SUPPLEMENTARY INFORMATION TO

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

Marco Bacci, Andreas Vitalis,* and Amedeo Caflisch*

University of Zurich

Department of Biochemistry

Winterthurerstrasse 190, CH-8057 Zurich

*To whom correspondence should be addressed: a.vitalis@bioc.uzh.ch (AV), caflisch@bioc.uzh.ch (AC)

Itemized Overview

Supplementary Methods

The temperature series used for the replica exchange (REX) simulations of the FS peptide was: 230, 234, 239, 244, 250, 256, 263, 270, 277, 284, 290, 297, 306, 315, 325 and 340 K (see 2.2.2 in the main text). No optimization was attempted, and the average acceptance rate for neighbor swaps across all replicas was 43.7%.

Table S1 lists the interatomic distances used to represent the FS peptide system for the purposes of progress index construction for r-PIGS (see 2.2.2), clustering (see 2.3.2), and principal component analysis (PCA, see 2.3.3). This implies use as the underlying metric for the assessment of similarity. For progress index construction for φ-PIGS (see 2.2.2), we used all backbone φ, ψ, and ω dihedral angles and side chain torsional angles of the arginine residues, from χ_1 to χ_4, which makes 76 dihedral angles in total.

Table S1: *Set of 145 interatomic distances used to perform pairwise distance evaluations.*

A1 backbone O	Backbone N atoms of: A5, A7, R9, A11, A13, A15, A17, R19, A21
A2 backbone O	Backbone N atoms of: A6, A8, A10, A12, R14, A16, A18, A20
A3 backbone O	Backbone N atoms of: A7, A8, R9, A11, A12, R14, A15, A17, A18, A20, A21
A4 backbone O	Backbone N atoms of: A8, R9, A10, A12, A13, A15, A16, A18, R19, A21
A5 backbone O	Backbone N atoms of: R9, A10, A11, A12, A13, R14, A15, A16, A17, A18, R19, A20, A21
A6 backbone O	Backbone N atoms of: A10, A11, A12, A13, R14, A15, A16, A17, A18, R19, A20, A21
A7 backbone O	Backbone N atoms of: A10, A11, A12, A13, R14, A15, A16, A17, A18, R19, A20, A21
A8 backbone O	Backbone N atoms of: A10, A11, A12, A13, R14, A15, A16, A17, A18, R19, A20, A21
R9 backbone O	Backbone N atoms of: A11, A12, A13, R14, A15, A16, A17, A18, R19, A20, A21
A10 backbone O	Backbone N atoms of: R14, A15, A16, A17, A18, R19, A20, A21
A11 backbone O	Backbone N atoms of: A15, A16, A17, A18, R19, A20, A21
A12 backbone O	Backbone N atoms of: A15, A16, A17, A18, R19, A20, A21
A13 backbone O	Backbone N atoms of: A15, A16, A17, A18, R19, A20, A21
R14 backbone O	Backbone N atoms of: A16, A17, A18, R19, A20, A21
A15 backbone O	Backbone N atoms of: R19, A20, A21
A16 backbone O	Backbone N atoms of: A20, A21
A17 backbone O	Backbone N atom of A21
R9 backbone N	Side chain CZ atom of R9
R14 backbone N	Side chain CZ atom of R14
R19 backbone N	Side chain CZ atom of R19
R9 side chain CZ	Side chain CZ atoms of: R14, R19
R14 side chain CZ	Side chain CZ atom of R19

Table S2 completes the summary of the clustering parameters exploited in our work. For the interpretation and impact of parameters, we refer the reader to the relevant literature cited in 2.1.3 and to the documentation for the software CAMPARI where these algorithms are implemented (http://campari.sourceforge.net).

Table S2: *Additional clustering parameters. Sections of the main text where this information is relevant are provided in parentheses.*

	Tree-based clustering (2.3.2)	Hierarchical clustering with mean linkage (2.3.2)	r-PIGS (2.2.2)	φ-PIGS (2.2.2)
Cluster radius at coarsest level	7 Å	-	5 Å	100 degrees
Cluster radius at finest level (size threshold)	1.8 Å	1.5 Å	1 Å	25 degrees
Tree height	10	-	8	8
Maximum search attempts for progress index construction	-	-	200	200

Supplementary Figures

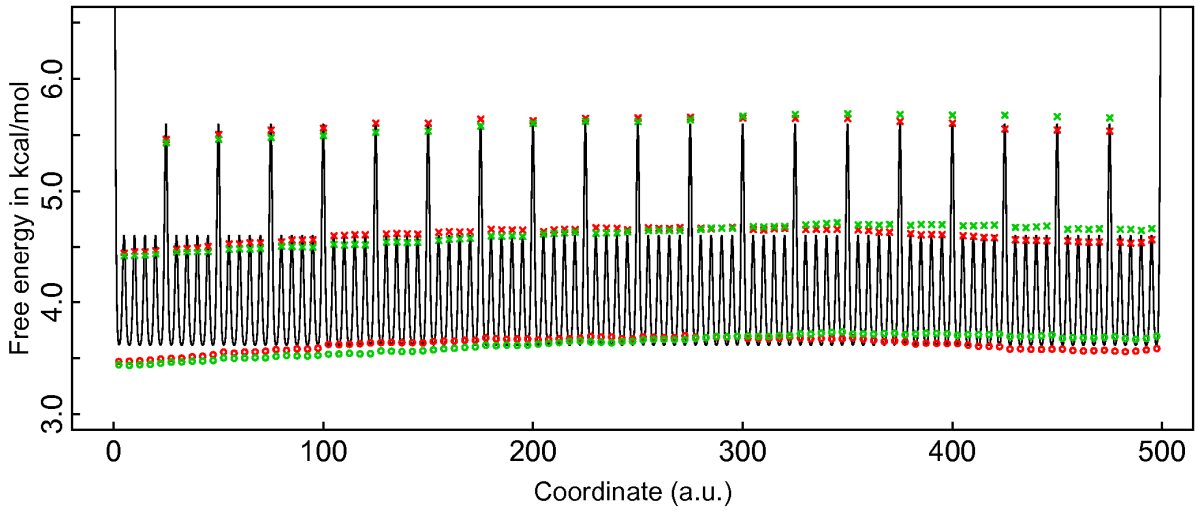

Figure S1: *One-dimensional potential for the toy model and measured distributions.* The analytical solution to the free energy is shown in black. There are 99 basins separated by barriers of height equal to either 1 or 2 kcal/mol. Resultant distributions from long Monte Carlo (MC) simulations are summarized by points at the bottom of the basins and on top of the barriers. Data from independent MC runs (N_r = 16, N_p = 16) are shown in green, and PIGS data are shown in red (N_r = 16, N_p = 8, f_p = 1000, n_O = 100).

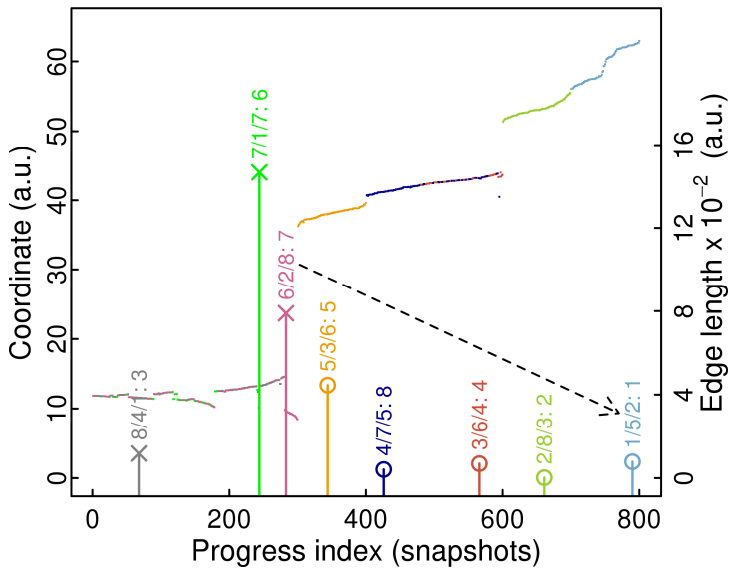

Figure S2: *Illustration of the reseeding procedure for the toy model (see Fig. S1).* For a run with parameters $N_r = 8$, $N_p = 4$, $f_p = 1000$, and $n_O = 100$, the 800 snapshots collected during a single interval have been ordered in the so-called progress index. Every replica contributes 100 snapshots. Colored dots correspond to the positions of the particles associated with these snapshots (left y-axis) and the replica they originated from. The vertical lines and symbols denote the positions of the final snapshots of the 8 replicas in the progress index along with the edge lengths of the spanning tree associated with them (right y-axis). Circles indicate replicas exempt from reseeding due to the quartile-based rejection criterion. Crosses indicate replicas for which this criterion is not fulfilled. The 3 individual rank orders and the composite rank are provided as well. They are, in order, based on position in the progress index, edge length, and minimum distance to any other final snapshot. In the example shown, the green and pink replicas are candidates for reseeding because their composite ranks (6 and 7) are larger than N_p and the quartile-based rejection criterion is not fulfilled. This is consistent with the fact that they sample the same area of phase space, between 10 and 15. The black, dashed arrow indicates the only reseeding actually performed for the interval in question, *i.e.*, the pink replica was reseeded with the final snapshot of the light blue replica (composite rank 1).

