

hexABC seeking the physical code of DNA

Received: 22 September 2025

Accepted: 4 June 2026

Cite this article as: Battistini, F., Wiczór, M., Hospital, A. *et al.* hexABC seeking the physical code of DNA. *Nat Commun* (2026). <https://doi.org/10.1038/s41467-026-74390-5>

Federica Battistini, Miłosz Wiczór, Adam Hospital, Marco Pasi, Juan Pablo Arcón, Israel Serrano-Chacón, Alba Sala, Subhamoy Deb, Agustín García-Doñate, Matthew Burman, Elliot W. Chan, Liwei Chang, Gabriela da Rosa, Jorge R. Espinosa, Gia Linh Hoang, Kazi A. Hossain, Michał Jurkowski, Romain Poupon, Rahul Sharma, Ran Sun, Thomas C. Bishop, Paolo Carloni, Thomas E. Cheatham III, Rosana Collepardo-Guevara, Jacek Czub, Pablo D. Dans, Sarah A. Harris, Charles Loughton, Rodrigo Galindo-Murillo, John H. Maddocks, Agnes Noy, Alberto Pérez, Daiva Petkevičiūtė-Gerlach & Modesto Orozco

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

hexABC seeking the physical code of DNA

Federica Battistini^{1,2}, Milosz Wieczor^{1&}, Adam Hospital^{1&}, Marco Pasi^{3&}, Juan Pablo Arcón^{1,4}, Israel Serrano-Chacón¹, Alba Sala^{1,4}, Subhamoy Deb¹, Agustín García¹, Matthew Burman⁵, Elliot W. Chan⁵, Liwei Chang⁶, Gabriela da Rosa⁷, Jorge R. Espinosa^{8,9}, Gia Linh Hoang¹⁰, Kazi A. Hossain¹¹, Michal Jurkowski¹¹, Romain Poupon³, Rahul Sharma¹², Ran Sun¹³, Thomas C. Bishop¹³, Paolo Carloni¹⁰, Thomas E. Cheatham III¹⁴, Rosana Collepardo-Guevara^{8,9}, Jacek Czub¹¹, Pablo D. Dans^{7,15}, Sarah A. Harris^{16,17}, Charles Laughton¹⁸, Rodrigo Galindo-Murillo¹⁹, John H. Maddocks¹², Agnes Noy⁵, Alberto Perez⁶, Daiva Petkevičiūtė-Gerlach²⁰, Modesto Orozco^{1,2*}.

1. Institute for Research in Biomedicine (IRB Barcelona), Baldori Reixac 10, Barcelona 08028, Spain

2. Departament de Bioquímica i Biomedicina, Facultat de Biologia, Universitat de Barcelona, Avgda Diagonal 647, Barcelona 08028, Spain

3. LBPA, UMR8113 CNRS, ENS Paris-Saclay, Université Paris-Saclay, Gif-sur-Yvette, France

4. NBD, Nostrum Biodiscovery, Josep Tarradellas 8–10, Barcelona 08019, Spain

5. School of Physics, Engineering and Technology, University of York, York YO10 5DD, UK

6. Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL, 32611, USA.

7. Department of Biological Sciences, CENUR Litoral Norte, Universidad de la República, Salto, Uruguay.

8. Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

9. Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

10. Computational Biomedicine, Institute for Neuroscience and Medicine INM-9, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany

11. Department of Physical Chemistry, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland

12. Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, EPFL SB MATH LCVMM, Station 8, CH-1015 Lausanne Switzerland

13. Chemistry and Physics Programs, Louisiana Tech University, PO Box 10348, Ruston, LA 71272, USA

14. Department of Medicinal Chemistry, College of Pharmacy, University of Utah, 2000 East 30 South Skaggs 306, Salt Lake City, Utah 84112, United States

15. Bioinformatics Unit. Institut Pasteur de Montevideo, Montevideo, Uruguay.

16. School of Mathematical and Physical Sciences, University of Sheffield, Hounsfield Road, Sheffield S3 7RH

17. School of Physics and Astronomy and Astbury Centre for Structural and Molecular Biology, University of Leeds, Leeds LS2 9JT

