Exploring protein-mediated compaction of DNA by coarse-grained simulations and unsupervised learning

Marjolein de Jager,[1,*] Pauline J. Kolbeck,[1,2] Willem Vanderlinden,[1,2,3] Jan Lipfert,[1,2] and Laura Filion[1]
[1]Soft Condensed Matter and Biophysics, Debye Institute for Nanomaterials Science, Utrecht University, Utrecht, the Netherlands; [2]Department of Physics and Center for NanoScience, LMU, Munich, Germany; and [3]School of Physics and Astronomy, University of Edinburgh, Scotland, United Kingdom

ABSTRACT Protein-DNA interactions and protein-mediated DNA compaction play key roles in a range of biological processes. The length scales typically involved in DNA bending, bridging, looping, and compaction ($\geq$1 kbp) are challenging to address experimentally or by all-atom molecular dynamics simulations, making coarse-grained simulations a natural approach. Here, we present a simple and generic coarse-grained model for DNA-protein and protein-protein interactions and investigate the role of the latter in the protein-induced compaction of DNA. Our approach models the DNA as a discrete worm-like chain. The proteins are treated in the grand canonical ensemble, and the protein-DNA binding strength is taken from experimental measurements. Protein-DNA interactions are modeled as an isotropic binding potential with an imposed binding valency without specific assumptions about the binding geometry. To systematically and quantitatively classify DNA-protein complexes, we present an unsupervised machine learning pipeline that receives a large set of structural order parameters as input, reduces the dimensionality via principal-component analysis, and groups the results using a Gaussian mixture model. We apply our method to recent data on the compaction of viral genome-length DNA by HIV integrase and find that protein-protein interactions are critical to the formation of looped intermediate structures seen experimentally. Our methodology is broadly applicable to DNA-binding proteins and protein-induced DNA compaction and provides a systematic and semi-quantitative approach for analyzing their mesoscale complexes.

SIGNIFICANCE DNA is central to the storage and transmission of genetic information and is frequently compacted and condensed by interactions with proteins. Their size and dynamic nature make the resulting complexes difficult to probe experimentally or by all-atom simulations. We present a simple coarse-grained model to explore $\sim$kbp DNA interacting with proteins of defined valency and concentration. Our analysis uses unsupervised learning to define the conformational states of the DNA-protein complexes and the pathways between them. We apply our simulations and analysis to the compaction of viral genome-length DNA by HIV integrase. We find that protein-protein interactions are critical to account for the experimentally observed intermediates and that our simulated complexes are in good agreement with experimental observations.

INTRODUCTION

DNA is central to the storage and transmission of genetic information, which critically involves a broad range of DNA-protein interactions. Both cellular and viral DNA are compacted by interactions with proteins, and recent evidence suggests that DNA often occupies cellular microenvironments or subcompartments, i.e., where DNA-protein interactions create condensates and membrane-less organelles (1–10). It has been shown that DNA-protein interactions are sufficient to compact DNA and create defined clusters (11–13). For example, vaccinia topoisomerase IB was found to induce the formation of DNA-protein filaments at low protein/DNA ratios by creating bridges between two segments of a single DNA molecule and the formation of DNA-protein clusters of multiple DNA molecules at high protein/DNA ratios (14). Recent work has highlighted that DNA bridging can explain the compaction of DNA by structural maintenance of chromosome (SMC) cohesin complexes in a phase diagram with an extended and a

*Correspondence: m.e.dejager@uu.nl

compacted phase (15). Similarly, DNA-protein interactions drive condensation involved in DNA repair (16) and the compaction of mitochondrial DNA in nucleoids (17,18).

As DNA looping and bridging—and ultimately compaction and clustering—typically involve length scales exceeding the DNA bending persistence length (40–50 nm, corresponding to ≈120–150 base pairs), characterizing the resulting mesoscale structures, either at high resolution experimentally (19,20) or by all-atom molecular dynamics simulations (21–24), becomes a challenging endeavor. Consequently, coarse-grained simulations can offer a highly complementary view that can test mechanisms and provide microscopic insights not available directly from experiments in the spirit of a computational microscope (25,26). Coarse-grained simulations have provided many insights into DNA topology and dynamics (27–29), and coarse-grained simulations of simple DNA-protein models have contributed to our understanding of the formation of protein bridges and the resulting bridging-induced compaction (30–35). As a more specific example, coarse-grained models (36) explained the liquid droplet formation of heterochromatin due to heterochromatin protein 1 (37,38).

In a recent experimental work, we observed DNA compaction by HIV integrase (IN) via a "rosette" intermediate (i.e., a central nucleo-protein core with extruding DNA loops) and introduced a coarse-grained model to explain this behavior (39). Here, we present this model in detail, as well as the analysis method that we used to understand the formation of DNA-protein complexes. Our coarse-grained model consists of DNA interacting with proteins, where both protein-DNA as well as protein-protein interactions are tuned to match experimental observations from (39). Specifically, the DNA is represented by a discrete worm-like chain, while proteins are treated as simple spherical particles. Our model is generic and can be readily extended to other proteins, as it reduces the protein-DNA interactions to a simple isotropic pair potential with a defined binding valency without making specific assumptions about the binding geometry. The protein-DNA binding strength is taken from experimental measurements.

In order to characterize the effect of protein-protein interactions on the protein-mediated compaction of DNA, we present an unsupervised machine learning pipeline to systematically and quantitatively classify DNA-protein complexes: first, we define a set of structural parameters; next, we reduce the dimensionality via principal-component analysis (PCA); and finally, we divide the conformations into distinct groups with a Gaussian mixture model (GMM). We apply the approach to the IN-DNA compaction data (39) and find that protein-protein interactions are critical to the formation of looped intermediated structures (rosettes) seen experimentally. We expect our methodology to be widely applicable to DNA-binding proteins and protein-induced DNA compaction.

MATERIALS AND METHODS

We simulate the DNA-protein systems using Monte Carlo (MC) simulations of coarse-grained models written in C. In this section, we introduce the specific representations of both the DNA and proteins and explain how we determine the binding strength between the two to match with experiments (39).

Coarse-grained model for DNA

To model the double-stranded DNA, we use the common discrete worm-like chain model (11,15,33,40–43) and perform MC simulations of a double-stranded DNA in the canonical ensemble (44). This model is also frequently referred to as the beads-on-a-string model, as the DNA is treated as a string of N beads of diameter σ connected via finitely extensible nonlinear elastic (FENE) springs. The springs are described by the potential

$$\phi_{\mathrm{FENE}}(r) = -\frac{1}{2}k\,r_0^2\log\left(1-\left(\frac{r}{r_0}\right)^2\right)\ \text{for}\ r \le r_0 \quad (1)$$

and infinity otherwise. Here, $k = 21k_BT/\sigma^2$ is the spring constant (15), r is the center-to-center distance between two beads, and $r_0 = 1.5\sigma$ is the maximum extension of the springs. Note that throughout this work, we express energies in terms of the thermal energy k_BT, where k_B is the Boltzmann constant and T is the temperature. Excluded volume interactions between the DNA beads are included via the Weeks-Chandler-Andersen (WCA) potential

$$\phi_{\mathrm{WCA}}(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12}-\left(\frac{\sigma}{r}\right)^{6}+\frac{1}{4}\right)\ \text{for}\ r \le 2^{\frac{1}{6}}\sigma \quad (2)$$

and zero otherwise. The interaction strength ϵ is set to $0.7k_BT$ (15). The stiffness of the DNA chain is included via a bending potential

$$\phi_{\mathrm{bend}}(\theta) = k_\theta(1-\cos\theta). \quad (3)$$

Here, k_θ is the spring constant for bending, and θ is the angle between successive springs. In order to imitate the experimental DNA, which has a persistence length of approximately 40 nm, we use DNA beads of $\sigma = 4$ nm, which corresponds to the effective diameter under physiological ionic strength (27), and obtain the desired persistence length of by taking $k_\theta = 11k_BT$ (see the supporting material).