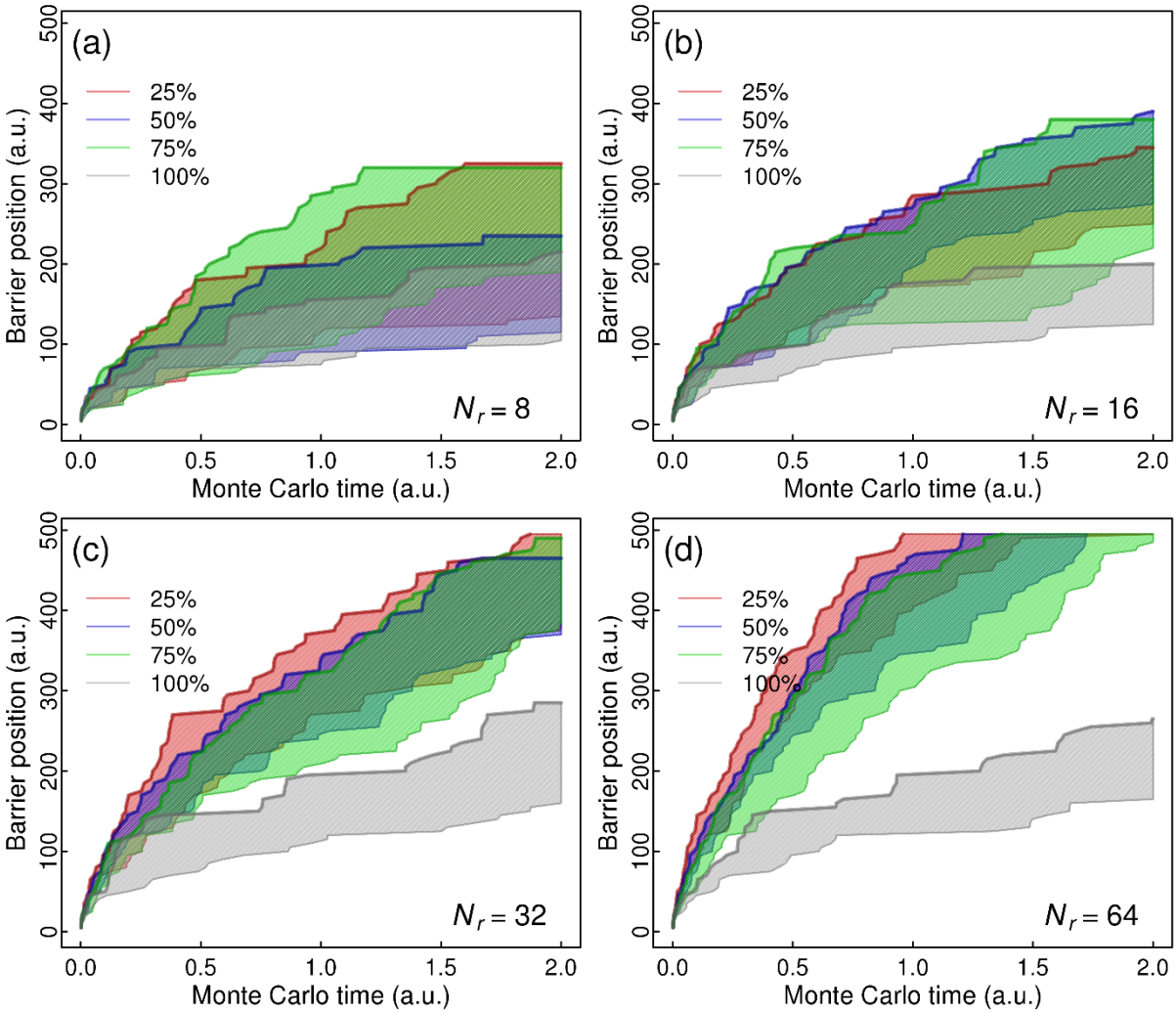

Figure S3: *Exploration rate for the model system for different values of* N_r and N_p. Here we simply rearrange the data of Fig. 1 in the main text (see 3.1). Shaded areas and solid lines are identical. The color scheme refers to four different values of the ratio N_p/N_r (100% corresponds to independent MC runs). **(a)** Data are shown for $N_r = 8$ and N_p/N_r ranging from 25% to 100%. **(b)** The same as (a) for $N_r = 16$ **(c)** The same as (a) for $N_r = 32$. **(d)** The same as (a) for $N_r = 64$.

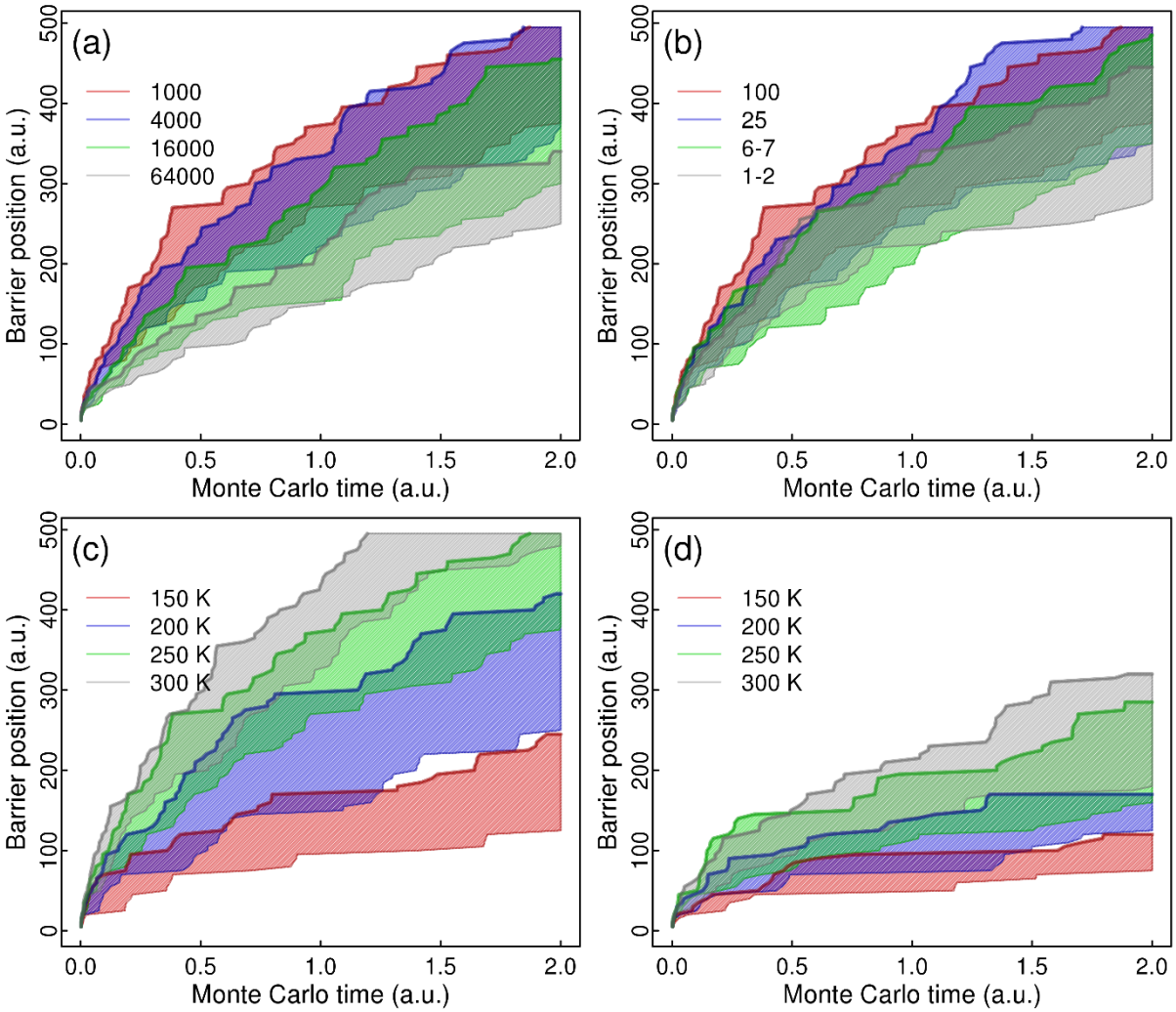

Figure S4: *Sensitivity to parameters f_p, n_o, and temperature for the one-dimensional model system.* The shaded areas and solid lines are obtained as explained in 3.1 for Fig. 1. **(a)** Data are shown for $T = 250$ K, $N_r = 32$, $N_p = 8$, and $n_O = 100$ with f_p being varied. **(b)** Data are shown for $T = 250$ K, $N_r = 32$, $N_p = 8$, and $f_p = 1000$ with n_O being varied. **(c)** Data are shown for $N_r = 32$, $N_p = 8$, $f_p = 1000$, and $n_O = 100$ with T being varied. **(d)** Data are shown for $N_r = 32$, $N_p = 32$, $f_p = 1000$, and $n_O = 100$ with T being varied.

