JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

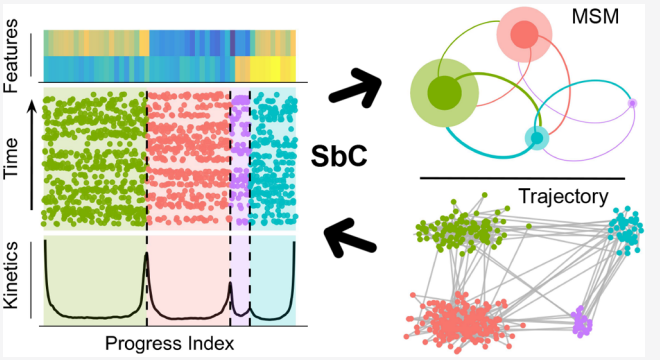
SAPPHIRE-Based Clustering

Francesco Cocina, Andreas Vitalis,* and Amedeo Caflisch

Cite This: https://dx.doi.org/10.1021/acs.jctc.0c00604 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular dynamics simulations are a popular means to study biomolecules, but it is often difficult to gain insights from the trajectories due to their large size, in both time and number of features. The SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot allows a direct visual inference of the dominant states visited by high-dimensional systems and how they are interconnected in time. Here, we extend this visual inference into a clustering algorithm. Specifically, the automatic procedure derives from the SAPPHIRE plot states that are kinetically homogeneous, structurally annotated, and of tunable granularity. We provide a relative assessment of the kinetic fidelity of the SAPPHIRE-based partitioning in comparison to popular clustering methods. This assessment is carried out on trajectories of *n*-butane, a β-sheet peptide, and the small protein BPTI. We conclude with an application of our approach to a recent 100 μs trajectory of the main protease of SARS-CoV-2.



1. INTRODUCTION

Molecular dynamics (MD) simulations are a powerful tool to analyze complex systems at atomic resolution.[1] Due to their scale, biomolecules like proteins undergo stochastic motion in aqueous solution. Because the data sets are large in terms of both sampling time and dimensionality, it is frequently impossible to identify the underlying conformational equilibrium from an inspection of MD trajectories (time series) alone.[2−4] There is often a wide range of time scales involved, and many of the atomistic details are primarily a source of noise. Thus, it is a natural goal to compress the trajectory into a finite number of states, which simplifies the comprehension of the system and might directly reveal the local structures that constitute the aforementioned equilibrium. To this end, various sophisticated clustering techniques have been developed to recognize both metastable and transition states for systems undergoing stochastic dynamics.[5−8] Other approaches expand these ideas by using proper objective functions to drive the agglomeration of results from an initial, fine partitioning of the phase space obtained by common clustering techniques.[9−14]

Most of the cited methods have either been applied to or directly rely on the implementation of kinetic models, such as Markov state models[15−17] (MSMs). MSMs have proven their merit for extracting thermodynamic, kinetic, and pathway information from suitable time series data. However, the deduction of an MSM from trajectories requires choosing many (hyper)parameters, which necessitates the ability to assess the quality of these models in comparative fashion. Recent developments have advanced this issue by introducing quantitative methods to assess the MSM performance in terms of resolving kinetics at a global level.[18,19]

In this work, we present a method that combines a compact visualization of trajectories with an efficient extraction of clusters. The SAPPHIRE plot (States And Pathways Projected with HIgh REsolution), introduced in refs 20 and 21, provides a comprehensive picture of all of the trajectory configurations. In this type of plot, the snapshots are rearranged and grouped according to their geometric similarity and subsequently annotated by suitable variables that highlight conformationally and/or kinetically homogeneous states as well as the dynamics between them. SAPPHIRE plots have been effectively applied to the analysis of both molecular systems[22,23] and neuronal networks.[24] We introduce here an algorithm, called SAPPHIRE-based clustering (SbC), for the identification of clusters, which relies on the annotations displayed by the SAPPHIRE plot. Unlike most common clustering techniques, SbC takes direct but nontrivial advantage of the temporal information provided by the time series.

The rest of the article is structured as follows. We briefly review the theory underlying the SAPPHIRE plot (Section 2.1) before describing in detail the SbC method (Section 2.2). The remainder of the Methods is dedicated to reviewing the tools

Received: June 12, 2020
Published: September 9, 2020

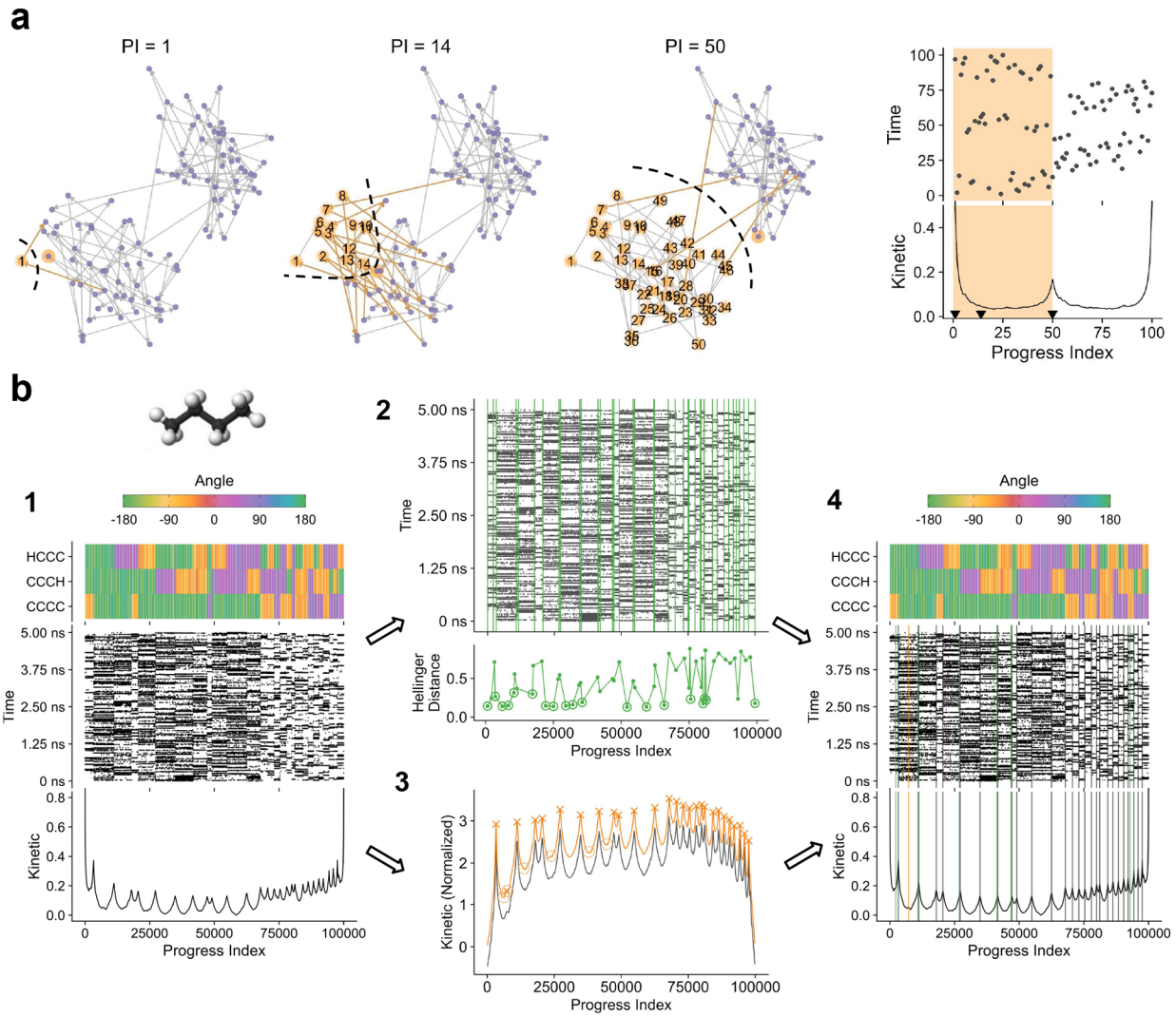


Figure 1. Progress index algorithm and Sapphire methodology. (a) Progress index construction. A time series traversing two different states is shown (violet dots connected by gray edges) in a two-dimensional phase space. PI values are assigned to the different snapshots following a minimum-linkage criterion (orange dots with numbering). The point that is next in line to be assigned in each image is outlined in orange. On the right, the resultant Sapphire plot displays the actual temporal index as a function of the assigned progress index values in the "Time" annotation. The "Kinetic" annotation decreases monotonically with the number of temporal edges that separate the points indexed by PI from the unassigned ones (orange edges in the three pictures to the left, which correspond to the little triangles). (b) Sapphire-based clustering (SbC). An example Sapphire plot for *n*-butane is shown in (1). Initially, clusters are separately identified in the Time annotation (2) and in the Kinetic one (3). In (2), the algorithm yields a set of partitions (vertical green lines) based on a 2D histogram of the dots. An initial partitioning is tested against a null hypothesis H_0 with a test statistic related to the Hellinger distance between adjacent states (lower panel). The partitions not compatible with H_0 are kept while the others are discarded (circled dots). In (3), putative states are identified by a peak identification procedure on a normalized and filtered kinetic annotation (orange line). Finally, in (4), the two sets of partitions from (2) and (3) are matched (black lines) and/or joined (green and orange lines).

and methods we chose for a quantitative comparison of the resultant MSMs as well as the data sets. Our tests are performed on three different systems: a toy model (Section 3.1) and two polypeptide systems, which are a medium-sized protein and a β-sheet-forming peptide (Sec. 3.2). We find that SbC yields results that are robustly competitive with other state-of-the-art techniques, thus establishing it as a useful tool for the quantitative investigation of time series. The article is completed by an illustration of the entire SbC workflow on recent MD simulations of the main protease of SARS-CoV-2 (Section 3.3) and by a concluding discussion.

