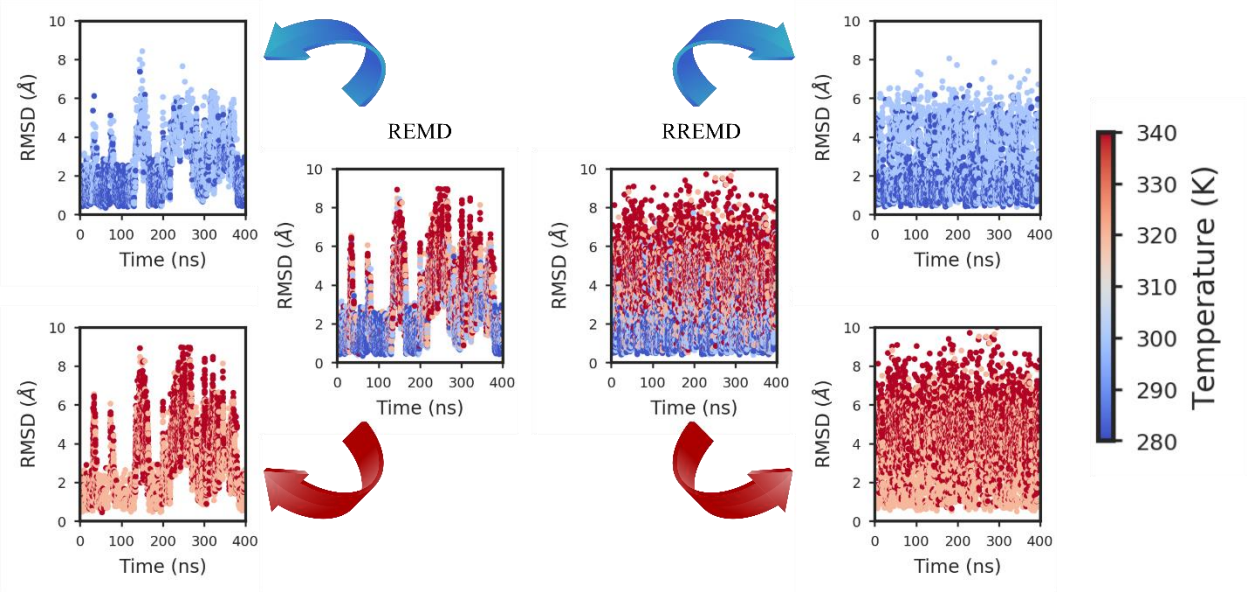


Figure 10. The distribution of temperature and RMSD of one replica during the first 400 ns of standard REMD (center/left) and B-RREMD (center/right) simulations for Trp-cage. The color of each point represents the temperature, while the position of each point on the Y-axis represents the RMSD to native 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 10**) 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 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 10**).

If the higher temperatures result in faster structural transitions, then the bottom left image in **Figure 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 appears that swapping to higher temperature does not help this replica to escape the basin in which it is trapped.

In contrast, a replica in RREMD simulation (center right image in **Figure 10**) 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 10**) and low (<2.0 Å) RMSD structures are routinely sampled even at high temperatures (bottom right image in **Figure 10**). 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 would 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.

If the MC steps are responsible for accelerated convergence, exchanging less often with the reservoir should slow down the convergence rate of RREMD. To explore this, we performed nB-RREMD simulations for Trp-cage in which we switched from exchange attempts with the reservoir each 2 ps for the data shown above, to attempting exchange with the reservoir every 50 ps. The exchange attempt frequency between the replicas was unchanged. As expected, less frequent MC steps slow convergence; 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 S8**).

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 in some cases could improve the accuracy of nB-RREMD simulations. **Figure 11** shows the effect of reservoir exchange frequency on the Trp-cage melting curves obtained using nB-RREMD simulations using structure reservoirs obtained using the three clustering methods.

When the exchanges with the reservoir were 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).

