Extracting collective motions underlying nucleosome dynamics via nonlinear manifold learning

Ashley Z. Guo,[1] Joshua Lequieu,[1] and Juan J. de Pablo[1,2]
[1)] *Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA*
[2)] *Materials Science Division, Argonne National Laboratory, Argonne, IL 60439*

(Dated: 12 December 2018)

The identification of effective collective variables remains a challenge in molecular simulations of complex systems. Here, we use a nonlinear manifold learning technique known as the diffusion map to extract key dynamical motions from a complex biomolecular system known as the nucleosome: a DNA-protein complex consisting of a DNA segment wrapped around a disc-shaped group of eight histone proteins. We show that without any *a priori* information, diffusion maps can identify and extract meaningful collective variables that characterize the motion of the nucleosome complex. We find excellent agreement between the collective variables identified by the diffusion map and those obtained manually using a free energy-based analysis. Notably, diffusion maps are shown to also identify subtle features of nucleosome dynamics that did not appear in those manually specified collective variables. For example, diffusion maps identify the importance of looped conformations in which DNA bulges away from the histone complex that are important for the motion of DNA around the nucleosome. This work demonstrates that diffusion maps can be a promising tool for analyzing very large molecular systems and for identifying their characteristic slow modes.

I. INTRODUCTION

The continued development of advanced sampling techniques has extended the reach of molecular simulations considerably, thereby enabling the study of molecular systems of substantial complexity.[1] Higher complexity is accompanied by the challenge of describing key molecular processes. Ideally, we desire for these complex dynamics to be represented by a few low-dimensional descriptors, but automatically identifying such descriptors and quantifying how well they capture the system's dynamics can be challenging.

A range of approaches is available to discover these low-dimensional descriptors from simulated trajectories of a particular system. One attractive option is to use a dimensionality reduction technique to furnish a low-dimensional embedding of data from molecular dynamics trajectories,[2] using algorithms such as principal component analysis (PCA[3]), isometric feature map (Isomap[4]), locally linear embedding (LLE[5]), sketch-maps,[6] and diffusion maps.[7,8] Diffusion maps have been widely applied to a variety of molecular systems, including all-atom miniprotein folding,[9] self-assembly of patchy colloids,[10] and coarse-grained protein models.[11] Furthermore, they have been adopted as part of multiple accelerated sampling algorithms, such as diffusion-map-directed MD (DM-d-MD[12]) and intrinsic map dynamics (iMapD[13]), and variations on the diffusion map itself have also been developed in order to address challenges in working with data with inhomogeneous densities and to reduce computational costs.[11,14]

While diffusion maps have been applied in diverse contexts, there remain interesting challenges in applying diffusion maps to large and complex macromolecular systems, which exhibit inherently rich dynamics. One such system is the nucleosome, a DNA-protein complex consisting of a DNA segment wrapped around a disc-shaped complex of eight histone proteins.[15] The nucleosome is the basic building block of eukaryotic chromatin, which packs into successively higher-order structures in order to form the mitotic chromosome. Nucleosome positions and proper packaging of DNA are important for healthy cellular function.[16,17]

Recent work has shown that DNA sequence is a key factor that governs nucleosome position, with different DNA sequences exhibiting different affinities for the histone octamer. The probability of nucleosome formation changes with this affinity and can span orders of magnitude across different DNA sequences. Several studies on DNA repositioning have been carried out, leading to the identification of two major repositioning mechanisms: (1) the loop propagation model,[18–23] in which a loop of DNA is formed on one side of the nucleosome and moves in an inchworm-like manner along the histone complex; and (2) the twist diffusion model,[24–27] in which a twist defect is introduced into the natural helicity of the DNA and diffuses in a corkscrew-like manner along the histone complex. Recent work by Lequieu et al.[28] investigated the relationship between DNA sequence and repositioning dynamics using a molecular model of the nucleosome; that study showed that different DNA sequences indeed rely on different mechanisms to reposition through pathways reminiscent of the proposed looping and twisting processes.

The simulations performed to reach these conclusions were considerably demanding, and required over 5 microseconds of unbiased simulation data, for 9 different DNA sequences. In the study of Lequieu et al. however, the order parameters used to characterize DNA motion were identified manually, and were necessarily influenced by human biases. As such, it is unclear if they can fully represent the true underlying dynamics of the

nucleosome. The order parameters used in Lequieu et al. were based on the two previously proposed repositioning mechanisms, and thus analysis of the simulations focused specifically on loop propagation and twist diffusion. It is conceivable that other motions within the nucleosome might play important roles in cellular function, and may have been overlooked in this prior study.

In this work, we exploit this wealth of molecular dynamics data to interrogate the dynamics of the nucleosome using the diffusion map. This approach represents a bias-free method for identifying the collective variables that dominate nucleosomal motions. We show that a diffusion map approach is effective for identifying the collective variables previously found by Lequieu et al. through a detailed free energy analysis. Notably, without any *a priori* information, the diffusion map can distinguish DNA sequences that reposition via loop propagation from those that reposition via twist diffusion. Furthermore, the diffusion map approach is able to identify subtle molecular motions involving looping conformations, in which DNA bulges away from the histone octamer, and DNA breathing, in which DNA spontaneously unwraps from the histone complex. By applying the diffusion map to nucleosome dynamics, we show that both dominant and subtle dynamical modes can be automatically extracted from molecular simulation data, thereby reinforcing the diffusion map as a useful tool for unraveling the behavior of complex biomolecular systems.

II. METHODS

A. MD Simulations of the Nucleosome

Molecular dynamics simulations were carried out with in-house codes using a coarse-grained representation of the 223 base pair nucleosome, as described in Lequieu et al.[28] The 3SPN.2C model is used to represent DNA and is the most recent version of the 3SPN model,[29–32] in which DNA is represented by three sites at the centers of mass of phosphate, sugar, and base of each DNA nucleotide. The 3SPN.2C model has been further parameterized to capture the correct melting behavior of double-stranded to single-stranded DNA, sequence effects, and salt effects. The AICG model is used to represent the histone proteins, using a single site per amino acid at the side chain center of mass.[33] Interactions between the DNA and histone proteins consist of excluded volume effects and electrostatic forces, calculated using Debye-Hückel theory. Molecular dynamics simulations were performed in the canonical ensemble using a Langevin thermostat and ionic strength of 150 mM, with frames saved for later analysis every 1 ns. Further details can be found in Lequieu at al.[28]

B. Diffusion maps

Diffusion maps are a type of nonlinear dimensionality reduction technique originally introduced by Coifman and co-workers.[7,8] Here, we briefly step through the algorithm to clarify and facilitate subsequent discussion. Specifically, we use the density-adapted diffusion map introduced by Wang and Ferguson[14] due to the inhomogeneous sampling of configurations in brute-force molecular dynamics simulations of the nucleosome.

