Biochimica et Biophysica Acta 1850 (2015) 889–902

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

A molecular simulation protocol to avoid sampling redundancy and discover new states ☆

Marco Bacci, Andreas Vitalis *, Amedeo Caflisch **

University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 18 June 2014
Received in revised form 21 August 2014
Accepted 25 August 2014
Available online 2 September 2014

Keywords:
Enhanced sampling
Molecular dynamics
Biomolecules
Simulation convergence
Transition path
Scalable algorithm

ABSTRACT

Background: For biomacromolecules or their assemblies, experimental knowledge is often restricted to specific states. Ambiguity pervades simulations of these complex systems because there is no prior knowledge of relevant phase space domains, and sampling recurrence is difficult to achieve. In molecular dynamics methods, ruggedness of the free energy surface exacerbates this problem by slowing down the unbiased exploration of phase space. Sampling is inefficient if dwell times in metastable states are large.
Methods: We suggest a heuristic algorithm to terminate and reseed trajectories run in multiple copies in parallel. It uses a recent method to order snapshots, which provides notions of "interesting" and "unique" for individual simulations. We define criteria to guide the reseeding of runs from more "interesting" points if they sample overlapping regions of phase space.
Results: Using a pedagogical example and an α-helical peptide, the approach is demonstrated to amplify the rate of exploration of phase space and to discover metastable states not found by conventional sampling schemes. Evidence is provided that accurate kinetics and pathways can be extracted from the simulations.
Conclusions: The method, termed PIGS for Progress Index Guided Sampling, proceeds in unsupervised fashion, is scalable, and benefits synergistically from larger numbers of replicas. Results confirm that the underlying ideas are appropriate and sufficient to enhance sampling.
General Significance: In molecular simulations, errors caused by not exploring relevant domains in phase space are always unquantifiable and can be arbitrarily large. Our protocol adds to the toolkit available to researchers in reducing these types of errors. This article is part of a Special Issue entitled "Recent Developments of Molecular Dynamics."
© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The microscopic foundation of life sciences is an appreciation of biomolecules as complex systems. Chemical reactions, binding and assembly phenomena, and conformational transitions can all span a broad range of time and length scales. The desire to understand these processes has produced an unprecedented amount of data obtained using many different techniques. Among these, computer simulations, while marred by the caveat that they model rather than project a physical reality, have been a useful tool due to the level of resolution they offer [1,2]. Indeed, recent work has convincingly demonstrated the appropriateness and potential accuracy of the underlying physical models [3]. Molecular dynamics (MD) simulations [4], the topic of this special issue, are the most common type of computer simulations used for biomolecules. They propagate suitable equations of motion numerically. Auxiliary constructs, such as thermostats, may be required to produce well-defined, thermodynamic ensembles, most often the canonical (NVT) or isothermal-isobaric (NPT) ones.

A single, continuous MD trajectory is often expected to yield correct equilibrium statistics and realistic dynamics, although this is far from a trivial issue, in particular with respect to the numerical discretization [5]. Faithful dynamics can cause undesired precision problems if the underlying rates are low, i.e., simulations that are too short may provide limited information carrying large biases toward the initial state. If a system cannot traverse all relevant, metastable states on the simulation time scale, computed time averages will differ from correct canonical averages. Pragmatically, initial portions of simulations are discarded heuristically as "equilibration periods," and statements about simulation precision must be restricted to specific observables [1]. Due to the above, a canonical MD sampling approach (CS) will often utilize resources inefficiently, and this inefficiency has motivated the development of enhanced sampling methods over the past few decades [6–9].

It is well beyond the scope of this introduction to provide an overview of all enhanced sampling methods, and we apologize for inevitable omissions. In the following, we highlight conceptual aspects of different

Abbreviations: CLD, Cartesian Langevin dynamics; CS, canonical sampling; MC, Monte Carlo; MD, molecular dynamics; MSM, Markov state model; PCA, principal component analysis; PES, potential energy surface; PIGS, Progress Index-Guided Sampling; REX, replica exchange; RMSD, root mean square deviation

☆ This article is part of a Special Issue entitled "Recent Developments of Molecular Dynamics."

* Corresponding author. Tel.: +41 44 635 5597.
** Corresponding author. Tel.: +41 44 635 5521.
E-mail addresses: a.vitalis@bioc.uzh.ch (A. Vitalis), caflisch@bioc.uzh.ch (A. Caflisch).

http://dx.doi.org/10.1016/j.bbagen.2014.08.013
0304-4165/© 2014 Elsevier B.V. All rights reserved.

classes of methods. Better MD integrators and in particular multiple time step methods can be viewed as advanced sampling methods as they allow larger (average) time steps to be used [5,10]. The same is true for efficient protocols for constrained dynamics [11,12] or mixed numerical and analytical schemes [13]. Dynamics can also be altered by scaling masses without introducing configurational bias [14].

We next want to mention knowledge-based approaches of two different kinds. The Rosetta [15] and similar methods essentially utilize database-derived, conformational biases in conjunction with hybrid sampling protocols to explore large regions of putatively relevant phase space. For polypeptides, this can be useful if the primary goal is to identify possible states of interest. Systematic coarse-graining [16,17] is a general strategy to achieve better sampling by increasing the ratio of simulation and CPU times and often also by smoothing the ruggedness of the potential energy surface (PES). One of its main challenges lies in maintaining protocols for subsequent fine-graining that could then yield physically realistic ensembles at the resolution of interest. Both ideas have been coupled to the replica exchange (REX) method discussed below [18,19]. We mention this aspect to highlight the difficulty in delineating classes of enhanced sampling methods, which often combine existing elements and ideas in innovative ways.

Rather than the smoothing incurred by coarse-graining, direct modifications to the PES can be used to control populations and exploration rates also for fine-grained systems. Two highly successful examples are umbrella sampling [20] and metadynamics [21] (or more generally, flat histogram methods) [22], both of which are usually meant to extract an estimate of an unbiased probability distribution or a density of states along a reaction coordinate. Steered MD simulations [23] use force rather than a potential, but can be viewed as comparable. All methods are excellent for generating potentials of mean force [24] that may even allow the prediction of coarse-grained dynamics. The major caveats of this class of methods are as follows. First, choice of reaction coordinates may not be straightforward. Second, there is no control over directions orthogonal to the chosen reaction coordinates [25]. This is a fundamental problem of low-dimensional projections, viz., different, kinetically and geometrically distal states mapping to the same value of the reaction coordinate. Third, the trajectories obtained are of limited scope because populations, microscopic rates, and pathways are all altered by the modifications to the PES. Detailed kinetic information is lost, and the effective, statistical weight of a large fraction of the data may be negligible. In practice, for large systems, the snapshot-based reweighting to recover equilibrium ensembles is too noisy [26]. This issue underlies similar approaches such as the accelerated molecular dynamics method [27].

Among multicanonical techniques, the most prominent one is the REX method [28,29], and most often temperature is used to globally scale the PES. By careful algorithm design, and by partitioning the data by temperature, each subset of the data can be analyzed as a *bona fide* canonical ensemble. The idea of the method is that excursions into higher temperatures facilitate barrier crossings, despite the absence of a dedicated geometric coordinate [30]. When following data at constant temperature, the perturbations incurred by the REX method consist entirely of swapping in compatible structures from neighboring conditions. It is precisely this design that gives the method its broad appeal [31] but also limits its price-to-performance ratio because the spacing and number of conditions to use cannot be considered free choices [32–34]. If data are required only at a single condition, multiple independent runs may utilize resources more efficiently. Due to its widespread use and simplicity, we choose the REX method for comparison purposes in this contribution. Relative to CS, REX poses difficulties when inferring kinetics and pathways [35,36] because rigorously only those stretches of the trajectories in between swaps can be mined for this purpose [37].

An elucidation of pathways is of great interest for obtaining a mechanistic understanding of complex systems. Given a notion of two states to connect, techniques like transition path sampling [38], the nudged elastic band method [39], or the string method [40] are exceptionally powerful tools to understand pathway heterogeneity, predict net rates, etc. Imposing a preconceived geometry on a path ensemble may be beneficial [41,42]. If the system displays a separation of time scales, it may be possible to construct Markov state models (MSMs) [43,44], which coarse-grain phase space into a set of kinetically homogeneous or metastable states. The initial set of states is often inferred from long CS simulations or some of the enhanced sampling methods outlined above [45]. The generalization of transition path sampling approaches to include all states in a network, i.e., the combination of these two ideas [46], can potentially provide a comprehensive picture of the thermodynamics and kinetics of a complex system at a given condition and at the level of the resolution of the states of the MSM. In this contribution, we suggest a method that addresses two of the underlying goals, viz., obtaining realistic pathway information and achieving fast phase space coverage.

In order to preserve pathway information, it seems necessary to sample from an unaltered PES. A possible strategy is to guide sampling by simply restarting simulations from interesting points, a process we refer to as reseeding. In distributed computing, a fluctuation-based heuristic was suggested to monitor relevant transitions [47]. Trajectories are selectively reseeded from those points indicating that a relevant transition has occurred. Recent work has used kinetic reaction coordinates to guide sampling toward new states [48,49]. Here, we suggest a different heuristic that rewards uniqueness of the current sampling domain of individual trajectories. The scheme is scalable, unsupervised, and explicitly parallel. The decision about reseeding a given trajectory depends on the regions of phase space sampled by other trajectories. The notion of uniqueness as a guide makes it most similar to the recent WExplore method [50] that defines states by spatial discretization [51,52] to inform the reseeding procedure.

We term our approach PIGS (Progress Index-Guided Sampling) as it relies on an efficient ordering of a slice of simulation snapshots, the so-called progress index [53]. The remainder of the text is structured as follows. We first introduce the algorithm and simulation protocols. We then provide a detailed set of results evaluating the performance of the scheme on two systems, viz., a 1D model and the FS peptide [54] in implicit solvent. We are able to demonstrate that PIGS, while minimally invasive at the level of pathways, amplifies the rate of exploration, i.e., we detect several metastable states that are not reached by either REX or CS on the same time scale. By definition, PIGS ensembles are thermodynamically biased, and we do not consider refinement or reweighting here. This issue is, among others, discussed in the final section of this manuscript.

2. Materials and methods

In this section, we introduce the algorithm and describe the general setup and sampling protocols.

2.1. The PIGS algorithm

Consider a set of N_r molecular simulations (replicas) of exactly the same system and under the same conditions that are propagated by a given base sampling algorithm, e.g., CS. The stochastic sampling algorithms we use here are either Metropolis Monte Carlo or Langevin dynamics. We set a deterministic interval, f_p, for attempting to reseed up to N_r–N_p of the simulations with the final configuration (machine precision) of any of the N_p remaining replicas. The decision whether to reseed a replica or not relies on a heuristic that utilizes data from all replicas and is history-free, i.e., only data from the last f_p steps enter the analysis. The number of snapshots to use per replica for an interval of length f_p is constant and referred to as n_O throughout. Thus, the scheme is scalable and explicitly parallel. It is easy to recognize that independent runs using the base sampler are obtained if $N_r = N_p$ or if $N_r = 1$. As we will see, the heuristic is designed as an unsupervised

learning protocol meant to optimize coverage while using as few perturbations to the base sampler as possible. This may not always be a useful motivation, and the protocol can be adjusted to reflect changes in motivation.