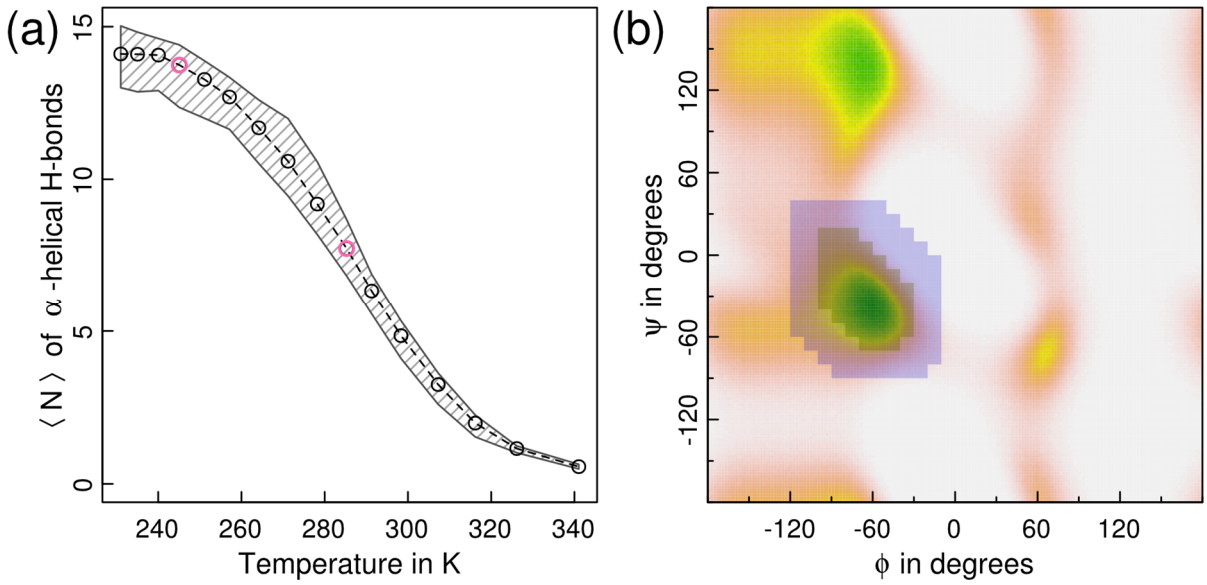

Figure S5: *The temperature-dependent helix-coil transition for the computational model in use and definition of the α-helical region for the state assignment explained in 2.3.1.* **(a)** Average number of α-helical hydrogen bonds as a function of the simulation temperature for REX runs starting from the straight α-helix. Shaded areas have been determined from a total of 12 blocks from 4 runs while discarding the first 78 ns. Pink circles correspond to the temperatures analyzed in detail, *viz*., 250 and 290 K. **(b)** The background shows a population histogram averaged over all residues and all runs for the REX data at 340 K. This Ramachandran map illustrates the realizable values for the φ/ψ-angles for the computational model. Definitions of α-helical (shaded area in gray) and boundary regions (shaded area in blue) are provided. These are required for the state assignment described in 2.3.1.

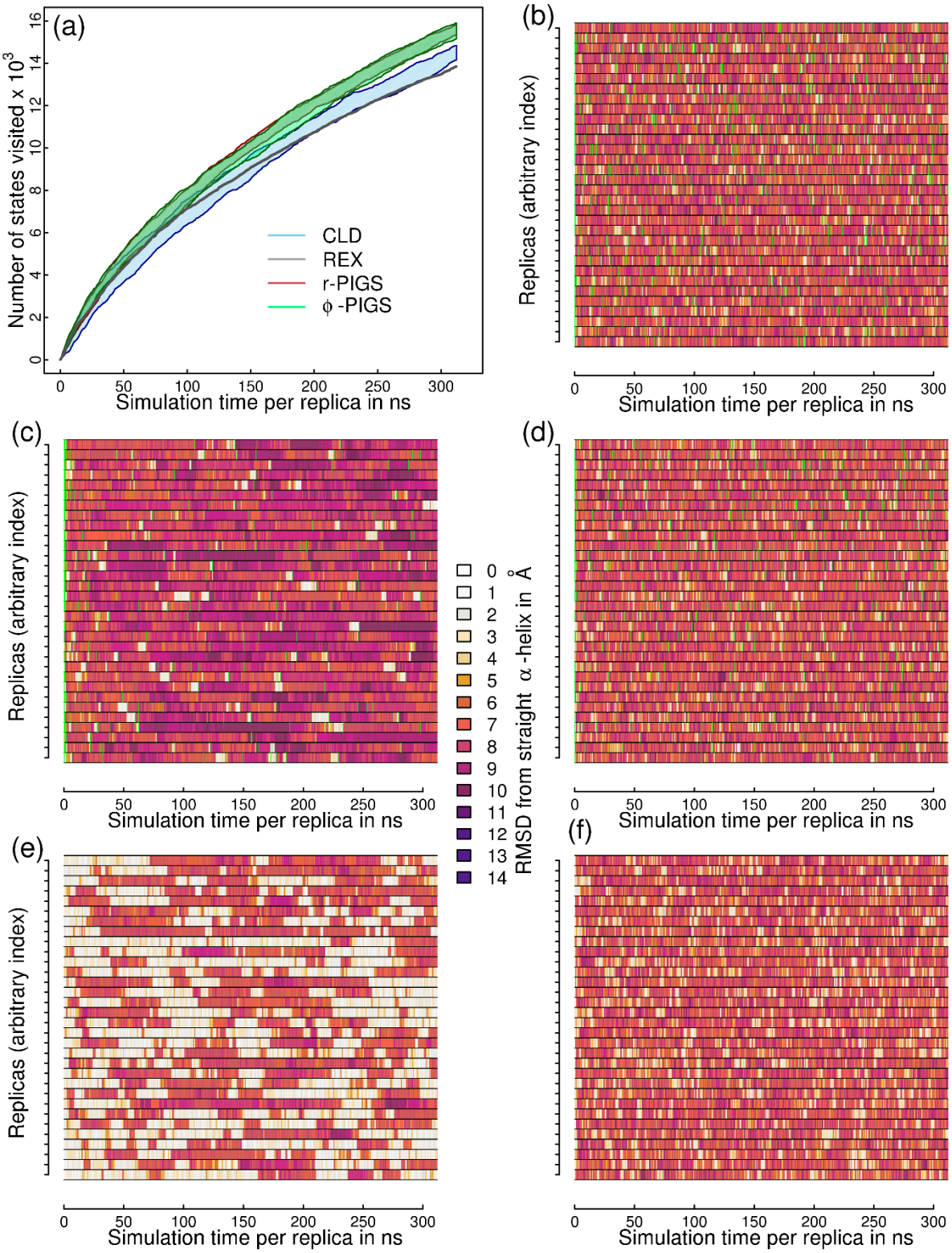

Figure S6: *Rate of exploration for different samplers and temperatures (complementary to Fig. 3 in the main text).* **(a)** Comparison for the exploration rate for the 4 sampling schemes at 290 K. Data favor REX since bootstrapping is applied to CLD and PIGS data as in Fig. 3(c). Data ranges for r- and φ-PIGS overlap almost exactly making the r-PIGS data difficult to see. **(b)** The complementary plot to Fig. 3(d) for r-PIGS at 290 K. **(c)** Same as (b) for φ-PIGS at 250 K. **(d)** Same as (b) for φ-PIGS at 290 K. **(e)** Same as (b) for CLD at 250 K (there are no reseedings for CLD). **(f)** Same as (e) for CLD at 290 K. The central color legend applies universally to panels (b)-(f).