First, pairwise distances d_{ij} are calculated between datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$. In this case, we use the root-mean-squared distance between translationally and rotationally aligned atomic coordinates between two molecular configurations. d_{ij} is then passed through a Gaussian kernel to construct matrix $\mathbf{A}$, which contains the now thresholded pairwise distances, with entries

$$A_{ij} = \exp\left(\frac{-d_{ij}^{2\alpha}}{2\epsilon}\right). \tag{1}$$

Here, ϵ is the kernel bandwidth and α rescales pairwise distances globally in order to smooth out density inhomogeneities in sampled configurations. We find that an α value of 0.3 works well for configurations from all three DNA sequences considered here. The kernel bandwidth ϵ defines the extent of the local neighborhood around each datapoint in which to consider pairwise distances to other points, and we use an ϵ of 3.0 for our data across all sequences. $\mathbf{A}$ is then row-normalized to form the Markov matrix

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}, \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix with entries

$$D_{ij} = \sum_j A_{ij}. \tag{3}$$

$\mathbf{M}$ is effectively a transition matrix, with entries M_{ij} corresponding to transition probabilities between configurations $\mathbf{x}_i$ and $\mathbf{x}_j$.

Finally, $\mathbf{M}$ is diagonalized in order to calculate its eigenvectors $\{\psi_i\}$ and associated eigenvalues $\{\lambda_i\}$. Due to the Markovian nature of $\mathbf{M}$, the top eigenvalue-eigenvector pair (ψ_0, λ_0) is trivial; this pair corresponds to the steady-state distribution of a random walk with $\lambda_0 = 1$.

By locating a gap in the eigenvalue spectrum between λ_k and λ_{k+1}, one can identify the top k non-trivial eigenvectors $\{\psi_i\}_{i=1}^{k}$ corresponding to slow diffusion modes of the system, which dominate over the fast modes corresponding to the remaining lower eigenvectors $\{\psi_i\}_{i>k}$. The original high-dimensional data can then be embedded in k dimensions by projecting the data onto the top k non-trivial eigenvectors,

$$\mathbf{x}_i \mapsto [\psi_1(i), \psi_2(i), \ldots, \psi_k(i)]\,. \tag{4}$$

In some cases, multiple gaps may emerge in the eigenvalue spectrum, in which case one must avoid only using eigenvectors up to the first gap, which may produce misleading results. The final low-dimensional embedding reflects the intrinsic manifold underlying the molecular system as extracted from the sampled molecular dynamics data.

Analysis of nucleosome simulations using the density-adapted diffusion map began with calculation of $\mathbf{A}$ for each DNA sequence studied, using Equation 1 and snapshots extracted from MD simulation trajectories. $\mathbf{M}$ was then calculated for each sequence as described above in Equations 2 and 3, followed by calculation of eigenvectors $\{\psi_i\}$ and eigenvalues $\{\lambda_i\}$ for each sequence's $\mathbf{M}$. The spectra of $\{\lambda_i\}$ were examined visually in order to identify gaps and determine non-trivial eigenvectors for each sequence-specific diffusion map embedding. Multiple collective variables (described in the following three subsections) were calculated for each simulation snapshot used to create the embeddings and then projected onto the non-trivial eigenvectors to create diffusion map embeddings of collective variables for each sequence. These diffusion map embeddings of collective variables were then used to identify sequence-specific correlations of collective variables with dominant dynamical modes of the nucleosome system.

C. Collective Variables Describing DNA Translocation and Rotation

DNA translocation relative to the histone dyad is characterized by S_T, defined as:

$$S_T = \left\langle \pm \arccos\left(\frac{\mathbf{P}\cdot\mathbf{P_0}}{\|\mathbf{P}\|\|\mathbf{P_0}\|}\right)\right\rangle. \tag{5}$$

Here, vector $\mathbf{P}$ points from the center of a base step to the center of the protein complex, and $\mathbf{P_0}$ is the corresponding value of $\mathbf{P}$ taken from a reference nucleosome crystal structure (PDB ID: 1KX5),[34] which was used to create initial structures for the nucleosome simulations. The average in Equation 5 is taken over -15, -5, +5, and +15 base steps relative to the histone dyad, located at the central position on the nucleosome (indicated by the triangle in Figure 1). If $(\mathbf{P}\times\mathbf{P_0})\cdot\hat{\mathbf{f}} \le 0$ then the positive sign is used (otherwise, negative), where vector $\hat{\mathbf{f}}$ points along the center of the nucleosomal DNA superhelix. Using this sign convention, positive S_T corresponds to forward translocation of DNA toward the 5' end, while negative S_T corresponds to reverse translocation toward the 3' end.

A second nucleosome repositioning order parameter is S_R, which characterizes DNA rotation:

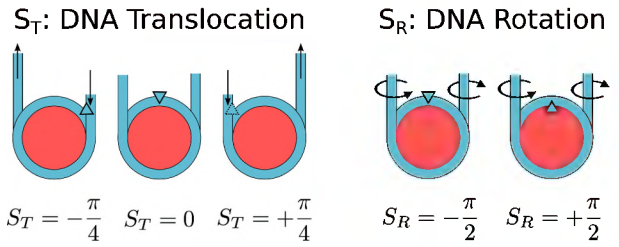


FIG. 1. Schematic of order parameters S_T and S_R, which characterize DNA translocation and DNA rotation, respectively. These order parameters are defined in Section II C.