2.1.1. Reseeding heuristic

The heuristic used in this manuscript is aimed to avoid redundant sampling of the same area of phase space by several replicas. For systems with either too much or too little degeneracy, this may not be an appropriate choice. The notion of degeneracy is defined only for the level of representation chosen, i.e., it can be arbitrarily focused on parts of the system, such as a flexible loop in a protein. The observations from all replicas are collected and ordered by a recent algorithm operating in near-linear time [53]. The ordered sequence, or progress index, corresponds to a specific path through an approximation to the minimum spanning tree. The path is obtained by always adding the closest available vertex. This corresponds to always adding the snapshot that has the smallest distance to any snapshot added previously. Distance is taken as the Euclidean distance for the chosen representation, e.g., dihedral angles. The algorithm corresponds to an unsupervised protocol stepping through regions of high sampling density one after another. The progress index contains data from all replicas, and the origin of each observation is known. We can therefore rank the final snapshots of each replica by the following criteria:

(a) The position in the progress index. Positions toward the end are more likely to be associated with low density (barrier) regions and to not correspond to the regions that are currently the most populated ones. This is because the starting snapshot is always taken as the representative of the largest cluster of an auxiliary clustering of the current data slice.
(b) The distance on account of which it was added. This distance is the shortest one to any other snapshot previously added. Long distances are another indicator of being in a low density region.
(c) The smallest distance to any other final snapshot. This is an indicator of uniqueness as related conformations tend to appear in close proximity in the progress index.

All 3 quantities are sorted in decreasing order. We construct a composite rank, $\zeta(R)$, as the sums of the individual ranks, and the replicas with the smallest composite ranks are assumed to be located in interesting regions of phase space. Since the goal is to improve phase space coverage, a conformation is deemed interesting if it is in a region that has low sampling density and is not explored by other replicas. The top N_p replicas by composite rank are guaranteed to continue unperturbed sampling. For every remaining replica, R_Y, we randomly pick one of the top N_p replicas, G_X, with uniform probability and evaluate:

$$p(R_Y \rightarrow G_X) = [\zeta(R_Y) - \zeta(G_X)]/(\Delta\zeta_{max} + 1). \quad (1)$$

Here, $\Delta\zeta_{max}$ is the maximum difference in summed ranks across all replicas. If a random number drawn uniformly from the unit interval is less than the probability in Eq. (1), the reseeding is putatively accepted. Note that the trajectory of a reseeded run is irrevocably terminated. Because the composite rank of the final snapshots may be an incomplete indicator of "interesting" we currently perform an additional check. Specifically, a reseeding accepted via Eq. (1) may still be rejected if the difference of the 25% and 75% quantiles (1st and 3rd quartiles) of the snapshots of R_Y in terms of their positions in the progress index is less than n_O. This rule is insensitive to outliers and measures progress index locality. If locality is high, we can infer that the snapshots from replica R_Y form a relatively unique subset, which is the motivation for the rejection of the reseeding.

2.1.2. Scalability

For scaling reasons, we assume that each of the N_r replicas collects a fixed number of observations of the system, n_O, in memory over any given stretch of f_p steps. This means that the total size of the data collected over a single stretch will scale at most linearly with N_r. Therefore, the cost of the reseeding scheme can be kept constant if we employ a parallelizable algorithm that operates in linear time. We currently compute the progress index and reseeding heuristic on a single node, and for this manuscript the parameters are such that the overall cost is small enough to not strongly affect the overall performance. Note that our discussion of parallelism is completely independent of any domain decomposition one may use to speed up the base sampler for every replica. For systems with too many degrees of freedom of little interest (e.g., simulations in explicit water), it is practically inevitable to use a reduced representation of the system that discards these degrees of freedom. This is not a particular feature of our protocol, however, and common choices such as coordinates of subsets of atoms or dihedral angles can all be used. We emphasize that there is no scaling issue with system size as the analysis scheme scales more favorably with the number of atoms than the base sampler in all but trivial cases. After the heuristic presented in Section 2.1.1 has been evaluated, the observations in memory are deleted, i.e., the history is forgotten completely. The algorithm itself poses no I/O requirements whatsoever. Because we restrict reseedings to the final configurations, the complete information required for reseeding a given replica is available in the system memory of a different replica, and this includes velocities and all other quantities required by the integrator in question. If we used a fully deterministic base sampler, e.g., MD with a Nosé-type thermostat [55], a stochastic component could be introduced by randomizing velocities after each successful reseeding. This is required to avoid sampling of (nearly) identical trajectories diverging solely because of numerical drift.

2.1.3. Parameters

The primary parameters are as stated, viz., N_r, f_p and N_p. We explore all 3 for the model system and N_r for the FS peptide (see Results). We note that N_r and f_p are shared with the REX method. For the rank analysis of final snapshots to make sense, n_O should be large enough (we have most often used values in the hundreds). The choice of representation as discussed in Sections 2.1.1 and 2.1.2 can be considered a parameter as well. It is algorithmically relevant exclusively as the distance function underlying the progress index construction. We therefore expect that it can be used to guide phase space exploration and coverage. We evaluate the performance of different representations for the FS peptide. It is important to emphasize that the requirement to represent the system at reduced dimensionality is shared with many similar methods, e.g., in the definition of states for either transition path sampling [38] or MSMs [43,46]. Lastly, the progress index requires minor auxiliary parameters for controlling the quality of the spanning tree and the preorganization of the data [53,56]. We do not consider these parameters essential to the performance of the method, and no sensitivity analyses are presented here.

2.2. Systems and simulation protocols

All simulations and most analyses were carried out with the CAMPARI molecular simulation package (http://campari.sourceforge.net), and the latest development version is available on request (campari.software@gmail.com). Further analyses were scripted in R and molecular graphics were generated using VMD [57].

2.2.1. Model system

We use a 1D model system to illustrate the PIGS algorithm. The potential as shown in Fig. S1 is constructed as a sum of Gaussians and spans 500 position units. Barriers of height 5 kcal/mol at positions 0 and 500 contain the particle. A rugged landscape is constructed by placing barriers of height 1 kcal/mol every 5 units except at multiples of 25 where a higher barrier of 2 kcal/mol is used. The system models diffusive evolution on a rugged PES with some hierarchy of time scales due to the presence of two relevant barrier heights. There are 100

well-defined free energy minima. The PIGS algorithm, at least as presented above, is not designed to enhance sampling in difficult systems with low degeneracy, i.e., with few relevant states separated by high, enthalpic barriers.

The model is explored using a Metropolis Monte Carlo (MC) algorithm at 250 K with step attempts drawn randomly and uniformly from an interval of size 0.4. The MC sampler ensures that ensemble concerns and memory effects are avoided altogether. Data for the position of the particle were collected every 10 steps. Simulations were generally run for 10^6 steps and all replicas started from the same positions with the particle set at 1.5. The particle never escaped the boundaries at 0 and 500. The exploration rate is measured by the time taken to reach positions of increasing values. For all combinations of parameters, 10 independent runs were performed. Possible values for N_r were 8, 16, 32, and 64. In each case N_p values of $N_r/4$, $N_r/2$, $3N_r/4$, and N_r were tested. The latter corresponds to unperturbed MC sampling. Default values for f_p and n_O were 1000 and 100, respectively. For specific settings, we also systematically varied f_p (1, 4, 16, and 64×10^3), n_O (100, 25, 6–7, 1–2), and the temperature (150, 200, 250, and 300 K).

2.2.2. FS peptide

The FS peptide, acetyl-A_5(AAARA)$_3$A-N'-methylamide [54], was simulated with the ABSINTH continuum solvent model [58] as in prior work [59]. The ABSINTH model is based on its own interaction model and Lennard–Jones parameters, and takes values for fixed, partial charges and some bonded potentials from the OPLS-AA/L force field [60]. Neutralizing counterions and a background of ~150 mM NaCl were also contained in the simulation droplet of radius 40 Å. Here, we used Cartesian Langevin dynamics (CLD) as the base sampler employing an impulse integrator according to Skeel and Izaguirre [61]. Using SHAKE [11], we constrained all bond lengths, i.e., the resultant data are comparable exactly to the "Flex." data in our prior work [59]. With a universal friction coefficient of 1 ps^{-1}, an integration time step of 3 fs, and masses of hydrogen atoms scaled by a factor of 4.0, we obtained robust integration despite the presence of truncation cutoffs at 12 Å (mean temperature errors of ~1 K). Polar interactions between groups carrying a net charge were not truncated but computed at monopole resolution. The simulation time per replica was 312 ns unless otherwise noted. Trajectory data for every replica were collected every 1.5 ps. The total data obtained exceed 10^7 snapshots. Simulations were run on the Schrödinger supercomputer at the University of Zurich. Each replica was run on a single core and took 6–8 days to complete. Except for the data discussed in Section 3.2.5, all simulations started from the same snapshot with the FS peptide forming a straight α-helix.

We considered two types of PIGS runs by using two different representations for the construction of the progress index [53], viz., the peptide's relevant dihedral angles (ϕ-PIGS) or a manually selected set of interatomic distances between peptide atoms (r-PIGS). Details are provided in the SI. We do not use the root mean square deviation (RMSD) of Cartesian coordinates here because of the additional cost introduced by the required alignment operator, and the r-PIGS data are a suitable replacement. Parameters f_p, N_p, and n_O were fixed and set to values of 10,000, $N_r/2$, and 200, respectively. N_r was 32 for nearly all the data presented. We compare these runs to CLD simulations of identical set up and length. Allocation of resources is not a straightforward task when trying to compare to REX. Here, we set up 4 independent runs with 16 temperatures each that include 250 and 290 K (see SI for details). Swaps between 15 randomly selected neighbor pairs were attempted every 10,000 steps (the same as the reseeding interval for PIGS). We wanted to achieve the same total number of cores (64) to obtain data at the two temperatures in all 4 cases. If we were interested in data at further temperatures, CLD and PIGS but not REX would have to use additional resources.

2.3. Data analysis for the FS peptide

Unless noted otherwise, all analyses were restricted to the data obtained at 250 K.

2.3.1. Exploration rate

In order to measure the exploration rate of the different sampling protocols, we defined a two state model (0 or 1) at the residue level. Given that we expect a high helix content for the FS peptide, we assume a residue to be in state 1 if its ϕ- and ψ-angles fall into the α-helical region of the Ramachandran plot and to be in state 0 otherwise. To reduce the spurious counting of fluctuations, we also defined a boundary region separating the two states that is shown in Fig. S5(b). If any residue of interest was found to reside in this boundary region, the corresponding structure did not count toward the exploration rate. To keep the total number of different 1/0 sequences tractable, and because of the low level of coupling of terminal residues, we discarded both the first and the last two residues from the two-state assignment. This yields $2^{17} = 131{,}072$ possible different configurations. For instance, (0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0) is the state in which all 17 central residues of the FS-peptide are outside of the α-helical region. The exploration rate is calculated as the number of unique states visited as a function of time. It should be noted that these results are sensitive to data set size, i.e., comparison between REX and the other samplers is not straightforward. Here, we performed bootstrapping on the CLD and PIGS data to overcome this difficulty.

2.3.2. Clustering

For clustering, the data from all 4 samplers (32 replicas at 250 K from CLD, ϕ-PIGS, and r-PIGS each, and 12 replicas from 4 independent runs at temperatures of 244, 250 and 256 K from the REX protocol) were concatenated. This is the only analysis for which we lump information from different temperatures together, and it is meant to avoid that states sampled predominantly by REX are missed simply on account of the low weight of the REX data. The time resolution was 3 ps yielding ~1.1×10^7 molecular configurations. We used a recently developed tree-based algorithm [56] particularly suited for the task as it scales linearly with data set size, produces clusters that are free of overlap, and reflects local sampling density. Additional information on required parameters is available in Table S2. To measure the level of similarity among molecular conformations, we used the same set of interatomic distances underlying the r-PIGS runs (see Section 2.2.2 and Table S1).

