The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 05 2023

Optimized reaction coordinates for analysis of enhanced sampling

Special Collection: Machine Learning Hits Molecular Simulations

Julian Widmer ; Cassiano Langini ; Andreas Vitalis ; Amedeo Caflisch

J. Chem. Phys. 159, 015101 (2023)
https://doi.org/10.1063/5.0149207

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Optimized reaction coordinates for analysis of enhanced sampling

Cite as: J. Chem. Phys. 159, 015101 (2023); doi: 10.1063/5.0149207
Submitted: 4 March 2023 • Accepted: 13 June 2023 •
Published Online: 5 July 2023

View Online Export Citation CrossMark

Julian Widmer, Cassiano Langini, Andreas Vitalis, and Amedeo Caflisch[a)]

AFFILIATIONS

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

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.

[a)]**Author to whom correspondence should be addressed:** caflisch@bioc.uzh.ch

ABSTRACT

Atomistic simulations of biological processes offer insights at a high level of spatial and temporal resolution, but accelerated sampling is often required for probing timescales of biologically relevant processes. The resulting data need to be statistically reweighted and condensed in a concise yet faithful manner to facilitate interpretation. Here, we provide evidence that a recently proposed approach for the unsupervised determination of optimized reaction coordinate (RC) can be used for both analysis and reweighting of such data. We first show that for a peptide interconverting between helical and collapsed configurations, the optimal RC permits efficient reconstruction of equilibrium properties from enhanced sampling trajectories. Upon RC-reweighting, kinetic rate constants and free energy profiles are in good agreement with values obtained from equilibrium simulations. In a more challenging test, we apply the method to enhanced sampling simulations of the unbinding of an acetylated lysine-containing tripeptide from the bromodomain of ATAD2. The complexity of this system allows us to investigate the strengths and limitations of these RCs. Overall, the findings presented here underline the potential of the unsupervised determination of reaction coordinates and the synergy with orthogonal analysis methods, such as Markov state models and SAPPHIRE analysis.

I. INTRODUCTION

In the pursuit of mechanistic explanation and prediction, molecular dynamics (MD) simulations can provide valuable information by virtue of their high spatial and temporal resolution.[1] Biological systems cover a broad spectrum in terms of length and time scales due to the size of biopolymers, e.g., proteins. With modern hardware and algorithms, it is possible to reliably sample processes that occur on short timescales for small systems. However, slower processes are frequently of particular interest, such as protein folding or molecular binding, for which the available computing resources are insufficient.

Enhanced sampling techniques are frequently applied as a remedy for this issue and exist in many different flavors.[2,3] Generally speaking, accelerating the sampling comes at the cost of introducing additional challenges during analysis. Thermodynamically and kinetically unbiased distributions are the sought-after quantities but are not directly available from enhanced sampling trajectories. Therefore, analyses should incorporate means for removing biases from datasets created by enhanced sampling schemes. Additionally and irrespective of the sampling methodology, the large amount of data—tens of thousands of atomic coordinates at millions of time points—must inevitably be reduced in size to become tractable (reviewed in Ref. 4). One of the dominant aspects of this reduction in dimensionality is to discard degrees of freedom, which are at most weakly coupled to the processes of interest. For example, in MD simulations of biomolecules, the solvent (water) almost always falls into this category. This is mirrored in many experimental observations of processes, such as binding or folding, which rely on probes that are specific to the biopolymer matter (such as a circular dichroism signal) or even just parts of it (such as the intrinsic fluorescence of tryptophan side chains). In addition, experimental signals are usually of low complexity and allow for the fitting of models with at best relatively few parameters, which must be reflected in analyses of simulation data if a comparison to experiment is desired.

Most commonly, one attempts to characterize a biomolecular system in terms of its metastable and transition states.[5] The interconversion between such states can be quantitatively described by rate constants, some of which might be experimentally determinable. Markov state models (MSMs) have emerged as a tool for integrating information from many short trajectories produced by a variety of enhanced sampling methods to yield such simplified descriptions.[6] In this popular approach, a number of modeling choices are imposed that pre-process and thereby reduce the information in the trajectories. Various clustering algorithms on a user-defined feature space lead to implicit criteria for the definition of states.[7,8] To achieve high accuracy, clustering should be very fine,[9,10] which complicates the statistical estimation of the transition matrix[11] and may additionally be detrimental to the interpretability of states, in particular in relation to experiment.

In an alternative approach, reaction coordinates (RCs) can capture information of the underlying system in just a few, or even one, dimension(s). While such an RC can be used for guiding exploration in MD,[12,13] we focus on RCs as a means for the analysis of existing trajectories. The difficulty in both cases lies in their construction, which should ensure that all the desired information from the system's constituents is captured to a maximal extent. This is, by definition, an insufficient condition for finding the best possible (so-called optimal) RC. It is insufficient because "desired information" has not been specified. One criterion, which is cited often but is weak in practice,[11] is to preserve the kinetics of the slowest process or processes in the system.[14] How does one diagnose sub-optimality of RCs in this context? For some classes of dynamical systems,[14] projections onto inappropriately chosen RCs will suggest spuriously fast dynamics in terms of the total squared displacement (TSD) on these RCs,[15] marking them as sub-optimal. In contrast, an "optimal RC" refers to an RC reproducing the slowest possible dynamics, given the reference system. Recently, a variational approach for the construction of such RCs has been proposed.[16,17] It requires minimal user intervention and thus provides a potential benefit relative to the system-specific expertise that is generally required for the construction of appropriate MSMs; see, e.g., Ref. 11. For determining an "optimal" RC according to this approach, the main user input consists in assigning subsets of snapshots to one of two boundary states, A and B.

We next describe the aforementioned way to determine projections on such RCs following Krivov.[14,16–18] The primary scheme is iterative starting from an initial guess, which is largely information-free. At each iteration, a random internal degree of freedom from a relevant space (usually, an interatomic distance) is selected and calculated for all time points in the trajectory (ensemble). The number of degrees of freedom considered can be much larger than what enters the construction of MSMs through clustering. Parameters for a flexible functional form (e.g., a polynomial) incorporate the time series of the internal degree of freedom into the current estimate of the RC such that the conditional TSD along the RC is minimal.[18] Due to the iterative reduction of the TSD, updating the RC in this manner is referred to as optimization in the sense discussed above,[14] and we adopt the same convention here. The associated optimization problem along with its solution is stated in Eqs. (9)–(12) in Ref. 17. The progressive nature of these updates can be understood as a complex composite function that increments the RC gradually in order to include information on slow, collective motions of the system while preserving the self-similarity of RC values for states that are geometrically self-similar. The snapshots making up the boundaries remain unchanged in this procedure; only the transition region is continuously updated such that the free-energy barrier along the RC is increased. The resulting RC can coincide with the committor function for the chosen boundary states,[14] i.e., capture the kinetics of the underlying transition in one dimension exactly.[16,19–22] In other words, for a system where only two states of interest have been chosen, an optimal RC defined on the unit interval with boundaries 0 and 1 should naturally be numerically equivalent to the committor function. This offers a stringent test that we exploit in this work.

More so than in MSMs, summarizing a MD trajectory by a 1D coordinate introduces assumptions and limitations to the processes that can still be described. Naturally, parallel pathways of a particular transition cannot be resolved. Conceptually, in the present approach, it is presumed that two boundary states can be defined meaningfully *a priori*. For ligand binding, for example, it is often natural to use a known crystal structure for defining one of the boundaries. This holds often but not always. If the definition of a boundary is not given directly by the specific problem, it can sometimes be supplied by intuition, orthogonal information on the system, prior analyses, or relevant experimental measurements. One problem with processes that are described as two-state models experimentally is that the "other" (e.g., unfolded, unbound) state is often conformationally heterogeneous. Therefore, it becomes more difficult and subjective to set the criteria defining it. A further conceptual limitation of the original method[16,18] is that the conditional minimization of Δr^2 applies if the dynamics of the original system are Markovian. When dealing with long, complex trajectories, these assumptions can lead to a "hen-egg-problem," where a model is required to comprehend the data, but comprehension of the data (e.g., meaningful state definitions, Markovian time step) is required to construct an accurate model in the first place.

In this contribution, we aim to put Krivov's approach to optimized RCs to a challenging test as follows. We elucidate how this easy-to-construct and largely unsupervised scheme performs in the analysis of challenging, realistic MD datasets (Fig. 1) and how the prerequisites and assumptions outlined above manifest in practice. Specifically, we are interested in what information can be extracted reliably and how sensitive quantitative readouts are to the fact that some assumptions are likely to be met only approximately. We test the compatibility of the unsupervised RC with two enhanced sampling techniques: progress-index guided sampling (PIGS)[23] and replica-exchange molecular dynamics (REX).[24,25] To do so, we assess the consistency of its results with those from established strategies for the analysis of MD trajectories. We also test, where possible, the interpretation of the resulting RC values as committor probabilities explicitly.

Biological processes of interest (e.g., folding or ligand unbinding) are often rare events on the microsecond timescales typically accessible by canonical sampling (CS) and hence call for enhanced sampling techniques. Many methods have been proposed for this task (see Refs. 2 and 3 for a general review and Refs. 26–28 for applications to ligand binding kinetics), which can be loosely classified into two groups: methods that bias the force or energy of the system and adaptive sampling methods. The latter, which include

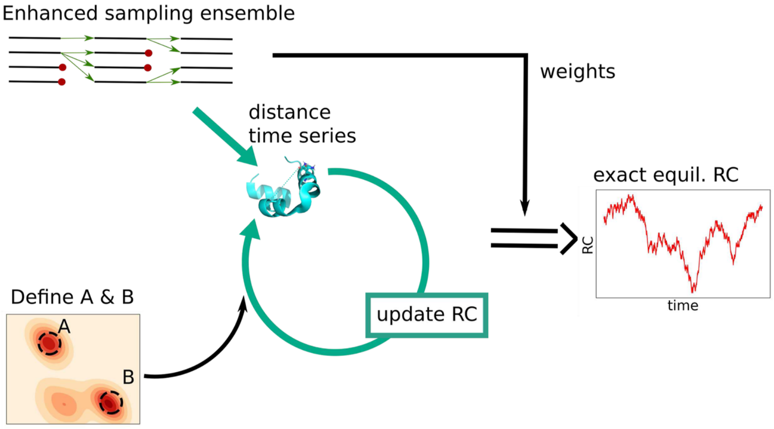

FIG. 1. Schematic of unsupervised RC optimization introduced in Ref. 17 applied to ensembles of enhanced sampling trajectories. Initially, two boundary states are supplied. Then, an RC capturing the exact kinetics is determined iteratively that incorporates information on the underlying system by selecting internal degrees of freedom at random. Using a reweighting scheme permits the extraction of the RC for the equilibrium process. The iterative parts are marked by turquoise arrows. In this work, weights are calculated using the weighted ensemble strategy[11,36] based on a preliminary RC determined using the same procedure, but they can also be calculated independently of the RC, for example, from the stationary distribution of an MSM.

PIGS, usually evolve multiple copies of the same system in parallel while periodically stopping or respawning some replicas based on information collected on the fly. REX is a parallel but not adaptive method that flattens the potential energy landscape through the use of higher-temperature replicas. It generates a multi-canonical ensemble, i.e., it simulates copies of the system at different temperatures and attempts to swap compatible conformations between neighboring temperatures by a well-defined protocol that generally preserves Boltzmann statistics. The sampling enhancement results from free-energy barriers being easier to overcome at higher temperatures. However, if the final analysis concerns a single temperature, the usable data are (effectively) much smaller than the actual sampling. Moreover, the trajectory continuity is broken for a specific condition whenever a swap occurs.

In PIGS, on the other hand, all independent replicas evolve at the same condition for a fixed stretch. The snapshots gathered from all the replicas for a single stretch are reordered according to the progress index (PI),[29] which groups them by geometric similarity. Based on the PI position of the final snapshot of each replica, a heuristic leads to the termination of copies deemed redundant and their replacement with more unique configurations conjectured to promote exploration. PIGS is unsupervised but relies on a user-selected set of features needed to calculate the PI. These features serve to focus the sampling enhancement on regions of the system that are relevant for the process under investigation.[30]

In the first set of results, we focus on data obtained for the FS-peptide (acetyl-$A_5(AAARA)_3A$-N-methylamide), which is predominantly α-helical at equilibrium at low enough temperature. It serves as a useful test because an estimate of the ground truth is available from CS. In the original work,[23] we obtained comparable amounts of sampling for CS, PIGS, and REX, and compared their abilities to drive the exploration of additional states. We also identified the initial state bias present in the PIGS data, which we later analyzed in detail.[11] The availability of CS trajectories capturing the reversible helix–coil transition allows us to assess how well RCs determined from enhanced sampling trajectories allow for inferences on the system's equilibrium properties.