In this work, we mainly consider four lengths of DNA, 150, 289, 408, and 774 beads, which correspond to roughly 1.8, 3.4, 4.8, and 9.1 kbp, respectively. The latter three lengths correspond to DNA constructs used in experiments probing DNA interactions with HIV IN (39) to enable qualitative comparison of our simulation results to experiments. Depending on the length of the worm-like chain, proper equilibration may require a significant amount of time. Hence, to speed up equilibration, each chain is initialized as a random walk with a fixed step size of σ and bending angles evenly distributed between $0 \le \theta \le \theta_{\max}$. The maximum bending angle $\theta_{\max}$ is set to $\theta_{\max} = 2\cos^{-1}(e^{-\sigma/l_p})$, such that the initial chain has the approximate desired persistence length (42), i.e., $l_p = 10\sigma$. We equilibrate each DNA chain for at least 5×10^7 MC cycles. Note that, on average, each DNA bead undergoes one trial move per MC cycle (44). During such a trial move, the DNA bead is displaced by a random vector, whose individual x, y, and z components are drawn from a uniform distribution between $\pm0.12\sigma$. For this value, the self-diffusion time of an isolated 4 nm DNA bead is ∼100 MC cycles. We validate our model for DNA by reproducing the theoretical radius of gyration for a worm-like chain (see the supporting material).

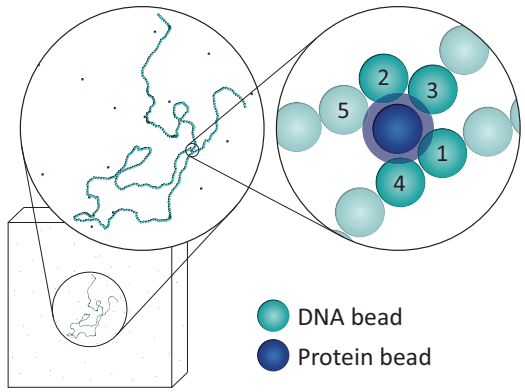

FIGURE 1 Schematic representation of the coarse-grained DNA-protein system. The zoom in shows the protein—with its isotropic binding potential—binding only the four nearest DNA beads. To see this figure in color, go online.

Coarse-grained model for proteins

Generally in DNA-protein systems, the proteins have a defined number of binding sites for DNA. For these kinds of systems, a patchy particle model is often used to simulate the coarse-grained multimer, see, e.g., (11,15,35,41,45–47). However, as the exact geometry of the binding sites is often either unknown or poorly defined due to conformational flexibility of the protein, we decided against the use of a patchy particle model to prevent any potential restrictions of the patch geometry on the compaction of DNA. Instead, we choose to reduce the complex structure of the protein to a single protein bead with an isotropic binding potential and restrict the binding valence of the protein to $n_{\max}$ DNA beads (Fig. 1). Similar valence models have frequently been used to approximate patchy particles, see, e.g., (48,49). In this section, we explain the mechanism behind this valence restriction.

For HIV IN, which is our main focus here, the active complex binding to DNA is thought to be a tetramer with four binding sites. The binding of HIV IN is relatively sequence unspecific, and its binding footprint was experimentally determined to be 10–20 bp (39), which corresponds roughly to one-tenth of the persistence length of DNA. For this reason, we model the DNA using beads of $\sigma = 4$ nm and the proteins as spheres with the same diameter σ. This ensures a binding footprint of 10–20 bp. However, one could easily simulate smaller or larger proteins. Preliminary tests with proteins of diameter 2σ gave qualitatively similar results to the ones reported in this work. This also suggests that weak deformability or softness of the protein bead would not qualitatively affect our results.

We assume a sequence-unspecific binding for the protein-DNA model, meaning that the protein can bind to any DNA bead, and ensure that the binding between the protein and DNA is short ranged, i.e., with an effective width of approximately 0.5 nm (50), by using the 18–36 Lennard-Jones (LJ) potential

$$\phi_{\text{bind}}(r) = 4\varepsilon_{\text{b}}\left(\left(\frac{\sigma}{r}\right)^{36} - \left(\frac{\sigma}{r}\right)^{18}\right) - \varepsilon_c \text{ for } r \le r_c \quad (4)$$

and zero otherwise. Here, ε_{b} is the binding strength, $r_c = 1.4\sigma$ is the distance of truncation, and ε_c is the energy shift such that the potential is zero at $r = r_c$. The valence restriction is established by introducing a bond swapping potential

$$\phi_{\text{swap}}(r) = \begin{cases} -\phi_{\text{bind}}(r_{\min}) & \text{for } r \le r_{\min}, \\ -\phi_{\text{bind}}(r) & \text{otherwise}, \end{cases} \quad (5)$$

where $r_{\min}$ is the minimum of the binding potential. This potential is inspired by the one of Sciortino (51) and provides the possibility of freely swapping between (potential) bonds, i.e., without any additional energy cost or gain, while preserving both the condition of detailed balance and the excluded volume interactions that occur for $r < r_{\min}$. To illustrate how this works in general, assume that one wants to restrict the maximum number of bonds per protein to $n_{\max}$. For each protein, one then wants to take only the energy gain of the $n_{\max}$ shortest bonds into account. To accomplish this, we first find all DNA beads within the distance r_c of a protein i and sort them according to their center-to-center distance. We then compute the total binding energy of protein i using

$$\phi^{(i)}_{\text{p-DNA}} = \sum_{j=1}^{n_{\text{bond}}} \phi_{\text{bind}}(r_{ij}) + \sum_{j=n_{\max}+1}^{n_{\text{bond}}} \phi_{\text{swap}}(r_{ij}), \quad (6)$$

where $\sum_j^{n_{\text{bond}}}$ is the sum over the sorted list of n_{bond} DNA beads j with $r_{ij} < r_c$. In this work, we use $n_{\max} = 4$ to match with the four binding sites of the active IN complex (39).

Lastly, we need to define the protein-protein interaction. As the exact protein-protein interaction is unknown, here, we (semi-)quantitatively tune the model such that the simulations reproduce the experimental observations from (39). We first look for a model that reproduces the protein-DNA structures observed in the experiments. Second, we further tune the model such that, given a specific DNA length, the critical protein concentration at which compaction sets in in the simulations approximately matches the critical concentration at which compaction arises in the experiments. Specifically, we consider two general cases for the protein-protein interaction. The first is a purely excluded volume interaction, which is included via the WCA potential with $\varepsilon = 0.7k_BT$, same as for the DNA beads (Eq. 2). Secondly, we consider the possibility of some form of mutual attraction between the proteins in addition to the excluded volume interactions. This is realized by changing the protein-protein interactions from the WCA potential into the regular 6–12 LJ potential with attraction strength ε_{pp}. The LJ potential is truncated and shifted at $r = 3\sigma$. We explore a couple of attraction strengths, from 0.7 to $5k_BT$, to find out what best fits the experimental observations.

