Supplementary information: Optimized reaction coordinates for analysis of enhanced sampling

Julian Widmer,[1] Cassiano Langini,[1] Andreas Vitalis,[1] and Amedeo Caflisch[1]
University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(*Electronic mail: caflisch@bioc.uzh.ch)

(Dated: 21 June 2023)

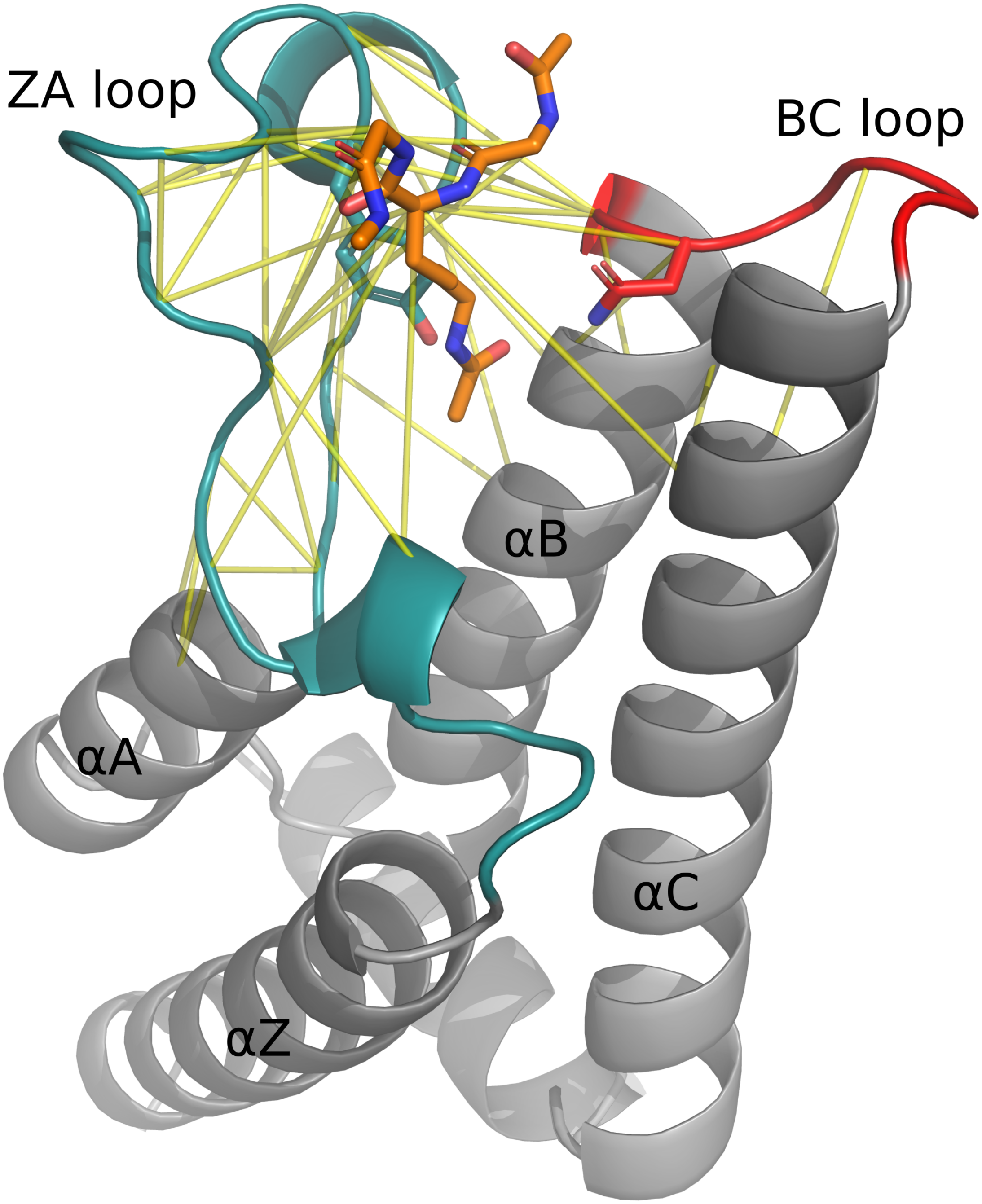

FIG. S1. **Structure of ATAD2 bromodomain with its cognate ligand and graphical illustration of MSM featurization**. Graphical representation of the bromodomain of ATAD2 in complex with the capped tripeptide GK*ac*G; coordinates are taken from PDB 4QUT. The ligand is in orange sticks whereas the architecture of the bromodomain is in cartoon and its most important structural elements are indicated, specifically the helix bundle (αZ, αA, αB, αC in grey), and the two major loops (ZA loop in cyan and BC loop in red). Tyr42 and Asn85, whose interaction with K*ac* is conserved, are in sticks. The pairs of residues used for featurization (see Table SI) are indicated as yellow lines connecting the Cα atoms of each pair. Note that the actual features between residue pairs contain multiple distances per pair (see IV D 1 for details).