In a second part, the RC is applied to a PIGS dataset of MD simulations of the bromodomain of ATAD2 in complex with a tripeptide containing an acetylated lysine residue (K*ac*). Bromodomains are epigenetic readers that recognize lysine acetylation marks on histone tails.[31] Their fold is well conserved and constituted by a left-handed bundle of four α helices (αZ, αA, αB, αC) connected by two major loops (ZA and BC loop) enclosing the largely hydrophobic K*ac*-binding pocket (see also Fig. S1).[32–34] Both of the aforementioned loops are involved in peptide binding, and the enhancement of their sampling leads to an accelerated unbinding rate of the peptide, which is the process of interest chosen here. The peptide and protein are both flexible enough to make the unbinding challenging to describe using simple geometric RCs; for reference, the properties of the bromodomain in complex with a longer peptide (16 residues) are discussed in Ref. 35. A carefully constructed MSM is used as reference for the analysis of unbinding simulations and, thus, to gauge the potential of the optimized RC on this challenging dataset, which lacks a computational ground truth. Geometric progress variables are used to establish that the process captured by the RC corresponds to an intuitive mechanism for the unbinding of the K*ac*-containing peptide. We further demonstrate the consistency of the two "orthogonal" models while highlighting inherent properties, advantages, and limitations of the RC applied to simulations of protein complexes.

For additional information on the methodologies we rely on in this work, we refer the reader to the original publications (see also Table I), which are too extensive to recapitulate in detail here.

TABLE I. Overview over the methodologies referred to in this study.

Acronym	Full name	Type	Summary
CS	Canonical sampling	Sampling	Brute-force MD
REX[25]	Replica exchange	Sampling	Use elevated temperature to flatten potential energy surface
PI[29]	Progress index	Analysis	Re-index snapshots by mutual, geometric similarity
PIGS[23]	Progress-index guided sampling	Sampling	Duplicate unique replicas according to PI to promote exploration
WE[36]	Weighted ensemble	Analysis	Recover equilibrium weights from biased trajectory ensembles
SAPPHIRE[37]	States and pathways projected with HIgh REsolution	Analysis	PI visualized with various annotations.

Relevant methods not yet touched upon are the weighted ensemble approach to statistical reweighting,[11,36] see Sec. IV B, and the SAPPHIRE analysis,[37] see Sec. IV D 2. In the Introduction, we have provided brief, qualitative descriptions of enhanced sampling techniques REX[25] and PIGS,[23] as well as Krivov's approach to the optimization of reaction coordinates.[16,18]

II. RESULTS

A. Conformational transition of the FS-peptide

The 21-residue FS-peptide undergoes a well-defined helix-coil transition as a function of temperature.[38] The ABSINTH implicit solvent model and force field paradigm[39] have been shown not only to reproduce this transition comparatively well but also to sample a diverse ensemble of partially helical and non-helical but collapsed states at low sampling temperature.[23,40] This ensemble reflects an underlying free energy landscape that is complex but possesses relatively low barriers. The latter property enabled the sampling of transitions in and out of the dominant state, the straight α-helix, with high fidelity with and without the use of advanced sampling methods. PIGS allowed us to identify a particular, low-likelihood, metastable, and non-helical state that we showed to be metastable in canonical sampling (CS) as well. The similarity of results obtained in simulations from two diametrically opposed states gave us the confidence to assert that, in the chosen model, the sampler achieved equilibrium.

Here, we first recognize that this dataset gives us access to robust references for comparison. In particular, direct estimates from CS are considered reliable. Since all three samplers we employed in the study (PIGS, REX, and CS) cover transitions from the fully helical configurations into helix-free, collapsed states, it, furthermore, appeared reasonable to define reference states that approximately describe these two limits. We were able to utilize the coordinate RMSD from the straight helix for this purpose since this conformation is relatively unique in shape and size within the ensemble. It is, however, a caveat that using a high RMSD threshold to define a boundary state as done here (see Sec. IV A) will simply map to a degenerate ensemble of possible candidate structures.

B. The RC achieves efficient kinetic sorting

In the following analysis, we rely on all of the CS and the PIGS sampling carried out at 250 K. For REX, only the trajectories at the relevant temperature of 250 K were considered for the construction of the RC. First, we seek to establish whether an optimized RC meaningfully describes the transition of the FS-peptide that we meant to analyze. If so, the RC should unveil the same underlying process for each of the three samplers, CS, REX, and PIGS, when supplied with identical definitions for boundary states. Furthermore, this process should offer a reasonable succession of states by RC value, even though it must be kept in mind that the RC-sorted snapshots are not a pathway. The residue-wise helicity and the RMSD from the straight helix will serve as annotations by virtue of being imperfect approximations to progress coordinates. Despite their sub-optimality,[14] these geometric descriptors can be expected to capture some key aspects of the underlying process.

Obtaining a reasonably converged RC is computationally feasible. The 3000 iterations detailed in Fig. S2 took about 2 h to run on a modern desktop with a 6-core, Intel i7-8700 CPU while processing >6M snapshots. It is a downside, discussed later, that the numerical optimization procedure does not permit an *a priori* stopping criterion but requires manual intervention. Nonetheless, the results in Fig. 2 highlight that the optimized RC largely meets the expectations posed to it for PIGS and CS: for low values of the RC, the RMSD to boundary state A (the straight helix) is low, while the helix content of residues is high. The helix content of C-terminal residues is lost gradually along the RC until an intermediate state is reached: Near $r = 0.85$, the two termini each form a helix with a bend in the middle. The alanine side chains form a hydrophobic core, while the arginine side chains remain solvent-exposed. Subsequently, helicity is lost entirely, although with low sampling weight, in favor of collapsed configurations near and within boundary state B. As shown in Fig. S3, the mapping of RC value to structure is generally well-defined although many of the non-helical residues in Fig. 2 are transiently helical, indicating the presence of kinetically but not structurally fully homogeneous states. This is expected for the relatively short FS-peptide since helical stretches can easily fluctuate in length.[40]

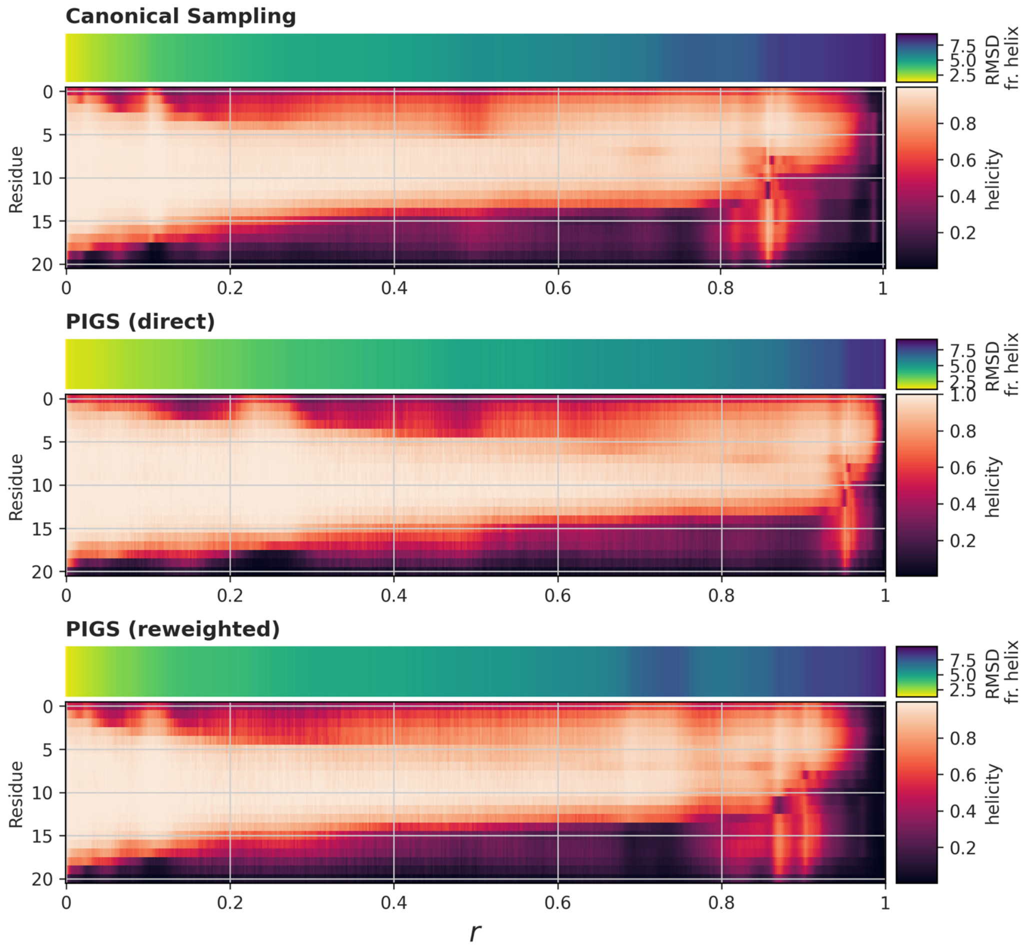

FIG. 2. Kinetic progress according to the optimized RC for FS-peptide. Trajectories from CS and PIGS are binned (binsize: 0.002) along the RC. The bin-averaged RMSD from the straight α-helix (top of each sub-panel) and the residue-wise helicity (bottom) are depicted in color (legends on the right). Color-coded measures for the heterogeneity within bins are given in Fig. S3.

This process, as characterized by the overall helicity and the RMSD is shared between CS, PIGS (Fig. 2), and also REX (Fig. S4). We note that the absolute RC values appear distorted for REX, for which there is no unequivocal explanation. REX data are challenging for the optimization since the continuous trajectory segments are very short if the swap probability is reasonably high (here, the average is only about 42 snapshots compared to PIGS with nearly 8000 snapshots), and the coverage of relevant conformational transitions by the data is unclear. This segment length is also clearly at odds with the autocorrelation lengths of the geometric features that are used to construct the RC (see Fig. S5, top).

The geometric properties suggest that the automatically determined RCs smoothly reflect the transition of the FS-peptide from a straight α-helical configuration to collapsed, non-helical configurations. This is true and consistent for all three samplers despite the fact that the RCs share only the definitions of the boundary states as inputs.

Next, we demonstrate how the optimized RC can be employed in reweighting ensembles obtained from adaptive sampling, which suffer from initial state bias. We proposed previously that the idea of statistical resampling and its implementation as the weighted ensemble (WE) scheme is the most promising strategy for this reweighting task.[11] In short, the goal is to recover equilibrium sampling weights from PIGS trajectories in post-processing. The method's performance is contingent on a reasonable guess of kinetic distance between replicas. Because the statistical weight of a terminated replica is lumped into that of one or more other replicas, errors will be accumulated if not at least one replica that can be considered nearby is available. Specifically, this idea of "kinetic proximity" implies that interconversion between their respective configurations at the point of termination is rapid. Normally, geometric similarity is used as a proxy, but we suggest here that the optimized RC offers an alternative sorting principle to find replicas corresponding to such kinetically nearby conformations. In Fig. 3 [(b) and (d)], the extracted, snapshot-wise weights are used to reweight the distribution of the radius of gyration, R_g, and the α-content. We observed intuitive correspondence between the RMSD from the straight helix and the α-content of snapshots sorted by the RC in Fig. 2. At a more quantitative level, the Kullback–Leibler divergence (KLD) reported in Fig. 3 measures the information lost when distributions are calculated from reweighting PIGS trajectories compared to CS. The low values indicate good agreement and suggest that the RC efficiently selects kinetically close replicas, more so than geometric descriptors can [(a) and (c) of Fig. 3]. The two samplers, CS and PIGS, propagated copies of the system entirely independently, and some minor differences in sampling are therefore unavoidable. For systems sampling perfectly parallel reaction channels, the RC alone would not be sufficient for ranking replicas according to kinetic distance. It is straightforward to include further descriptors of the system that are able to discern those channels for ranking the replicas. The distance between replicas is then calculated on a feature space with more than one dimension. In addition to introducing the RC as a sorting principle, we also amend here the original strategy by allowing the weight of a terminated trajectory to be lumped into multiple target trajectories, weighted by distances (see Sec. IV B and Fig. S6). Thus, we conclude that the RC in combination with the WE strategy allows for an accurate reconstitution of equilibrium distributions from PIGS sampling, which, as shown previously, is an exceptionally challenging task for other methods, most prominently Markov state models.[11]

C. The RC as an estimator for the committor

As touched upon in Sec. I, optimized RCs of this type are proposed to approximate a Markov chain's committor function,[16] which preserves the kinetics of a potentially high-dimensional system as a one-dimensional coordinate.[22] We, therefore, consider here the RC as a statistical estimator for the committor. This allows for

