Performance and Free Energy Estimation for Solvated Polypeptides and Proteins Using Partial Infinite Swapping

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

†*Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland.*

‡*Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin.*

¶*Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.*

E-mail: m.meuwly@unibas.ch

Abstract

Partial infinite swapping (PINS) is a powerful enhanced sampling method for complex systems. In the present work thermodynamic observables are determined from reweighting at the post-processing stage for folding of $(Ala)_{10}$ in implicit and explicit solvent and for Xenon migration in myoglobin. In every case free energy surfaces are determined using PINS with an accuracy comparable to Molecular Dynamics and Parallel Tempering simulations but at considerably reduced computational cost. Round trip times through the ensemble of temperature space are shown to be almost one order of magnitude shorter for PINS compared to PT simulations for $(Ala)_{10}$ in implicit solvent which suggests that PINS is more efficient for sampling diverse structures. Consistent with NMR experiments on shorter $(Ala)_7$ poly-alanine peptides, simulations of $(Ala)_{10}$ in explicit solvent highlight the essential role played by the environment in stabilizing extended conformations which are unfavourable in implicit solvent. Additional low-energy regions are $\beta-$hairpins, 1 to 2 kcal/mol above the minimum energy structure. For Xenon migration in Myoglobin, PINS finds stabilization energies of the experimentally known Xenon-pockets to range from -4.6 to -6.2 kcal/mol, in accord with experiment. Furthermore, the barrier heights between neighboring pockets have been determined to be ≈ 4 kcal/mol. By starting simulations from individual pockets, PINS finds sampling of the entire pocket network on the time scale of 3 ns using 32 replicas whereas with MD, 100 ns are not sufficient to access all pockets. Hence, PINS with reweighting is found to be both, a quantitatively accurate and computationally efficient method for studying complex biological molecules in solution.

Introduction

Molecular Dynamics (MD) and Monte Carlo methods (MC) are widely used for characterizing biological processes using computer simulations. Although the computational resources are continuously increasing, sampling the large conformational space available for proteins is very challenging. However, in order to provide an atomistically refined picture for processes

such as protein folding or large conformational changes with functional relevance, such rare events must be sampled. As they usually occur on time scales of the order of microseconds (or longer), directly sampling them with unbiased MD or MC simulations is difficult. Hence, enhanced sampling methods are a possible way forward as such approaches increase the probability for accessing low-probability configurations.

An important aspect which is subject to continuous improvement efforts is the ability to sample rare-events, a particular challenge for complex systems. For systems in which configuration space is well connected, standard techniques (e.g. Metropolis-Hastings[1–3]) are efficient. However, for situations in which configuration space decomposes into poorly connected subregions or where barriers between neighbouring states are high, enhanced sampling is required. Several such methods for rare event sampling have been developed in the past. They include parallel tempering (PT),[4–6] replica exchange (RE)[7] (including the recent development of asynchronous variants such as ASyncRE[8]), spatial averaging[9–11], umbrella sampling (US),[12] metadynamics,[13] and many more. The methods either use a bias to steer the system between regions in configuration space (US, metadynamics) or they expand thermodynamic state space as is done for PT or RE based methods. This contrasts with conventional stochastic methods which typically use random walks for generating a statistical sampling of the desired equilibrium probability distribution.

Another method which has recently been developed is Partial Infinite Swapping (PINS)[14–18] which is based on the PT/RE algorithms. PINS uses a symmetrisation strategy for combining probability distributions at different temperatures, so that they become more connected and thus easier to sample than the original ones. The present work discusses the statistical reweighting to extract thermodynamic information from PINS simulations[15,17] and applications to two systems: the alanine decapeptide (or deca-alanine) which becomes a challenging system particularly in explicit solvent,[19–26] and Xenon migration in Myoglobin,[27–34] a system

for which experimental data to compare with is available and which requires extensive direct sampling of the free energy surface.

Computational Methods

Considering a Canonical (NVT) ensemble, the probability $\rho(\mathbf{X})$ of observing a system in state $\mathbf{X}$ is related to its potential energy $V(\mathbf{X})$ through

$$\rho(\mathbf{X}) = \frac{1}{Z} e^{-\beta V(\mathbf{X})} \tag{1}$$

where $\mathbf{X} = X_1, ..., X_k$ is a $k-$dimensional vector (where $k = 3$ for MC or $k = 6$ for MD), populating a subset D of the configuration space $\mathbb{R}^{kN}$, Z is the canonical partition function $Z \sim \int^{D \subset \mathbb{R}^{kN}} e^{-\beta V(\mathbf{X})} d\mathbf{X}$, and $\beta = 1/k_{\mathrm{B}}T$ is the inverse temperature and k_{B} the Boltzmann constant.

Parallel Tempering (PT) (also known as Replica Exchange (RE)) methods[4–6] were successfully applied to investigating a wide range of chemical and biological systems. In PT K replicas are followed and the partition function Z of the overall ensemble is:

$$Z = \prod_{i=1}^{K} \frac{q_i}{M!} \int d\mathbf{X}_i \; e^{-\beta_i V(\mathbf{X}_i)} \tag{2}$$

where $q_i = \prod_{k=1}^{M} (2\pi m_k k_B T_i)^{3/2}$ is obtained by integrating out the momenta of the M particles with mass m_k, $V(\mathbf{X}_i)$ is the potential energy for the coordinates $\mathbf{X}_i$, and $\beta_i = 1/k_{\mathrm{B}}T_i$ is the reduced temperature for replica i. In the simulations, replicas are exchanged between two adjacent temperatures $T_i \leftrightarrow T_j$ with probability

$$P_{acc}(i \leftrightarrow j) = \min\{1, e^{(\beta_i - \beta_j)(V(\mathbf{X}_i) - V(\mathbf{X}_j))}\} \tag{3}$$

The K temperatures are usually distributed non-linearly between T_1 and T_K with $T_K > T_1$. Here T_1 is the desired simulation temperature and the ratio $\frac{T_{i+1}}{T_i}$ is kept constant which yields a constant acceptance value of $P_{acc}(i \leftrightarrow j)$, see Ref.[6] for a discussion on the choice of temperatures and the impact on P_{acc}. It is also worth mentioning that when exchanging coordinates between replicas velocities are also exchanged, but rescaled to the temperature of the destination replica.

Infinite Swapping limit for Parallel Tempering simulations

The infinite swapping (INS) method also uses an expanded ensemble built from a number of replicas at different temperatures.[14] Contrary to PT, INS is based on the fully symmetrized distribution of configurations in temperature space, whereas PT only occasionally enriches the local temperature with configurational information from simulations at a higher temperature. Formally, INS is based on a mathematical analysis of the convergence rate of PT simulations as a function of the temperature swap attempt frequencies.[14,16,18] It was proven[16,18] that this convergence rate is a monotonically increasing function of the swap rate, and thus optimal sampling is reached in the *infinite swapping* limit (i.e. swap at every MD time step).

In other words, INS provides optimal sampling for a given replica by using information from all other temperatures used in the simulation. This can be achieved by allowing exchanges between all replicas at each time step. For K replicas the number of permutations of a set of configurations $\mathbf{X}$ is $K!$ if all possible exchanges were attempted and the probability $\rho_k(\mathbf{X})$ of permutation k is

$$\rho_k(\mathbf{X}) = \frac{p_k(\mathbf{X})}{\sum_{k=1}^{K!} p_k(\mathbf{X})}. \tag{4}$$

with

$$p_k(\mathbf{X}) = \prod_{i=1}^{K} e^{-\beta_i V(\mathbf{x}_{k,i})} \tag{5}$$

and $\mathbf{x}_{k,i}$ is the configuration of replica i corresponding to the assignment of configurations to temperatures in permutation k. The permutation k with the highest acceptance probability p_k is found by evaluating and comparing all $\rho_k(\mathbf{X_i})$. However, for large systems this includes a large number of permutations, and it is computationally too expensive to consider all $K!$ probabilities.

For putting INS to practical use, the partial infinite swapping (PINS) algorithm was introduced.[14–17] PINS uses a partitioning strategy whereby temperature space is divided into blocks, and local (but full) symmetrisation is used within each block. More precisely, the current implementation uses the "dual-chain" approach[15], where the $K-$temperature set is partitioned into blocks of temperatures in two different ways, one for each chain. The two blocks must have a complementary structure without a boundary between the blocks defined for the two chains. This is required in order to achieve sampling of the overall temperature space for all the replicas. For a set of 12 temperatures, a possible partitioning for the two chains $(a|b)$ is $(3, 6, 3|4, 4, 4)$, where the common boundaries for chain a are between T_3 and T_4, T_9 and T_{10}, and for chain b they are between T_4 and T_5 and T_8 and T_9, respectively. On the other hand, the partitioning $(3, 3, 6|6, 3, 3)$ is not valid, as chains a and b share a common boundary between T_6 and T_7.

Similar to standard PT simulation, PINS requires K replicas, and the temperatures $\{T_1, \cdots, T_K\}$ at which they are run. The user also provides a frequency of attempted exchanges between replicas. The sampling efficiency of PT simulations is optimal with attempted exchanges at each MD step, see above. However, this requires communication of the coordinate vectors, using technologies such as message passing (MPI) which is a bottleneck for the simulation of large systems as inter-node communication is usually slower than computation. It is thus required to attempt exchanges as often as possible, but not too often to avoid inter-node communication saturation. For a concrete application the best choice depends on (i) the system

size because the smaller the system, the larger the ratio of communication time/calculation time, and (ii) hardware/software considerations, mainly the maximum communication speed possible between two replicas running on different compute nodes.

Statistical reweighting

An important aspect in making practical use of PINS for concrete applications is the evaluation of unbiased statistical averages. This point has not been considered specifically in previous work but is essential when comparing efficiency and accuracy of a particular computational methodology and comparing with experiment for specific observables.[17] PINS provides data at all temperatures T_j which are used for computing properties at a given thermodynamic state j. When computing averages of observables this requires reweighting of the data collected at different T_j. This step can either be performed during the simulation ("on the fly"), or at the end of the simulation ("post-processing"). For the current implementation it was decided to employ post-processing because on the fly processing would result in non-negligible computational overhead when studying large systems. This implementation is particularly efficient if various system properties are potentially of interest since it allows to decide after the simulation which evaluations are to be carried out. Finally, the post-processing step is optional (if no thermodynamic estimate is desired), compared to biased methods where the unbiasing step is a requirement in order to computer accurate properties.

For the reweighting the list of permutations attempted at a given swapping step of the simulation is required. This list can be rebuilt at the post-processing stage from the following information: (i) the total number of simulation steps, and the swapping frequency at which the PINS algorithm was applied (100 steps was found to be a good compromise between sampling and performance[17]), (ii) the number of temperatures K, (iii) the dual chain parameters, i.e. the number of temperature blocks, and the number of temperatures within

each block (see above). It is also necessary to save the potential energy $V(\mathbf{X}_i)$ of each atomic configuration when swapping. With this information it is possible to calculate a system property $C(i, T_j)$ at temperature T_j based on its replica-specific value $c(\mathbf{x}_{k,i})$, with $\mathbf{x}_{k,i}$ being the coordinates of the replica assigned to T_j in permutation k and step i. The permutations to be considered here belong to the block containing T_j in the dual-chain configuration of step i, which results in

$$C(i, T_j) = \sum_{k=1}^{P_{block}} \rho_k(\mathbf{X_i}) c(\mathbf{x}_{k,i}), \tag{6}$$

with P_{block} being the number of permutations in the given block of the current chain and $\rho_k(\mathbf{X_i})$ calculated using Eq. 4 for the given block. For PINS to be computationally efficient it is essential that $\sum P_{block} << K!$ is fulfilled. As an example, for the $(3, 6, 3|4, 4, 4)$ dual-chain mentioned above, $\sum P_{block} = (3! + 6! + 3!) + (4! + 4! + 4!)$ i.e. $P = 804$ permutations, to be compared with $K! = 12! \approx 4.8 * 10^8$ permutations for full INS swapping.

