JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

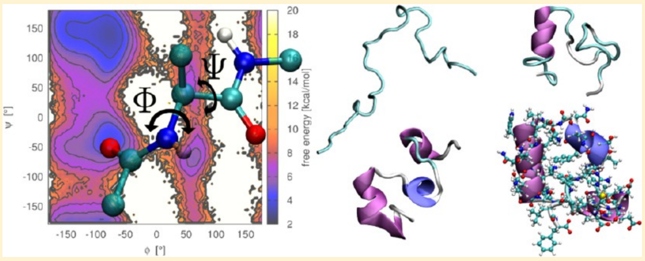
Overcoming the Rare Event Sampling Problem in Biological Systems with Infinite Swapping

Nuria Plattner,*,[†] J. D. Doll,*,[‡] and Markus Meuwly*,[¶]

[†]Department of Mathematics and Computer Science, Free University Berlin, Arnimallee 6, 14195 Berlin, Germany
[‡]Department of Chemistry, Brown University, Providence, Rhode Island 02912, United States
[¶]Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

S Supporting Information

ABSTRACT: Infinite swapping (INS) is a recently developed method to address the rare event sampling problem. For INS, an expanded computational ensemble composed of a number of replicas at different temperatures is used, similar to the widely used parallel tempering (PT) method. While the basic concept of PT is to sample various replicas of the system at different temperatures and exchange information between the replicas occasionally, INS uses the symmetrized distribution of configurations in temperature space, which corresponds to the infinite swapping limit of PT. The effect of this symmetrization and the enhanced information exchange between replicas is evaluated for three different biological systems representing different sampling problems in biology: (1) blocked alanine dipeptide, which is a small system and therefore optimal to evaluate sampling efficiency quantitatively, (2) Villin headpiece, which is used as a test case for the protein folding process, and (3) neuroglobin, which is used to evaluate the effects of enhanced information exchange between replicas for sampling the substate space of a folded protein. For these three test systems, PINS is compared to PT, and it is found that in all cases the sampling with PINS is substantially more efficient.



■ INTRODUCTION

Atomistic molecular dynamics (MD) simulations of biological systems are useful tools for the understanding of biological processes.[1−3] Despite the increasing availability of computational resources, many biologically relevant processes cannot be studied easily with atomistic simulations. The problem arises due to the different time scales on which biological processes occur. The maximal time step of a MD simulation is dictated by the motions of the fastest degrees of freedom in the system, i.e., bond vibrations, and therefore limited to femtoseconds. The time scale on which functional biological processes take place is however much longer, e.g., microseconds to seconds for protein−protein association or protein folding. The biological processes of interest are therefore dictated by events that occur infrequently on the simulation time scale relevant for biological processes. For the calculation of quantities characterizing these processes, such as, for example, folding rates or association constants, these rare events have to be simulated for a statistically significant number of times, which is difficult from direct MD simulations. Several enhanced sampling strategies suitable to overcome the rare event sampling problem in the sampling process and reduce the overall computation time have been developed in the past and have been successfully applied to biological systems, including parallel tempering (PT),[4] umbrella sampling,[5] and metadynamics.[6] Replica exchange (RE)[7] is a widely utilized method for dealing with rare event sampling problems. In RE, an expanded computational ensemble composed of systems corresponding to different values of one or more control variables is used. Parallel tempering (PT)[4,8] is an RE method utilizing temperature as the control variable and has been shown to be useful for many studies of biological systems.[9,10] In a PT simulation with N replicas in the canonical ensemble, the partition function of the combined ensemble is given by

$$Q = \prod_{i=1}^{N} \frac{q_i}{M!} \int dX_i e^{-\beta_i V(X_i)} \tag{1}$$

where $q_i = \prod_{k=1}^{M} (2\pi m_k k_B T_i)^{3/2}$ comes from integrating out the momenta of the M particles in the system, m_k is the particle mass, X_i is the configuration of replica i defined by the positions of the M particles, $V(X)$ is the systems potential function, and $\beta_i = 1/k_B T_i$ is the inverse temperature T_i of replica i divided by the Boltzmann constant k_B. Configurations are exchanged between replicas i and j with acceptance probability

$$P_{acc}(i \leftrightarrow j) = \min\{1, e^{(\beta_i - \beta_j)(V(X_i) - V(X_j))}\} \tag{2}$$

PT can be combined with different methods for sampling the conformational space, including different Monte Carlo (MC) move strategies and MD simulations.[4,11]

Received: May 2, 2013

dx.doi.org/10.1021/ct400355g | *J. Chem. Theory Comput.* XXXX, XXX, XXX−XXX

The recently developed infinite swapping (INS) method also uses an expanded computational ensemble composed of a number of replicas at different temperatures.[12,13] However, instead of considering the sampling of individual replicas at various temperatures and occasional information exchange between the replicas as is done in PT, INS uses the symmetrized distribution of configurations in temperature space. INS is based on a mathematical analysis of the PT convergence as a function of the swap rate, i.e., the rate of attempting configuration exchanges between temperatures. The mathematical analysis proves that the rate of convergence of PT is a monotonically increasing function of the swap rate. Hence, optimal sampling is achieved in the infinite swapping limit, which is a mathematically rigorous result.[13] The symmetrized distribution of configurations in temperature space used in INS corresponds mathematically to the infinite swapping limit of PT. Therefore INS allows for an optimal exchange of information between temperatures and uses information of all replicas for evaluating statistical properties at individual temperatures.[12] In the PT algorithm, the most aggressive attempt to exchange information between replicas would imply attempting to exchange all configurations between all replicas at each step. However, in PT such an exchange scheme would result in a large amount of computing time being spent in attempted moves between temperatures with very few moves in configuration space. The INS algorithm addresses this problem from a different perspective. Instead of attempting individual exchanges between replicas, the probability of possible outcomes of a general exchange attempt is used to propagate the simulation. For a system of N replicas, there are $N!$ permutations of configurations in temperature space which could be the outcome of attempting all possible exchanges. For a given set of configurations X, the probability $\rho_k(X)$ of permutation k is given by

$$\rho_k(X) = \frac{p_k(X)}{\sum_{k=1}^{N!} p_k(X)} \tag{3}$$

with $p_k(X)$ given by

$$p_k(X) = e^{-\beta_1 V(x_{k,1})} e^{-\beta_2 V(x_{k,2})} \ldots e^{-\beta_N V(x_{k,N})} \tag{4}$$

where $x_{k,n}$ denotes the configuration of replica n corresponding to the assignment of configurations to temperatures in permutation k. In the INS algorithm, the permutation probabilities are used to select the permutation of configurations in temperature space. The selected permutations are then used to sample the configuration space and the permutation probabilities are used to calculate the statistical system properties for each temperature based on information contributed by all replicas. Because for large numbers of temperatures it would be computationally too expensive to calculate the probabilities of all $N!$ permutations, the algorithm proposed for general use is partial infinite swapping (PINS), where symmetrization of all replicas is achieved by partitioning temperature space into locally symmetrized temperature blocks at each step.[12] The INS and PINS methods have been demonstrated to be substantially more efficient than PT for Lennard–Jones (LJ) clusters.[12] Here, the performance of PINS combined with MD is compared to PT for biological systems. While in principle the rare event sampling problem for biological systems and LJ clusters is similar, the connectivity between the atoms in macromolecular systems changes the characteristics of the underlying free energy landscape. Furthermore, the present work presents a concrete implementation into a widely used molecular simulation package, which also allows rigorous testing and comparison of the methods. The main quantity of interest in this study is the structural change of biomolecules.