2. METHODS

2.1. Progress Index and Sapphire Plot. For a full description of the progress index algorithm and Sapphire plot, we refer the reader to refs 20 and 21. For the sake of clarity, the algorithm is also illustrated schematically in Figure 1a. We start by considering a set of data points in a particular, usually high-dimensional space. This space is defined by the features we extract from the raw data set, and we refer to it as the feature space below. The progress index method rearranges the time frames (snapshots) into a new order, called progress index (PI), such that neighboring points are structurally similar in the selected space. Specifically, similarity must be defined in this space, and here, as in the original work, we always use the Euclidean distance as the metric of dissimilarity. While PI = 1

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

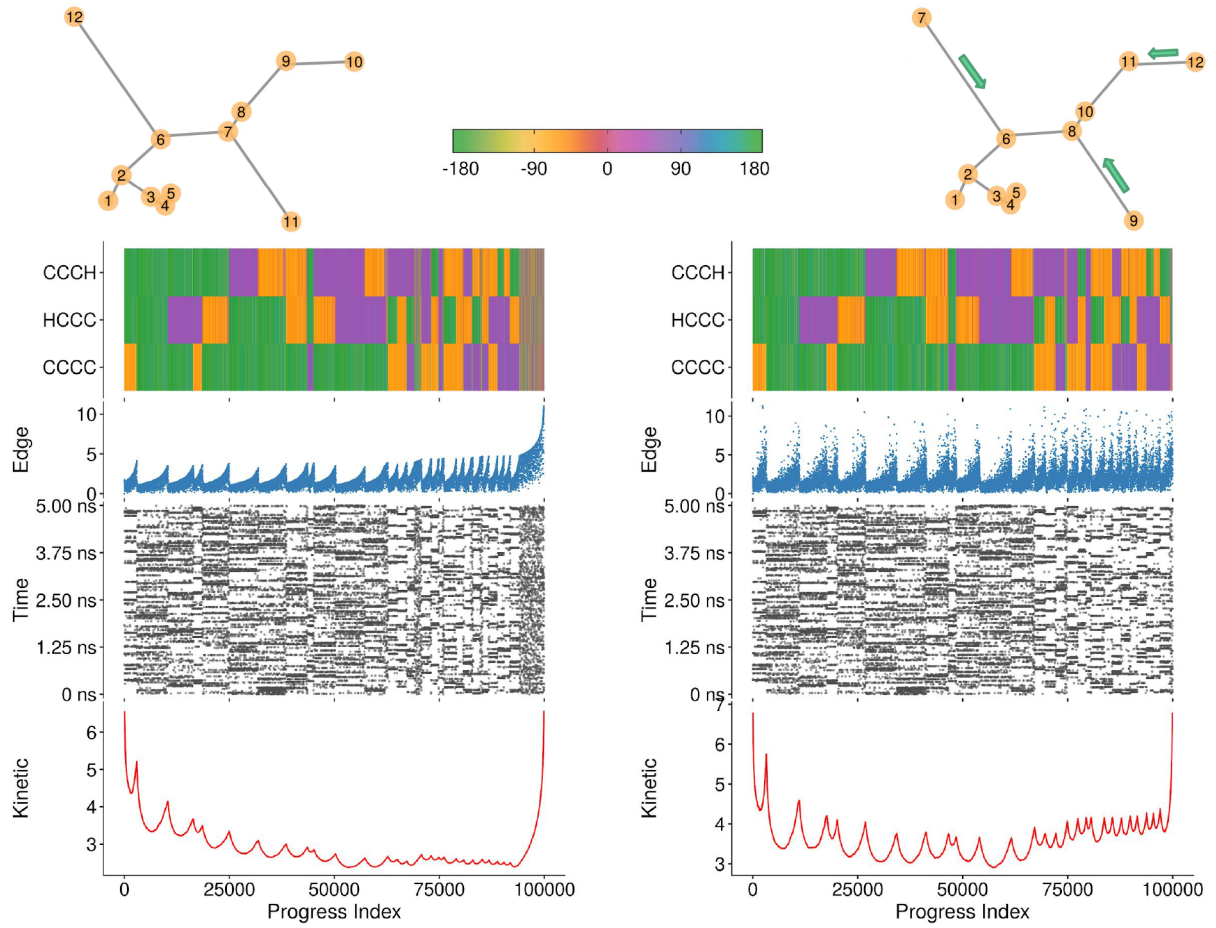


Figure 2. Sapphire plots of *n*-butane for two different values of the amount of leaves pooling, LP (0 and 10). In practice, the successive assignment of PI values shown in Figure 1a utilizes a spanning tree of the complete graph of snapshots. An LP value of n implies collapsing terminal branches of length n successively onto the $(n + 1)$th node. When the $(n + 1)$th node is added to the PI, all the pooled snapshots from connected branches are added immediately afterward.[25] An example of PI reindexing via LP is shown on the mSTs above the Sapphire plots (LP = 0 and 1 on the left and on the right, respectively). The Sapphire plots show, from bottom to top, the kinetic, temporal, edge length, and structural annotations (dihedral angles, color bar on top). The effect of LP is visible in the distribution of edge lengths ("Edge"). These are the edges on account of which a snapshot was added. The points found at the rightmost side of the plot without LP (left) are uniformly reassigned to their respective parent basins when LP is performed (right). A subsampling by a factor of 5 was applied along the PI axis due to reasons of plotting resolution.

is an arbitrary, initial choice, all subsequent indexes are assigned one-by-one by following a single-linkage criterion between the group of indexed points and those that have not yet been indexed.

In practice, the algorithm relies on the minimum spanning tree (mST) of the complete graph of snapshots where the edge lengths correspond to the geometric distances between pairs of snapshots. The mST is used for finding the snapshots to be indexed next. In most of the cases, we adopt an approximate version of the mST, called short spanning tree (sST).[20] Its construction makes use of a hierarchical clustering technique[6] whose parameters were tuned automatically according to *ad hoc* criteria. The overall settings are such that all PIs shown in this work are either exact (Section 3.1 and Section 3.3) or nearly exact (elsewhere).[25] A potential adjustment of the PI indexing can be attained by the prior pooling, or aggregation, of the mST leaves into the parent vertex. In detail, by setting n as the only parameter for this technique of "leaves pooling" (LP), the n outer vertices of any branch will be folded inward and collapsed onto the $(n + 1)$th parent node (Figure 2, top). Once the parent node is added by the PI algorithm, all the pooled nodes are indexed consecutively to it. In the case of different branches collapsing onto the parent node, leaves connected by the shortest edges are added first.[25]

The progress index algorithm is implemented in the software CAMPARI (http://campari.sourceforge.net/). A wrapper of the original Fortran code has been used in the analysis (R package "CampaRi"). This includes also an implementation of the SbC algorithm and is available on our public GitLab repository (https://gitlab.com/CaflischLab).

Sapphire Plot Annotations. A simple and informative feature that can be plotted with respect to the PI is the original time indexing. We will refer to this as the temporal or "Time" annotation. Another helpful variable is the "Kinetic" annotation. Given a PI = n, this annotation is inversely related to the number of transitions between the entire sets of points to the left and to the right of n. Often, it is more informative to count the number of transitions traversing a PI-local neighborhood of n.[20] When this is the case, we refer to this is as the "local" kinetic annotation, rather than as the aforementioned "global" one. The local kinetic annotation will be the default choice unless stated otherwise. The size of the local PI neighborhood is set to 10% of the data set size

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

throughout the analysis, which is an empirical rule found to work sufficiently well for the present work.

2.2. Sapphire-Based Clustering Algorithm. The SbC method employs only the temporal and the kinetic annotations of the Sapphire plot, that is, only those variables that are derived purely from the mapping from (simulation) time to the position in the PI sequence. The algorithm relies on the property of the PI of stepping sequentially through groups of nearby points in the feature space. In practice, we will identify the states by placing barriers, or partitions, along the PI sequence. In the initial steps, the algorithm extracts the clusters independently from the temporal and kinetic annotations, respectively, and only at the very end both sets are merged.

2.2.1. Clustering from the Temporal Annotation Alone. Because of the properties of the PI ordering, the temporal annotation consists of a "blocky" scatter plot. Putative states are identifiable visually because each block is, ideally, a visit of a kinetically homogeneous state in time. Along the PI axis (x-axis), putative states are delineated while on the Time axis (y-axis) the transitions between them and the possible recurrence of visits are highlighted.

2-D Histogram. To account for this particular structure of the temporal annotation and to reduce the computational effort, an underlying 2-D histogram is created, and only the bin frequencies are used during the analysis (see Figure S1a). In principle, the bin size on the x-axis, ΔPI, has to be selected according to the smallest cluster size that we want to identify, whereas the size on the y-axis, Δt, has to be related to the smallest residence time that we want to resolve.

Temporal Stretches. As a first step, we identify on each row of the histogram a stretch of consecutive bins (Figure S1a). Ideally, such a stretch indicates a visit of the putative state in that particular Δt window. Given f, the vector of frequencies of a selected row, two neighboring bins, i and j, are considered to have a similar and finite density if both bins receive nonzero counts and their frequencies are within 50% of each other's value. We denote the minimum and the maximum indices in one of these stretches as r and s, respectively. This allows us to define the center of "mass" (sampling weight), m, of the stretch delimited by r and s as $m = \sum_{k=s}^{r} f_k\, k / \sum_{k=s}^{r} f_k$.

For the subsequent steps, each stretch is assigned a weight that is meant to describe the fidelity with which it captures a residence interval of only one state. If the progress index groups snapshots by the underlying state perfectly, and the states are perfectly homogeneous, we expect the stretches to be balanced, i.e., $m \simeq \frac{s+r}{2}$. Second, we want to heuristically penalize stretches that are too long along the progress index axis since these are likely to include multiple states due to a faulty detection of similarity in the previous procedure. We account for these properties by computing the following weight

$$w = \frac{\min(m-s, r-m)}{\max(m-s, r-m)} \cdot \left(1 - \frac{r-s}{n_{\mathrm{PI}}}\right)$$

where n_{PI} is the number of bins along the PI axis.

Initial Partitioning. We assemble the PI values of the left and right extremities of the stretches of bins with similar frequencies, s and r, into the sets S and R, respectively. If a putative state is visited several times, we would expect that histograms of S and R, weighted by the aforementioned quality measure, should be sharply peaked at the transition points on either side of that state. However, in practice, these peaks will not be perfectly sharp and, more importantly, skewed to the right for the set S and to the left for R. The reasons for this skew have to do with the PI itself, with the actual sequence of state visits in the trajectories, and with the binning. Due to the expected difference between these two distributions, both sets are initially analyzed separately. Cumulative distributions, χ, are computed by integrating from left to right R and from right to left S. We then convolute χ with a Haar-like wavelet $H = [1, 1, 1, -1, -1]$, and eventually we follow a naive peak identification criterion on the resulting smoothed profiles (see Figure S1b). A point is identified as a peak and, thus, as a partition between two states if it is strictly larger than its two adjacent points. A first rough clustering of the trajectory is obtained by joining the two sets of partitions obtained from R and S.