At a size threshold of 1.8 Å, the algorithm yielded about 250,000 clusters. Of these, ~100,000 were made of single snapshots, and 10,700 had a population equal to or greater than 100. The representative snapshots of the 600 most populated clusters (representing roughly half of the data set size) served as input data to an additional grouping step, which we performed with a bottom-up, hierarchical algorithm and a size threshold of 1.5 Å yielding 101 clusters. The motivation for this step was to obtain a set of clusters that are geometrically not directly adjacent to one another. For subsequent analyses, we focused on the 101 clusters associated only with those snapshots identified by the hierarchical scheme to be representative of each of them.

2.3.3. Network and transition paths

The layout for a 2D visualization of the complex network obtained by the clustering was generated with the Kamada–Kawai algorithm [62]. Uniqueness of a cluster with respect to a sampling protocol was defined as follows. Given the number of all snapshots constituting an individual cluster, we identify the sampler that produced the trajectory containing every snapshot in the cluster (CLD, r-PIGS, ϕ-PIGS, or REX). The number for a given sampler is divided by the total number of snapshots to define uniqueness with respect to a sampling protocol. A uniqueness of 100% for CLD would mean that the cluster is sampled only in CLD trajectories.

The smaller number of REX snapshots (all subsequent analysis is restricted to a single temperature of interest and does not include adjacent temperatures as in Section 2.3.2) was corrected by naive rescaling of these data by a factor of 32/4. The edges of the network, i.e., the observed transitions among the 101 clusters, were assigned by parsing – replica by replica and protocol by protocol – the continuous stretches within trajectories. For CLD, these stretches corresponded to the entire trajectories of individual replicas, whereas for PIGS and REX they corresponded to the stretches between two actual reseedings or swaps. Clearly, these stretches also contain snapshots that do not belong to any of the 101 clusters that we refer to by number in the following. For transition path analyses, these points were all assumed to constitute a global boundary region, and successful transition paths were those segments connecting directly two of the states without encountering any other numbered state in between. We also carried out principal component analysis (PCA) [63] on the molecular conformations sampled during four specific transitions that were well-populated by a majority of algorithms. PCA was performed over a subset of trajectory frames containing all snapshots for the two end states and all snapshots on successful transition paths. The coordinates used for PCA were the same interatomic distances used for the r-PIGS data set and in clustering (see Table S1 for details).

3. Results

We have evaluated the PIGS algorithm for two different systems, viz., a toy model and the FS peptide in implicit solvent. All data are compared to the unperturbed base sampler (either MC or CLD). Both systems explore phase space in a predominantly diffusive manner and visit a significant number of metastable states, and we chose them because of this. For the FS peptide, we also compare to the REX method. We emphasize that there are no hidden costs associated with the PIGS simulations, i.e., setup and post-processing are no different from the other cases. Note, however, that post-processing would incur extra costs if we attempted to reweight the resultant distributions for PIGS (and REX [64]).

3.1. Model system

The model system (Fig. S1) serves primarily as a conceptual test for the algorithm. We use it to illustrate parameter sensitivity in detail, which is generally intractable for more complex systems. As described in Section 2.2.1, the system, while always starting from the leftmost state, uses an MC propagator with small step sizes to explore a 1D rugged surface containing 99 metastable states. Because there is only a single dimension, construction of the reseeding heuristic should not be marred by dimensionality or accuracy concerns. Specifically, the ordering of snapshots is expected to track the coordinate itself. Fig. S2 contains an example for the reseeding procedure that illustrates the description in Section 2.2.1.

Comparison of Fig. 1(a)–(c) to (d) demonstrates that PIGS offers a substantial speed-up of exploration for this toy model. This is irrespective of the chosen value for N_p. The shaded areas in the plots represent the envelopes defined by 10 independent runs for a given set of parameters. These estimates of data spread confirm that differences are significant. As expected, small values of N_r offer the fewest benefits throughout, and with 64 replicas the speed-up is always maximal for the cases studied. We emphasize that this is

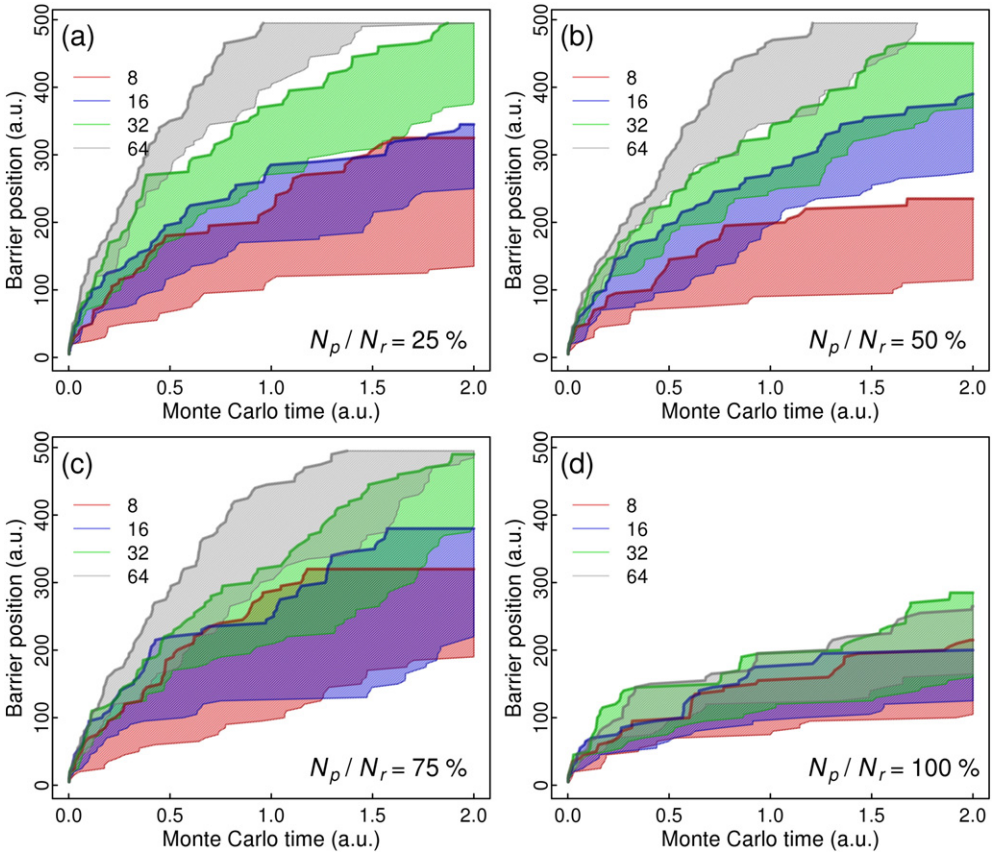

Fig. 1. Exploration rate for the model system for different values of N_r and N_p. There are 10 runs for each combination of parameters. For every run, we recorded the earliest sampling time at which each of the 99 barriers has been crossed. This crossing time is taken as the minimum value across all replicas. The envelopes defined by the resultant 10 curves are indicated by the shaded areas. Thick solid lines highlight the fastest exploration rates across all runs. (a) Data are shown for $N_p = N_r / 4$ and 4 different values of N_r. (b) The same as (a) for $N_p = N_r / 2$. (c) The same as (a) for $N_p = 3N_r / 4$. (d) The same as (a) for $N_p = N_r$, which corresponds to unperturbed MC sampling.

not a trivial consequence of the number of replicas used since, owing to the diffusive nature of the system, there is no robust speed-up seen in Fig. 1(d). Fig. S3, which plots the same data in a different arrangement, highlights that there is no choice of N_p that clearly outperforms the explored alternatives. This result motivated the use of $N_p = N_r / 2$ for all simulations of the FS peptide.

We performed additional analyses of parameter sensitivity for the model system. For these, we fixed the values of N_r and N_p to 32 and 8, respectively. In general, the variability of the speed-up seems to decrease with decreasing N_p, and this motivated the choice. While we cannot study the effects of coarse-graining representation for the 1D system, we did evaluate robustness with respect to the reseeding interval, f_p, and the amount of data collected per replica for construction of the reseeding heuristic, n_O (Section 2.1.1). As seen in Fig. S4, there is surprisingly little dependency on both parameters over two orders of magnitude. To us, this indicates that the heuristic is, under admittedly favorable circumstances, able to efficiently diversify the sampling domains of individual replicas, in particular if N_r becomes larger. This implies that one of our design goals, i.e., obtaining long enough stretches of unperturbed evolution to study pathways, is achievable. We emphasize that the length of continuous trajectory segments is often much larger than f_p.

We conclude the discussion of the toy model by pointing out that all speed-up depends on starting all replicas in the same position. This mimics a common situation in molecular simulations of complex systems, where a well-defined reference state, e.g., a crystal structure, is given but no diverse ensemble of starting structures can be accessed easily. Conversely, for polymers of low sequence complexity or mostly disordered systems, this problem may indeed be solvable [65], and this might override exploration benefits offered by PIGS or other reseeding protocols.

3.2. FS peptide

At low enough temperatures, the FS peptide samples a diverse ensemble of states rich in α-helical content. At 290 K, which is close to the melting temperature for the model in use (see Fig. S5(a)), fluctuations are large, and interconversion is fast. The straight α-helix is the dominant helix-rich state. The size of fluctuations spanning coil-like and compact helical states motivated inclusion of this temperature. Conversely, at 250 K, collapse is a major driving force, and a diverse ensemble of two-helix bundles and partially helical states is seen. Under these conditions, interconversion is slowed down significantly. Note that there is no explicit temperature dependency of parameters and of course no phase transition in the implicit solvation model [58]. From prior work [59], we hypothesized that the time scale of the simulations would be sufficient to obtain converged data at 250 K via CLD, albeit barely so.

For the FS peptide, we compare PIGS to CLD and REX. As outlined in Section 2.2.2, we tried to make the comparison fair in terms of resources invested by utilizing a total of 64 replicas for each protocol. For REX, this implies that a large portion of the data are not directly useful as they correspond to temperatures different from 250 and 290 K. Importantly, we did not perform any parameter optimization for PIGS. The results provide indirect evidence regarding the robustness of the approach. The reseeding/swapping interval for PIGS and REX was 30 ps, but, as shown below, the actual rate was much lower for PIGS. In contrast to the toy model, the FS peptide does allow us to explore the impact of geometric representation on the algorithm, and we obtained independent data sets based on either interatomic distances (r-PIGS) or dihedral angles (ϕ-PIGS). The results are structured as follows. First, we provide an overview of the system's complexity (Section 3.2.1). We then quantify rates of exploration starting from a straight α-helix for all samplers (Section 3.2.2). This is followed by an analysis of bias in both configurational statistics (Section 3.2.3) and pathway information (Section 3.2.4). We conclude with an investigation of the dependency on starting structure (Section 3.2.5).

3.2.1. Conformational landscape at low temperature

To familiarize the reader with the system, Fig. 2 displays a network representation of 101 clusters corresponding to distinct and highly sampled geometrical states visited by the FS peptide at 250 K. As explained in Section 2.3.2, these conformations are obtained from a clustering of the composite data set of runs starting from a straight α-helix. A wide variety of conformations are illustrated by the cartoon representations, which portray the central snapshot (taken as the representative structure for a given cluster) of various states of interest. From Fig. 2, we infer that the CLD sampling protocol visits a well-connected but limited subset of states. Consistent with the initial condition and the assumed conformational preferences of the peptide, this subset is dominated by helix-rich structures (cartoons for states labeled 1, 2, 30, and 72) and few excursions into non-helical, collapsed states, e.g., state 45. None of these states is sampled exclusively by CLD. In contrast, the network for the ϕ-PIGS data, panel (b), indicates access to a wider range of characteristic structures, many of which are not sampled by other protocols. Among these, there are several, collapsed globules, e.g., states 5 and 29, and even structures with significant β-content, viz., states 9 and 10. As analyzed in more detail below, the connectivity of the overlapping regions of the network appears qualitatively similar to that observed in CLD. However, in ϕ-PIGS, a greater number of densely sampled states are achieved by an equivalent amount of data, and therefore the weights of states and transitions in the helical region are reduced. Indeed, transitions reaching states that are unique to ϕ-PIGS are poorly sampled indicating that they are associated with large free energy barriers. In many cases, these states are visited by just one trajectory.