18. School of Pharmacy and Biodiscovery Institute, University of Nottingham, University Park, Nottingham NG7 2RD, UK.

19. Ionis Pharmaceuticals Inc., 2855 Gazelle Court, Carlsbad, CA 92010, United States

20. Department of Applied Mathematics, Kaunas University of Technology, Studentų 50-318, Kaunas, Lithuania

& These authors contributed equally

* Corresponding author modesto.orozco@irbbarcelona.org

Abstract

We present a *tour de force* of atomistic molecular dynamics simulations involving the coordinated effort of 14 research groups of the Ascona B-DNA Consortium (ABC). This initiative provides a complete characterization of the 2,080 DNA hexamers embedded in 190 carefully selected 20-mer duplexes, each simulated in replicate for at least 10 microseconds in explicit solvent. The consortium generates 0.25 petabytes of data, capturing millisecond-scale ensembles at the oligomer level and dynamics up to 10^{-1} seconds at the base-pair level. Analysis yields a comprehensive description of sequence-dependent DNA properties, including rare events such as backbone transitions, reversible base-pair changes, and partial unfolding. Processing these atomistic ensembles reveals a hidden physical code of DNA, helping explain rules of genome composition and evolution beyond coding regions. This community effort delivers unprecedented, validated FAIR data to support coarse-grained and AI models of DNA at cellular scale.

Introduction

At the start of the millennium there was already a bulk of evidence showing that DNA is a highly polymorphic and flexible polymer whose physical properties are defined by a complex choreography of sequence-dependent correlated movements sampling a vast variety of sub-states, some of them stabilized by non-canonical interactions^{1–13}. The sequence-dependent physical properties of DNA modulate specific and non-specific protein binding to DNA,^{14,15} the effect of epigenetic variants^{16–22}, mismatches^{23–27}, lesions^{5,28–36}, the placement of nucleosomes^{5,30–37}, the binding of effector proteins^{14,38} and even the compaction of chromatin^{21,39,40}. Unfortunately, the combinatorial explosion of possible sequences severely limits our understanding of how sequence determines the physical properties of DNA, hampering a genome-scale representation of DNA physics⁴¹ and the understanding of evolutionary pressure beyond simple maintenance of coding information.

The first attempts to describe sequence-dependent properties of DNA relied on the analysis of the variability in experimental structures deposited in the Protein Data Bank (PDB)^{1,5,6,42} and were focused on the description of the equilibrium and harmonic deformability of the 10 unique base pair steps (bps). These pioneer studies delivered the mesoscopic model of sequence-dependent physical properties of the DNA duplex⁶. However, it soon became evident that an independent bps model was not accurate enough and at least tetramer effects should be considered. This motivated the Ascona B-DNA Consortium (ABC)^{3,8,10–12} to use atomistic molecular dynamics (MD) simulations to characterize the physical properties of 136 unique DNA tetramers^{7,8,10,12,43}. Leading-edge at the time, trajectories were collected using consensus protocols and were stored and made available, something that was instrumental in further force-field refinement as well as for the development of coarse-grained and mesoscopic models of DNA^{44–54} and chromatin^{37,40}. However major limitations have now become apparent in these models, for example: poor description of correlated movements^{55–57}, limited knowledge on neighbour effects beyond the tetramer⁵⁸, poor representation of rare events, such as fraying, opening, unfolding, or the formation of alternative base-pairing schemes. Consequently, our understanding of the sequence-dependent properties of DNA is incomplete and unable to reliably describe and correlate DNA properties at the genome scale.

We present here an extensive simulation project, where 14 experienced groups joined efforts to describe the physical properties of DNA at the hexamer level. Through an unprecedented simulation tour de force, we provide not only the complete catalogue of harmonic and non-harmonic deformations for the entire set of 2080 unique DNA double-stranded hexamers, but describe previously unknown dynamic features of DNA, define sequences showing anomalous behaviour, and determine a similarity tree between DNA sequences in terms of their structure, dynamics and interaction properties. The massive amount of data collected here under FAIR (Findable Accessible Interoperable and Reusable) principles⁵⁹, validated by a large part of the

DNA-community, and accessible through a powerful database infrastructure, provides a set of information for training AI methods and the development of mesoscopic and coarse-grain models of DNA.