For calculating the 2D free energy surface $F(x, y)$ (FES) depending on two variables x and y (illustrated later with an application to the alanine-decapeptide) this is applied as follows: values of x and y are first computed for each configuration from each trajectory. From a normalised 2D histogram the free energy is therefore

$$\Delta F^s(x_i, y_j) = -RT^s \ln(\rho(x_i, y_j)) \tag{7}$$

where F^s and T^s are the single state free energy estimate and the corresponding temperature, respectively, and $\rho(x_i, y_j)$ is the probability density for cell (i, j) on the 2D grid. An estimate based on data generated at *multiple* thermodynamic states (m superscript) is obtained from

$$\Delta F^m(x_i, y_j) = \Delta F^s(x_i, y_j) \cdot \sum_{k=1}^{K} \rho_k(x_i, y_j) \tag{8}$$

where (i, j) runs over the discretised cells, and m over all single states (defined by Eq. 7).

This procedure is easily adapted and extended to other thermodynamic properties which can be estimated from a discretised grid. As will be demonstrated in the applications, this reweighting improves the FES estimates in low probability regions, such as transition states, which are particularly important for conformational transitions.

For the statistical reweighting in a post-processing step separate analysis software was written which is currently available for the CHARMM implementation of PINS. For completeness it is also mentioned that all simulations carried out in the present work employ an updated implementation of PINS because CHARMM underwent profound changes in its architecture. Hence, it was decided to implement PINS with a dual-chain approach into the most recent CHARMM c41 release. The current implementation is fully parallelized following the MPI implementation of the ENSEMBLE module and compatible with domain decomposition which provides additional computational speedup. However, most algorithmic aspects are those from the original implementation.[17]

Ala_{10} in Implicit and Explicit Solvent

Alanine decapeptide $(Ala)_{10}$ is a chain of 10 residues. In vacuum it is known to fold into a regular α-helix due to the stabilizing effects of the hydrogen bonds. However, in explicit solvent its structure is more debatable. In fact, NMR experiments together with MD simulations on shorter solvated $(Ala)_3$ to $(Ala)_7$ peptides suggest that they populate predominantly polyproline II conformations with some $\beta-$sheet structures but $\alpha-$helical conformations are not observed.[35] Hence, for simulations in solution a distribution of candidate structures is expected.

$(Ala)_{10}$ has also been used as a test system in the recent development of several optimised MD

sampling techniques such as unconstrained MD (explicit solvent),[19] adaptive steered molecular dynamics (in vacuum),[20] Multi-Replica and Multiple-Walker Adaptive Biasing Force (in vacuum),[21] or Simulated Tempering (explicit solvent),[22] and thus is a suitable benchmark system. The thermodynamic stability of the α-chain and folding or unfolding pathways were also investigated in vacuo[24] and in explicit TIP3P solvent[26,36], and free enthalpy differences between the α-chain and two less stable π- and 3_{10}- chains were also investigated (coarse-grained water model).[25] However a recent study[26] of $(Ala)_{10}$ in solution showed that folding/unfolding is much more complex than the previously reported "accordion-like scheme"[37] in explicit solvent, and indeed an extended set of non-helical and compact states is usually observed. This system was investigated with the new PINS CHARMM implementation, and results are compared with Molecular Dynamics (MD) and Parallel Tempering (PT) simulations running with (when possible) identical simulation parameters for direct comparison.

Simulations in Implicit solvent

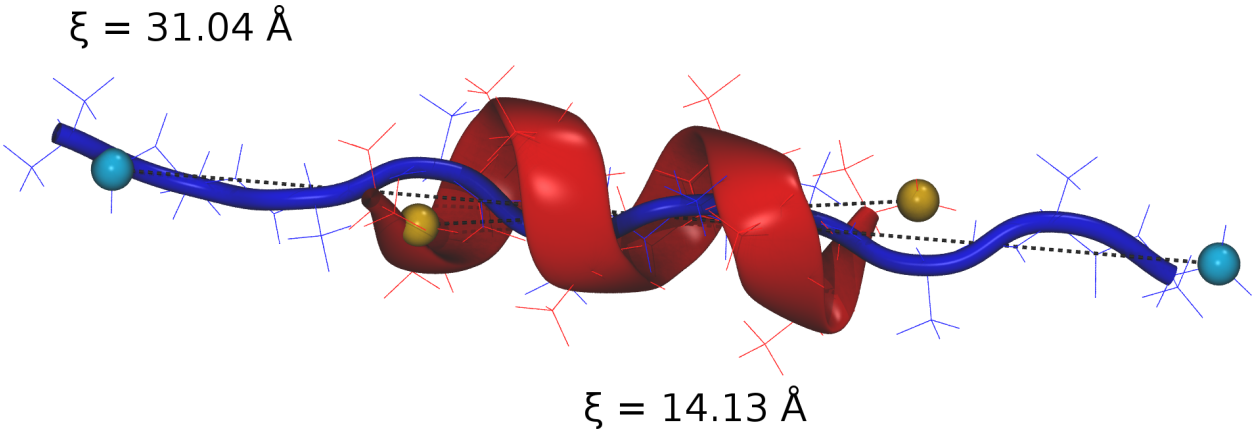

Figure 1: (Ala_{10}): extended starting structure (blue), and folded structure (red) obtained after 100 ns of MD with GENBORN implicit solvent. In cyan and orange, the carbonyl-carbon atoms define the end-to-end distance ξ in Å, used for following compactification and building ΔF surfaces. The extended structure has $\xi = 31.04$ Å, and the α-helical structure is characterised by a $\xi = 14.13$ Å.

All simulations were run with CHARMM c41 and the CHARMM Force Field version 36

which includes CMAP corrections.[38,39] Simulations in implicit solvent used the GENBORN model and started from an extended structure (Figure 1, in blue). A time step of 1 fs was used, and three independent simulations of 100 ns each were first performed with each method (MD, PT and PINS). For all structural comparisons the $\alpha-$helical structure (red in Figure 1) was the reference for computing the RMSD.

The temperature was 300 K for the MD simulation, and for the PT and PINS simulations an ensemble of 16 replicas at the following temperatures was used: 300.00, 329.52, 361.58, 396.42, 434.24, 475.31, 519.92, 568.35, 620.92, 677.99, 739.94, 807.32, 880.32, 959.70, 1045.71, 1139.03 K. These temperatures were chosen according to a temperature prediction algorithm[40] available as a free web-service, which generates a temperature set optimised for obtaining a desired exchange acceptance ratio, which was chosen to be 40% in the present case. The dual-chain PINS approach with two chains of 3 blocks $(6, 6, 4|4, 6, 6)$ was used. In the following MD, PT and PINS simulations are compared with regards to i) the simulation time required to reach a compact state ii) structural diversity of the ensemble generated iii) round trip time analysis (PT and PINS only) to visit the entire temperature manifold considered and iv) 1- and 2-dimensional free energy surfaces along meaningful progression coordinates.

Time to reach a compact state: First, the compactification of Ala_{10} for an individual trajectory using MD (Figure 2A), PT (Figure 2B) and PINS (Figure 2C) was followed by considering the RMSD(t). The red label in Figure 2 indicates the simulation time τ required to reach a compact state, defined as a structure with RMSD < 2 Å compared to the reference structure. This occurs by 11.99 ns for MD (Figure 2A), whereas for PT and PINS this threshold is reached within 0.40 ns and 0.12 ns (respectively Figures 2B and 2C). Thus, for this case PINS reaches a compact state two orders of magnitude more rapidly than MD, and approximately three times faster than PT.

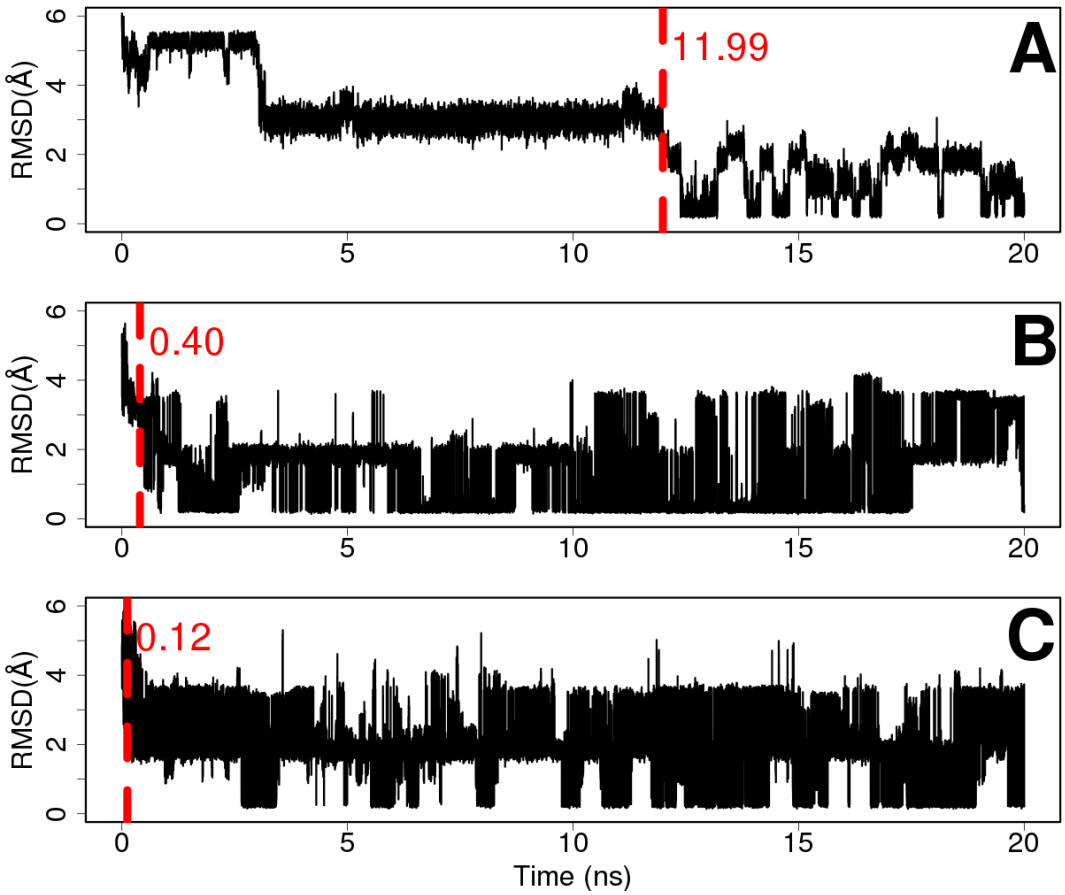

Figure 2: RMSD fluctuations of (Ala_{10}). 20 ns (from a total of 100 ns, see section SI1 for an illustration of the full 100 ns) of MD (A), PT (B) and PINS (C), with GENBORN solvent. The red vertical lines indicate the time at which RMSD$\leq$ 2 Å (partially helical compact state). The reference structure is the α-helix from Figure 1.

By repeating such simulations one hundred times for each of the three methods (for a maximum of 20 ns for MD, and 5 ns for PT and PINS), the distribution $P(\tau)$ of times required to sample a structure with RMSD $<$ 2 Å is obtained (see Figure 3). This confirms the results from a single run (Figure 2). For MD (red) the maximum of $P(\tau)$ is at around 12 ns with a broader distribution, whereas for PT (green) τ ranges from 0.5 to 2.0 ns, with a peak around 1.0 ns, and for PINS (blue) from 0.25 to 2.0 ns with a peak at 0.5 ns. Hence, PINS converges to the target structure more rapidly than PT, although the difference in this case is small.