The use of a symmetrized distribution of configurations in temperature space has two implications for the sampling in PINS compared to PT: (a) Information of all replicas is used to calculate the temperature-specific properties at each step. Therefore, there is a simple statistical advantage of PINS over PT due to the larger number of data points available. (b) The probability of each replica to change its temperature, which is at the same time the probability of each replica to contribute to different temperatures, is higher due to the fact that instead of occasional pairwise exchanges of configurations between temperatures, new permutations of configurations in temperature space are selected as frequently as possible. (The details of selecting these frequencies are explained in the Computational Methods section.)

For point (a), it is clear that this concept is general and leads to an improvement of the sampling for arbitrary systems. The benefits of point (a) are similar to the reweighting of PT data described in the literature.[14] However, technically the PINS algorithm is different from the reweighting procedure described in the literature[14] because the permutation probabilities are used in the PINS method directly to estimate the statistical properties at each temperature. The importance of point (b) is system dependent. Point (b) affects the structural change of each replica quantitatively due to a more uniform distribution of structural change between all replicas and a larger probability for each replica to contribute to many different temperatures. If, for example, some replicas more extensively sample the higher temperatures of the ensemble, their structures change more rapidly than for the rest of the replicas, which as a consequence of the unequal distribution, have more extensively sampled low temperatures. Therefore, more frequent temperature changes lead to a more uniform distribution of structural change between all replicas. A systematic analysis of the occupancy of different temperatures in the computational ensemble by different replicas in PT and INS has been carried out for LJ clusters.[15] Furthermore, point (b) affects the sampling qualitatively because sampling over a long time at a given temperature can lead to sampling a different biological process, e.g., a different reaction pathway, than frequent temperature changes. For a long sampling time at high temperatures a protein could, for example, completely unfold, while if the temperature is changed frequently, it would only partially unfold before returning to the same or a different folded state upon changing to lower temperatures. For the higher temperatures to reach their equilibrium distribution, they have to sample the unfolded state space, whereas the replicas at low temperatures have to sample the folded and partially folded state space. Generating unfolded conformations at the higher temperatures is easy, whereas in contrast sampling different folded or partially folded conformations starting from an unfolded or random initial structure is difficult. In order for PT or INS to be efficient, the high temperature sampling needs to be useful for the convergence of the lower temperatures. This is only the case if the sampling at high and low temperatures is connected, i.e., if each replica samples both the folded and unfolded state space. The importance of the temperature change frequency depends therefore on the difference between low- and high-temperature processes in the given system. The effect of point (b) can be further elucidated by comparing the INS algorithm used here to a recently published new interpretation of the INS algorithm.[16] In this reformulation of INS, MD simulations are carried out over a mixture potential, which is a function of the temperatures of the different replicas present in the simulation. The mixture potential replaces the temperature changes for each replica. The single-temperature properties are then calculated from the sampled data using an estimator, which involves all replicas and is the equivalent to point (a). The importance of the temperature change frequency can therefore be interpreted as the effect of a temperature dependent mixture potential on the sampling process.

These aspects of PINS will be considered for three biological systems: the blocked alanine dipeptide, Villin headpiece, and neuroglobin (Ngb). The three systems will be used to analyze the different aspects of the sampling processes in PINS and PT. The blocked alanine dipeptide is a small system, and therefore, all the local minima of its free energy landscape and the pathways connecting them are known and can be sampled at low computational cost, which makes it a good test system to quantitatively compare the overall sampling efficiency of PT and PINS. For the Villin headpiece and Ngb, the complete sampling of the free energy landscape in the folded and unfolded state requires substantially more sampling. Therefore, the analysis will be focused on comparing the structural change in these two systems for PINS and PT simulations. For the Villin headpiece, the structural change of interest is the folding of the protein. Simulations will therefore be started from the unfolded state and the performance in reaching partially folded and folded structures will be compared for PT and PINS. The usefulness of the PT method for the folding of Villin headpiece has been tested in the past,[17,18] and it has been demonstrated that it is possible to fold the Villin headpiece protein in a 200 ns PT simulation with 20 replicas.[18] For Ngb, the structural changes of interest in this study are changes within the substate space of the folded state, and the effect of changing the temperature of each replica more frequently on the structural change observed in the protein will be evaluated. In order to apply PINS to the systems described above, it is useful to work with a MD version of the PINS algorithm because MD is much more common for simulations of biological systems than the MC methods used for previous evaluations of the PINS sampling efficiency.[12] For this purpose, a MD version of PINS is described in the next section.

■ COMPUTATIONAL METHODS

Combining Infinite Swapping with Molecular Dynamics. The PINS method has been previously described as a sampling method using MC moves to sample conformational space. After selecting a given permutation of configurations in temperature space, (i.e., after assigning a temperature to each replica), a MC move is carried out in configuration space at all temperatures.[12] In the application of PINS to LJ clusters, smart Monte Carlo (SMC) moves[19] were used to propagate the system in configuration space. In SMC, one MC move consists of a number of MD steps. The SMC-PINS algorithm used before is therefore to some extent already a combination of PINS with MD. The difference to the combination of PT with MD[11] consists in the fact that in SMC-PINS new momenta are assigned corresponding to the temperature instead of scaling the existing momenta in PT MD. PINS MD can therefore be implemented in analogy to the procedure proposed for PT MD, where upon exchange of configurations i and j between temperatures T_n and T_m, the new momenta p^i_{new} and p^j_{new} are given by

$$p^i_{\text{new}} = p^i_{\text{old}}\sqrt{\frac{T_n}{T_m}} \tag{5}$$

$$p^j_{\text{new}} = p^j_{\text{old}}\sqrt{\frac{T_m}{T_n}} \tag{6}$$

Hence, the momenta are rescaled according to the ratio of the old and new temperature.[11] For INS, the exchanges between temperatures are not pairwise. Therefore, rescaling of momenta needs to be handled more generally as a function of the old and new permutation k_{old} and k_{new} of configurations in temperature space. For this, the momenta $p_i(k_{\text{new}})$ of each replica i are rescaled upon selection of a new permutation k_{new} according to

$$p_i(k_{\text{new}}) = p_i(k_{\text{old}})\sqrt{\frac{T_i(k_{\text{new}})}{T_i(k_{\text{old}})}} \tag{7}$$

where $T_i(k_{\text{new}})$ denotes the temperature of replica i in the new permutation k_{new}, and $T_i(k_{\text{old}})$ denotes the temperature of replica i in the old permutation k_{old}. In order to obtain a more stable dynamics and re-randomize the momenta of the particles, the rescaling of the velocities is combined in this study with occasional reassignments of the velocities at temperature $T_i(k)$, which are independent of the frequency of selecting new permutations. The numerical stability of the combined velocity rescaling and reassignment procedure has been tested here for two of the systems described below by comparing to distributions of plain MD simulations and the velocity distributions were found to be equivalent.