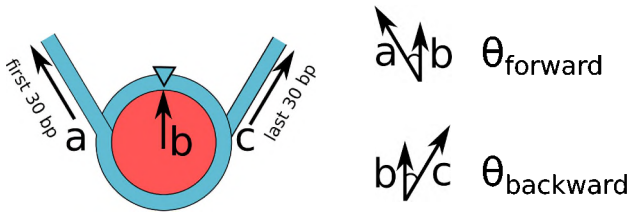


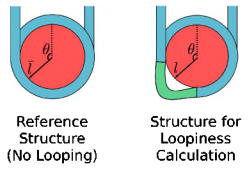
FIG. 2. Schematic of order parameters θ_{forward} and θ_{backward} characterizing DNA breathing, in which strands of DNA spontaneously unwrap and rewrap from the histone complex. Each angle is calculated from two vectors: vector $\mathbf{b}$ points from the histone complex center of mass to the dyad, and vectors $\mathbf{a}$ and $\mathbf{c}$ point along either end of the DNA strand, from the 30th base pair to the first and endmost base pair.

$$S_R = \left\langle \pm \arccos\left(\frac{\mathbf{P}\cdot\mathbf{B}}{\|\mathbf{P}\|\|\mathbf{B}\|}\right)\right\rangle, \tag{6}$$

where vector $\mathbf{B}$ points from the center of a given base step on the sense strand to its complementary base step on the anti-sense strand. $\mathbf{P}$ and the average denoted by the angle brackets are as defined for S_T. If $(\mathbf{P}\times\mathbf{B})\cdot\mathbf{D} \le 0$, then the positive sign is used (otherwise, negative). $\mathbf{D}$ is a vector in the 5' to 3' direction along the sense strand of the DNA. If $S_R = -\frac{\pi}{2}$, the minor groove of the DNA double helix is oriented toward the histone core, whereas when $S_R = \frac{\pi}{2}$, the minor groove is oriented away from the histone complex.

D. Collective Variable Describing DNA Breathing

DNA breathing, which involves spontaneous unwrapping and rewrapping of DNA from the nucleosome, was characterized by two angle parameters, θ_{forward} and θ_{backward}, shown in Figure 2. Each angle is calculated between a vector from the center of mass of the histone to the dyad, which is relatively immobile, and a vector from the 30th DNA base pair to the first and endmost DNA base pair, which moves significantly as DNA unwraps.



$$\Delta l(\theta) = \langle l(\theta)\rangle - \langle \bar{l}(\theta)\rangle$$
$$\Delta l^*(\theta) = \max(\Delta l(\theta) - 8\text{Å}, 0)$$
$$\text{loopiness} = \int_\theta \Delta l^*(\theta) d\theta$$

FIG. 3. Schematic of the loopiness order parameter, which characterizes the extent to which DNA bulges away from the histone octamer. Calculation of this order parameter is described in Section II E.

E. Collective Variable Describing DNA Looping

DNA looping, which involves DNA bulging away from the histone octamer, was characterized using a *loopiness* order parameter (Figure 3). To calculate loopiness, we first calculate two values for each ith DNA base pair: the distance of the base pair to the histone center of mass, l_i, and the location of the base pair relative to the dyad, θ_i. For ease of calculation, we compute the average distance from a base pair to the histone center of mass as a function of location θ, denoted as $\langle l(\theta)\rangle$. In order to normalize $\langle l(\theta)\rangle$, we then calculate the corresponding value of this average distance in the complete absence of looping, $\langle \bar{l}(\theta)\rangle$, which is calculated from a nucleosome simulation performed in very low salt concentration for a strongly binding DNA sequence. We then normalize $\langle l(\theta)\rangle$ using $\langle \bar{l}(\theta)\rangle$ by calculating deviation from the loop-free case $\Delta l(\theta) = \langle l(\theta)\rangle - \langle \bar{l}(\theta)\rangle$; in cases where there is no DNA looping, Δl is approximately 0 across all locations θ, and in cases where DNA loops form, $\Delta l > 0$. To eliminate baseline noise, we threshold Δl by subtracting a threshold value of 8Å, which corresponds to the Debye length at 150 mM at which DNA-histone attraction has largely decayed. The post-threshold looping parameter Δl^* is then integrated along the entire circumference around the histone octamer (over all θ) in order to obtain our final *loopiness* order parameter.

III. RESULTS AND DISCUSSION

We apply the diffusion map to a subset of these trajectories from three representative DNA sequences: sequence A, a strongly binding sequence that primarily repositions by loop propagation; sequence B, a moderately binding sequence that exhibits a combination of loop propagation and twisting; and sequence C, a weakly binding sequence that primarily repositions by twisting. These sequences are tabulated in Table I with their respective binding strengths and sequence identities. Figure 4 summarizes the loop propagation and twisting models of nucleosome repositioning, along with the respective repositioning behaviors for all three sequences studied and the collective variables used to describe repositioning, which will be introduced later in the text.

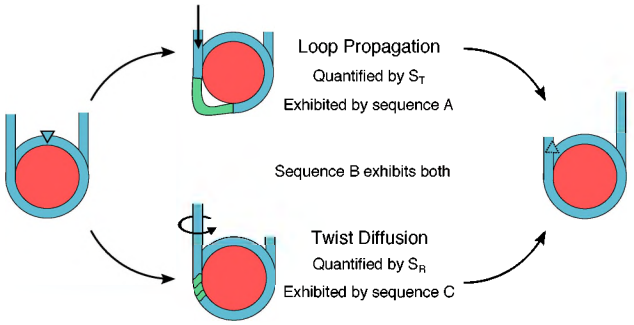


FIG. 4. Schematic showing two proposed nucleosome repositioning mechanisms: loop propagation and twist diffusion. The histone complex is represented in red, and DNA in blue. S_T and S_R quantify loop propagation and twist diffusion, respectively; definitions for these order parameters are introduced in the Methods section. Individual repositioning propensities for sequences A, B, and C are also shown.

TABLE I. DNA sequences used in this work, along with their binding strengths and sequence names used in the literature.