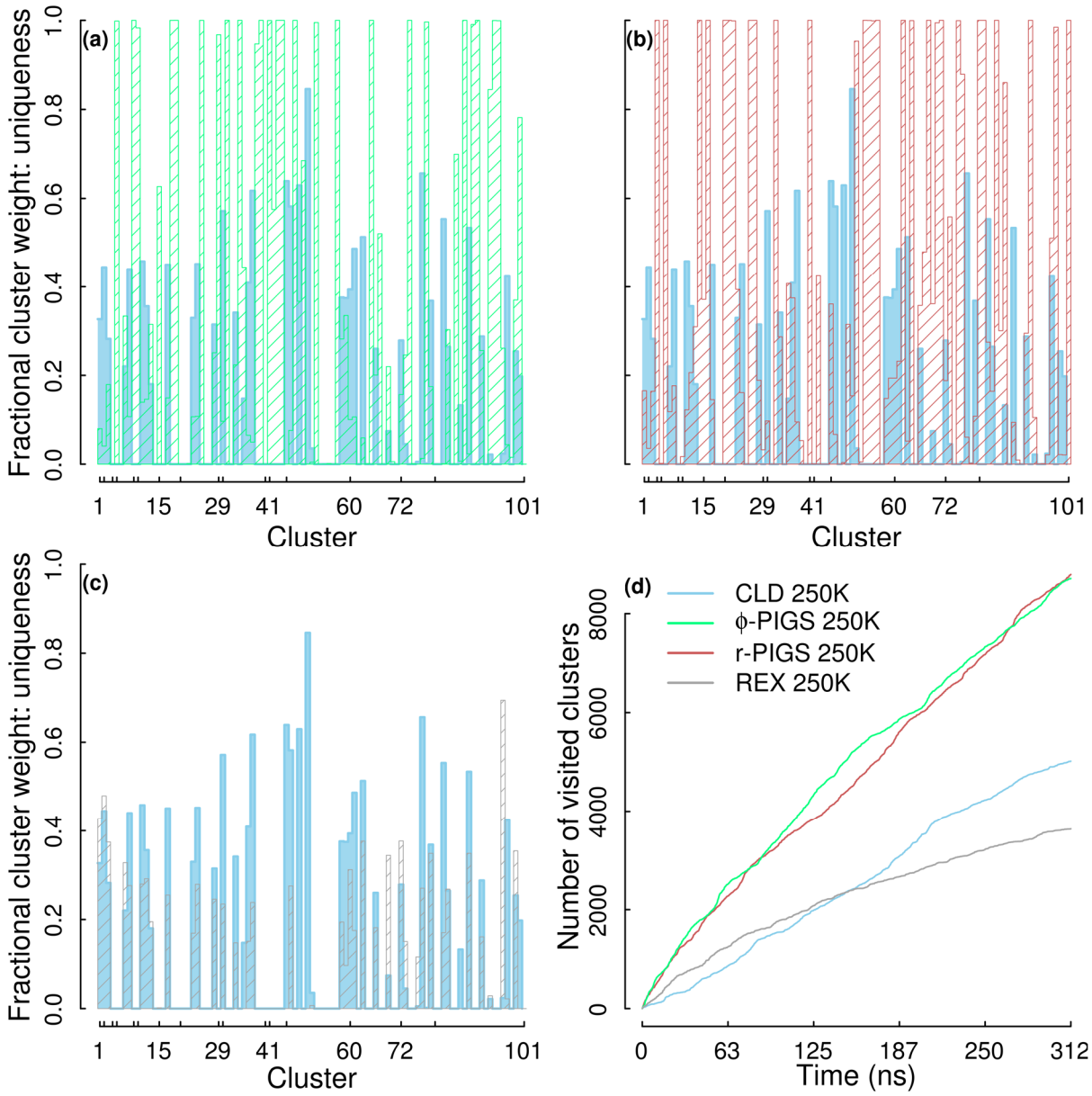

Figure S7: *Relative weights for the 101 numbered states at 250 K and cluster-based exploration rate.* These data quantify cluster uniqueness as used in Fig. 2 in the main text. A state explored by just one sampler has a uniqueness value equal to 1.0. REX data are scaled by a factor of 32/4 for uniqueness computations. **(a)** Comparison between CLD (cyan) and φ-PIGS (green). 56 and 21 states are not visited by CLD and φ-PIGS, respectively. **(b)** The same as (a) for CLD and r-PIGS (red). 13 states are not visited by r-PIGS, and the joint set sampled by φ- and r-PIGS is complete. **(c)** The same as (a) for CLD and REX (gray). 62 states are not sampled by REX. Neither REX nor CLD provide a unique state. **(d)** Exploration rate measured by visitation counts for clusters with size of at least 100 as a function of time. The PIGS schemes cover more than 80% of the roughly 10^4 clusters, and the advantage over CLD and REX is consistent with Fig. 3(c). Among these 10^4 clusters, only 26 are unique to REX, which compares to 84 for CLD and about 1500 each for φ-PIGS and r-PIGS.

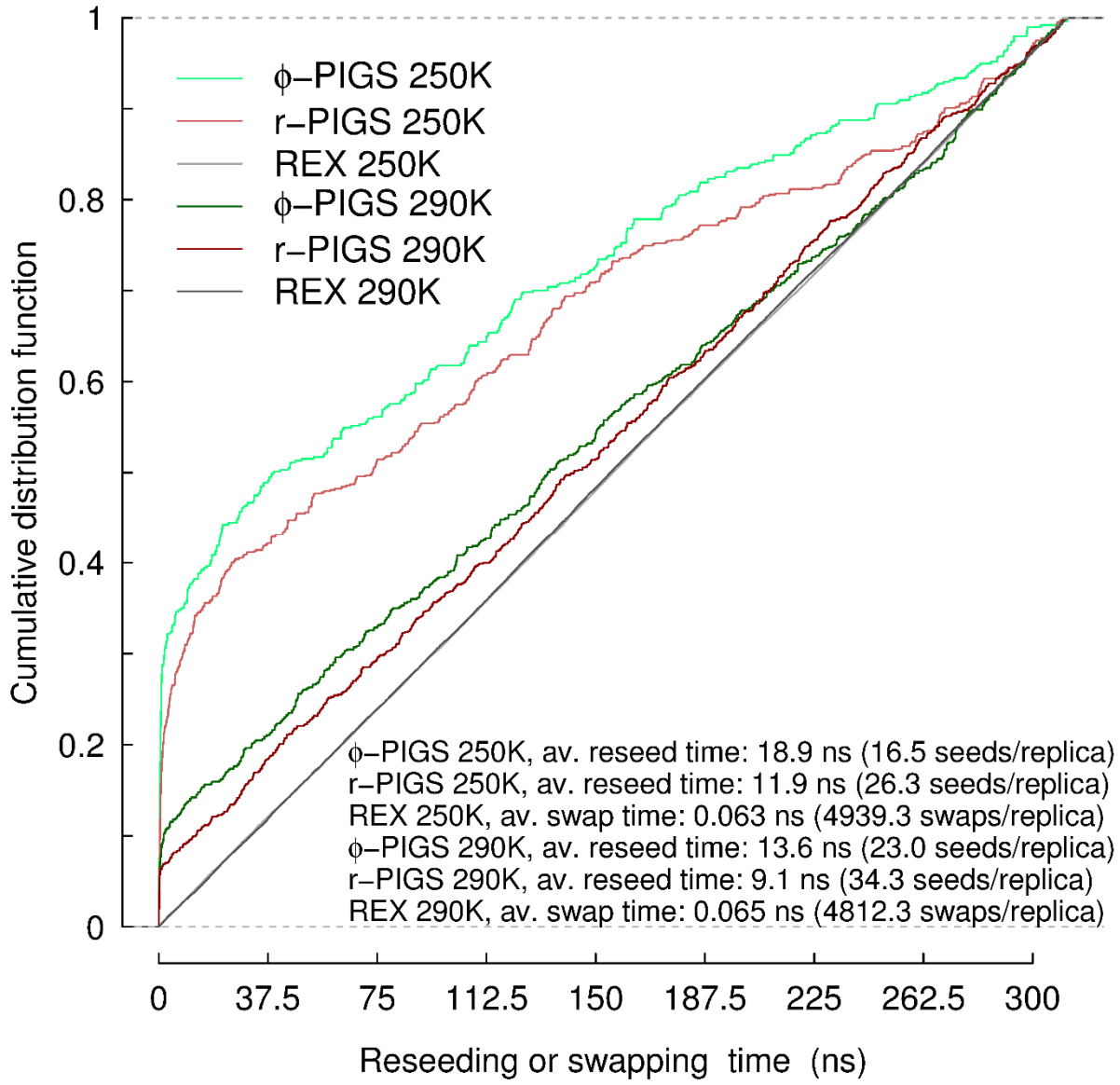

Figure S8: *Cumulative probability densities for reseedings (PIGS) and swaps (REX).* For PIGS, reseedings are concentrated in the first part of the simulations when sampling overlap is likely to be high. Later on, the large number of metastable states appears to support a high level of uniqueness for different replicas, and the reseeding rate is reduced. This allows the sampling of long stretches of unperturbed dynamics as indicated. The drop in reseeding rate may be used to identify the simulation time at which most PIGS replicas have diverged. Conversely, REX maintains a constant swapping rate throughout (acceptance rate near 50%).