Diversity of structures sampled: Secondly, it is of interest to assess the gain in terms of

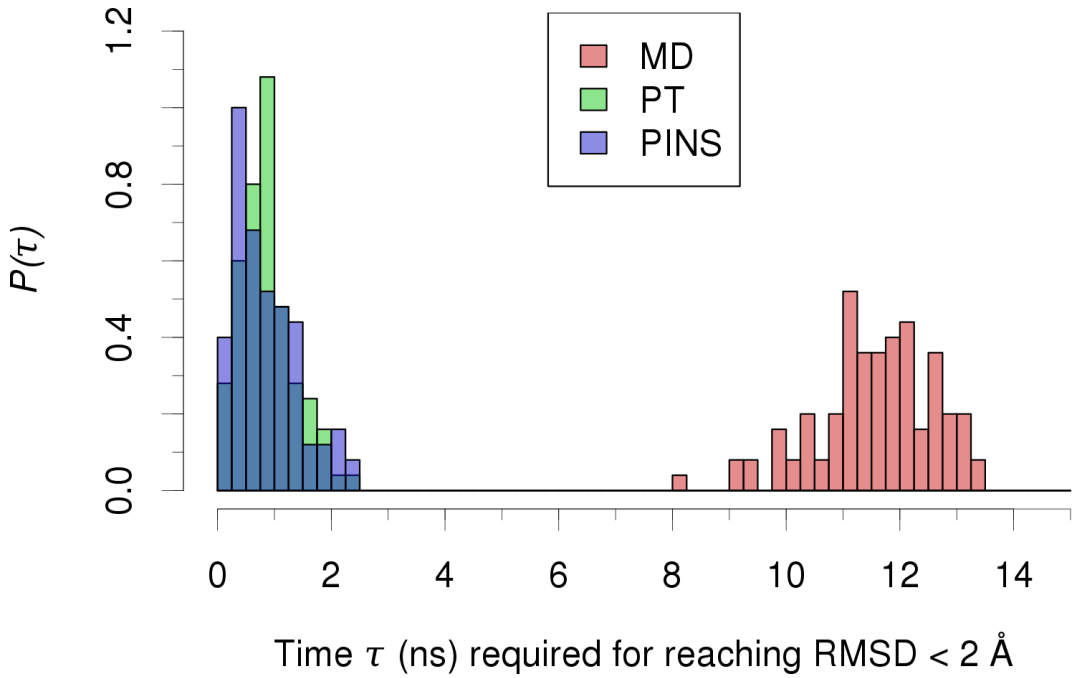

Figure 3: Histograms of the required time τ before reaching a RMSD of less than 2 Å, for 100 simulations of MD (red), PT (green), and PINS (blue). Simulations of 20 ns maximum for MD, and 5 ns for PT-PINS, all using the implicit GENBORN solvent model.

structural diversity of sampling provided by PINS compared to PT. A visual analysis of Figures 2B and 2C shows that the RMSD fluctuations from PINS (bottom) are usually larger compared to PT (middle), indicative of a more exhaustive sampling in RMSD space. In particular, taking RMSD $<$ 2 Å as the threshold, many more recrossings are found with PINS compared to PT or MD which leads to a more diverse set of structures which may be advantageous for generating diverse sets of structures.

For a more comprehensive analysis it is advantageous to cluster the sampled structures. The classification of recurring structural motifs can be based on different measures. In the following RMSD is used together with $k-$means.[41–43] Section SI2 provides details on the methodology followed for performing the clustering, and justifies the choice of $k = 6$ clusters, used for all the MD PT and PINS analysis.

Table 1 summarizes the results of clustering the data shown in Figure 2. The centers are

Table 1: $K-$means clustering with $k = 6$ centers applied to RMSD fluctuations from Figure 2. Clusters are sorted by increasing RMSD. PT (81 %) and PINS (71 %) both show an increased sampling of the low RMSD centers (RMSD < 2 **Å) compared to MD (65 %). See Section SI2 for a justification of $k = 6$.**

MD		PT		PINS	
centers (Å)	pop. (%)	centers (Å)	pop. (%)	centers (Å)	pop. (%)
0.4	15.8	0.3	38.7	0.3	22.8
1.1	9.2	1.2	3.7	1.2	4.2
1.8	39.7	1.9	40.7	1.9	45.3
2.2	19.4	2.8	2.8	2.2	10.5
3.0	12.8	3.5	8.9	3.0	3.6
5.1	3.0	4.2	5.2	3.6	13.6

sorted by increasing RMSD, and the columns contain the relative population of each center. For a representative clustering, 2×10^5 snapshots (taken from the first 20 ns shown in Figure 2) were analyzed. Performing the clustering over such a long time interval was necessary because MD simulations often remain trapped in metastable configurations for extended times (see Fig. 2 A between 5 and 10 ns), resulting in an overweighting of the corresponding cluster. Tests with fewer data showed that none or only a few transitions are observed, and with a larger amount of data the population of the low-RMSD cluster from MD simulations monotonically increases as mentioned above.

It is found that PT (39 %) and PINS (23 %) lead to a larger population of the lowest RMSD cluster around 0.3–0.4 Å compared to MD (16 %). It is also interesting to note that the cluster center with the largest RMSD is centered at 5.1, 4.2 and 3.6 Å for MD, PT, and PINS, respectively, which confirms that PT and PINS lead to an enrichment of compact configurations. Furthermore, PT and PINS lead to very similar cluster centers for the three most compact states whereas the next three cluster centers are more compact for PINS compared to PT, highlighting that PINS favors compact states. This suggests that PINS samples a larger number of stable and metastable structures than PT or MD simulations. This is supported by Figure 2 where the bottom panel (PINS) shows a larger amplitude in

the RMSD-fluctuations than the top and middle panels which correspond to MD and PT simulations, respectively.

Next, all snapshots (i.e. 60 ns of simulation) from MD, PT and PINS were clustered together which yields a different set of cluster centers. Then, the structures from each method (MD, PT, PINS) were projected onto the cluster centers which yields their population for each method (see Table S1 from Section SI2). Again, PT and PINS yield a larger population of the most compact state. Furthermore, the transition between compact (RMSD < 2.0 Å) and extended (RMSD > 3.0 Å) structures is more frequently sampled, as is also evident from Figure 2.

Round-Trip Times: Another useful performance measure for comparing PT and PINS simulations is the round-trip time t_r which reports on the number of moves in multi-temperature simulations to traverse the entire temperature ensemble.[15,44–46] Figures 4A and B show the occupation traces for PT and PINS, respectively, for one typical 100 ns long simulation using 16 temperatures. The replica considered is replica 1, initially starting at the lowest temperature (300 K). For PT the average round-trip time is approximately 10 ns and several full cycles over the entire 100 ns are observed for this replica. Contrary to that the PINS simulation show 100 round-trip events, i.e. an average t_r of 1 ns.

Next, a statistical analysis of 100 PT and PINS simulations, each 5 ns in length, was carried out. For PINS a total of 485 round-trips was found during the aggregate of 500 ns simulations which yields an average round trip time of $t_r = 1.1$ ns, representative of the single simulation discussed above. The distribution of round trip times (red) is shown in Figure 5A.

On the contrary, for PT only 14 round-trips were observed. Hence, for most cases 5 ns were not sufficient for one full round trip. For a better estimate of the round trip times for PT,

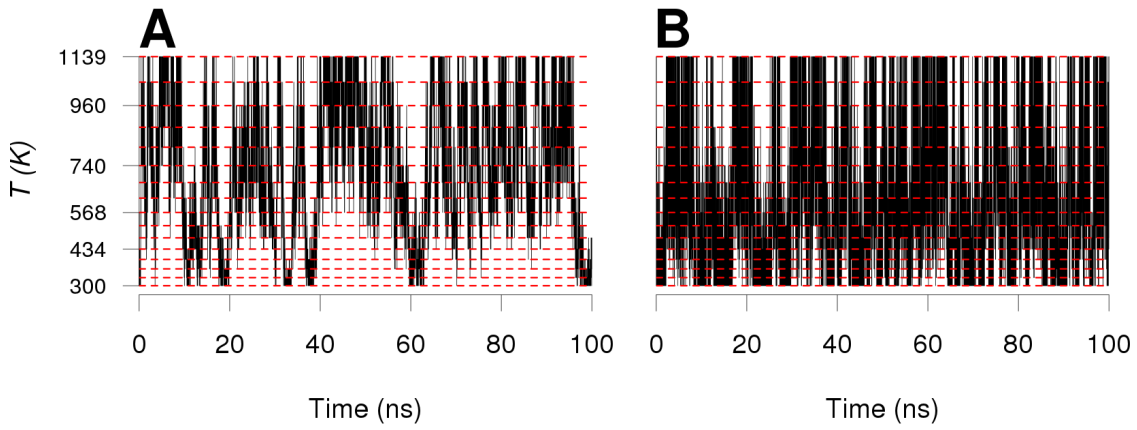

Figure 4: Traces for replica 1 ($T = 300$ K initially) for PT (A) and PINS (B). Red dashed lines correspond to the 16 simulation temperatures.

100 simulations, each 20 ns in length were carried out. This yields 221 events and an average $t_r = 8.7$ ns. The distribution $P(t_r)$ of round trip times is also shown in Figure 5A (black trace).

Table 2: Statistical analysis for the round-trip time t_r for replica 1, i.e. $T_1 \rightarrow T_{16} \rightarrow T_1$.

	PINS	PT
Total sim. time	100×5 ns	100×20 ns
Observations	485	221
Mean t_r (ns)	1.1	8.7
Std. dev. on t_r (ns)	0.8	4.0

A final performance measure considered here is the autocorrelation function of the occupation trace.[15] This quantity reports on how rapidly a given replica α is propagated away from the temperature trace T_N it started off from. Let N_m^α be the index N of T_N for replica α at simulation step m. The normalized autocorrelation function $C^\alpha(s)$ of replica α is

$$C^\alpha(s) = \frac{\langle (N_m^\alpha - \mu)(N_{m+s}^\alpha - \mu) \rangle}{\sigma^2}$$

where s is the lag time between observations, $\mu = \langle N^\alpha \rangle$ and σ^2 are the mean and variance of N^α, respectively, and $\langle ... \rangle$ denotes an expected value. Because the number of samples is large

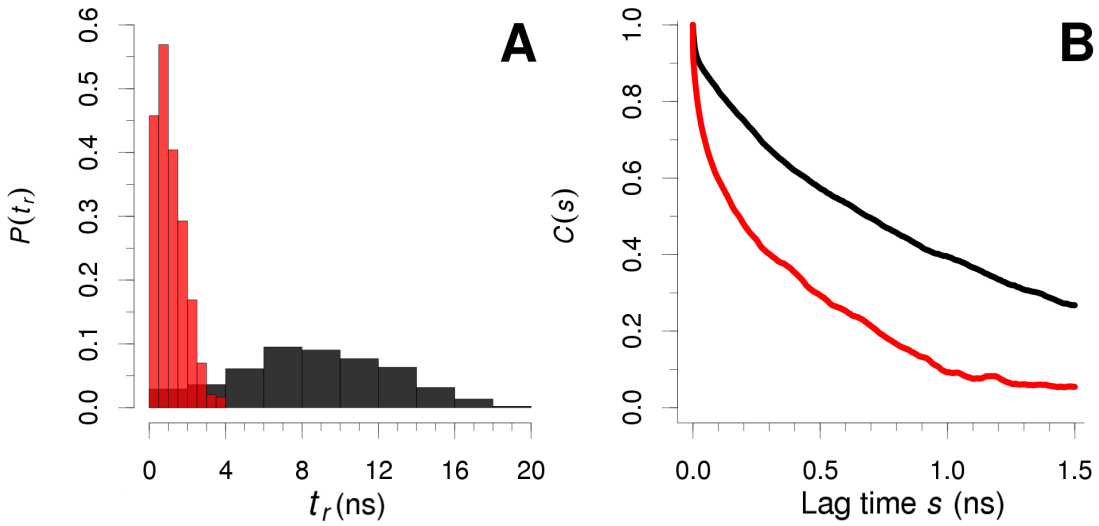

Figure 5: Panel A: Distribution of round-trip times (t_r) $T_1 \rightarrow T_{16} \rightarrow T_1$ for the 100×5 ns PINS simulations (red) and the 100×20 ns PT simulations (black). Panel B: Autocorrelation $C(s)$ for the Pins (red) and PT (black) simulations. Analyses are performed for the replica initially at $T = 300$ K.