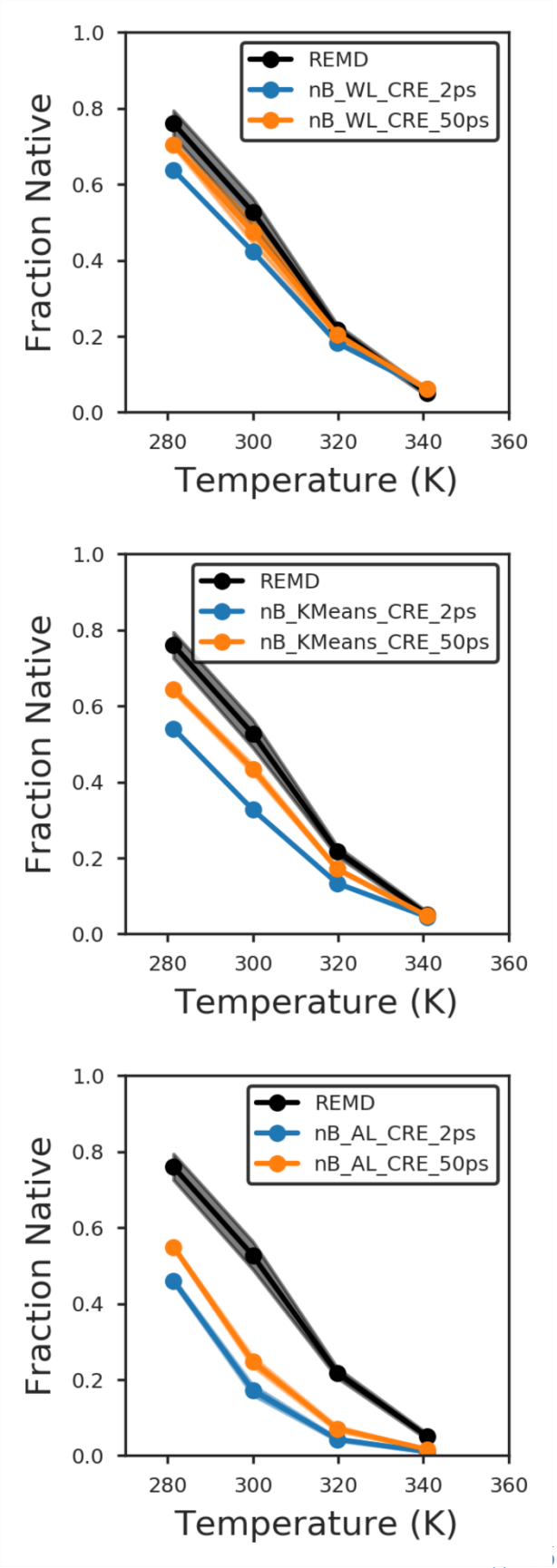

REMD
nB_WL_CRE_2ps
nB_WL_CRE_50ps
Fraction Native
Temperature (K)
REMD
nB_KMeans_CRE_2ps
nB_KMeans_CRE_50ps
Fraction Native
Temperature (K)
REMD
nB_AL_CRE_2ps
nB_AL_CRE_50ps
Fraction Native
Temperature (K)

Figure 11. 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 6**). 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.

ConclusionsDespite the significant increase in convergence speed compared to standard REMD, 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 ported the AMBER RREMD code on to the GPUs and explored protocols to build reservoirs to accelerate the convergence of REMD simulations. Results were evaluated using three systems for which reliable ensembles could be obtained using standard methods.

Specifically, we explored 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 also showed that RREMD simulations using the same reservoir will converge to the same, and sometimes, wrong answer. Therefore, it is critical to ensure that the reservoirs themselves are well converged before using them to run RREMD simulations. Nonetheless, we also showed that performing long MD simulations at a single high temperature followed by short B-RREMD simulations at all temperatures will save considerable computing resources compared to performing long REMD simulations at all temperatures.

We have also shown that generating structures for a non-Boltzmann reservoir will require significantly less time since the majority of the relevant minima are sampled significantly earlier in simulations than sampling of these same minima with the correct relative populations. Therefore, 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 were the most reliable. 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 cluster average energies tend to slightly favor more native-like structures compared to the corresponding simulations using the energy of the single representative snapshot. Nonetheless, the overall ensembles obtained from both approaches were similar.

The clustering protocols described here for building 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, appear too large 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.[18a, 18b, 18d] Nevertheless, we have also shown that the imperfections in a non-Boltzmann reservoir can be slightly ameliorated by exchanging less frequently with reservoir.