Protein-DNA binding strength

The key parameter describing the protein-DNA interactions is the binding strength ε_{b}. Frequently, this binding energy has been treated as a free parameter in simulations of protein-DNA interactions, or it is estimated roughly from affinity measurements. Here, however, we use the binding strength taken from experimental measurements performed in (39). In order to obtain the protein-DNA binding strength, they used atomic force microscopy images of short double-stranded DNA constructs in the presence of relative low concentrations of protein. By counting the number of protein molecules bound to the DNA constructs, the binding probability can be directly determined and is, in turn, compared quantitatively to simulations to obtain ε_{b}. Note that the binding probability is computed by dividing the total number of bound protein molecules by the total number of DNA constructs considered and the number of possible binding sites on a DNA construct. The short DNA constructs and low IN concentration ensure that both multiple IN copies overlapping and protein-protein interactions are negligible. The binding probability increases monotonously with increasing binding strength with a crossover between simulations and experiments at a binding strength between 4.5 and $5.0k_BT$ (39). We, therefore, take $\varepsilon_{\text{b}} = 5k_BT$ in this work.

Note that for the experiments reported in (39), the IN concentrations are expressed in terms of the monomer concentration, [IN]. However, since the complexes actually binding to the DNA are thought to be tetramers consisting of four IN proteins, the protein concentrations in the simulations need to be adjusted to directly compare the experiments. Hence, we will express the protein concentration in terms of the monomer concentration throughout this work. Yet, keep in mind that the simulated protein beads, which represent the IN tetramers, will have a concentration equivalent to one-fourth of the monomer concentration.

Semi-grand canonical simulations to model DNA-protein mixtures

The experiments reported in (39) were performed at relatively low DNA concentrations, i.e., around 0.5 ng/μL, such that the observed protein-DNA condensates typically contained only a single DNA chain and the IN buffer was not depleted during the condensation. To reproduce this biologically relevant scenario in the simulations, we consider simulations of a single DNA strand treated in the canonical ensemble and treat the proteins in the grand canonical ensemble. As a substantial number of proteins can be involved in the protein-induced compaction of DNA, this ensures that the free proteins are not depleted in our simulation box. On top of the regular trial moves of a canonical ensemble, in the grand canonical ensemble, one needs trial insertions and removals of proteins (44). In order to prevent these insertions and removals from disturbing the DNA dynamics and allow us to explore the intermediate structures during compaction, we restrict the part of the simulation box where we perform insertion and deletion moves. Specifically, we take a spherical volume of radius R around the center of mass of the DNA chain and prohibit the insertion or removal of proteins in this volume. The radius R is taken as the distance to the furthest DNA bead plus an additional padding of $5r_c$. Mirroring the experimental conditions, we will consider protein concentrations in the nanomolar to micromolar range. In the supporting material, we provide more details on the semi-grand canonical simulations and demonstrate that the concentration of free protein is not depleted during the compaction of DNA.

RESULTS AND DISCUSSION

We examine the protein-induced compaction of DNA by performing a large set of simulations for the different protein-protein interactions (i.e., attractive and nonattractive) and for a large range of protein concentrations. For each protein-protein interaction and protein concentration, we perform 10–12 independent simulations. To initialize the combined DNA-protein systems, we use an equilibrated DNA chain and insert the desired concentration of protein at random positions in the box. After equilibrating the total system for 10^6 MC cycles with the DNA-protein interaction turned off, we gradually turn on the DNA-protein interaction within 100 MC cycles and simulate for a maximum of another 5×10^8 MC cycles. Each simulation took around 1–20 days on a single, modern CPU core.

During these simulations of the DNA-protein mixtures, we observe a range of different structures and conformations, depending on protein concentration and the form and strength of their mutual interactions. Some examples of typical DNA-protein conformations are shown in Fig. 2, *A–G*. Even though we can distinguish between some of these conformations by eye, it is difficult to classify all of them using a single order parameter like, e.g., the radius of gyration of the molecule. Hence, in order to systematically analyze and categorize the conformations formed in the simulations, we design an unsupervised machine learning pipeline.

The remainder of the results and discussion section is split in two. In the first part, we explain and set up the unsupervised machine learning pipeline for the classification of DNA-protein complexes. In the second part, we investigate the possible compaction pathways for the protein-mediated compaction of DNA by applying the trained machine learning pipeline to our simulations and classifying for each the structure of the DNA-protein complex as a function of time.

Unsupervised machine-learned classification

In short, the unsupervised machine learning pipeline for the classification of protein-DNA complexes operates as follows. First, it receives a multidimensional set of order parameters describing the geometrical characteristics of each conformation. It then applies a dimensionality reduction scheme, which extracts its most important features. Lastly, a clustering algorithm identifies distinct groups of conformations in the resulting lower-dimensional space. For the implementation of the dimensionality reduction and the grouping, we use the Python package scikit-learn (52). In this section, we explain the specifics of our classification approach in more detail.

Structural order parameters

To start, we first define a set of order parameters that capture the geometrical characteristics of each conformation. Note that for this initial selection of parameters, it is irrelevant whether parameters are independent or covary strongly. In the subsequent processing steps, these correlations are taken into account or can be removed if needed. So, while it is desirable to define parameters that capture a broad range of conformational features, it is not critical in our approach to a priori pick uncorrelated parameters. By looking at examples of different conformations (e.g., Fig. 2, *A–G*), we compose a set of 15 parameters, which can be grouped into different categories.

1) Global conformation of the DNA chain: the radius of gyration R_g, the normalized asphericity b/R_g^2, and the anisotropy κ of the DNA chain.
2) Bending angles of the DNA chain: the average and standard deviation of the bending angle, $\langle\theta\rangle$ and $s(\theta)$, respectively.
3) Proteins bound: the total number of proteins bound to DNA N_b, the fraction of DNA beads bound to at least one protein x_o, and the average and standard deviation of the number of proteins bound per DNA bead, $\langle n_b\rangle$ and $s(n_b)$, respectively.
4) Bare DNA segments: the total number of unoccupied DNA segments N_s and the average and standard deviation of the size of these unoccupied DNA segments, $\langle n_s\rangle$ and $s(n_s)$, respectively.
5) Protein clusters: the total number of clusters of bound protein N_c and the average and standard deviation of the size of these clusters, $\langle n_c\rangle$ and $s(n_c)$, respectively.

Exact definitions of the different parameters can be found in the supporting material.