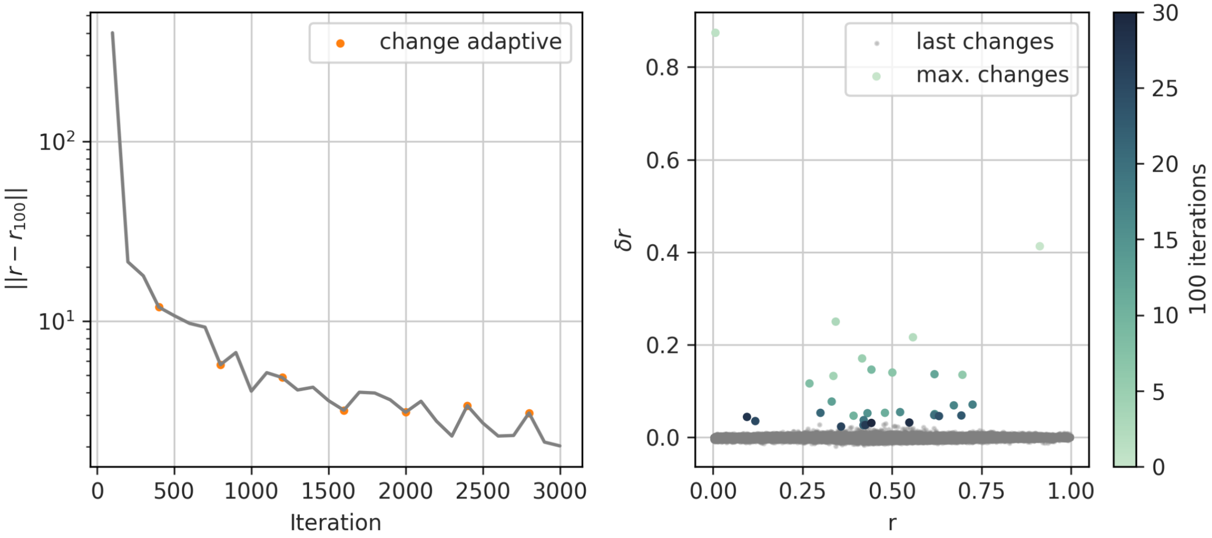

FIG. S2. **Optimization progress of the RC describing the transition of the straight, α-helical conformation of FS-peptide to distal (non-helical, collapsed) conformations**. Data are shown for PIGS, which is used as an illustration. Left: The norm of the RC compared to itself but 100 iterations earlier is plotted as a function of the number of iterations. Orange dots mark iterations, at which optimization is focused on another region according to the derivative of Δr^2 (see IV A for details). Right: Absolute differences per trajectory snapshot across 100 iterations are plotted as a function of the RC itself. Large, colored dots indicate the maximal, unsigned difference across snapshots plotted for the current value of the RC for that snapshot. The difference is taken compared to the RC 100 iterations prior. Darker colors indicate later iterations (one value per 100 iterations). Small, grey dots indicate the signed changes in RC for all snapshots of the final RC across the last 100 iterations of the optimization. These data show that the vast majority of snapshots undergo only small changes in RC during later stages of the optimization, and that these changes are largely free of systematic trends.

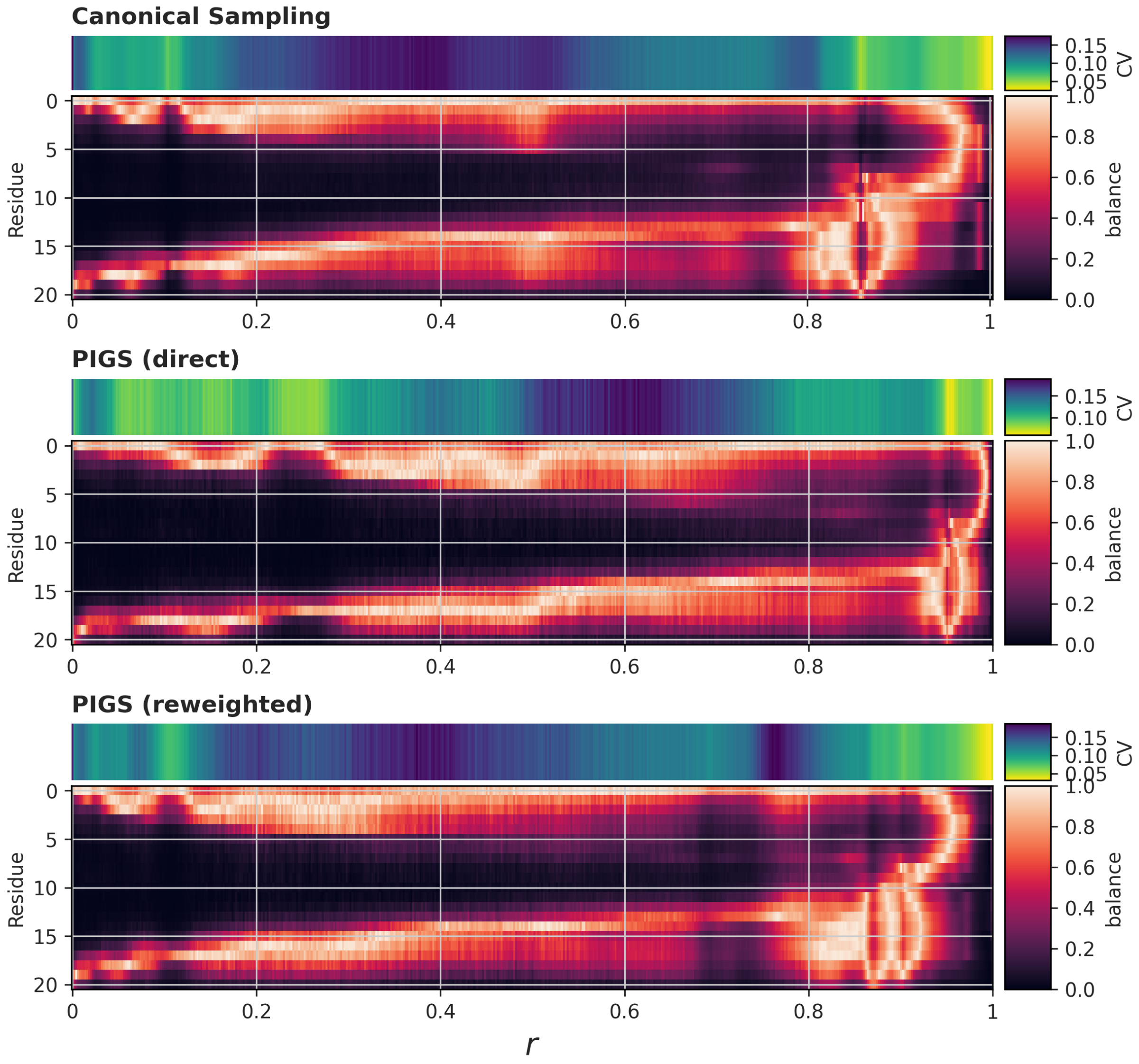

FIG. S3. **Kinetic progress according to the optimized RC for FS-peptide**. This figure is equivalent to Fig. 2, except that measures for intra-bin variability are shown instead of the means to aggregate snapshots in each bin. The heterogeneity in RMSD is quantified by the coefficient of variation (CV), whereas a balance measure for helicity is used as detailed in section IV B, low values indicating low variance within bins.