Selection of Clusters. Given an initial set of partitions from the temporal annotation, we test the significance of these partitions by comparing each pair of adjacent candidate states with their reshuffled versions. In detail, given a PI range restricted to a pair of adjacent clusters, we randomly shuffle the PI values within each histogram row Δt. The Hellinger distance between the two resulting time histograms is computed. It measures the similarity of two distributions as the L^2 norm of the difference of the individual square root vectors divided by $\sqrt{2}$. This procedure is repeated 50 times, thus delivering a distribution of Hellinger distances that represent a numerical null model in which the points form a single state. A one-sided Grubbs test is used with a significance level $\alpha = 0.005$. The Grubbs test has the null hypothesis that the data contain zero outliers and is applied to an individual data point (here, the actual Hellinger distance) relative to an assumed normal distribution (here, the one derived from the numerical null model). If the actual Hellinger distance is indeed deemed to be a (right) outlier, the partition is kept; otherwise, the two clusters are joined together; see Figure 1b, panel 2, for an example. Both the Hellinger distance and the Grubbs test were chosen for performance reasons after a broad search trialing several comparable methods.

2.2.2. Clustering from the Kinetic Annotation Alone. The basic idea is much simpler for the kinetic annotation than for the temporal one because the peaks of the kinetic annotation are expected to directly highlight the transition points between states. For this, it is largely inconsequential that their actual values cannot be used to quantitatively infer free energy differences between basins.[20] A two-pronged approach is beneficial because the kinetic annotation can easily delineate states that might be obfuscated in the temporal annotation, for example, if adjacent states in the PI are also adjacent in time.

One of the disadvantages of the global kinetic annotation is that failures in identifying small states can occur in regions of inherent curvature, i.e., at the extremities of the Sapphire plot. In order to prevent this, we subtract from the kinetic annotation the parabolic curve derived from assuming a random exploration of the phase space.[20] From now on, for simplicity, the resulting curve will also be called the kinetic annotation and denoted as k.

Smoothing Filter. To deal with the rugged surface of the kinetic annotation curve and to comply with the resolution of the previous analysis, we employ a Savitzky–Golay filter with a window length equal to $2 \times \Delta$PI. ΔPI is exactly the same bin width as that chosen for the 2D histogram in Section 2.2.1. We will use as polynomial degrees of the filter both one, which corresponds to a moving average filter, and two. Subsequently,

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

we perform the naive peak identification introduced in Section 2.2.1 with a window size equal to 2 × ΔPI.

Peaks Check and Projection. A simple heuristic is implemented to test the goodness of the identified peaks. We denote as m_i and m_{i+1} the absolute minima of k_{sm}, the smoothed kinetic annotation, in the two adjacent states delimited by the three peaks (l_{i-1}, l_i, l_{i+1}). We then compute the following ratio

$$D_i = \frac{\langle k_{sm}(l_i) - k_{sm}(m_{i-1}),\ k_{sm}(l_i) - k_{sm}(m_{i+1})\rangle}{\langle k_{sm}(l_{i-1}) - k_{sm}(m_{i-1}),\ k_{sm}(l_{i+1}) - k_{sm}(m_{i+1})\rangle}$$

where $\langle\cdots\rangle$ indicates the average of the two arguments. If the ratio D_i is lower than 0.05, the partitioning represented by peak l_i is discarded; see Figure 1b, panel 3, for an example. The remaining peaks are projected back onto the original kinetic annotation, k. Each maximum l_i is shifted to the absolute maximum of k in the interval $[l_i - \Delta PI, l_i + \Delta PI]$. As a final step, we check whether any two suggested partition boundaries are closer than ΔPI; if so, the one with the smaller value in k is discarded.

The temporal and kinetic annotations provide two separate clustering results. They differ conceptually in their resolution: in particular, the partitions obtained from the Time annotation are placed discretely with a step size of ΔPI whereas the kinetic ones are identified within snapshot resolution. To make them homogeneous, we first shift the Time partitions to the absolute maximum of the kinetic function within a neighborhood of 2 × ΔPI. The two sets are then merged, and the barriers are matched if they are closer than ΔPI, retaining only the one with the highest kinetic annotation. All the partitions, matched or not, are used in the analysis. This is important; i.e., we are not looking for a consensus set but rather for an exhaustive set. This is because both annotations on their own can lead to false negatives as outlined above.

2.3. Theoretical Framework: Markov State Models and VAMP Scores. Markov state models (MSM) are a powerful framework to extract thermodynamic and kinetic properties of a system from MD simulation data.[15,16] In brief, MSMs utilize discrete trajectories to infer a transition matrix $\mathbf{T}(\tau)$ that succinctly describes the propagation of the system across different states. Each element T_{ij} indicates the conditional probability of reaching state i, starting from j, in a lag time τ. Applying $\mathbf{T}(\tau)$ to a probability distribution vector at time t, one can obtain the probability distribution at $t + \tau$. This master-equation process is thus memoryless, i.e., Markovian, and the chosen lag time defines an intrinsic lower bound for the time scales that can be resolved.

The field of kinetic modeling has recently been extended by the introduction of variational principles, which are used to find the representation, or model, that optimally approximates the slow dynamical processes of a system.[26,27] One of the main advantages of this approach comes from the availability of specific scores that allow for an objective comparison between different models and, in turn, for a properly guided selection of hyperparameters.[18,28] Among these scores, we adopted in the analysis those defined by the so-called variational approach for Markov processes (VAMP).[19] In brief, a Markov process can be described by the Koopman equation,[29,30] which generalizes the MSM master equation mentioned above. The Koopman equation consists of the application of a linear operator, the Koopman operator, on a suitably transformed feature space, in order to provide the time evolution of a system in another (transformed) feature space. The optimization of the top singular values of the operator, related to the slowest modes and summarized in the VAMP score, is thus largely related to the search of suitable transformations of the input feature space. VAMP scores can not only provide comparisons between different discretizations of a trajectory but can also operate directly on nondiscretized featurizations of the trajectory. This allows, e.g., a comparison of the inherent ability of different features like dihedral angles or contact probabilities to capture the selected modes (see refs 27 and 28.).

2.4. MD Trajectories, Procedures, and Settings. *2.4.1. MD Trajectories and Preprocessing.* The first system investigated is the 58-residue bovine pancreatic trypsin inhibitor (BPTI) simulated in its native state. We analyzed here the 1.03 ms MD trajectory of Shaw et al.[31] with a sampling time of 25 ns. The sine and cosine of 271 dihedral angles, which exclude only those χ-angles where multiple values map to the same conformation due to the presence of symmetrical substituents (like χ_2 in phenylalanine), were extracted for further preprocessing. Despite this manual feature selection, as shown in prior work,[22] feature weights or dimensionality reduction are still needed to yield an informative feature space. Here, we use the latter (see below).

The second system we analyze is Beta3S, a 20-residue peptide that folds into a three-stranded, antiparallel β-sheet as its native state.[32] The data are a concatenation of 10 MD trajectories, each of 2 μs length and recorded with a time resolution of 0.2 ns. Here, we used the sine and cosine of all of the available 103 dihedral angles.

For both systems, the extracted features were transformed with tICA[33,34] while also applying a published kinetic mapping scheme.[35] To reduce dimensionality, we retained only the first 10 tICA components for subsequent analysis. The autocorrelation lags used for BPTI and Beta3S were 500 and 4 ns, respectively.

In the final part of the article, we examine a trajectory of the *apo* form of the main protease of SARS-CoV-2[36] (PDB entry 6Y84).[37] The total sampling length is 100 μs, and snapshots were saved every 1 ns. This enzyme is a homodimer with each chain composed of 306 residues. We preselected 865 of the available dihedral angles manually, which represent well over 50%. Without preselection, the density in the feature space is not sufficiently informative even after an aggressive dimensionality reduction. We ignored the following classes of dihedral angles: all dihedral angles in the 10 terminal residues on both ends, all χ-angles in charged residues except χ_1, all ω-angles, the χ_2-angles in Ser, Thr, and Cys, the $\chi_{2/3}$-angles in Tyr, and lastly the χ_2- and χ_1-angles in Phe/Leu and Val, respectively. The omissions for Phe, Leu, Val, and Tyr are due to symmetry. As for the previous systems, we utilized the data set of the sine and cosine of these angles for further preprocessing and analyses.

2.4.2. Clustering Methods. The discrete state space obtained by SbC is compared to major methods employed in the current literature. In detail, we compare to k-means, k-medoids, and a hierarchical clustering with Ward's method.[14,38] For simplicity, we will refer to the latter as "Ward".

2.4.3. Markov State Models and VAMP Scores. Transition matrices were inferred via maximum likelihood estimation from the MD trajectories, and we did not impose detailed balance. We chose MSM lag times for BPTI and Beta3S of 500 and 20 ns, respectively. The VAMP-2 scores account for the

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

first 10 singular values, and they were computed also for the nondiscretized input features, i.e., sine and cosine values of the selected dihedral angles. All VAMP computations were cross-validated with a 10- and 50-fold splitting of the discrete trajectories for BPTI and Beta3S, respectively. Only the averages of the test-set scores are shown in the results. pyEMMA 2.5[39] was used for Markov state modeling and calculations of related quantities.

2.4.4. Mean First Passage Times. Mean first passage times (MFPTs) were evaluated between the two macrostates with the largest statistical weights. In detail, for BPTI we used the two biggest macrostates identified by Shaw et al.[31] For Beta3S, the two macrostates were the three-stranded antiparallel β-sheet configuration (the folded state) and an ensemble of states rich in α-helical content. The latter was defined as the set of snapshots featuring a sum of at least five α- or π-assignments according to DSSP across residues.[40] The annotations thus provide a set of labels from which we can estimate MFPTs from the trajectories without further processing. This estimation, also called "direct" estimation, was performed by counting the periods of transitions between the two sets of clusters and averaging over the first passage times observed.[41,42]

Conversely, for a clustering result, the labels derived from annotations are not necessarily homogeneous within a given cluster. Thus, we require a strategy on how to assign macrostate labels to the clusters of a given discrete trajectory. To do so, for both of the two macrostates, we selected the minimum set of clusters that contains at least 80% of the macrostate snapshots. Here, minimum refers to the number of clusters in the set. These two sets of clusters should largely encompass the two macrostates, but they should not overlap with other macrostates. MFPTs are expected to be insensitive to the exclusion of some snapshots in a basin since there is a large separation of time scales for relaxation within and between states. In contrast, MFPTs can change dramatically upon the accidental inclusion of snapshots from a different basin (effectively causing a kinetic shortcut). The MFPTs between the two sets of clusters were calculated from the MSMs by solving the related linear system of equations.[32,43] In addition, the two sets of clusters were used to redefine the macrostate labels: unlabeled snapshots in selected clusters were labeled, and labeled snapshots outside of the selected clusters were unlabeled. On these modified trajectories of labels, the direct calculation mentioned above was performed as well.