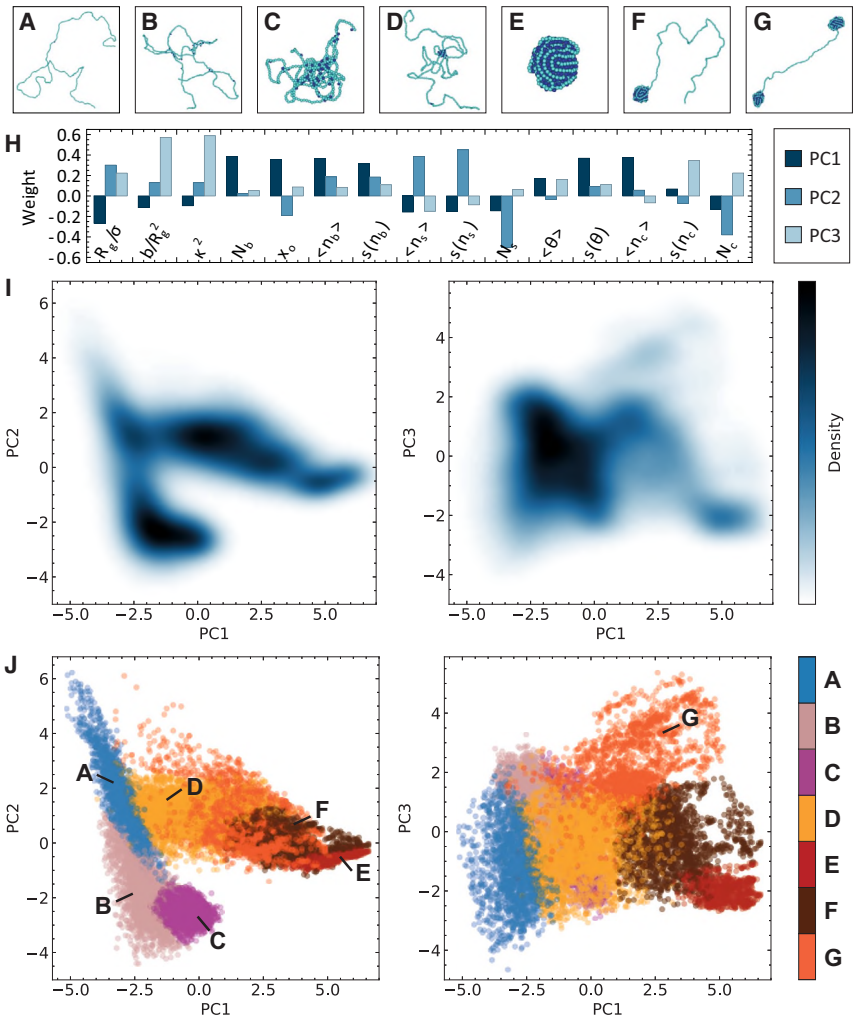

FIGURE 2 (*A-G*) Example conformations observed in simulations of DNA of 408 beads. Note that these are closeups and that the total three-dimensional simulation box is many times larger. The DNA beads are indicated by light blue spheres and the proteins are indicated by dark blue spheres. For visualization purposes, the free proteins are not displayed. (*H*) The weight of each order parameter in the composition of the first three principal components (PCs). (*I*) Density distribution of the training dataset in the PC1-PC2 plane and the PC1-PC3 plane. (*J*) Grouping according to the Gaussian mixture model (GMM). The seven groups are indicated with different colors, and they correspond to the typical conformations found in (*A*)–(*G*), i.e., (*A*) no compaction (*blue*), (*B*) bridging (*pink*), (*C*) bridging-induced compaction (*purple*), (*D*) rosette (*yellow*), (*E*) full compaction (*red*), (*F*) fully compacted complex with a bare tail (*brown*), and (*G*) multiple fully compacted complexes (*orange*). To see this figure in color, go online.

Dimensionality reduction

To extract the important features of the 15-dimensional space of order parameters, we need to reduce its dimensionality. There are many options, both linear and nonlinear, for unsupervised dimensionality reduction, e.g., PCA (53–56), autoencoders (57–60), and manifold learning methods (61–65). For our problem, we found that PCA, despite being purely linear, robustly provides a satisfactory separation of the various states of compaction. To check for the effects of linearity, we have also confirmed that using a simple nonlinear neural-network-based autoencoder, like the one in (59), does not improve our ability to classify structures for this problem. Since PCA is computationally efficient, deterministic, and parameter free, we here choose to use it for the remainder of this manuscript.

It is important to note that, in general, dimensionality reduction schemes require a balanced dataset, consisting of a fairly equal representation of the various possible states, for the scheme to perform well. Hence, as a first test to demonstrate our classification method, we focus on the DNA of 408 beads for which we compose a training set of nearly 15,000 configurations. This dataset contains roughly equal numbers of noncompacted configurations, configurations in different states of compaction mediated by mutual nonattractive proteins, and the same for compaction mediated by mutual attractive proteins with an attraction strength of $2.0k_BT$. We later explain how to extend this to include configurations of DNA of different lengths.

After constructing this balanced set of configurations, we normalize the distribution of each order parameter using standard scaling before feeding it to PCA. This ensures that each order parameter has an average value of zero and a variance of one, such that the variations in each parameter are treated as equally important. We determine the number of relevant principal components (PCs) by looking at the proportion of variance explained of each PC. In this case, the first three PCs combined capture more than 75% of the total variance of the dataset. Further analysis on the proportion of variance explained using the elbow method (66) also confirms the use of the first three PCs (see the supporting material).

The weight of each order parameter in the composition of the first three PCs is shown in Fig. 2 *H*, and Fig. 2 *I* shows the density distribution of the training dataset. We can already, by eye, distinguish some groups in this density distribution. For example, we see a distinct group on the top left of the PC1-PC2 plane from which two separate branches grow. The weights reveal that configurations in this group have a large radius of gyration and not many bound proteins; hence, this group most likely contains configurations of noncompacted DNA. Note that the weights also reveal that some order parameters are indeed highly correlated and therefore most likely redundant. For example, the asphericity and anisotropy have very similar weights, as well as the average and standard deviation of the number of proteins bound per DNA bead. At this stage, one could spend some time sieving out the redundant order parameters. However, since the resulting classification was already sufficient in our case, we did not do this here.

Identifying the distinct DNA-protein conformations

To finish up the classification, we use a clustering algorithm to divide the three-dimensional distribution of PCs into distinct regions. As for the dimensionality reduction, there are many options for clustering (67), e.g., K-means, spectral clustering and GMMs. We find that, in our case, the GMM provides a satisfactory distinction of the different groups in the PC landscape. We want to stress that the term "clustering" here means the classification of DNA-protein configurations into distinct groups with similar geometric features and should not be confused with the condensation or compaction of DNA-protein complexes, which can also be called clustering.

To determine the number of groups, we first look for a minimum in the Bayesian information criterion (68), which indicates how well a GMM fits the distribution while simultaneously penalizing the number of groups to prevent overfitting. However, as an alternative criterion to safeguard against overfitting, we additionally look for an elbow in the clustering entropy (69).

$$S_K = -\sum_{i=1}^{D}\sum_{j=1}^{K} P_{ij}\log\left(P_{ij}\right), \tag{7}$$

where D is the size of the dataset that needs to be grouped, K is the number of groups, and P_{ij} is the probability of data point i to belong to group j. We find that the optimum number of groups is seven (see the supporting material).

By definition, the GMM provides a "soft grouping," i.e., the probability for a given configuration to belong to any of the groups. To turn this into a discrete grouping, we assign each configuration to the group it is most likely to belong to. Fig. 2 *J* shows the resulting grouping using seven groups for the GMM. Comparing Fig. 2, *J* and *I*, we see that these groups nicely correspond to the groups visible in the density profile. Furthermore, by looking at various configurations belonging to these seven groups, we can identify them. Each group can be identified by one of the (deliberately chosen) example conformations depicted in Fig. 2, *A*–*G*, i.e., (Fig. 2 *A*) no compaction, (Fig. 2 *B*) bridging, (Fig. 2 *C*) bridging-induced compaction, (Fig. 2 *D*) rosette, (Fig. 2 *E*) full compaction, (Fig. 2 *F*) fully compacted complex with a bare tail, and (Fig. 2 *G*) multiple fully compacted complexes connected by bare segment(s) of DNA.