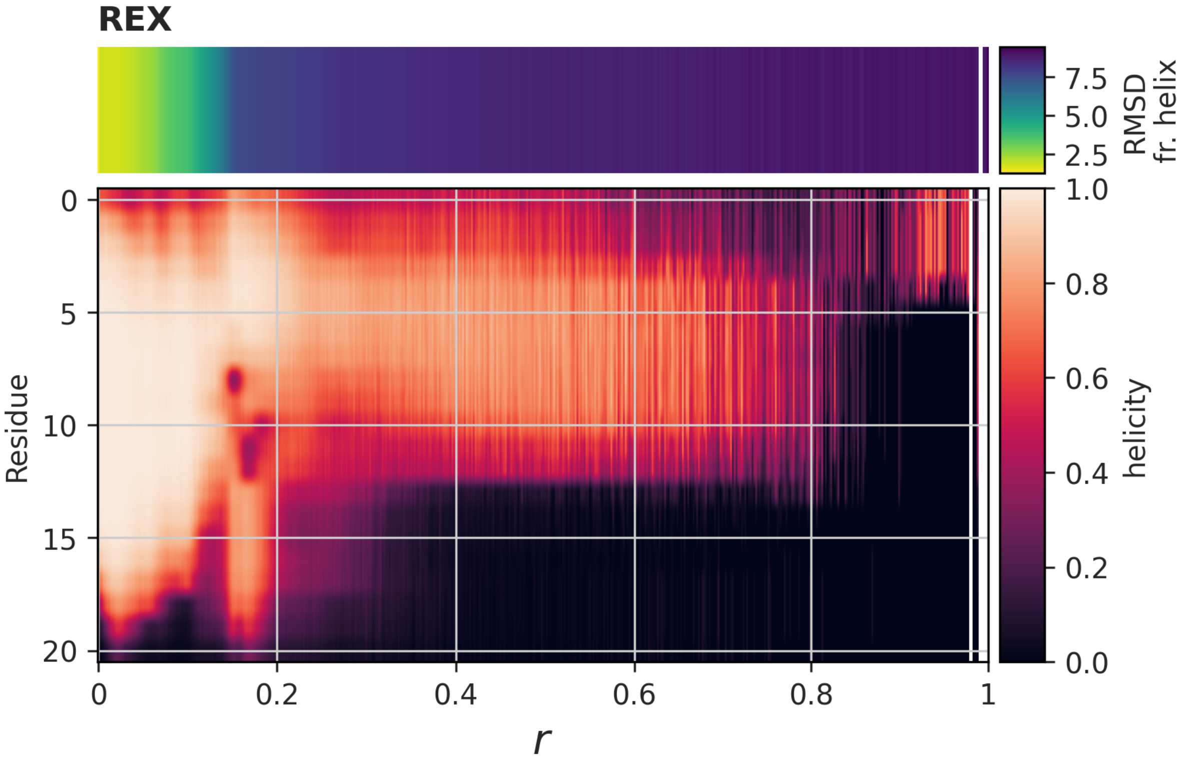

FIG. S4. **Annotated RC for the transition from the straight α-helix to distal (non-helical, collapsed) conformations for REX**. This figure is the same as an individual panel in Fig. 2. It highlights that, depending on the data set, absolute RC values are much less stable than the ranking of snapshots by RC. Here, most of the spectrum of RC values is covered by a specific set of partially helical states, indicating either that the interpretability of the RC as the committor is damaged in this case or that, as suggested in prior work, REX achieves thermodynamic equilibrium but does not sample conformational transitions at equilibrium at fixed condition,[1] or both. REX data are challenging for the framework because swaps occur frequently at the target condition, and it is not clear how well the many short trajectories (average length of only 42 snapshots per continuous segment) connect the two end states by geometrically continuous pathways.