Comparison of Fig. 2(c) and (b) reveals that the choice of coarse-grained representation has little influence on the qualitative performance of the PIGS algorithm. We stress, however, that the non-helical states visited by r- and ϕ-PIGS, respectively, have low overlap. As in panel (b), the network for r-PIGS indicates less sampling in the helical region but appears to retain information about the relative likelihood of transitions between states. Conversely, the REX data for 250 K, panel (d), suggest a sampling domain that is very similar to CLD but with altered weights of transitions between states. While REX data are underrepresented in the clustering, we note that only 24 of the 10^4 largest clusters (84% of the data) are unique to REX. Thus, in summary, Fig. 2 provides a qualitative picture of the conformational landscape as explored by the different protocols. Using CLD as the reference, it appears that PIGS causes a substantial thermodynamic bias, whereas REX yields biased pathway information. We return to these issues in Sections 3.2.3 and 3.2.5.

3.2.2. Rate of exploration

While Fig. 2 indicates that PIGS provides quicker coverage of phase space, this is a qualitative observation that depends heavily on the clustering procedure and our choices. Due to the α-helical nature of the peptide, we pursued a different type of coarse-graining relying on a two-state model at the residue level (see Section 2.3.1). This allows the definition of 2^{17} states, and we can measure the rate at which new states are discovered. These rates are plotted in Fig. 3 for different sampling algorithms and the two temperatures of interest.

As seen in Fig. 3(a), at 290 K, CLD and both PIGS schemes produce a very similar rate of exploration. At a sampling time of 300 ns per replica, ~30% of all states have been visited. REX uses 8 times less data in terms of absolute numbers, and this lowers the measured rate. Correspondingly, the data at 250 K shown in panel (b) show a similar ratio of ~4 between REX and CLD. However, both ϕ-PIGS and r-PIGS discover new states at a much higher rate at this lower temperature. We note that the swapping protocol used in REX creates a mixing of information, which complicates straightforward comparison. We

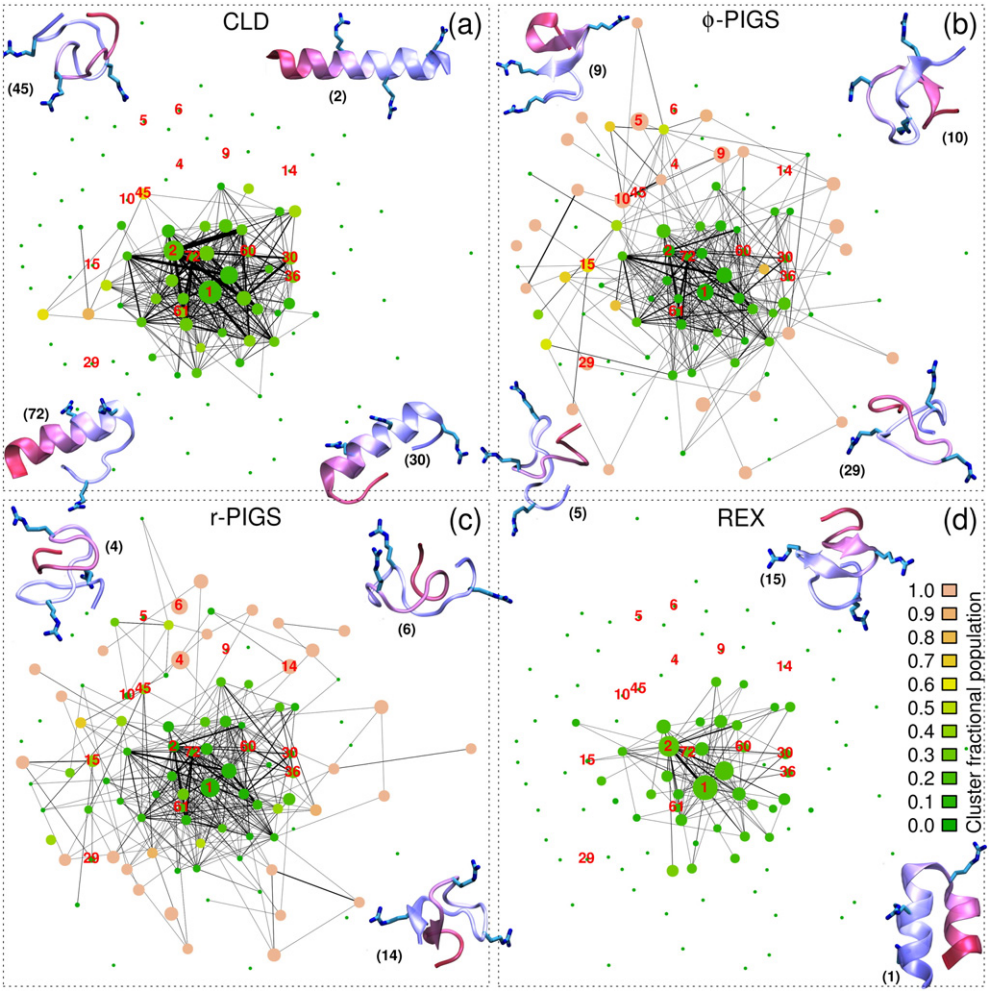

Fig. 2. Complex network representation for the FS-peptide at 250 K. Data were gathered from all sampling protocols to define states (see Section 2.3.2). Node diameters are proportional to the 4th root of the fractional cluster population for each sampler. The size is buffered so that nodes not visited by a given protocol remain visible. The color scheme marks how unique a state is with respect to a given protocol (see exact definition in Section 2.3.3). Links represent direct transitions as observed along continuous stretches of single trajectories (see Section 2.3.3). Their width is proportional to the square root of the normalized number of transitions for each sampler (transition probability). Selected, numbered states are depicted in cartoon representation. (a) Conformational network when using only data from CLD runs. (b) The same as (a) for ϕ-PIGS. (c) The same as (a) for r-PIGS. (d) The same as (a) for REX.

therefore construct another extreme comparison that should favor REX by bootstrapping the CLD and PIGS data to the number of replicas utilized for REX. This analysis is shown in Fig. 3(c) for 250 K, and the data clearly support the notion that REX offers no benefits over CLD for this system and chosen initial condition. Conversely, both PIGS schemes outperform both REX and CLD reliably. Given that the two-state model is constructed based on dihedral angle values, it may be surprising that r-PIGS appears to be more efficient than ϕ-PIGS at low temperature. However, it should be kept in mind that the choice of representation for r-PIGS emphasizes backbone degrees of freedom, which may explain the small difference. Fig. S6(a) provides the corresponding analysis at 290 K.

Panel (d) of Fig. 3 contains an illustration of the r-PIGS scheme at 250 K. The color annotation reveals that the RMSD to the straight α-helix is low for only a single replica at any given time. It is seen clearly that this particular conformation is almost always exited from by virtue of a reseeding to a dissimilar structure rather than by unperturbed evolution. It also emerges that reseedings of two or more replicas tend to coincide in time. This suggests that the intended detection of overlapping sampling domains, which underlies the reseeding heuristic, is successful. Fig. S6 displays analogous plots for some of the other data sets. In particular, panel (b) of Fig. S6 emphasizes the fast inherent dynamics at 290 K, which leads to more reseedings and less bias. The same reasoning as for the straight α-helix holds for any other basin of attraction. As a result, the average population of any given state is unlikely to exceed $1/N_r$, which will generally result in a thermodynamic bias. This is discussed next.

3.2.3. Configurational and ensemble bias

We are interested in assessing to what extent the equilibrium distributions of physical observables differ from protocol to protocol. In reseeding approaches, the only obvious source of bias is that from terminating simulations and reseeding with biased initial conditions. Initial state bias is also inherent to CS. Therefore, the question of bias is one of time scales, and it relates to the diffuse assumption of recurrence (ergodicity) on an observed sampling domain over a finite sample size [1,66,67]. Thus, it is expected that differences are reduced when interconversion between states is faster, i.e., we expect all distributions to be more similar to one another at 290 K than at 250 K.

Panel (a) of Fig. 4 compares histograms of the radius of gyration at both temperatures for the four samplers. At 250 K, the CLD and REX protocols generate the same statistics, which is expected. The right peak represents primarily straight α−helical conformations and accounts for about half of the overall population. On the contrary, both PIGS data sets sample distributions that are very similar to one another but have depleted density at 9.5 Å. This is consistent

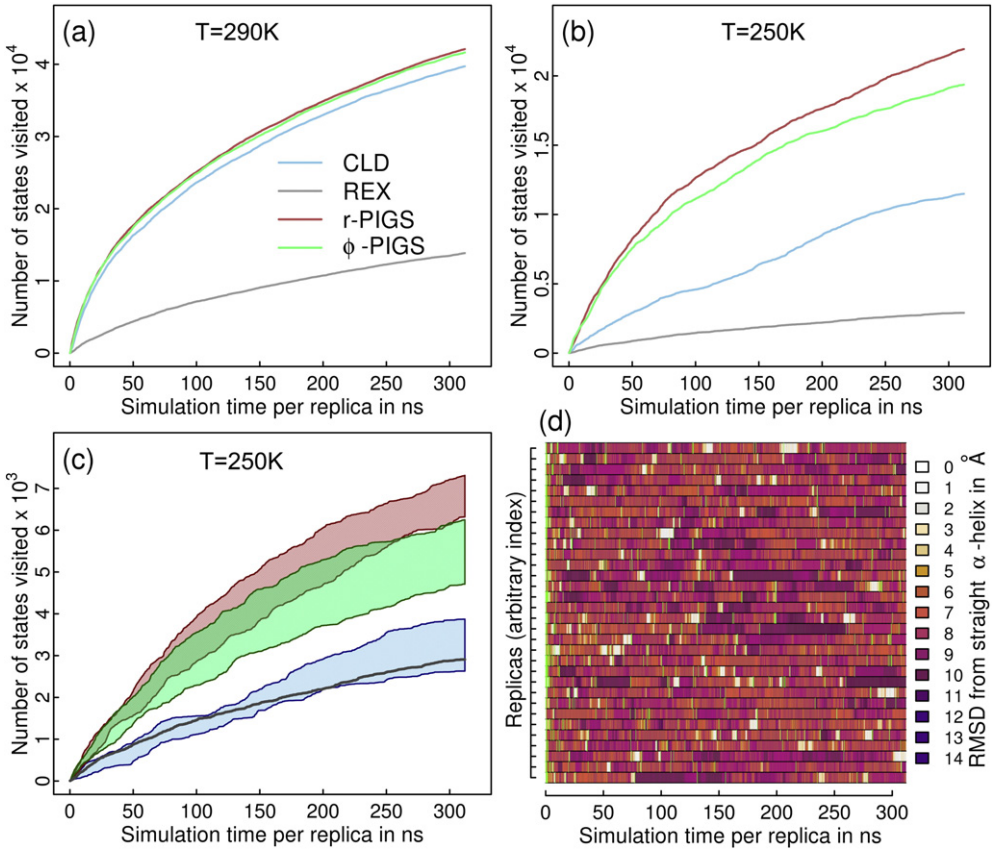

Fig. 3. Rate of exploration for different samplers and temperatures. (a) We plot the number of states visited (see Section 2.3.1 for details) as a function of time at 290 K. The REX curve relies on a data set that is smaller by a factor of 8. (b) The same as (a) for 250 K. The same legend as in (a) applies. (c) To make data comparable to REX, the CLD and PIGS runs are randomly bootstrapped to 10 sets of 4 replicas without replacement. The resultant envelopes over the 10 sets are plotted as shaded areas. The same legend is used as in (a), and data are shown for 250 K only. (d) To illustrate the reseeding, we plot the RMSD from a straight α-helical conformation for r-PIGS at 250 K. The color annotation is shown separately for every replica and for every 20th snapshot (net time resolution is 30 ps). Vertical bars (green) indicate the times of all actual reseedings for this data set.

with the previous discussion in the context of Fig. 3(d). Importantly, the bimodal nature of the distributions sampled by CLD and REX is retained and density is simply shifted to collapsed structures with radii of gyration around 7.0 Å. These include the disordered globules seen in Fig. 2. At 290 K (smooth, dashed lines), the bias introduced by the reseeding procedure of the PIGS samplers is considerably reduced owing to the faster underlying dynamics. However, bias remains discernible for this observable. We perform this analysis not only to reveal bias but also to point out that fundamental characteristics are preserved despite the bias.