Sequence Name	Binding Strength	Name in Literature
A	Strong	c3[a]
B	Moderate	TRGC[b]
C	Weak	TTAGGG[c]

[a] See Segal et al. 2006.[35]
[b] See Moyzis et al. 1988[36] and Morin 1989.[37]
[c] See Shrader and Crothers 1989.[38]

By applying the density-adapted diffusion map on configurations for sequences A, B, and C as described in the previous section, we obtain the eigenvalue spectra shown in Figure 5. Snapshots for the diffusion map analysis were extracted from the molecular dynamics trajectories at evenly spaced intervals (every 40 ns for sequences A and B, and every 25 ns for sequence C), for a total of 16,207 snapshots from sequence A, 14,917 from sequence B, and 10,713 from sequence C. Sequences A and B exhibit similar hierarchical eigenvalue spectra, indicated by multiple spectral gaps. Both sequences exhibit gaps between ψ_3 and ψ_4, and between ψ_6 and ψ_7, suggesting that dynamics are dominated by a combination of three major slow modes (ψ_1 to ψ_3) and three moderate modes (ψ_4 to ψ_6). The eigenvalue spectrum for sequence C exhibits a large, distinct gap after ψ_1 and a smaller gap after ψ_3, which indicates that one major slow mode dominates the system, followed by two moderate modes.

A. DNA Translocation

First, we check if the diffusion map is able to recover the two nucleosome repositioning order parameters studied in Lequieu et al.[28] We begin with S_T, the order parameter characterizing DNA translocation relative to the histone dyad. Figure 6 shows two- and three-dimensional

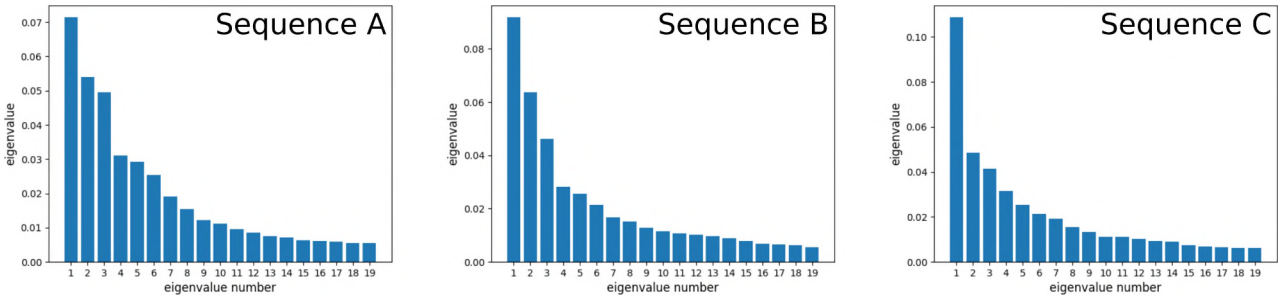


FIG. 5. Eigenvalue spectra for sequences A, B, and C. Note that sequences A and B exhibit hierarchical character, indicated by multiple gaps in the eigenvalue spectra. Both spectra show three dominant eigenvalues, followed by three moderate eigenvalues, suggesting that three major slow dynamical modes dominate the system, while three less significant modes still contribute to the system dynamics.

diffusion map embeddings for all three sequences studied, using the first three non-trivial eigenvectors and colored by S_T. In all three sequences, DNA translocation is found to be well parameterized by either the slowest (ψ_1) or second slowest (ψ_2) dynamical mode identified by the diffusion map, indicating that S_T correlates with slow modes across binding affinities. The correlation of S_T with either ψ_1 or ψ_2 in all three sequences supports the idea that there will always be some degree of translocational motion in the nucleosome repositioning process, regardless of the preference for a particular DNA sequence to reposition by either looping or twisting.

B. DNA Rotation

Figure 7 shows diffusion map embeddings of S_R, which quantifies DNA rotation, for all three sequences studied, using the top three non-trivial eigenvectors and colored by S_R. There is no correlation of S_R with these top three eigenvectors for sequences A and B; further analysis confirms that S_R is not well parameterized by any of the top six eigenvectors for these sequences. This is expected, since A and B exhibit relatively strong binding affinities and are more likely to reposition by a looping mechanism as opposed to a twisting mechanism.

In contrast, sequence C, a weakly binding sequence that primarily repositions by rotation, exhibits a periodic banded structure, which appears more clearly in the two-dimensional embedding of sequence C in ψ_1 and ψ_3 (Figure 7d). Furthermore, we can construct an effective free energy landscape from the diffusion map embedding of sequence C by collecting a histogram of sequence C datapoints in ψ_1 and S_R, normalizing by the total number of datapoints so that the resulting probability in the bins sum to 1, and taking the negative logarithm of these probabilities. This effective free energy landscape is plotted in Figure 7e; this is reminiscent of the free energy landscape calculated for sequence C using conventional methods (umbrella sampling and WHAM) found by Lequieu et al.,[28] plotted in S_T vs S_R and reproduced in Figure 7f; this is consistent with our earlier finding that S_T correlates with ψ_1 for this sequence.

The order parameters characterizing DNA translocation and rotation emerge in the same non-trivial eigenvector for sequence C, consistent with prior observations that sequence C repositions via twisting. In contrast, only translocation is extracted from the underlying MD data for sequences A and B, consistent with prior observations that sequences A and B do not reposition through DNA twisting. Through analysis of all three sequences, we observe that the diffusion map approach identifies a slow mode that correlates with DNA translocation across all binding strengths. DNA rotation emerges in the same slow mode if the sequence exhibits repositioning by rotation as well, suggesting that this particular non-trivial eigenvector corresponds to a more general repositioning motion consisting of a combination of translocation and, if the sequence exhibits it, rotation.