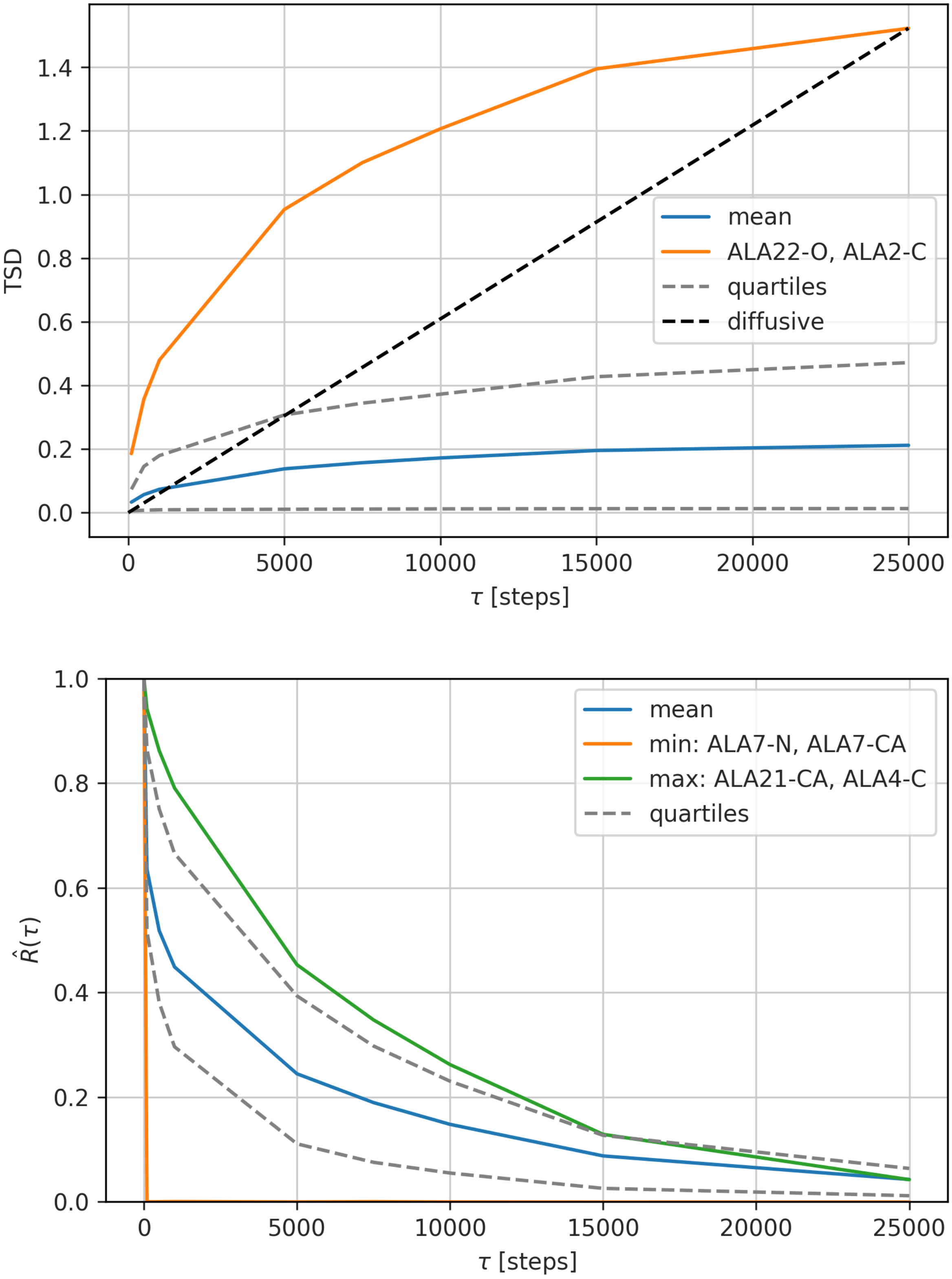

FIG. S5. **Characterization of input features entering the construction of the RC for FS-peptide**. Top: The autocorrelation was calculated for 100 randomly chosen input features between pairs of heavy atoms from FS-peptide (CS) as a function of various lag times. Naturally, features describing long-range conformational changes tend to retain longer memory. Bottom: The total-squared displacement of the same features. An interatomic distance between Ala2 and Ala22 is shown as the feature with the steepest increase. The black line represents free diffusion where the TSD is proportional to the time step Δt. Since features are defined on a domain bounded by covalent geometry, the TSD cannot grow linearly with time.

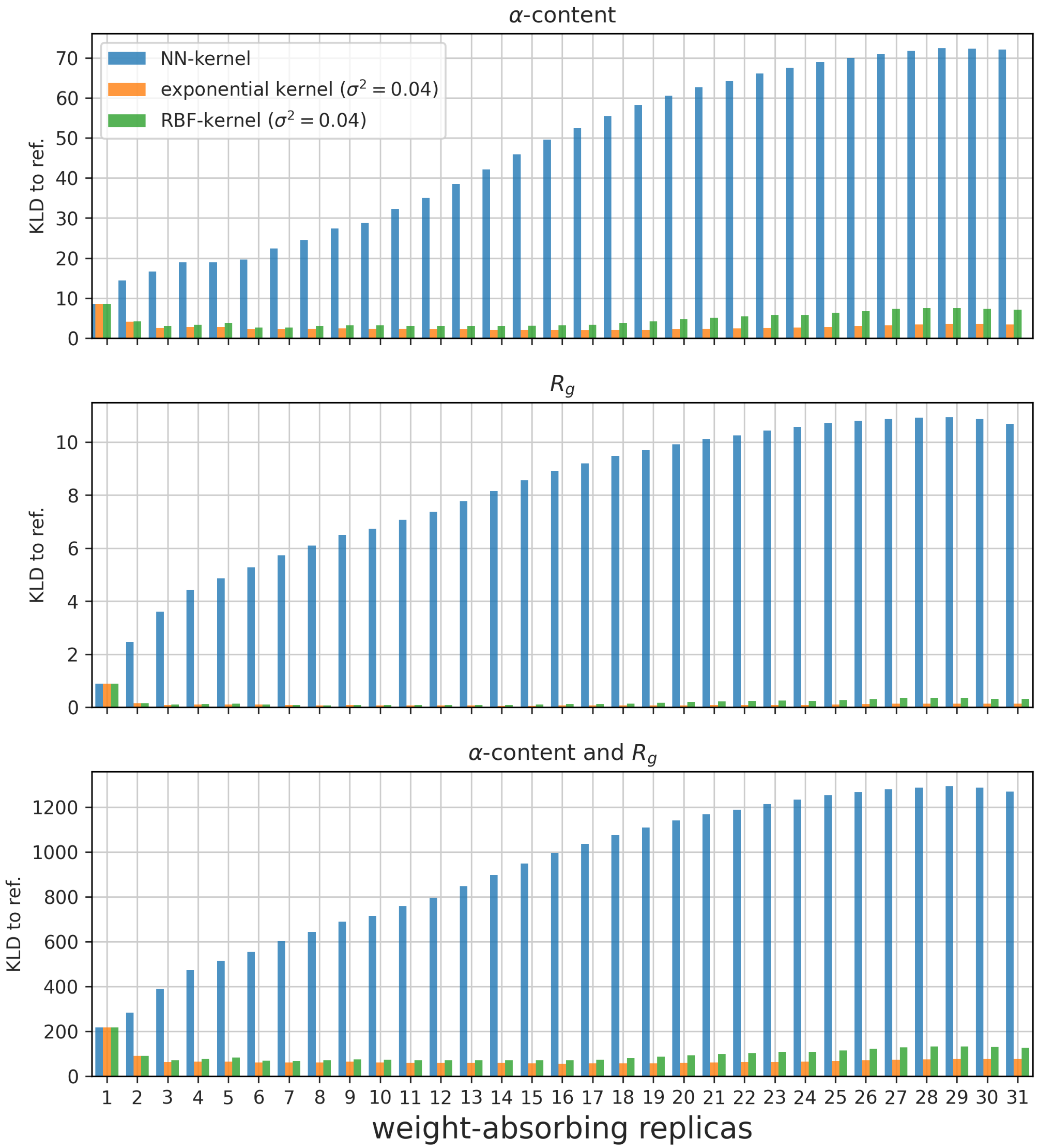

FIG. S6. **Kernel-based weight-splitting for the weighted ensemble method**. The KLD of the WE-reweighted distributions of α-content (top) and R_g (middle) as well as their joint distribution (bottom) to the canonical reference is shown for three different kernels as a function of the number of replicas (parameter n in IV B) that absorb the weight of a terminated replica. The NN-kernel is a default (negative control) model where the weight is split uniformly to the n closest, surviving replicas. The other two kernels are localized and thus depend on a width parameter (σ^2), which means that distal replicas receive only negligible weights irrespective of n. Note that for $n > 16$, it is possible that the effective n is lower because the number of surviving replicas is not large enough. This is handled dynamically per individual reseeding cycle.