The preservation of the temperature-dependent balance between chain entropy, helix stability, and solvent properties is illustrated by an observable that depends less on the population of the straight α-helix, viz., histograms of residue–residue contacts. Fig. 4(b) shows that, unlike the radius of gyration histograms, contact number distributions are similar for all protocols at either temperature. At 290 K, the differences are generally marginal. At 250 K, the peaks are shifted slightly to the left for CLD and REX with respect to the PIGS algorithms, and the variance is lower, which hints at the reduced diversity of the sampled ensemble. This suggests that the different topologies sampled by PIGS replace the local, $i \rightarrow i + 4$ contacts of α-helices with other, most likely nonlocal contacts. This suggestion is confirmed by Fig. 4(c). At 250 K, the difference in contact probabilities between CLD and PIGS reveals a dramatic shift from contact patterns indicative of helix-rich states (e.g., cartoons 1 and 2 in Fig. 2) to nonhelical states. This is true for both ϕ- and r-PIGS. The greater frequency of distal contacts for the PIGS data reflects the presence of disordered, globular conformations. These states are long-lived, which is vaguely suggested by Fig. 3(d), and which we confirm in Section 3.2.5. Finally, Fig. 4(d) shows the same analysis as panel (c) for 290 K. Consistent with panels (a) and (b), the magnitude of the systematic difference in equilibrium distributions introduced by the PIGS reseeding heuristic is considerably reduced at the higher temperature. This is because interconversion among states is faster (see Fig. S6).

It is important for us to dispel concerns about the interpretability of the thermodynamic ensemble generated by the base sampler when coupled with the reseeding heuristic. Table 1 indicates that we do not observe major differences in thermodynamic quantities. Specifically, the average kinetic and potential energies and associated fluctuations are generally within error. The invariance of kinetic energies at both temperatures indicates that integrator stability is not compromised by the reseeding in PIGS or the swaps in REX. The quantitative significance of this is also inferred by comparison to data at adjacent temperatures for the REX scheme. For potential energies, there are significant differences at 250 K between REX and PIGS illustrated also by the greater overlap in the average potential energies between PIGS data at 250 K and REX data at 256 K. This is not surprising due to the depletion of energetically favorable conformations rich in α-helix content. The generally larger fluctuations in potential energy for PIGS reflect the greater diversity of the conformations sampled by PIGS. We note that these implicit solvent simulations in the canonical ensemble can exhibit non-Gaussian tails of the potential energy spectrum. For these, it is difficult to assess what is correct, i.e., variance estimates may change significantly even by simply increasing the length of CLD simulations. While the trends appear similar at 290 K compared to 250 K, no significant differences between samplers are observed.

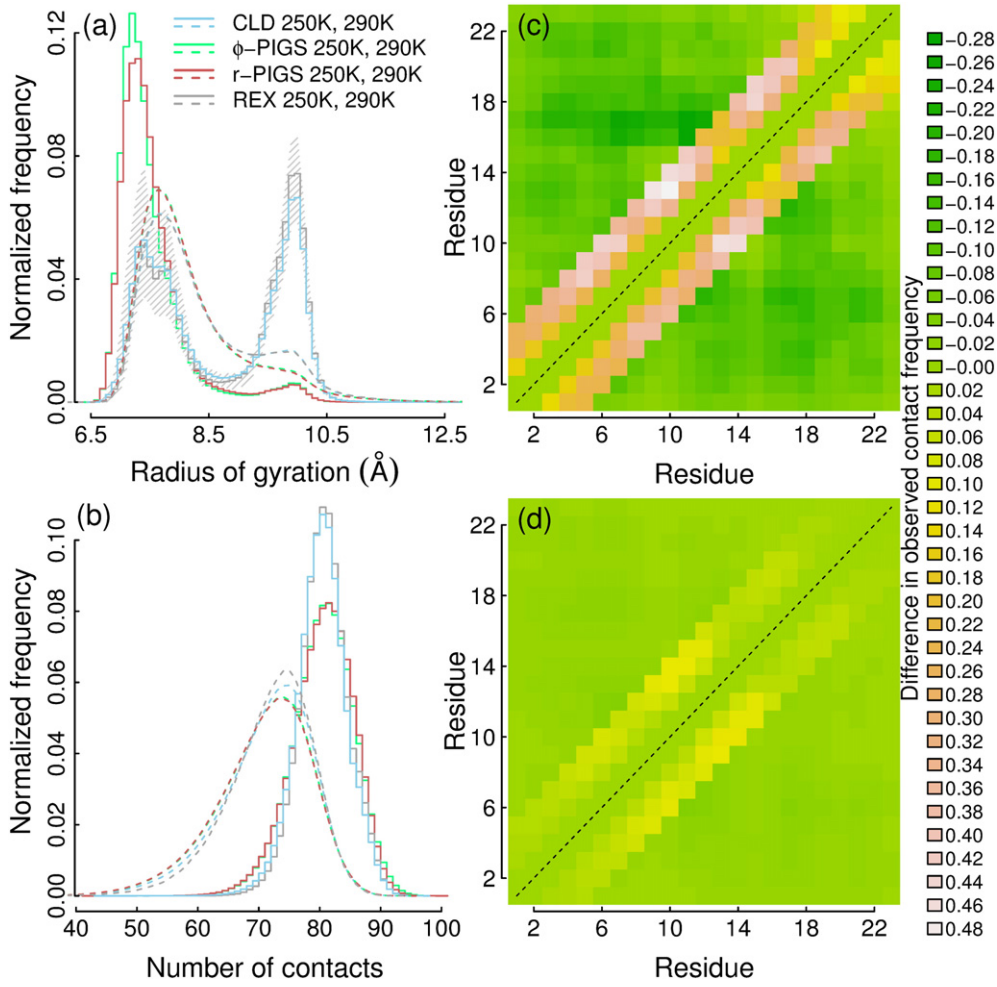

Fig. 4. Comparison of equilibrium statistics for radii of gyration and contacts. (a) Histograms of radii of gyration for the four different data sets at 250 K (solid steps) and 290 K (dashed, continuous lines). The bin width for the radius of gyration is 0.1 Å. Min/max ranges across 16 blocks are indicated for the REX data by the shaded region. (b) Histograms for the number of contacts at 250 K (solid steps) and 290 K (dashed, continuous lines). A residue–residue contact was counted if the distance between any two atoms of the two residues in question was below 5.5 Å. The same legend applies as in (a). (c) Difference in the average contact maps for CLD and PIGS at 250 K. The upper left half-matrix refers to ϕ-PIGS, and the lower right half-matrix plots the same data for r-PIGS. (d) The same as (c) for 290 K.

In summary, the bias introduced by PIGS primarily influences the equilibrium statistics of physical observables that are sensitive to conformation as discussed for Fig. 4. As mentioned before, we do not address approaches toward snapshot-based reweighting or rigorous MSM construction for these data in this contribution. Figs. S7(a)–(c) highlight the population bias at the level of the states shown in Fig. 2. However, Fig. 2 also suggests that important state connectivity information is preserved by PIGS, and this is analyzed next.

3.2.4. Transition path analyses

Describing transition pathways has always been a major interest and challenge of protein science as it can suggest mechanisms for, among others, folding, enzymatic activity, signal transduction, or ligand binding. Since PIGS does not bias the PES, and since we discard any transition path that is not continuous, i.e., that has been interrupted by a reseeding event, we expect differences among transition paths relative to CLD to be caused predominantly by a lack of proper sampling. Figs. S8 and 3(d) emphasize that the actual number of reseeding events is small.

Table 1
Average kinetic and potential energies and their standard deviations for the different samplers.

Sampler	Mean kinetic energy (kcal/mol)	Standard deviation of the kinetic energy (kcal/mol)	Mean potential energy (kcal/mol)	Standard deviation of the potential energy (kcal/mol)
CLD at 250 K	170.17 ± 0.05	9.22 ± 0.02	− 4562.4 ± 1.8	9.42 ± 0.33
ϕ-PIGS at 250 K	170.17 ± 0.05	9.22 ± 0.02	− 4559.9 ± 2.5	9.82 ± 0.48
r-PIGS at 250 K	170.17 ± 0.04	9.22 ± 0.02	− 4559.9 ± 2.5	9.93 ± 0.48
REX at 250 K	170.17 ± 0.05	9.21 ± 0.02	− 4563.4 ± 0.5	9.30 ± 0.11
REX at 256 K	174.24 ± 0.05	9.44 ± 0.02	− 4558.9 ± 0.4	9.66 ± 0.10
CLD at 290 K	197.39 ± 0.05	10.69 ± 0.02	− 4527.9 ± 1.7	12.81 ± 0.35
ϕ-PIGS at 290 K	197.39 ± 0.06	10.69 ± 0.02	− 4526.7 ± 1.5	12.81 ± 0.31
r-PIGS at 290 K	197.40 ± 0.06	10.69 ± 0.02	− 4526.6 ± 1.4	12.85 ± 0.32
REX at 290 K	197.38 ± 0.05	10.69 ± 0.02	− 4528.7 ± 0.7	12.44 ± 0.23
REX at 297 K	202.16 ± 0.05	10.95 ± 0.02	− 4521.3 ± 0.8	12.99 ± 0.20

Statistical errors are reported for each quantity and were obtained by averaging over the mean values for different numbers of blocks per temperature (256 for CLD and PIGS; 32 for REX).

At 250 K, the average lengths of uninterrupted simulation stretches are 18.9 and 11.9 ns for ϕ- and r-PIGS, respectively, which is more than two orders of magnitude larger than the values seen for REX despite f_p being identical. Moreover, the average reseeding frequency is notably higher at the beginning of a PIGS run than toward the end owing to the identical starting condition.

When speaking of transition pathways, it is practically inevitable to avoid definitions of states [38]. Defining states may involve somewhat arbitrary decisions. For Fig. 2 we used the approach described in Sections 2.3.2 and 2.3.3 to identify and delineate a set of states, and we analyze only direct transitions between any given pair of states, which is akin to the local equilibration idea in MSM construction. Fig. 2 and Movie S1 display only such direct transitions as links in a network, and suggest a fundamental similarity between CLD and PIGS for the common sampling domain. Conversely, REX results differ in terms of relative, statistical weights of transitions. This can be expected due to a bias toward shorter time scales, which results from discarding stretches of trajectory interrupted by swaps. However, the analysis may be affected by further subtleties of the REX protocol.