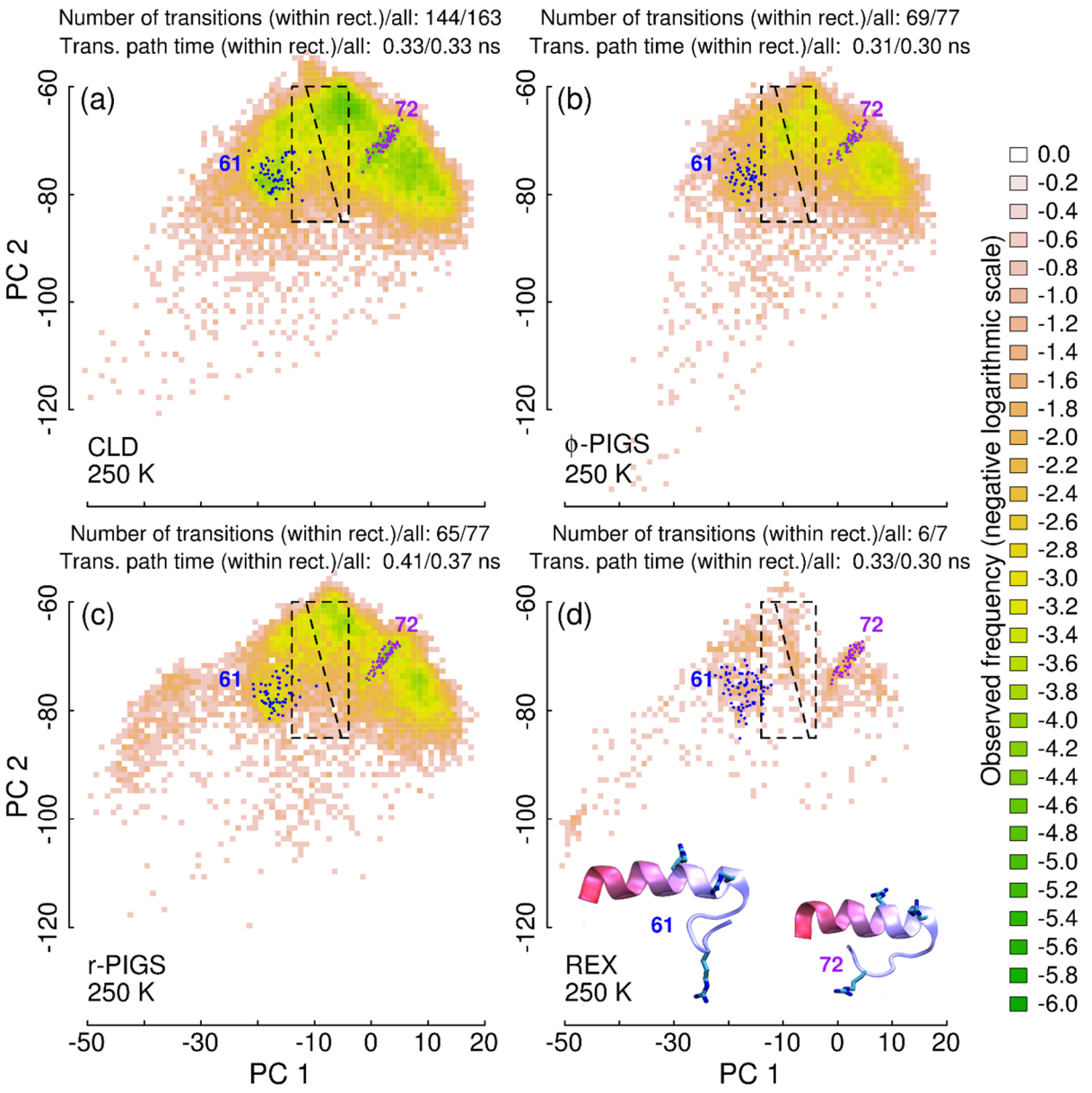

Figure S9: *Analysis of transition paths between states 61 and 72 at 250 K.* Data and plots are analogous to Fig. 6 in the main text (see 3.2.4). The negative logarithms of the individual histograms for the four different protocols are plotted as color maps. Bins with no counts are shown in white. Bin widths are 1 Å for both principal components (see Table II in the main text for further details). 50 points each belonging explicitly to states 61 and 72 are shown as blue and purple dots, respectively. In each panel, they represent the state as sampled by the protocol in question. 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.

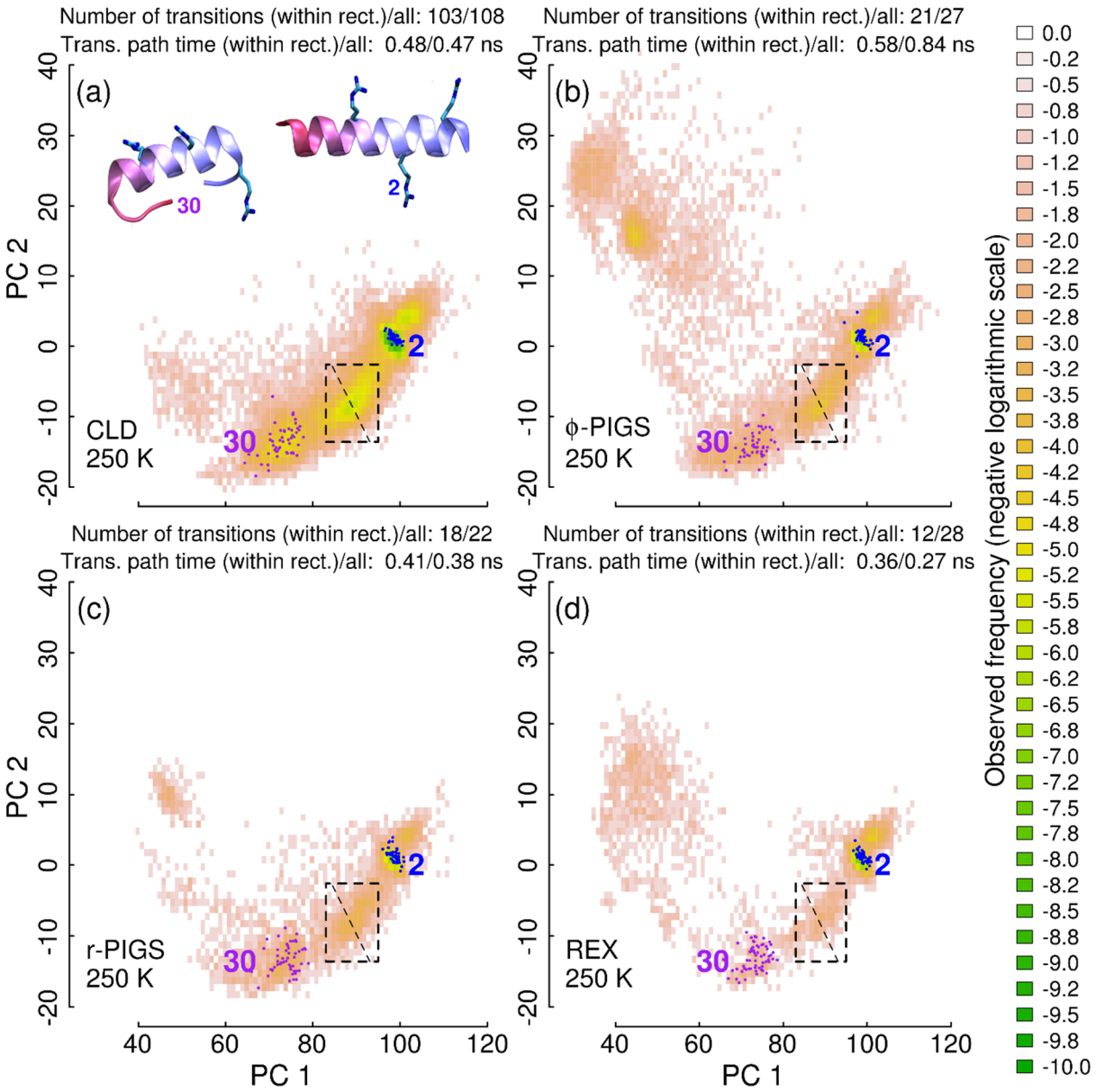

Figure S10: *Analysis of transition paths between states 2 and 30 at 250 K.* This figure is identical to Fig. S9 except that data for the transition paths between states 2 and 30 are shown.