In analogy to the number of MD steps used previously for one SMC move,[12] the number of MD steps between new assignments of configurations in temperature space needs to be chosen. Because the basic idea of INS and PINS is to achieve faster convergence compared to PT by sampling at the infinite swapping limit,[13] it is desirable to choose this number as small as possible. The details of choosing the number of MD steps in practice are explained in the next section. In summary, the combination of MD to PINS proposed here is straightforward and very similar to the previously used SMC-PINS algorithm.[12]

Implementation into the CHARMM Program. In order to use this MD version of PINS for the simulation of biomolecules, the method has been implemented into the CHARMM program.[20,21] CHARMM already contains provisions for PT simulations with MD and can therefore be used to directly compare the efficiency of PT and PINS simulations.[21] In analogy to the existing PT implementation in CHARMM, PINS has been implemented using a simple strategy for parallel computing in which each computer node performs the calculation of MD trajectories at a given temperature and communication between the computer nodes is required upon selection of a new permutation of configurations in temperature space. This communication includes collecting the current potential energies of each node required for calculating the permutation probabilities $\rho_k(X)$ as well as the distribution of the coordinates and momenta of the new replica i assigned to each node in the new permutation k. Technically, this communication step defines the lower limit of how many MD steps should be carried out between the selection of new permutations because if too few steps are chosen the overall computational effort will be dominated by the communication time between the nodes. In practice, this trade-off between energy evaluation and communication time depends on the available hardware and on the system because the communication effort needs to be small compared to the evaluation of the forces and potential energies needed for the MD steps. For the systems used in this study, the communication time is only considerable in the case of blocked alanine dipeptide compared to the very fast energy evaluations for this system. For the Villin headpiece and neuroglobin, the communication effort has been found to be completely negligible if 50 or more MD steps are carried out between the assignment of new permutations and within a few percent of computational overhead for 10 MD steps. The mixture potential-based PINS version[16] explained in the Introduction section may have some advantages for reducing the communication effort because only the potential energy needs to be communicated between replicas in contrast to the present implementation where energies, coordinates and momenta need to be communicated. Depending on the system and the available hardware, the mixture potential version of PINS[16] may therefore be more efficient. Apart from this small difference in the communication however, the computational effort for the two versions of PINS is very similar because the number of calculations to be carried out for each new assignment of replicas to temperatures should roughly correspond to the number of calculations required to determine the mixture potential.

The PINS implementation described above does not include the use of the permutation probabilities $\rho_k(X)$ to calculate specific system properties for each temperature at each step, as intended in the PINS algorithm.[12] For the following evaluations, the calculation of temperature-specific properties is carried out as a post-processing step based on the potential energies of the given configuration X and the selected permutation k. This separation of the statistical data analysis from the sampling process is due to the fact that for biological systems there are many different properties of interest and it is not always known beforehand which property is important, e.g., which coordinate needs to be selected to optimally describe a structural transition. In order to validate the implementation in CHARMM, simulations were carried out for a LJ-13 cluster, and the results were compared to the initial PINS results. It was found that the potential energies and temperature change probabilities for each replica are distributionally equivalent to the previous results.[12]

System and Simulation Setups. *Blocked Alanine Dipeptide.* For the blocked alanine dipeptide (N-methylalanyl-acetamide), the residue sequence ACE-ALA-CBX in the CHARMM19 force field was used[22] in combination with the analytical continuum solvent potential (ACE) as implicit solvent.[23,24] PT and PINS simulations were carried out using MD with a time step of 1 fs for an ensemble of 12 temperatures between 300 and 1400 K. The temperature gaps were chosen exponentially resulting in the following temperatures for the 12 replicas: 300, 345, 397, 456, 525, 603, 694, 798, 918, 1055, 1213, and 1396 K. For the PT simulations, swap rates between 0.01% and 100% (i.e., one attempt after 10000 MD steps to one attempt after each step) were used. PINS was used with a dual-chain structure[12] in which each of the chains contained a total of 12 temperatures. PINS requires a minimum of two chains with a complementary block structure, i.e., with no common boundaries between the blocks in the first and second chain. The complementary block structure is necessary to allow sampling of the entire temperature space for all replicas. In the first chain, these 12 temperatures were partitioned into a total of three blocks that consisted of (in order of increasing temperatures) three, six, and three temperatures, respectively, while the second chain was partitioned into three blocks of four temperatures each. The handoff between the two chains (i.e., the change between the two different block structures) and the selection of new permutations of configurations in temperature space was performed every 100 MD steps.

Villin Headpiece. For the Villin headpiece (a thermostable subdomain of chicken Villin), the CHARMM27 force field[22,25] was used again in combination with ACE as implicit solvent.[23,24] The folding of Villin headpiece has been tested in a recent study using several force fields including CHARMM27, and it was found that all force fields reproduce the experimental native state structure of the Villin headpiece reasonably well.[26] Initial coordinates for the native states were taken from the protein database NMR structure 1VII.[27] In order to generate an initial structure for the unfolded state, MD simulations were carried out at 1000 K for 2 ns, and the final structure was taken as a starting structure for the unfolded state. Twenty temperatures were used for the PT and PINS simulations. The temperatures were taken from the literature to allow comparison with other studies[18] and are as follows: 273, 283, 293, 300, 307, 313, 319, 325, 332, 339, 348, 360, 374, 390, 406, 422, 440, 460, 480, and 500 K. For the PT simulations, a 1% swap rate was used, which is higher than the 0.1% swap rate used in earlier PT simulations.[18] PINS was used with a dual-chain structure[12] in which each of the chains contained a total of 20 temperatures. In the first chain, these 20 temperatures were partitioned into a total of four blocks that consisted of (in order of increasing temperatures) four, four, six, and six temperatures, respectively. The second chain was also partitioned into a total of four blocks, but the ordering of the blocks was taken to be the reverse of that of the first (i.e., six, six, four, and four). The handoff between the two chains and the selection of new permutations of configurations in temperature space was performed every 10 MD steps. In all simulations, structures were recorded every 100 MD steps.