(10^5 to 10^6), the computation of $C^\alpha(s)$ can become computationally demanding. In practice, it is possible to use the Wiener–Khintchine theorem[47,48], which relates autocorrelation and power spectrum, in order to determine $C^\alpha(s)$ over the desired time interval:

$$C^\alpha = \frac{1}{\sigma^2} \mathcal{F}^{-1} \left(|\mathcal{F}(N^\alpha - \mu)|^2 \right). \tag{9}$$

Here $\mathcal{F}$ is the Fast Fourier Transform (FFT) of the quantity in brackets. Figure 5B shows the autocorrelation function estimated for PT and PINS for replica 1, initially at 300 K. Equation 9 is used to estimate C^α over several ten ns, although only the first few ns are sufficient for PINS to reach values of $C(s) \approx 0$. PINS always reaches quasi-uncorrelated values for every replica at least twice as rapidly compared to PT. Typically, an autocorrelation value of $C(s) \leq 0.2$ is reached within 0.2 to 0.5 ns for PINS, whereas it usually takes 1.2 to 2.5 ns for PT.

One- and Two-Dimensional Free Energy Profiles: In a next step the end-to-end distance ξ between the carbonyl carbon atoms of the first and last residue was analysed (see Figure

1). This coordinate was already used previously for monitoring the progress of folding.[26,37] The α-helical structure was assigned to $\xi = 15.2$ Å or $\xi = 14.2$ Å, and extended structures were associated with $\xi = 33.0$ Å or $\xi = 32.0$ Å, respectively. The structures shown in Figure 1 correspond to $\xi = 14.1$ Å (red α-helix) and $\xi = 31.0$ Å (blue). Compact structures are usually defined as configurations with $\xi \leq 16.75$ Å.[26]

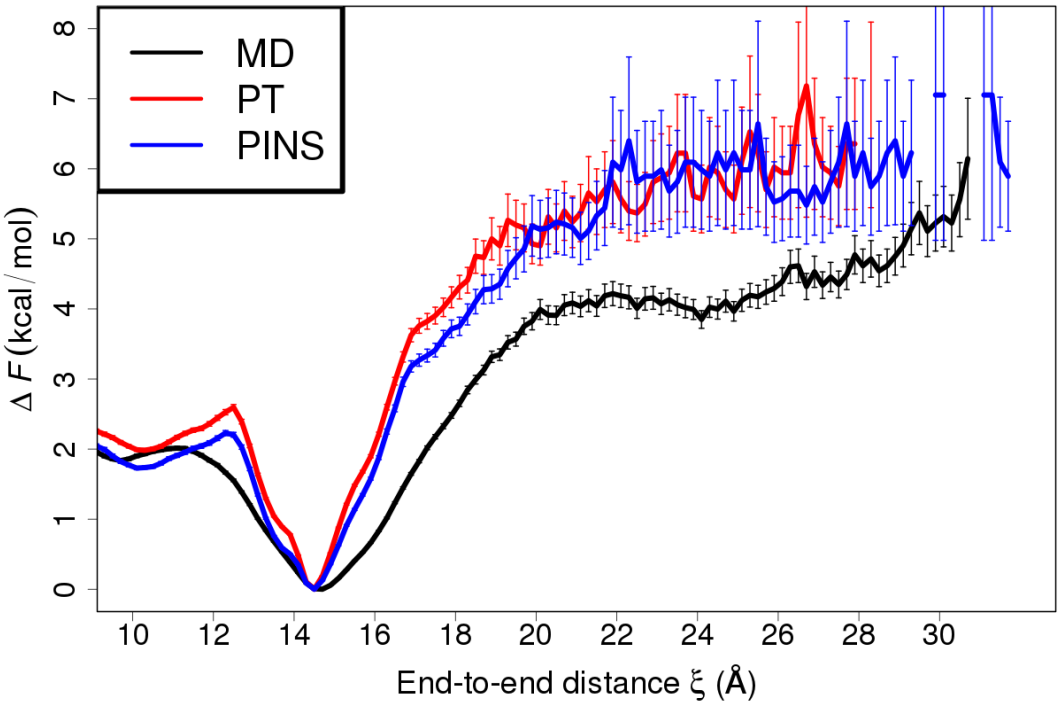

Figure 6: Free energy profile (ΔF in kcal/mol) built using the end-to-end distance ξ between carbonyls' carbon from first and last residue of Ala_{10} (see Figure 1) in implicit GENBORN model. Estimated for a total simulation time of 100 ns of MD (black), PT (red) and PINS (blue). The error bar is the statistical 95% confidence interval.

Figure 6 shows free energy profiles from MD (black), PT (red) and PINS (blue) simulations. They were generated from the 100 ns simulations by extracting and binning the end-to-end distance ξ from which the Helmholtz Free Energy was estimated according to $\Delta F(\xi) = -RT \ln(\rho(\xi))$, where $\rho(\xi)$ is the normalized density. The error bars correspond to the statistical 95% confidence interval. From the present simulations, minima were found at $\xi = 14.7$ Å (MD) and $\xi = 14.5$ Å (PT and PINS), respectively. The extended states

($16.0 \leq \xi \leq 28.0$ Å) are associated with free energies rising up to 6.0 kcal/mol above the minimum.

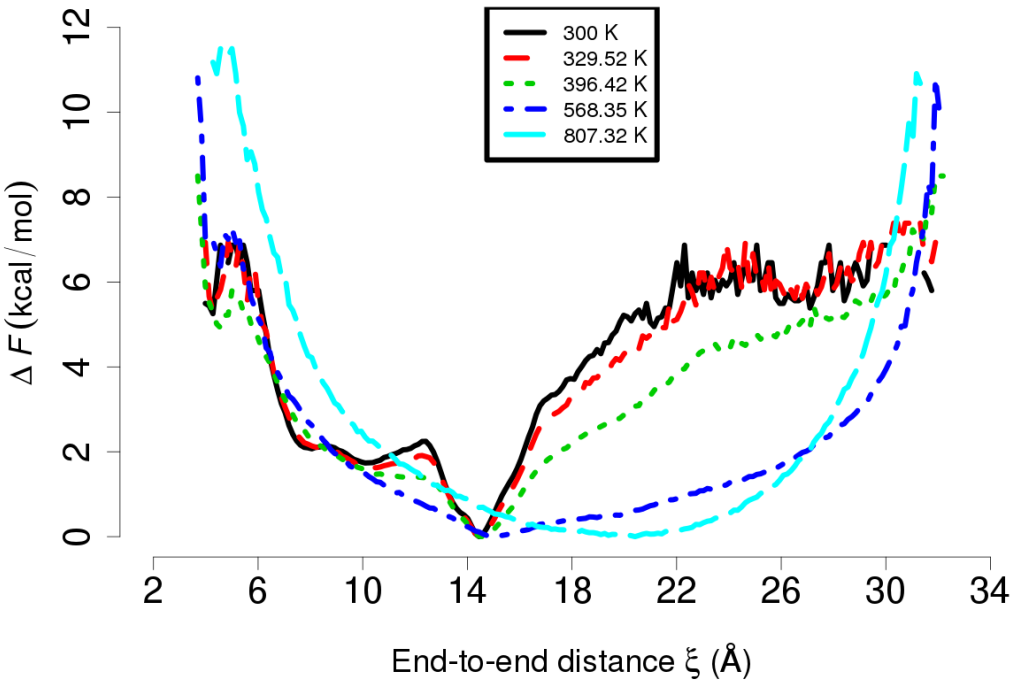

Figure 7: Free energy profile (ΔF in kcal/mol) built using the end-to-end distance between carbonyls' carbon from first and last residue of Ala_{10} (see Figure 1) in implicit GENBORN model. Estimated for 5 temperatures from the PINS simulation of 100 ns long.

PINS simulations can be used for investigating the stability of Ala_{10} in implicit solvent at higher temperatures. Figure 7 reports ΔF at five different temperatures. First, ΔF curves for 300 K and 329 K are fairly similar. This suggests that the decapeptide is stable at ambient temperatures. The α-helical structure, for $\xi = 14.5$ Å is still found to be the most stable state at 329 K. At $T = 396$ K (green curve on Fig 7), this is still the case, but it is observed that extended states ($\xi > 15$ Å) start being more sampled and thus more stable. At $T = 568$ K (blue curve of Fig 7), the funnel-like structure centered around the α-helix minimum disappears. The lowest value of ΔF is still found at $\xi = 15$ Å, but the free energy curve flattens considerably. All configurations characterised by $\xi \in [10; 27]$ Å are within 2 kcal/mol of the minimum, so frequent conformational changes will occur in this range of

end-to-end distances. When considering even higher temperatures (e.g. 807 K, cyan curve in Figure 7), the most stable configuration occurs around $\xi = 21$ Å, which is a non-helical, extended structure. Finally, it is also found that at higher temperatures the free energy curves flatten considerably.

2-dimensional FESs provide further insight into the relative stabilities of native and intermediate states. For this it is necessary to introduce meaningful progression coordinates describing the process of interest. They were chosen as the end-to-end distance ξ and the degree of α-helical content α.[26] The coordinate ξ describes the compactness of a structure and α quantifies the amount of $\alpha-$helical content (see Section SI4 for more details).

From simulations in implicit solvent a 2D histogram $P(\alpha, \xi)$ was built along ξ and α as progression coordinates, see Figure S4 from Section SI3. However, as the 2d FES is sampled less extensively in the transition regions a multi-variate Kernel Density Estimation (KDE)[49,50] was used for estimating the probability distribution matrix. KDE methods provide an accurate density estimation, combined with an intrinsic interpolation step, compensating the poor sampling of some of the bridging regions and higher energy areas (see Section SI3 for details). For MD the simulation time is 100 ns; for PT and PINS the simulations were 6 ns long and included 16 replicas (i.e. 96 ns of total simulation time). The MEP finding method (see Sections SI3 and SI5) was used to determine paths between important (local) minima.