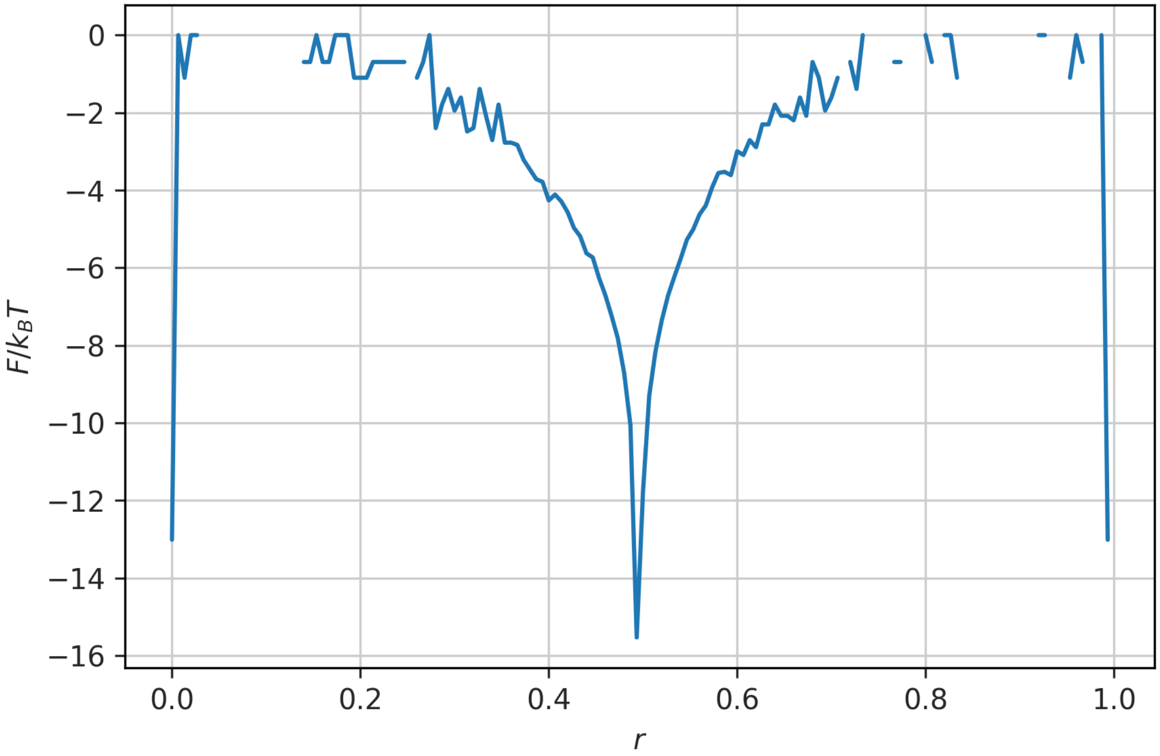

FIG. S7. **Behavior of the RC for FS-peptide in a pathological limit**. Here, the boundary states were chosen such that they practically coincide. The snapshots corresponding to the original boundary state representing the straight, α-helical conformation were split in half and assigned as the new boundary states A and B. The histogram-based FEP demonstrates that the RC values remain close to the initial guess (all intermediate states are given the value 0.5 initially) and reveals no interpretable kinetic structure. It is not clear whether the appearance of the FEP reflects an artificial, slow process 'discovered' by the RC or whether the RC should be regarded as non-converged.

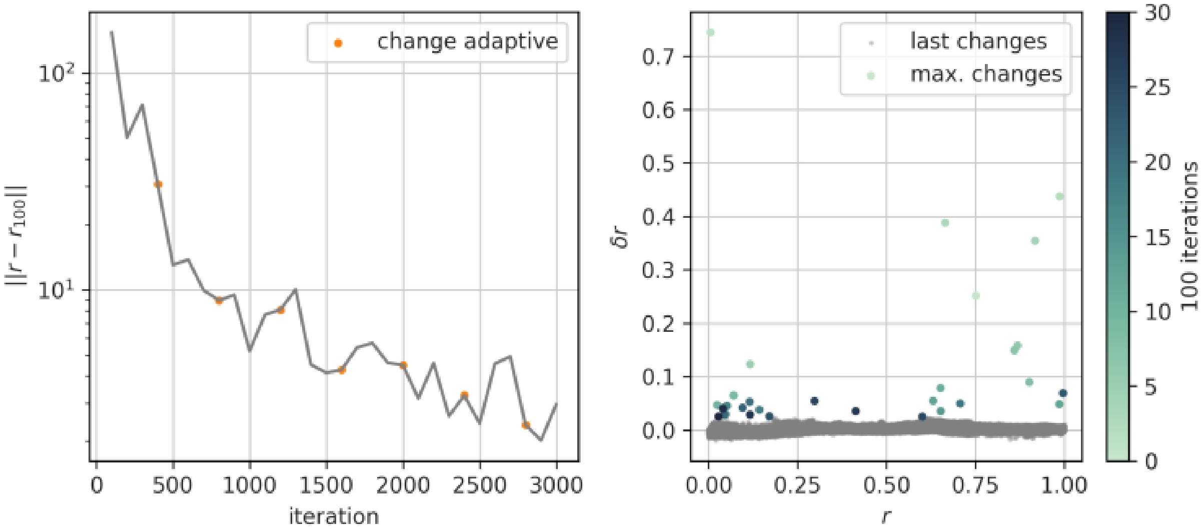

FIG. S8. **Optimization progress for the construction of the RC describing the unbinding of GK*ac*G from ATAD2**. This figure is completely analogous to Fig. S2.