Neuroglobin. For the PINS simulations of Ngb, initial structures and force field parameters from previous work were used.[28,29] Murine Ngb was built initially from the X-ray structure of mNgbCO (PDB code 1W92[30]) and solvated in a water box of dimensions 43 Å × 62 Å × 58 Å. For the PINS simulations, a cutoff of 12 Å was used together with periodic boundary conditions. An ensemble of the following 32 temperatures was used for the PINS simulations: 284, 286, 288, 290, 292, 294, 296, 298, 300, 302, 304, 306, 308, 311, 313, 315, 317, 319, 322, 324, 326, 328, 331, 333, 335, 338, 340, 343, 345, 347, 350, and 352 K. For the PT simulations, a 1% swap rate was used. PINS was used with a dual-chain structure[12] in which each of the chains contained a total of 32 temperatures. In the first chain, these 32 temperatures were partitioned into a total of six blocks in which the first four consisted of six temperatures, with the next two involving five and three temperatures, respectively. The second chain was partitioned similarly, but the ordering of the blocks was taken to be the reverse of that of the first. The handoff between the two chains and the selection of new permutations of configurations in temperature space was performed every 10 MD steps. In all simulations, structures were recorded every 10 MD steps.

RESULTS

Blocked Alanine Dipeptide. The blocked alanine dipeptide is used as a test system to compare the sampling efficiency of PINS and PT. For blocked alanine dipeptide, it is easy to sample the conformational space, and the highest temperature of 1396 K is more than sufficient to overcome the barriers between the different local minima. The exchange probability between neighboring temperatures is between 60% and 80%. The simulations are analyzed using Ramachandran plots[31] of the different conformations.

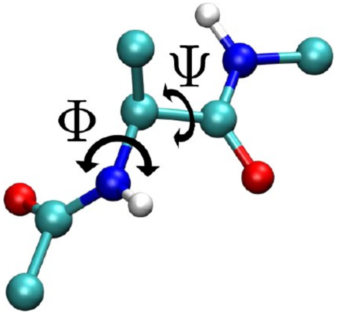


Figure 1. Blocked alanine dipeptide with the two dihedral angles (ϕ,ψ) used in the Ramachandran plots.

For dipeptides, a single pair of Ramachandran dihedral angles (ϕ and ψ, Figure 1) is sufficient to describe all relevant features of conformational space. The population at 300 K of each conformation in (ϕ,ψ) space is calculated from the simulation data, with conformations being recorded every 10 MD steps. From this analysis, the population density $\rho(\varphi,\psi)$ is obtained and converted into a free energy $G(\phi,\psi) \propto - RT\ln(\rho(\phi,\psi))$, where T is the absolute temperature in K and R is the ideal gas constant. Because some unfavorable regions of conformational space are not sampled, the free energy is set to a threshold value in these regions which is higher than the rest of the free energy landscape. These regions are the white patches in Figures 2−4. In previous work, it has been mathematically proven[13] and empirically demonstrated[12] that PINS and PT converge to the same distribution. The comparison of sampling efficiency in this work is based on these earlier findings. First, results on PT sampling are reported in Figure 2. The simulation data is obtained from three independent 5 ns trajectories (short sampling) and three trajectories of 50 ns (long sampling) with a swap rate of 1%. In agreement with an earlier study carried out with the identical force field parameters and implicit solvent model (see System and Simulation Setup section) as the present study,[32] the free energy landscape of blocked alanine dipeptide shows four local minima in the (ϕ,ψ) projection: C_{7eq} minimum at (−86.4°, −136.8°), α_R minimum at (−79.2°, 43.2°), α_L minimum at (50.4°, 50.4°), and C_{7ax} minimum at (57.6°, −79.2°) in our simulations. For both sampling times in Figure 2, all the four local minima are sampled, but for the short sampling, the least favorable region around α_L is rarely visited.

As a next step, the effect of the PT swap rate is evaluated using the three independent 50 ns trajectories. The results in Figure 3 show a gradual improvement of the sampling as the swap rate is increased from 0.01% to 10%. For the lowest swap rate, the α_L minimum at (50.4°, 50.4°) is not sampled at all. With increasing swap rate, the sampling of the C_{7ax} and the α_L minimum is improved as well as the sampling of the transitions between the local minima. Overall, the results show that the convergence of the PT sampling is improved with increasing swap rates. This agrees with results from theoretical work that are the basis for INS[13] and the findings of earlier studies for LJ clusters.[12]

While these results confirm the general concepts of INS, a further evaluation is required to quantify the improvement in sampling efficiency of PINS compared to PT with the maximal swap rate of 100%. Such a comparison is shown in Figure 4. Results from three times 5 ns are used, and PT with swap rates of 10% and 100% are compared with PINS. The results show that PINS sampling performs significantly better than PT with a 100% swap rate. The sampling covers a larger part of phase space in general, and specifically, the transition states between the minima are sampled much better. If the PINS results from this short sampling time are compared with the PT results obtained from the ten times longer sampling in Figure 3, the PINS results are still significantly superior, in particular for the transition state regions. The overall conclusion from the blocked alanine dipeptide simulations is therefore that the convergence of PINS is at least an order of magnitude better than that of PT. The convergence of PT can be improved gradually by increasing the swap rate. However, even with the

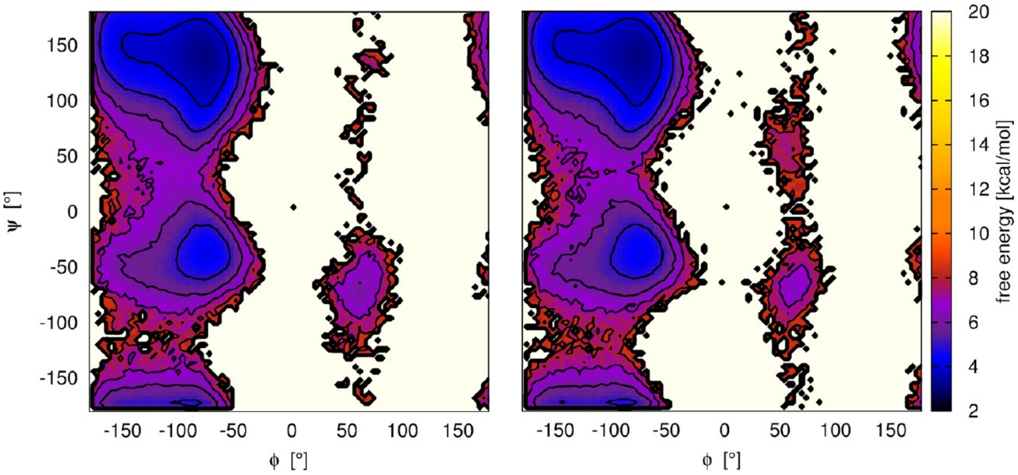


Figure 2. Change of free energy landscape at 300 K with sampling time for blocked alanine dipeptide in (ϕ,ψ) space from a 12 temperature PT simulation after a total of 15 ns of sampling time (left) and a total of 150 ns of sampling time (right).