The 2d FESs from MD, PT and PINS in implicit solvent are reported in Figures 8A, C and E, respectively. All simulations started in the fully extended state (characterised by $\alpha \approx 0.25; \xi \approx 28$ Å), and all find the global minimum to be an α-helix (point 1 in Figures 8A, C and E). For all cases a broad basin with a variable number of stable additional α-helical structures (1 to 3 for MD; 1 and 2 for PT; 1 to 5 for PINS) is found. β-hairpin structures (5 and 6 for MD; 3 and 4 for PT; 6 for PINS) appear at $\xi = 5$ Å and $\xi = 7$ Å for

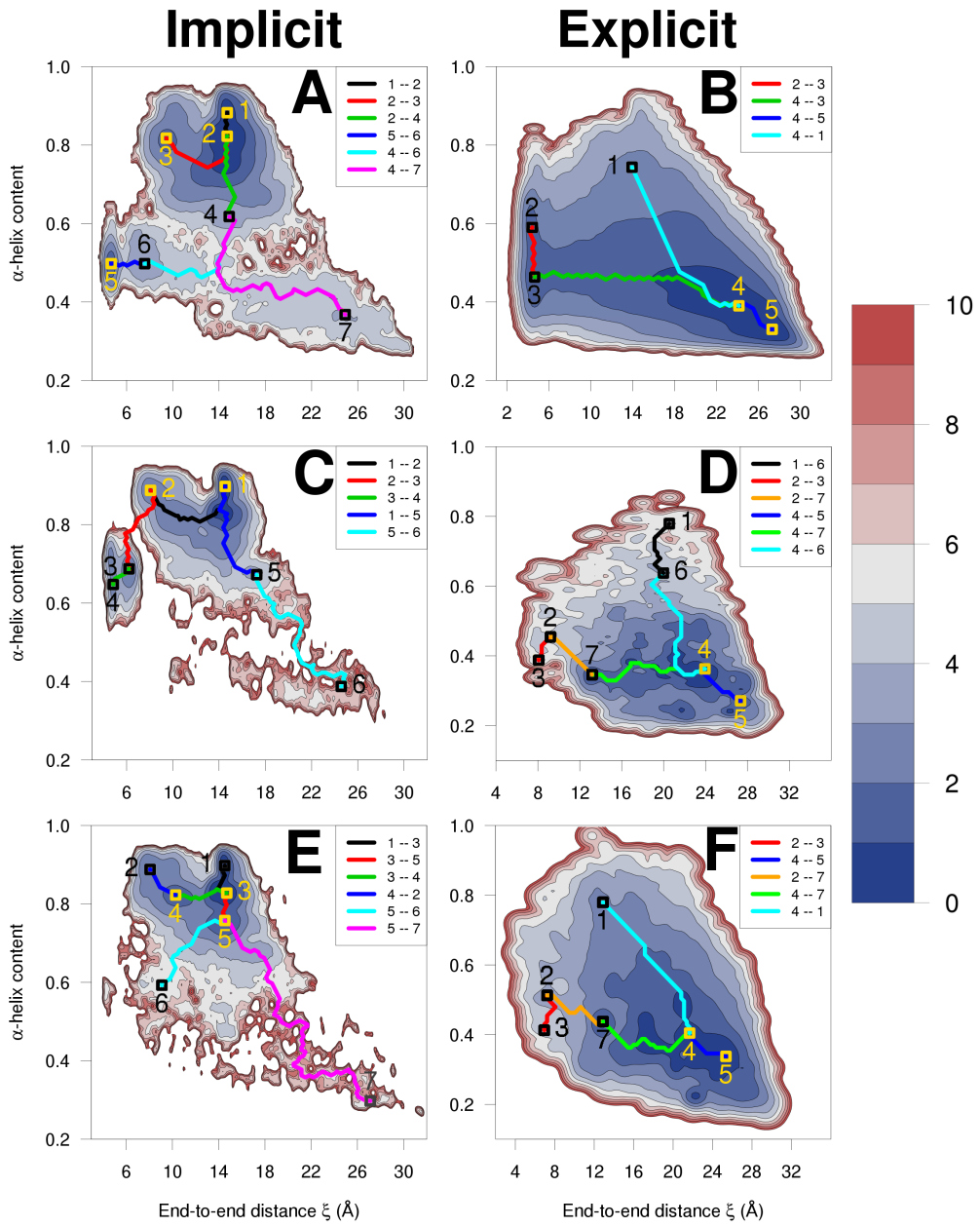

Figure 8: Ala_{10} FES (ΔF in kcal/mol) from simulations at 300 K in implicit (left column) and in explicit (TIP3P) water (right column) from the replica initially at 300 K. Panels A and B for MD (100 ns for implicit, and 4 μs for explicit), C and D for PT (16 replicas of 6 ns for implicit, 32 replicas of 75 ns for explicit), E and F for PINS (16 replicas of 6 ns for implicit, 32 replicas of 75 ns for explicit). The 2D FESs were built using Kernel Density Estimation (KDE). Compared with standard 2D histograms (see Figure S4 Section SI3), KDE yields a smoother surface with more connected areas. In panels B, D, and F state 1 corresponds to an $\alpha-$helix, states 2 and 3 to a hairpin and states 4 and 5 to an extended polyproline (PP-II) conformation. The effect of post-processing is illustrated in Section SI6

MD simulations but only at one of these values for PT and PINS, respectively. Furthermore, the connectivity of the α-helical and β-hairpin basins for MD and PINS is sampled, whereas for PT these two regions appear to be poorly connected: the path connecting states "2" and "3" (in Fig. 8C) was only found by allowing the path finding algorithm to extrapolate a possible path even if the saddle point was poorly sampled. Despite the fact that less than 10 % of the data was analyzed for PINS (6 ns) compared to MD (100 ns), PINS provides a representative picture of the 2D-FES which is less obvious for PT.

Previously, the folding of $(Ala)_{10}$ in implicit solvent (Generalized Born Implicit Solvent) was investigated using the CHARMM22 force field.[51] It was found that compact structures are favoured and the free energy curve varies over 5 kcal/mol between helical and fully extended structures. This is comparable to the present work although no clearly preferred minimum energy structure was found there contrary to the present simulations (see Figure 6). These differences will be discussed further below. Even earlier work has focused on differences between possible helical structures of $(Ala)_{10}$ without, however, considering the entire folding landscape including unfolded and extended states.[52] These restrained simulations found that the $\alpha-$ and $\pi-$helical structures are only separated by ≈ 1 kcal/mol whereas the $3_{10}-$helix is destabilized by more than 10 kcal/mol relative to the two other structures. Such high-energy regions (around 10 kcal/mol above the minimum) exhibiting helical structures in implicit solvent are also found in the present work where the $3_{10}-$helix is at $\xi \approx 17$ Å and $\alpha \approx 0.6$.

Ala_{10} in Explicit Solvent

Next, the performance of PINS was assessed for studying the compactification of deca-alanine in explicit solvent using the TIP3P water model.[53] The same initial unfolded configuration, Figure 1 (blue), was used. It was solvated in a cubic box of size 40.5 Å, heated and equilibrated to a temperature of 300 K for MD (or to the target temperature for PT and PINS) for

100 ps. The time step was always 1 fs and SHAKE[54] was used for bonds involving H-atoms. For MD, 100 independent simulations were performed, each 40 ns in length. Thus the total simulation time is 4 μs (compared to 2.5 μs from Ref.[26]). The PT and PINS simulations used 32 temperatures, between 300.00 and 380.87 K and 50 independent simulations, 1.5 ns each were carried out which yields a total aggregated simulation time of 1.6 μs. The PINS dual-chain structure used 6 temperature blocks $(6,6,6,6,5,3|3,5,6,6,6,6)$. The Particle Mesh Ewald method[55,56] was used, combined with domain decomposition (DOMDEC)[57] for MD simulations. The non-bonded energy cutoff-cuton were set to, respectively, 9 Å and 7.5 Å, and the non-bonded lists were built using a heuristic algorithm with a buffer of 11 Å. These settings follow the official documentation of CHARMM with DOMDEC.

ξ-based ΔF profile for MD Figure 9 shows the 1-dimensional free energy curve obtained from binning the end-to-end distance ξ using 4 μs of MD simulations in explicit solvent. The 4 configurations shown are examples of typically sampled conformations ($5.0 \leq \xi \leq 25.0$) during the simulations. Their free energy is within $\Delta F \leq 1$ kcal/mol of the global minimum, illustrating the large number of metastable configurations possible for solvated Ala_{10}.

2D ΔF surfaces for MD PT and PINS in explicit water: Two-dimensional FESs $F(\alpha, \xi)$ for simulations in explicit water are reported in Figure 8B which uses data from 4 μs of MD simulations. Figures 8D and F show FESs from PT and PINS simulations based on 75 ns of data which is approximately 50 times less than for the MD simulations. This is because only the data of the lowest-temperature replica was analyzed for PT and PINS. Overall, 1.6 μs (32 replicas of 75 ns each) of data is available, though.

MD, PT and PINS simulations in explicit solvent find the lowest energy minima (points 4 and 5 in Figure 8) for extended conformations, indicated by values of $\alpha \in [0.2; 0.5]$ and $\xi \in [22; 28]$. Such configurations correspond to those shown in Figure 9 and are expected due

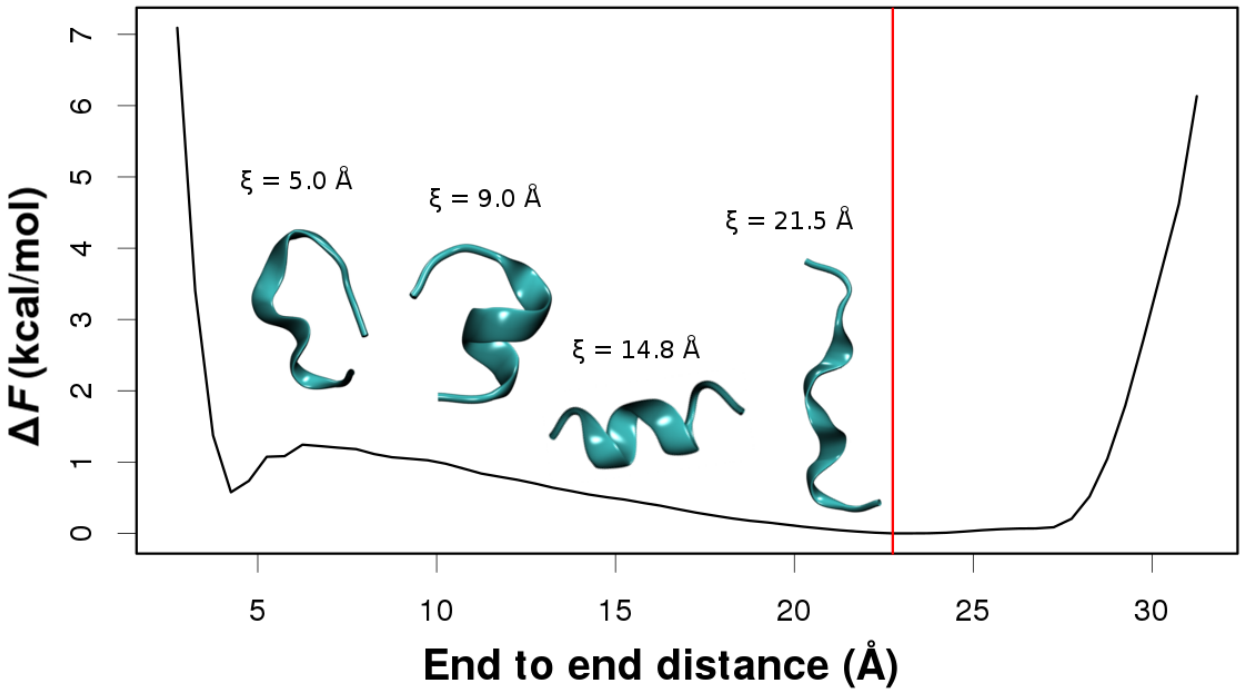

Figure 9: Free Energy as a function of ξ for Ala_{10} in explicit TIP3P water from 4 μs MD simulation. The red vertical line marks the point where $\Delta F = 0$ kcal/mol for $\xi = 22.75$ Å, i.e. the most sampled extended (non-helical) state. The 4 displayed configurations are examples of structures for which $\Delta F \leq 1$ kcal/mol.

to the stabilising interactions between the solvent (water) molecules and the polypeptide. A second set of stable conformations are $\beta-$hairpin structures (points 2 and 3), characterized by $\alpha \in [0.4; 0.6]$ and $\xi \in [4; 8]$. With PT and PINS this area is sampled but find the states at 3 kcal/mol (for PINS) and 5 kcal/mol (for PT) above the lowest-energy basin, compared to 1 kcal/mol from MD simulations. The $\alpha-$helix is found by MD and PINS (point 1) for ($\xi \approx 12$ Å, $\alpha \approx 0.8$). Contrary to that, PT does not find such compact, helical structures but rather extended and poorly sampled structures (1 and 6 in Figure 8D) at $\xi \approx 20$. Finally for PT and PINS an extra intermediate (state 7), "bridging" the extended (states 4 and 5) and β-hairpin (2 and 3) sets, was located around ($\xi \approx 12$ Å, $\alpha \approx 0.4$). This state, which is not found in MD simulations and located along the path connecting states 3 and 4 in the MD simulations is a possible, but marginally stabilized (1.6 kcal/mol) intermediate.