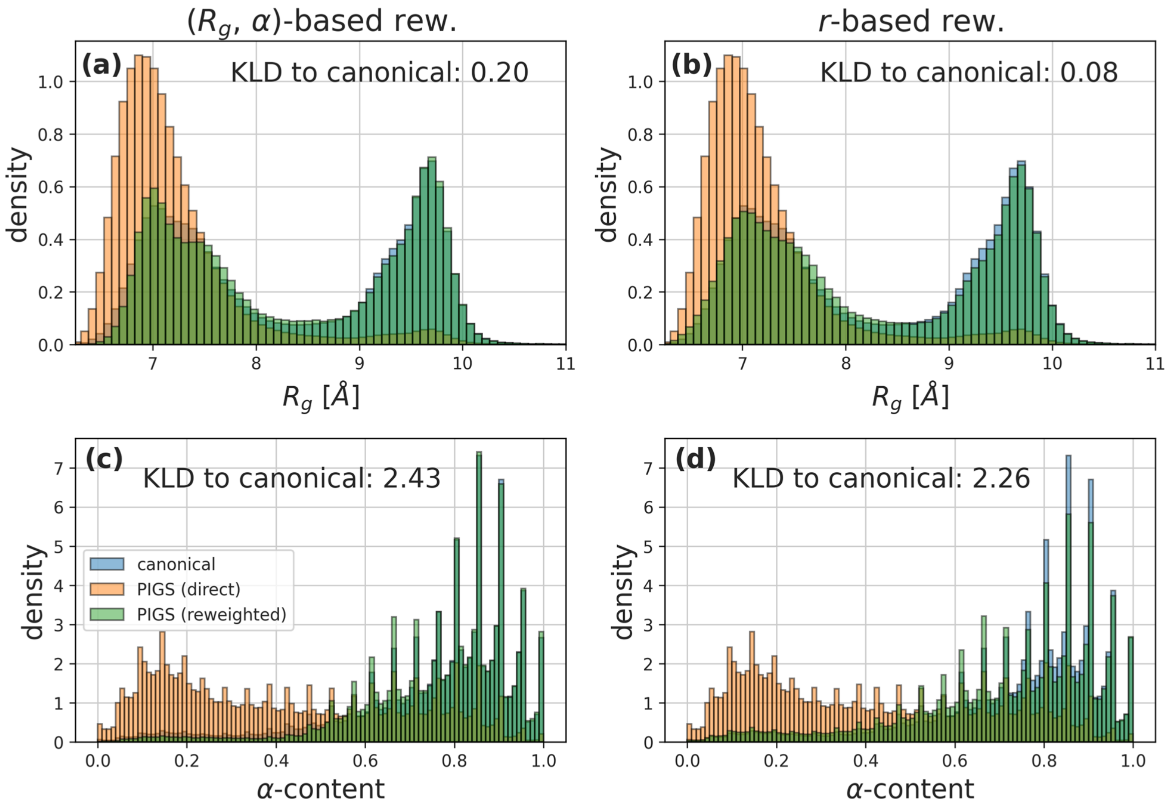

FIG. 3. Reweighting of PIGS data to remove initial state bias. The distributions for R_g and the α-content of FS-peptide are compared between CS and PIGS. The inset lists KLD values that quantify the distance of the reweighted PIGS distribution to the canonical reference. The KLDs between unweighted PIGS and CS are 11.3 for the R_g and 81.3 for the α-content. The left column [(a) and (c)] uses ($R_g \times \alpha$) as a proxy for kinetic distance as in Ref. 11, while the right column [(b) and (d)] shows results for using the RC to encode kinetic distance. The α-content was computed as in Ref. 23.

comparisons to two references. First, direct counting yields the maximum likelihood estimate (MLE) of the same quantity: We simply enumerate and normalize the fraction of transition paths initiated in a small range around an RC value that reach state B before state A for CS.[41] Figure 4 reveals that the optimized RC is perfectly rank-correlated with the MLE but that it overestimates the MLE over a large range of intermediate RC values. Agreement is best near both boundaries.

Since both quantities (MLE and RC) derive from the same data, we attempted to further assess the generalizability of the RC. To this end, the RCs for the CS and PIGS ensembles (cf. Fig. 2) were binned in the range $[0.025, 0.975]$. New trajectories were computed by launching independent simulations from 1024 unique starting structures for each bin. The MLE was computed by tracking the fraction of the 1024 that committed to state B before A. The number of trajectories that did not commit to either state over the simulation length of 30 ns was below 1%. Following the training paradigm in machine learning, these trajectories are test data. For the RC describing CS, MLEs of the committor for the training trajectories are close to MLEs for unseen trajectories for RC values smaller than 0.7. Indeed, various non-helical configurations (i.e., configurations in the vicinity of boundary B of this study) are not visited by CS in the training simulations.[23]

The MLE for PIGS follows a similar trend as the MLE for CS trajectories. The deviations from the MLE become more pronounced at higher values of the RC while, with one exception, preserving the ranking. This might indicate an incomplete reweighting, similar to what was generally observed for this system in prior work,[11] but the definite reasons remain elusive.

In summary, the near-perfect rank-correlation with the MLE on the training set trajectories corroborates the finding that the optimized RC is a reliable tool for ranking intermediate snapshots by their kinetic proximity to the chosen boundary states. In contrast, RC values are unreliable for the estimation of absolute commitment probabilities. As the relaunched trajectories exceed the sampling time of the training set roughly 30-fold (~600 μs instead of 20 μs), some discrepancy is expected, in particular for the restarts. However, not only do the training data show similar deviations but, more importantly, most of the deviations are systematic: they are stronger in sparsely populated (transition) regions of the RC (see also Fig. 5), and the RC values indicate that snapshots are further away from folded state A than they actually are (RC values exceed $\hat{q}_{MLE}$ for the test set consistently). We suspect that this is due to the algorithmic minimization of the TSD, which implies a focus on high-population regions on the RC.[18] Sampling quality in the original data is less of a concern, which is evident from the fact that the MLEs for training and test sets for CS agree much more with each other than with the RC.

D. The reweighted RC accurately estimates kinetics from PIGS-trajectories

While geometric descriptors indicate that the optimized RCs sort snapshots kinetically in a mutually consistent manner, the underlying state probabilities are not consistent between CS and PIGS if the weights are not taken into account (Figs. 3 and 5). By virtue of penalizing the redundancy of sampling self-similar,

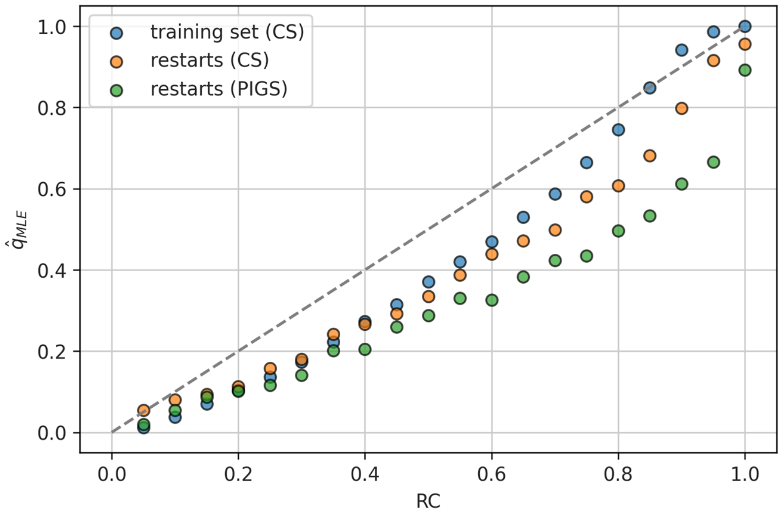

FIG. 4. The maximum likelihood estimate (MLE) of the committor plotted against the RC. In all cases, the y axis indicates MLE values, $\hat{q}_{MLE}$, estimated as binomial probabilities by direct counting (see the text). The "training set" denotes the set of CS trajectories analyzed in Fig. 2. From 19 intervals along the RC (binsize: 0.05), 1024 unique configurations each were extracted to start a new set of 30 ns-simulations, which are labeled "restarts." Trajectories that did not commit are too few to matter. The errors in the estimation of the binomial probabilities would be negligible for the restarts if we considered the trajectories as independent samples (which they are not since the 1024 configurations are partially time-correlated). The restarts from PIGS snapshots were redistributed according to their reweighted RC values (Fig. 2, bottom), which still left at least 500 committing trajectories per bin. The gray dashed line of slope 1 is added as a visual reference.

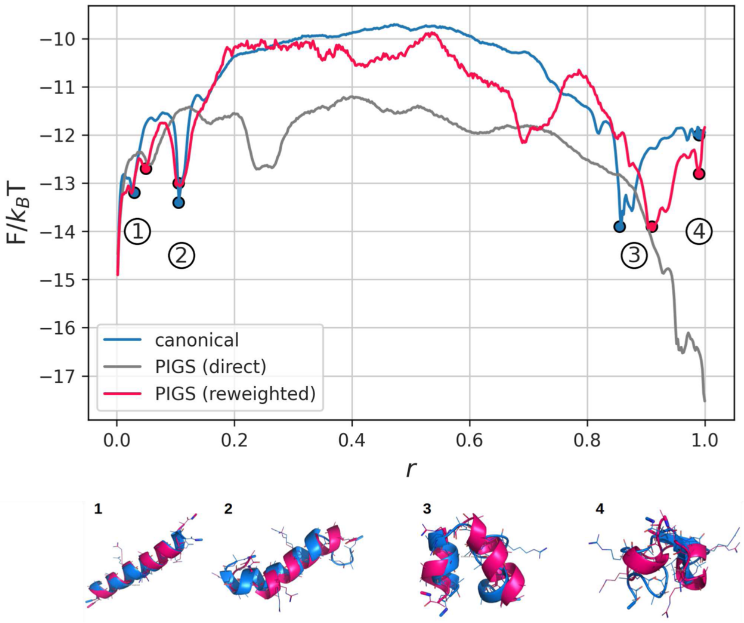

FIG. 5. Free-energy-profile for the RC describing the transition between the straight α-helix and the collapsed state for FS-peptide. Top: The reaction coordinate values were binned, and the histogram-based free energy profile is shown for CS, PIGS, and PIGS after including the WE correction. The weights were first derived from a preliminary RC and subsequently used during optimization, with the final RC providing the basis for the weights applied to the histogram (see Sec. IV B for details). Bottom: Pictorial representation of the process captured by the RC for FS-peptide trajectories. From the numbered circles marked on the top panel, one structure each for PIGS and CS is shown jointly after alignment (red: PIGS, blue: CS).

helix-rich states, PIGS effectively promotes the visitation of the non-helical boundary state. This results in an overpopulation of such conformations in the RC derived from PIGS trajectories compared to CS and, by extension, an overestimation of the rate from A to B and, possibly, and underestimation of the rate from B to A. Applying the WE-weights to weight Δr^2 during optimization[17] controls the absolute value of the RC for a given snapshot. Applying weights in the calculation of the histogram-based free-energy profile (FEP) serves to recover the equilibrium population in essentially the same way as in Fig. 3. We find that, when both strategies are applied to PIGS of the FS-peptide (Fig. 5), not only the relative population of boundaries and intermediates is recovered but also the main barrier height of ~5 k_BT is reconstituted to very good agreement. The main discrepancy between the CS and PIGS data is a state around $r = 0.7$ where the N-terminal half of the peptide is helical and the C-terminus is more variable. We did not investigate the origin of this disagreement further, but the presence of additional states that are metastable in PIGS but are not or only transiently seen in CS is not *per se* surprising.

In principle, it is achievable that the RC contains accurate kinetic information of the underlying system, given the selected boundary states. The optimization can result in the RC coinciding with the committor function q, which has been referred to as the optimal RC.[18] As can be gleaned from Fig. 2 and Fig. S2, the values of the RC can be subject to continuous change over the course of optimization. Consequently, the absolute values of the RC are not generally stable. Since the committor function has to, by definition, associate fixed values with specific configurations of the system, the interpretation of the RC as the committor cannot hold universally but only incidentally (compare Fig. 4). Therefore, we wondered to what extent the reweighting gives access to quantitatively accurate estimates of mean first-passage times (MFPTs). If we do assume that the optimized r is a reasonable approximation to q, then the MFPT τ and the mean transition-path time τ_{TPT} can be calculated for both directions as[42]

$$\tau_{AB} = \langle 1 - r \rangle / J_{AB}^{eq}(r), \tag{1a}$$

$$\tau_{BA} = \langle r \rangle / J_{AB}^{eq}(r), \tag{1b}$$

$$\tau_{TPT} = \langle r(1 - r) \rangle / J_{AB}^{eq}(r). \tag{1c}$$

The equilibrium flux J_{AB}^{eq} between boundaries is estimated by numerically integrating the cut function, $Z_{C,1}(r)$, introduced in Ref. 43. CS trajectories are considered the ground truth, which allows kinetic quantities to be estimated by direct counting. This is again contingent upon choosing boundary states via simple geometric descriptors, which is a limitation, but does not depend either on theoretical arguments or on any reaction coordinate. MFPT estimates based on the optimized RCs are compared to this reference in Fig. 6.

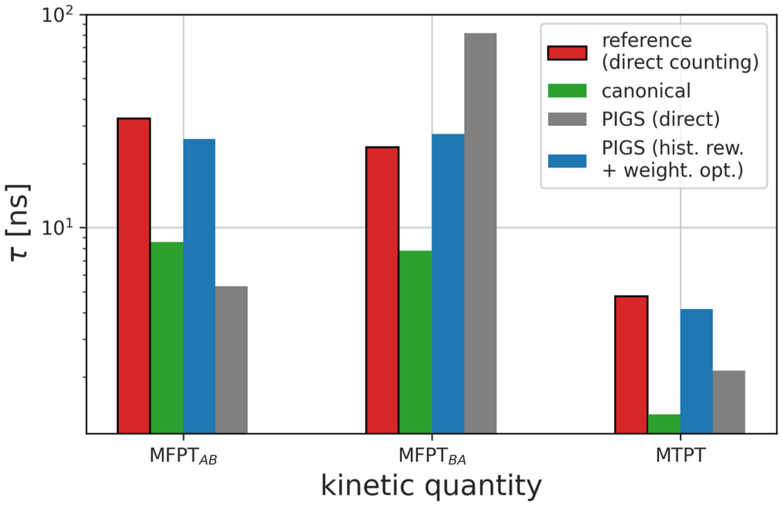

FIG. 6. Kinetic quantities for the transition between the straight helix (state A) and the collapsed state (state B) for FS-peptide. Mean first-passage times and mean transition path times are estimated based on the RC using Eqs. (1). Direct counting on the ensemble of canonical trajectories was used to derive the reference (ground truth) values shown.