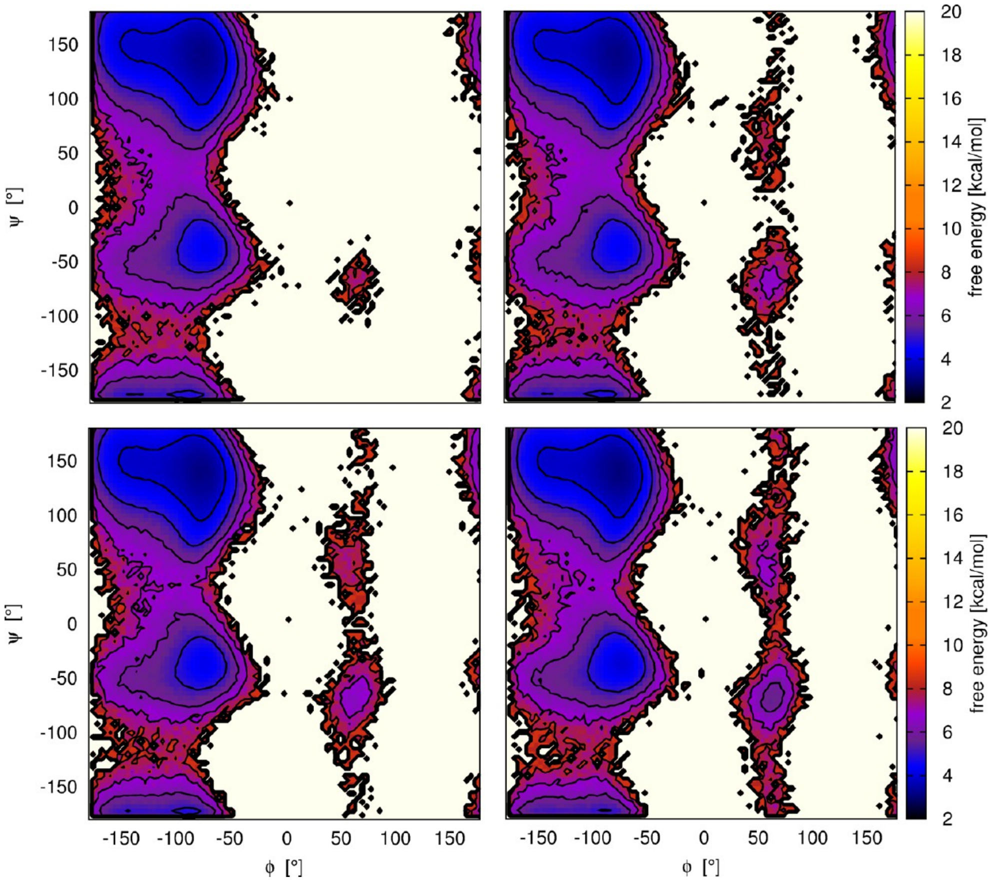


Figure 3. Change of the free energy landscape at 300 K as a function of the PT swap rate for blocked alanine dipeptide in (ϕ,ψ) space obtained from 12 temperature PT simulations after a total of 150 ns of sampling time: 0.01% swap rate (top left), 0.1% swap rate (top right), 1% swap rate (bottom left), and 10% swap rate (bottom right).

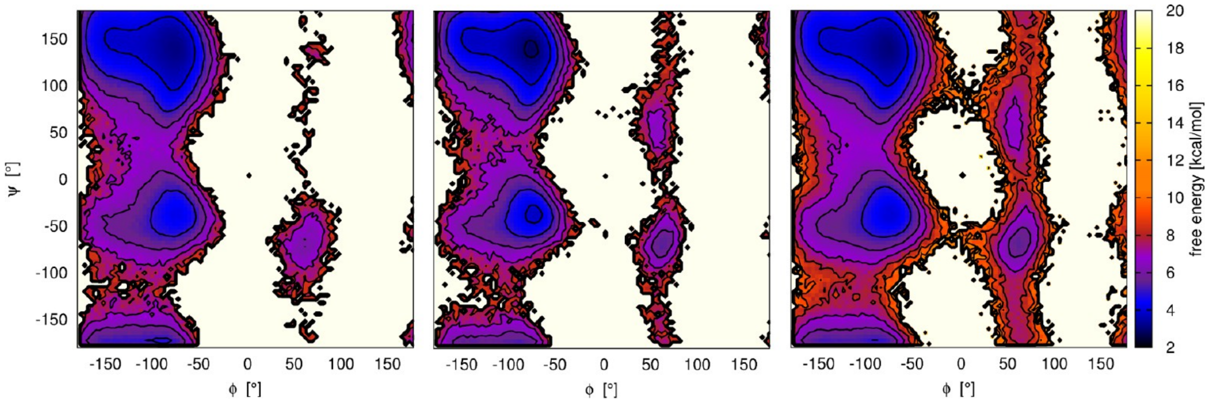


Figure 4. Comparison of the free energy landscape obtained with PINS and PT with different swap rates for blocked alanine dipeptide in (ϕ,ψ) space after a total of 15 ns of sampling time: PT with 1% swap rate (left), PT with 100% swap rate (center), and PINS (right).

maximum swap rate its performance is not comparable to that of PINS.

Folding of Villin Headpiece. *MD Simulations at Different Temperatures and PT Simulations.* For the Villin headpiece, the force field and solvation model was first tested with MD simulations at different temperatures starting from the native state (PDB code 1VII). The initial structure was optimized, and MD simulations were carried out for 1 ns at 300, 500, 700, and 1000 K. It was found that the protein remains in the folded state at 300 K, whereas it partially unfolds at 500 and 700 K and completely unfolds at 1000 K. An additional trajectory of 10 ns at 300 K confirmed the stability of the folded state over longer sampling times. It is known from an earlier study that PT can be used to fold the Villin headpiece in about 200 ns, which is substantially faster than the microsecond long folding process at 300 K.[18] Because a different force field, the AMBER FF03,[33] was used in the earlier PT study of Villin headpiece,[18] PT simulations were first carried out for 200 ns in order to test whether it is possible to find folded or partially folded structures with CHARMM27. The simulations show a gradual evolution of the structures observed at 300 K toward the native state and therefore demonstrate that correct folding of the Villin headpiece is highly likely with the present simulation setup. The details of these structures are provided in the Supporting Information. In order to assess the structural change between 50 and 150 ns quantitatively, three different properties of the structures at 300 K are analyzed:

- *The root-mean-square distance (RMSD) of the C_α atoms between the native state and the given structure including all except the two terminal residues.* To calculate the RMSD between a given structure a and the native state structure n (PDB code 1VII), the two structures are first reoriented to their minimum overall RMSD, and the RMSD $(a,n) = ((\Sigma_{i=1}^{N} r_i^2/N))^{1/2}$ is calculated with N being the number of C_α atoms and r being the distance between the two structures for atom i based on the reoriented coordinates.
- *The number of native contacts $N_{nc}^{C_\alpha}$ of the C_α atoms in the given structure.* To obtain $N_{nc}^{C_\alpha}$, the distance matrix of the C_α atoms in the given structure and the native state structure is calculated. $N_{nc}^{C_\alpha}$ is then defined as the number of distances that are equal in the two distance matrices. Because during a MD simulation these distances are never exactly equal, a tolerance criterion is required. This tolerance criterion is set to 10%.
- *The number of native contacts N_{nc}^{h} of the heavy atoms.* N_{nc}^{h} is calculated in analogy to $N_{nc}^{C_\alpha}$ based on the distance matrix of the heavy atoms, except that the tolerance is 20%. The optimal value of the tolerance has been evaluated by comparing the change in N_{nc} during MD simulations of the folded and unfolded state at 300 K and selecting the tolerance that allows the clearest distinction between the two simulations.