Comparison of the 2-dimensional landscapes from simulations in implicit (Figures 8A, C,

E) and explicit solvent (Figures 8B, D, F) highlight profound differences. Simulations in implicit solvent find compact helical structures ($\xi \approx 15$Å) preferred over extended structures ($\xi \approx 28$Å). This finding is i) consistent with recent simulations using NAMD with either the CHARMM22/CMAP or CHARMM36 (which includes CMAP corrections by default) force fields together with a range of biasing techniques (adaptive biasing force (ABF), umbrella sampling (US), replica exchange MD (REMD))[26] but ii) at variance with other, even more recent, simulations using NAMD together with the CHARMM22 force field and various flavors of steered MD simulations.[51] Contrary to the simulations using CHARMM22,[51] which find that in explicit, implicit and in vacuo simulations prefer more compact structures ($\xi \leq 15$ Å), simulations using CHARMM22/CMAP and CHARMM36 suggest that both, compact and extended structures are metastable states on the free energy surface.[51] This is consistent with the finding that the CHARMM force field without CMAP corrections yields an overpopulation of $\alpha-$helical structures which was also found for other force fields.[58] The 2-dimensional FES from all simulations in solution find that extended structures are favoured in solution which supports findings from NMR experiments on short Ala-peptides (up to $(Ala)_7$),[35] and previous studies[59,60] that described those extended structures as Polyproline II (PP2) helices.

It is of interest to briefly compare the present results with those from ABF and replica exchange MD-umbrella sampling using CHARMM36 and explicit solvent.[26] The simulations were carried out over a range $12.5 \leq \xi \leq 31.5$ Å which by construction excludes $\beta-$hairpin structures. For the 1-dimensional free energy curve along ξ (see Figure 9 compared to Figures 3 and 8 from Ref.[26]) it is found that they all have more or less pronounced minima for helical and extended conformations, separated by a flat and extended plateau. The free energy is typically within 1 to 2 kcal/mol of the global minimum. One-dimensional free energy curves from biased (ABF or umbrella sampling) simulations in Ref.[26] without additional constraints do not exhibit clearly preferred structures (e.g. $\alpha-$helical or extended). However, when integrating the 2-dimensional free energy surface along the $\alpha-$helical content (Figure

8 in Ref.[26]) two clearly distinguishable basins appear: one in an $\alpha-$helical conformation and the other in an extended state, separated by a 1 kcal/mol barrier. Such a free energy curve along ξ is supported by the extensive 4 μs simulations from the present work, see Figure 9.

The present results (Figure 8) can also be compared with Figure 6 (bottom) from Ref.[26]. Path ($1-4-5$ - between $\alpha-$helical and extended) in the present work (Figure 8B) is also found in Figure 6 (bottom).[26] However, the least free-energy path from replica exchange MD-umbrella sampling simulations finds a barrier of 3.5 kcal/mol between the $\alpha-$helical and the extended structure whereas the present work identifies this as a downhill process (Figures 8B, D, and F), i.e. $\alpha-$helical structures are unstable in explicit solvent. In addition, there are further low-energy states found in the present work (states 2, 3 (MD, PT, PINS) and state 7 (PT and PINS)).

To summarise, PINS (Figure 8F) provides results similar to what is obtained from extensive MD sampling (Figure 8B), but with a total simulation time of 4 μs for MD, compared to an aggregated 1.6 μs from PINS of which only 75 ns were included in the analysis. The results support experimental results on shorter $(\mathrm{Ala})_7$ and earlier simulations in implicit and explicit solvation which yield flat free energy surfaces along the minimum energy path, exhibiting stabilized $\alpha-$helical and extended structures. However, the energetic ordering of these two states depends on the description of the solvent. In explicit water, extended structures are preferred over compact, $\alpha-$helical structures and the presence of solvent makes alternative helical structures such as $\pi-$ or $3_{10}-$helices much more favourable. This was also found in previous work on rat and human Amylin.[61] Finally, it is demonstrated that the CMAP correction is required for meaningful description of the conformational free energy landscape for $(\mathrm{Ala})_{10}$.

Xe Migration in Myoglobin

As a third application of PINS with reweighting the free energy surface of ligand diffusion in globular proteins is considered. Because the physiological role of such pockets has as yet not been clarified beyond doubt their characterization is essential. Ligand migration takes place on extended time scales and usually involves appreciable barriers separating the various metastable states. As a recent example, Xenon migration in Cytochrome ba_3 oxidase has been found to involve rate coefficients for exchange between neighboring sites on the order of 1 s^{-1}.[62]

Myoglobin (see Figure 11, left) is one of the best characterized proteins, both experimentally and by using various types of simulation techniques, and serves as a model system for studying ligand binding, unbinding, and migration.[63] While the pockets accessible to guest atoms (Xenon) and small molecules (O_2, NO, CO) are well characterized from experiment[27–30] and theory/computer simulation[31–34], the stabilization energies in these pockets, the pathways between them and the energy barriers separating them are more debatable. Experimentally, CO-migration was followed using Laue diffraction and the integrated electron density of the CO-associated features were found to change over 6 orders of magnitude in time between 10^{-9} and 10^{-3} s with signal decay only starting after 10^{-5} s.[64]

A full characterization of ligand migration requires direct sampling of the entire free energy surface. A considerable step towards this goal has been the analysis of several trajectories of 90 ns (with 8 CO molecules each) to identify ligand entry pathways from the solvent. Despite such a serious effort no free energy profiles were presented because most transitions between pockets are still rare events and occur only once per trajectory.[32] Since such extensive sampling is computationally expensive, application of enhanced sampling methods is of great interest.

The use of Xenon as a guest molecule is motivated by the fact that it diffracts well (54 electrons) in X-ray studies. Furthermore, Xenon only interacts via Van der Waals interactions with its environment which - in addition to its large mass - further slows down diffusion inside and between the cavities which makes the use of rare event sampling techniques mandatory. Hence, Xe diffusion in Mb is a typical example of a topical and high-dimensional system for which PINS offers potential advantages over other sampling techniques.

Computational setup: The CHARMM-GUI[65,66] interface was used for generating an initial structure, based on the Protein Data Bank entry 4NXA[67]. The CHARMM c36 force field[39] was used together with CHARMM version c41. A cubic box of size 67 $Å^3$, containing 8596 TIP3P[53] water molecules was used for solvating the system. The non-bonded parameters for the Xe atom were $\varepsilon = -0.423$ kcal/mol and $R_{\mathrm{min,Xe}}/2 = 2.05$ Å, which are comparable to those used in previous work ($\varepsilon = -0.494$ kcal/mol and $R_{\mathrm{min,Xe}}/2 = 2.24$ Å).[68]. The Particle Mesh Ewald[55,56] algorithm is used for treating the non-bonded interactions (cuton–cutoff of respectively 10–12 Å), bonds involving hydrogen were constrained using SHAKE[54], and a time step of 2 fs was used. The system was heated and equilibrated for 100 ps for MD at a temperature of 300 K. The same heating–equilibration protocol is followed for PINS, which uses a dual chain approach of 32 replicas $(6, 6, 6, 6, 5, 3|3, 5, 6, 6, 6, 6)$, and to each replica a temperature between 300.00 and 393.95 K is assigned. Simulations were started using a set of configurations in which one Xe atom is initially assigned to one of the 4 experimentally known pockets (Xe1 to Xe4, see Figure 11 right part). For each of the 4 systems MD simulations 100 ns long were carried out. For PINS each replica was simulated for 3.0 ns resulting in a total aggregated simulation time of 96 ns, in order to compare similar amount of data from MD and PINS.

Results: For both, MD and PINS, trajectories were aligned relative to the crystal structure. In order to ascertain the long-time stability of the system, the RMSD relative to the crys-

tal structure was determined and was found to be below 2 Å throughout, confirming the observed stability and structural integrity of the protein. For the analysis, the distance between the Xe atom and the center of each pocket (Xe1 to Xe4 as found in the X-ray reference structure[27]) was determined for all configurations and all trajectories which provides a first clustering. Then the Xe atom in a particular snapshot was assigned to the pocket for which the distance between the current location and the center of each pocket is lowest. This yields a discretized trajectory. From this it is possible to estimate the relative occupation (q_{PINS} or q_{MD}) of each pocket Xei along the trajectory of interest. In order to compare the relative efficiency of sampling, a boost factor R of PINS over MD is defined as $R = \frac{q_{\mathrm{PINS}}}{q_{\mathrm{MD}}}$. Figure 10 shows R for the Xe atom in any of the 4 different starting positions. As sampling of the protein interior is of concern here, events in which the Xe atom remains in the initial pocket are not considered. Furthermore, situations in which Xe escapes to the solvent are also discarded. Red bars with a "∞ symbol" correspond to transitions which are not sampled at all using conventional MD (i.e. $q_{\mathrm{MD}} = 0$), and for which PINS finds transitions. The results show that PINS increases the sampling efficiency by a factor of 2 to 10. Furthermore, for 3 of the 4 simulations one transition which was not sampled using MD is sampled with PINS (pockets Xe4, Xe2, Xe3 when starting from Xe2, Xe3, Xe4, respectively). Hence, overall PINS samples transitions more effectively.

As R not only reports on the number of transitions but also on the actual occupation of particular pockets, the transition count matrices were also determined (Table 3). These matrices were built using the full data from the MD and PINS simulations with snapshots taken every 1 ps. With PINS more frequent transitions from or to pocket Xe3 are found, which is rarely and poorly sampled with MD. Another interesting observation is that PINS simulations allow direct Xe1$\leftrightarrow$Xe4 transitions. Analysis of the 300 K to 350 K replicas supports that this only occurs for replicas run at higher temperatures. Finally, it is also noticed that the transition matrices are near-symmetric.