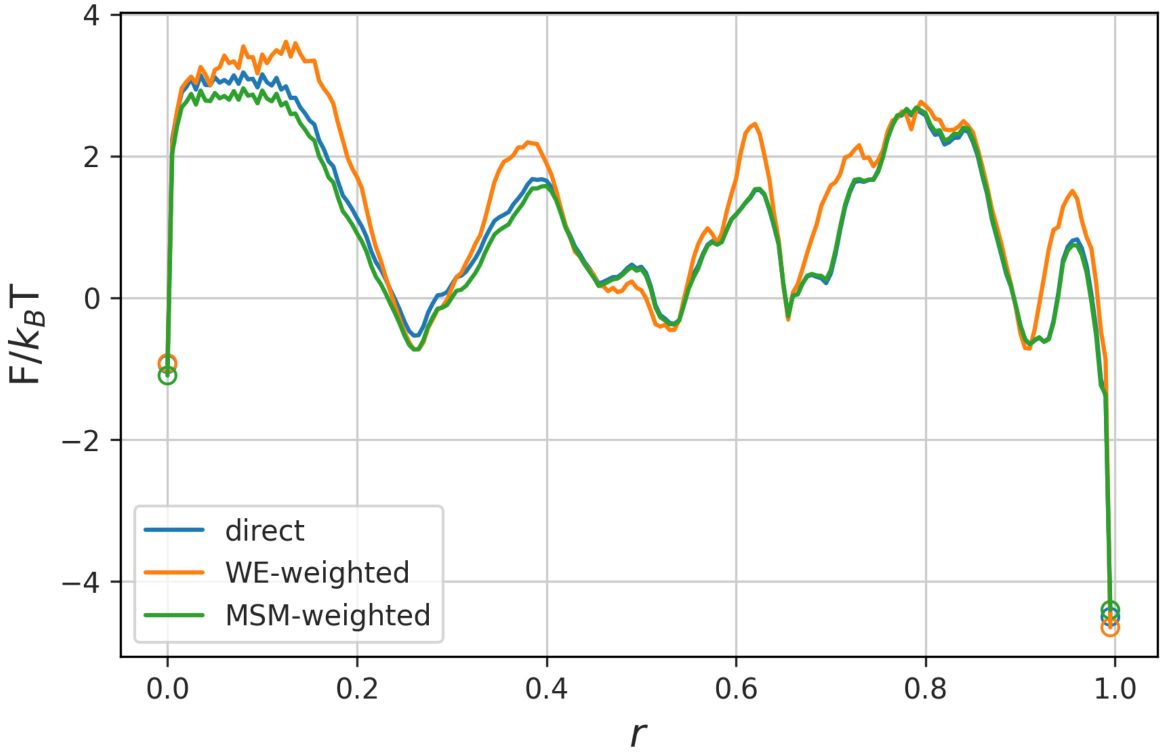

FIG. S9. **Effect of equilibrium reweighting strategies on the RC-derived free energy profile**. The histogram-based FEP of the RC for the bromodomain-peptide unbinding system is reproduced from Fig. 9 (direct) and compared to reweighted FEPs. The WE strategy ("WE-weighted") applied the reweighting only to the histogram with the WE weights derived from the same RC. This means that weights played no role during the optimization itself (compare IV B). The MSM strategy derived the weights for the histogram analogously from the stationary probabilities calculated from the MSM's transition matrix. At the boundaries, circles are added for visual clarity.

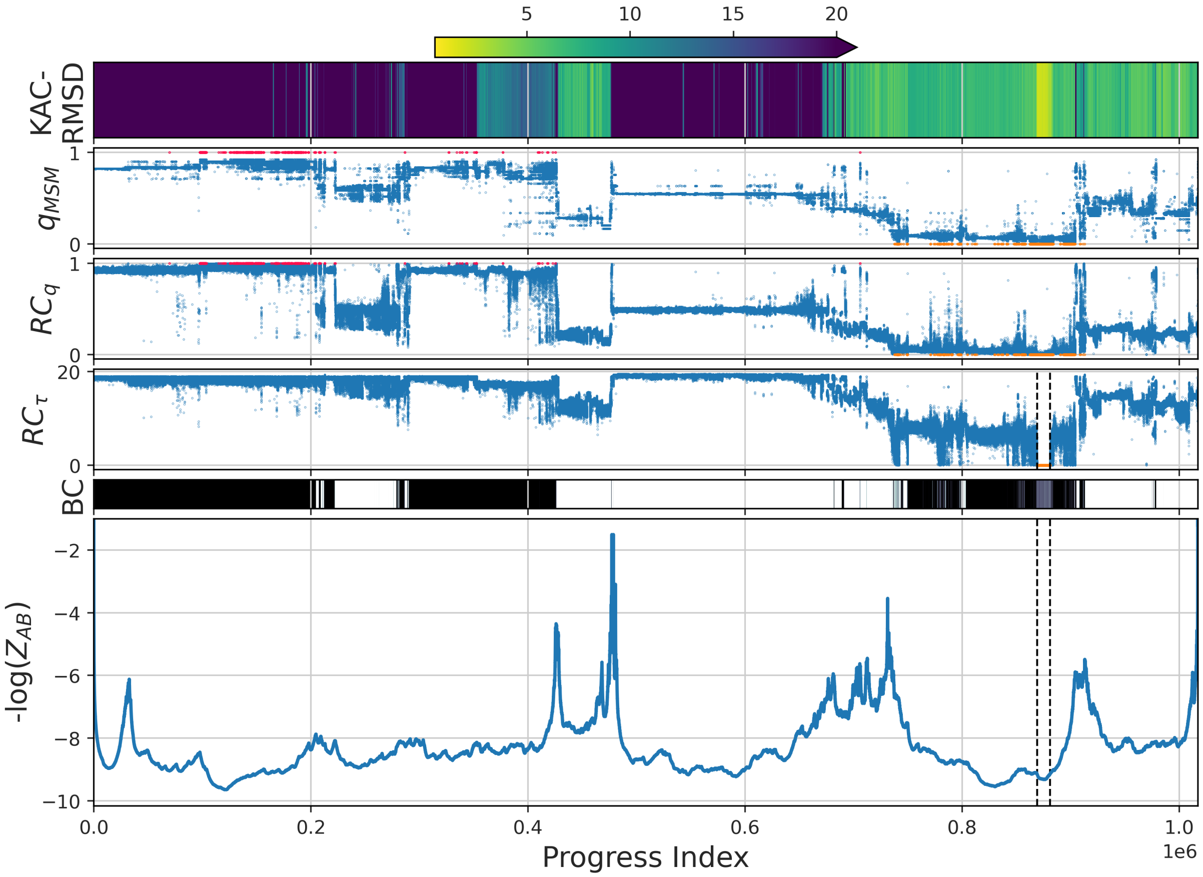

FIG. S10. **SAPPHIRE plot of the ATAD2-peptide system for an alternative definition of boundary states**. This figure is the same as Fig. 7 except that RC_q was calculated based on boundary states identical to the MSM and that its optimization was run only for 750 (rather than 3000) iterations. See main text and Fig. S12 for further information.