The results of the three measures are reported in Figure 5. They show that the structural change toward the folded state is visible in the change of all the three analysis criteria. It correlates with the increase in $N_{nc}^{C_\alpha}$ and N_{nc}^{h} and the decrease in RMSD.

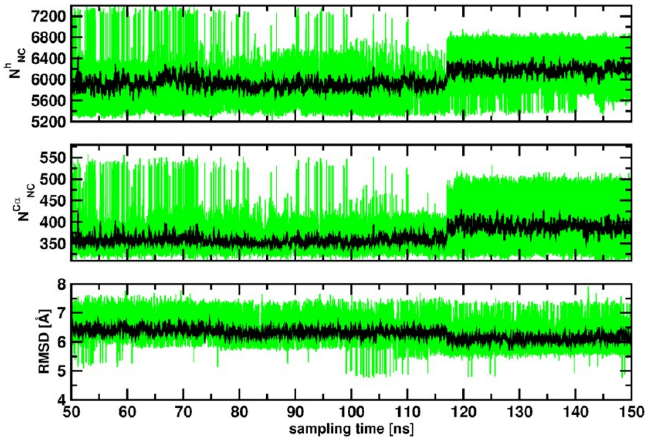


Figure 5. Analysis of the structural change at 300 K of Villin headpiece between 50 and 150 ns of PT simulation time. Number of native contacts of all heavy atoms N_{nc}^{h} (top), number of native contacts of the C_α atoms $N_{nc}^{C_\alpha}$ (center), and RMSD between C_α atoms in PT structure and native structure (bottom). The green lines show the results for each recorded structure; the black lines are averages over 100 structures.

Comparison of PT and PINS Simulations. For comparing the folding process in the PT and PINS simulations, the same methods as explained above are used to analyze the structural change. The structural change is compared for the lowest temperature (273 K) in Figure 6 and for 300 K in Figure 7. The results show that structures close to the native state are found substantially faster in the PINS simulations. This can clearly be seen for the structures at the lowest temperature in Figure 6, where an RMSD below 5 Å is reached after 20 ns. The corresponding structure contains two of the three helices found in the native state and a tertiary structure very close to the native state. A Figure of this structure is provided in the Supporting Information. As the simulation progresses, the different structural analyses in Figure 6 show that the structures remain similarly close to the native state on average. After about 70 ns, a gradual broadening of all three structural properties is observed. This broadening is related to the fact that initially only one of the replicas finds a very favorable structure that stays predominantly at the lowest temperature for the following 30 to 40 ns. As the simulation progresses, more replicas find favorable structures with an increasingly high probability of

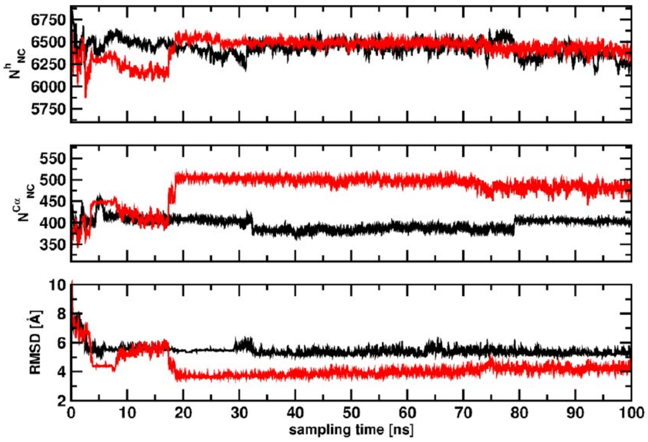


Figure 6. Comparison of structural change at the lowest temperature during 100 ns of PT (black lines) and PINS (red lines) simulation. Number of native contacts of all heavy atoms N_{nc}^{h} (top), number of native contacts of the C_α atoms $N_{nc}^{C_\alpha}$ (center), and RMSD between C_α atoms in PT structure and native state structure (bottom).

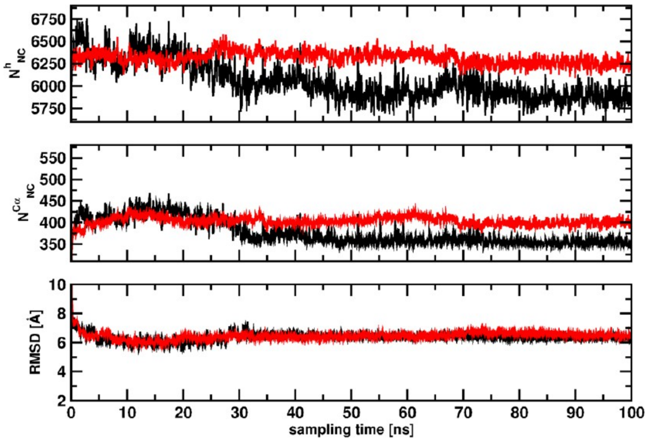


Figure 7. Comparison of structural change at 300 K during 100 ns of PT (black lines) and PINS (red lines) simulation. Number of native contacts of all heavy atoms N_{nc}^{h} (top), number of native contacts of the C_α atoms $N_{nc}^{C_\alpha}$ (center), and RMSD between C_α atoms in PT structure and native state structure (bottom).

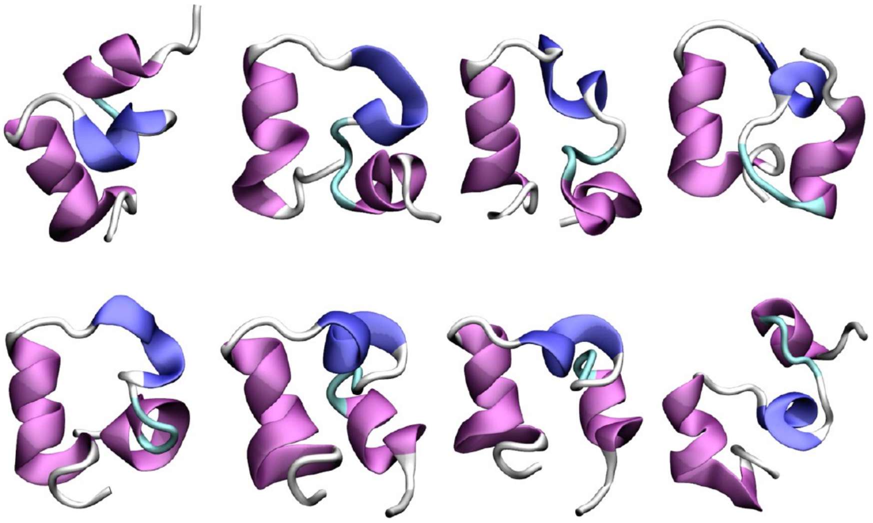


Figure 8. Villin headpiece structures from different replicas found between 90 and 100 ns of PINS simulation time at 300 K.