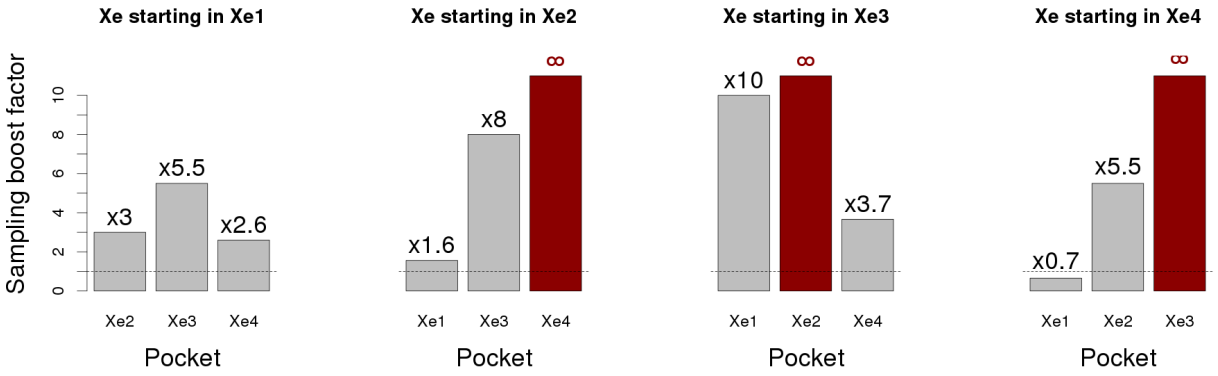

Figure 10: Ratio between the relative occupation of each pocket ($y-$axis) defined as $R = \frac{q_{\mathrm{PINS}}}{q_{\mathrm{MD}}}$. Values of $R > 1$ mean that PINS is more efficient than MD. $R = \infty$ denotes situations where the corresponding pocket was not sampled by MD simulations starting from a given initial pocket.

Table 3: Transition matrix estimated for MD (left) and PINS (right), using 4 simulations of respectively 100 and 96 ns, each starting in one of the four pockets. The transition boost provided by PINS is evident, and the effect of high temperature replicas allows for example direct jumps Xe1↔Xe4, unobserved with MD.

	MD				PINS			
	Xe1	Xe2	Xe3	Xe4	Xe1	Xe2	Xe3	Xe4
Xe1	·	66	8	0	·	240	30	610
Xe2	68	·	38	66	234	·	40	112
Xe3	12	40	·	0	34	40	·	2
Xe4	0	54	0	·	620	94	4	·

The capability of PINS for sampling low probability (transition) regions connecting the pockets can be directly visualised. For that, coordinates of the Xe atom are extracted, and the normalized probability distribution $\rho(x, y, z)$ at a given point (x, y, z) is evaluated on a 3D grid with resolution 0.5 Å, see Figure 11 for simulations with Xe initially in Xe4. The densities shown are for $\rho(x, y, z) = 10^{-5}$. This analysis confirms the results from Figure 10 and Table 3, i.e. that the Xe3 pocket is poorly sampled by MD, whereas PINS explores this region of the protein. It is also demonstrated that PINS samples the transition region more extensively, e.g. the Xe4↔Xe2 and Xe1↔Xe2 transitions (see upper- and bottom-right parts

of the isosurfaces).

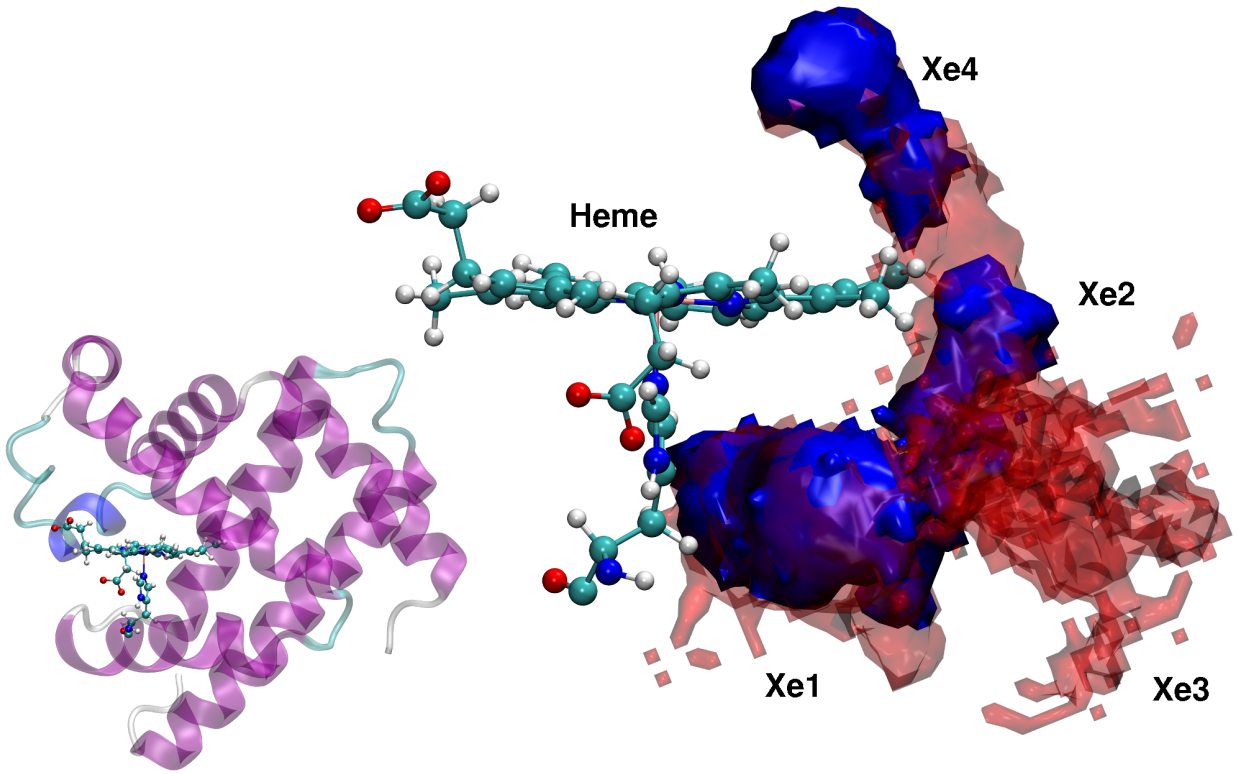

Figure 11: Left: Myoglobin with its heme functional group. Color code for the protein secondary structure is purple and blue for α and 3-10 helices, cyan and white for turn and coil, respectively. Right: Isosurface of normalised probabilities ($\rho = 10^{-5}$) to find the Xe atom at a given grid point, and definition of the 4 Xe pockets. Blue surface for MD, red for PINS. Built using the 100 ns and 96 ns long simulations. For simulations starting in pocket Xe4. PINS samples pocket Xe3 not explored with conventional MD. The transition channels Xe4 $\leftrightarrow$Xe2 and Xe1$\leftrightarrow$Xe2 are also more widely sampled when using PINS than with standard MD.

From the probability distribution functions $P(x, y, z)$ the relative stabilization energies of Xe in the 4 pockets can be determined and are summarized in Table 4. The MD results are obtained from inverting $P(x, y, z) \propto \exp(-\beta \Delta F_{\text{stab}}(x, y, z))$ and for PINS the post-processing procedure, as described in the Methods section, was applied (see Equation 8). The PINS results compare quite favourably with those from an earlier study[68] (values also reported in Table 4 for comparison) based on a 5 ns simulation, with relative absolute differences ranging between 0.1 and 0.6 kcal/mol. Binding free energies from the present MD simulations are

Table 4: Stabilization free energy ΔF_{stab} (kcal/mol) for the 4 Xe pockets, estimated for the MD and PINS simulations, and compared with the Implicit Ligand Sampling results from Ref.[68]. The 95 % confidence interval was estimated using bootstrapping, dividing data in 10 sets.

	ΔF_{stab} (kcal/mol)			
	MD	PINS	Cohen et al.[68]	Exp.[68]
Xe1	-4.4 ± 0.1	-6.2 ± 0.1	-6.4	-5.1
Xe2	-2.6 ± 0.4	-4.6 ± 0.3	-5.2	-4.5
Xe3	-3.7 ± 0.3	-5.6 ± 0.1	-5.1	-4.6
Xe4	-4.3 ± 0.3	-5.6 ± 0.2	-5.5	-4.4

somewhat too low which may be related to under-sampling in the MD simulations although the aggregate simulation time is 400 ns (i.e. 100 ns per initial Xe placement). It should be recalled that implicit ligand sampling[68] carries out simulations without the guest molecule present (i.e. the empty protein) and coupling between protein and ligand dynamics is absent. Also, there is little guarantee that large energy barriers will be sampled accurately which leads to overestimated energy barriers. Given the considerably larger amount of data from the present simulations (aggregate of 400 ns for PINS) compared to the previous study[68] (5 ns of MD with Implicit Ligand Sampling), it is expected that the present stabilization energies ΔF_{stab} are more representative.

Finally by extracting the free energy along the path connecting two pockets, it is also possible to estimate the transition barrier free energies from the PINS simulations. For the Xe1↔Xe2 transition the barrier height is estimated to be 4.4 kcal/mol and for the Xe2↔Xe4 transition it is 3.9 kcal/mol, corresponding to typical transition times on the sub-nanosecond time scale according to transition state theory. This is also reflected in the MD transition count matrix from Table 3 (left), where 120 transitions are found in total for the Xe1↔Xe2 transition and the number of transitions for the Xe2↔Xe4 is slightly larger (134), indicative of a lower free energy barrier. These results were confirmed for the Xe2↔Xe4 transition by using umbrella sampling simulations[12]. The progression coordinate for this transition was the distance between the center of gravity of the Phe138 carbon atoms and the Xe-atom. This coordinate

was found to be useful in previous simulations of transition paths for CO between these two pockets for which a barrier height of 6.0 kcal/mol or larger was found depending on the initial protein structure.[69] For Xe, which is expected to interact less strongly with the protein environment, a barrier height of 4.5 ± 0.4 kcal/mol was obtained using WHAM[70]. This agrees favourably with the estimate of 3.9 kcal/mol from PINS simulations, which further validates the implementation and analysis protocol.

Summary

The present work describes the implementation, analysis and application of PINS to two systems of different complexity: finding compact structures of deca-alanine in implicit and explicit solvent and Xenon migration in Myoglobin. For deca-alanine, compactification to the expected $\alpha-$helical structure in implicit solvent was found to occur more rapidly by one order of magnitude with PINS and PT compared to MD simulations. For $(\mathrm{Ala})_{10}$ in solution conflicting earlier results[26,51] were resolved and are most likely related to the need to use the CMAP correction with CHARMM22 or CHARMM36 when studying populations of different peptide conformational states. The present work (see Figures 8B, D, F) confirms experimental results from NMR spectroscopy on solvated alanine-repeats up to $(\mathrm{Ala})_7$ which found predominantly extended PP-II structures, some hairpin structures but no $\alpha-$helical conformations.[35]

The third application considered Xenon atom migration in the internal cavities of Mb. PINS extensively samples the 4 experimentally known Xe pockets and the transition regions between them. This contrasts with MD simulations which provide little information about barrier crossings for comparable simulation times. PINS yields estimates for Xe-binding free energy comparable to alternative methods such as implicit ligand sampling.[68] The height of

the Xe4↔Xe2 transition barrier was estimated to be ≈ 3.9 kcal/mol from the PINS-unbiasing procedure and was confirmed using umbrella sampling simulations.

Finally, it is pointed out that PINS could be further generalised. In Equations 4 and 5 the probability density $\rho(\mathbf{X})$ and the partition function Z were defined by only considering the potential energy $V(\mathbf{X})$. Instead of $V = E_{\text{pot}}$ it is possible to use a classical Hamiltonian $\mathcal{H} = E_{\text{kin}} + E_{\text{pot}}$ where E_{kin} and E_{pot} are the kinetic and potential energy, respectively. Then it is possible to define K Hamiltonians instead of K temperatures for the replicas, each Hamiltonian thus containing e.g. different biasing potentials. The two Equations 4 - 5 are still valid, so from the algorithmic point of view the only necessary modification is to broadcast the total Hamiltonian between the replicas instead of the potential energy.

Acknowledgement

This work was supported by the Swiss National Science Foundation through grants 200021-7117810, the NCCR MUST (to MM). JDD wishes to acknowledge support from the National Science Foundation award DMS-1317199, from the DARPA EQUiPS award W911NF-15-2-0122, and for continuing discussions with P. Dupuis.