Two pathways for the protein-mediated compaction of DNA

In order to investigate the effect of the protein-protein interactions on the time evolution of the protein-mediated compaction of DNA, we apply our trained classification method to our simulations of DNA of 408 beads. For each simulation, we classify at regular time intervals the structure of the DNA-protein complex such that we can easily interpret its time evolution. To demonstrate the strength of the machine-learned classification, in Fig. 3, *A* and *B*, we show two typical simulation trajectories, the first for mutually nonattractive proteins (Fig. 3 *A*) and the second for mutually attractive proteins (Fig. 3 *B*). We clearly see that the two simulations follow two very different paths in the landscape of the PCs. Moreover, taking the classification of the GMM into account, we find that these are even two completely separate pathways, i.e., a pathway that goes via bridging to bridging-induced compaction for nonattractive protein-protein interactions (blue-pink-purple classification) and a pathway that goes via rosette to full compaction for attractive protein-protein interactions (blue-yellow-brown-red classification). (Note that here we observe two clear pathways that we connect to nonattractive and attractive protein-protein interactions. However, note that this is only true for sufficiently strong attraction.) The latter pathway corresponds to the structures observed in experiments (39).

By examining the time evolution of the classification of all 10–12 simulations per protein concentration and protein-protein interaction, we find that none of these simulations exhibit a crossover between the bridging and rosette states. This confirms that the bridging-induced compaction and rosette to full compaction are two completely separate pathways and, moreover, that the compaction pathway is fully determined by the protein-protein interaction. To illustrate that this result is independent of the protein concentration, we select the most typical simulation per concentration and show the associated time evolution of the classification in Fig. 3, *C* and *D*. These trajectories are selected based on the inverse of the time the system needs to start (irreversible) compaction and the time spent in either the bridging or rosette state. A trajectory is coined "most typical" when these (inverse) times best match with the average (inverse) times for the system at that protein concentration. One can

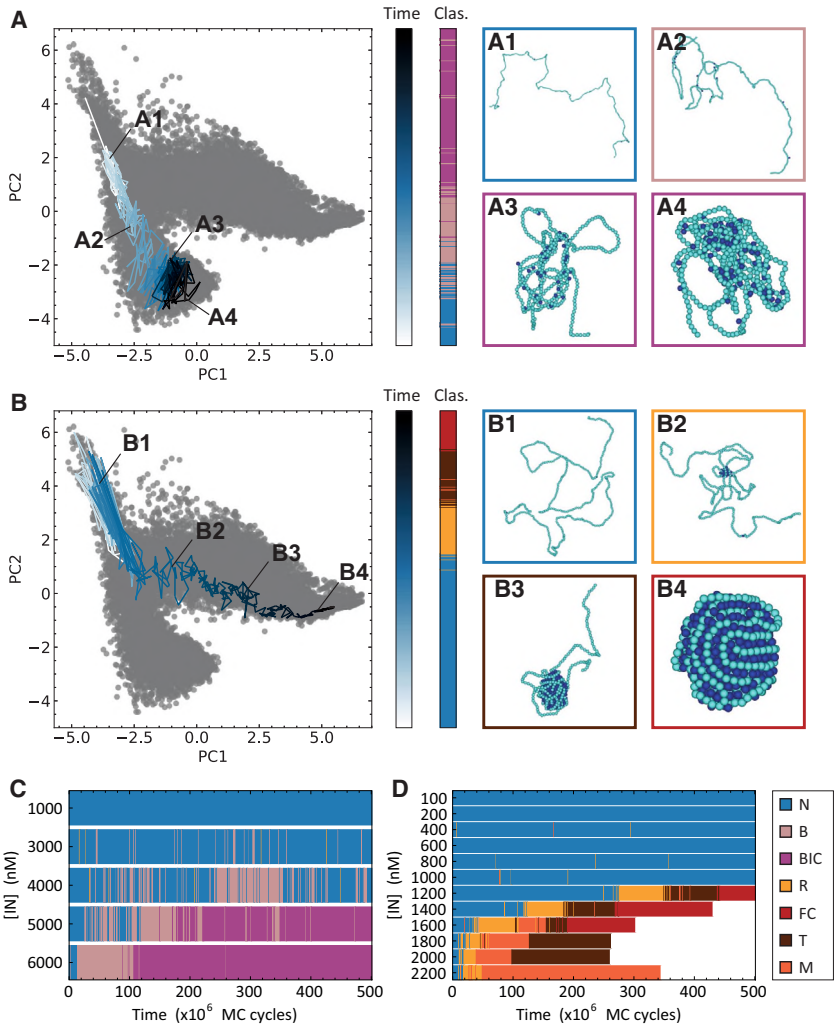

FIGURE 3 Typical simulations of DNA of 408 beads in (*A* and *C*) a system with nonattractive protein-protein interactions and (*B* and *D*) a system with attractive protein-protein interactions of strength $\beta\varepsilon_{pp} = 2.0$. (*A*) and (*B*) show, for protein concentrations of, respectively, [IN] = 5000 and [IN] = 1200 nM, the simulation trajectories on top of PC1-PC2 distribution of the training dataset. The trajectories are colored with a blue gradient indicating the simulation time (*first colorbar*), and the second colorbar is the classification according to the GMM. The figure blocks (*A1–A4*) and (*B1–B4*) each show four characteristic configurations during the simulation. (*C* and *D*) Time evolution of the classification as a function of protein concentration. Each bar represents the most typical run for each protein concentration studied. The seven options for the classification are (see Fig. 2) no compaction (N), bridging (B), bridging-induced compaction (BIC), rosette (R), full compaction (FC), fully compacted complex with a bare tail (T), and multiple fully compacted complexes (M). When a bar ends in white, it means that the simulation was terminated early, as the most interesting behavior had already happened. To see this figure in color, go online.

clearly see that there is no crossover between the pathway of bridging-induced compaction (observed in Fig. 3 *C*) and the pathway of rosette to full compaction (observed in Fig. 3 *D*).

Furthermore, it is noteworthy that from these simulations, we can conclude that protein-protein attractions promote DNA compaction; Fig. 3, *C* and *D*, clearly show that DNA compaction in systems with $2k_BT$ protein-protein attraction sets in at lower protein concentrations than in systems without protein-protein attraction. Although this is not surprising—protein-protein attractions naturally lead to a higher affinity to compact—it is interesting to explore the effect of the attraction strength on the onset of DNA compaction.

Role of DNA length and protein-protein attraction strength

To explore the effect of protein-protein attraction strength in the protein-induced compaction of DNA, we focus on attraction strengths of 1.5 and $2.0k_BT$ and protein concentrations from 100 to 2400 nM. Additionally, we investigate the role of DNA length by considering the DNA of 150, 289, 408, and 774 beads. Importantly, given the dependence of some of the structural order parameters used as input in the classification on DNA length (e.g., the number of bound proteins or the length and number of bare DNA segments), we need to adjust our classification to accommodate different DNA lengths. In order to use one pipeline for the classification of systems with different DNA lengths, we first obtain a balanced training dataset for each DNA length, which we then normalize separately from one another. This ensures that even though the important features differ in absolute values, they are treated as roughly equal in each normalized dataset. Note that we were able to do this because no new or distinctly different structures emerged as a result of varying the DNA length and protein-protein attraction strength.