Applying Eqs. (1) to the RC for CS, the model underestimates the reference values. One possible explanation suggested by Krivov[18] is insufficient optimization of the RC. The disagreement between MFPTs for CS and PIGS stems from a differing estimate for the flux J_{AB}^{eq} (direct counting: 0.037 ns^{-1}; CS: 0.061 ns^{-1}; PIGS: 0.019 ns^{-1}). Indeed, when setting the flux to the same value, model predictions are within 1% of one another. It appears that the equilibrium flux for PIGS is, despite applying weights during the optimization, underestimated, which compensates for the sub-optimality of the RCs.[14]

It must be noted that the RC does not carry explicit information about transition paths for (un-)folding; it merely encodes kinetic distance to the boundaries in one dimension. If the source data contain (continuous) reactive trajectories, as is the case especially for CS, such transition paths can be isolated and labeled with both the RC and geometric properties. Similarly, it is not straightforward to study faster processes with the same RC or to obtain an interpretable RC for a process that is clearly not the slowest mode in the system. The scenario posed by the FS-peptide given our choice of boundary states is amenable to the method: the largest barrier is not substantially affected by the many faster transitions in and out of intermediate, metastable states. If we forcefully violate these conditions, for example, by choosing the two boundary states to be more or less the same kinetically, the RC will too no longer describe any meaningful process (Fig. S7). This is consistent with the logic of finding the slowest path between the two states, which is now spurious.

We thus stipulate that in a favorable scenario on a realistic, complex system, the unsupervised determination of RCs can (i) achieve efficient kinetic sorting for enhanced sampling trajectories, (ii) gives access to an RC for the equilibrium process (free of initial state bias) when combined with reweighting, and (iii) allows for the estimation of MFPTs to within an order-of-magnitude or better for both PIGS and CS trajectories.

E. The RC captures the unbinding process of a K*ac*-containing peptide from ATAD2

The complex of capped GK*ac*G with the ATAD2 bromodomain was sampled using PIGS with no ground truth reference available. 24 968 short trajectories were generated according to two diversification schemes focused on either the ZA loop or the BC loop of the bromodomain (see Sec. IV D and Ref. 30). We shall refer to these two subsets as BC PIGS and ZA PIGS, respectively. This system constitutes a challenging test for unsupervised RC optimization.

Due to the large amount of encounter complex-like configurations, the RMSD is not reliable for defining a clearly delineated bound state. When choosing an RMSD-based cutoff tightly, many configurations that are visually very close to the initial state are not contained, which is hard to justify. A more loose cutoff leads to the inclusion of partially detached structures. We note that such problematic boundary definitions do not necessarily interfere with the construction of an RC; instead, they lead to an RC for a process with poorly defined meaning. As shown for FS-peptide (Fig. S7), in the extreme case of boundary states chosen deliberately to derive from the same kinetic basin, virtually all intermediate points collapse at $r = 0.5$. To prevent issues of this type, we defined the bound state for the unbinding problem with a more sophisticated criterion based on an independent analysis. Specifically, we selected continuous snapshots from a part of a SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot,[37] which relies on the progress index (PI).[29] In short, geometrically similar snapshots are adjacent to one

12 July 2023 10:54:26

another on the PI (see Sec. IV D 2 for details), which was exploited for the definition of the bound state (Fig. 7, vertical dashed lines). State B is defined in terms of the ligand RMSD (>25 Å). While similar definitions, e.g., in terms of specific interatomic distances, proved to have little impact on the resulting RC, we acknowledge that problems might arise from this issue and further explore the topic below.

With the boundary states in place, the optimization for this complex PIGS dataset proceeded similarly to the one for FS-peptide by convergence indicators; see Fig. S8. From 2000 distances between randomly chosen heavy atom pairs (which are combinatorially predominantly intra-protein distances), distances between the ligand and the protein were most informative on the RC resulting from this optimization as measured by their mutual information (MI) (Fig. 8). In addition, the RMSD of the peptide from the initial pose suggests that the RC represents the progression between boundary states in a largely systematic and anticipated manner (Fig. 9).

The observable intermediate states on the FEP in Fig. 9 are almost exclusively found in either BC PIGS (around $r = 0.25$, orange shading) or ZA PIGS (r between 0.42 and 0.75, green and red shading) data but not both. This suggests that the different diversification schemes of PIGS lead to the sampling of (partially) unique states (as was already observed in Ref. 30 for the bromodomains alone), which hint at the presence of multiple unbinding pathways explored by PIGS.

It has been noted that the ZA loop is prone to assuming disordered states enabling a plethora of possible binding poses.[35,44–46] Indeed, this is also apparent in unbinding simulations. The RC in Fig. 9 suggests that basin 1 (orange) is kinetically closer to the bound state with a closed ZA loop (basin 0, blue shading) than to the unbound state (basin 5). The ligand, on the other hand, is already subject to more variability as the hydrogen bond between inserted, modified Lys, and the conserved Asn85 is broken, which may be a result of the diversified BC loop. Basins 2 and 3, composed of fuzzy encounter complexes,[47] can be interpreted to be the main intermediates that permit recognition of acetylated histone tails. The flexibility of the ZA loop allows for a wide range of poses for the peptide, in many of which the contact between K*ac* and Asn85 is intact or close to intact, especially in basin 2. Basins 1, 2,

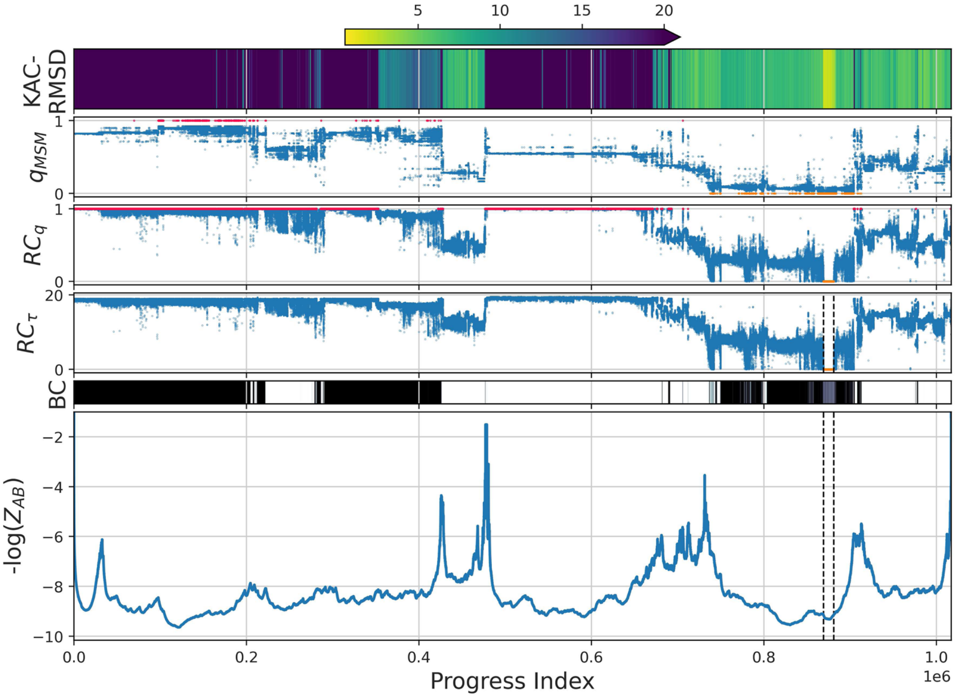

FIG. 7. SAPPHIRE plot of PIGS data on ATAD2 in complex with capped GK*ac*G. Various data values are sorted by the progress index, which groups snapshots according to mutual conformational similarity. Bottom: Kinetic annotation of the PI based on the number of transitions between snapshots to the left and snapshots to the right of a given progress index. This annotation is the equivalent of a (qualitative) FEP, and peaks tend to delineate geometrically homogeneous regions, given the featurization (see Sec. IV D 2 for details). The bound state A, which was used for RC optimization, is emphasized by vertical dashed lines. Top: The RMSD of the ligand to the initial structure in Å is color-coded (legend on top). Directly below are three comparable annotations: the committor estimated from an MSM (see Sec. IV D 3), q_{MSM}; the RC computed for two boundary states, RC_q; and the RC computed with only one boundary state (the bound form) imposed, RC_τ. The RC_τ is given in units of 1000 time steps of 1.5 ps each. In each case, snapshots declared as bound (state A, $r = 0$) are colored orange, while snapshots declared as unbound (state B, $r = 1$) are colored red. Finally, the "BC" annotation indicates in black that snapshots originate from BC PIGS and in white that snapshots originate from ZA PIGS.

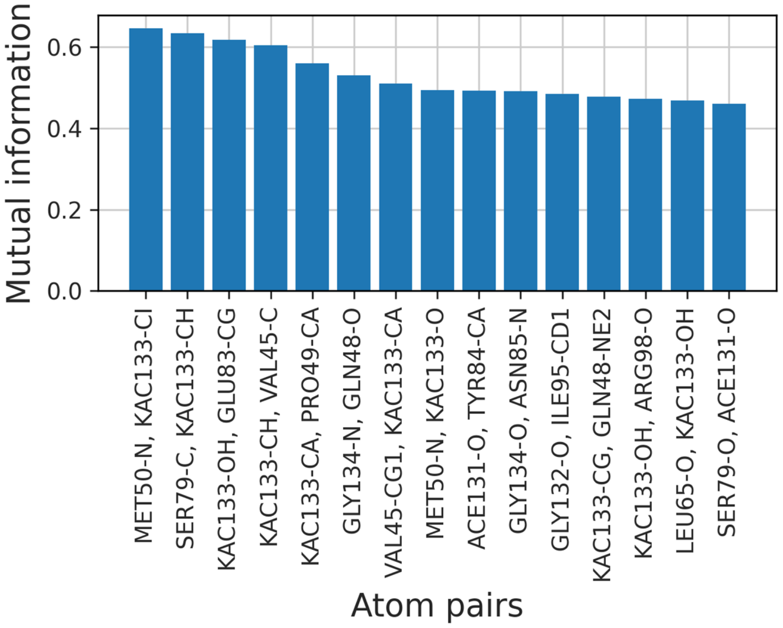

FIG. 8. Correlation of individual features with the RC for the GK*ac*G-ATAD2 system. The bar plot shows the 15 interatomic distances possessing the highest mutual information content with the RC for the unbinding of the peptide from ATAD2. These distances are from a set of 2000 randomly selected trial distances between heavy atoms that were drawn independently of the distances used in the construction of the RC.

and 3 are distinguished by barriers of ~$2k_BT$ each; in both basins 2 and 3, the peptide is typically no longer close to the crystal binding pose, but the latter basin features a higher degree of opening of the ZA loop. Basin 4 around $r = 0.9$ (violet shading) consists of encounter complexes with a closed ZA loop that precludes contact of K*ac* with the hydrophobic contact area lining the binding pocket; the peptide instead typically associates with the outside of the pocket. Finally, basin 5 collects structures, where the peptide ligand is (kinetically) very close to the fully unbound state and (geometrically) distant to the binding site. In this overall picture, it is tempting to associate basin 1 with largely unproductive, partial unbinding events (rebinding necessary for unbinding) and basin 4 with largely unproductive encounters (unbinding necessary for binding), but the RC itself unfortunately does not provide such pathway information.

Taken together with the MI analysis in Fig. 8, Fig. 9 suggests that unsupervised RC optimization achieves a meaningful kinetic ranking for a complex system sampled using a sophisticated enhanced sampling scheme. The imperfect match with the RMSD of the ligand highlights that simple geometric descriptors, such as the RMSD, are, in many cases, deficient in capturing kinetic distance, which is known and expected.[14] Somewhat surprisingly and differently from FS-peptide (Fig. 5), efforts to reweight the FEP toward equilibrium produced only minor changes; see Fig. S9. This hints either at a lack of flux imbalance in the PIGS datasets or a breakdown of the reweighting procedure, e.g., because "nearby" replicas for absorbing the weight are generally all too distant. Since there is no ground truth available, we did not investigate this phenomenon further.

F. The RC is consistent with the committor from an MSM

MSMs are an established tool for analyzing ensembles of short trajectories but require many (hyper)parameter choices in construction. Given two boundary states, the committor for the transition between the two is one of the most useful properties that can be estimated from an MSM. In contrast, the RC aims to approximate a conceptually identical committor with minimal user intervention. It is therefore appropriate to compare findings derived from the (unsupervised) RC with those from a carefully constructed MSM. No meta-information is shared between the two analyses such that they can be considered independent beyond using the same source data.