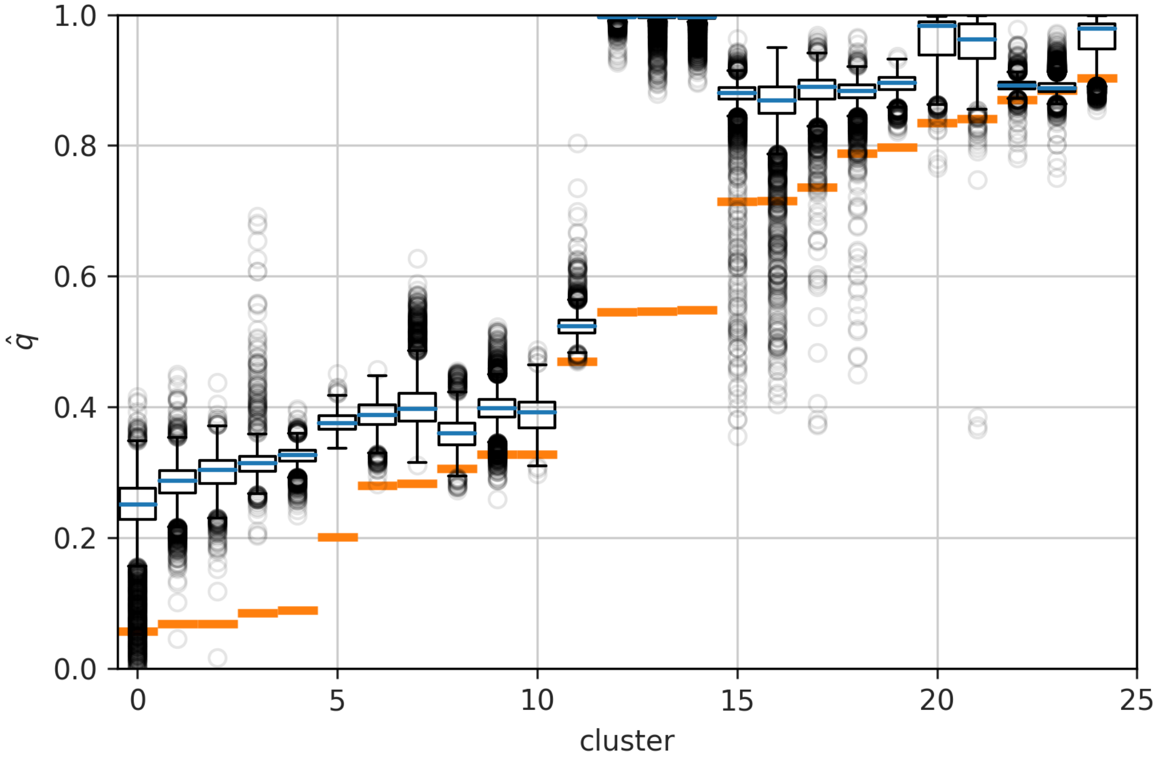

FIG. S11. **Comparison of the RC- and MSM-derived committor estimates at an earlier stage of optimization**. This figure is the same as Fig. 10 except that data correspond to a RC optimized for only 750 iterations instead of 3000 iterations.