C. DNA Breathing

Next, we examine whether the diffusion map approach can be used to identify key nucleosome dynamics beyond the translocational and rotational repositioning mechanisms studied in Lequieu et al.[28] One particularly interesting aspect of nucleosome dynamics is DNA breathing, which involves unwrapping of nucleosomal DNA from the histone complex. Single-molecule FRET experiments have shown that nucleosomal DNA can spontaneously unwrap and rewrap from the histone octamer, allowing transcription factors, enzymes, and other proteins to interact with previously unaccessible portions of DNA that were buried by the histone complex.[39,40]

Figure 8 shows two-dimensional diffusion map embeddings for all three sequences, colored by the average of θ_{forward} and θ_{backward}, which captures breathing on both sides of the nucleosome. The average breathing order parameter correlates with ψ_2 for sequence A, and with ψ_1 for sequences B and C. Interestingly, in each sequence, the average breathing order parameter correlates with

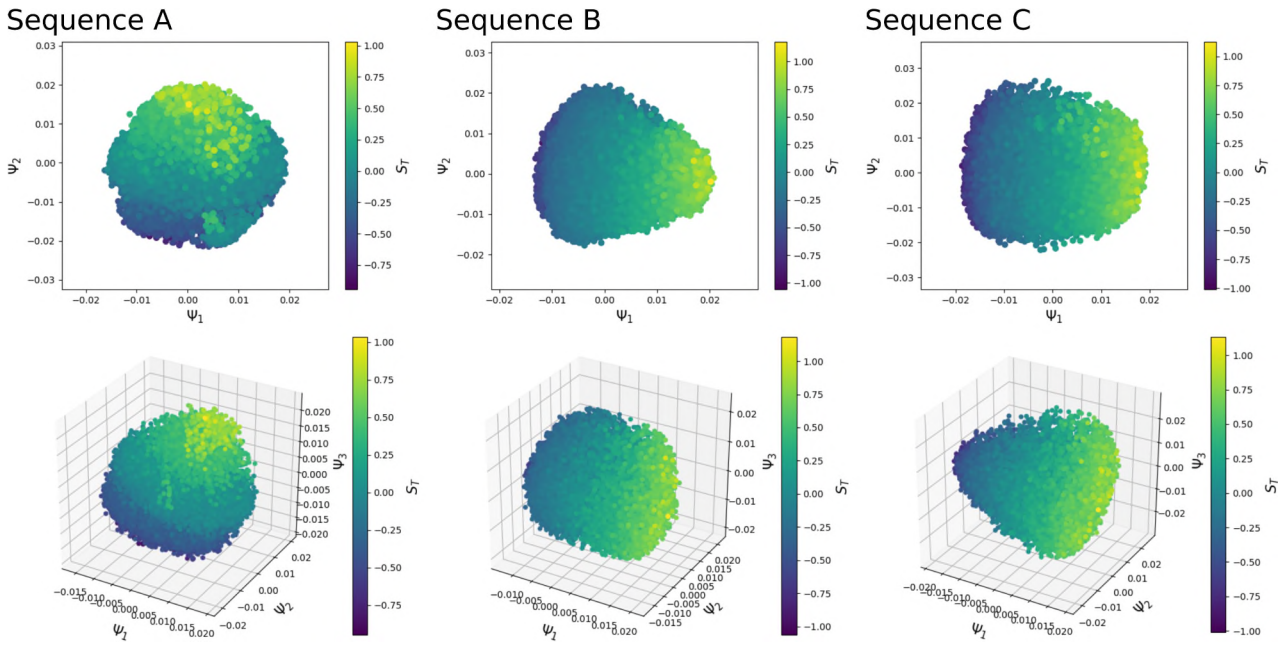


FIG. 6. 2- and 3-dimensional diffusion map embeddings of S_T for all sequences. DNA translocation correlates with ψ_2 for sequence A, indicated by the gradient in S_T along the vertical ψ_2 axis. DNA translocation correlates with ψ_1 for sequences B and C, indicated by the gradient in S_T along the horizontal ψ_1 axis.