Next, we combine these separately normalized sets and obtain a new classification by retraining both the PCA and GMM on the combined dataset. To demonstrate that the new

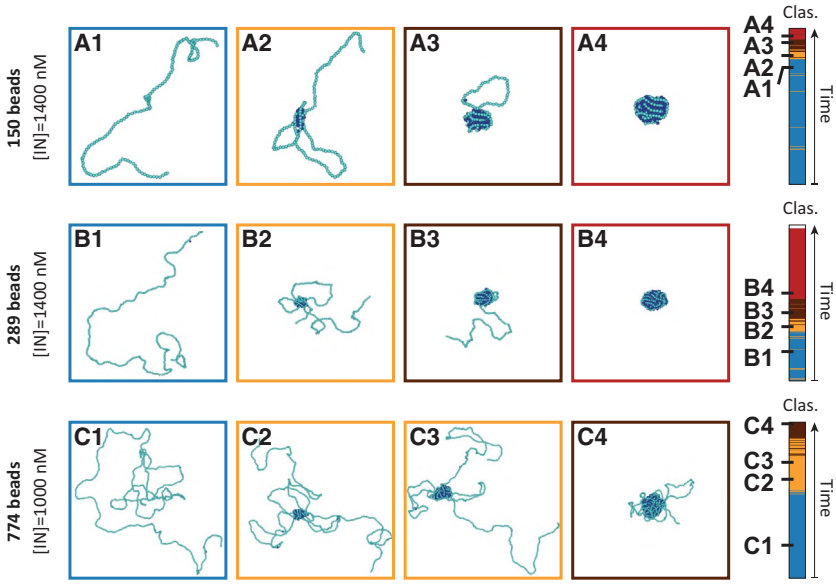

FIGURE 4 Short time series of a typical simulation of DNA of (*A1–A4*) 150, (*B1–B4*) 289, and (*C1–C4*) 774 beads. In all cases, the DNA is in solution with mutual attractive proteins ($\beta\varepsilon_{pp} = 2.0$). The colorbars indicate the classification as a function of time (see Fig. 3). To see this figure in color, go online.

classification is indeed able to identify conformations like the rosette conformation for the different DNA lengths, we show a short time series of a typical trajectory for the DNA of 150, 289, and 774 beads in Fig. 4. Even though the rosette conformation, for example, looks undeniably different for different DNA lengths, the classification is able to identify it as the same conformation for all DNA lengths.

Using the newly trained pipeline, we classify all 12 simulations per DNA length and protein-protein attraction strength. To illustrate the influence of both DNA length and protein-protein attraction strength on the protein-mediated compaction of DNA, we again select the most typical simulation per concentration and system and show the associated time evolution of the classification in Fig. 5. Note that, as mentioned before, these typical trajectories are selected based on the inverse of the time the system needs to start (irreversible) compaction and the time spent in the rosette state. There are four key observations that can be deduced from Fig. 5. First, we find that all trajectories leading to compaction go through a transient rosette state (indicated in yellow), independent of the DNA length or protein concentration. Second, we observe that the time spent in a rosette conformation increases with the DNA length. Third, it is evident that stronger protein-protein attractions facilitate DNA compaction at lower protein concentrations for all investigated DNA lengths. Lastly, we see that an increase of DNA length similarly enables compaction at lower protein concentrations.

Note that the experiments demonstrated the same dependency on the DNA length (39). Additionally, by tuning the protein-protein attraction strength to $1.5 - 2.0 k_B T$, we approximately matched the protein concentrations at which compaction sets in to the experimental protein concentrations at which full compaction sets in. Preliminary tests with $\varepsilon_{pp} > 3.0 k_B T$ revealed structures not observed in experiments, such as fully compacted DNA-protein structures covered in additional proteins or (large) aggregates of free protein.

In order to translate some of our observations into quantitative measures, we compute the rate of compaction as a function of protein concentration for the different systems. Assuming that the start of compaction in our simulations is a rare event that has a censored exponential distribution, i.e., not necessarily all simulations managed to compact, we can approximate the rate λ with (70).

$$\lambda = \frac{m}{\sum_{i=1}^{n} t_i}, \tag{8}$$

where the sum is taken over the $n = 12$ simulations performed for a specific system, m is the number of simulations in which the DNA managed to compact, and t_i is the time of simulation i at which (irreversible) compaction starts. Practically, t_i is the last instance of a conformation classified as not compacted (indicated by *blue* in Fig. 5). Note that the sum over all $n = 12$ simulations also includes the time of the simulations in which the DNA did not manage to compact. The resulting rates are given in Fig. 6. Here, we have normalized the rates with the DNA length, such that a rough collapse onto a master curve for both protein-protein attraction strengths is revealed. The trends of these master curves suggest that the compaction of DNA is an activating event with a rate given by

$$\lambda \Big/ N = C_0 \exp\left(-\frac{C_1}{c_p}\right), \tag{9}$$

where c_p is the protein concentration and C_0 and C_1 are parameters that depend on the protein-protein attraction strength.

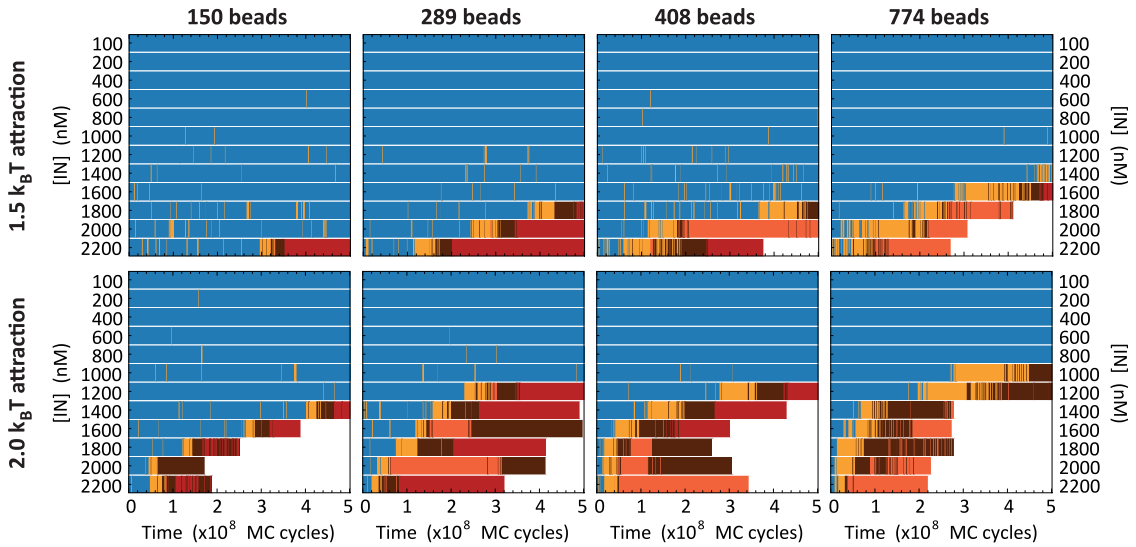

FIGURE 5 The most typical runs for a range of protein concentrations for the four DNA lengths studied. The top row gives the results for a mutual protein interaction with attraction strength $\beta\varepsilon_{pp} = 1.5$ and the bottom row $\beta\varepsilon_{pp} = 2.0$. See Fig. 3 for the interpretation of the classification that the colors of the bars represent. To see this figure in color, go online.

Note that although the translation to real time is not as direct as for, e.g., molecular dynamics simulations, the obtained rates can still be roughly related to real time given that the self-diffusion time of an isolated 4 nm DNA bead is ~100 MC cycles in our simulations and corresponds to roughly 25 ns. Moreover, the relative trends obtained from comparing different systems give legitimate insights into the workings of protein-mediated compaction of DNA.