3. RESULTS AND DISCUSSION

3.1. Application to *n*-Butane. We apply the SAPPHIRE analysis to an *n*-butane trajectory of 10^5 snapshots with a sampling time of 50 fs. The system has three degrees of freedom, namely the dihedral angles associated with the three carbon−carbon bonds. For the sake of PI construction, we measure the geometric distance as the Euclidean distance in this three-dimensional space. The system has access to $3^3 = 27$ metastable states, representing all the possible combinations of the three stable configurations of each angle (centered at 180°, 60°, and −60°, respectively). In Figure 2, we show two applications of the PI algorithm, with corresponding SAPPHIRE plots, with values of leaves pooling (LP)[25] equal to 0 and 10. A substantial difference between the two profiles is found in the presence of the large and structurally inhomogeneous "state" visible on the rightmost side of the left panel. The annotations provided by the SAPPHIRE plot clearly identify this region and, among them, the "Edge" annotation is particularly revealing. This annotation represents the length of the mST edge by which a snapshot was added in the PI. The left panel of Figure 2 shows that higher values of Edge are present in the aforementioned PI region on the far right as well as on the right fringe of each basin. The right fringe region corresponds to the transition and peripheral regions of the metastable states to the left. For high values of LP, these low-density regions are redistributed homogeneously within their parent metastable states (right panel of Figure 2). Conversely, for LP = 0 (left panel), points that are further away from their parent basin than *ca.* 5 distance units are skipped and not added until the very end. This is why the aforementioned state is not actually a state but rather a collection of points from many different low-density regions.

The consequences of this reordering are apparent from the kinetics of the MSM constructed based on the SbC results. In Figure 3 the first two implied time scales are shown both for

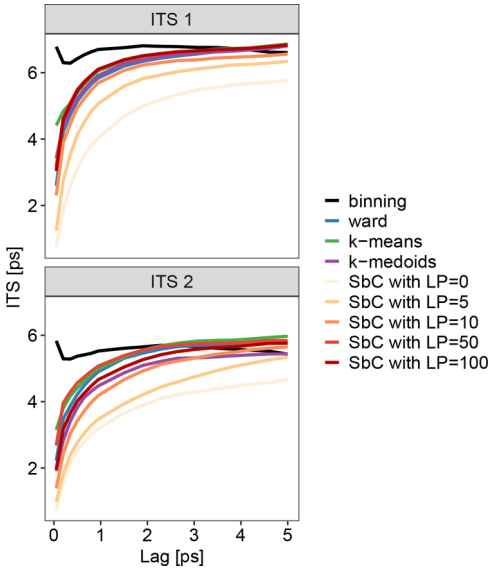


Figure 3. Implied time scales (ITS) of *n*-butane from SbC for different LP values as well as from other clustering methods. Values computed on a binned phase space (12 bins per angle) are shown as a reference. The other implied time scales are computed from MSMs based on 27 clusters. For SbC, we analyzed the global kinetic annotation with a moving average filter. For each LP value, SbC was repeated 50 times with random settings of n_{PI} and n_T, both ranging from 100 to 500. Results are shown for one partitioning selected at random from the subset of results yielding exactly 27 states.

SbC with different LP values and for other clustering methods as a comparison. All the MSMs are built with 27 states (see the caption of Figure 3 for details). As a reference, a gold standard MSM is constructed using the binned phase space (12 bins for each angle). As Figure 3 shows, the accurate resolution of time scales is challenging even for a toy model if the number of states is small. The discretization error is not fully overcome by any of the tested methods, especially at shorter lag times. For SbC, it is evident that an LP of about 10 is sufficient to improve the kinetic fidelity of resultant MSMs to approach the same performance as that of the other methods. This is mostly due to the elimination of the diffuse low-density region that was previously acting as a kinetic shortcut for many transitions between states. A comparison of the two panels of Figure 2 clearly hints at this effect because the lower height of the

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

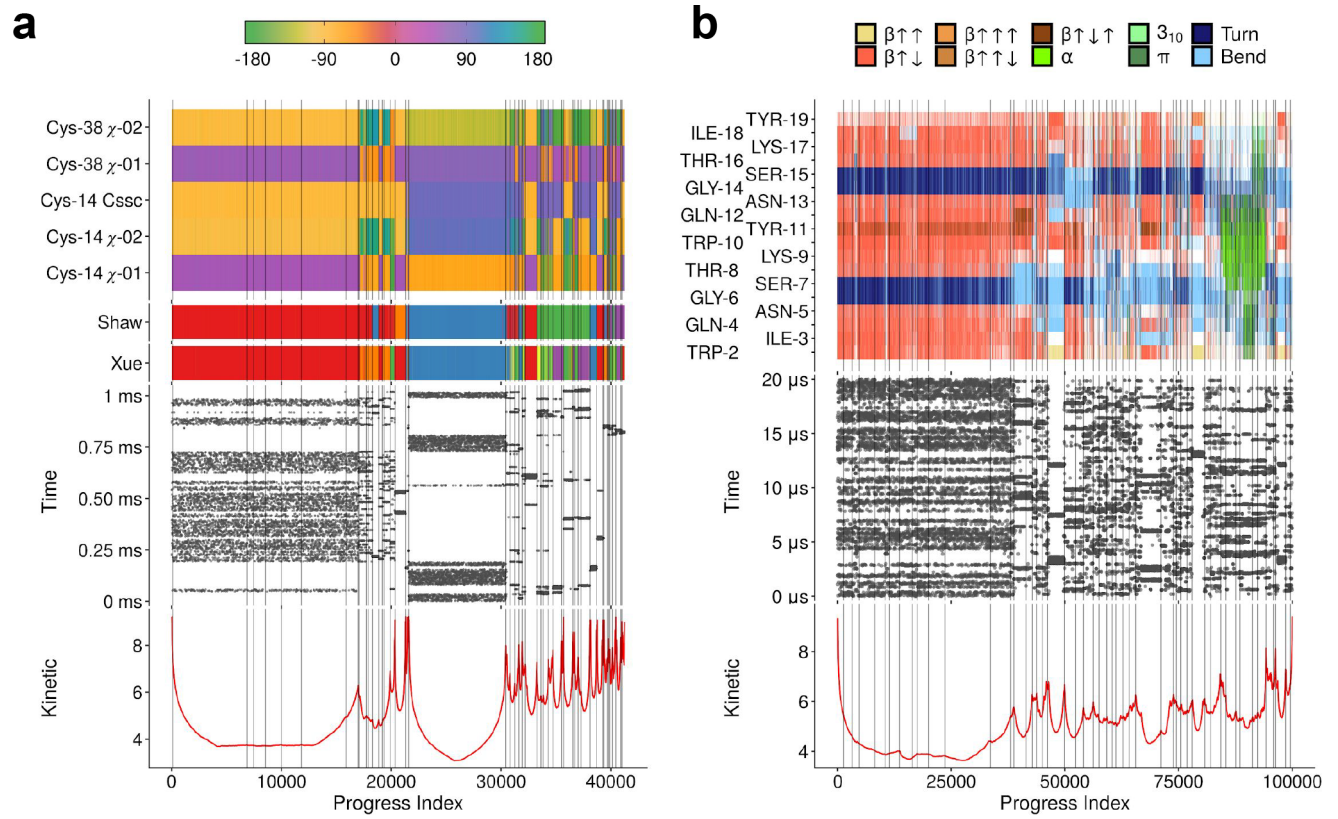


Figure 4. SAPPHIRE plot of MD trajectories for BPTI (a) and Beta3S (b). An example of SbC is shown in both figures by the gray vertical lines. The kinetic and temporal (Time) annotations are the red curve and the gray dots (bottom and center), respectively. The annotations stacked on top differ in the two panels. For BPTI, we show as a reference the clustering results obtained by Shaw et al.[31] (red, blue, green, magenta, orange, and yellow for states 0−5) and Xue et al.[44] (M1, blue; M2, yellow; M3, magenta; m_{C14}, red; m_{C38}, orange; other states, green). Above, the five dihedral angles of the Cys14−Cys38 disulfide bridge are plotted (color bar on top). For Beta3S, we show instead the DSSP annotation per residue (color legend on top). Due to limitations in plotting resolution, a subsampling by factors of 2 and 5 was applied along the PI axis for BPTI and Beta3S, respectively.

barriers between states in the kinetic annotation reflects these kinetic shortcuts in the absence of LP. This is an illustrative example of the beneficial effect of LP on the performance of SbC-derived MSMs. In light of this, we paid attention for the other systems to set a LP value sufficiently large to adequately capture the kinetics. As a practical note, we emphasize here that the values for LP should be dimensionality-dependent (the smaller, the larger).

3.2. Application to BPTI and Beta3S. We subsequently applied the SAPPHIRE methodology to two simulation data sets for proteins. The first is a very long MD (1.03 ms) simulation of the folded state ensemble of bovine pancreatic trypsin inhibitor (BPTI),[31] a 58-residue protein stabilized by three disulfide bridges. The second is a concatenation of 10 MD simulations (20 μs cumulative)[32] of the reversible folding of Beta3S, a 20-residue peptide adopting a three-stranded β-sheet conformation along with several misfolds. We analyzed snapshots saved every 25 and 0.1 ns for BPTI and Beta3S, respectively. In both cases, the distance measure used for the construction of the PI is the Euclidean distance in the 10-dimensional space of those tICA components with the largest autocorrelation values at lags of 500 and 4 ns for BPTI and Beta3S, respectively. tICA was applied to data sets of sine and cosine values of extracted dihedral angles; see Section 2.4.1. SAPPHIRE plots of the two trajectories are shown in Figure 4 along with an example of the partitioning of the plot into coarse clusters. The structural annotations were chosen to highlight the main structural changes of the systems across the different states. For BPTI (panel a), we add a comparison to the original kinetic clustering results derived by Shaw et al.[31] along with the partitioning suggested in Xue et al.[44]

It is well-known that the slow dynamics of the simulated folded state of BPTI are largely driven by the disulfide bridge Cys14−Cys38 (Figure 4a, upper panel), which is sufficient to distinguish most of the states identified by SbC. The large basin on the left splits further, which is not evident from the disulfide bridge annotation. The shape of the kinetic trace suggests that this state is not perfectly homogeneous; i.e., there are comparatively fast interchanges between substates that must be mapped to other degrees of freedom. The recognition of such shallow maxima could be avoided by adjusting the peak detection heuristic described in Section 2.2.2, but we did not deem it necessary for the results presented here.