Figure 7 shows the RC and the estimate of the committor from the MSM, q_{MSM}, sorted by the PI. Visually, close correspondence between the two independently determined coordinates, q_{MSM} and RC_q, can be established. In addition, geometrically homogeneous sections on the PI tend to have self-similar values in both coordinates, i.e., they are considered kinetically homogeneous by both approaches as expected. For example, the geometric basin around the bound state collects a large number of snapshots, mainly from BC PIGS. This basin has local structure that is, however, only weakly resolved by the kinetic annotation of the PI. This structure is more

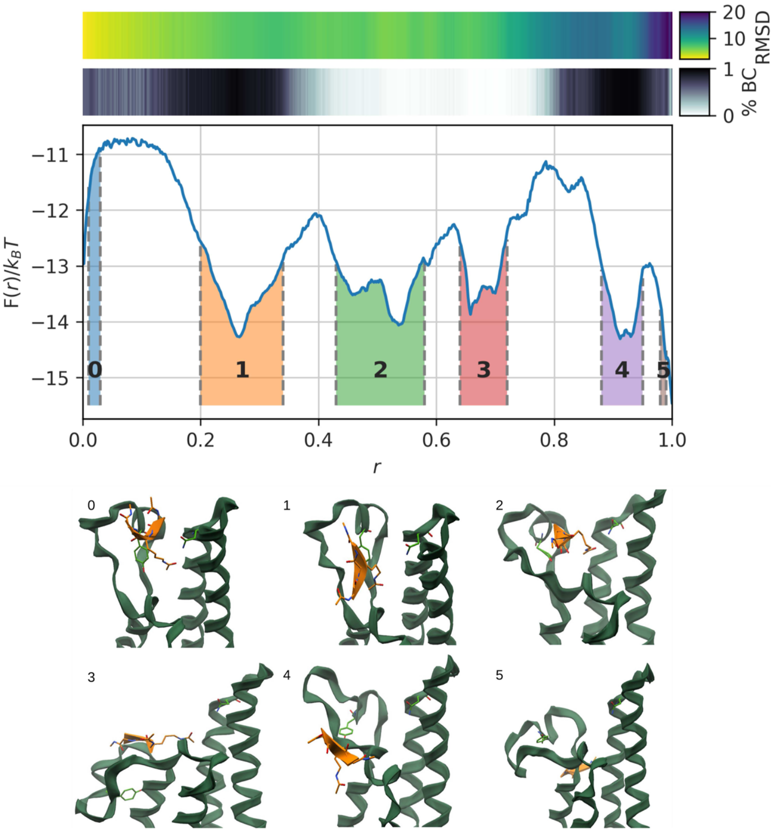

FIG. 9. Reaction coordinate for the unbinding of capped GK*ac*G from ATAD2 with annotations. The top bar color-codes the RMSD of the peptide to the initial structure with the value of the RC shown below. The second bar indicates the fraction of snapshots originating from BC PIGS in each bin (the remainder coming from ZA PIGS). The main subpanel contains a histogram-based free energy profile on the RC describing the dissociation process. Below, the numbered cartoons depict hand-picked, representative structures from the basins marked by numbers and colors on the FEP. The K*ac*-containing ligand is shown in orange, and the ATAD2 bromodomain is in green. The conserved residues Tyr42 and Asn85 of ATAD2 are shown in stick representation in lighter green. The protein is oriented such that the ZA loop (containing Tyr42) is on the left and the BC loop (containing Asn85) is on the right.

evident from the RC_q values within the basin, which highlight that a large fraction of snapshots is kinetically quite distant from the actual bound state, an observation mirrored largely in the RMSD values. On the other hand, this collection of states is clearly closer to the bound form than the non-specific, encounter complex-like configurations collected primarily from ZA PIGS at PI > 900 000 and PI between 420 000 and 440 000. These states feature similar RMSD values but are distinguished from the crystal-like basin by both committor values and the PI.

To further corroborate the interpretability of RC_q, an RC inspired by the MFPT [Ref. 17, SI, Eqs. (26) and (27)], RC_τ, for which only the ligand-bound form is used as a boundary state, was constructed. This quantity is strongly correlated with RC_q (Pearson: 0.98; Spearman: 0.93), supporting the notion that the choice of boundary states and the constructed RC_q capture the process of interest. While this supports the idea that the unbound state defined via its RMSD is kinetically most distant from the complex, we were curious what impact the choice of boundary states has. First, to permit a maximally meaningful comparison between methods, correlations of RC_q with q_{MSM} were calculated for snapshots that are not part of a boundary in either model. This resulted in a Pearson correlation of 0.93 and a Spearman correlation of 0.91, which reveals a high degree of consistency in both linear relationship and ranking of the two orthogonal methods for estimating the committor. Next, we repeated the analysis of Fig. 7 by imposing the boundary states utilized by the MSM onto the RC construction. As shown in Fig. S10, this preserves the qualitative interpretations from above while improving the quantitative agreement specifically with q_{MSM} while leading to larger disagreements with RC_τ. The correlation coefficients between RC_q and q_{MSM} are 0.972 (Pearson) and 0.967 (Spearman) in this case, and, on average, they differ by 0.082 per snapshot. This is the expected behavior with one important caveat:

in this case, the optimization had to be stopped earlier (after 750 iterations), an issue we return to below.

A key difference between the RC and the MSM is that the former assigns a snapshot-wise RC value, while the latter relies on discretized states. Naturally, this lends the RC higher resolution. If the two methods offer consistent estimates, the RC-values for members of a cluster are expected to be distributed tightly around its q_{MSM}-value. Figure 10 suggests that this is generally the case for the present analyses, but the value around which the RC is scattered tends to be shifted toward higher values compared to q_{MSM}, which is visually evident from Fig. 7 as well.

While extensive optimization of the RC can be responsible for such shifts due to the lack of a well-defined stopping criterion, which we discussed above, this does not seem causative for the discrepancy. First, the progress of the optimization suggests a relatively stable plateau (Fig. S8); second, even after relatively few iterations, RC-values are offset with respect to q_{MSM} (Fig. S11). Given the improved agreement in Fig. S10 when enforcing the MSM-derived definitions of boundary states, we conclude that an exact match of states might be required for q_{MSM} and RC_q to approximate each other semi-quantitatively. When using 750 iterations to construct this RC, the agreement is improved between the RC and the MSM not only in Fig. S10 but also when looking more in detail at individual clusters (Fig. S12, top panel). Matching the boundary states also leads to improved agreement of MFPTs derived from the MSM and the RC. For the latter, τ_{AB} is estimated to be on the order of 75 ns, whereas τ_{BA} is estimated to be threefold slower. When using the MSM-states, the RC gives MFPTs of 158 ns for unbinding and 171 ns for the reverse process. This is in very good agreement with MFPTs obtained by solving the system of equations using the MSM's underlying transition matrix where τ_{AB} = 123 ns and τ_{BA} = 137 ns.

When continuing the optimization of the RC using the MSM's boundary states beyond 750 iterations, at which point q_{MSM} and RC_q exhibit strong correspondence, up to 3000 iterations, intermediate points collapse into the boundary states in terms of their committor values (Fig. S12, bottom panel). This does not affect the sorting of snapshots (up to floating point precision) but unfortunately renders the numerical values meaningless. Consequently, rate constants cannot be estimated accurately based on this RC. When choosing the MSM boundaries for the RC, the bound state contains 2.12 as many snapshots compared to the unbound state (15 650 snapshots bound, 7366 unbound). Conversely, in the original, RMSD-based definition, there are 12 000 bound configurations, but 398 344 snapshots in the unbound reference state, which is a dramatically different ratio of 0.03. We can see at least three factors contributing to this sensitivity, which are all linked. First, in any rate analysis, the boundary states must be defined carefully. The very large unbound state used for RC_q might be kinetically homogeneous but certainly does not consist of geometrically self-similar members. Second, the sampling must allow for a reasonable inference of the kinetics between those two states. In absurd scenarios, such as Fig. S7, the profile cannot be expected to be insightful. Similarly, if the two end states are connected by very few transitions, the RC might optimize toward lumping almost everything into one of the two states. Hints of this problem are evident in both Figs. S12 and S13. Third, the RC construction might lack an appropriate theoretical framework to compensate for imbalanced states in terms of flux, population, or congruity. In other words, the method relies on hidden assumptions that have yet to be formalized. At the moment,

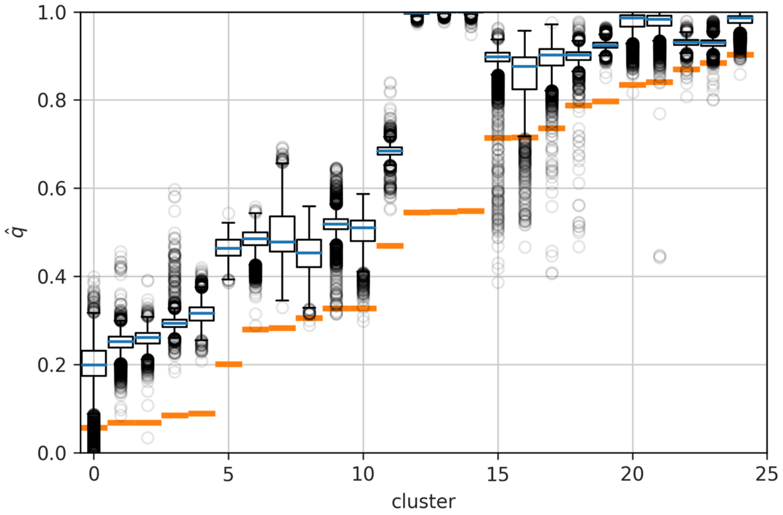

FIG. 10. Comparison of the committor calculated from the MSM and the RC encoding the unbinding of capped GKacG from ATAD2. The 25 clusters containing the most snapshots that are not part of the boundary states are shown. They are sorted by their q_{MSM}-value in increasing order and collectively account for roughly one third of non-boundary snapshots. The orange horizontal bars indicate the estimate of the committor from the MSM. The snapshot-wise values of the RC after 3000 iterations for the constituents of the respective cluster are represented by the boxplots. The boxes extend from the first to the third quartile with the median shown as a blue line. The whiskers extend at most 1.5 times the inter-quartile range. Individual data points beyond this range are shown as circles.

it appears likely that all three factors are in play. Comparing Figs. S13 and S8 allows for the relatively straightforward diagnosis that, with the MSM boundary states imposed, the optimization does not proceed as expected after 500–1000 iterations. It therefore appears indispensable to monitor the convergence behavior and discard RCs that fail this visual inspection.

In summary, in favorable circumstances, the RC can exhibit a high degree of consistency with the MSM and good correspondence with independent geometric descriptors of unbinding. Information from enhanced sampling trajectories is condensed into a single kinetic coordinate, which permits the estimation of rate constants. The good quantitative agreement supports the idea of the RC being an adequate alternative to MSMs for modeling the kinetics of the unbinding of K*ac* from ATAD2 and, by extension, for problems of similar complexity. However, the necessity to explore different definitions for boundary states and different levels of optimization revealed in this example poses a challenge. Somewhat similar to MSMs,[11] it appears that the selection of an appropriate RC is not automatic and that their conceptual "optimality" must be questioned. This is especially related to the theoretical underpinnings and the convergence properties of the optimization, which evidently can lead to largely meaningless RCs. Additional research on modifications and extensions that formalize and control this behavior would be highly desirable. Despite the caveats, our findings underline the potential of optimized RCs in the analysis of challenging trajectory ensembles.

III. CONCLUDING DISCUSSION

We have tested the performance of a recently proposed algorithm[17] to determine RCs on trajectory ensembles generated by different sampling paradigms, including advanced sampling methodologies that carry biases. We investigated two different processes: the (un)folding of an α-helical peptide and the unbinding of a short peptide ligand from its cognate protein domain. We established that such RCs are able to summarize information from thousands of disconnected enhanced-sampling trajectories in order to describe the transition of a system between two appropriate boundary states. They allow us to estimate quantitatively the rates for equilibrium folding (FS-peptide) and ligand (un-)binding (ATAD2 and capped GK*ac*G) from these complex datasets. The principal input required from the user is nothing more than the definition of the two boundary states based on *a priori* criteria, which constitutes the method's foremost advantage. In principle, the definition of boundary states is arbitrary and depends on the research question; the RC is restricted to describing a process conditional on that definition. While such a two-state model is mirrored as a common approximation in analyzing many experiments, it may be prohibitively difficult to define both states for a given system. For example, in binding processes, the unbound state can be delineated in many ways. To address this at least in part, we briefly discussed an RC analogous to a MFPT rather than a committor, which requires the definition of only a single boundary. This allowed us to assess the impact of the ambiguity in the definition of the unbound state for the peptide–protein system on the resulting RC.

When distilling information from ensembles of trajectories into a single dimension, the challenge is to limit the loss in accuracy that offsets gains in interpretability. In recent years, reduction, or at least the clear recognition, of the impact of subjective choices has been an important focus in the field of MD simulations.[37,48–52] Low-dimensional RCs find application in both analyzing simulations and in guiding enhanced sampling; we only explore the former aspect here even though unsupervised RCs of this type have been proposed for the latter.[17] Naturally, sampling biases of this general type have been developed and used in many guises,[13,53–55] the most similar being free-energy guided sampling.[56] Recent years have seen the application of deep-learning methods to omit feature preprocessing or selection and to achieve modeling accuracy by choice of a suitable objective function.[13,57–59] The determination of RCs scrutinized here can be regarded as similar in spirit. Many competing methods address the problem through maximizing the time autocorrelation of collective variables rather than by conditionally minimizing the total squared displacement.[57,58,60–63] The two objectives need not be formally equivalent, making a comparison of the optimized RCs we study here to such alternative estimators difficult *a priori*. Furthermore, while low-dimensional model systems are indispensable for establishing the formal correctness of a method, we were interested here in assessing performance in scenarios resembling real use cases: we generally deal with non-ideal, complex systems that produce noisy data of low statistical quality. Viewing the RC as an estimator per snapshot, we are specifically interested in its bias and variance under such suboptimal conditions.