Results

Stability and convergence of the simulations. The 190 duplexes simulated here are shown in Supplementary Information Table S1, where the labels r01, r02, r03 and r04 refer to the different replicas, and labelled (*) simulations extended for double simulation times. RMSD analyses indicate that MD simulations are sampling well the expected B-fiber geometry without any major drift along the trajectories as shown in RMSD distributions for the first and last 200 ns of trajectories (Supplementary Figure S1).

The slightly larger values obtained for the 20-mer compared with the 16-mer highlights the presence of fraying effects (see Supplementary Figure S1). Only in a very few cases, these fraying effects affect the central 16-mer (see below and Supplementary Figure S1) in a stochastic manner, appearing only in one of the replicas (see Supplementary Figure S1). Overall, we obtained stable trajectories in the vast majority of simulations, and additional replicas and extension of trajectories was required only in 3% of the cases, suggesting good convergence of trajectories in the 10 μ s range. Projection of the 190x2 average structures on the PCA map (1st and 2nd mode, PC1 and PC2) of the meta-trajectory obtained by joining the individual 190x2 simulations for the central 16-mer segments shows that replicas are nearly indistinguishable, and that the impact of sequence on the duplex dynamics is not negligible (see Supplementary Figure S2A-B). Among the different sequences, the analysis shows that seq134 (GCGTCCCGAAGGCCAAGAGC, see Supplementary Table S1) displays most accurately the average characteristics of a DNA duplex, while duplexes seq047 and seq144 are despite their similar composition (see Supplementary Table S1), the ones showing the most diverse conformations. Projection of the average structures on the PCA map of seq134 shows that the structural variability sampled in a single trajectory is larger than that inferred by looking at sequence dependent variability of the average structures in the meta-trajectory (see Supplementary Figure S2), suggesting that flexibility descriptors extracted from the analysis of structural variability in experimental databases may underestimate the true flexibility.

Global conformation of the duplex. The integration of all trajectories in a global ensemble allowed us to obtain (time and sequence averaged) helical properties of DNA duplexes. The results (see Supplementary Table S2) fit very well accepted fiber data⁶⁰, except for a lower twist (33.9 degrees in average), which in fact agrees better with the average value obtained from the analysis of PDB structures (33.6 degrees), and the value inferred from the analysis of the electrophoretic behaviour of long DNA fibers (c.a. 34 degrees)⁶¹. Persistence lengths (PLs; see Supplementary Figure S3) computed here using the Noy-Golestanian formalism⁶² yielded average values slightly larger than the accepted experimental estimate (around 60 nm vs around 50 nm). This discrepancy may arise from force-field errors or from limitations inherent in mapping experimental persistence lengths onto atomistic observables. A bias toward larger persistence lengths associated with the

use of the Noy–Golestanian algorithm cannot be neglected, as slightly lower values were reported by Maddocks and co-workers⁶³ using a different persistence-length definition applied to BSC1-based coarse-grained ensembles (45–70 nm, with an average of approximately 56 nm). In any case, even major changes are expected by considering or not terminal base pairs in the calculations. Thus, when we repeated the same computations considering terminal base pairs, PL values below 50 nm were obtained (Supplementary Table S3), warning against quantitative comparisons between experimental and theoretical PLs, or even between theoretical predictions, where not detailed description of the calculations are provided. Furthermore, a recent study found a very good agreement between flexibility descriptors derived from MD-Parmbsc1 simulation and optical tweezer experiments, arguing against a systematic overestimation of stiffness in current simulations⁶³. Results in Supplementary Figure S3 show that there is a large dispersion of PL values derived for the 190 different hexABC sequences. We defer to the literature⁶⁴ for a full explanation of this fact, which is primarily due to the strong contribution of intrinsic bends to estimates of (apparent) PL at short length scales. The presence of short A-tracts, such as d(A·T)₄, is known to induce high localized intrinsic curvature of the double helix nearby, which is why seq021 with a single A-tract has a rather low PL of 43 nm. In the presence of multiple A-tracts the phasing between them strongly influences the overall intrinsic bend of the oligomer, so that seq149, which has two A-tracts separated by 3 bps, has a rather high PL of 71–72 nm, while seq077 with two A-tracts separated by one bp has an average persistence length; see Supplementary Table S1 and Supplementary Figure S3. These data highlight the complexity of A-tract-induced curvature, in which both the internal A-tract sequence and the surrounding sequence context contribute subtly to the resulting global curvature^{65,66}. They also indicate that persistence length alone cannot be taken as a simple proxy for DNA flexibility.