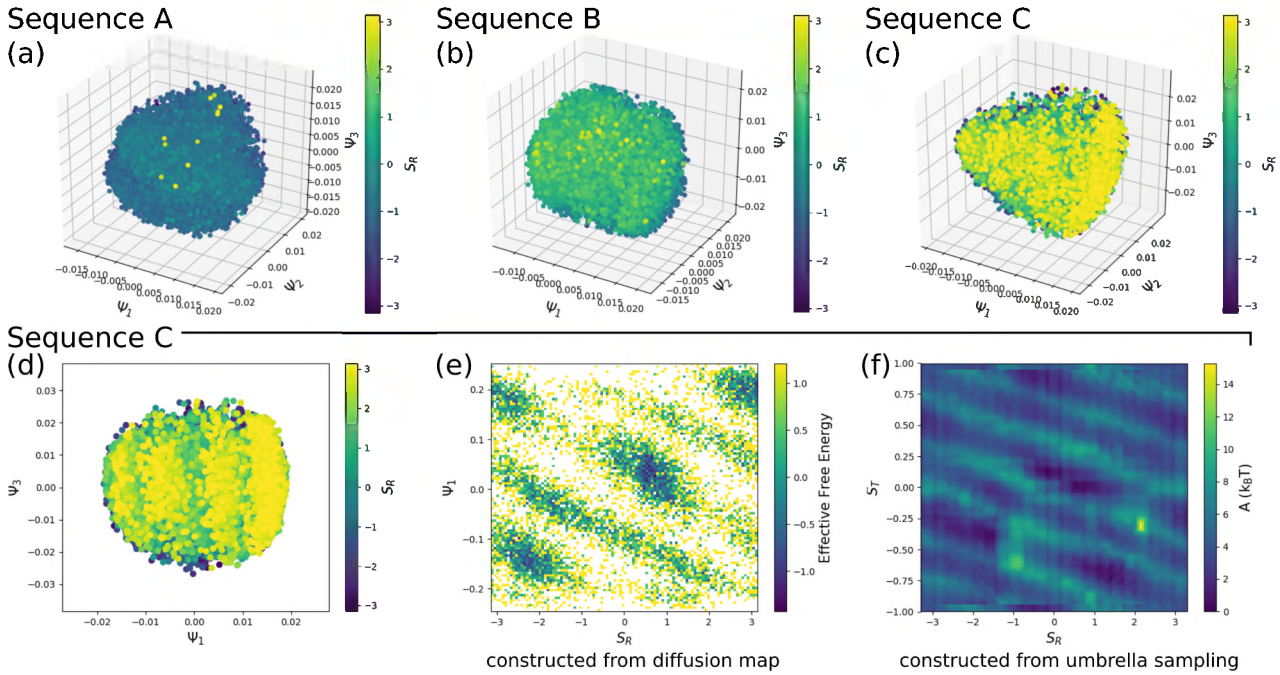


FIG. 7. (a-c): 3-dimensional diffusion map embeddings of S_R, the order parameter that characterizes DNA rotation, for all sequences. For clarity, datapoints with greater values of S_R are shown at higher layers of the plot. There is no correlation of S_R with top non-trivial eigenvectors for sequences A and B; (d) 2-dimensional diffusion map embedding of S_R for sequence C. ψ_1 correlates with cycles of DNA rotation, as indicated by the periodic bands of S_R along ψ_1; (e) effective free energy constructed from the diffusion map embedding for sequence C. Effective free energy is calculated by histogramming datapoints for sequence C in ψ_1 and S_R, normalizing each histogram bin by the total number of datapoints to calculate probabilities, and then taking the negative log of each bin. The resulting density plot exhibits a periodic banded structure reminiscent of the free energy landscape for sequence C constructed by conventional methods by Lequieu et al.,[28] which is plotted in (f) in S_T vs S_R. Note that S_T was previously found to correlate with ψ_1 for sequence C; the diffusion map has effectively unfurled the same previously calculated free energy landscape.

the same eigenvector as S_T (and S_R, in the case of sequence C); this is evident in the visual similarities between Figures 6 and 8. The shared correlations of the average breathing order parameter with S_T and S_R suggest that repositioning dynamics and breathing dynamics are closely tied. The embeddings generated by the diffusion map approach capture both of these motions within the same non-trivial eigenvector, implying that these two types of dynamics are innately part of the same characteristic dynamic mode exhibited by the nucleosome. Although the diffusion map is unable to provide an explicit nonlinear mapping from the high-dimensional input to low-dimensional coordinates, and interpretation of the low-dimensional coordinates is limited to correlating the top eigenvectors of $\mathbf{M}$ with various descriptors of the system, this perceived deficiency may also be interpreted as an advantage, since it provides a tool for identifying multiple CVs that may be coupled together in the same slow dynamical mode, as we have just observed with S_T and S_R for sequence C.

D. DNA Looping

In Figure 5, sequences A and B were found to exhibit hierarchical eigenvalue spectra, with three dominant non-trivial eigenvectors (ψ_1, ψ_2, ψ_3) and three moderate non-trivial eigenvectors (ψ_4, ψ_5, ψ_6). Our analysis thus far, using the diffusion map approach, has focused on motions correlating with the top group of non-trivial eigenvectors. We now examine the significance of the moderate non-trivial eigenvectors in sequences A and B (and why this feature is absent from the eigenvalue spectrum for sequence C).

Figure 9 shows two-dimensional diffusion map embeddings of the loopiness order parameter, described in the Methods section, for sequences A, B, and C using the moderate eigenvectors ψ_4, ψ_5, and ψ_6. Protrusions are observed in all three embeddings for sequence A and the embedding of sequence B in ψ_5 and ψ_6. Through visual inspection of configurations corresponding to points within and outside of the protruding lobe, we find that the protrusion corresponds to configurations exhibiting DNA loops. In more "loopy" configurations, DNA bulges away from the histone complex, and gaps are formed between the DNA and histone octamer. Loopy configurations are necessary for the loop propagation involved in DNA translocation characterized by the order parameter S_T, as described earlier, with translocation dominating in more strongly binding sequences.

The emergence of loopiness in ψ_4, ψ_5, and ψ_6 in sequences A and B is consistent with their relative propensities for translocation. For strongly binding sequence A, loopiness emerges in multiple higher eigenvectors compared with moderately binding sequence B; loopy configurations for sequence A are clearly isolated in ψ_4 through ψ_6. In contrast, loopiness only emerges in ψ_6 for sequence B, which exhibits a lower propensity for translocation compared to sequence A. Furthermore, weakly binding sequence C repositions entirely by rotation and does not exhibit any moderate eigenvectors. In fact, loopiness does not emerge in any of the top 12 eigenvectors for sequence C.

DNA loop formation is important well beyond the context of the mechanics of loop propagation, with implications in chromatin remodeling and spontaneous nucleosome migration,[23] and we show that the diffusion map can automatically identify this subtle mode. Furthermore, we find that looping is embedded in higher-order eigenvectors, which diffusion map studies often bypass while focusing on the first several dominant eigenvectors. These top eigenvectors often extract the dynamic modes corresponding to collective variables more easily identified by hand, as the present study shows with S_T and S_R. We show that thorough examination of higher-order modes can provide valuable insight into more subtle dynamics of complex systems that may be easier for humans to miss.

IV. CONCLUSIONS

Diffusion maps were used to extract key motions underlying nucleosome dynamics from MD trajectories of nucleosome repositioning for three representative DNA sequences, spanning different binding strengths (and consequently, different repositioning dynamics). Translocational and rotational motions, which had been previously identified through a detailed free energy analysis by Lequieu et al.,[28] were confirmed by the diffusion map approach. Translocational motions were found to correlate with dominant slow modes across the three DNA sequences examined here. Rotational motions were only found to emerge in the weakest binding sequence studied, emerging in the same slow mode that correlates with translocation.

In addition to finding the previously reported translocational and rotational order parameters, the diffusion map analysis was also used to extract DNA breathing and looping motions. Measures of DNA breathing, in which DNA spontaneously unwraps from and rewraps around the histone complex, were found to correlate with the same eigenvectors that correlate with DNA translocation and rotation, suggesting that DNA repositioning and DNA breathing are inherently part of the same dynamical mode. Sequences that exhibit DNA sliding were found to exhibit hierarchical eigenvalue spectra, with looping configurations isolated in the moderate eigenvectors corresponding to eigenvalues between the first and second spectral gap. The dominance of DNA sliding over twisting is further reflected in the order in which loopiness appears in these moderate eigenvectors. Weakly binding sequence C, which primarily repositions by twisting, neither exhibited a hierarchical eigenvalue spectra nor any eigenvectors that correlated with loopiness.