How can the method be characterized in terms of machine learning (ML)? The optimization proceeds iteratively and constructs a guess that at each step becomes a weighted outer product of its own powers with those of a randomly selected input feature (distance between biopolymer atoms, normalized to its maximum). This is not equivalent to how polynomial kernels are used in ML, most often support vector machines,[64,65] and there is no simple analogy for the workflow in common neural network architectures. If we assume that the features we draw from are sufficient to describe arbitrary processes, the method is unsupervised beyond imposing one or two boundary states. Cross-validation would theoretically be possible with the sliding window logic of MSMs,[66] but practically we found that using lag times other than the sampling time step proved problematic for ensembles of short trajectories generated by PIGS and REX. It would also require storing the functional form of the RC, which is not implemented at the moment. Since the method fits new parameters at each iteration, it is inherently prone to overfitting. Normally, ML approaches protect themselves from a lack of robustness by regularizing the solution space in a controlled manner.[67,68] This does not happen here, which we consider the primary culprit for the lack of stability during optimization (see Figs. S12 and S13). Instead, the properties of the features control the solution space, which leads to the conditional optimization under some homogeneity of RC and geometry, which is desirable. Unfortunately, it is unclear to us how well this conditional optimization protects from limits where the RC values are essentially flat (e.g., 0.5 as in Fig. S7 or 0/1 as in Fig. S12, bottom panel), especially if the number of transitions in the data is low. Flat RC values are a largely albeit trivially homogeneous limit.

In this context, it is important to highlight that the multiplicative construction during iterations, which hinders interpretability,[69] does not allow the RC to remain rigorously unchanged. It does, on the other hand, correctly limit the impact of distances with low variance such as those between covalently bound atoms. In fact, the inhomogeneity of iterations caused by the reliance on randomly selected features challenges the notion of convergence conceptually, for example, when studying a slow but localized process in a large system, which only a vanishing fraction of distances carries information on. The choice of iterations also causes numerical difficulties: artificial clipping is needed to prevent RC values <0 or >1. Overall, it might be preferable to characterize the method as an unusual, stochastic, and conditional optimization procedure where the choice of an individual distance at each step to solve a step-wise optimization problem is akin to an extreme form of stochastic learning on a minimal random subspace. That said, no ensemble learning[70,71] is explored here, which could compensate for the global stochasticity at finite numbers of iterations, and no systematic analysis of the numerical procedure as such has been undertaken.

These caveats aside, we demonstrated above that the integration of the relevant information from randomly selected distances (if any) can be integrated into an RC that faithfully describes the system's slow, collective interconversion between the two selected boundary states. This property was leveraged to recover equilibrium populations from PIGS simulations using the WE strategy,[36] which requires a heuristic to distribute the statistical weight of terminated trajectories onto nearby states. We proposed and tested here the use of the RC to serve as this heuristic. For algorithms that do not change the potential energy surface, such as PIGS, equilibrium reweighting is commonly attempted by constructing a MSM.[72,73] However, Bacci *et al.* demonstrated that this approach can lead to incomplete reweighting and have suggested a WE to be superior in this regard, at least for PIGS and similar data.[11] This result is corroborated here for FS-peptide where the computational ground truth is known (Figs. 3 and 5).

When testing a methodology operating on simulation data, it is important to distinguish the computational ground truth from the experimental one. The rates we estimate here are noisy quantities and carry a sampling error. Beyond that, their accuracy relative to experiment is conditional upon the chosen model, which is not relevant in assessing the performance of the RC methodology. Known force field limitations[74,75] or difficulties in predicting experimental signals can make quantitatively correct rate predictions impossible.[76] To give an example, rate constant estimates for the recognition of benzamidine by trypsin vary over several orders of magnitude in absolute value, with the main consensus being that k_{off} is estimated to be orders of magnitudes faster compared to k_{on} (summarized in Ref. 27). Even within the same force field, estimates are determined by an interplay between the sampler, the extent of sampling, and the estimator itself. We are interested here in the properties of the sampler and, first and foremost, the estimator and thus chose to compare results for the same model/data using reference estimators: for FS-peptide, we compared to estimates from direct counting from CS (Figs. 6 and 4), while for the unbinding problem, we compared to estimates from a carefully constructed MSM (Figs. 7 and 10). From these comparisons, we conclude that the RC can preserve the kinetics from the available sampling in a way that is quantitatively similar to reference methods. We reemphasize that its main advantage over an MSM lies in the simplicity of its construction, for which little prior knowledge in terms of data or system was necessary.

A one-dimensional progress variable cannot fully inform about pathways unless there is no variance in path space. Even though one-dimensional projections can contain pathway information indirectly,[29,37] the RC-based FEPs are, in their intended limit, histograms of committor values. Similar to cut-based FEPs on either committor or MFPT,[77] they sort states by kinetic distance, which means that different states can overlap and off-pathway events are difficult to grasp.[77] While an extension of RCs to faster modes has been proposed,[78] the current method can always be exploited synergistically as part of a toolbox that also contains orthogonal analyses, which potentially involve more modeling choices but summarize other properties of the system or the same properties at higher resolution. In addition to being an independent tool, our findings highlight the RC's potential as a possible means for model validation. For estimation of the committor, discretization as a major source of modeling error[10,79] is largely eliminated, the explicitly defined boundary states notwithstanding. Thus, committors obtained from MSMs can be directly compared, especially if the boundary states are matched exactly (Fig. S10). Such comparisons should focus on measuring rank correlations rather than absolute values, which is a caveat (Fig. 4).

In order to make the method available as a routine tool for learning from data or for providing a reference for independent methodologies, some problems will have to be addressed. First, the semantic meaning of the RC strongly depends on the employed definitions of states, and care must be taken in choosing appropriate boundary states independently of RC-based analyses. For imperfect sampling, which is the general case, only heuristic stopping criteria for the iterative RC construction are available, and continuing the optimization can cause RC values for most intermediate snapshots to collapse to the closer boundary,[17] which is a type of overfitting as discussed above. The missing overall regularization of solutions and the feature-dependent impact of individual iterations in the method are major concerns. Overfitting will usually cause the RC to deviate from the actual committor and will consequently make the interpretation of RC-values as probabilities[80] questionable. The dependence on optimization progress (Fig. S12) seems to be exacerbated in cases where the raw data violate detailed balance as evident for PIGS for ATAD2 or where their connectivity might be compromised as for the case of the REX data for the FS-peptide (Fig. S4). Here, intermediate snapshots tended to collapse onto the more populated boundary (helix). While it was shown for this system that REX alters the way and extent by which conformational transitions are sampled,[23] the impact of this should ideally be small, given that REX data appear to be at thermodynamic equilibrium when compared to CS. Assumptions about equilibrium enter the procedure because the objective function is generally constructed as a linear sum, which we attempted to correct for FS-peptide PIGS data using weights (see Sec. IV B). Furthermore, the Markovianity of input features is an approximation or assumption that is not generally met (Fig. S5), which similarly challenges the definition of the objective function.

While adjustments of the boundary conditions and exploiting adaptive sampling time steps have been proposed to address this,[18] finite sampling is an unavoidable source of error in the estimation of the exact committor in practice. Thus, we think that it would be better to develop the method in the direction of preserving qualitative insights similar to Ref. 29 that contain quantitative insights whenever the data and the method allow it. This will require revisiting both the optimization logic and the objective function while not touching the main strength of the method: its largely intervention-free construction. The qualitative insights are contained more in the RC's capability of ordering snapshots according to kinetic proximity, and this is what we have predominantly focused on in this work. Conceptually, the objective function is written based on Markovian dynamics and, to correct for errors in this, RC-dependent diffusion, which arises naturally from projection to a single dimension,[43,81] and memory terms[82] should be taken into account explicitly.

We conclude that RC optimization constitutes a valuable tool for summarizing complex ensembles of MD trajectories. This is particularly remarkable for ensembles from advanced sampling methodologies, which are a focus of current research,[58,83–85] as they afford reaching timescales of biological relevance.

IV. METHODS

A. RC optimization

The definition of boundary states is a crucial choice for RC optimization. For FS-peptide, state A was defined as the ensemble of structures possessing a heavy-atom RMSD of <1.5 Å from the fully extended helix. Snapshots with an RMSD of >9.1 Å instead make up boundary B. For further settings of the optimization procedure, we loosely followed recommendations from Ref. 17. Specifically, the RC was initialized to 0.5 for all intermediate snapshots and updated with 3000 random interatomic distances at a sampling time step of 1.5 ps. The initial guess does not influence the RC beyond the very early stages of optimization, as changes to the RC are drastic in this regime (Figs. S2, S8, and S13). Empirically, uniform (0.5) and random initializations work equally well. To minimize user intervention, all atoms, including hydrogen atoms, were considered.

Conceptually, at each iteration, the RC evolves trying to incorporate information on the slow dynamics of the system by decreasing the objective function $\Delta r^2 = \sum_k [r(k\Delta t + \Delta t) - r(k\Delta t)]^2$ over all k snapshots of the trajectory. A basis function $f(r_i, d_i;\ \alpha_i)$ updates the RC at iteration i, r_i, in a multiplicative manner and conditional upon the randomly chosen interatomic distance d_i. The parameter α_i is chosen to minimize Δr^2_{i+1} by solving a least-squares problem such that $r_{i+1} = r_i + f(r_i, d_i; \alpha_i^*)$, where α_i^* minimizes Δr^2_{i+1} while keeping the RC values at the boundaries fixed.[16] By using only geometric features (distances), an implicit precondition of homogeneity of the RC for geometrically similar structures is built in but the exact mechanism is hard to spell out. We point out that, clearly, linearly interpolating the RC between boundary states as necessary will generally be vastly superior in terms of objective function but will violate this principle.

Note that we record only the numerical values of the RC but not the function itself such that unseen configurations cannot easily be mapped to an RC value. Because the method is iterative and changes at early stages of the optimization are more drastic, it is not straightforward to interpret coefficients as importance measures for specific features. Specifically, the RC was updated in three steps: first, the new interatomic distance and the previous RC were combined using a fourth-degree polynomial. Second, the region on r with the highest density in Δr^2 was updated using an eighth-degree polynomial by applying a Laplacian envelope with a scale parameter drawn randomly from a uniform distribution on the interval $[0, 1]$ at each iteration. This strategy was proposed in Ref. 18 and was devised for providing greater emphasis on the region on r where the representation by the RC is least accurate. This region was (re)determined every 400 iterations. Third, r itself was updated with an eighth-degree polynomial. The parameters for the basis function used in the second and third steps were chosen analogously to the procedure described above but for a function $g(r_i; \beta_i)$, which takes as input only the RC itself. Due to the assumptions inherent to the construction, only snapshots pairs that are actually neighboring in time must contribute to Δr^2. This information has to be supplied by the user. We did not attempt to analyze the properties of this numerical procedure analytically or by systematic, numerical exploration.

For REX, only the trajectories at the relevant temperature of 250 K that match CS and PIGS were considered for constructing the RC. Compared to PIGS and CS, this is a substantially smaller amount of sampling (32 × 208 000 = 6 656 000 snapshots each for PIGS and CS compared to 4 × 208 000 frames for REX).

For ATAD2, PIGS was run to diversify the configuration of the ZA loop and the BC loop, both of which are in contact with the bound peptide in the crystal structure (see Sec. IV D for details and Fig. S1). We determined an RC to capture the unbinding of the acetylated lysine-containing peptide for both sets of 64 replicas combined (Fig. S8) at a sampling time step of 20 ps. State A was defined to be at PI values 869 000–881 000 (see Fig. 7 and Sec. IV D 2), whereas state B was chosen via RMSD relative to the initial, bound state: the value across all ligand heavy atoms had to exceed 25 Å. This RMSD was calculated using Gromacs 2020.3[86] after aligning the system to the set of protein heavy atoms that are part of neither the ZA nor the BC loop. The RC was updated as described above for 3000 iterations but with a polynomial of degree 16 (instead of 8) to update suboptimal regions and r itself.

Calculations for determining the RCs were performed running Python 3.8.6 and libraries NumPy[87] (version 1.21.0) and TensorFlow[88] (2.4.3, CPU only) on a desktop machine. Functions for determining optimal parameters were adapted from the code presented in Ref. 17 (available at https://github.com/krivovsv/NPNE). Trajectory files were read, and distances were calculated using the MDtraj package (version 1.9.6).[89]

B. Weighted ensemble for FS-peptide PIGS trajectories

The goal of the weighted ensemble strategy[11,36] is to track the weight changes incurred by the splitting and termination of trajectories, which happens frequently in PIGS (but never in REX or CS). In detail, at the start of PIGS, each of the 32 replicas is assigned a uniform weight of 1/32. At each reseeding event, the weight of every terminated trajectory is distributed to the n kinetically closest copies of the system. As the RC encodes kinetic distance between snapshots, we hypothesized that the application of a kernel estimate to differences in RC constitutes an effective way to devise a splitting rule that represents kinetic proximity. Among the three kernels tested, an exponential kernel performed best (Fig. S6) while also offering better numerical stability compared to a radial basis function (RBF) kernel. In the end, we chose to distribute the weight of a terminated replica to the $n = 16$ surviving copies closest to it at the time of reseeding, with the fractional contributions determined by the probability density of the kernel function at their individual RC difference values. The choice of n is a free parameter but of limited impact for localized kernels (see Fig. S6), and we chose it in accordance with the number of protected replicas during PIGS, which means that there were always at least 16 copies available to absorb the weight.