To assess the impact of base composition on global DNA flexibility, we grouped duplexes (16-bp central segments) into A/T-rich and G/C-rich sequences (base pair fraction >70%) and analysed their mechanical properties separately. Persistence length, RMSD, and force constants associated with the four dominant deformation modes (obtained from covariance matrix diagonalization, PC1 to PC4) show no systematic differences between the two groups (Supplementary Figures S4–S5). A/T-rich duplexes exhibit a slightly larger dispersion in all metrics, whereas G/C-rich duplexes show marginally increased flexibility along specific collective motions. Overall, base composition alone (A/T vs G/C content) does not provide a reliable predictor of global duplex flexibility. To move beyond a simple G/C vs A/T classification, we analysed sequence-dependent variations in RMSD, PL, and the force constant associated with the first principal component (PC1 in Principal Component analysis; PCA) for the central 16 bps of each sequence. Local sequence effects were quantified using a k-mer analysis (k = 4), considering tetramers present at least once within the central region and represented by a minimum of three sequences. Results in Supplementary Figure S6 show an inverse correlation between stiffness along PC1 and RMSD, indicating that both metrics describe similar deformation physics. Duplexes containing tetramers such as AGGG, TATA, ATAG, and CCTA display enhanced flexibility (high RMSD and low PC1 force constant), whereas AAAA, AATT, and AAAT are associated with increased stiffness. In general, analysis of

the most flexible sequences reveals enrichment in C-rich motifs (CpC/GpG) and in TpA, CpA/TpG steps, whereas the stiffest duplexes preferentially contain ApT, ApA/TpT and TpC/GpA motifs, typically within mixed A/T/C contexts. Notably, TpA steps are prevalent in flexible duplexes, while ApT motifs dominate rigid ones, revealing a marked directional asymmetry within A/T-rich sequences. Poly(A)-containing sequences show only moderate PL, indicating that A-tracts alone do not confer high stiffness. Overall, these results demonstrate that local sequence context and base-step orientation, rather than global A/T or G/C content, govern the mechanical response of DNA duplexes to deformation.

To understand the origin of sequence dependence on flexibility, stacking energies (intra-strand, at the base level; and total, at the base pair step level) were computed, explicitly accounting for hexameric sequence context. The results, which agree reasonably well with QM data in solution for isolate base or base pair step⁶⁷ show a marked dependence on neighbouring bases (Supplementary Figure S7). In particular, base pair steps known to display large variability in quantum calculations, such as CpG, also show the strongest context-dependent modulation of stacking energies in our analysis. These findings are further supported by melting temperature estimates obtained with OligoCalc (<https://oligocalc.eu/>), where sequences with the same number of hydrogen bonds display markedly different stabilities; for example, (CpG)₅ exhibits a T_m of 49°C, whereas (CpC/GpG)₅ melts at 33°C, highlighting the dominant contribution of stacking interactions to duplex stability.