For Beta3S (Figure 4b), the states are characterized globally, which we highlight using a comprehensive DSSP annotation.[40] The most populated state is the native fold, i.e., the three-stranded β-sheet configuration (large basin on the left). The dominant misfold is a mainly α-helical set of states (right region, visible in green in terms of secondary structure annotation). The remaining states are partially folded variants of the native state, e.g., single β-hairpins. Just as we observed for BPTI, the largest state is not fully homogeneous. As we showed previously,[45] the native basin has kinetically well-defined substates differing in specific χ-angles that are easily

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

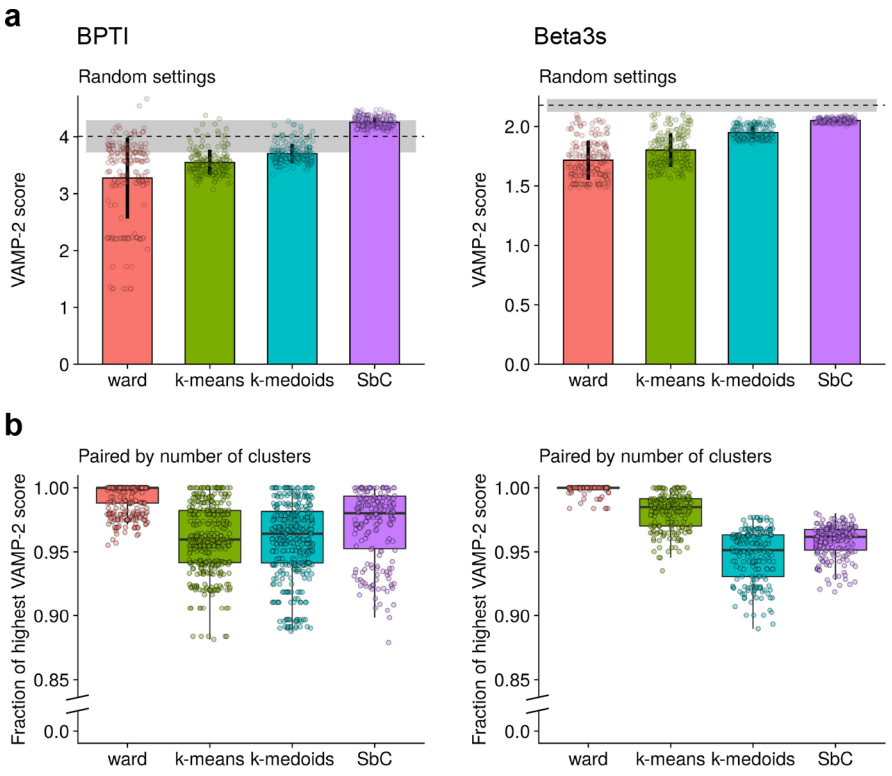


Figure 5. VAMP-2 scores of BPTI and Beta3s for different clustering methods and settings. The first 10 singular components are used for the calculation. The dots are based on mean test scores from a 10- and 50-fold cross validation for BPTI and Beta3S, respectively. For each method, dots are splayed out horizontally to improve readability, but this horizontal (sub)axis has no meaning. (a) VAMP-2 scores for four methods when using variable clustering settings. Clustering parameters were sampled 200 times for each method (see Section 3.2.1 for details). Bars and whiskers represent average and standard deviation of mean test scores across the different clustering settings. The dotted line and the gray stripe represent VAMP scores obtained directly from the original features (sine and cosine of dihedrals). They correspond to the mean and standard error interval, respectively. (b) Relative VAMP-2 scores for matched numbers of clusters. Ward, k-means, and k-medoids were performed with the same number of clusters obtained by SbC for each of the 200 SbC samples. The dots represent the fractions of the highest score among the four methods. Tukey-style box plots are plotted underneath. Note that Ward is a stable algorithm that will always give the same results for the same number of clusters.

obscured.[6] Here, the use of tICA and reduction to 10 dimensions ensure their discovery.

3.2.1. VAMP Analysis. We next compared our results from SbC with other clustering methods by using the so-called VAMP scores (see Section 2.4.3) within the MSM framework. These scores are expected to quantify how well the slow dynamics of the system are approximated by the underlying MSM. The MSM was built on the state space partitioning delivered by the clustering methods. The type of algorithm and the clustering resolution are known to be dominant determinants of the accuracy of an MSM.[16,46] First, we evaluated the VAMP scores for each method on a set of 200 different partitionings. For SbC, we tested random values for three different options, namely, a number of bins along the PI, n_{PI}, and Time, n_T, ranging from 200 to 700, and the choice of the smoothing filter, either a moving average or a Savitzky–Golay filter with degree two. For BPTI, we obtained a number of clusters ranging from 17 to 92, with a median value of 50. Regarding Beta3S, we identified 23–132 states with a median value equal to 67. As is evident from Figure S2, the primary determinant of the number of clusters discovered is n_{PI}, suggesting that SbC results with many clusters differ from those with few primarily in terms of clusters with low overall sampling weight. The target numbers of clusters for the other methods were also varied: we chose random values between 20 and 2000 for BPTI and between 20 and 5000 for Beta3S (for k-medoids only between 20 and 1500 for computational reasons in both systems). We adopted MSM lag times of 500 and 20 ns for BPTI and Beta3S, respectively.

In Figure 5a, we show the VAMP scores computed across the aforementioned settings. We also display, as a reference, the VAMP scores computed on the original data set of sine/cosine values of dihedral angles (no tICA and no dimensionality reduction). With little differences among the various methods, SbC performs slightly better on both systems. Taking into account also results from Husic and Pande,[38] we analyzed the VAMP scores within the low range of clusters found by SbC. For each SbC partitioning, we set the corresponding number of clusters as the target for the other methods, reclustered the data, and recomputed the VAMP scores. Results in Figure 5b partially alter the trends between methods, revealing, as in ref 38, that a low number of states is favored by VAMP. Nonetheless, SbC remains the best performing method along with Ward for BPTI, and it yields values within 95% of the top scores for Beta3S.

3.2.2. Analysis of Mean First Passage Times. The VAMP scores account cumulatively for the first slowest modes of the system providing, therefore, little or no specific information about individual processes, which might be of particular interest. In our case, for BPTI, we wanted to focus on the transitions between the two largest regions of the native state, namely, the red and blue states of the Shaw et al. classification in Figure 4a. Regarding Beta3S, we investigated the transitions between the native fold (the three-stranded β-sheet) and the

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

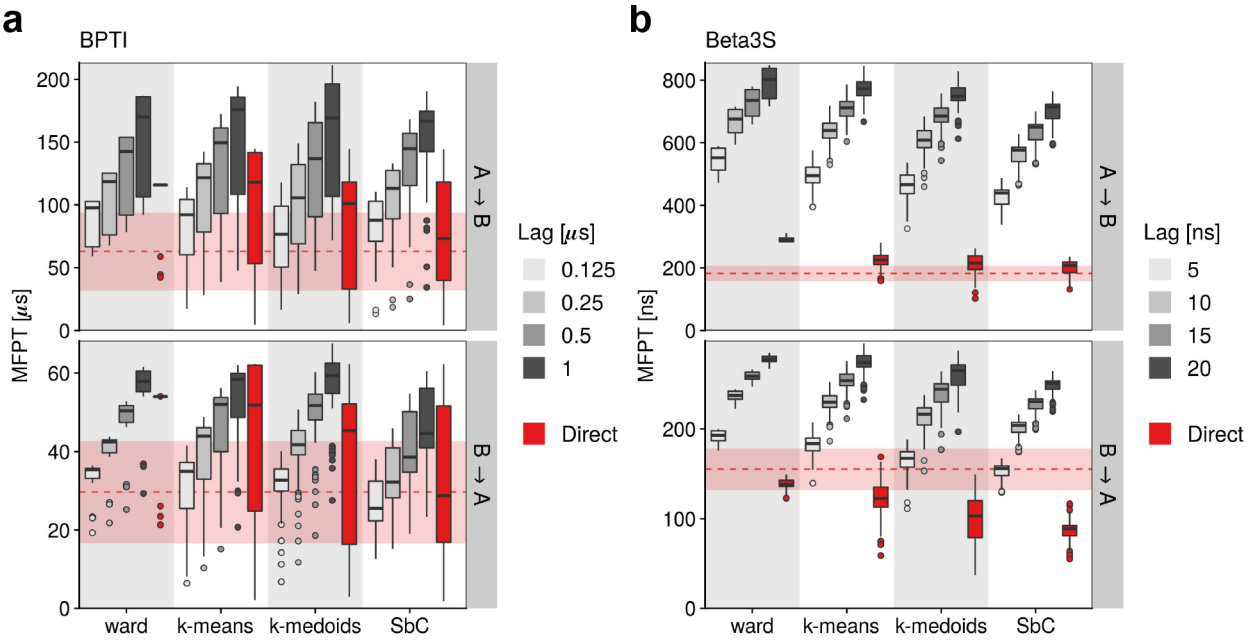


Figure 6. MFPTs of BPTI (a) and Beta3S (b) between the selected macrostates computed in three different ways and for both directions (top and bottom). For BPTI, A and B indicate, respectively, the red and blue state of the "Shaw" annotation in Figure 4a. For Beta3S, A and B represent, respectively, the three-stranded β-sheet macrostate and the α-helix-rich region (see Figure 4b and Section 2.4.4 for details). The gray, Tukey-style box plots summarize the results of MSM-based MFPT calculations across the same set of discrete trajectories analyzed in Figure 5b, i.e., 200 trajectories per method where the number of clusters was matched to the particular value found by SbC. The red box plots summarize MSM-free MFPTs obtained by reassigning labels A and B based on the selected clusters. Finally, the dashed lines indicate the MFPTs estimated on the original sets of labels, and the shaded regions indicate the corresponding standard error computed over the set of first passage times.

region high in α-helix content (see Figure 4b). Methodological details on MFPT computations can be found in Section 2.4.4.