The RC optimization employed here conditionally minimizes the total, squared displacement, which, for a Markov chain, can be written as

$$\Delta r^2 = \sum_{i,j} P(j|i)\pi_i(r_i - r_j)^2, \tag{2}$$

where π_i denotes the Boltzmann-weight of state i.

PIGS trajectories use a geometric distance between snapshots to promote exploration while not altering the conditional probabilities $P(j|i)$. Therefore, the only reason that the sampling weights are not equivalent to equilibrium weights is due to the repeated splitting/terminating of trajectories in a manner that creates a non-Boltzmann distribution of starting configurations for each short trajectory. To correct this initial state bias, we exploit the WE strategy described above to calculate snapshot-wise equilibrium weights π_i. These are then applied to provide a weight for contributions to Δr^2 during optimization. Specifically, we apply the weight at time $k\Delta t$ for reweighting the transition $r(k\Delta t) \rightarrow r(k\Delta t + \Delta t)$. This choice is expected to have no relevant impact as the weights for most neighbor pairs are identical since changes can only occur at reseeding points. Similarly, we use the weighted ensemble weights to remove initial state bias from histogram-based FEPs, in which case we ensured that the total weight integrates to 1. The helicity of the FS-peptide was determined by the DSSP-algorithm[90] as implemented in CAMPARI v4 (http://campari.sourceforge.net). To capture the variability of the binary helicity within each bin of Fig. 2, a balance measure was calculated according to $1 - (|n_1 - n_0|)/(n_1 + n_0)$, where n_0 and n_1 denote the number of non-α-helical and α-helical snapshots in a bin, respectively. Values close to 1 indicate perfect balance between helical and non-helical configurations, whereas perfectly homogeneous bins result in 0. Cartoons of representative protein structures were generated by PyMOL 2.4.1.[91]

C. Simulation restarts for generating test data for FS-peptide

The data labeled "restarts" in Fig. S4 were generated identically to the CS settings in the original work:[23] they were independent, canonical Langevin dynamics simulations at 250 K with the ABSINTH implicit solvation model.[39] The only differences were their starting structures (1024 per RC bin, extracted by systematically subsampling available snapshots) and the short simulation length of 30 ns.

D. PIGS simulation of ATAD2

The bromodomain of ATAD2 was simulated in complex with a tri-peptide with sequence GK*ac*G; this little motif can be found at multiple positions along the histone sequence, e.g., in the original construct H4K12*ac* of PDB structure 4QUT,[92] from which the initial coordinates of the system are taken. The N- and C-termini of the peptide and of the protein were capped with acetyl and N-methylamide groups, respectively. Asp and Glu side chains were negatively charged, Arg and Lys were positively charged, and histidines were kept neutral in the Nε tautomer. The system was solvated in a cubic box of 85 Å side length, and K^+ and CL^- ions were added to neutralize the complex and approximate an ionic strength of 150 mM. Parameters for the system were taken from the CHARMM36[93] force field with modified TIP3P water[94] and a custom patch for the non-standard residue K*ac*. In simulations, all covalent bonds were constrained by the LINCS algorithm;[95] non-bonded interactions (both electrostatic and van der Waals) were cut off at 12 Å, and long-range electrostatic interactions were calculated by the generalized reaction field method.[96] A first equilibration in the NPT ensemble at 1 bar and 310 K (using Berendsen pressure coupling[97] and the velocity rescaling thermostat[98]) was run for 1 ns in order to allow the volume of the box to adjust. The box side was then fixed to its average value obtained during the first relaxation (84.305 Å), and the system was further equilibrated in the NVT ensemble at 310 K (with velocity rescaling coupling) for additional 0.5 ns. The following production simulations kept the same settings used in the NVT equilibration. Each PIGS run consisted of 64 replicas, attempting reseeding every 100 ps, each time protecting the 32 top-ranked replicas from being terminated. Trajectory coordinates were saved every 0.2 ps for the calculation of the PIGS heuristic and every 20 ps for analysis (PI and MSM construction). The simulations were run with GROMACS 2016,[86] whereas the reseeding heuristic was calculated with CAMPARI v3b, and the two softwares were interfaced by a custom Python script. Two disjoint sets of dihedral angles were chosen to separately target the ZA and BC loop for diversification, giving rise to the ZA PIGS and BC PIGS sets of simulations. Additional details on the simulation protocol and the full list of degrees of freedom for PIGS enhancement are given in Ref. 30, which introduced equivalent runs for the ATAD2 bromodomain in its *apo* form. The total sampling amount is 157.6 ns/copy (10.1 μs cumulative sampling) for ZA PIGS and 160.3 ns/copy (10.3 μs) for BC PIGS.

1. Featurization of the system

To be able to construct the progress index or an MSM from the data, we need to design a featurization of the system, which should be representative of the process under study, i.e., ligand unbinding. The use of inter-residue distances (possibly intra-receptor, intra-ligand, and between receptor and ligand) is a reasonable choice. In order to identify the most relevant of such distances, we made use of contact maps, which keep track of the frequency of contacts between each pair of residues; a contact is considered formed whenever any atoms of the two residues are closer than 5 Å. For each set of simulations separately (ZA PIGS and BC PIGS), we compared two contact maps: the first was calculated on the 1000 snapshots closest to the crystal pose, in terms of the RMSD of all Cα atoms of the bromodomain and ligand upon alignment on the Cα atoms of the bromodomain alone; the second contact map considered all the remaining snapshots. We then selected the inter-residue contacts whose difference in frequency between the two contact maps was greater than or equal to 0.15. The initially identified contact pairs were further reduced by intersecting the set of pairs from ZA PIGS and BC PIGS and by manual selection. Out of the final 43 pairs, 27 correspond to intra-bromodomain and 16 correspond to peptide-bromodomain contacts. The selected contacts are listed in Table S1 and annotated in the structure of the complex in Fig. S1. Each residue pair (i,j) was further expanded into four inter-residue distances, considering the minimum distance between the following pairs of atom sets: backbone$_i$–backbone$_j$, backbone$_i$–sidechain$_j$, sidechain$_i$–backbone$_j$, and sidechain$_i$–sidechain$_j$. The resulting 172 distances were smoothed with a sigmoid of the form

$$f(x) = 1 - \frac{1}{1 + \exp\frac{-(x-5.0)}{0.75}} \tag{3}$$

in order to mimic a continuous approximation of a binary contact metric. The application of a sigmoid provides the additional advantage that when the ligand is unbound, the variance and thereby impact of large, noisy peptide-bromodomain distances on conformational distance are squashed to 0. As the set of 172 distances is somewhat redundant due to covalent geometry constraints, its dimensionality was reduced to 30 components by principal component analysis (PCA).[99] The featurization (contact distance calculation and PCA) was carried out with CAMPARI v4. Cartoons of representative protein structures were generated using an in-house tool and PyMOL 2.4.1.[91]

2. *Progress index calculation*

A qualitative overview of the system's thermodynamics and kinetics can be visually rendered by means of a SAPPHIRE plot[37] as in Fig. 7 where each snapshot is reordered along the x axis according to the progress index.[29] Briefly, given a set of features and a geometric similarity criterion, starting from an arbitrary snapshot, the next one to be added to the progress index is the closest to any of the snapshots already added. The creation of the progress index relies on nearest-neighbor distances and can be solved exactly by determining a minimum spanning tree. For scalability, an approximate version of the algorithm is available that relies on a heuristic search of the closest snapshots guided by a multi-resolution clustering of the conformations into a tree-based structure.[100] A parallel version of this algorithm is implemented in CAMPARI since v3.[101] The PI, by grouping together snapshots from dense regions of the conformational space, creates a 1D ordering of the sampled regions and can be used to plot a pseudo-free energy profile. Additional snapshot-based, geometric annotations can highlight properties of the different free energy basins. For our specific system, the geometric distance between snapshots was calculated using the Euclidean distance of the set of 30 features defined above. Our own tree-based clustering algorithm[100] was used to organize the conformations from all trajectories (ZA PIGS and BC PIGS) at 16 resolution levels; the finest level, with a cluster radius of 0.25 Å, contained 36 562 clusters, whereas an intermediate resolution (cluster radius 0.47 Å, 1255 clusters) was used for network-related analyses.

3. *MSM construction and related analyses*

The transition counts along the trajectories were used to determine the MSM transition matrix by maximum *a posteriori* estimation (MAP) using a Dirichlet prior with flat concentration parameters equal to $1 + 1/N$, where N is the number of states (corresponding to the maximum likelihood estimate if one assumes an additional pseudo-count of $1/N$ for every transition).[11,102] The resulting network is fully connected but detailed balance is not imposed. Transitions are counted using a sliding window with a lag time of 100 ps, and they take into consideration the PIGS reseeding history. To reduce the bias related to the sampling enhancement, clusters were reweighted using the steady-state distribution from the MSM.

For MSMs, boundary states A and B are defined at the level of clusters. The definition should be stringent rather than too generous to exclude kinetic shortcuts. Therefore, we defined the bound state A ($q_{MSM} = 0$) as the clusters containing the initial snapshots of the ZA PIGS replicas (four clusters, 15 650 snapshots) and the unbound state B ($q_{MSM} = 1$) as the largest cluster whose centroid has a bromodomain–peptide distance of >25 Å (one cluster, 7366 snapshots). This distance was defined as the minimum distance between residues from the ZA and BC loop regions and residues from the peptide. The MFPTs from the MSM were calculated by solving the linear system for Markov chains[103] using PyEMMA v2.5.[104] The committor values were calculated using transition path theory (TPT)[105–107] as implemented in CAMPARI.

SUPPLEMENTARY MATERIAL

Figures S1 along with Table S1 (related to Fig. 9), S2–S4 (related to Fig. 2), S5 and S7 (related to Fig. 1), S6 (related to Fig. 3), S8 and S9 (related to Fig. 9), S10 (related to Fig. 7), and S11–S13 (related to Fig. 10) are included in a single file as the supplementary material.

ACKNOWLEDGMENTS

We thank Sergei Krivov for his insight and help in getting started with his method. Furthermore, we acknowledge the fruitful discussions with Alessia Del Panta Ridolfi and Francesco Cocina. We acknowledge access to computing resources at the Swiss National Supercomputing Centre, Switzerland. This work was supported, in part, by a grant of the Swiss National Science Foundation to A.C. (Grant No. 310030-212195) and by a computing grant from CSCS ("s636") to A.V. and A.C.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Julian Widmer: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Cassiano Langini**: Data curation (equal); Formal analysis (equal); Investigation (supporting); Visualization (supporting); Writing – original draft (supporting); Writing – review & editing (equal). **Andreas Vitalis**: Conceptualization (equal); Data curation (equal);

Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Amedeo Caflisch**: Conceptualization (lead); Funding acquisition (lead); Project administration (lead); Supervision (lead); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available at Zenodo under the DOI: 10.5281/zenodo.7688787. Code is openly available under https://gitlab.com/CaflischLab/optimalrc.

REFERENCES

[1] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: A computational microscope for molecular biology," Annu. Rev. Biophys. **41**, 429–452 (2012).

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

[3] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, "Principles and overview of sampling methods for modeling macromolecular structure and dynamics," PLOS Comput. Biol. **12**, e1004619 (2016).

[4] Y. Wang, J. M. Lamim Ribeiro, and P. Tiwary, "Machine learning approaches for analyzing and enhancing molecular dynamics simulations," Curr. Opin. Struct. Biol. **61**, 139–145 (2020).

[5] W. E and E. Vanden-Eijnden, "Towards a theory of transition paths," J. Stat. Phys. **123**, 503 (2006).

[6] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**, 2386–2396 (2018).

[7] F. Cocina, A. Vitalis, and A. Caflisch, "Sapphire-Based clustering," J. Chem. Theory Comput. **16**, 6383–6396 (2020).

[8] B. E. Husic and V. S. Pande, "Ward clustering improves Cross-Validated Markov state models of protein folding," J. Chem. Theory Comput. **13**, 963–967 (2017).

[9] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).

[10] N. Djurdjevac, M. Sarich, and C. Schütte, "Estimating the eigenvalue error of Markov state models," Multiscale Model. Simul. **10**, 61–81 (2012).

[11] M. Bacci, A. Caflisch, and A. Vitalis, "On the removal of initial state bias from simulation data," J. Chem. Phys. **150**, 104105 (2019).

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

[13] J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary, "Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)," J. Chem. Phys. **149**, 072301 (2018).

[14] S. V. Krivov, "On reaction coordinate optimality," J. Chem. Theory Comput. **9**, 135–146 (2013).

[15] S. V. Krivov, "Is protein folding sub-diffusive?," PLOS Comput. Biol. **6**, e1000921 (2010).