The diffusion map approach is particularly useful in

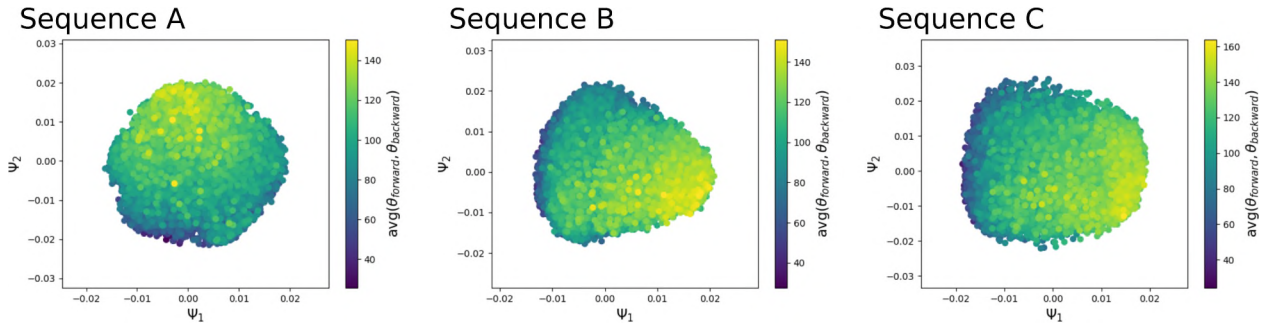


FIG. 8. Two-dimensional embeddings of the average of θ_{forward} and θ_{backward}, which characterizes DNA breathing, for all sequences. The average breathing order parameter for sequence A correlates with ψ_2, as indicated by the gradient in the breathing order parameter along the vertical ψ_2 axis; ψ_2 also correlates with the order parameter characterizing DNA translocation, S_T, for sequence A, as seen in Figure 6. The average breathing order parameter for sequences B and C correlate with ψ_1, as indicated by the gradient in the breathing order parameter along the horizontal ψ_1 axis; this eigenvector also correlates with S_T for these two sequences, again as seen in Figure 6. For sequence C, this eigenvector also correlates with S_R, as seen in Figure 7.

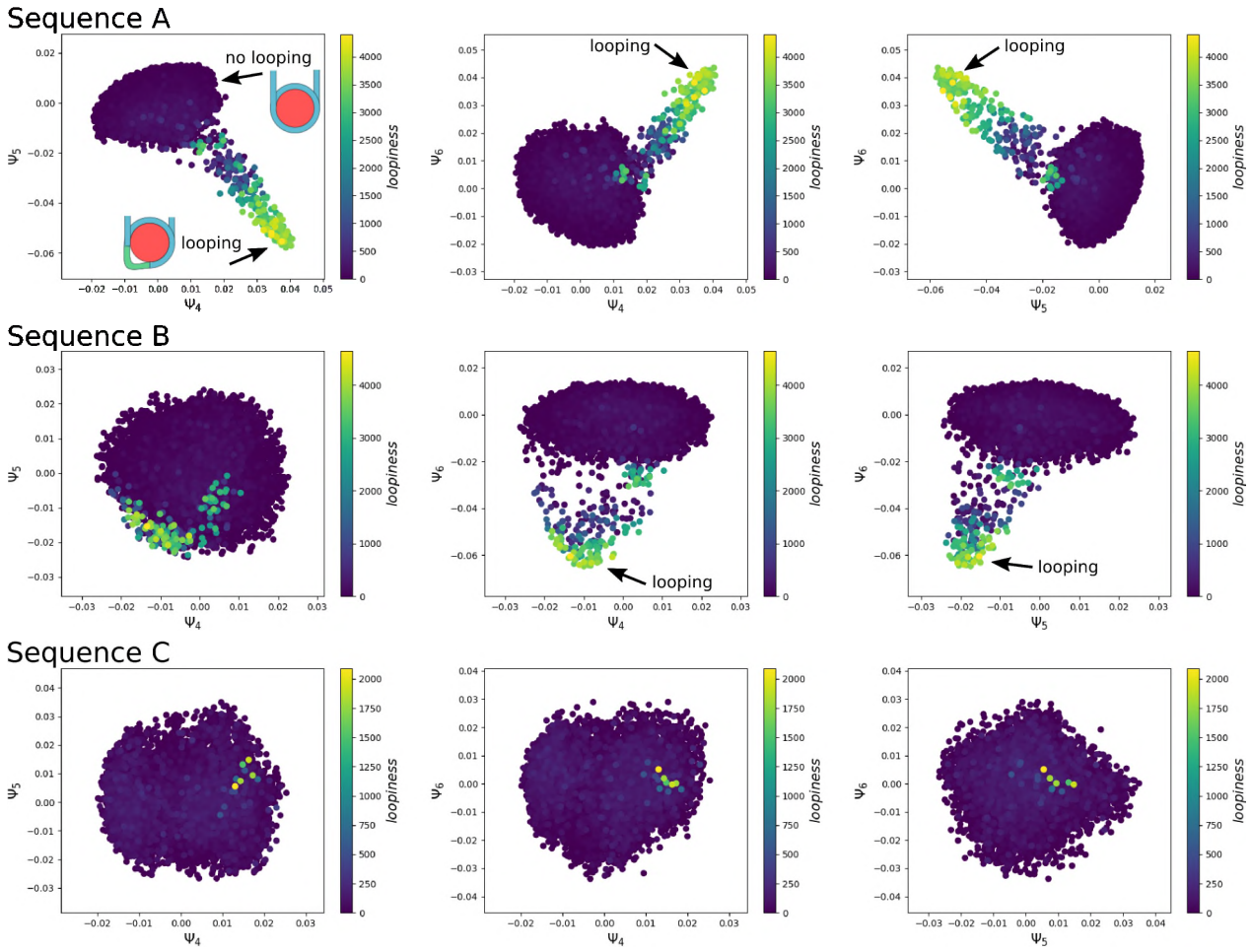


FIG. 9. Two-dimensional diffusion map embeddings of loopiness for all sequences, plotted by moderate eigenvectors ψ_4, ψ_5 and ψ_6. For sequence A, more loopy configurations are isolated by all three moderate eigenvectors. For sequence B, loopy conformations are only isolated by ψ_6. Loopy configurations are not extracted for sequence C.