Rare events in duplex dynamics. The exhaustive 10⁻¹ s ensemble allowed us to explore rare events usually inaccessible to standard MD simulations. Overall hydrogen bonding (HB) in the central 16-mer duplex is very well preserved, with most of the duplexes displaying a fraction Watson-Crick (WC) pairing maintained along the simulation of 95%-100% (Figure 1A). We found that central d(G·C) and d(A·T) pairs are in an open conformation compatible with solvent exchange in 0.06% (d(G·C)) and 0.25% (d(A·T)) of the meta-ensemble generated by combining the 190x2 simulations (see Figure 1B). Opening events are rare, short-lived (0.16 ns on average for the central 16mer, 0.87 ns overall) and stochastic, often coupled to terminal fraying, the most common distortion found in DNA (see examples in Figure 1C, D). Overall, terminal d(C·G) pairs are opened around 9% of the time (see Figure 1B), a value that agrees perfectly with opening estimates from NMR experiments⁶⁸. The simultaneous opening of both ends is rare (1.2 %), but slightly more probable than expected (1.0%) for fully independent events. If only long-lived (more than 10 ns) events are considered, the figures reduce to 6% opening of one end, and 0.5% for both ends (0.36% expected for independent events). Interestingly, while most of the fraying events are short-lived, reversible and restricted to the terminal base pair step, a small portion last much longer (up to 5 μs for Seq186_r01, Figure 1C). Furthermore, in some cases openings are propagated to the central d(A·T) pairs, leading to reversible partial unfolding (up to 8 base pairs are lost in Seq152_r01 and 7 in Seq040_r01; see Figure 1D and 2). It is worth noting the stochastic nature of these large fraying events (that never appear in more than one of the replicas; see Figure 1D and Supplementary Figure S1) and their reversibility (see Figures 1 and 2).

Fraying at terminal (loss of all the hydrogen bonds, at 5' or 3') d(G·C) base pairs ([G]C/[C]G) occurs predominantly in GCG/CGC trimers. The same trimer preference is observed for opening of the second base pair at duplex ends (G[C]/C[G]), suggesting that perturbations at terminal base pairs propagate to their nearest neighbours.

Opening of the third base pair in the terminal trimer GCX/CGX' is rare, but becomes more likely when this position is occupied by an d(A·T) or d(T·A) base pair (Supplementary Figure S8). The average opening d(A·T) and d(T·A) at the third base pair (around 2%) is slightly above the average opening of the same pairs in central position of the duplex (see Figure 1), suggesting a moderate next-to-nearest-neighbour effect. Central trimers exhibit consistently reduced fraying compared to terminal regions, confirming higher internal stability (Figure 1), yet show motif-specific opening predominantly in A/T-rich triads (ATA/TAT and TAA/TTA), indicating A/T content as the main driver of central base-pair opening (Supplementary Figure S9).

Opening of central d(A·T) steps leads in some cases to the formation of Hoogsteen pairs (see examples in Figure 2), instead of the canonical Watson-Crick ones. The WC→H transitions are very rare, as they were found in only 3 of the c.a. 400 individual simulations (more than 500 if those used to refine the simulation protocol are considered). Furthermore, when WC→H transitions happen in one replica, they are not detected in any of the other replicas collected for the same oligo. On average, d(A·T) base pairs predominantly adopt the Watson-Crick (WC) pairing scheme, as defined by an apparent equilibrium constant: $K_{eq} = \frac{[H]}{[WC]} \approx 10^{-6}$, indicating a nearly negligible role of the Hoogsteen pairing in defining the global characteristics of the duplex. However, WC→H transitions are not so rare in certain A/T contexts, such as d(TAAT), where (combining all the ensembles collected in this study) we obtain an apparent equilibrium constant $K_{eq} = 0.013$, very close to the value $K_{eq} = 0.010$ obtained from averaging NMR estimates on the same triads⁶⁹. It is important to note that even for those sequences the overall occurrence of Hoogsteen pairing is too low to significantly affect global metrics. To get further support for these theoretical results, we synthesized a short oligo containing one of the segments where WC→H transitions appeared spontaneously in simulations, and collected the ¹H NMR spectrum, which was compared with that obtained for a control duplex with the same length and A/T vs G/C composition. Imino signals (Supplementary Figure S10) corresponding to the central d(A·T) segment are wider and displaced to very unusual positions, indicative of polymorphism in the H-bonding pattern, something that was confirmed by melting experiments (Supplementary Figure S10). Finally, while our simulations are too short to determine the real time-scale of WC→H transitions, as we are just collecting the tails of the transition rate distribution⁷⁰, a few conclusions can be reached: i) once WC→H happens the new H-bonding scheme is quite stable^{71,72}, generating little distortion in the B-DNA helix^{73,74}; ii) at least two mechanisms for WC→H transition are possible: one following the local paradigm suggested by Al-Hashimi and co-workers^{69,75}, and another, involving partial unfolding of an A/T segment (see different cases in Figure 2)