CONCLUSION

To conclude, we presented a simple coarse-grained model for DNA-protein mixtures that treats the protein-DNA interactions as an isotropic binding potential with an imposed binding valency without specific assumptions about the binding geometry. While we have designed this model to capture (semi-)quantitatively the behavior of DNA mixed with HIV IN, this approach for protein-DNA binding presents a generic solution for the many cases in which the exact geometry of the binding sites is either unknown or poorly defined due to conformational flexibility of the protein. Additionally, we designed a simple, fast, and effective unsupervised machine learning model for the classification of the DNA-protein complexes into different conformational states. We applied our model to recent data on the compaction of viral genome-length DNA by HIV IN and found that protein-protein attractions are critical to the formation of looped intermediated structures ("rosettes") observed experimentally. Not only is our model applicable to a broad range of different protein and nucleic acid systems, but the additional unsupervised learning method for the classification of the intricate complexes formed in such systems can offer key insights into the variety of conformational states and their formation pathways.

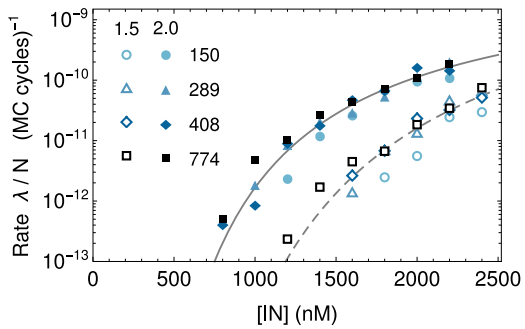

FIGURE 6 The rate of compaction divided by the DNA length as a function of the protein concentration. The shapes of the markers indicate the four different lengths of DNA, and the two different attraction strengths $\beta\varepsilon_{pp} = 1.5$ and $\beta\varepsilon_{pp} = 2.0$ are indicated by the open and closed markers, respectively. The lines indicate fits to Eq. 9. To see this figure in color, go online.

DATA AND CODE AVAILABILITY

A data package containing the data supporting the findings of this study as well as the self-developed C codes for performing the simulations and C and Python codes for the structural analysis is openly available on Zenodo at https://doi.org/10.5281/zenodo.12770995.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2024.07.023.

AUTHOR CONTRIBUTIONS

Experimental work of P.J.K., W.V., and J.L. initiated the research question. The model, simulation methods, and classification algorithm were designed by L.F. and M.d.J. L.F. supervised the project. M.d.J. wrote all the code and performed the simulations and analysis. All authors contributed to the interpretation and the manuscript.

ACKNOWLEDGMENTS

We would like to thank Frank Smallenburg and Rinske Alkemade for many useful discussions. We acknowledge funding from the Vidi research program with project number VI.VIDI.192.102, which is financed by the Dutch Research Council (NWO) and Utrecht University.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPORTING CITATIONS

References (71–73) appear in the supporting material.

REFERENCES

1. Liu-Yesucevitz, L., A. Bilgutay, …, B. Wolozin. 2010. Tar DNA binding protein-43 (TDP-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. *PLoS One.* 5:e13250.
2. Hyman, A. A., C. A. Weber, and F. Jülicher. 2014. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* 30:39–58.
3. Bergeron-Sandoval, L.-P., N. Safaee, and S. W. Michnick. 2016. Mechanisms and consequences of macromolecular phase separation. *Cell.* 165:1067–1079.
4. Frykholm, K., L. K. Nyberg, and F. Westerlund. 2017. Exploring DNA–protein interactions on the single DNA molecule level using nanofluidic tools. *Integr. Biol.* 9:650–661.
5. André, A. A. M., and E. Spruijt. 2018. Rigidity rules in DNA droplets: Nucleic acid flexibility affects model membraneless organelles. *Biophys. J.* 115:1837–1839.
6. Sazer, S., and H. Schiessel. 2018. The biology and polymer physics underlying large-scale chromosome organization. *Traffic.* 19:87–104.
7. Sawyer, I. A., J. Bartek, and M. Dundr. 2019. Phase separated microenvironments inside the cell nucleus are linked to disease and regulate epigenetic state, transcription and RNA processing. *Semin. Cell Dev. Biol.* 90:94–103.
8. Choi, J.-M., A. S. Holehouse, and R. V. Pappu. 2020. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* 49:107–133.
9. Weinmann, R., L. Frank, and K. Rippe. 2023. Approaches to characterize chromatin subcompartment organization in the cell nucleus. *Curr. Opin. Struct. Biol.* 83:102695.
10. Uversky, V. N. 2023. Biological Liquid–Liquid Phase Separation, Biomolecular Condensates, and Membraneless Organelles: Now You See Me, Now You Don't. *Int. J. Mol. Sci.* 24:13150.
11. Brackley, C. A., S. Taylor, …, D. Marenduzzo. 2013. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. USA.* 110:E3605–E3611.
12. Jiang, K., N. Humbert, …, F. Westerlund. 2021. The HIV-1 nucleocapsid chaperone protein forms locally compacted globules on long double-stranded DNA. *Nucleic Acids Res.* 49:4550–4563.
13. Gien, H., M. Morse, …, M. C. Williams. 2022. HIV-1 nucleocapsid protein binds double-stranded DNA in multiple modes to regulate compaction and capsid uncoating. *Viruses.* 14:235.
14. Moreno-Herrero, F., L. Holtzer, …, N. H. Dekker. 2005. Atomic force microscopy shows that vaccinia topoisomerase IB generates filaments on DNA in a cooperative fashion. *Nucleic Acids Res.* 33:5945–5953.
15. Ryu, J.-K., C. Bouchoux, …, C. Dekker. 2021. Bridging-induced phase separation induced by cohesin SMC protein complexes. *Sci. Adv.* 7:eabe5905.
16. Chappidi, N., T. Quail, …, S. Alberti. 2024. PARP1-DNA co-condensation drives DNA repair site assembly to prevent disjunction of broken DNA ends. *Cell.* 187:945–961.e18.
17. Farge, G., M. Mehmedovic, …, M. Falkenberg. 2014. In vitro-reconstituted nucleoids can block mitochondrial DNA replication and transcription. *Cell Rep.* 8:66–74.
18. Isaac, R. S., T. W. Tullius, …, L. S. Churchman. 2024. Single-nucleoid architecture reveals heterogeneous packaging of mitochondrial DNA. *Nat. Struct. Mol. Biol.* 31:568–577.
19. Jones, S., P. Van Heyningen, …, J. M. Thornton. 1999. Protein-DNA interactions: a structural analysis. *J. Mol. Biol.* 287:877–896.
20. Marsh, J. A., and S. A. Teichmann. 2015. Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* 84:551–575.
21. Orozco, M., A. Pérez, …, F. J. Luque. 2003. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* 32:350–364.
22. Pérez, A., F. J. Luque, and M. Orozco. 2012. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.* 45:196–205.
23. Noy, A., T. Sutthibutpong, and S. A Harris. 2016. Protein/DNA interactions in complex DNA topologies: expect the unexpected. *Biophys. Rev.* 8:145–155.
24. Yoo, J., D. Winogradoff, and A. Aksimentiev. 2020. Molecular dynamics simulations of DNA–DNA and DNA–protein interactions. *Curr. Opin. Struct. Biol.* 64:88–96.
25. Lee, E. H., J. Hsin, …, K. Schulten. 2009. Discovery through the computational microscope. *Structure.* 17:1295–1306.
26. Dror, R. O., R. M. Dirks, …, D. E. Shaw. 2012. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* 41:429–452.
27. Vologodskii, A. V., and N. R. Cozzarelli. 1994. Conformational and thermodynamic properties of supercoiled DNA. *Annu. Rev. Biophys. Biomol. Struct.* 23:609–643.
28. Schöpflin, R., H. Brutzer, …, G. Wedemann. 2012. Probing the elasticity of DNA on short length scales by modeling supercoiling under tension. *Biophys. J.* 103:323–330.
29. Ott, K., L. Martini, …, U. Gerland. 2020. Dynamics of the buckling transition in double-stranded DNA and RNA. *Biophys. J.* 118:1690–1701.
30. Barbieri, M., M. Chotalia, …, M. Nicodemi. 2012. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA.* 109:16173–16178.
31. Johnson, J., C. A. Brackley, …, D. Marenduzzo. 2015. A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction. *J. Phys. Condens. Matter.* 27:064119.
32. Brackley, C. A., J. Johnson, …, D. Marenduzzo. 2016. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Res.* 44:3503–3512.
33. Brackley, C. A., B. Liebchen, …, D. Marenduzzo. 2017. Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains. *Biophys. J.* 112:1085–1093.
34. Nguemaha, V., and H.-X. Zhou. 2018. Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation. *Sci. Rep.* 8:6728.