enabling the discovery of key dynamical motions directly from MD data without defining *a priori* what exactly these motions might be. Although interpretation of dominant dynamical modes is aided by embedding user-specified order parameters in the diffusion map, as done in this work, these order parameters need not be supplied in order to calculate the non-trivial eigenvectors corresponding to these dominant modes, nor specially created in order to interpret a specific non-trivial eigenvector. For example, one might interpret a particular eigenvector by visually examining snapshots of the simulation drawn from different areas of the diffusion map, or use a generalized collective variable instead (ex. fit an eigenvector as a function of atomic coordinates from each simulation snapshot in the diffusion map). Considering the importance of sequence dependence in nucleosome dynamics, diffusion maps provide an attractive solution for rapid screening and identification of key dynamics across sequences in more complex scenarios, for example in higher order chromatin structures or comparing across mutated sequences. Even in the single nucleosome case studied in this work, there remain several significant eigenvectors for which the corresponding dynamics are unknown; we are actively working on elucidating these dynamics. More generally, this work emphasizes the possibilities of uncovering unintuitive properties in MD data that may be missed by more traditional approaches. Here, we are able to confirm both previously known and new order parameters using a small subset (and only 3 out of 9 total sequences) of the MD trajectories previously used in a detailed free energy analysis, attesting to the usefulness and efficiency of applying diffusion maps to previously simulated complex systems.

V. ACKNOWLEDGEMENTS

We thank Andrew Ferguson, Joshua Moller, Aron Coraor, and Hythem Sidky for helpful discussions. Work was completed with the support of the University of Chicago Research Computing Center.

[1] H. Sidky, Y. Colón, J. Helfferich, B. Sikora, C. Bezik, W. Chu, F. Giberti, A. Guo, X. Jiang, J. Lequieu, J. Li, J. Moller, M. Quevillon, M. Rahimi, H. Ramezani-Dakhel, V. Rathee, D. Reid, E. Sevgen, V. Thapar, M. Webb, J. Whitmer, and J. De Pablo, J. Chem. Phys. **148** (2018).
[2] M. Duan, J. Fan, M. Li, L. Han, and S. Huo, J. Chem. Theory Comput. **9**, 2490 (2013).
[3] I. Jolliffe, *Principal Component Analysis* (Springer Verlag, 1986).
[4] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**, 2319 (2000).
[5] S. Roweis and L. Saul, Science **290**, 2323 (2000).
[6] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. **108**, 13023 (2011).
[7] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. **102**, 7432 (2005).
[8] R. R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. **21**, 5 (2006).
[9] S. B. Kim, C. J. Dsilva, I. G. Kevrekidis, and P. G. Debenedetti, J. Chem. Phys. **142**, 085101 (2015).
[10] A. W. Long and A. L. Ferguson, J. Phys. Chem. B **118**, 4228 (2014).
[11] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134** (2011).
[12] W. Zheng, M. A. Rohrdanz, and C. Clementi, J. Phys. Chem. B **117**, 12769 (2013).
[13] E. Chiavazzo, R. Covino, R. R. Coifman, C. W. Gear, A. S. Georgiou, G. Hummer, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. , 201621481 (2017).
[14] J. Wang, M. A. Gayatri, and A. L. Ferguson, J. Phys. Chem. B **121**, 4923 (2017).
[15] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).
[16] B. Hendrich and W. Bickmore, Hum. Mol. Genet. **10**, 2233 (2001).
[17] S. R. Bhaumik, E. Smith, and A. Shilatifard, Nat. Struct. Mol. Biol. **14**, 1008 (2007).
[18] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart, Phys. Rev. Lett. **86**, 4414 (2001).
[19] I. M. Kulić and H. Schiessel, Phys. Rev. Lett. **91**, 3 (2003).
[20] Y. Lorch, B. Davis, and R. D. Kornberg, Proc. Natl. Acad. Sci. **102**, 1329 (2005).
[21] R. Strohner, M. Wachsmuth, K. Dachauer, J. Mazurkiewicz, J. Hochstatter, K. Rippe, and G. Längst, Nat. Struct. Mol. Biol. **12**, 683 (2005).
[22] P. Ranjith, J. Yan, and J. F. Marko, Proc. Natl. Acad. Sci. **104**, 13649 (2007).
[23] M. Pasi and R. Lavery, Nucleic Acids Res. **44**, 5450 (2016).
[24] J. M. Gottesfeld, J. M. Belitsky, C. Melander, P. B. Dervan, and K. Luger, J. Mol. Biol. **321**, 249 (2002).
[25] R. K. Suto, R. S. Edayathumangalam, C. L. White, C. Melander, J. M. Gottesfeld, P. B. Dervan, and K. Luger, J. Mol. Biol. **326**, 371 (2003).
[26] T. J. Richmond and C. A. Davey, Nature **423**, 145 (2003).
[27] I. M. Kulić and H. Schiessel, Biophys. J. **84**, 3197 (2003).
[28] J. Lequieu, D. C. Schwartz, and J. J. D. Pablo, (2017).
[29] T. A. Knotts, N. Rathore, D. C. Schwartz, and J. J. De Pablo, J. Chem. Phys. **126** (2007).
[30] E. J. Sambriski, D. C. Schwartz, and J. J. De Pablo, Biophys. J. **96**, 1675 (2009).
[31] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. De Pablo, J. Chem. Phys. **139** (2013).
[32] G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. De Pablo, J. Chem. Phys. **141** (2014).
[33] W. Li, P. G. Wolynes, and S. Takada, Proc. Natl. Acad. Sci. **108**, 3504 (2011).
[34] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, J. Mol. Biol. **319**, 1097 (2002).
[35] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J. P. Z. Wang, and J. Widom, Nature **442**, 772 (2006).
[36] R. K. Moyzis, J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. Meyne, R. L. Ratliff, and J. R. Wu, Proc. Natl. Acad. Sci. **85**, 6622 (1988).
[37] G. B. Morin, Cell **59**, 521 (1989).
[38] T. E. Shrader and D. M. Crothers, Biophysics (Oxf). **86**, 7418 (1989).
[39] G. Li, M. Levitus, C. Bustamante, and J. Widom, Nat. Struct. Mol. Biol. **12**, 46 (2005).
[40] H. S. Tims, K. Gurunathan, M. Levitus, and J. Widom, J. Mol. Biol. **411**, 430 (2011).