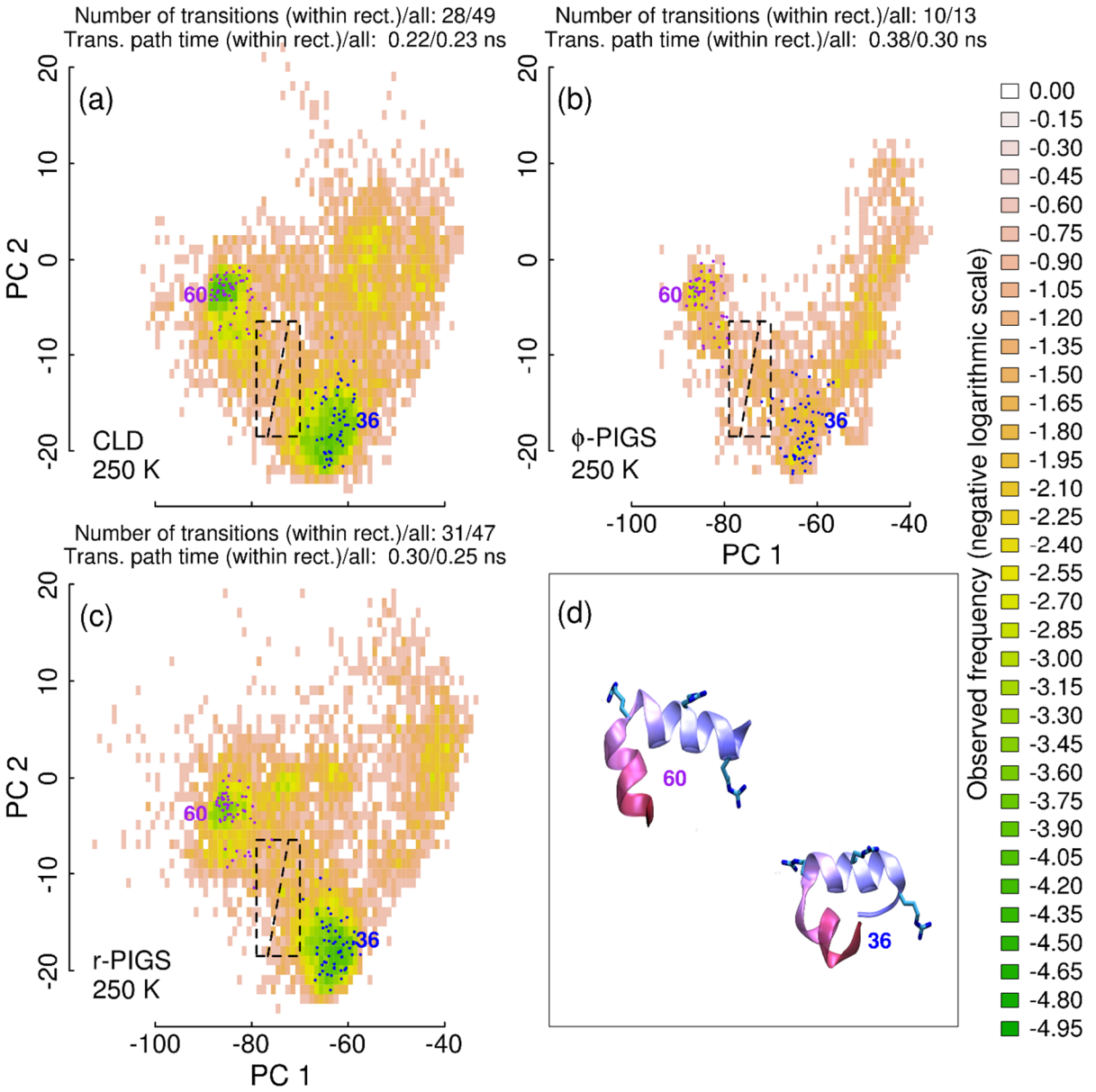

Figure S11: *Analysis of transition paths between states 36 and 60 at 250 K.* This figure is identical to Fig. S9 except that data for the transition paths between states 36 and 60 are shown. REX does not sample this transition, and panel (d) is left empty except for cartoon images of the end states.

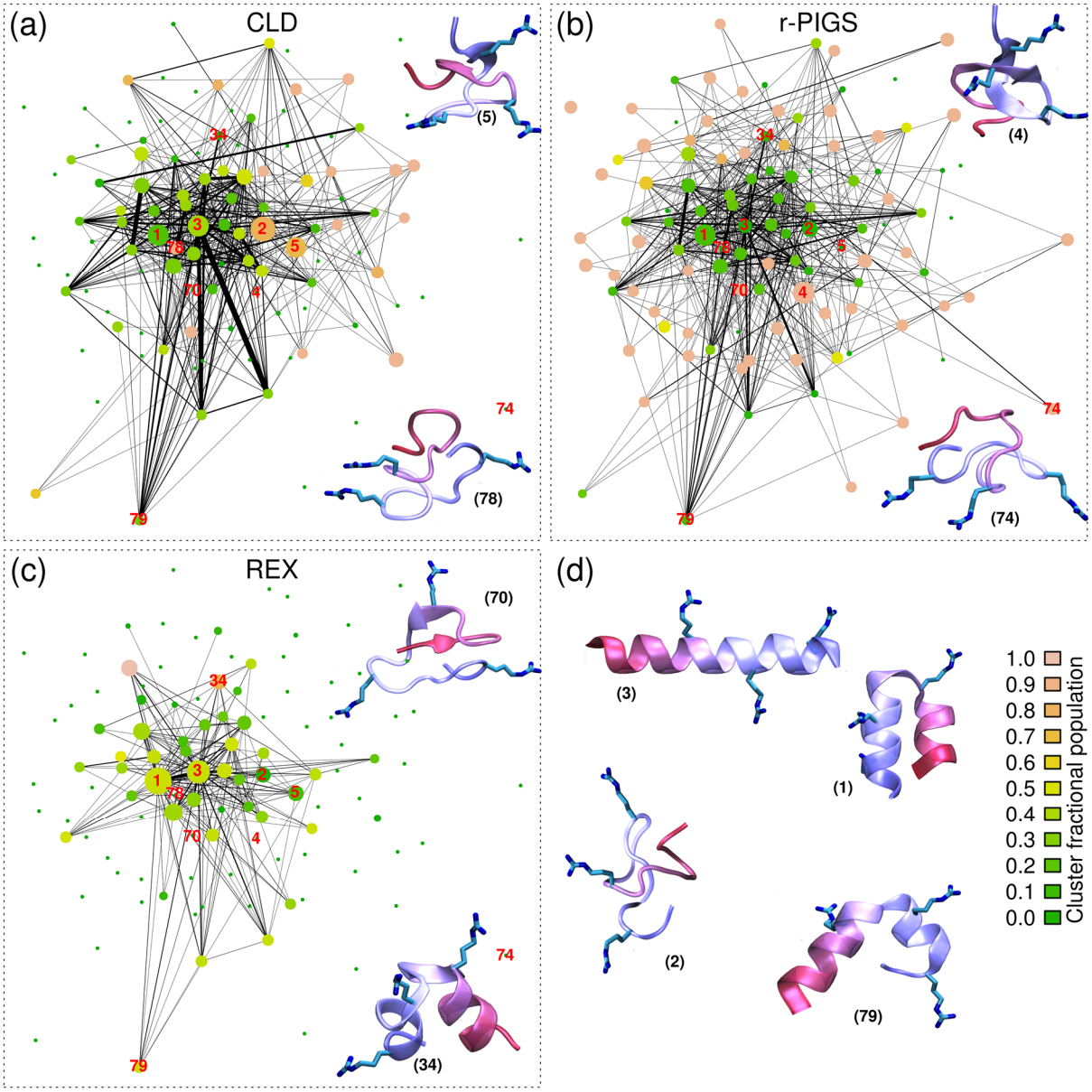

Figure S12: *Conformational space networks for the FS peptide simulations starting from the globule at 250 K*. We refer the reader to Fig. 2 in the main text for information on coloring scheme, layout, and the relations between the underlying statistics and plotting parameters. The numbering of states indicated next to the structure images is completely independent of that in Fig. 2, and any agreement is by chance. **(a)** Network for CLD. Due to the metastability of the starting state (here, state 2), it is sampled much more by CLD than by REX or r-PIGS. Note the larger amount of unique states sampled by CLD compared to Fig. 2(a). **(b)** The same as (a) for r-PIGS. **(c)** The same as (a) for REX. **(d)** Additional images of states and color legend.