We calculated MFPTs between the respective sets of reference states for both systems. We computed them not only from MSMs but also directly from the discrete trajectories where the state labels were assigned based on the selected sets of clusters (see Section 2.4.4). In the latter case, MFPTs can be estimated with maximal time resolution, i.e., with a lag equal to the sampling step. Because both estimates rely on the clustering, they are subject to the discretization errors resulting from the partitioning of phase space. Generally speaking, discretization errors lead to an underestimation of the time scales,[46] and their magnitude can be reduced by adopting a finer partitioning, especially across the transition regions.[16] On the other hand, this negative bias can be compensated by using higher values for the MSM lag time since rapid transitions between nearby states will be neglected.[42]

In Figure 6, we show MFPT values from SbC and other clustering methods for both types of computation, viz., through MSMs (gray scale) and from the trajectory of cluster indices (direct estimation in red). We analyzed the same set of clusterings used in Figure 5b, that is, with the number of states paired to SbC results. As an additional point of reference, the MFPTs estimated directly from the original set of labels are included as well (dashed horizontal lines). These numbers are not based on clustering and are a function of only the time sequences of the labels A and B assigned to the snapshots in the two selected macrostates. For BPTI (Figure 6a), regardless of the computation type, Ward exhibits generally smaller variances, proving to be robust with respect to the number of clusters. Ward is a strictly hierarchical agglomeration scheme, which means that the difference in two clusterings of N *vs* $N + 1$ clusters is that an additional pair of the $N + 1$ clusters has been merged, the others all being identical. This guaranteed mutual similarity is not present in the other techniques. For SbC, this is true not only because of the use of the sST but primarily because of the ability to choose n_{PI} and n_T; see Figure S2.

For all methods, the direct estimates from the clustered trajectories are generally consistent with the values computed directly on the labeled trajectory (dashed lines). This is valid in particular for SbC where median lines fall within the standard error of the reference values in both directions.

As expected, the MSM estimates increase steadily with higher lag times overcoming eventually, on average, the direct estimates. Already at lag times equal to 125 ns, for all the methods except Ward, the lower tails observed on the direct estimates are reduced, hinting at the presence of very short transitions and, thus, at problems with the underlying discretization. As expected, the data points corresponding to these lower tails largely correspond to small numbers of clusters ($\leq$22 upon visual inspection). Focusing on the same lag time used in VAMP scoring, τ = 500 ns, the overlap between the MSM values and either of the direct estimates is weak for almost all methods and both directions. However, this is, as alluded to above, expected: direct estimates are likely to be underestimates because the data are inferred from trajectories of finite length, and, more importantly, fast transitions, which might be indicative of "shortcuts" arising from utilizing discrete labels,[47] are filtered by choosing long lag times.

In Figure 6b, we show the MFPT estimates for Beta3S. They refer to transitions between the three-stranded β-sheet and the α-helical state. All the MFPTs estimates have smaller relative errors than those of BPTI. This is likely due to the higher

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

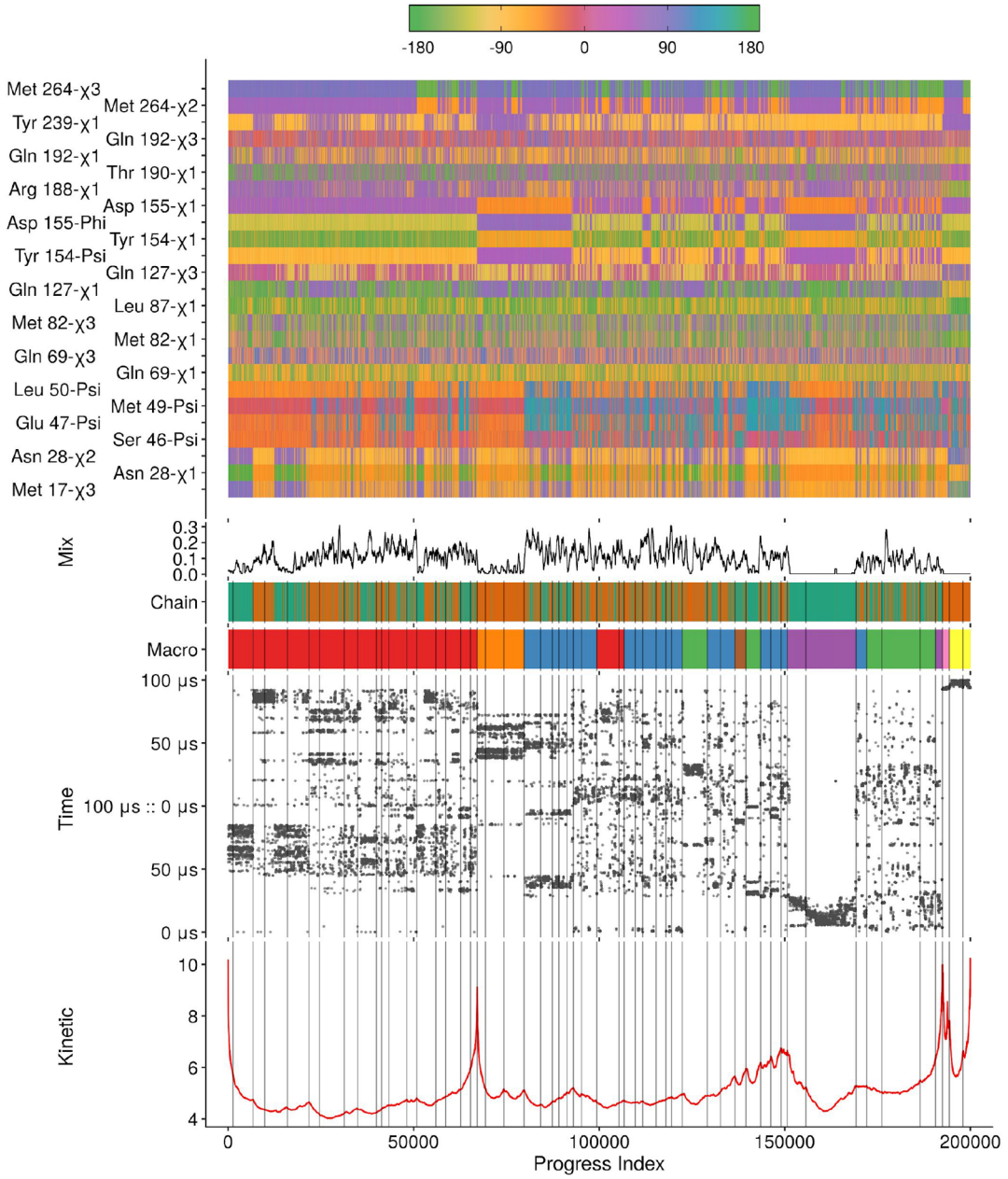


Figure 7. SAPPHIRE plot created from MD simulations[36] of the *apo* form of 3CL^{pro} (PDB entry 6Y84).[37] We isolated from the trajectories the two chains composing the homodimer and concatenated them in time. The 10 PCs with largest variance when ignoring the first one were calculated from a data set of 865 dihedral angles (sine and cosine) and retained for PI analysis. We adopted an LP value of 15. The dihedral angles offering the highest loadings to these 10 PCs are shown (structural annotation with color bar on top). The "Mix" annotation is the normalized number of interchanges between the two chains ("Chain" annotation below) in a centered PI window of 500 snapshots. The results of SbC are shown as vertical lines. In total, 54 clusters are identified when using 300 and 200 as the numbers of bins along the PI and Time axes, respectively (see Section 2.2). The "Macro" annotation indicates the eight macrostates identified by applying the PCCA+ algorithm to the SbC result using a lag time of 100 ns (see Figure 8 for visualizations of snapshots representing these states). For reasons of plotting resolution, a regular subsampling by a factor of 10 was applied along the PI axis.

number of transitions between the two macrostates under investigation (compare Time annotations in Figure 4).

While the shortest lag times evaluated here match the direct estimates best, there are clear indications that this reference is of limited use for Beta3S. First, the different types of direct estimates do not match consistently. Figure 6 relies on low numbers of clusters, but the free energy landscape of Beta3S is rich in many smaller yet mutually similar states (compare Figures 4a and b). Thus, a finer clustering is necessary in order to capture correctly the macrostate boundaries derived from the labels, in particular for the native β-sheet conformation of Beta3S. This is demonstrated in Figure S3 where the two types of MFPT estimates reconcile when using a sufficiently high number of clusters. Second, and this is the major concern, the MFPTs are very similar for both directions while the equilibrium state weights are not: the latter display a ratio of *ca.* 4; see Figure 4b and Figure S4. The MSM-based MFPTs reproduce this asymmetry much better at all lag times. The primary reason for the failure of direct estimates is the same one discussed for BPTI, i.e., shortcuts introduced by using