[16] P. V. Banushkina and S. V. Krivov, "Nonparametric variational optimization of reaction coordinates," J. Chem. Phys. **143**, 184108 (2015).

[17] S. V. Krivov, "Nonparametric analysis of nonequilibrium simulations," J. Chem. Theory Comput. **17**, 5466–5481 (2021).

[18] S. V. Krivov, "Protein folding free energy landscape along the committor—The optimal folding coordinate," J. Chem. Theory Comput. **14**, 3418–3427 (2018).

[19] A. Berezhkovskii and A. Szabo, "One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions," J. Chem. Phys. **122**, 014503 (2005).

[20] A. M. Berezhkovskii and A. Szabo, "Committors, first-passage times, fluxes, Markov states, milestones, and all that," J. Chem. Phys. **150**, 054106 (2019).

[21] E. Weinan, W. Ren, and E. Vanden-Eijnden, "Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes," Chem. Phys. Lett. **413**(1–3), 242–247 (2005).

[22] A. M. Berezhkovskii and A. Szabo, "Diffusion along the splitting/commitment probability reaction coordinate," J. Phys. Chem. B **117**, 13115–13119 (2013).

[23] 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).

[24] R. H. Swendsen and J.-S. Wang, "Replica Monte Carlo simulation of spin-glasses," Phys. Rev. Lett. **57**, 2607 (1986).

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

[26] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of molecular dynamics and related methods in drug discovery," J. Med. Chem. **59**, 4035–4061 (2016).

[27] N. J. Bruce, G. K. Ganotra, D. B. Kokh, S. K. Sadiq, and R. C. Wade, "New approaches for computing ligand–receptor binding kinetics," Curr. Opin. Struct. Biol. **49**, 1–10 (2018).

[28] M. Bernetti, M. Masetti, W. Rocchia, and A. Cavalli, "Kinetics of drug binding and residence time," Annu. Rev. Phys. Chem. **70**, 143–171 (2019).

[29] N. Blöchliger, A. Vitalis, and A. Caflisch, "A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems," Comput. Phys. Commun. **184**, 2446–2453 (2013).

[30] M. Bacci, C. Langini, J. Vymětal, A. Caflisch, and A. Vitalis, "Focused conformational sampling in proteins," J. Chem. Phys. **147**, 195102 (2017).

[31] L. Zeng and M.-M. Zhou, "Bromodomain: An acetyl-lysine binding domain," FEBS Lett. **513**, 124–128 (2002).

[32] P. Filippakopoulos and S. Knapp, "The bromodomain interaction module," FEBS Lett. **586**, 2692–2704 (2012).

[33] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J.-P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson *et al.*, "Histone recognition and large-scale structural analysis of the human bromodomain family," Cell **149**, 214–231 (2012).

[34] J.-R. Marchand and A. Caflisch, "Binding mode of acetylated histones to bromodomains: Variations on a common motif," ChemMedChem **10**, 1327–1333 (2015).

[35] C. Langini, A. Caflisch, and A. Vitalis, "The ATAD2 bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes," J. Biol. Chem. **292**, 19121 (2017).

[36] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, "The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures," J. Chem. Phys. **132**, 054107 (2010).

[37] N. Blöchliger, A. Vitalis, and A. Caflisch, "High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations," Sci. Rep. **4**, 6264 (2014).

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

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

[40] A. Vitalis and A. Caflisch, "50 years of Lifson–Roig models: Application to molecular simulation data," J. Chem. Theory Comput. **8**, 363–373 (2012).

[41] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, "Estimation of protein folding probability from equilibrium simulations," J. Chem. Phys. **122**, 184901 (2005).

[42] P. V. Banushkina and S. V. Krivov, "Optimal reaction coordinates," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **6**, 748–763 (2016).

[43] S. V. Krivov, "Numerical construction of the p(fold) (committor) reaction coordinate for a Markov process," J. Phys. Chem. B **115**, 11382–11388 (2011).

[44] S. Steiner, A. Magno, D. Huang, and A. Caflisch, "Does bromodomain flexibility influence histone recognition?," FEBS Lett. **587**, 2158–2163 (2013).

[45] Y. Zhou, M. Hussain, G. Kuang, J. Zhang, and Y. Tu, "Mechanistic insights into peptide and ligand binding of the ATAD2-bromodomain via atomistic simulations disclosing a role of induced fit and conformational selection," Phys. Chem. Chem. Phys. **20**, 23222–23232 (2018).

[46] L. Raich, K. Meier, J. Günther, C. D. Christ, F. Noé, and S. Olsson, "Discovery of a hidden transient state in all bromodomain families," Proc. Natl. Acad. Sci. U. S. A. **118**, e2017427118 (2021).

[47] P. Tompa and M. Fuxreiter, "Fuzzy complexes: Polymorphism and structural disorder in protein–protein interactions," Trends Biochem. Sci. **33**, 2–8 (2008).

[48] G. A. Tribello, M. Ceriotti, and M. Parrinello, "A self-learning algorithm for biased molecular dynamics," Proc. Natl. Acad. Sci. U. S. A. **107**, 17509–17514 (2010).

[49] N. Blöchliger, A. Caflisch, and A. Vitalis, "Weighted distance functions improve analysis of High-Dimensional data: Application to molecular dynamics simulations," J. Chem. Theory Comput. **11**, 5481–5492 (2015).

[50] S. R. Hare, L. A. Bratholm, D. R. Glowacki, and B. K. Carpenter, "Low dimensional representations along intrinsic reaction coordinates and molecular dynamics trajectories using interatomic distance matrices," Chem. Sci. **10**, 9954–9968 (2019).

[51] D. Wang and P. Tiwary, "State predictive information bottleneck," J. Chem. Phys. **154**, 134111 (2021).

[52] H. Wu and F. Noé, "Variational approach for learning Markov processes from time series data," J. Nonlinear Sci. **30**, 23–66 (2020).

[53] S. Doerr and G. De Fabritiis, "On-the-Fly learning and sampling of ligand binding by High-Throughput molecular simulations," J. Chem. Theory Comput. **10**, 2064–2069 (2014).

[54] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello, "Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps," Proc. Natl. Acad. Sci. U. S. A. **112**, E386–E391 (2015).

[55] H. Jung, R. Covino, and G. Hummer, "Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations," arXiv:1901.04595 [physics.chem-ph] (2019).

[56] T. Zhou and A. Caflisch, "Free energy guided sampling," J. Chem. Theory Comput. **8**, 3423 (2012).

[57] A. Mardt, L. Pasquali, H. Wu, and F. Noé, "VAMPnets for deep learning of molecular kinetics," Nat. Commun. **9**, 5 (2018).

[58] L. Bonati, G. Piccini, and M. Parrinello, "Deep learning the slow modes for rare events sampling," Proc. Natl. Acad. Sci. U. S. A. **118**, e2113533118 (2021).

[59] R. Ketkaew and S. Luber, "DeepCV: A deep learning framework for blind search of collective variables in expanded configurational space," J. Chem. Inf. Model. **62**, 6352–6364 (2022).

[60] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," Phys. Rev. Lett. **72**, 3634–3637 (1994).

[61] Y. Naritomi and S. Fuchigami, "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions," J. Chem. Phys. **134**, 065101 (2011).

[62] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations," J. Chem. Phys. **146**, 154104 (2017).

[63] W. Chen, H. Sidky, and A. L. Ferguson, "Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets," J. Chem. Phys. **150**, 214114 (2019).

[64] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," PLOS Comput. Biol. **4**, e1000173 (2008).

[65] S. Klus, A. Bittracher, I. Schuster, and C. Schütte, "A kernel-based approach to molecular conformation analysis," J. Chem. Phys. **149**, 244109 (2018).

[66] R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics," J. Chem. Phys. **142**, 124105 (2015).

[67] D. Calvetti and E. Somersalo, "Inverse problems: From regularization to Bayesian inference," WIREs Comp Stats. **10**, e1427 (2018).

[68] N. G. Polson and V. Sokolov, "Bayesian regularization: From Tikhonov to horseshoe," WIREs Comp Stats. **11**, e1463 (2019).

[69] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," IEEE Access **8**, 42200–42216 (2020).

[70] O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Min. Knowl. Discov. **8**, e1249 (2018).

[71] S. Kotsiantis, "Combining bagging, boosting, rotation forest and random subspace methods," Artif. Intell. Rev. **35**, 223–240 (2011).

[72] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé, "Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias," J. Chem. Phys. **146**, 094104 (2017).

[73] H. Wan and V. A. Voelz, "Adaptive Markov state model estimation using short reseeding trajectories," J. Chem. Phys. **152**, 024103 (2020).

[74] Y. Mao and Y. Zhang, "Thermal conductivity, shear viscosity and specific heat of rigid water models," Chem. Phys. Lett. **542**, 37–41 (2012).

[75] L. A. Abriata and M. Dal Peraro, "Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization," Comput. Struct. Biotechnol. J. **19**, 2626–2636 (2021).

[76] W. F. van Gunsteren, J. Dolenc, and A. E. Mark, "Molecular simulation as an aid to experimentalists," Curr. Opin. Struct. Biol. **18**, 149–153 (2008).

[77] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, "One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: New insights into the folding process," J. Phys. Chem. B **112**, 8701–8714 (2008).

[78] S. V. Krivov, "Blind analysis of molecular dynamics," J. Chem. Theory Comput. **17**, 2725–2736 (2021).

[79] S. V. Krivov and M. Karplus, "Diffusive reaction dynamics on invariant free energy profiles," Proc. Natl. Acad. Sci. U. S. A. **105**, 13841–13846 (2008).

[80] G. R. Bowman, V. S. Pande, and F. Noé, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer Science & Business Media, 2013).

[81] R. J. d. Oliveira, "Coordinate-Dependent Drift-Diffusion reveals the kinetic intermediate traps of Top7-Based proteins," J. Phys. Chem. B **126**, 10854–10869 (2022).

[82] B. Kowalik, J. O. Daldrop, J. Kappler, J. C. F. Schulz, A. Schlaich, and R. R. Netz, "Memory-kernel extraction for different molecular solutes in solvents of varying viscosity in confinement," Phys. Rev. E **100**, 012126 (2019).

[83] S.-T. Tsai, Z. Smith, and P. Tiwary, "SGOOP-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations," J. Chem. Theory Comput. **17**, 6757–6765 (2021).

[84] J. Rydzewski and O. Valsson, "Multiscale reweighted stochastic embedding: Deep learning of collective variables for enhanced sampling," J. Phys. Chem. A **125**, 6286–6302 (2021).

[85] Z. Belkacemi, P. Gkeka, T. Lelièvre, and G. Stoltz, "Chasing collective variables using autoencoders and biased trajectories," J. Chem. Theory Comput. **18**, 59–78 (2022).

[86] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1**, 19–25 (2015).

[87] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," Nature **585**, 357–362 (2020).

[88] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," software available from tensorflow.org, 2015.

[89] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," Biophys. J. **109**, 1528–1532 (2015).

[90] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," Biopolymers **22**, 2577–2637 (1983).

[91] Schrödinger, LLC, The PyMOL molecular graphics system, version 2.4.1, 2015.

[92] Y. Morozumi, F. Boussouar, M. Tan, A. Chaikuad, M. Jamshidikia, G. Colak, H. He, L. Nie, C. Petosa, M. De Dieuleveult *et al.*, "ATAD2 is a generalist facilitator of chromatin dynamics in embryonic stem cells," J. Mol. Cell Biol. **8**, 349–362 (2016).

[93] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, Jr., "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles," J. Chem. Theory Comput. **8**, 3257–3273 (2012).
[94] S. R. Durell, B. R. Brooks, and A. Ben-Naim, "Solvent-induced forces between two hydrophilic groups," J. Phys. Chem. **98**, 2198–2202 (1994).
[95] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).
[96] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, "A generalized reaction field method for molecular dynamics simulations," J. Chem. Phys. **102**, 5451–5459 (1995).
[97] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684–3690 (1984).
[98] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).
[99] C. C. David and D. J. Jacobs, "Principal component analysis: A method for determining the essential dynamics of proteins," Methods Mol. Biol. **1084**, 193–226 (2014).
[100] A. Vitalis and A. Caflisch, "Efficient construction of mesostate networks from molecular dynamics trajectories," J. Chem. Theory Comput. **8**, 1108–1120 (2012).
[101] A. Vitalis, "An improved and parallel version of a scalable algorithm for analyzing time series data," arXiv:2006.04940[cs.DC] (2020).
[102] G. R. Bowman, "Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty," J. Chem. Phys. **137**, 134111 (2012).
[103] T. J. Sheskin, "Computing mean first passage times for a Markov chain," Int. J. Math. Educ. Sci. Technol. **26**, 729–735 (1995).
[104] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PYEMMA 2: A software package for estimation, validation, and analysis of Markov models," J. Chem. Theory Comput. **11**, 5525–5542 (2015).
[105] P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition path theory for Markov jump processes," Multiscale Model. Simul. **7**, 1192–1219 (2009).
[106] A. Berezhkovskii, G. Hummer, and A. Szabo, "Reactive flux and folding pathways in network models of coarse-grained protein dynamics," J. Chem. Phys. **130**, 205102 (2009).
[107] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci. U. S. A. **106**, 19011–19016 (2009).