We explored 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. This remains a challenge of the reservoir approach.

Our results demonstrated 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 may be useful to quickly test the accuracy of new force fields, and design novel peptides. Future work should focus on using the alternate methods for building reservoirs.

ASSOCIATED CONTENT

Supporting Information

The detailed simulation setup for each system is provided in the Supporting Information along with the temperatures used for all replicas.

AUTHOR INFORMATION

Corresponding Author

*E-mail: carlos.simmerling@stonybrook.edu. Phone: +1 (631) 632-5424.

Author Contributions

Koushik Kasavajhala – Carried out the work, wrote the manuscript.

Kenneth Lam – Helped with numerous discussions on the topic and provided initial scripts for plotting Figure 8.

Carlos Simmerling – Principal Investigator.

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by NIH grant GM107104. We gratefully acknowledge support from Henry and Marsha Laufer.

Notes

The authors declare no competing financial interest.

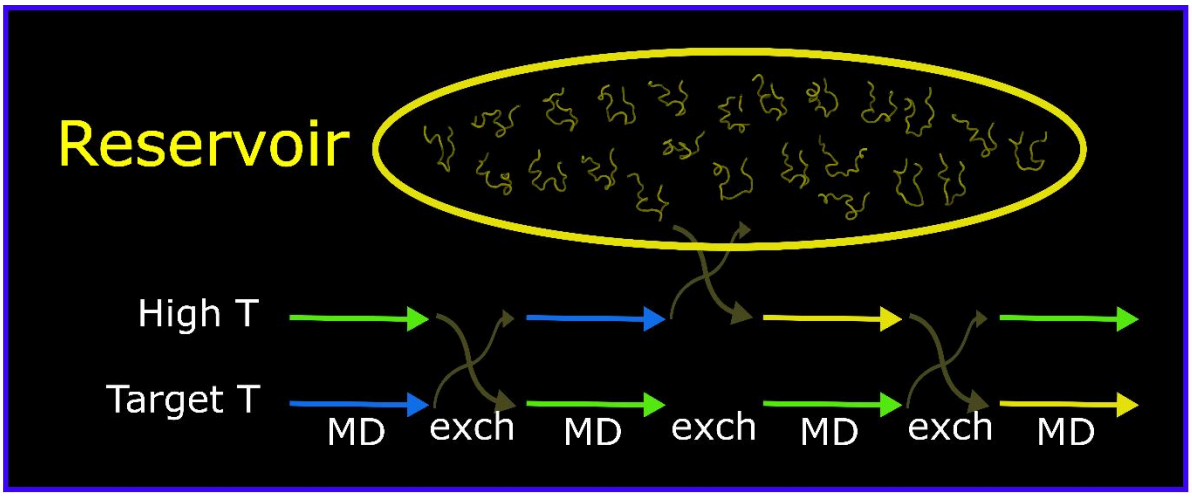
ACKNOWLEDGEMENTS

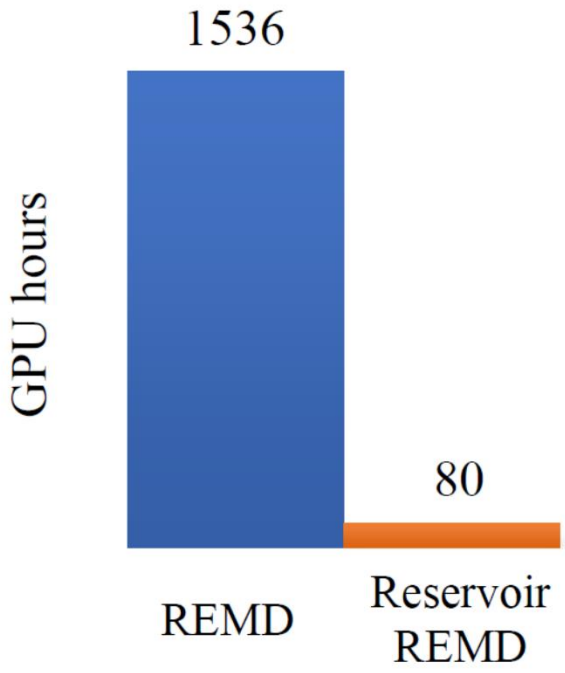
We thank Dr. Daniel Roe for making custom modifications of the *cpptraj* code. These were valuable for testing the method, but are not required to reproduce any of the work shown.