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

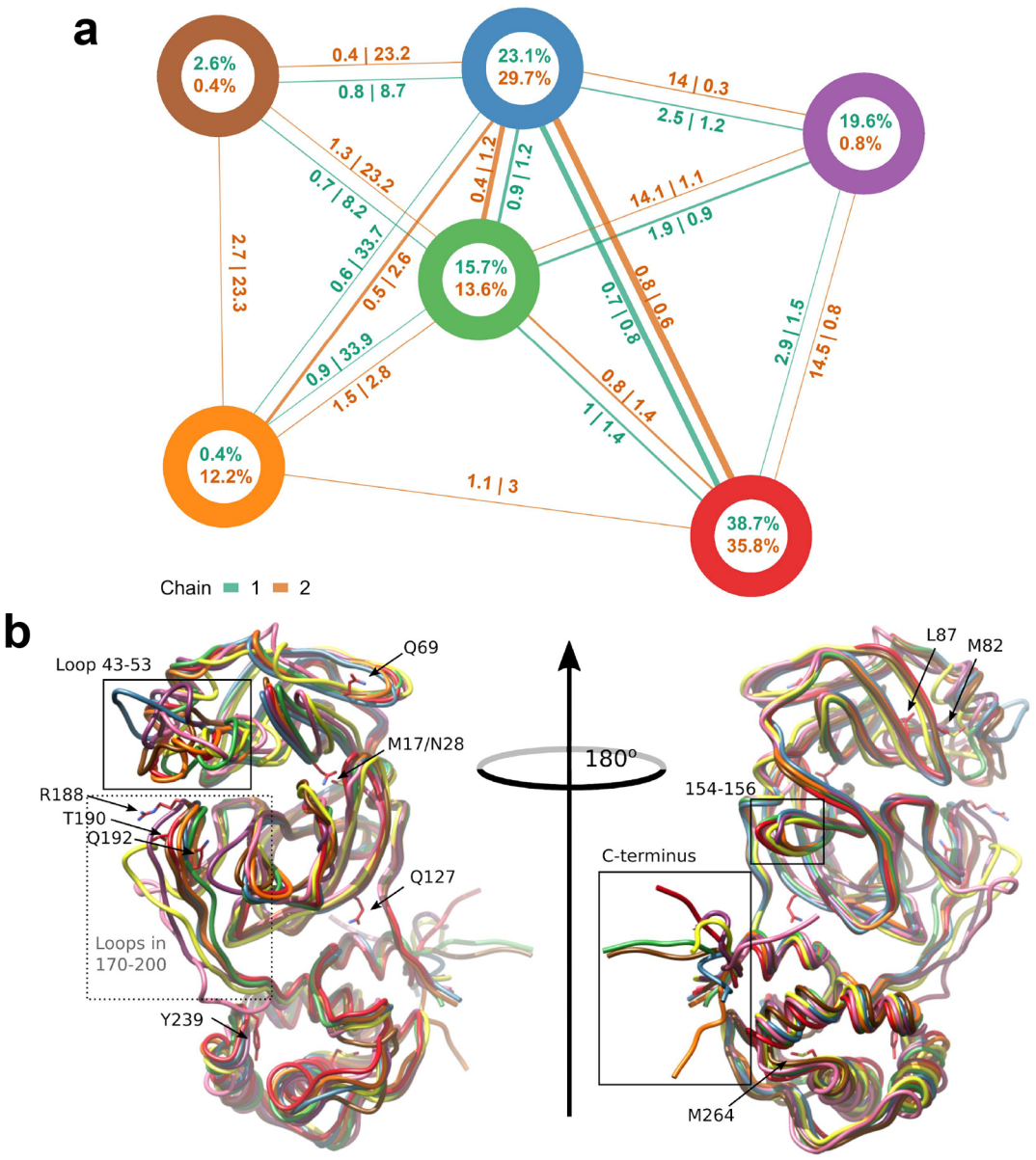


Figure 8. Kinetic network of 3CL^{pro}. (a) We show the six macrostates annotated in Figure 7 that are shared by both chains. The colors are the same. The relative weights of the macrostates and the MFPTs between them are added as labels to the vertices and edges, respectively. These data are obtained for both chains separately (color code at the bottom of (a)). MFPTs are shown for both directions in units of microseconds: the first value refers to the transition from left to right, the second to the opposite direction. Edges corresponding to less than 25 counts are omitted as is a transition between the brown and red states (chain 2 only, MFPTs 0.7/23.4). The MFPTs were inferred from a nonreversible MSM with a lag time of 100 ns. The edge widths are proportional to the number of transitions between the macrostates. (b) The snapshots closest to the centroid of each of the eight macrostates in terms of the metric used in Figure 7 were identified and are shown using the same color scheme. Note that the yellow and light pink states are not part of the network in (a). Chains were aligned using Cα atoms. Protein backbone conformations are depicted as tubes. Heavy atoms in side chains of specific residues appearing in Figure 7 are shown but only for the red macrostate. Residue stretches or individual side chains are annotated by legends and either arrows or boxes. The images on the left and right differ by a rotation of 180° around the vertical axis. Some side chains produce variance that does not appear to be structurally important like the contact pair Met17/Asn28.

labels that fail to resolve kinetically distinct states. For Beta3S, this is a well-known problem when using secondary structure annotations to identify states.[32,47] Notably, it is not possible to infer stringently from Figure 6 which of the tested lag times is most appropriate. However, the data suggest that shorter lag times are better for both systems. Tests for Markovianity do not provide unequivocal evidence for the use of either the shortest or longest lag time evaluated (Figures S5 and S6). In recent work, we observed similarly that indirect tests of Markovianity tend to suggest lag times that are too large when trying to match steady state probabilities with a ground truth.[48]

3.3. Analysis of SARS-CoV-2 Main Protease. In the final part of this work, we examine a publicly available MD trajectory of the unliganded main protease of SARS-CoV-2[36] (PDB entry 6Y84).[37] This system serves here to illustrate the entire SbC workflow to highlight the ease with which a very large and complicated data set can be mined to extract results ready for human comprehension. The total sampling length and time resolution of the data are 100 μs and 1 ns,

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

respectively. This enzyme, also called M^{pro} or $3CL^{pro}$, is a homodimer with the two chains oriented perpendicular to each other (see Figure S7). Each chain is composed of 306 residues grouped into three domains.[49]

For the Sapphire pipeline, the dynamics of the two chains were treated separately by concatenating the two subsystem trajectories. The resultant data set comprised 865 dihedral angles extracted from the central 286 of the 306 residues per chain arranged into 2×10^5 snapshots. The data set was preprocessed by applying a principal component analysis to the sine and cosine values of the dihedrals and retaining the 10 principal components (PCs) with largest variance but skipping the first one. The first component is omitted since it primarily separates the two chains by featuring high loadings for dihedral angles that rarely move but are consistently different between chains 1 and 2. The distance metric used for constructing the PI was the Euclidean distance in this 10-dimensional space. Results obtained with other preprocessing pipelines are shown in the Supporting Information.

In Figures 7 and 8, we present the results of our analysis; the salient aspects are summarized as follows. First, the Sapphire analysis indicates that sampling has not reached convergence. As a symmetric homodimer, we expect both chains to access the same conformational states. However, there are phase space regions visited only by one of the two, e.g., the rightmost PI area in Figure 7. The sampling overlap between chains decreases dramatically if the first PC is included (see Figure S8) or if instead we use the 10 tICA modes with largest autocorrelation values (again, skipping the first one, see Figure S9). Second, the relatively coarse view offered by the 10 PCs allows SbC to identify 44 states visited by both chains and 10 states sampled by only one of the two chains (4 and 6 for chains 1 and 2, respectively). The latter are defined as states that have more than 95% of the cluster weight coming from only one chain. The major differences between states at the backbone dihedral angle level are localized in the 43−53 stretch of a loop that is part of the S2 pocket.[49,50] The intrinsic flexibility of this loop might play a role in substrate binding and/or product release. In addition, there is a set of side chains in direct contact with this loop (residues 82, 87, 188, 190, and 192), which contribute strongly to the selected PCs. The remaining residues that do so are more isolated, e.g., Met17 and Asn28, which form a tertiary contact; see Figure 8b. Third, we summarize these 54 states by lumping them via PCCA+ ("Macro" annotation in Figure 7) into eight states. The six states with finite sampling weights in both chains were isolated and used to derive the simplified kinetic network shown in Figure 8. The network, which uses a lag time of 100 ns and is split by a chain, reveals that the chains share several transitions with often similar rates (MFPTs in the low-to-sub microsecond regime). This holds as long as the transition is between well-sampled states in the respective chain. Importantly, setting the caveats about the limited amount of sampling aside, the resultant kinetic network is human-comprehensible and could be integrated into a larger kinetic model of the active cycle of this protease. In turn, kinetic models of this type are needed to make predictions about the efficacy of interfering with its function through pharmaceutical or other strategies.

4. CONCLUSION AND OUTLOOK

We have introduced a method for time-series clustering that derives from an unsupervised data-mining technique developed to efficiently analyze high-dimensional data sets such as those generated by MD simulations. The construction of the Sapphire plot,[20,21,25] which is inherent to SbC, allows for an effective visual inspection of the putative states and pathways of a system at maximal (snapshot) resolution. The clustering is obtained from partitioning along the PI using both kinetic and temporal annotations (Figure 1). It is an advantage that the nature of the identified clusters is directly apparent from the Sapphire plot provided that suitable annotations are used (see, for example, Figure 4, in particular for Beta3S). While the applications here are on conformational equilibria, Sapphire plots and thus SbC can be used equally well to characterize other stochastic systems such as binding equilibria[22,23] or neuronal networks.[24]

Even though the sampling density in a selected feature space, often explored through spanning trees, is the foundation for the identification of metastable and transition states,[6−8,51] a direct inclusion of temporal information has proven beneficial. Temporal information can be incorporated into clustering techniques in different ways: directly[31] or by the inclusion of time-based feature weights either directly[45,47] or through tICA.[33,34] The latter category has been complemented by the introduction of purely kinetic modeling techniques that rely, for example, on basis sets[26] and provide a variational principle for the systematic evaluation of MSMs. The clustering algorithm is one of the hyperparameters of a traditional MSM, and it has been argued that methods optimizing variance- or mean-based criteria, such as Ward and k-means, are best-suited to maintain kinetic fidelity in the resultant MSMs.[38]

In this work, we showed that a Sapphire plot with its standard kinetic and temporal annotations is sufficient to derive a clustering that yields MSMs whose kinetic performance from tICA-transformed data is competitive relative to the application of Ward or k-medoids. We readily acknowledge that this conclusion is based on matching the numbers of clusters; see Figures 5b and 6. It is a caveat that SbC has only indirect control over the number of states it produces, in particular through the numbers of bins on the PI and time axes. We demonstrated that the pooling of spanning tree leaves (LP), which is a property of the Sapphire plot itself, can be beneficial for deriving MSMs because it acts as an effective lumping technique for low-density fringe regions surrounding states (Figures 2 and 3). While an automatic selection of the SbC (hyper)parameters might be of interest, the heuristic is in part specific to the goal of our analysis. The simplest and most general rule is that the bin sizes need to be set according to the minimum sampling weights and residence times of the states to be identified. For LP, our general recommendation is to choose the smallest possible value that avoids finding a diffuse set of mutually unrelated points from different regions of low sampling density on the right of the Sapphire plot. However, this might hinder the detection of actual low-density states, and a smaller or zero value may offer additional insights in this scenario.