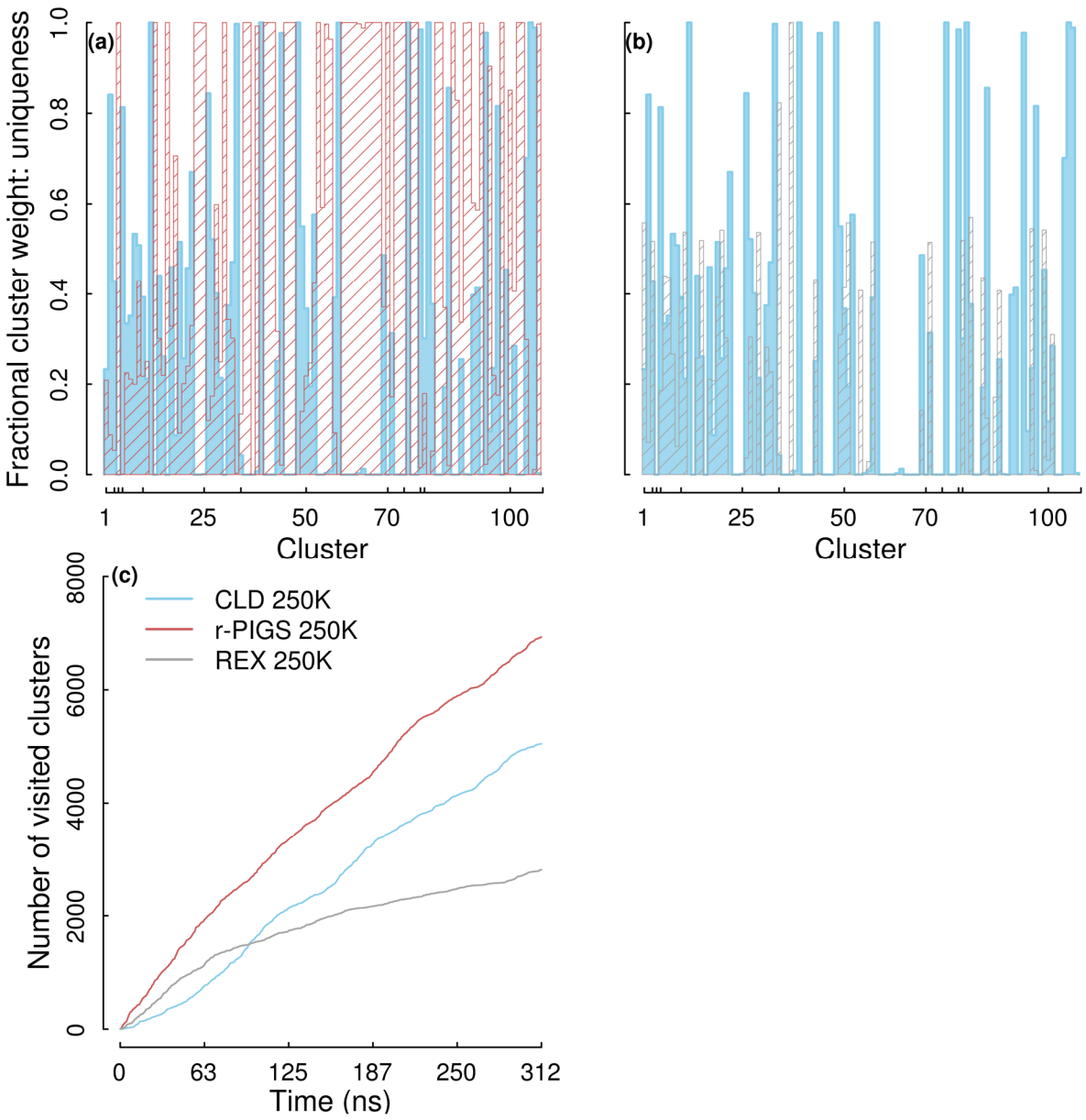

Figure S13: *Relative weights for the 108 numbered states at 250 K obtained for simulations started from the globule (see 3.2.5) and cluster-based exploration rate*. This figure is analogous to Fig. S7. **(a)** Comparison between CLD (cyan) and r-PIGS (red). 39 and 6 states are not sampled by CLD and r-PIGS, respectively. **(b)** The same as (a) for CLD and REX (gray), which does not sample 63 of the 108 states. **(c)** Exploration rate measured by visitation counts for clusters with size of at least 100 as a function of time. The trends are overall very similar to Fig. S7(d) with the relative performance of CLD slightly improved over long times. The initial burst of exploration for REX is amplified relative to Fig. S7(d).