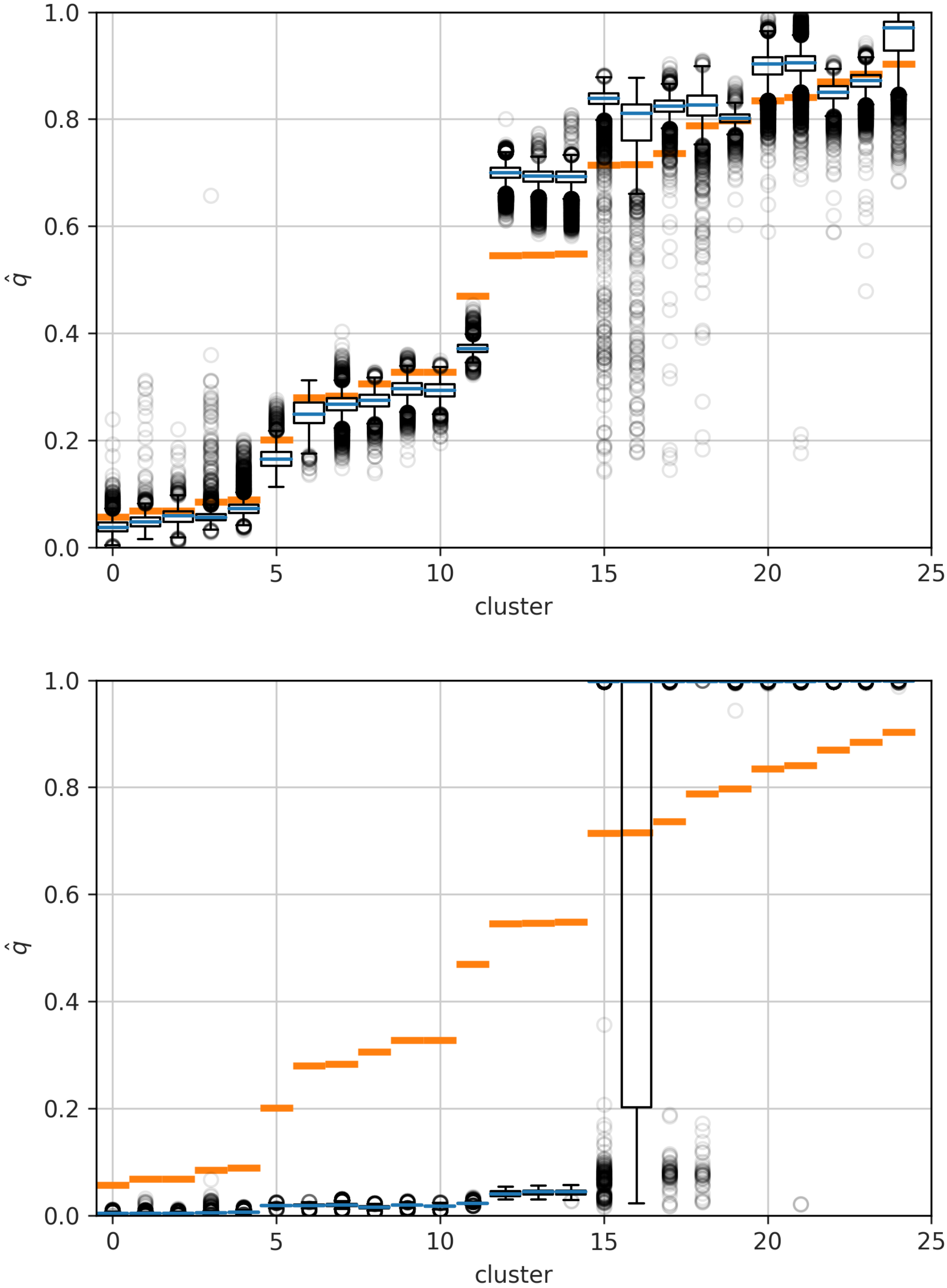

FIG. S12. **Comparison of the RC- and MSM-derived committor estimates when the set of boundary states is chosen identically.** Both panels are the same as Fig. 10 in the main text except that the definition of boundary states for calculating the RC is identical to the MSM and that two different stages of the optimization are shown. Top: Comparison of q_{MSM} with the RC after 750 iterations. Bottom: Comparison of q_{MSM} with the RC after 3000 iterations.

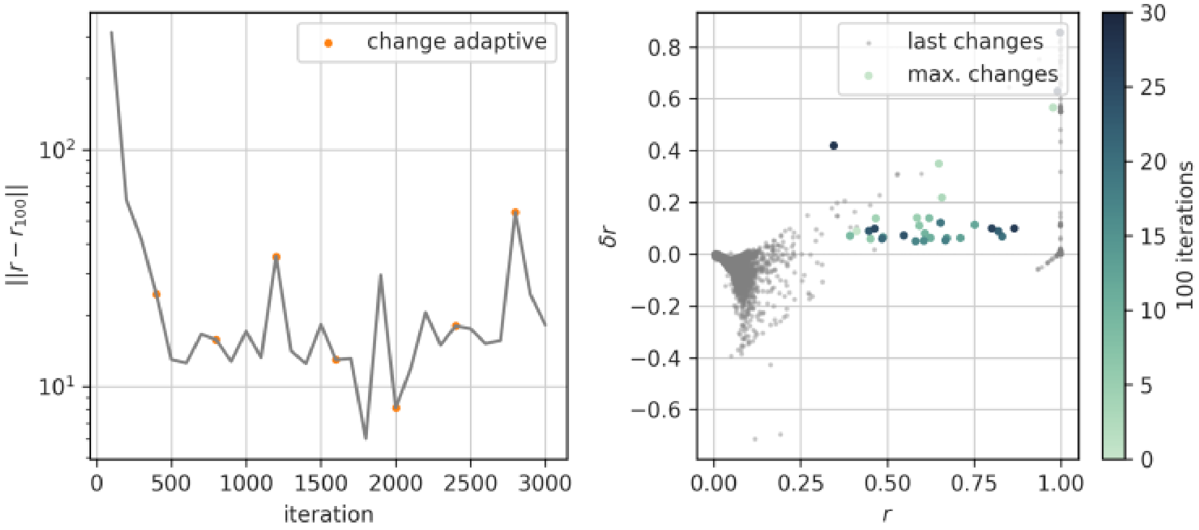

FIG. S13. **Optimization progress for the construction of the RC describing the unbinding of GK*ac*G from ATAD2 with state definitions identical to the MSM**. This figure is completely analogous to Fig. S2. The variability of the RC as optimization progresses (left panel) combined with the large and non-random changes of RC values (right panel), which highlight how the majority of snapshots are pulled towards zero committor value, can serve as a diagnostic for spurious RCs obtained from a breakdown of the optimization procedure. The iteration was stopped at a point where the behavior still seemed as expected (750 iterations, left panel).

Res *i*	**Res *j***
VAL 34	VAL 29
KAC 133	PHE 30
ASP 51	LYS 32
SER 53	LYS 32
KAC 133	LYS 32
ASP 51	PRO 33
SER 53	PRO 33
SER 54	PRO 33
VAL 39	VAL 34
TYR 42	VAL 34
ASP 51	VAL 34
GLY 132	VAL 34
KAC 133	VAL 34
GLY 134	VAL 34
GLU 38	ASP 35
VAL 39	ASP 35
KAC 133	ASP 35
VAL 39	PRO 36
TYR 42	PRO 36
VAL 43	PRO 36
GLY 134	GLU 38
TYR 84	VAL 39
GLY 132	VAL 39
KAC 133	VAL 39
GLY 134	VAL 39
VAL 45	ASP 41
ILE 46	ASP 41
TYR 84	ASP 41
PRO 49	TYR 42
MET 50	TYR 42
ASN 80	TYR 42
KAC 133	TYR 42
LEU 76	ILE 46
LEU 76	PRO 49
TYR 84	SER 79
ASN 85	ASN 80
KAC 133	ALA 81
ILE 95	LEU 82
KAC 133	TYR 84
GLY 134	TYR 84
KAC 133	ASN 85
ARG 96	ASP 87
KAC 133	ILE 95

TABLE SI. **Selected pairs of residues used for the featurization of the ATAD2 bromodomain-peptide complex**. The residue numbering refers to the construct used in the simulations; residues in the range $[1,96]$ are in the bromodomain while Gly132-K*ac*133-Gly134 corresponds to the peptide (see Fig. S1 for a graphical illustration).

[1] M. Bacci, A. Vitalis, and A. Caflisch, "A molecular simulation protocol to avoid sampling redundancy and discover new states," Biochim. Biophys. Acta **1850**, 889–902 (2015).