Classical simulations performed here do not allow to explore WC→H transitions in d(G·C) pairs, as a stable Hoogsteen d(G·C) requires the protonation of cytosine. However, mining our vast ensemble, a few events of partial opening of a central d(G·C) pair were found, leading to guanosine in the syn conformation and the cytosine exposed to the solvent (Supplementary Figure S11) ready to accept a proton. As no protonation is possible in classical MD, the Hoogsteen d(G·C) pairing is not stable in our simulations, the guanosine moves to the anti-conformation, and the WC pairing scheme is recovered (last case in Figure 2). However, we checked further if upon protonation the Hoogsteen base pairing would be stable. To do so, on the open structure we force protonation of cytosine and we run a simulation. A fully stable duplex is obtained (see Supplementary Figure S11), in agreement with Al-Hashimi results on similar oligos^{69,75}, proving that once exposed to the solvent the cytosine could be protonated and then it could form a stable Hoogsteen pairing.

The hexamer helical space. Our extensive ensembles allowed characterizing the complete hexameric configurational space of DNA. The central base steps of the 2080 unique hexamers sample for 97.8% canonical helical structures. Disruptions are in most cases related to stochastic fast breathing events breaking H-bond schemes (Figure 1), and just in a few cases non-helical conformations last for a significant time. Removing those snapshots provides a slightly reduced ensemble, which can be processed to generate helical ensembles for the 6 degrees of freedom (translational and rotational base pair step parameters: shift, slide, rise, tilt, roll and twist) at the central pair of the 2080 unique hexamer environments. Analysis of the entire 6-dimensional dataset demonstrated that simple harmonic approximations are not able to fully describe the helical space sampled by DNA, due to its multimodality (see Figure 3). In fact, only around 15% of the hexamers (identical results in both replicas) can be accurately represented by a single 6-dimensional Gaussian (a single helical state, unimodal in Figure 3A). Most of the hexamers are represented by 2 (bimodal, 48%) and 3 (trimodal, 31%) Gaussian distributions, and more than 3 clusters are required only in a very small (quadrимodal, around 6%) number of cases. The intrinsic multimodality of the helical space^{1-3,77} is already visible in the 2D and 1D projections of the helical space (see Methods for details and examples in Figure 3B-D). The individual helical parameters showing higher bimodality are: twist, slide and shift (as an example see the twist distribution in Figure 3B-D and Supplementary Figures S12-S14). As a general trend, bimodality is frequent in CpG, CpA and TpA steps (twist, slide and shift respectively), but tetramer effects are quite large (see reference examples in Supplementary Figures S12 and S13), confirming the need to introduce neighbour effects to accurately describe helical space. Hexamer effects are important in just a few cases (see Figure 3 and Supplementary Figure S14), and octamer effects are expected to be negligible (see Supplementary Figure S15). These conclusions are a refinement of, but compatible with the conclusions regarding bimodality of helical space coordinates presented previously¹⁰.

Finally, we computed the force-constant matrix associated with the set of base pair step helical parameters⁹ for the complete hexameric helical space (available in the Supplementary Material).

Consistent with the global analyses of flexibility detailed above, hexamers containing A-tracts are generally more rigid in the elastic regime than other sequences. However, the primary determinant distinguishing flexible from rigid hexamers is the presence of TpA versus ApT steps: TpA steps are associated with increased flexibility, whereas ApT steps confer the highest rigidity. This distinction is quantitatively captured by the product of the diagonal elements of the force-constant matrix describing helical motions, which is systematically lower for TpA-containing hexamers and higher for those containing ApT steps (see Supplementary Figure S16).