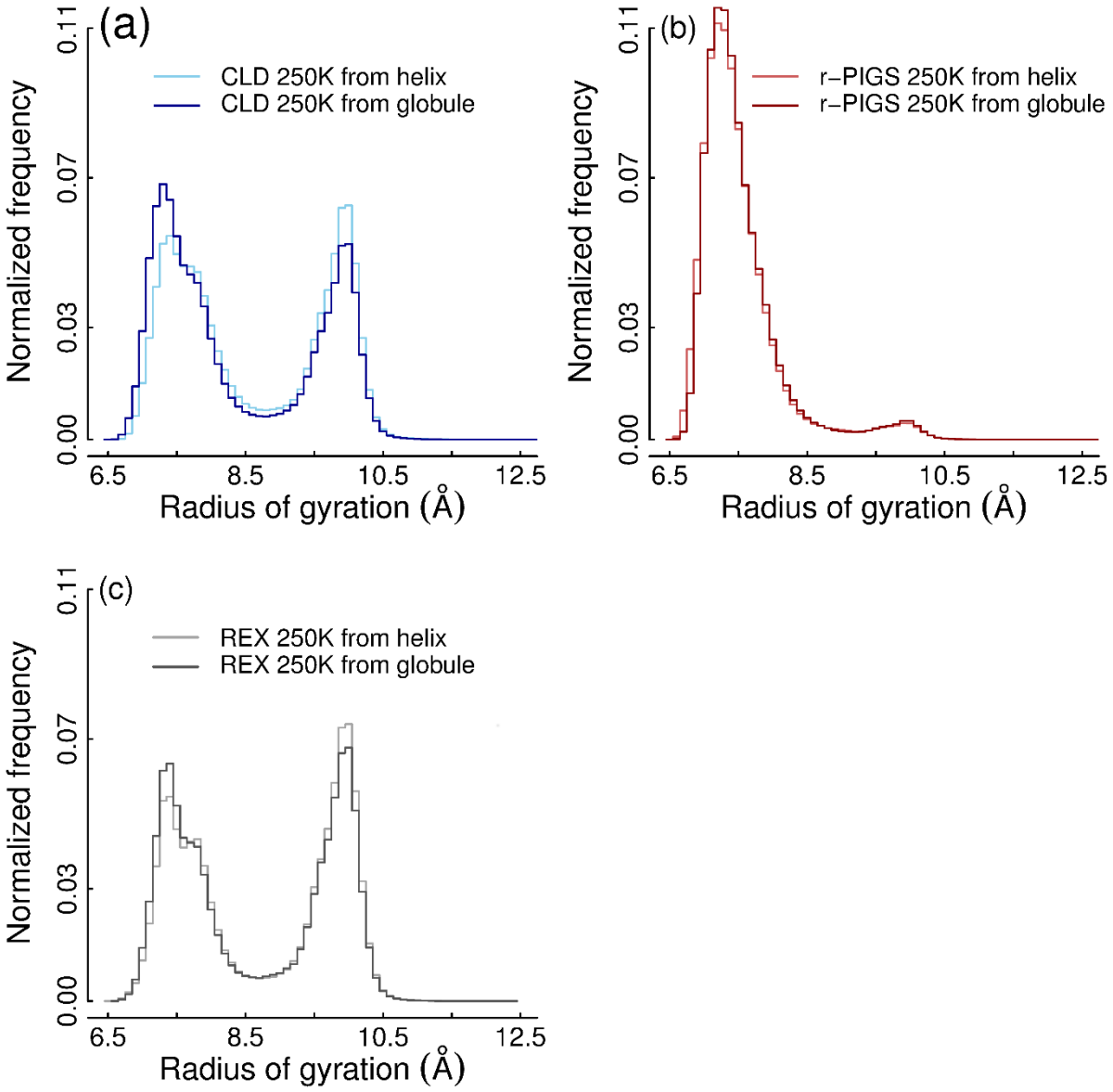

Figure S14: *Distributions for the radii of gyration*. We compare distributions of the radii of gyration (bin width of 0.1 Å) between the sets of simulations starting either from the straight α-helix or from the globule (see 3.2.5 in the main text). To allow for relaxation of initial state bias, the first half of the data, corresponding to 158 ns, has been discarded in each case. **(a)** The data for CLD. The curves indicate an influence of the different starting structures as indicated by the small differences in peak heights. **(b)** The same as (a) for r-PIGS. The curves are more similar to one another than for CLD. **(c)** The same as (a) for REX. The same conclusion holds as for CLD.

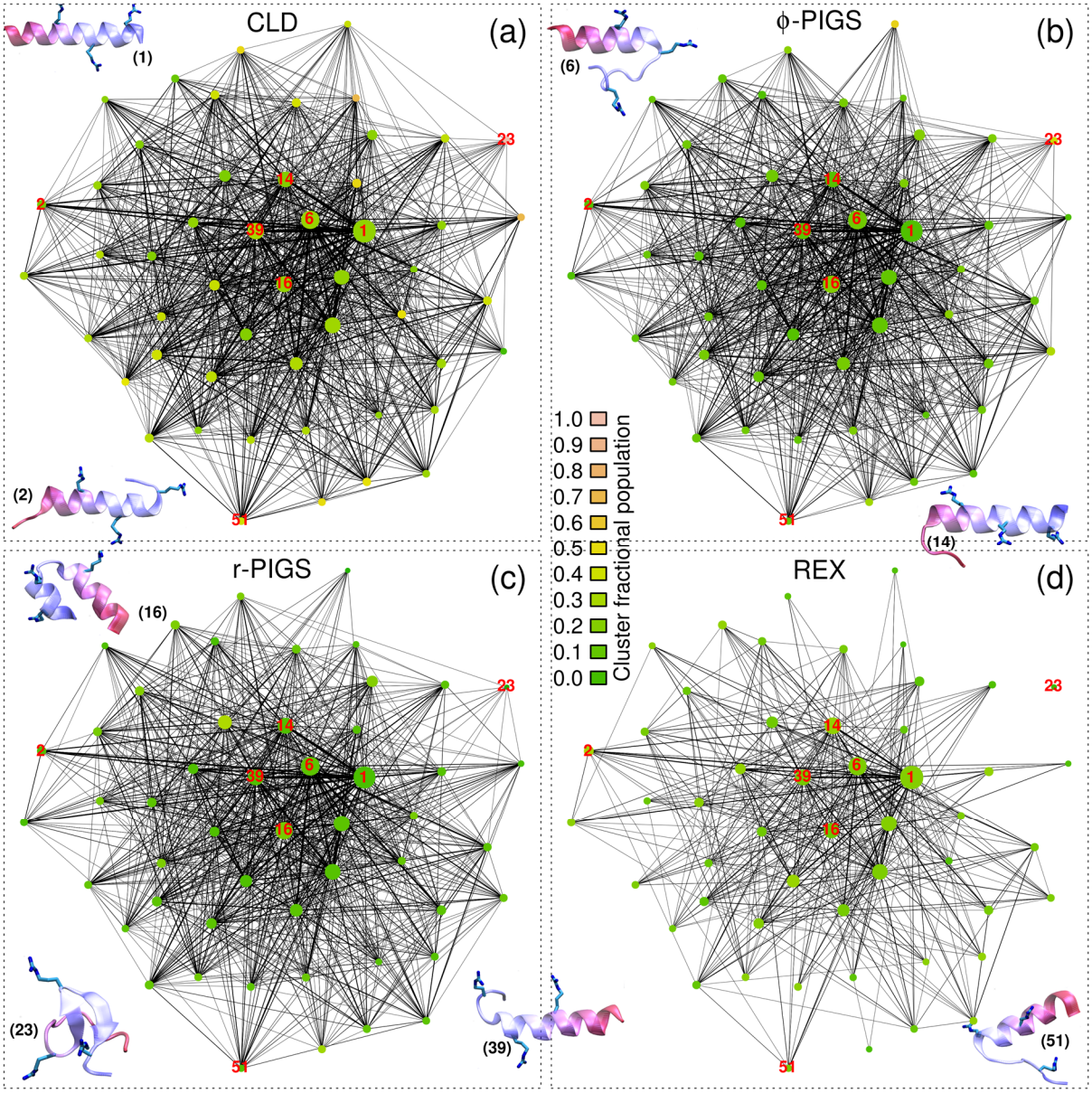

Figure S15: *Conformational space network for the FS peptide at 290 K.* This figure is completely analogous to Fig. 2 in the main text, and we followed the same protocol as described in 2.3.2. The numbering of states indicated next to the structure images is completely independent of that in Fig. 2. The data at this temperature show much more heterogeneity, and the initial clustering of the 11.232x10^6 snapshots yielded *ca*. 1.9x10^6 clusters, about ⅔ of which are single snapshots. The top 600 clusters in this case encompass only *ca*. 20% of the total data (compared to *ca*. 45% at 250 K). The hierarchical clustering of their centroids produced the 51 clusters shown in the network. For clustering only, the initial data concatenation contained data for REX from two adjacent temperatures, *viz*., 284 and 297 K. The network only reflects data at 290 K, however, in analogy to Fig. 2. The cartoons illustrate that almost all of the identified states are rich in α-helix. No qualitative differences are found between individual panels with the exception that REX loses a number of transitions globally. Most uniqueness values are close to 0.25.

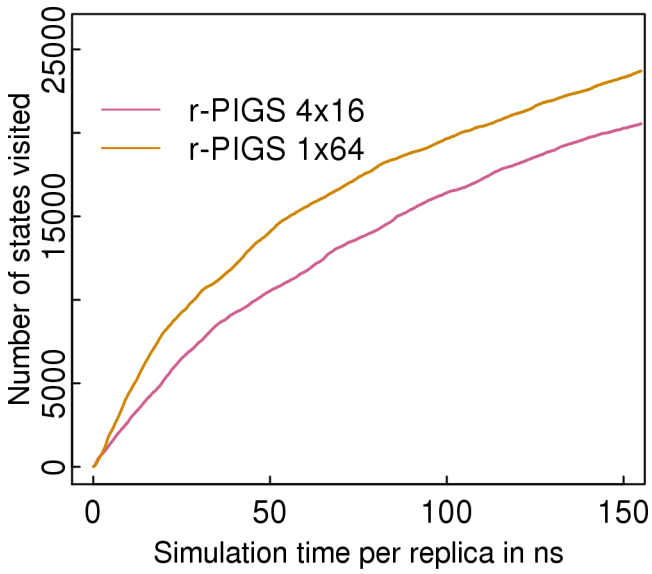

Figure S16: *Rate of exploration for varying allocation of resources for the FS peptide at 250 K*. We analyze the exploration rate of r-PIGS by comparing two data sets with simulation times of *ca*. 150 ns per replica. The comparison is between a single run across 64 cores and 4 independent runs using 16 cores each. Exploration rate is measured as the number of states as in Fig. 3 in the main text (see 2.3.1). It is evident that there is a synergistic effect from the greater parallelism allowing for a faster degree of exploration.

Movie S1: Evolution of the complex network for the four samplers. The set of states and layout is the same as in Fig. 2. The color coding for the nodes describes uniqueness of a state to a specific sampler at a given point in time. The movie illustrates qualitatively how all methods perform similarly in exploring the region of phase space that is directly adjacent to the initial structure (straight α–helix), which takes roughly 100 ns. Thereafter, discovery of new, long-lived metastable states is mainly accomplished by PIGS. This is a qualitative and potentially misleading measure of the exploration rate as the analysis is limited to 101 states derived from the data themselves. As in Fig. 3 in the main text, the REX method seems to offer no advantage over CLD here. This is most likely because the starting structure, a straight α–helix, is not surrounded by large energetic barriers.