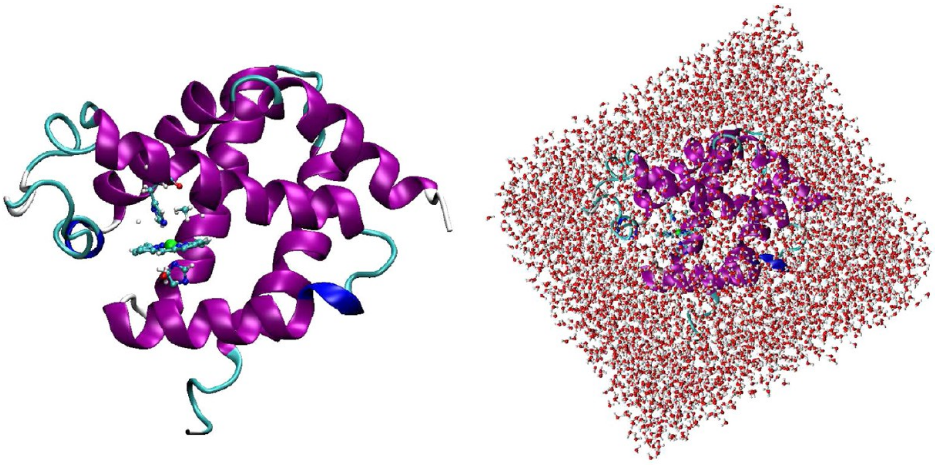


Figure 9. Neuroglobin protein with and without solvation box.

being at the lowest temperature. This broadening of the structural space can also be observed for the structures found at 300 K. Structures from different replicas found between 90 and 100 ns of simulation time at 300 K are shown in Figure 8. The variety of structures with two or three of the secondary structure elements demonstrates that a large number of replicas contribute to the folded state ensemble at 300 K, and therefore, the information generated by the PINS simulation could potentially not only be used to characterize different folded and partially folded structures but also the different pathways leading from the unfolded to the folded state. The evaluation in Figure 6 clearly shows that PINS is significantly more efficient than PT in finding favorable structures. For this evaluation, the use of data from all temperatures for the calculation of single-temperature properties (point (a)) in the Introduction has not been considered. The fact that favorable structures are found faster in PINS compared to PT sampling is therefore explained on one hand by the more uniform distribution of the structural change to all replicas, which is a consequence of the frequent temperature changes, and on the other hand by the fact that favorable structures are assigned faster to the low temperature part of the computational ensemble.

Effect of Temperature Change Frequency on Sampling Substates of the Native State in Neuroglobin. In the previous paragraph, the sampling efficiency of PINS has been tested for the sampling problem of finding folded and partially folded states of a protein from an unfolded initial structure. Another sampling problem in biological systems consists in finding various substates of the native state. Also for this problem, the frequency of changing the temperature of each replica affects the overall efficiency of the sampling. In the extreme case where some replicas remain always at high temperatures and others remain always at low temperatures, high-temperature replicas will unfold and be essentially useless for the purpose of finding additional substates of the native state, while low-temperature replicas will not show any faster structural change than a conventional single-temperature MD simulation. Furthermore, a faster exchange of replicas between different temperatures leads to a larger contribution of the conformational space sampled by each replica to the single-temperature properties. In order to quantify these effects, Neuroglobin is used as a test system. The structure of neuroglobin and the solvated neuroglobin system are shown in Figure 9.

The conformational space sampled during 500 ps in a PT simulation is compared to the sampling during a 500 ps PINS simulation. While 500 ps sampling time is most likely not sufficient to sample the substate space of native neuroglobin exhaustively, it is sufficient to compare the performance of PT and PINS. The conformational space sampled at each temperature is analyzed using a regular spatial clustering algorithm of all the 5000 structures recorded at each temperature during the PT and PINS simulations. The regular spatial clustering is carried out with a minimum RMSD metric and a cutoff of 1.5 Å. The number of cluster centers for a set of structures is determined as follows: the first structure of the set is defined as the first cluster center. For each of the following structures, the minimum RMSD to all existing cluster centers is then evaluated. If the RMSD is larger than 1.5 Å, the structure is taken as a new cluster center, otherwise it is assigned to an existing cluster center. The number of cluster centers determined by this algorithm is therefore an unambiguous way to quantify the size of the conformational space corresponding to a given set of structures. Regular spatial clustering was carried out for the 5000 structures at each temperature of the PT and PINS simulations using the program EMMA.[34] Two different atom selections were used and compared: Selection (A) includes all the protein atoms including the Heme group and all protein hydrogens. Selection (B) includes only the C_α atoms of the protein. The results are shown in Table 1.

Table 1. Number of Cluster Centers Determined by Regular Spatial Clustering at Each Temperature

	selection A		selection B	
temperature [K]	PT	PINS	PT	PINS
283.7171	6	18	1	3
285.7031	11	19	1	3
287.7030	10	28	1	6
289.7169	11	30	3	4
291.7450	13	33	3	4
293.7871	17	38	3	5
295.8437	19	40	2	6
297.9146	19	43	3	5
300.0000	16	44	3	7
302.1000	18	48	4	7
304.2147	21	51	3	7
306.3442	20	50	4	7
308.4886	19	53	4	8
310.6480	19	56	6	7
312.8226	23	56	5	11
315.0123	26	62	5	10
317.2174	26	66	5	8
319.4379	28	68	5	7
321.6740	27	76	4	9
323.9257	23	72	4	9
326.1932	30	77	5	8
328.4766	29	80	5	9
330.7759	30	83	6	11
333.0913	33	90	5	8
335.4226	34	89	6	11
337.7709	35	90	5	8
340.1353	34	94	7	11
342.5162	40	87	6	11
344.9139	44	80	8	9
347.3283	44	72	7	9
349.7596	38	64	7	10
352.2079	25	52	7	8
all temperatures	168	161	16	14

The evaluation shows that the number of cluster centers as determined from PINS is larger than that from PT for all temperatures. In most cases, the number of clusters from PINS is twice as large compared to PT. As a consequence, the conformational space at each temperature is sampled twice as fast with the PINS algorithm. It should be noted that the overall number of cluster centers is, however, very similar for PT and PINS. This is due to the fact that the amount of overall sampling in conformational space by all replicas is the same for PINS and PT, i.e., the same number of new conformations is generated overall by all replicas. The difference for the calculation of single-temperature properties is on one hand that each replica in PINS contributes to a larger number of temperatures, and on the other hand, the sampling of new conformations is distributed more equally between all temperatures.