Table 2 provides quantitative evidence for this qualitative result. We find that the statistical weight of prominent transitions differs significantly for the REX protocol with respect to all other samplers. With the exception of the transition between state 2 and 30, the ranks for REX differ considerably. If this were solely a consequence of losing data on longer paths, one would not expect certain transitions to be sampled much more in absolute numbers, e.g., from state 2 to 72. These data hint that REX may alter state connectivity in more fundamental ways, for example by emphasizing transitions that are naturally shorter, yet are not particularly probable in unperturbed trajectories. In Table 2, we also provide further statistics for transition path times. Generally, the PIGS data sets seem to contain more outliers toward long transition times than CLD, which is indicated by greater values for both averages and standard deviations. This is noticeable in particular for the transition between state 2 and 72. Conversely, the average transition path times are shorter for REX when compared to CLD, which is consistent with the reasoning above.

Panels (a)–(c) of Fig. 5 reveal that three prominent transitions yield cumulative distribution functions for transition path times that show little difference between CLD and either PIGS scheme. For the transition between state 61 and 72, panel (a), REX samples only a few events that appear to follow a similar distribution. The transition between state 36 and 60, panel (c), is depleted entirely for REX despite these transition path times being shortest among the ones investigated. This is further evidence toward the notion that REX fundamentally alters network properties. The expected results for REX, viz., oversampling of short transition paths, are seen for the remaining transitions, which both involve the straight α-helix as one of the two states. In Fig. 5(d), the transition between state 2 and 72, which is explored often by REX, yields distributions that appear to differ quantitatively for all 4 cases. The corresponding statistics for the transition path times in Table 2 confirm this heterogeneity.

Besides insufficient sampling, we can think of two main sources for the type of heterogeneity seen in Fig. 5(d). First, there might be different types of paths available and their likelihoods may depend on the sampling protocol. Second, the definitions of clusters as states could mask heterogeneous subpopulations, i.e., a given state as sampled by PIGS may differ subtly from the set of snapshots classified to be the same state but sampled by CLD.

In order to address these issues, we performed principal component analyses on the partitions of data representing the two states and all direct transitions connecting them (see Section 2.3.3 for details). In Fig. 6, we can see that a projection of the data onto the two principal components accounting for the largest total variance gives an informative plot. The two end states are connected by a region of continuous density. For CLD, panel (a), this appears to be the main area of probability flow and we defined an *ad hoc* separator as indicated by the rectangle. Partially overlapping regions of high density outside of this rectangle are seen for both PIGS data sets. We stress that the underlying histogram effectively weighs paths by their lengths in time. This suggests that some of the direct transitions involve extensive dwell times in areas not characterized by us as states. In fact, we can count the number of paths passing through the rectangle, and these statistics are reported in Fig. 6. Indeed, the majority of direct transitions for panels (a)–(c) pass through the separator. The drastic exception is REX for which 70/78 transitions appear to skip directly from one state to the other.

The corresponding analyses for the other three cases in Table 2 and Fig. 5 is provided as Figs. S9–S11. We note that in all cases the variance encapsulated by the first two components is sufficiently large to justify this projection approach (see Table 2). Similarly, the distributions of the points corresponding to the end states themselves are always similar for all four data sets indicating that ambiguities in state annotation do not explain heterogeneity. In summary, there is little reason to believe that these PIGS simulations bias transition pathways for the FS peptide. This is consistent with the low average reseeding rate shown in Fig. S8. Conversely, the REX data must be treated with caution. Obtaining converged data on transition paths from equilibrium simulations is difficult for complex systems, and insufficient sampling may mask systematic differences.

3.2.5. Dependence on starting structure

As a final point of analysis, we wish to address the following, possible concerns. First, one may wonder whether those clusters

Table 2
Transition path statistics.

Transition	Sampler	Number of events	Rank	Mean time (ns)	Standard deviation (ns)	Variance in 1st and 2nd PC
2–30	CLD	108	11	0.47	0.37	75.6%
	ϕ-PIGS	27	18	0.84	1.23	
	r-PIGS	22	22	0.38	0.32	
	REX	28	12	0.27	0.25	
2–72	CLD	27	39	0.86	0.63	74.2%
	ϕ-PIGS	12	40	5.08	5.80	
	r-PIGS	13	31	1.83	2.17	
	REX	78	6	0.33	0.31	
36–60	CLD	49	20	0.23	0.52	55.1%
	ϕ-PIGS	13	36	0.30	0.66	
	r-PIGS	47	8	0.25	0.48	
	REX	0	–	–	–	
61–72	CLD	163	9	0.33	0.41	46.9%
	ϕ-PIGS	77	3	0.31	0.32	
	r-PIGS	77	3	0.37	0.44	
	REX	7	44	0.30	0.26	

All data were combined for both possible directions. The 'Rank' column indicates the rank of the transition in question in lists sorted individually for each sampler by numbers of events. The last column reports the percentage of the standard deviation of the underlying data that is encapsulated by the first two principal components (see notes on PCA in Section 2.3.3).

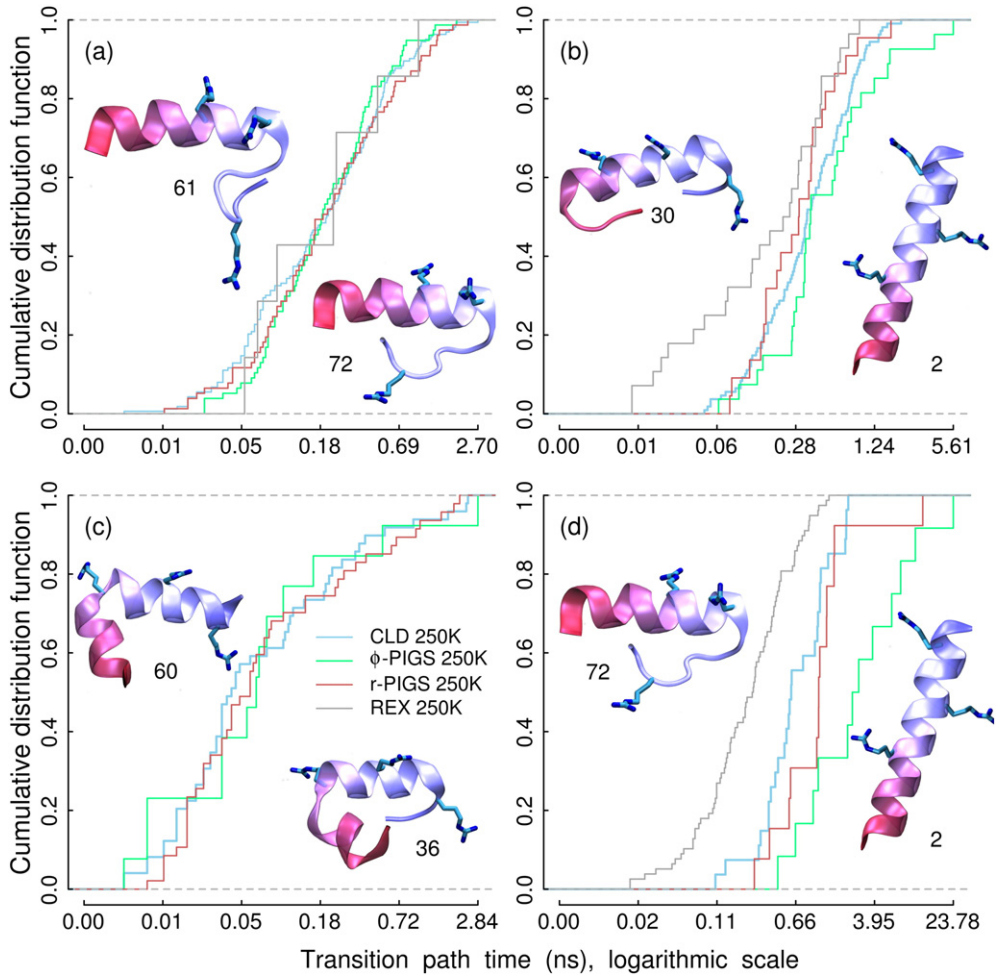

Fig. 5. Cumulative probability densities for transition path times for 4 selected transitions. Each panel reports two representative cartoons (central snapshots) for the clusters used to define the states for the transition in question. The legend in (c) applies to all panels. (a) Cumulative probability density for the times it takes to reach one state from the other irrespective of direction. These are the data for states 61 and 72. (b) The same as (a) for transitions between state 2 and 30. (c) The same as (a) for transitions between state 60 and 36. (d) The same as (a) for transitions between state 2 and 72.

sampled exclusively by PIGS as seen in Figs. 2 and S7 actually correspond to proper metastable states. Second, we want to know how much the resultant distributions depend on the initial state for all 3 samplers. Both questions can be addressed by repeating simulations with a different starting configuration. Here, we restricted ourselves to a temperature of 250 K and r-PIGS, REX, and CS. The starting configuration is taken as a collapsed globule sampled uniquely by ϕ-PIGS in the original data set (similar to cartoon (5) in Fig. 2).

Fig. 7 shows an analysis comparable to that in Fig. 3 for these data, i.e., we quantify exploration rate and memory loss. We suppose that the starting structure is a state that is metastable and kinetically distant from the α-helical domain. We do not know its correct statistical weight due to CLD and REX not having found this structure. Fig. 7(a) plots, for all replicas, the RMSD to the starting structure as a function of time for CLD. Clearly, the state is metastable in CLD and the life time distribution appears to have long tails as indicated by one of the replicas never escaping from it. This confirms our aforementioned assumptions. The net sampling weight should be low based on these data. For r-PIGS, which did not encounter this structure before either, we obtain considerably faster escape. After ~50 ns, the picture is analogous to that in Fig. 3(d), i.e., on average only a single replica continues to explore this basin at a given point in time. This highlights again the efficacy of the reseeding heuristic in avoiding overlap of sampling domains.

Panel (c) of Fig. 7 plots the early time course of the maximum and minimum deviations from the globular starting structure and the α-helix, respectively. The first replica to sample the straight helix in CLD does so after ~25 ns. The corresponding number for both r-PIGS and REX is below 10 ns. Similar conclusions hold for the maximum deviation from the starting conformation. The performance of REX is remarkable as it relies on considerably less data. The stochasticity of escape suggested by panel (a) seems to be circumvented. We conjecture that the starting structure, due to its slightly inferior energetic stability, is rapidly and systematically swapped toward higher temperatures. The collapse constraint is substantially weakened, and helix-rich states form quickly. These are then swapped back to low temperatures. Essentially, the high temperature replicas provide kinetic shortcuts as intended by the protocol. However, this comes at the cost of not sampling appropriate pathways.