There are several benefits of SbC. First, it is a clustering technique that can produce clusters of arbitrary size and shape in the selected feature space. Second, it is a scalable technique; i.e., computational cost increases nearly linearly with data set size. This is because the computational time added by SbC to the Sapphire algorithm is insensitive to data set size (see Figure S10) whereas the computation of the PI based on a sST scales as $O(N \log N)$ with the number of snapshots and linearly with the number of features in the data set.[20] Third, and this is

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

probably the main benefit of SbC, it is immediately possible to both visualize the states and diagnose the result. For example, from the left panel of Figure 2, it is obvious that there is an ill-defined state that serves as a kinetic hub, which here is an artificial shortcut. Similarly, from Figure 7, it is immediately clear that any kinetic predictions will be noisy because there are few transitions between states. For a toy model of sufficiently low dimensionality, SbC is expected to find a number of states that is more or less the same as the true number of metastable states, see Figure 2. For complex systems, the true number of states is generally unknown. Coarse-grained models might aim to achieve human comprehension of every state of the system, which usually limits the analysis to 5−10 macrostates.[12,13,52] SbC produces more clusters, but their structural differences are almost always clear; see Figures 4 and 7. Forcefully reducing the number of inherent states implies either lumping[53−55] or pruning.[56,57] This can happen either in postprocessing or directly at the level of feature selection and transformation. While the use of time-based information changes the results for Beta3S only slightly, those for BPTI are dramatically different,[45] which we find here in similar form for $3CL^{pro}$; compare Figure 7 to Figure S9. The choice and processing of features are the most critical steps in understanding MD data, and the impact of these steps on the inferences drawn remains the biggest caveat in the field.[28,45]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00604.

Illustration of the procedure used for clustering from the temporal annotation (Figure S1); number of clusters obtained by SbC with various settings (Figure S2); additional kinetic measures for Beta3S (Figures S3−S6); cartoon of $3CL^{pro}$ (Figure S7); SAPPHIRE plots of $3CL^{pro}$ with alternative preprocessing (Figures S8−S9); computational cost of SbC (Figure S10) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Andreas Vitalis − *Biochemistry Department, University of Zurich, Zurich CH-8057, Switzerland;* orcid.org/0000-0002-5422-5278; Email: a.vitalis@bioc.uzh.ch

Authors

Francesco Cocina − *Biochemistry Department, University of Zurich, Zurich CH-8057, Switzerland;* orcid.org/0000-0003-0514-7418

Amedeo Caflisch − *Biochemistry Department, University of Zurich, Zurich CH-8057, Switzerland;* orcid.org/0000-0002-2317-6792

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00604

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank David E. Shaw for sharing the trajectory data and the state annotation for BPTI. In particular, we acknowledge Marco Bacci for his contributions and suggestions during earlier stages of the work as well as Davide Garolini for interesting discussions and for the development of the R package "CampaRi". This work was supported financially by an excellence grant of the Swiss National Science Foundation (31003A_169007) to A.C.

■ REFERENCES

(1) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(2) Moon, K. R.; van Dijk, D.; Wang, Z.; Gigante, S.; Burkhardt, D. B.; Chen, W. S.; Yim, K.; van den Elzen, A.; Hirn, M. J.; Coifman, R. R.; Ivanova, N. B.; Wolf, G.; Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **2019**, *37*, 1482−1492.

(3) Fan, J.; Han, F.; Liu, H. Challenges of Big Data Analysis. *Natl. Sci. Rev.* **2014**, *1*, 293−314.

(4) Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality reduction: A comparative. *J. Mach. Learn. Res.* **2009**, *10*, 13.

(5) Jain, A.; Stock, G. Identifying Metastable States of Folding Proteins. *J. Chem. Theory Comput.* **2012**, *8*, 3810−3819.

(6) Vitalis, A.; Caflisch, A. Efficient Construction of Mesostate Networks from Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2012**, *8*, 1108−1120.

(7) Yao, Y.; Sun, J.; Huang, X.; Bowman, G. R.; Singh, G.; Lesnick, M.; Guibas, L. J.; Pande, V. S.; Carlsson, G. Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys.* **2009**, *130*, 144115.

(8) Sittel, F.; Stock, G. Robust Density-Based Clustering To Identify Metastable Conformational States of Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 2426−2435.

(9) Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **2005**, *398*, 161−184.

(10) Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147−179.

(11) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *J. Chem. Phys.* **2013**, *138*, 174106.

(12) Bowman, G. R.; Meng, L.; Huang, X. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.* **2013**, *139*, 121905.

(13) Martini, L.; Kells, A.; Covino, R.; Hummer, G.; Buchete, N.-V.; Rosta, E. Variational Identification of Markovian Transition States. *Phys. Rev. X* **2017**, *7*, 031060.

(14) Husic, B. E.; McKiernan, K. A.; Wayment-Steele, H. K.; Sultan, M. M.; Pande, V. S. A Minimum Variance Clustering Approach Produces Robust and Interpretable Coarse-Grained Models. *J. Chem. Theory Comput.* **2018**, *14*, 1071−1082.

(15) Keller, B.; Hünenberger, P.; van Gunsteren, W. F. An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles. *J. Chem. Theory Comput.* **2011**, *7*, 1032−1044.

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

(17) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386−2396.

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

(19) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **2020**, *30*, 23−66.

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

(21) Blöchliger, N.; Vitalis, A.; Caflisch, A. High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* **2015**, *4*, 6264.
(22) Blöchliger, N.; Xu, M.; Caflisch, A. Peptide Binding to a PDZ Domain by Electrostatic Steering via Nonnative Salt Bridges. *Biophys. J.* **2015**, *108*, 2362−2370.
(23) Langini, C.; Caflisch, A.; Vitalis, A. The ATAD2 bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes. *J. Biol. Chem.* **2017**, *292*, 16734−16745.
(24) Garolini, D.; Vitalis, A.; Caflisch, A. Unsupervised identification of states from voltage recordings of neural networks. *J. Neurosci. Methods* **2019**, *318*, 104−117.
(25) Vitalis, A. *An Improved and Parallel Version of a Scalable Algorithm for Analyzing Time Series Data*. 2020; arXiv:2006.04940. arXiv.org ePrint archive. https://arxiv.org/abs/2006.04940 (accessed June 11, 2020).
(26) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.
(27) Scherer, M. K.; Husic, B. E.; Hoffmann, M.; Paul, F.; Wu, H.; Noé, F. Variational selection of features for molecular kinetics. *J. Chem. Phys.* **2019**, *150*, 194108.
(28) Husic, B. E.; McGibbon, R. T.; Sultan, M. M.; Pande, V. S. Optimized parameter selection reveals trends in Markov state models for protein folding. *J. Chem. Phys.* **2016**, *145*, 194103.
(29) Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proc. Natl. Acad. Sci. U. S. A.* **1931**, *17*, 315−318.
(30) Lasota, A.; Mackey, M. C. *Probabilistic Properties of Deterministic Systems*; Cambridge University Press: New York, 1985.
(31) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.
(32) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: new insights into the folding process. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.
(33) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.
(34) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(35) Noé, F.; Clementi, C. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.
(36) Shaw, D. E. *Research, Molecular Dynamics Simulations Related to SARS-CoV-2*. D. E. Shaw Research Technical Data, 2020; http://www.deshawresearch.com/resources_sarscov2.html.
(37) Owen, C. D.; Lukacik, P.; Strain-Damerell, C. M.; Douangamath, A.; Powell, A. J.; Fearon, D.; Brandao-Neto, J.; Crawshaw, A. D.; Aragao, D.; Williams, M.; Flaig, R.; Hall, D. R.; McAuley, K. E.; Mazzorana, M.; Stuart, D. I.; von Delft, F.; Walsh, M. A. *SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19)*. 2020; DOI: 10.2210/pdb6y84/pdb.
(38) Husic, B. E.; Pande, V. S. Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J. Chem. Theory Comput.* **2017**, *13*, 963−967.
(39) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.
(40) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(41) Rao, F.; Settanni, G.; Guarnera, E.; Caflisch, A. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.* **2005**, *122*, 184901.
(42) Suárez, E.; Adelman, J. L.; Zuckerman, D. M. Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. *J. Chem. Theory Comput.* **2016**, *12*, 3473−3481.
(43) Grinstead, C. M.; Snell, J. L. *Introduction to Probability*, 2nd ed.; American Mathematical Society: Providence, RI, 1997.
(44) Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. Microsecond time-scale conformational exchange in proteins: using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. *J. Am. Chem. Soc.* **2012**, *134*, 2555−2562.
(45) Blöchliger, N.; Caflisch, A.; Vitalis, A. Weighted Distance Functions Improve Analysis of High-Dimensional Data: Application to Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 5481−5492.
(46) Djurdjevac, N.; Sarich, M.; Schütte, C. Estimating the Eigenvalue Error of Markov State Models. *Multiscale Model. Simul.* **2012**, *10*, 61−81.
(47) Krivov, S. V.; Karplus, M. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13841−13846.
(48) Bacci, M.; Caflisch, A.; Vitalis, A. On the removal of initial state bias from simulation data. *J. Chem. Phys.* **2019**, *150*, 104105.
(49) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **2020**, *368*, 409−412.
(50) Zhang, L.; Lin, D.; Kusov, Y.; Nian, Y.; Ma, Q.; Wang, J.; von Brunn, A.; Leyssen, P.; Lanko, K.; Neyts, J.; de Wilde, A.; Snijder, E. J.; Liu, H.; Hilgenfeld, R. α-Ketoamides as Broad-Spectrum Inhibitors of Coronavirus and Enterovirus Replication: Structure-Based Design, Synthesis, and Activity Assessment. *J. Med. Chem.* **2020**, *63*, 4562−4578.
(51) Keller, B.; Daura, X.; van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **2010**, *132*, 074110.
(52) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807−17813.
(53) Evans, D. A.; Wales, D. J. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **2004**, *121*, 1080−1090.
(54) Carr, J. M.; Wales, D. J. Global optimization and folding pathways of selected alpha-helical proteins. *J. Chem. Phys.* **2005**, *123*, 234901.
(55) Carr, J. M.; Wales, D. J. Folding pathways and rates for the three-stranded beta-sheet peptide Beta3s using discrete path sampling. *J. Phys. Chem. B* **2008**, *112*, 8760−8769.
(56) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.
(57) Bello-Rivas, J. M.; Elber, R. Exact milestoning. *J. Chem. Phys.* **2015**, *142*, 094102.

https://dx.doi.org/10.1021/acs.jctc.0c00604
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX