PNAS Nexus, 2026, **5**, pgag167
https://doi.org/10.1093/pnasnexus/pgag167
Published: 15 May 2026
Research report

Searching for sequence features that control DNA cyclizability

Margarita Gordiychuk[1], Jonghan Park[2], Aakash Basu[3,4], Taekjip Ha[3,5,6,7,8,9], William Bialek[10,11]*, and Yaojun Zhang[1,5,10]*

[1]Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Medicine, Yonsei University College of Medicine, Seoul, South Korea
[3]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[4]Department of Biosciences, Durham University, Durham DH1 3LE, United Kingdom
[5]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA
[6]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA
[7]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA
[8]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115, USA
[9]Howard Hughes Medical Institute, Boston, MA 02115, USA
[10]Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[11]Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave, New York, NY 10016, USA
***To whom correspondence should be addressed: Email: wbialek@princeton.edu (W.B.); Email: yaojunz@jhu.edu (Y.Z.)
Edited By Ozlem Keskin

Abstract

The mechanical properties of DNA molecules are crucial for many biological processes, from DNA packaging to transcription. While the mechanics of long DNA typically follow the worm-like chain polymer model, multiple studies have shown that the mechanics of short DNA, at the length scale of DNA–protein interactions, depend strongly on their sequence content. Motivated by recent high-throughput measurements of sequence-dependent DNA cyclizability—the DNA's tendency to mechanically bend and form a loop—we developed a statistical-mechanics framework to systematically explore how cyclizability depends on the collective contributions of an increasing number of nucleotides in the sequence. By applying the method to datasets of randomly generated and biologically derived sequences, we identified a minimal pairwise model that describes the sequence-dependence of DNA cyclizability. The pairwise model enabled the extraction of characteristic sequence features that control DNA cyclizability and predicted the most and least cyclizable sequences, which we validated through all-atom molecular dynamics simulations. Our work advances current understanding of sequence-dependent DNA mechanics and its role in various biological processes, with implications for the growing field of DNA nanofabrication.

Keywords DNA cyclizability, sequence-dependent DNA mechanics, statistical-mechanics modeling

Significance statement

It has long been recognized that the functions of biological molecules, such as proteins and nucleic acids, are encoded by their sequences. However, identifying the relevant sequence features that control their functional behaviors has proven challenging due to the high dimensionality of sequence space. Here, we develop an interpretable statistical-mechanics framework to understand how the sequences of short DNA segments determine their probability of looping. Our results reveal key sequence features that promote or suppress DNA looping and predict sequences with extreme looping probabilities. This work provides mechanistic insight into the sequence dependence of DNA mechanics, with implications for natural and synthetic systems. The developed framework is also widely applicable to problems in which high-dimensional inputs determine system-level functional outputs.

Received: October 16, 2025. **Accepted:** May 4, 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of National Academy of Sciences.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

The double-helical DNA is not just a passive repository of genetic information but also an active physical entity that interacts with diverse protein machinery to regulate essential biological processes, such as genome packaging, transcription, DNA replication, repair, and recombination. Increasingly, the mechanical properties of DNA are recognized to play critical roles in these processes (1, 2). For instance, DNA mechanics can influence where nucleosomes form (3, 4) and how DNA is packaged in cells (5, 6) or viruses (7), how DNA interacts with transcription factors (8, 9) and how probable it is to establish enhancer–promoter interactions via DNA looping (10, 11), as well as how DNA mismatches recruit repair machinery (12).

The mechanical behavior of DNA is commonly described by the worm-like chain (WLC) polymer model (13, 14), which treats DNA as an elastic rod with a bending persistence length of ∼50 nm (∼150 base pairs). While the WLC model successfully captures the average DNA mechanics at long length scales, it predicts that DNA segments shorter than the persistence length are essentially rigid rods. However, such prediction is contradicted by observations of prevalent spontaneous large-angle bends in DNA at short length scales (tens of base pairs) (15, 16). More importantly, single-molecule assays over the years have shown that DNA segments can exhibit complex mechanical behaviors highly dependent on their nucleotide sequences (17–19). Epigenetic modifications (eg methylation) and DNA mismatches can further induce noncanonical DNA structures that significantly alter DNA mechanics, thereby modulating their functional outcomes (12, 20, 21).

A key experimental approach to characterizing DNA mechanics is through the measurement of cyclizability, which quantifies a polymer's ability to bend and form a loop. DNA cyclization has been investigated from a thermodynamic perspective as quantified by the Jacobson–Stockmayer (J) factor (22–24). Multiple studies have demonstrated that sequence content can significantly affect the J-factor of short DNA (18, 25, 26). However, a systematic characterization of how DNA cyclizability depends on sequence features has been lacking due to the limited number of probed sequences. Recently, a novel high-throughput sequencing-based method called loop-seq was developed (27) (Fig. 1), which enabled the simultaneous measurement of cyclizability across hundreds of thousands of sequences. This extensive dataset has inspired several deep neural network-based approaches to understand sequence-dependent DNA cyclizability and predict the DNA mechanics at the genome level (9, 28–31). Yet, while these machine-learning approaches achieve high predictive accuracy, they often operate as "black boxes," providing limited physical insight into how DNA sequence encodes cyclizability.

Here, we develop a statistical-mechanics framework to understand the sequence-dependence of DNA cyclizability using high-throughput loop-seq measurements. We first model cyclizability as a sum of collective contributions from an increasing number of nucleotides, from which we identify a minimal pairwise model that captures the observed sequence dependence. We then analyze the interaction matrix from the pairwise model to extract sequence features that encode cyclizability, and predict the most and least cyclizable sequences. Finally, we validate key model predictions using all-atom molecular dynamics simulations and discuss potential extensions to higher-order models.

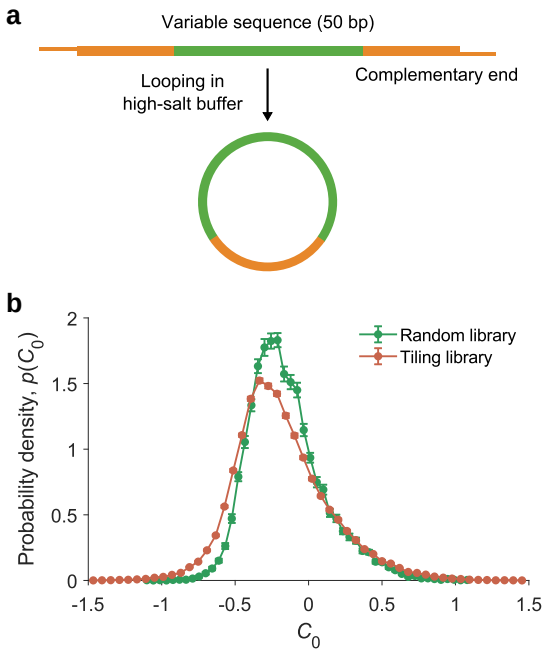

Figure 1 High-throughput measurements of sequence-dependent DNA cyclizability (27). a) Experimental construct. b) The probability distribution of the intrinsic cyclizability C_0 for the Random and Tiling libraries. The mean and SD are calculated using random sampling of halves of the data.

Results

High-throughput data on DNA cyclizability

Recent loop-seq experiments (27) have chosen hundreds of thousands of DNA sequences that span yeast genome and random sequences, and estimated the intrinsic cyclizabilities of these sequences by measuring the probability that they close on themselves into a loop, Fig. 1a. In brief, selected variable sequences of length $N = 50$ were flanked by fixed double-stranded adapters (length 25) and single-stranded complementary overhangs (length 10), and immobilized on a bead. The looping reaction was initiated in high-salt buffer for 1 min, which reduces electrostatic repulsion along DNA to promote loop formation. Afterward, the unlooped molecules were digested by an exonuclease that attacks only free ends. The remaining population of looped molecules was sequenced and compared with the control ensemble in which the digestion step was omitted. Cyclizability was defined as the natural logarithm of the probability ratio for finding a sequence in the looped versus control ensemble. Observations on a small number of sequences showed that this measure correlates very well with direct single-molecule measurements of cyclizability quantified by fluorescence resonance energy transfer. Please refer to Ref. (27) for experimental details.

The measured cyclizability depends periodically on the location of the bead attachment. The intrinsic cyclizability C_0 was defined as the mean over this variation. The values of C_0 were further refined using a mathematical operation to eliminate the bias introduced by the adapters and overhangs. After these adjustments, C_0 becomes independent of tether location and sequence orientation. Please refer to Ref. (9) for data processing details. The loop-seq data contains multiple libraries of sequences. In this study, we utilize two of them: the Random library, which consists of 12,404

randomly generated sequences, and the Tiling library, which consists of 81,801 sequences that tile around 576 selected genes in the *Saccharomyces cerevisiae* genome. The distributions of C_0 across the sequences in the two libraries are shown in Fig. 1b.

A statistical mechanics approach to sequence-dependent cyclizability

The intrinsic cyclizability of a sequence is a collective variable determined by contributions from all base sites. To model the cyclizability, we developed a statistical mechanics approach in which cyclizability is systematically expressed in terms of collective contributions from an increasing number of bases. To make it concrete, we first convert nucleotide sequences into Potts variables (also known as one-hot encoding). We represent DNA sequences as $\{S_i^\alpha\}$, where $S_i^\alpha = 1$ if the base at site i is of type α, and $S_i^\alpha = 0$ otherwise. The index $i = 1, 2, \ldots, N$, where N is the sequence length, and $\alpha = 1, 2, 3, 4$, corresponding to A, T, C, G. In the matrix form, every column of $\{S_i^\alpha\}$ is a unit vector denoting the encoded nucleotide type. For example, $\{S_i^\alpha\}$ for sequence AGTCGTT$\cdots$AA is

$$\begin{matrix} & & & & & & & & i\rightarrow & \\ \left(\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \end{matrix}\right) & & & & & & & & & \downarrow \alpha. \end{matrix}$$

We then express the cyclizability as a cluster expansion, accounting for one-base, two-base, three-base contributions, and so on, in a hierarchical order:

$$\begin{aligned} C_0 = \langle C_0\rangle &+ \sum_{i,\alpha} W_i^\alpha \left(S_i^\alpha - \langle S_i^\alpha\rangle\right) \\ &+ \frac{1}{2}\sum_{i\neq j,\alpha\beta} J_{ij}^{\alpha\beta}(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle) \\ &+ \frac{1}{6}\sum_{i\neq j\neq k,\alpha\beta\gamma} G_{ijk}^{\alpha\beta\gamma}(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)(S_k^\gamma - \langle S_k^\gamma\rangle) \\ &+ O(S^4), \end{aligned} \tag{1}$$

where W_i^α, $J_{ij}^{\alpha\beta}$, and $G_{ijk}^{\alpha\beta\gamma}$ are the coupling constants.

To provide a rationale for Eq. (1), we note that cyclizability is a global quantity which, in principle, would require a complete expansion in sequence variables to yield an exact description. In practice, however, experience from other contexts, such as maximum-entropy-type models, suggests that low-order contributions often capture much of the relevant dependence for many observables (32–34). Motivated by this perspective, we treat Eq. (1) as a systematic hierarchy and examine its terms order by order in the following, with the goal of identifying a minimal model that captures the essential sequence features governing cyclizability.

A linear model

The simplest model for how the cyclizability depends on sequence is linear, where the cyclizability is a weighted sum of contributions from the individual nucleotides:

$$C_0 = \langle C_0\rangle + \sum_{i,\alpha} W_i^\alpha \left(S_i^\alpha - \langle S_i^\alpha\rangle\right). \tag{2}$$

Here, W is analogous to the position weight matrices that appear in models of transcription factor binding (35–37). Without loss of generality, we can set $\sum_\alpha W_i^\alpha = 0$ at every site i since $\sum_\alpha S_i^\alpha = 1$, see "Details of theoretical methods, (1)" section. If the linear terms are sufficient, then we can extract the elements of W by performing a least squares fit of the data with respect to Eq. (2), or, in the particular case when the sequences are random, by computing a correlation function:

$$W_i^\alpha = 4\langle (C_0 - \langle C_0\rangle)\left(S_i^\alpha - \langle S_i^\alpha\rangle\right)\rangle, \tag{3}$$

where the average is over all sequences, see "Details of theoretical methods, (3)" section for a derivation. Fig. 2 shows the estimate of W that results from computing this correlation for the Random library. The results are consistent with $W = 0$, suggesting that there is no linear term in the dependence of C_0 on the sequence. Indeed, when we estimate W from 90% of the data and predict C_0 for the remaining 10% using Eq. (2), the correlation coefficient between predictions and measurements is essentially zero, $r = 0.05 \pm 0.03$.

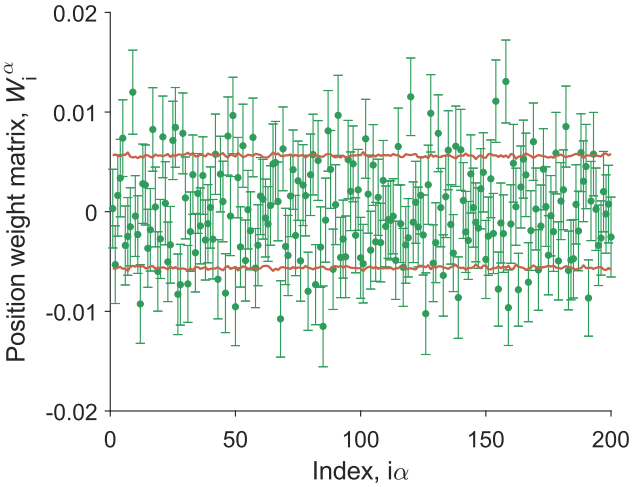

Figure 2 The matrix W_i^α for the Random library computed using Eq. (3). Points: mean and SD across random halves of the data. Lines: ±1SD across random halves of shuffled data. The index iα runs from 1 to 200, corresponding to 1A, 1T, 1C, 1G,..., 50A, 50T, 50C, 50G.

A pairwise model

If Eq. (2) does not work, because the data are consistent with $W = 0$, the next simplest model is then a pairwise model, where the intrinsic cyclizability is a weighted sum of contributions from pairs of nucleotides:

$$C_0 = \langle C_0\rangle + \frac{1}{2}\sum_{i\neq j,\alpha\beta} J_{ij}^{\alpha\beta}(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle). \tag{4}$$

Again, without loss of generality, we can set $\sum_\alpha J_{ij}^{\alpha\beta} = \sum_\beta J_{ij}^{\alpha\beta} = 0$ since $\sum_\alpha S_i^\alpha = 1$, see "Details of theoretical methods, (2)" section. As with the linear model, we can extract the elements of J by performing a least squares fit of the data with respect to Eq. (4), or, in the case of Random library, directly recover the underlying interaction parameters by computing correlation functions over random sequences:

$$J_{ij}^{\alpha\beta} = 16\langle (C_0 - \langle C_0\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle, \tag{5}$$

see "Details of theoretical methods, (4)" section for a derivation. For ease of presentation, we further combine the indices $(i, \alpha) \to i\alpha$ and $(j, \beta) \to j\beta$, so that $J_{ij}^{\alpha\beta} \to J_{i\alpha,j\beta}$. Figure 3 shows the estimates of $J_{i\alpha,j\beta}$ from computing the above correlation for the Random library. Compared to the J matrix of the shuffled data in Fig. S1, the J matrix in Fig. 3 shows clear signals near the diagonal line, highlighting the importance of nearest neighbor interactions. However, when we estimate J from 90% of the data and predict C_0 for the remaining 10% using Eq. (4), the correlation coefficient between predictions and measurements is $r = 0.45 \pm 0.02$, which is not high.

Physical constraints on the J matrix

At first glance, it seems that the pairwise terms are not sufficient to capture the sequence dependence of intrinsic cyclizability, as indicated by the low correlation between the predictions and measurements of the Random library. Therefore, it seems necessary to include higher-order terms in the model. However, a close inspection reveals that there are 9,800 independent elements in the J matrix, comparable to the total number of sequences (12,404) in the Random library. The limited number of sequences can lead to a low signal-to-noise ratio in the inferred J matrix, reducing its predictive power. To improve the performance of the pairwise model, we apply physical constraints to the J matrix to reduce the number of variables.

First, $J_{ij}^{\alpha\beta}$ is a coupling parameter between the bases at positions i and j, we expect it to depend on separation but not on absolute position, leading to translational invariance:

$$J_{ij}^{\alpha\beta} = J_{i'j'}^{\alpha\beta}, \quad \text{if } j - i = j' - i'. \tag{6}$$

Further, the double helical structure of DNA suggests that a sequence and its reverse complement should have the same cyclizability, leading to reverse-complement invariance:

$$J_{ij}^{\alpha\beta} = J_{N-j+1,N-i+1}^{\beta'\alpha'} = J_{ij}^{\beta'\alpha'}, \tag{7}$$

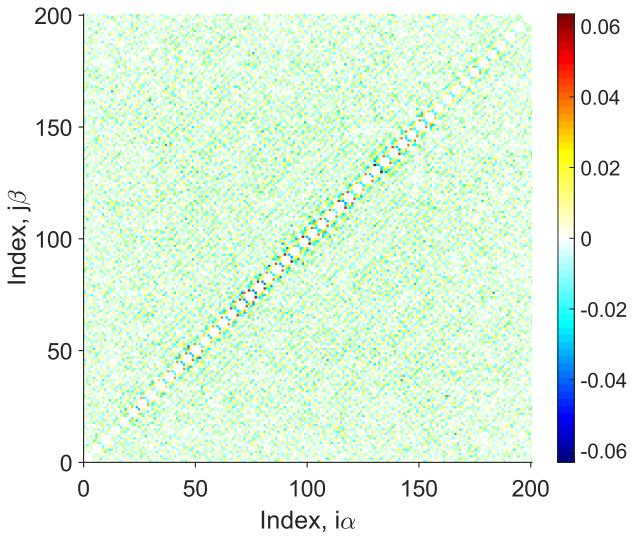

Figure 3 The interaction matrix $J_{ij}^{\alpha\beta}$ of the pairwise model for the Random library computed from Eq. (5). The indices $i\alpha$ and $j\beta$ follow the order 1A, 1T, ..., 50C, and 50G.

where α' and β' denote the complements of α and β (ie A $\leftrightarrow$ T and C $\leftrightarrow$ G), and translational invariance is used in deriving the last step. We note that by imposing translational and reverse-complement invariances, the number of independent elements in the J matrix is reduced from 9,800 to 294.

We impose translational invariance by replacing each matrix element of J in Fig. 3 with the average of all elements at the same separation of $j - i$,

$$J_{i,j>i}^{\alpha\beta} \to \frac{1}{N-j+i} \sum_{k=1}^{N-j+i} J_{k,k+j-i}^{\alpha\beta}. \tag{8}$$

We impose reverse-complement invariance by

$$J_{ij}^{\alpha\beta} \to \frac{1}{2}(J_{ij}^{\alpha\beta} + J_{ij}^{\beta'\alpha'}). \tag{9}$$

Figure 4a shows the J matrix for the Random library after imposing the symmetry conditions. A positive (negative) component $J_{ij}^{\alpha\beta}$ indicates that the nucleotide pair (α, β) separated by a distance $j - i$ promotes (suppresses) cyclization. For example, the dinucleotides TA and GC increase cyclizability, whereas TC and GA decrease it, since J_{12}^{TA} and J_{12}^{GC} are positive and J_{12}^{TC} and J_{12}^{GA} are negative (Fig. 4a, inset). The "symmetrized" J not only shows

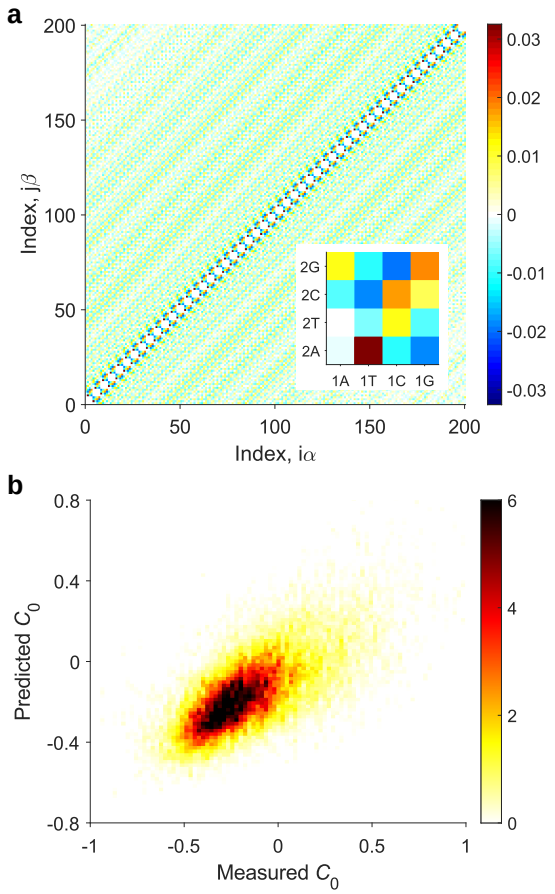

Figure 4 The pairwise model with translational and reverse-complement invariances for the sequence-dependence of intrinsic cyclizability. a) The interaction matrix $J_{ij}^{\alpha\beta}$ for the Random library after imposing the translational and reverse-complement invariances. Inset: nearest-neighbor interaction parameters $J_{ij}^{\alpha\beta}$ for $j - i = 1$. b) Joint probability distribution of predicted and measured cyclizability C_0 across the ensemble of sequences, Pearson correlation coefficient $r = 0.68 \pm 0.02$.

clear signals near the diagonal, suggesting strong nearest-neighbor interactions but also displays a set of stripes separated at half-helical (∼5 bp) and helical (∼10 bp) period of DNA, suggesting a role of longer-ranged interactions collectively contribute to DNA cyclizability.

Imposing translational and reverse-complement invariances raises the signal-to-noise ratio of the inferred J matrix, and consequently raises the predictive performance of the model. Predictions versus measurements of intrinsic cyclizability are shown in Fig. 4b as a joint density plot, which are obtained from multiple random 90/10 splits of the Random library into training and testing data. The correlation between predictions and measurements is now $r = 0.68 \pm 0.02$.

Sequence features that control DNA cyclizability

The predictive power of the model would improve further if we incorporate the three-base interaction terms in Eq. (1). Yet, to keep the model minimal, we defer these extensions and instead search for sequence features that control DNA cyclizability within the pairwise model. To identify sequence features, we apply eigenvalue decomposition (or principal component analysis) to the J matrix:

$$J_{ij}^{\alpha\beta} = \sum_n \lambda_n w_i^\alpha(n) w_j^\beta(n), \tag{10}$$

where n is the mode index and the eigenvectors are orthonormal,

$$\sum_{i,\alpha} w_i^\alpha(n) w_i^\alpha(m) = \delta_{nm}. \tag{11}$$

We show in Fig. 5a the spectrum of eigenvalues $\{\lambda_n\}$ in rank order and compare with data that have been shuffled to break any correlations between sequence and cyclizability. We first notice that, in both the real and shuffled data, there are some true zero eigenvalues. These arise because we have $\sum_\alpha S_i^\alpha = 1$ at each site i, by definition. In the shuffled data, we see a spreading of the eigenvalues, which arises because the J matrix is estimated from a finite sample (38, 39). Last and most importantly, in the real data, the eigenvalues stand out from the shuffled background with high signal-to-noise ratios at both large positive and negative values, especially at the two largest eigenvalues. We show in Fig. 5b the first and last two eigenvectors, which pick out the most and least cyclizable modes of sequence variation. The first two eigenvectors come as a quadrature pair, showing an approximate ten-base periodicity that aligns with the pitch of the double helix DNA. The last two modes, with their eigenvalues less clearly distinguished from the background noise (Fig. 5a), are less clearly periodic and exhibit boundary effects.

We can decompose the sequence variations into modes defined by the eigenvectors:

$$S_i^\alpha = \langle S_i^\alpha \rangle + \sum_n f_n w_i^\alpha(n), \tag{12}$$

where

$$f_n = \sum_{i,\alpha} w_i^\alpha(n)\left(S_i^\alpha - \langle S_i^\alpha \rangle\right). \tag{13}$$

Equation (4) is then simplified to

$$C_0 = \langle C_0 \rangle + \frac{1}{2}\sum_n \lambda_n f_n^2, \tag{14}$$

where the sum is over all the modes. In Fig. 5c, we show the dependence of the cyclizability C_0 on f_n at the extremes of the spectrum. To avoid overfitting, we estimate $J_{ij}^{\alpha\beta}$ and subsequently the eigenvectors $w_i^\alpha(n)$ from random half of the sequences, and then probe C_0 versus f_n in the other half. The mean behavior is quadratic for each mode, consistent with the prediction of Eq. (14), supporting the validity of the model in Eq. (4).

Equation (14) further enables the separation of contributions from individual modes to cyclizability. Including only the first two modes, ie sum over $n = 1, 2$ in Eq. (14), results in $r = 0.60 \pm 0.02$ in the predictive performance, suggesting that these two modes make the largest contribution, but including all modes provides better predictions, Fig. S2.

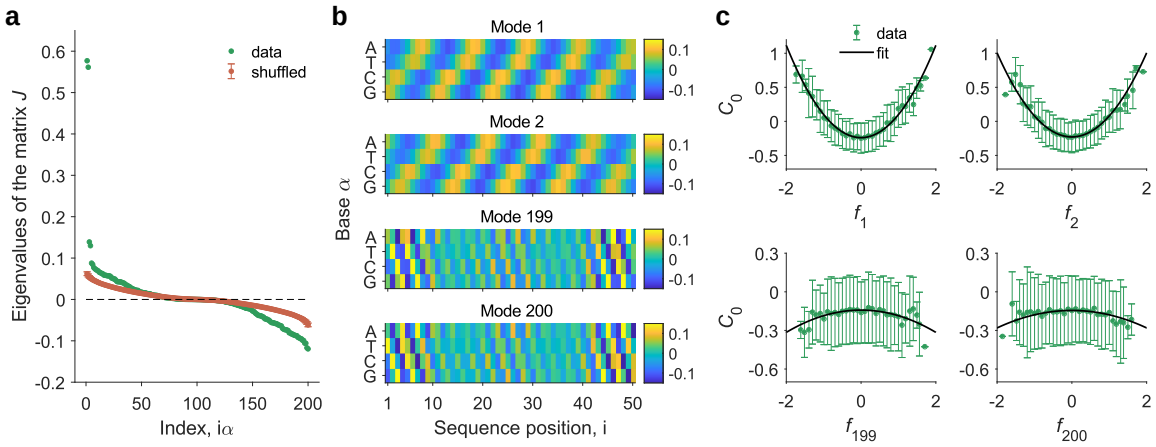

Figure 5 The pairwise interaction model captures sequence features for cyclizability. a) Eigenvalues $\{\lambda_n\}$ of the interaction matrix J from the Random library, ranked in descending order, compared with results from shuffled data of the Random library. Points and error bars in the shuffled data are the means and SD calculated across multiple random 50/50 splittings of the data. b) The first two eigenvectors, $w_i^\alpha(1)$ and $w_i^\alpha(2)$, and the last two eigenvectors, $w_i^\alpha(199)$ and $w_i^\alpha(200)$, of the matrix J. Scale is set by normalization, Eq. (11). c) Cyclizability as a function of sequence projection in the feature space. Error bars are SD of the test dataset. Lines are quadratic fits as expected from Eq. (14).

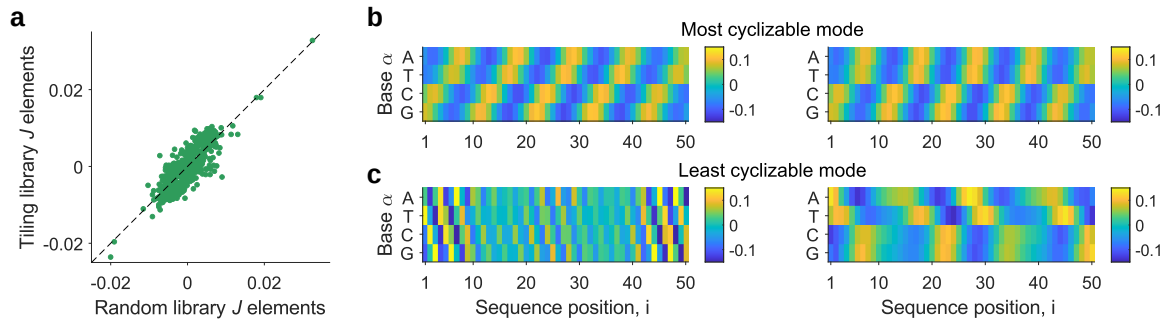

Figure 6 Comparison of Random and Tiling libraries. a) Elements of the J matrix of the Random library versus the Tiling library, with Pearson's $r = 0.90$. b) The most and least cyclizable modes of the Random library (left) and the Tiling library (right).

Comparison of Random and Tiling libraries

Above, we adopted a statistical-mechanics approach to analyze cyclizability of sequences in the Random library. Yet biologically relevant sequences are not random. For example, the yeast genome has a base composition of approximately 62% A/T and 38% C/G (40), which deviates significantly from random sequences. To test if our findings from the Random library extend to biological sequences, we analyze a second available dataset, the Tiling library, which comprises sequences from the genome of *S. cerevisiae* (27).

In the analysis of the Random library, we utilized the statistical properties of random sequences to compute the matrices W and J directly from data, using Eqs. (3) and (5), respectively. These two equations however no longer hold for the Tiling library. Nevertheless, as briefly mentioned before, it is possible to extract the matrices with a least-squares fit directly applied to Eqs. (2) and (4). To validate the least-squares approach, we first apply it to the Random library. The resulting W and J matrices match almost exactly those obtained from Eqs. (3) and (5), with Pearson correlation coefficients $r = 0.99$ and 0.98, respectively (Fig. S3). We then apply the least-squares approach to the Tiling library. In Fig. S4, we show the extracted W which again is consistent with $W = 0$. The correlation coefficient between predictions and measurements of the linear model is $r = 0.07 \pm 0.01$, slightly higher than that of the Random library. In Fig. S5, we show the extracted J matrix, which shares many similarities with the one from the Random library. The correlation coefficient between predictions and measurements of the pairwise model is $r = 0.68 \pm 0.01$, comparable to that of the Random library. For a direct comparison between the Tiling and Random libraries, the matrix elements of J for the two libraries were plotted against each other in Fig. 6a, which are highly correlated with $r = 0.90$. The most and least cyclizable modes of the two libraries were compared in Fig. 6b. While the most cyclizable modes are consistent between the two datasets, the least cyclizable modes are clearly different.

We can further construct the most and least cyclizable sequences from the eigenvectors of the two libraries. Because the eigenvectors are orthonormal, increasing the projection of a sequence onto one eigenvector necessarily decreases its projection onto others, the highest and lowest values of C_0 are thus predicted to occur in sequences that have maximal squared projection onto the first and last modes, respectively. Mathematically, this means we look for sequences $\{S_i^\alpha\}$ that maximize f_1^2 and f_{200}^2 in Eq. (13). We note that because λ_1 and λ_2 (similarly, λ_{199} and λ_{200}) come almost as a degenerate pair, sequences that maximize f_2^2 and f_{199}^2 will have comparable cyclizabilities as the most and least cyclizable ones. The predicted most and least cyclizable sequences for the two libraries are listed in Table 1. The most cyclizable sequences are identical for both libraries, which consists of 5–6 bp tracts of TA-rich segments (eg TTAAA) followed by 5–6 bp tracts of GC-rich segments (eg GGGCCC) periodically. In contrast, the least cyclizable sequence for the Random library consists of short ∼3 bp segments with nucleotides in the order of ATCG, while the sequence for the Tiling library exhibits a periodic structure similar to the most cyclizable sequence but with longer 8–10 bp tracts of AT-rich segments where As appear in front of Ts (eg AAAAATTTTT).

Table 1 Predicted DNA sequences with highest and lowest intrinsic cyclizabilities.

Most cyclizable sequence (Random and Tiling libraries)
TAAAGGCCCTTTAAGGGCCCTTAAAGGGCCTTTAAGGGCCCTTTAAGGGC
Least cyclizable sequence (Random library)
GATGATCGTCGTCGACGATCATCGTCGTCGACGACGATGATCGTCGACGA
Least cyclizable sequence (Tiling library)
ATTTGGGCCCAAAAATTTTTGGCCAAAATTTTGGGCCAAAAAATTTGGCG

Molecular dynamics simulations of most and least cyclizable sequences

To support our predictions of the most and least cyclizable sequences in Table 1, we performed all-atom molecular dynamics simulations of these sequences in a 1M NaCl solution at constant temperature (300 K) with periodic boundary conditions. The simulations were carried out using the NAMD software (41) with the CHARMM36 force field (42), see "Details of molecular dynamics simulations" section for details of the simulation setup. We show in Fig. 7a–c representative snapshots of DNA for each sequence (taken at 72 ns). Remarkably, the most cyclizable sequence curves into a semicircle (Fig. 7a), whereas the least cyclizable sequences for both libraries remain straight (Fig. 7b and c). To quantify the bending differences, we extracted the helical axis of the DNA molecule using the Curves+ program (43) and computed the end-to-end distance at each frame along the trajectories. While all sequences were initialized in an extended

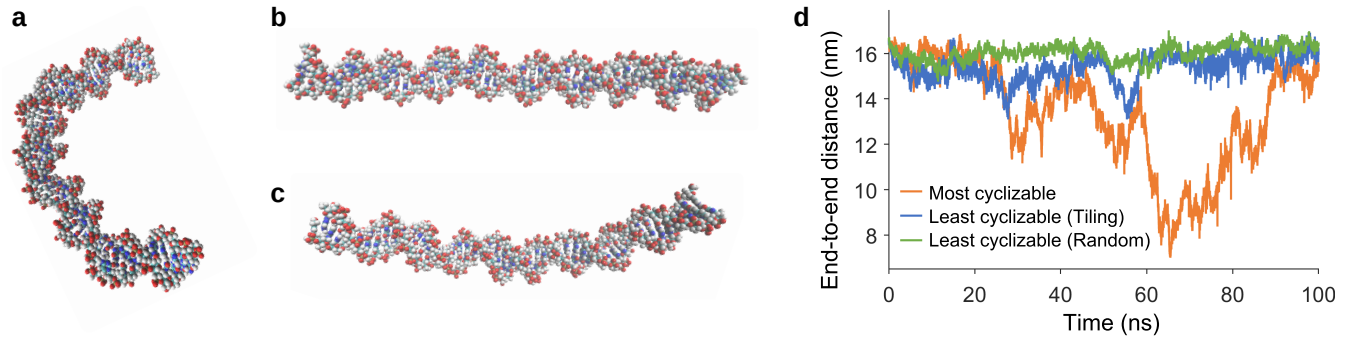

Figure 7 All-atom simulations of DNA molecules with sequences listed in Table 1 support our model predictions. a) Representative snapshot of the most cyclizable sequence for both the Random and Tiling libraries. b) Representative snapshot of the least cyclizable sequence for the Random library. c) Representative snapshot of the least cyclizable sequence for the Tiling library. d) End-to-end distance versus time plot for the corresponding DNA molecules quantified from the all-atom trajectories.

state, the end-to-end distance of the most cyclizable sequence decreases noticeably after around 20 ns, compared to those of the two least cyclizable sequences, which remain relatively extended throughout the run (Fig. 7d). Corresponding videos can be found in the Zenodo data repository (DOI: 10.5281/zenodo.18698120). Observations from the end-to-end distance plot and simulation videos further suggest that the least cyclizable sequence of the Tiling library is slightly more dynamically flexible than that of the Random library. Overall, these results strongly support the predictions of our statistical model.

Discussion

The mechanical properties of DNA play an important role in many biological processes. Recent experiments have revealed that the mechanics of short DNA molecules depends strongly on their sequences. However, a systematic characterization of the sequence features that control DNA mechanics has been lacking. In this work, we developed a statistical mechanics approach to analyze high-throughput measurements of DNA cyclizability of randomly generated and biologically derived sequences. We found that a minimal model with pairwise interactions between nucleotides is sufficient to account for the sequence dependence of cyclizability, leading to the prediction of the most and least cyclizable sequences. We further validated these findings through all-atom molecular dynamics simulations of predicted sequences.

How do our findings compare to those in the literature? Early studies on the sequence dependence of DNA mechanics primarily focused on the influence of dinucleotide pairs due to limited data (44, 45). Recent development in high-throughput methods have enabled investigations that extend nucleotide interactions to longer length scales. In many respects, our findings align with the consensus knowledge in the field (2, 46). Specifically, our predicted most cyclizable sequence is consistent with findings from recent SELEX experiments, where the selected highly loopable sequences are enriched with alternating short segments of A/T and C/G separated by half a helical pitch of DNA (47). Mechanistically, both dynamic flexibility and static bending contribute to the cyclizability of a sequence. Our all-atom simulations suggest that the increased cyclizability of the most cyclizable sequence is mainly due to constructive bending toward one direction, driven by intrinsic curvatures encoded in the sequence. This is analogous to the previously reported globally curved structures formed by sequences with in-phased A-tract repeats (48–50). Our predicted least cyclizable sequence from the Random library is enriched in CpG steps, which was reported to increase the rigidity of the sequence (51, 52). The least cyclizable sequence from the Tiling library is enriched in out-phased A-tracts, also known to be very rigid (18). While our findings are largely consistent with previous studies, our systematic framework allows us to extract sequence features more comprehensively, providing a deeper and more quantitative understanding of the determinants of DNA cyclizability.

One limitation of the pairwise model is that its predictive power is lower than that of recent deep neural network models. The Pearson correlation coefficient between predictions and measurements is $r \approx 0.7$ for the pairwise model, compared with $r \approx 0.8$–0.9 for deep neural networks (28–31) and notably $r = 0.96$ in the most recent development (9). This indicates that higher-order interaction terms are necessary to fully capture DNA cyclizability. In a separate (unpublished) study, we incorporated three-body interaction terms in Eq. (1) and improved the predictive power to $r \approx 0.92$, validating that cyclizability is a highly collective property. Because sequence features inferred from the pairwise model may inherit a similar level of inaccuracy, we checked their robustness by comparing extreme sequences predicted by our model to those inferred from a deep-learning approach (9) (Table S1). Despite the neural network retaining higher-order correlations, both approaches identify a common coarse-grained organization: high-cyclizability sequences are enriched for alternating short (5–6 bp) AT-rich tracts and GC-rich tracts, whereas low-cyclizability sequences contain longer (8–10 bp) AT-rich tracts with comparatively shorter GC-rich interruptions. These shared motif-level trends suggest that such coarse-grained features inferred from the pairwise model are robust. However, there are differences in finer-scale nucleotide ordering within tracts. For example, our most cyclizable sequence has T preceding A (eg TTAAA), whereas the deep-learning approach predicts A preceding T (eg AAAAAT) or pure A/T tracts (eg AAAAAA). Moreover, the GC-rich segments in the low-cyclizability sequences from the deep-learning approach are mostly CG dinucleotide, whereas our sequence contains GC-rich segments with G preceding C in most cases (eg GGGCC). These finer-scale differences motivate feature

interpretation in higher-order models. Nevertheless, the pairwise model stands out for its simplicity and physical interpretability. It captures leading-order sequence determinants, and offers a starting point for systematic higher-order extensions.

Moving beyond sequence-dependent DNA mechanics, our method establishes a general framework to address a broad class of problems, where the input lives in a high-dimensional space and the output represents a functional property of the system. With the continued advancement of high-throughput technologies, we anticipate the method developed here to have broad applications. Representative problems in this class include the search for relevant sequence features of DNA, RNA, and proteins that determine their structures, interactions, and functions, as well as the search for relevant stimulus features that govern the responses of sensory neurons.

Methods

Details of theoretical methods

1. Justification of setting $\sum_\alpha W_i^\alpha = 0$. Assume we find the matrix $\tilde{W}_i^\alpha$ for the linear model, however $\sum_\alpha \tilde{W}_i^\alpha = c_i \neq 0$. Let $W_i^\alpha = \tilde{W}_i^\alpha - c_i/4$, the second term on the right of Eq. (2) becomes:

$$\begin{aligned}\sum_{i,\alpha} \tilde{W}_i^\alpha \left(S_i^\alpha - \langle S_i^\alpha\rangle\right) &= \sum_{i,\alpha}\left(W_i^\alpha + \frac{c_i}{4}\right)\left(S_i^\alpha - \langle S_i^\alpha\rangle\right)\\ &= \sum_{i,\alpha} W_i^\alpha \left(S_i^\alpha - \langle S_i^\alpha\rangle\right),\end{aligned} \tag{15}$$

where the c_i term is gone because $\sum_\alpha S_i^\alpha = \sum_\alpha \langle S_i^\alpha\rangle = 1$ for any given i. Therefore, W_i^α is also a solution to the linear model and we have $\sum_\alpha W_i^\alpha = 0$.

2. Justification of setting $\sum_\alpha J_{ij}^{\alpha\beta} = \sum_\beta J_{ij}^{\alpha\beta} = 0$. Similar to the linear model, assume we find the matrix $\tilde{J}_{ij}^{\alpha\beta}$ for the pairwise model, however $\sum_\alpha \tilde{J}_{ij}^{\alpha\beta} = c_{ij}^\beta \neq 0$ and $\sum_\beta \tilde{J}_{ij}^{\alpha\beta} = d_{ij}^\alpha \neq 0$. Let $J_{ij}^{\alpha\beta} = \tilde{J}_{ij}^{\alpha\beta} - \tilde{c}_{ij}^\beta - \tilde{d}_{ij}^\alpha$, where $\tilde{c}_{ij}^\beta$ and $\tilde{d}_{ij}^\alpha$ are the solutions of $\sum_\alpha (\tilde{c}_{ij}^\beta + \tilde{d}_{ij}^\alpha) = c_{ij}^\beta$ and $\sum_\beta (\tilde{c}_{ij}^\beta + \tilde{d}_{ij}^\alpha) = d_{ij}^\alpha$, the second term on the right of Eq. (4) becomes:

$$\begin{aligned}&\frac{1}{2}\sum_{i\neq j,\alpha\beta} \tilde{J}_{ij}^{\alpha\beta}(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\\ &= \frac{1}{2}\sum_{i\neq j,\alpha\beta}(J_{ij}^{\alpha\beta} + \tilde{c}_{ij}^\beta + \tilde{d}_{ij}^\alpha)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\\ &= \frac{1}{2}\sum_{i\neq j,\alpha\beta} J_{ij}^{\alpha\beta}(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle),\end{aligned} \tag{16}$$

where again the $\tilde{c}$ and $\tilde{d}$ terms are gone because $\sum_\alpha S_i^\alpha = \sum_\alpha \langle S_i^\alpha\rangle = 1$. Therefore, $J_{ij}^{\alpha\beta}$ is also a solution to the pairwise model and we have $\sum_\alpha J_{ij}^{\alpha\beta} = \sum_\beta J_{ij}^{\alpha\beta} = 0$.

3. Derivation of Eq. (3) for the Random library. The sequences in the Random library were chosen from a uniform distribution, therefore $\langle S_i^\alpha\rangle = 1/4$. For $i \neq j$, nucleotides at different positions are independently chosen, so $\langle S_i^\alpha S_j^\beta\rangle = \langle S_i^\alpha\rangle\langle S_j^\beta\rangle = 1/16$. For $i = j$, when $\alpha \neq \beta$, S_i^α and S_i^β cannot both be 1, so $\langle S_i^\alpha S_j^\beta\rangle = 0$; and when $\alpha = \beta$, $S_i^\alpha S_i^\beta = (S_i^\alpha)^2 = S_i^\alpha$ for $S_i^\alpha \in \{0, 1\}$, so $\langle S_i^\alpha S_j^\beta\rangle = 1/4$. Together, these results can be summarized in a compact form:

$$\begin{aligned}&\langle (S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle\\ &\quad = \langle S_i^\alpha S_j^\beta\rangle - \langle S_i^\alpha\rangle\langle S_j^\beta\rangle = \frac{1}{4}\delta_{ij}\delta^{\alpha\beta} - \frac{1}{16}\delta_{ij}.\end{aligned} \tag{17}$$

The correlation function between C_0 in Eq. (2) and S_i^α is then

$$\begin{aligned}&\langle (C_0 - \langle C_0\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)\rangle\\ &\quad = \sum_{j,\beta} W_j^\beta \langle (S_j^\beta - \langle S_j^\beta\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)\rangle = \frac{W_i^\alpha}{4},\end{aligned} \tag{18}$$

which is Eq. (3).

4. Derivation of Eq. (5) for the Random library. Similar to the above derivation, the correlation function between C_0 in Eq. (4), S_i^α, and S_j^β are

$$\begin{aligned}\langle (C_0 - \langle C_0\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle &= \frac{1}{2}\sum_{k\neq l,\gamma\eta} J_{kl}^{\gamma\eta}\\ \times \langle (S_k^\gamma - \langle S_k^\gamma\rangle)(S_l^\eta - \langle S_l^\eta\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle.\end{aligned} \tag{19}$$

The second line in the above equation is nonzero only if $k = i$ and $l = j$ (case I) or $k = j$ and $l = i$ (case II).

For case I, we have

$$\begin{aligned}&\langle (S_k^\gamma - \langle S_k^\gamma\rangle)(S_l^\eta - \langle S_l^\eta\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle\\ &\quad = \langle (S_i^\gamma - \langle S_i^\gamma\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)\rangle\langle (S_j^\eta - \langle S_j^\eta\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle\\ &\quad = \frac{1}{16}(\delta^{\gamma\alpha} - \frac{1}{4})(\delta^{\eta\beta} - \frac{1}{4}),\end{aligned} \tag{20}$$

where Eq. (17) was used to derive the last line. Similarly, for case II, we have

$$\begin{aligned}&\langle (S_k^\gamma - \langle S_k^\gamma\rangle)(S_l^\eta - \langle S_l^\eta\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle\\ &\quad = \frac{1}{16}(\delta^{\gamma\beta} - \frac{1}{4})(\delta^{\eta\alpha} - \frac{1}{4}).\end{aligned} \tag{21}$$

Substituting results in Eqs. (20) and (21) into Eq. (19), we have

$$\begin{aligned}&\langle (C_0 - \langle C_0\rangle)(S_i^\alpha - \langle S_i^\alpha\rangle)(S_j^\beta - \langle S_j^\beta\rangle)\rangle\\ &\quad = \sum_{\gamma\eta}\frac{J_{ij}^{\gamma\eta}}{32}(\delta^{\gamma\alpha} - \frac{1}{4})(\delta^{\eta\beta} - \frac{1}{4}) + \frac{J_{ji}^{\gamma\eta}}{32}(\delta^{\gamma\beta} - \frac{1}{4})(\delta^{\eta\alpha} - \frac{1}{4})\\ &\quad = \frac{1}{32}J_{ij}^{\alpha\beta} + \frac{1}{32}J_{ji}^{\beta\alpha} = \frac{1}{16}J_{ij}^{\alpha\beta},\end{aligned} \tag{22}$$

which is Eq. (5).

Details of molecular dynamics simulations

All-atom molecular dynamics simulations of 50-bp DNA sequences in Table 1 were conducted following the outlined steps below, which were adapted from a previously developed pipeline for quantifying the flexibility of DNA molecules (21).

1. Construction of initial simulation files. The initial structure of the dsDNA molecule was built from the nucleotide sequence using the web 3DNA 2.0 server (53), which generated the starting structure in the B-DNA form with the double

helix arranged in a straight line. The resulting PDB file was loaded into the VMD program (54), where the DNA was solvated with water molecules. To ensure that the simulation box dimensions were sufficiently large to accommodate the DNA, we used a 220 Å × 80 Å × 80 Å solvation box for the 50-bp DNA (contour length ∼170 Å), which allowed for padding of at least 24.5 Å between the DNA and the box edges. We next added 1 M concentration of NaCl to neutralize the system (1 M NaCl was chosen instead of $MgCl_2$ to mimic the experimental salt conditions). To prevent the DNA ends from fraying during simulation, we introduced two additional bonds between the ends of the two DNA strands, following the standard protocol (21). The PSF and PDB files containing the topological and structural information of all molecules inside the simulation box were then exported from VMD.

2. Energy minimization and equilibration. The system with the DNA, water, and ions was simulated in NAMD (41) using the CHARMM36 force field (42), TIP3P water model (55), and ion parameters (56). Electrostatic interactions were computed using the particle-mesh Ewald method with a 1.5 Å grid spacing (57). Simulations were performed with a 2-fs integration time step and periodic boundary conditions. During the initial energy minimization and equilibration steps, we restrained the DNA molecule to its original coordinates, which allows the surrounding water and ions to equilibrate without disrupting the double-stranded DNA due to large forces involved in these processes. Energy minimization was conducted for 10,000 fs to resolve steric clashes and bad contacts. This was followed by a 1-ns constant-volume simulation and a 10-ns constant-pressure (1 bar) simulation both at 298 K to ensure the proper equilibration of the solvent and ions. The constant temperature was maintained via Langevin dynamics, and the constant pressure was achieved using a modified Nosé–Hoover method (58) with Langevin piston control (59), using a piston period of 100 fs and a decay time of 50 fs.
3. Final simulation run and trajectory recording. After energy minimization and equilibration, we conducted the final constant-pressure simulation with unrestrained DNA for ∼100 ns. Trajectories of all molecules were saved every 9.6 ps for a total of ∼10,420 recordings. Snapshots and videos of the DNA molecules were built in VMD using these trajectories. An independent run with a different random seed was performed for each sequence to confirm the reproducibility of simulation results. Representative snapshots and the end-to-end distance-versus-time plot of DNA are shown in Fig. 7.

Acknowledgments

We are grateful to Ana Damjanovic, Albert Lau, and Mankun Sang for insightful discussions and constructive feedback on the manuscript.

Supplementary material

Supplementary material is available at *PNAS Nexus* online.

Competing interest

The authors declare no competing interests.

Funding

M.G. and Y.Z. were supported by a startup fund at Johns Hopkins University and by NIH award R35GM162296. Y.Z. acknowledges support from Sloan Foundation through a Sloan Research Fellowship (FG-2025-25076). W.B. and Y.Z. were supported in part by the National Science Foundation through the Center for the Physics of Biological Function (PHY-1734030) and by NSF grant PHY-1607612. A.B. was a Simons Foundation Fellow of the Life Sciences Research Foundation, and T.H. is an Investigator at the Howard Hughes Medical Institute.

Preprints

This manuscript was posted on a preprint: https://doi.org/10.1101/2025.01.02.631081.

Data availability

The analysis code supporting the findings of this study is available on GitHub (https://github.com/Zhang-Lab-JHU/DNA-cyclizability-code), which includes scripts for the statistical mechanics model and molecular dynamics simulations, along with the corresponding input cyclizability data. All-atom simulation outputs are deposited in Zenodo (DOI: 10.5281/zenodo.18698120), which includes DNA trajectories stripped of solvent and ions, and the associated helical axis trajectories.

References

1. Basu A, Bobrovnikov DG, Ha T. 2021. Dna mechanics and its biological impact. *J Mol Biol*. 433:166861.
2. Marin-Gonzalez A, Vilhena J, Perez R, Moreno-Herrero F. 2021. A molecular view of DNA flexibility. *Q Rev Biophys*. 54: e8.
3. Segal E, *et al.* 2006. A genomic code for nucleosome positioning. *Nature*. 442:772.
4. Kaplan N, *et al.* 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. 458:362.
5. Parmar JJ, Padinhateeri R. 2020. Nucleosome positioning and chromatin organization. *Curr Opin Struct Biol*. 64:111.
6. Oberbeckmann E, Quililan K, Cramer P, Oudelaar AM. 2024. In vitro reconstitution of chromatin domains shows a role for nucleosome positioning in 3D genome organization. *Nat Genet*. 56:483.
7. Garcia HG, *et al.* 2007. Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers*. 85:115.
8. Afek A, *et al.* 2020. DNA mismatches reveal conformational penalties in protein–DNA recognition. *Nature*. 587:291.
9. Park J, Prokopchuk G, Popchock AR, Hao J, Liao T-W, Yan S, Hedman DJ, Larson JD, Walther B, Becker NA. 2024. Probing mechanical selection in diverse eukaryotic genomes through

accurate prediction of 3D DNA mechanics [preprint], bioRxiv, https://doi.org/10.1101/2024.12.22.629997.
10. Matthews K. 1992. Dna looping. *Microbiol Rev*. 56:123.
11. Mulligan PJ, Chen Y-J, Phillips R, Spakowitz AJ. 2015. Interplay of protein binding interactions, DNA mechanics, and entropy in DNA looping kinetics. *Biophys J*. 109:618.
12. Sharma M, Predeus AV, Mukherjee S, Feig M. 2013. Dna bending propensity in the presence of base mismatches: implications for DNA repair. *J Phys Chem B*. 117:6194.
13. Marko JF, Siggia ED. 1995. Stretching DNA. *Macromolecules*. 28:8759.
14. Bustamante C, Marko JF, Siggia ED, Smith S. 1994. Entropic elasticity of λ-phage DNA. *Science*. 265:1599.
15. Wiggins PA, *et al.* 2006. High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat Nanotechnol*. 1:137.
16. Cloutier TE, Widom J. 2004. Spontaneous sharp bending of double-stranded DNA. *Mol Cell*. 14:355.
17. Moreno-Herrero F, Seidel R, Johnson SM, Fire A, Dekker NH. 2006. Structural analysis of hyperperiodic DNA from *Caenorhabditis elegans*. *Nucleic Acids Res*. 34:3057.
18. Vafabakhsh R, Ha T. 2012. Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. *Science*. 337:1097.
19. Ngo TT, Zhang Q, Zhou R, Yodh JG, Ha T. 2015. Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell*. 160:1135.
20. Pérez A, *et al.* 2012. Impact of methylation on the physical properties of DNA. *Biophys J*. 102:2140.
21. Ngo TT, *et al.* 2016. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat Commun*. 7:10813.
22. Jacobson H, Stockmayer WH. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J Chem Phys*. 18:1600.
23. Shimada J, Yamakawa H. 1984. Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromolecules*. 17:689.
24. Levene SD, Giovan SM, Hanke A, Shoura MJ. 2013. The thermodynamics of DNA loop formation, from *J* to *Z*. *Biochem Soc Trans*. 41:513.
25. Hagerman PJ. 1986. Sequence-directed curvature of DNA. *Nature*. 321:449.
26. Geggier S, Vologodskii A. 2010. Sequence dependence of DNA bending rigidity. *Proc Natl Acad Sci U S A*. 107:15421.
27. Basu A, *et al.* 2021b. Measuring DNA mechanics on the genome scale. *Nature*. 589:462.
28. Li K, Carroll M, Vafabakhsh R, Wang XA, Wang J-P. 2022. DNAcycp: a deep learning tool for DNA cyclizability prediction. *Nucleic Acids Res*. 50:3142.
29. Khan SR, Sakib S, Rahman MS, Samee MAH. 2023. Deepbend: an interpretable model of DNA bendability. *iScience*. 26:00.
30. Back G, Walther D. 2023. Predictions of DNA mechanical properties at a genomic scale reveal potentially new functional roles of DNA flexibility. *NAR Genom Bioinform*. 5:lqad097.
31. Jiang W-J, *et al.* 2023. Assessing base-resolution DNA mechanics on the genome scale. *Nucleic Acids Res*. 51:9552.
32. Schneidman E, Berry MJ, Segev R, Bialek W. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*. 440:1007.
33. Figliuzzi M, Barrat-Charlaix P, Weigt M. 2018. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol Biol Evol*. 35:1018.
34. Meshulam L, Bialek W. 2025. Statistical mechanics for networks of real neurons. *Rev Mod Phys*. 97:045002.
35. Berg OG, von Hippel PH. 1987. Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J Mol Biol*. 193:723.
36. Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics*. 16:16.
37. Kinney JB, Tkačik G, Callan Jr CG. 2007. Precise physical models of protein–DNA interaction from high-throughput data. *Proc Natl Acad Sci U S A*. 104:501.
38. Potters M, Bouchaud J-P. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.
39. We have verified that the maximum and minimum eigenvalues in the shuffled data vary as $1/\sqrt{M}$, where M is the number of sequences in our sample, as expected from random matrix theory.
40. Meyer SA, Phaff H. 1969. Deoxyribonucleic acid base composition in yeasts. *J Bacteriol*. 97:52.
41. Phillips JC, *et al.* 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys*. 153:044130.
42. Hart K, *et al.* 2012. Optimization of the CHARMM additive force field for DNA: improved treatment of the BI/BII conformational equilibrium. *J Chem Theory Comput*. 8:348.
43. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K. 2009. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res*. 37:5917.
44. Olson WK, Gorin AA, Lu X-J, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc Natl Acad Sci U S A*. 95:11163.
45. Pasi M, *et al.* 2014. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res*. 42:12272.
46. Basu A, *et al.* 2022. Deciphering the mechanical code of the genome and epigenome. *Nat Struct Mol Biol*. 29:1178.
47. Rosanio G, Widom J, Uhlenbeck OC. 2015. In vitro selection of DNAs with an increased propensity to form small circles. *Biopolymers*. 103:303.
48. Crothers DM, Haran TE, Nadeau JG. 1990. Intrinsically bent DNA. *J Biol Chem*. 265:7093.
49. Marin-Gonzalez A, *et al.* 2020. Understanding the paradoxical mechanical response of in-phase A-tracts at different force regimes. *Nucleic Acids Res*. 48:5024.
50. Stefl R, Wu H, Ravindranathan S, Sklenář V, Feigon J. 2004. DNA A-tract bending in three dimensions: solving the dA_4T_4 vs. dT_4A_4 conundrum. *Proc Natl Acad Sci U S A*. 101:1177.
51. Pongor CI, Bianco P, Ferenczy G, Kellermayer R, Kellermayer M. 2017. Optical trapping nanometry of hypermethylated CPG-island DNA. *Biophys J*. 112:512.
52. Shon MJ, Rah S-H, Yoon T-Y. 2019. Submicrometer elasticity of double-stranded DNA revealed by precision force-extension measurements with magnetic tweezers. *Sci Adv*. 5:eaav1697.
53. Li S, Olson WK, Lu X-J. 2019. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res*. 47:W26.

54. Humphrey W, Dalke A, Schulten K. 1996. VMD—visual molecular dynamics. *J Mol Graph*. 14:33.
55. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 79:926.
56. Beglov D, Roux B. 1994. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *J Chem Phys*. 100:9050.
57. Darden T, York D, Pedersen L. 1993. Particle mesh Ewald: an $N \log(N)$ method for Ewald sums in large systems. *J Chem Phys*. 98:10089.
58. Martyna GJ, Tobias DJ, Klein ML. 1994. Constant pressure molecular dynamics algorithms. *J Chem Phys*. 101:4177.
59. Feller SE, Zhang Y, Pastor RW, Brooks BR. 1995. Constant pressure molecular dynamics simulation: the Langevin piston method. *J Chem Phys*. 103:4613.