Fig. S12 provides an overview that is analogous to Fig. 2. We can see that REX hardly samples unique states. In fact, the sampling domain is as restricted as that when starting from the α-helix. Conversely, there is a large increase in states sampled uniquely by CLD (see also Fig. S13). Probably, these are encountered "along the way," i.e., they result from undirected exploration of a larger fraction of phase space than when starting from the helix. The picture for r-PIGS continues to be the richest. These observations are quantified in Fig. 7(d). Plotting the data in direct comparison to those in Fig. 3(a), we can confirm the better coverage of CLD when starting from the globule. The indifference of REX is also visible straight away. Lastly, the exploration rate for r-PIGS, while superior throughout, lags behind when comparing to the run starting from the helix. We believe that this result simply reflects that PIGS is

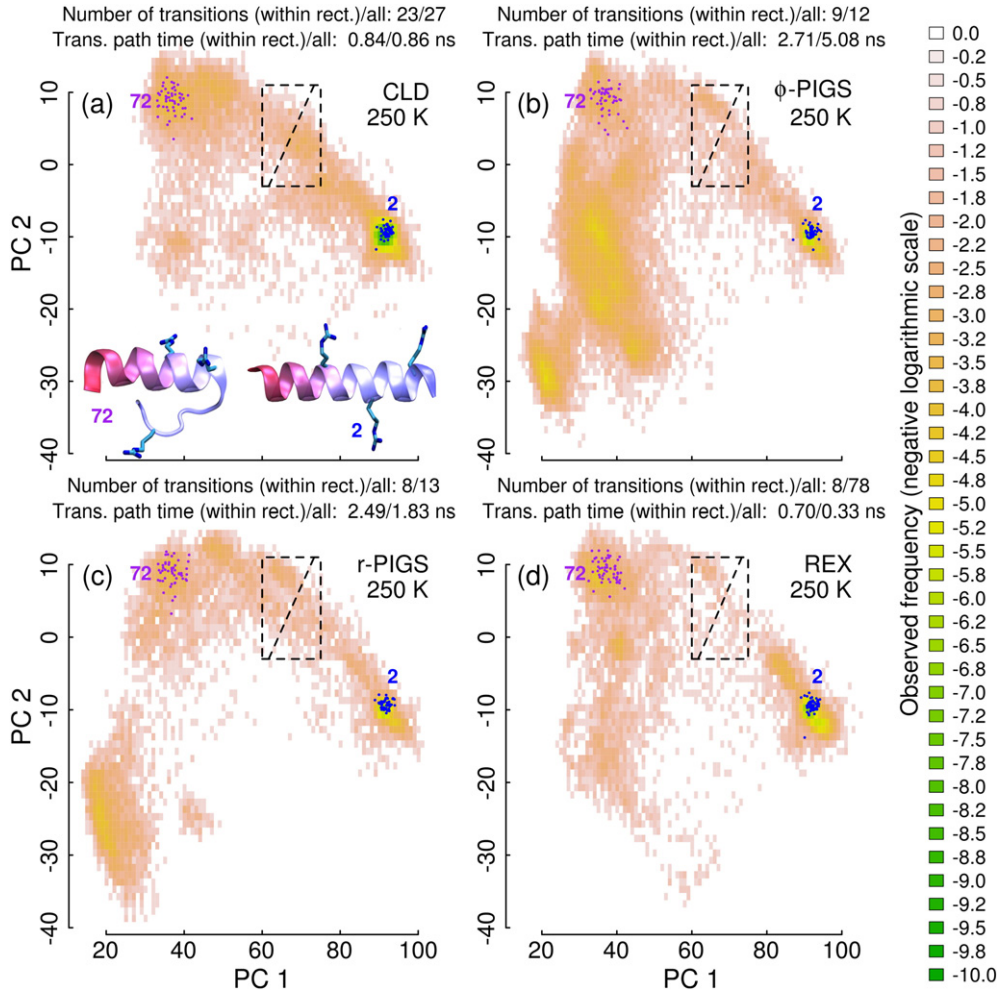

Fig. 6. Projection of transition paths between state 2 and 72 onto the space defined by the first two principal components (see Section 2.3.3). The negative logarithms of the individual histograms for the four different protocols are plotted as color maps. Bins with no counts are shown in white, and the bin width is equal to 1 Å for both components. Cartoons show representative structures for the two states connected by the transition. A total number of 50 points belonging explicitly to states 2 and 72 as sampled by the respective protocol are shown as blue and purple dots, respectively. The rectangle in each panel is identical and highlights an area defined as a separator region for CLD. Relevant statistics are also reported. (a) Histogram and statistics for the subset of the transition path data sampled by CLD. (b) The same as (a) for ϕ-PIGS. (c) The same as (a) for r-PIGS. (d) The same as (a) for REX.

not designed to escape from narrow minima surrounded by large, energetic barriers. This also emerges from comparison of Figs. 7(b) and 3(d).

We conclude the results by pointing out that the structural ensembles obtained after discarding the first half of the data exhibit a high level of similarity between runs started from different initial conformations. Analyzed in terms of the radius of gyration, Fig. S14 shows that this holds for all samplers. Small deviations in the expected direction, e.g., less population of the straight helix when starting from the globule, are seen for CLD and REX. In view of Figs. 7 and S12, it may be surprising that REX retains bias. The thermodynamic bias seen for r-PIGS in Fig. 4(a) is quantitatively preserved in Fig. S14(b). This level of convergence suggests that most time scales that we probe in the simulations are faster than the 312 ns of simulation time per replica.

4. Discussion and conclusions

In this contribution, we introduce a reseeding heuristic for molecular simulations. For two different systems, the heuristic is demonstrated to increase the rate of exploration (Figs. 1, 3, 7, S3, S4, and S6) leading to better coverage of phase space (Figs. 2 and S12). This is despite the effective reseeding rate being surprisingly low (Figs. 3(d), 7(b), and S8). Consequently, the algorithm is able to preserve useful information about pathways of interconversion (Figs. 2, 5, 6, and S9–S12). If time scales that are orders of magnitude slower than f_p dominate the system dynamics (data at 250 K), even a small number of reseeding events will accumulate considerable thermodynamic bias (Figs. 4 and S14). The bias is reduced considerably for the FS peptide at 290 K (Fig. 4). As demonstrated by Figs. 3(a) and S6(a), this corresponds to a case where either REX or PIGS provides no benefit over CLD. Since CLD preserves equilibrium sampling weights and pathway information maximally, the use of PIGS and in particular of REX is wasteful (see Fig. S15 for the corresponding state network). This manuscript is not concerned with the removal of the thermodynamic bias introduced by PIGS. We note that the bias is not of a unique type, i.e., it can be rephrased as the task of extracting equilibrium information from short trajectories of heterogeneous lengths whose starting points are not drawn with the correct Boltzmann weights. Strategies for this problem are available [45,68,69].

We believe that the favorable properties of our approach can be summarized as follows.

- It is unsupervised beyond specifying fixed parameters. In this context, the demonstrated robustness with respect to parameter choices is important (see Fig. S4 and the comparison of ϕ-PIGS and r-PIGS for the FS peptide).
- PIGS is parallel in a way that provides synergistic benefits (see Figs. 1 and S3). This means that the width of available computing resources

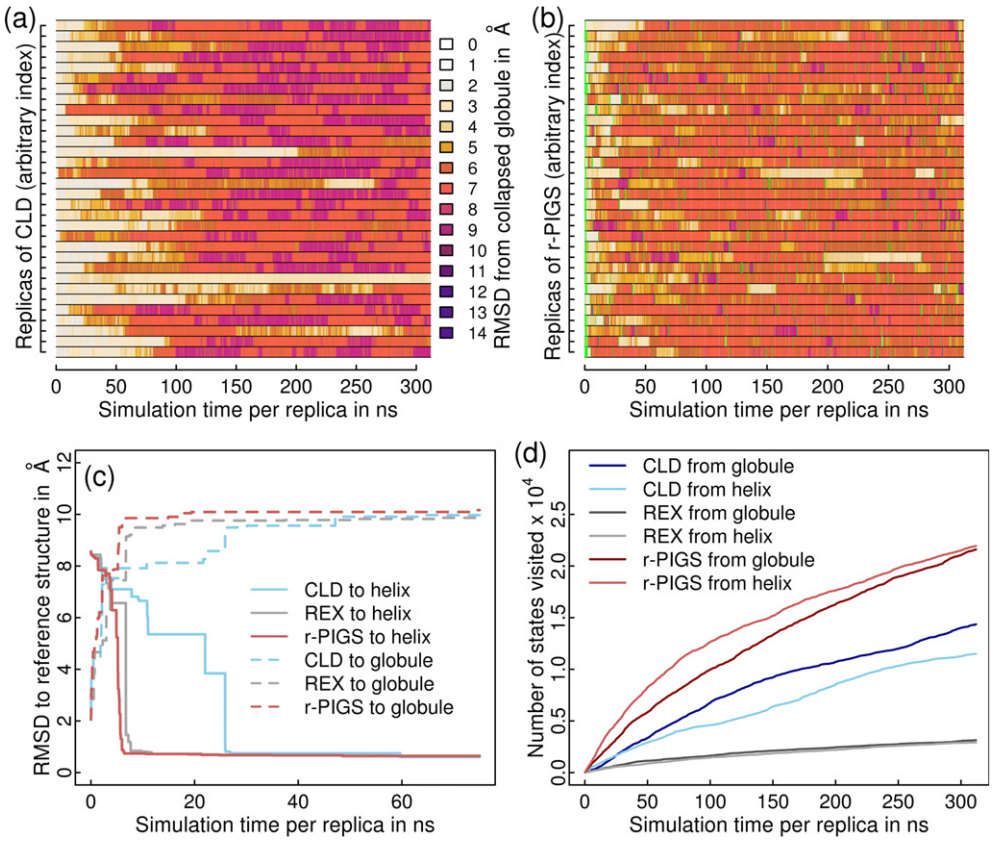

Fig. 7. Dependence on starting conformation at 250 K. (a) We plot the RMSD values as a function of time with respect to the starting structure, viz., a globular state free of secondary structure. The time resolution is 30 ps. Data are shown for all 32 replicas for CLD with the color code indicated on the right. (b) The same as (a) for r-PIGS. The same color legend applies. In addition, the actual reseeding times are indicated as green, vertical bars as in Fig. 3(d). (c) We show as solid lines the minimum RMSD to the straight α-helix as a function of time for CLD, r-PIGS, and REX. These data are cumulative, i.e., they correspond to the overall minimum across all replicas or runs (REX) up to the indicated time. The dashed lines represent analogous values for the maximum RMSD to the starting structure. (d) Exploration rates for CLD, REX, and r-PIGS. The plot includes the corresponding data from Fig. 3(b) to facilitate comparison.

can be meaningfully exploited. In this regard, we have also performed a test for the FS peptide (see Fig. S16) confirming the results for the model system. Scalability with the number of replicas is critical for this property to hold, and we establish it conceptually.

- It does not differ in I/O load from conventional sampling, and all communication requirements are restricted to the reseeding points as in REX.
- Despite relying on few reseeding events, PIGS speeds up exploration rates for the systems under study. The data in Section 3.2.5 suggest that this advantage is slightly reduced for energetic barriers. This is the opposite behavior compared to REX [30], which makes the methods somewhat complementary. Pathway information is preserved, which is directly useful for MSM construction and the extraction of rates.
- We believe that the heuristic is both conceptually and practically useful for a wide spectrum of molecular simulations. We expect the choice of representation, although not tested directly here, to be of value in targeting specific research questions for complex systems.

As such, we hope that PIGS will be a useful addition to the toolkit available to molecular dynamics practitioners. We are currently investigating the possibility of post-processing the trajectories sampled by PIGS to extract realistic estimates of both equilibrium statistics and long time scale dynamics without having to perform additional simulations. At present, the thermodynamic bias is a caveat inasmuch as one accepts the CLD or REX results as free of initial state bias (see Fig. S14). As discussed in Section 3.2.3, this question is ultimately linked to diffuse notions of recurrence that have to pragmatically replace the fundamental idea of the ergodic hypothesis. For the FS peptide, the fact that the sampling overlap between data sets is low for metastable states that are far from the main sampling domain indicates that simulation convergence holds at most for low-resolution projections such as the radius of gyration or net helicity. The dilemma of having to restrict conclusions to an actual sampling domain [66], whether discovered by unbiased simulations or defined by reaction coordinates, is a profound one. Indeed, for systems without analytical results, we can at most falsify an assertion of convergence [70].