35. Joseph, J. A., J. R. Espinosa, …, R. Collepardo-Guevara. 2021. Thermodynamics and kinetics of phase separation of protein-RNA mixtures by a minimal model. *Biophys. J.* 120:1219–1230.

36. Ancona, M., and C. A. Brackley. 2022. Simulating the chromatin-mediated phase separation of model proteins with multiple domains. *Biophys. J.* 121:2600–2612.

37. Larson, A. G., D. Elnatan, …, G. J. Narlikar. 2017. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature.* 547:236–240.

38. Strom, A. R., A. V. Emelyanov, …, G. H. Karpen. 2017. Phase separation drives heterochromatin domain formation. *Nature.* 547:241–245.

39. Kolbeck, P. J., M. de Jager, …, W. Vanderlinden. 2024. HIV integrase compacts viral DNA into biphasic condensates. Preprint at bioRxiv. https://www.biorxiv.org/content/early/2024/03/17/2024.03.15.585256.

40. Formanek, M., L. Rovigatti, …, A. J. Moreno. 2021. Gel Formation in Reversibly Cross-Linking Polymers. *Macromolecules.* 54:6613–6627.

41. Hafner, A. E., J. Krausser, and A. Šarić. 2019. Minimal coarse-grained models for molecular self-organisation in biology. *Curr. Opin. Struct. Biol.* 58:43–52.

42. Auhl, R., R. Everaers, …, S. J. Plimpton. 2003. Equilibration of long chain polymer melts in computer simulations. *J. Chem. Phys.* 119:12718–12728.

43. Kremer, K., and G. S. Grest. 1990. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.* 92:5057–5086.

44. Frenkel, D., and B. Smit. 2002. Understanding Molecular Simulation: From Algorithms to Applications, 2nd edition. Academic Press, San Diego.

45. Brackley, C. A. 2020. Polymer compaction and bridging-induced clustering of protein-inspired patchy particles. *J. Phys. Condens. Matter.* 32:314002.

46. Samanta, R., and V. Ganesan. 2018. Influence of protein charge patches on the structure of protein–polyelectrolyte complexes. *Soft Matter.* 14:9475–9488.

47. Bianchi, E., B. Capone, …, P. D. J. van Oostrum. 2017. Limiting the valence: advancements and new perspectives on patchy colloids, soft functionalized nanoparticles and biomolecules. *Phys. Chem. Chem. Phys.* 19:19847–19868.

48. Zaccarelli, E., I. Saika-Voivod, …, F. Sciortino. 2006. Gel to glass transition in simulation of a valence-limited colloidal system. *J. Chem. Phys.* 124:124908.

49. Rovigatti, L., G. Nava, …, F. Sciortino. 2018. Self-dynamics and collective swap-driven dynamics in a particle model for vitrimers. *Macromolecules.* 51:1232–1241.

50. Tse, C., L. Wickstrom, …, N. Deng. 2020. Exploring the free-energy landscape and thermodynamics of protein-protein association. *Biophys. J.* 119:1226–1238.

51. Sciortino, F. 2017. Three-body potential for simulating bond swaps in molecular dynamics. *Eur. Phys. J. E.* 40:3–4.

52. Pedregosa, F., G. Varoquaux, …, E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

53. Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *Mag. J. Sci.* 2:559–572.

54. van Damme, R., G. M. Coli, …, M. Dijkstra. 2020. Classifying Crystals of Rounded Tetrahedra and Determining Their Order Parameters Using Dimensionality Reduction. *ACS Nano.* 14:15144–15153.

55. Gardin, A., C. Perego, …, G. M. Pavan. 2022. Classifying soft self-assembled materials via unsupervised machine learning of defects. *Commun. Chem.* 5:82.

56. Coslovich, D., R. L. Jack, and J. Paret. 2022. Dimensionality reduction of local structure in glassy binary mixtures. *J. Chem. Phys.* 157:204503.

57. Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford University Press.

58. Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep Learning. MIT Press.

59. Boattini, E., M. Dijkstra, and L. Filion. 2019. Unsupervised learning for local structure detection in colloidal systems. *J. Chem. Phys.* 151:154901.

60. Boattini, E., S. Marín-Aguilar, …, L. Filion. 2020. Autonomously revealing hidden local structures in supercooled liquids. *Nat. Commun.* 11:1–9.

61. Reinhart, W. F., A. W. Long, …, A. Z. Panagiotopoulos. 2017. Machine learning for autonomous crystal structure identification. *Soft Matter.* 13:4733–4745.

62. Adorf, C. S., T. C. Moore, …, S. C. Glotzer. 2020. Analysis of self-assembly pathways with unsupervised machine learning algorithms. *J. Phys. Chem. B.* 124:69–78.

63. Statt, A., D. C. Kleeblatt, and W. F. Reinhart. 2021. Unsupervised learning of sequence-specific aggregation behavior for a model copolymer. *Soft Matter.* 17:7697–7707.

64. Reinhart, W. F. 2021. Unsupervised learning of atomic environments from simple features. *Comput. Mater. Sci.* 196:110511.

65. Allegra, M., E. Facco, …, A. Mira. 2020. Data segmentation based on the local intrinsic dimension. *Sci. Rep.* 10:16449.

66. Salvador, S., and P. Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In* 16th IEEE International Conference on Tools with Artificial Intelligence IEEE, pp. 576–584.

67. Press, W. H., W. T. Vetterling, …, B. P. Flannery. 2007. Numerical Recipes: The Art of Scientific Computing, 3rd edition. Cambridge University Press.

68. Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

69. Baudry, J.-P., A. E. Raftery, …, R. Gottardo. 2010. Combining mixture components for clustering. *J. Comput. Graph Stat.* 9:332–353.

70. Deemer, W. L., and D. F. Votaw, Jr. 1955. Estimation of parameters of truncated or censored exponential distributions. *Ann. Math. Stat.* 26:498–504.

71. Hays, J. B., M. E. Magar, and B. H. Zimm. 1969. Persistence length of DNA. *Biopolymers.* 8:531–536.

72. Benoit, H., and P. Doty. 1953. Light scattering from non-Gaussian chains. *J. Phys. Chem.* 57:958–963.

73. Robertson, R. M., S. Laib, and D. E. Smith. 2006. Diffusion of isolated DNA molecules: Dependence on length and topology. *Proc. Natl. Acad. Sci. USA.* 103:7310–7314.