The authors declare no competing financial interest.

Supporting Information Available

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.xxxxxxx :

- Details, figures and tables concerning the RMSD analysis and the clustering method-

ology.

- Details concerning the "kernel density estimation" and "minimum energy path" finding methods.
- Details concerning the α-helical content, used as a reduced coordinate for producing the 2-dim free energy surfaces.
- Supplementary figures associated to Fig. 8.

References

(1) Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341.

(2) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(3) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109.

(4) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(5) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(6) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *PCCP* **2005**, *7*, 3910–3916.

(7) Kofke, D. A. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **2002**, *117*, 6911–6914.

(8) Xia, J.; Flynn, W. F.; Gallicchio, E.; Zhang, B. W.; He, P.; Tan, Z.; Levy, R. M. Large-scale asynchronous and distributed multidimensional replica exchange molecular simulations and efficiency analysis. *J. Comp. Chem.* **2015**, *36*, 1772–1785.

(9) Doll, J. D.; Gubernatis, J. E.; Plattner, N.; Meuwly, M.; Dupuis, P.; Wang, H. A spatial averaging approach to rare-event sampling. *J. Chem. Phys.* **2009**, *131*, 104107.

(10) Plattner, N.; Doll, J. D.; Meuwly, M. Spatial averaging for small molecule diffusion in condensed phase environments. *J. Chem. Phys.* **2010**, *133*, 044506.

(11) Hédin, F.; Plattner, N.; Doll, J. D.; Meuwly, M. Spatial Averaging: Sampling Enhancement for Exploring Configurational Space of Atomic Clusters and Biomolecules. *J. Chem. Theory Comput.* **2014**, *10*, 4284–4296.

(12) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Chem. Phys.* **1977**, *23*, 187–199.

(13) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.

(14) Plattner, N.; Doll, J. D.; Dupuis, P.; Wang, H.; Liu, Y.; Gubernatis, J. E. An infinite swapping approach to the rare-event sampling problem. *J. Chem. Phys.* **2011**, *135*, 134111.

(15) Doll, J. D.; Plattner, N.; Freeman, D. L.; Liu, Y.; Dupuis, P. Rare-event sampling: occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods. *J. Chem. Phys.* **2012**, *137*, 204112.

(16) Dupuis, P.; Liu, Y.; Plattner, N.; Doll, J. D. On the Infinite Swapping Limit for Parallel Tempering. *Multiscale Model. Simul.* **2012**, *10*, 986–1022.

(17) Plattner, N.; Doll, J. D.; Meuwly, M. Overcoming the Rare Event Sampling Problem in Biological Systems with Infinite Swapping. *J. Chem. Theory Comput.* **2013**, *9*, 4215–4224.

(18) Zhang, B. W.; Dai, W.; Gallicchio, E.; He, P.; Xia, J.; Tan, Z.; Levy, R. M. Simulating Replica Exchange: Markov State Models, Proposal Schemes, and the Infinite Swapping Limit. *J. Phys. Chem. B* **2016**, *120*, 8289–8301.

(19) Hénin, J.; Chipot, C. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.* **2004**, *121*, 2904–2914.

(20) Ozer, G.; Keyes, T.; Quirk, S.; Hernandez, R. Multiple branched adaptive steered molecular dynamics. *J. Chem. Phys.* **2014**, *141*, 064101.

(21) Comer, J.; Phillips, J. C.; Schulten, K.; Chipot, C. Multiple-Replica Strategies for Free-Energy Calculations in NAMD: Multiple-Walker Adaptive Biasing Force and Walker Selection Rules. *J. Chem. Theory Comput.* **2014**, *10*, 5276–5285.

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

(23) Apostolakis, J.; Ferrara, P.; Caflisch, A. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.* **1999**, *110*, 2099–2108.

(24) Uribe, L.; Gauss, J.; Diezemann, G. Comparative Study of the Mechanical Unfolding Pathways of alpha- and beta-Peptides. *J. Phys. Chem. B* **2015**, *119*, 8313–8320.

(25) Lin, Z.; Riniker, S.; Gunsteren, W. F. v. Free Enthalpy Differences between alpha-, pi-, and 310-Helices of an Atomic Level Fine-Grained Alanine Deca-Peptide Solvated in Supramolecular Coarse-Grained Water. *J. Chem. Theory Comput.* **2013**, *9*, 1328–1333.

(26) Hazel, A.; Chipot, C.; Gumbart, J. C. Thermodynamics of Deca-alanine Folding in Water. *J. Chem. Theory Comput.* **2014**, *10*, 2836–2844.

(27) Tilton, R.; Kuntz, I. D.; Petsko, G. A. Cavities in Proteins: Structure of a Metmyoglobin-Xenon Complex Sol ved to 1.9 AT. *Biochem.* **1984**, *23*, 2849–2857.

(28) Olson, J. S.; Phillips, G. N. Kinetic Pathways and Barriers for Ligand Binding to Myoglobin. *J. Biol. Chem.* **1996**, *271*, 17593–17596.

(29) Scott, E. E.; Gibson, Q. H.; Olson, J. S. Mapping the Pathways for O_2 Entry Into and Exit from Myoglobin. *J. Biol. Chem.* **2001**, *276*, 5177–5188.

(30) Schotte, F.; Lim, M.; Jackson, A.; Smirnov, V.; Soman, J.; Olson, J.; Phillips, G.; Wulff, M.; P., A. Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science* **2003**, *300*, 1944–1947.

(31) Elber, R.; Karplus, M. Enhanced Sampling in Molecular Dynamics: Use of the Time-Dependent Hart ree Approximation for a Simulation of Carbon Monoxide Diffusion through Myoglobin. *J. Am. Chem. Soc.* **1990**, *112*, 9161–9175.

(32) Ruscio, J. Z.; Kumar, D.; Shukla, M.; Prisant, M. G.; Murali, T. M.; Onufriev, A. V. Atomic level computational identification of ligand migration pathways between solvent and binding site in Myoglobin. *Proc. Natl. Acad. Sci.* **2008**, *105*, 9204–9209.

(33) Bossa, C.; Anselmi, M.; Roccatano, D.; Amadei, A.; Vallone, B.; Brunori, M.; Di Nola, A. Extended Molecular Dynamics Simulation of the Carbon Monoxide Migration in Sperm Whale Myoglobin. *Biophysical J.* **2004**, *86*, 3855–3862.

(34) Plattner, N.; Doll, J. D.; Meuwly, M. Spatial averaging for small molecule diffusion in condensed phase envir onments. *J. Chem. Phys.* **2010**, *133*, 044506.

(35) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.

(36) Esque, J.; Cecchini, M. Accurate Calculation of Conformational Free Energy Differences in Explicit Water: The ConfinementSolvation Free Energy Approach. *J. Phys. Chem. B* **2015**, *119*, 5194–5207.

(37) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynskis equality. *J. Chem. Phys.* **2003**, *119*, 3559–3566.

(38) Brooks, B.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(39) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theo. Comp.* **2012**, *8*, 3257–3273.

(40) Patriksson, A.; van der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.

(41) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.

(42) Lloyd, S. Least squares quantization in PCM. *IEEE T. Inform. Theory* **1982**, *28*, 129–137.

(43) J. A. Hartigan, M. A. W. Algorithm AS 136: A K-Means Clustering Algorithm. *J. Roy. Stat. Soc. C-App.* **1979**, *28*, 100–108.

(44) Trebst, S.; Troyer, M.; Hansmann, U. H. E. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* **2006**, *124*.

(45) Katzgraber, H. G.; Trebst, S.; Huse, D. A.; Troyer, M. Feedback-optimized parallel tempering Monte Carlo. *J. Stat. Mech.-Theory E.* **2006**, *2006*, P03018.

(46) Kouza, M.; Hansmann, U. H. E. Velocity scaling for optimizing replica exchange molecular dynamics. *J. Chem. Phys.* **2011**, *134*.

(47) Wiener, N. Generalized harmonic analysis. *Acta Math.* **1930**, *55*, 117–258.

(48) Khintchine, A. Korrelationstheorie der stationären stochastischen Prozesse. *Math. Ann.* **1934**, *109*, 604–615.

(49) Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Statist.* **1956**, *27*, 832–837.

(50) Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* **1962**, *33*, 1065–1076.

(51) Bureau, H. R.; Merz, D. R., Jr.; Hershkovits, E.; Quirk, S.; Hernandez, R. Constrained Unfolding of a Helical Peptide: Implicit versus Explicit Solvents. *PLOS ONE* **2015**, *10*.

(52) Lin, Z.; Liu, H.; Riniker, S.; van Gunsteren, W. F. On the Use of Enveloping Distribution Sampling (EDS) to Compute Free Enthalpy Differences between Different Conformational States of Molecules: Application to 3(10-), alpha-, and pi-Helices. *J. Chem. Theo. Comp.* **2011**, *7*, 3884–3897.

(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(54) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327 – 341.

(55) Hockney, R.; Eastwood, J. *Computer Simulation Using Particles*; CRC Press, 1988.

(56) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(57) Hynninen, A.-P.; Crowley, M. F. New faster CHARMM molecular dynamics engine. *J. Comput. Chem.* **2014**, *35*, 406–413.

(58) Best, R. B.; Buchete, N.-V.; Hummer, G. Are current molecular dynamics force fields too helical? *Biophys. J.* **2008**, *95*, L7–L9.

(59) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.

(60) 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*, 1846–1857.

(61) Reddy, A. S.; Wang, L.; Singh, S.; Ling, Y. L.; Buchanan, L.; Zanni, M. T.; Skinner, J. L.; de Pablo, J. J. Stable and Metastable States of Human Amylin in Solution. *Biophys. J.* **2010**, *99*, 2208–2216.

(62) Luna, V. M.; Fee, J. A.; Deniz, A. A.; Stout, C. D. Mobility of Xe Atoms within the Oxygen Diffusion Channel of Cytochrome ba(3) Oxidase. *Biochem.* **2012**, *51*, 4669–4676.

(63) Frauenfelder, H.; McMahon, B. H.; Fenimore, P. W. Myoglobin: The hydrogen atom of biology and a paradigm of complexity. *Proc. Natl. Acad. Sci.* **2003**, *100*, 8615–8617.

(64) Srajer, V.; Ren, Z.; Teng, T.; Schmidt, M.; Ursby, T.; Bourgeois, D.; Pradervand, C.; Schildkamp, W.; Wulff, M.; Moffat, K. Protein conformational relaxation and ligand migration in myoglobin: A nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochem.* **2001**, *40*, 13802–13815.

(65) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

(66) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y. et al. CHARMM-GUI Input Generator for NAMD,

GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12*, 405–413.

(67) Abraini, J. H.; Marassio, G.; David, H. N.; Vallone, B.; Prangé, T.; Colloc'h, N. Crystallographic Studies with Xenon and Nitrous Oxide Provide Evidence for Protein-dependent Processes in the Mechanisms of General Anesthesia:. *Anesthesiology* **2014**, *121*, 1018–1027.

(68) Cohen, J.; Arkhipov, A.; Braun, R.; Schulten, K. Imaging the Migration Pathways for O2, CO, NO, and Xe Inside Myoglobin. *Biophysical J.* **2006**, *91*, 1844–1857.

(69) Plattner, N.; Meuwly, M. Quantifying the Importance of Protein Conformation on Ligand Migration in Myoglobin. *Biophys. J.* **2012**, *102*, 333–341.

(70) Grossfield, Alan, WHAM: the weighted histogram analysis method, version 2.0.9. Accessed on 8th June 2016; `http://membrane.urmc.rochester.edu/content/wham`.

Graphical TOC Entry