In simulations of folded and/or assembled (macro)molecules, starting structures cannot be made random. This increases the chance of masking errors because one of the most stringent and useful falsification tests of convergence is taken away, i.e., assessing the similarity of results when using truly independent initial configurations [65,71]. Unquantifiable errors of this type lead to ambiguity, which is arguably inherent to most applied molecular dynamics studies. It is desirable to reduce this ambiguity, and this motivates the development and application of methods capable of discovering new metastable states in unsupervised fashion. PIGS is one such algorithm.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbagen.2014.08.013.

Acknowledgment

AV acknowledges support from the Holcim Foundation through a habilitation fellowship. AC was partially supported by a grant from the Swiss National Science Foundation. The authors acknowledge infrastructural support from the University of Zurich and its Schrödinger supercomputer. We thank the editors of this special issue for the invitation to contribute.

References

[1] W.F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D.P. Geerke, A. Glättli, P.H. Hünenberger, M.A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N.F.A. Van Der Vegt, H.B. Yu, Biomolecular modeling: goals, problems, perspectives, Angew. Chem. Int. Ed. Engl. 45 (2006) 4064–4092.

[2] T. Schlick, R. Collepardo-Guevara, L.R. Halvorsen, S. Jung, X. Xiao, Biomolecular modeling and simulation: a field coming of age, Q. Rev. Biophys. 44 (2011) 191–228.

[3] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–520.

[4] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Biol. 9 (2002) 646–652.

[5] S.D. Bond, B.J. Leimkuhler, Molecular dynamics and the accuracy of numerically computed averages, Acta Num. 16 (2007) 1–65.

[6] R. Elber, Long-timescale simulation methods, Curr. Opin. Struct. Biol. 15 (2005) 151–156.

[7] H. Lei, Y. Duan, Improved sampling methods for molecular simulation, Curr. Opin. Struct. Biol. 17 (2007) 187–191.

[8] D.M. Zuckerman, Equilibrium sampling in biomolecular simulations, Annu. Rev. Biophys. 40 (2011) 41–62.

[9] K. Klenin, B. Strodel, D.J. Wales, W. Wenzel, Modelling proteins: conformational sampling and reconstruction of folding kinetics, Biochim. Biophys. Acta 1814 (2011) 977–1000.

[10] T. Schlick, E. Barth, M. Mandziuk, Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation, Annu. Rev. Biophys. Biomol. Struct. 26 (1999) 181–222.

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

[12] A. Jain, N. Vaidehi, G. Rodriguez, A fast recursive algorithm for molecular dynamics simulation, J. Comput. Phys. 106 (1993) 258–268.

[13] M.E. Tuckerman, G.J. Martyna, B.J. Berne, Molecular dynamics algorithm for condensed systems with multiple time scales, J. Chem. Phys. 93 (1990) 1287–1291.

[14] R. Pomès, J.A. McCammon, Mass and step length optimization for the calculation of equilibrium properties by molecular dynamics simulation, Chem. Phys. Lett. 166 (1990) 425–428.

[15] R. Das, D. Baker, Macromolecular modeling with Rosetta, Annu. Rev. Biochem. 77 (2008) 363–382.

[16] E. Brini, E.A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, N.F.A. van der Vegt, Systematic coarse-graining methods for soft matter simulations—a review, Soft Matter 9 (2013) 2108–2119.

[17] M.G. Saunders, G.A. Voth, Coarse-graining methods for computational biology, Annu. Rev. Biophys. 42 (2013) 73–93.

[18] H. Li, G. Li, B.A. Berg, W. Yang, Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces, J. Chem. Phys. 125 (2006) 144902.

[19] E. Lyman, F.M. Ytreberg, D.M. Zuckerman, Resolution exchange simulation, Phys. Rev. Lett. 96 (2006) 028105.

[20] G.M. Torrie, J.P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling, J. Comput. Phys. 23 (1977) 187–199.

[21] A. Laio, M. Parrinello, Escaping free-energy minima, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 12562–12566.

[22] S. Singh, M. Chopra, J.J. de Pablo, Density of states-based molecular simulations, Annu. Rev. Chem. Biomol. Eng. 3 (2012) 369–394.

[23] S. Park, K. Schulten, Calculating potentials of mean force from steered molecular dynamics simulations, J. Chem. Phys. 120 (2004) 5946–5961.

[24] B. Roux, The calculation of the potential of mean force using computer simulations, Comp. Physiol. Commut. 91 (1995) 275–282.

[25] C. Neale, T. Rodinger, R. Pomès, Equilibrium exchange enhances the convergence rate of umbrella sampling, Chem. Phys. Lett. 460 (2008) 375–381.

[26] W. Sinko, Y. Miao, C.A.F. de Oliveira, J.A. McCammon, Population based reweighting of scaled molecular dynamics, J. Phys. Chem. B 117 (2013) 12759–12768.

[27] D. Hamelberg, J. Mongan, J.A. McCammon, Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules, J. Chem. Phys. 120 (2004) 11919–11929.

[28] R.H. Swendsen, J.S. Wang, Replica Monte Carlo simulation of spin glasses, Phys. Rev. Lett. 57 (1986) 2607–2609.

[29] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141–151.

[30] H. Nymeyer, How efficient is replica exchange molecular dynamics? An analytic approach, J. Chem. Theory Comput. 4 (2008) 626–636.

[31] K. Ostermeir, M. Zacharias, Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins, Biochim. Biophys. Acta 1834 (2013) 847–853.

[32] D.M. Zuckerman, E. Lyman, A second look at canonical sampling of biomolecules using replica exchange simulation, J. Chem. Theory Comput. 2 (2006) 1200–1202.

[33] E. Rosta, G. Hummer, Error and efficiency of replica exchange molecular dynamics simulations, J. Chem. Phys. 131 (2009) 165102.

[34] W. Zhang, J. Chen, Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions, J. Chem. Theory Comput. 9 (2013) 2849–2856.

[35] W. Zheng, M. Andrec, E. Gallicchio, R.M. Levy, Simulating replica exchange simulations of protein folding with a kinetic network model, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 15340–15345.

[36] S. Muff, A. Caflisch, ETNA: equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations, J. Phys. Chem. B 113 (2009) 3218–3226.

[37] D.A. Beck, G.W. White, V. Daggett, Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations, J. Struct. Biol. 157 (2007) 514–523.

[38] P.G. Bolhuis, C. Dellago, D. Chandler, P.L. Geissler, Transition path sampling: throwing ropes over rough mountain passes, in the dark, Annu. Rev. Phys. Chem. 53 (2002) 291–318.

[39] G. Henkelman, B.P. Uberuaga, H. Jónsson, A climbing image nudged elastic band method for finding saddle points and minimum energy paths, J. Chem. Phys. 113 (2000) 9901–9904.

[40] W. E., W. Ren, E. Vanden-Eijnden, Finite temperature string method for the study of rare events, J. Phys. Chem. B (2005) 6688–6693.

[41] A.K. Faradjian, R. Elber, Computing time scales from reaction coordinates by milestoning, J. Chem. Phys. 120 (2004) 10880–10889.

[42] L. Maragliano, A. Fischer, E. Vanden-Eijnden, G. Ciccotti, String method in collective variables: minimum free energy paths and isocommittor surfaces, J. Chem. Phys. 125 (2006) 024106.

[43] G.R. Bowman, K.A. Beauchamp, G. Boxer, V.S. Pande, Progress and challenges in the automated construction of Markov state models for full protein systems, J. Chem. Phys. 131 (2009) 124101.

[44] J.D. Chodera, F. Noé, Markov state models of biomolecular conformational dynamics, Curr. Opin. Struct. Biol. 25 (2014) 135–144.

[45] X. Huang, G.R. Bowman, S. Bacallado, V.S. Pande, Rapid equilibrium sampling initiated from nonequilibrium data, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19765–19769.

[46] N. Singhal, C.D. Snow, V.S. Pande, Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin, J. Chem. Phys. 121 (2004) 415–425.

[47] V.S. Pande, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, E.J. Sorin, B. Zagrovic, Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing, Biopolymers 68 (2002) 91–109.

[48] T. Zhou, A. Caflisch, Free energy guided sampling, J. Chem. Theory Comput. 8 (2012) 2134–2140.

[49] W. Zheng, M.A. Rohrdanz, C. Clementi, Rapid exploration of configuration space with diffusion-map-directed molecular dynamics, J. Phys. Chem. B 117 (2013) 12769–12776.

[50] A. Dickson, C.L. Brooks III, WExplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm, J. Phys. Chem. B 118 (2014) 3532–3542.

[51] C. Schütte, A. Fischer, W. Huisinga, P. Deuflhard, A direct approach to conformational dynamics based on hybrid Monte Carlo, J. Comput. Phys. 151 (1999) 146–168.

[52] E. Vanden-Eijnden, M. Venturoli, Markovian milestoning with Voronoi tesselations, J. Chem. Phys. 130 (2009) 194101.

[53] N. Blöchliger, A. Vitalis, A. Caflisch, A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems, Comp. Physiol. Commut. 184 (2013) 2446–2453.

[54] D.J. Lockhart, P.S. Kim, Internal Stark effect measurement of the electric field at the amino terminus of an α helix, Science 257 (1992) 947–951.

[55] S.D. Bond, B.B. Laird, B.J. Leimkuhler, The Nosé–Poincaré method for constant temperature molecular dynamics, J. Comput. Phys. 151 (1999) 114–134.

[56] A. Vitalis, A. Caflisch, Efficient construction of mesostate networks from molecular dynamics trajectories, J. Chem. Theory Comput. 8 (2012) 1108–1120.

[57] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38.

[58] A. Vitalis, R.V. Pappu, ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions, J. Comput. Chem. 30 (2009) 673–699.

[59] A. Vitalis, A. Caflisch, 50 Years of Lifson–Roig models: application to molecular simulation data, J. Chem. Theory Comput. 8 (2012) 363–373.

[60] G.A. Kaminski, R.A. Friesner, J. Tirado-Rives, W.L. Jorgensen, Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides, J. Phys. Chem. B 105 (2001) 6474–6487.

[61] R.D. Skeel, J.A. Izaguirre, An impulse integrator for Langevin dynamics, Mol. Phys. 100 (2002) 3885–3891.

[62] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Inf. Process. Lett. 31 (1989) 7–15.

[63] T. Ichiye, M. Karplus, Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations, Proteins Struct. Funct. Bioinforma. 11 (1991) 205–217.

[64] Y. Sugita, A. Kitao, Y. Okamoto, Multidimensional replica-exchange method for free-energy calculations, J. Chem. Phys. 113 (2000) 6042–6051.

[65] A. Vitalis, R.V. Pappu, Methods for Monte Carlo simulations of biomacromolecules, Annu. Rep. Comput. Chem. 5 (2009) 49–76.

[66] S. Asmussen, P.W. Glynn, H. Thorisson, Stationarity detection in the initial transient problem, ACM Trans. Model. Comput. Simul. 2 (1992) 130–157.

[67] A. Grossfield, D.M. Zuckerman, Quantifying uncertainty and sampling quality in biomolecular simulations, Annu. Rep. Comput. Chem. 5 (2009) 23–48.

[68] C. Komalapriya, M. Thiel, M.C. Romano, N. Marwan, U. Schwarz, J. Kurths, Reconstruction of a system's dynamics from short trajectories, Phys. Rev. E. 78 (2008) 066217.

[69] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19011–19016.

[70] A. Grossfield, S.E. Feller, M.C. Pitman, Convergence of molecular dynamics simulations of membrane proteins, Proteins Struct. Funct. Bioinforma. 67 (2007) 31–40.

[71] W.F. van Gunsteren, A.E. Mark, Validation of molecular dynamics simulation, J. Chem. Phys. 108 (1998) 6109–6116.