■ DISCUSSION AND CONCLUSIONS

The details of applying the INS concepts to biological systems using a combination of the PINS algorithm with MD simulations have been explained, and the performance of PINS has been tested for three different biological systems. The three systems have been selected with a focus on different aspects of sampling problems in biology. For blocked alanine dipeptide, the efficiency of sampling the free energy landscape with the current implementation of PINS can be directly compared with PT for different simulation conditions, specifically the swap rate. It is important to note that this is done within the same simulation package, which minimizes the bias in comparing the computational approaches. The general idea of INS, that is, the fact that more exchange between replicas improves the convergence, has been confirmed. Furthermore, it has been shown that PINS converges significantly faster than PT even if PT is used with very high swap rates. For the Villin headpiece, it has been shown that partially folded and folded conformations are obtained significantly more rapidly with PINS than with PT. This difference in sampling efficiency can be attributed on one hand to the fact that favorable conformations are being transferred faster to the low temperatures of the computational ensemble and on the other hand to the more uniform distribution of structural change between all replicas, which is a consequence of the higher probability for each replica to change its temperature. The effect of the temperature change frequency for each replica and the enhanced exchange of replicas between temperatures has been further investigated using neuroglobin, which was used as a test case for sampling the substate space of the native state of a protein. It was found that more frequent exchange of replicas between temperatures leads to a more efficient sampling of the conformational space for each temperature. For the parameters used in these tests, the conformational space was explored about twice as fast by the PINS simulations compared to PT for all temperatures. Overall, the systems and simulations presented here show that PINS is a useful method to address the rare event sampling problem in biological systems and that it is more efficient than PT.

While these results are overall promising, a number of open questions concerning the use of the INS concepts and the PINS algorithm for biological systems remain and will need to be addressed in future work. So far, PINS has been used to calculate thermodynamic properties of biological systems. However, in general, kinetic properties are also of interest, e.g., the calculation of transition times and rates between different local minima of the blocked alanine dipeptide or the characterization of different folding pathways for the Villin headpiece. For PT, a number of methods have been proposed that allow the use PT efficiently in order to calculate the kinetic properties of biological systems. These methods include a combination of PT with transition path sampling[35] and the reweighting of kinetic data of multiple temperatures to obtain transition times and transition rates at 300 K.[36,37] These methods cannot be used in their current form with the PINS algorithm; therefore, further method developments will be required to make the INS concept useful for the calculation of kinetic properties.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Information related to the Villin headpiece simulations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: nuria.plattner@fu-berlin.de (N.P.); jimmie_doll@brown.edu (J.D.D); m.meuwly@unibas.ch (M.M.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the stimulating discussions with Professor P. Dupuis and Dr. J. Chodera concerning the topics in the present paper. N.P. thanks the Swiss National Science Foundation for postdoctoral support and J.D.D. wishes to acknowledge grant support of this research through the DOE Multiscale and Optimization Program No. DE-SC0002413. M.M. gratefully acknowledges support by the Swiss National Science Foundation through the NCCR-MUST and Grant 200021-117810.

■ REFERENCES

(1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−652.
(2) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151−156.
(3) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6679−6685.
(4) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910−3916.
(5) Torrie, G. M.; Valleau, J. *J. Comput. Phys.* **1977**, *23*, 187−199.
(6) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562−12566.
(7) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.
(8) Geyer, C. J.; Thompson, E. A. *J. Am. Stat. Assoc.* **1995**, *90*, 909−920.
(9) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57−83.
(10) Klenin, K.; Strodel, B.; Wales, D. J.; Wenzel, W. *Bioch. Biophys. Acta* **2011**, *1814*, 977−1000.
(11) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(12) Plattner, N.; Doll, J. D.; Dupuis, P.; Wang, H.; Liu, Y.; Gubernatis, J. E. *J. Chem. Phys.* **2011**, *135*, 134111.
(13) Dupuis, P.; Liu, Y.; Plattner, N.; Doll, J. D. *SIAM J. Multiscale Model. Simul.* **2012**, *10*, 986−1022.
(14) Chodera, J. D.; Shirts, M. R. *J. Chem. Phys.* **2011**, *135*, 194110.
(15) Doll, J. D.; Plattner, N.; Freeman, D. L.; Liu, Y.; Dupuis, P. *J. Chem. Phys.* **2012**, *137*, 204112.
(16) Lu, J.; Vanden-Eijnden, E. *J. Chem. Phys.* **2013**, *138*, 084105.
(17) Lin, C.-Y.; C.-K., H.; Hansmann, U. H. E. *Prot. Struct. Funct. Genet.* **2003**, *52*, 436−445.
(18) Lei, H.; Wu, C.; Liu, H.; Duan, Y. *Proc. Natl. Acad. Sci.* **2007**, *104*, 4925−4930.
(19) Rossky, P. J.; Doll, J. D.; Friedman, H. L. *J. Chem. Phys.* **1978**, *69*, 4628−4633.
(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(21) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(22) MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(23) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.
(24) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comput. Chem.* **2001**, *22*, 1857−1879.
(25) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III *J. Chem. Theory Comput.* **2004**, *25*, 1400−1415.
(26) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−L49.
(27) Mcknight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol.* **1997**, *4*, 180−184.
(28) Lutz, S.; Nienhaus, K.; Nienhaus, G. U.; Meuwly, M. *J. Phys. Chem. B* **2009**, *113*, 15334−15343.
(29) Lutz, S.; Meuwly, M. *Faraday Discuss* **2011**, *150*, 375−390.
(30) Vallone, B.; Nienhaus, K.; Matthes, A.; Brunori, M.; Nienhaus, G. U. *Proc. Natl. Acad. Sci.* **2004**, *101*, 17351−17356.
(31) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. J. M. *J. Mol. Biol.* **1963**, *7*, 95−99.
(32) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. *Proc. Natl. Acad. Sci.* **2007**, *104*, 1817−1822.
(33) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(34) Senne, M.; Trendelkamp-Schroer, B.; Mey, A.; Schutte, C.; Noe, F. *J. Chem. Theory. Comput.* **2012**, *8*, 2223−2238.
(35) Vlugt, T. J. H.; Smit, B. *Phys. Chem.Comm.* **2001**, *2*, 1−7.
(36) Chodera, J. D.; Swope, W. C.; Noé, F.; Prinz, J.-H.; Shirts, M. S.; Pande, V. S. *J. Chem. Phys.* **2011**, *134*, 244107.
(37) Prinz, J.-H.; Chodera, J. D.; Pande, V. S.; Swope, W. C.; Smith, J. C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 244108.