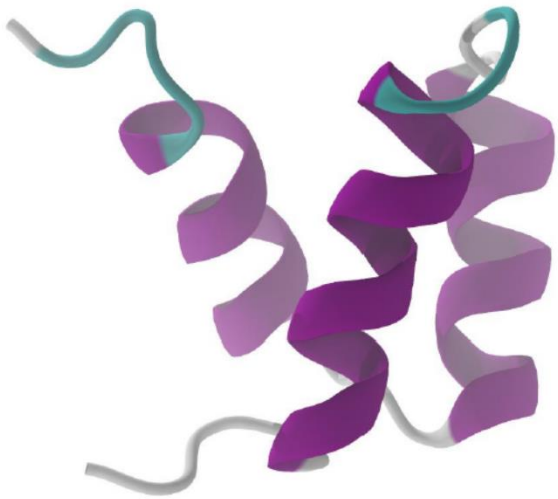
ABBREVIATIONS

GB, generalized Born; MD, Molecular Dynamics; REMD, 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.

TOC Figure







Homeodomain
(52 residues)

References

1. (a) Pietrucci, F., Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics* **2017,** *2*, 32-45; (b) 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; (c) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q., Enhanced sampling in molecular dynamics. *J Chem Phys* **2019,** *151* (7), 070902.
2. Torrie, G. M.; Valleau, J. P., Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *Journal of Computational Physics* **1977,** *23*, 187.
3. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc Natl Acad Sci U S A* **2002,** *99* (20), 12562-6.
4. Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K., Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys J* **1997,** *72* (4), 1568-1581.
5. (a) Hansmann, U. H. E., Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* **1997,** *281* (1-3), 140-150; (b) Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **1999,** *314* (1-2), 141-151; (c) Earl, D. J.; Deem, M. W., Parallel tempering: Theory, applications, and new perspectives. *Phys Chem Chem Phys* **2005,** *7* (23), 3910-3916.
6. (a) 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); (b) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M., Resolution exchange simulation. *Phys Rev Lett* **2006,** *96* (2), 028105; (c) 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; (d) 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; (e) 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; (f) 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; (g) 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; (h) Damjanovic, A.; Miller, B. T.; Okur, A.; Brooks, B. R., Reservoir pH replica exchange. *Journal of Chemical Physics* **2018,** *149* (7).
7. 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.
8. (a) Sindhikara, D.; Meng, Y. L.; Roitberg, A. E., Exchange frequency in replica exchange molecular dynamics. *Journal of Chemical Physics* **2008,** *128* (2); (b) Trebst, S.; Troyer, M.; Hansmann, U. H. E., Optimized parallel tempering simulations of proteins. *Journal of Chemical Physics* **2006,** *124* (17); (c) Nadler, W.; Hansmann, U. H. E., Generalized ensemble and tempering simulations: A unified view. *Phys Rev E* **2007,** *75* (2).
9. 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.
10. 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.

11. Andricioaei, I.; Straub, J. E.; Voter, A. F., Smart darting Monte Carlo. *Journal of Chemical Physics* **2001,** *114* (16), 6994-7000.
12. Opps, S. B.; Schofield, J., Extended state-space Monte Carlo methods. *Phys Rev E* **2001,** *63* (5).
13. Brown, S.; Head-Gordon, T., Cool walking: A new Markov chain Monte Carlo sampling method. *J Comput Chem* **2003,** *24* (1), 68-76.
14. (a) 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; (b) 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.
15. 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.
16. 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.
17. 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.
18. (a) 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; (b) 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); (c) 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); (d) 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; (e) 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.
19. 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.
20. Ward, J. H., Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* **1963,** *58* (301), 236-244.
21. Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H., Designing a 20-residue protein. *Nat Struct Biol* **2002,** *9* (6), 425-430.
22. 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.
23. (a) 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; (b) 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. (a) 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; (b) 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); (c) 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.