Correlation effects. Contrary to assumptions used in early mesoscopic models⁶, correlations in the helical movements between neighbouring base pair steps (i and $i+1$) are far from being negligible^{51,52,55,56,78}. In general, intra-parameter (same base pair helical parameter) correlations between neighbouring base pair steps $i \rightarrow i+1$ are negative (anti-correlations) and in some cases quite significant, while inter-parameter correlations are less important, and can be either positive or negative (see examples in Figure 4A and Supplementary Figure S17). Overall, negative correlations dominate versus the positive ones ($\sum_{k,l} r_{i \rightarrow i+1} = -1.5$, with i standing for a given bps and k, l for the different helical parameters) and just a few $i \rightarrow i+2$ correlations, and $i \rightarrow i+3$ correlations are nonnegligible ($|r| > 0.1$; see Figure 4A), which means that local perturbations do not propagate at large distances along the helix. A significant part of such correlation is due to the coupling between the populations of different helical states in neighbouring bps (see Figure 4B). Thus, simultaneously sampling of clusters i and j in two consecutive bps ($P_{i \cup j}$) is in most cases different to that expected by a non-correlated model ($P_i \cdot P_j$; where P_i is the probability of state i in one step and P_j the probability of state j in the neighbouring bps), leading to coupling values ($C_{ij} = \frac{P_{i \cup j}}{P_i \cdot P_j}$) different from 1 (the uncorrelated reference value) in 89% of the cases (13.5% $C_{ij} < 0.5$ and 13% $C_{ij} > 1.5$) (see Figure 4B). Interactive plots showing couplings between clusters of neighbouring steps along all the sequences using Sankey's diagrams are available in Supplementary Material (Supplementary Material – Correlations).

Frustration in the helical space. Neighbour effects can generate frustration in the central bps, i.e. a given bps may sample helical regions in one sequence context, which are impeded for the same bps in a different one. This can enhance specificity, in cases where certain neighbouring bases prevent a given bps from adopting the bioactive conformation, which is, in contrast, the most stable for other neighbouring contexts. This can be quantitatively analysed using a harmonic simplification, the frustration energies (deformation energies, see Methods) like those shown in Figure 5, where we estimate the energy cost required to move the geometry of a central bps in a given sequence environment to that which is the optimum one for the same bps in all the other potential sequence environment. Results in Figure 5A illustrate that tetramer environment effects, that is, how the same central base pair step behaves when embedded in different tetrameric contexts (i.e., varying the two flanking base pairs immediately adjacent to it) are moderate in average (< 0.5 kcal/mol, i.e. around 1 kBT) but can be remarkably large in specific cases. For instance, we

detected frustration energies up to 3.5 kcal/mol, corresponding to the energy cost required for the central CpC step of the d(TCCCAT)·d(ATGGGA) hexamer to adopt the equilibrium geometry of the same bps in the d(CACCTT)·d(AAGGTG) one. Such sizable frustration energies highlight the influence of the tetrameric environment on local DNA mechanics and caution against using base-pair-level descriptors that neglect these context-dependent effects. In contrast, hexameric environment effects, which account for next-to-nearest neighbour interactions (i.e., how the same tetrameric core behaves when embedded in different hexameric sequences), are significantly milder. On average, the corresponding frustration energies are below 0.05 kcal/mol (within thermal noise), and only rarely do hexamer-hexamer comparisons exceed 0.3 kcal/mol. This behaviour is illustrated, for example, by the central CpG step shown in Figures 5B-C, where the influence of the extended hexameric environment is negligible compared to that of the tetrameric one.

The backbone torsion angles. MD simulations of all the double-stranded DNA hexamers lead to the expected sampling of backbone conformations reproducing well the expected values from experimental data (see dial plots in Supplementary Figure S18A-B). The lengths of the dynamics allow the DNA backbone to sample nonstandard regions. Most glycosidic torsion angles sample the traditional anti region (see χ (chi) torsion angle in Supplementary Figure S18C), but exploration of the syn conformations is not rare ($\chi \approx 0^\circ$ to $+90^\circ$), typically linked to terminal fraying and to the formation of Hoogsteen pairings (see Figure 2). For example, while average values indicate a South puckering, dynamics allow the DNA to sample phase angles in the entire South-East region (see Supplementary Figure S18D) with non-negligible sampling of the C3'endo conformation. If a traditional two-state puckering model is used, North states account for 18% of the total, very close to the recent estimates obtained by the analysis of experimental DNA backbones⁷⁹. The largest and most likely relevant variability is related to the ϵ/ζ concerted movement that leads to the well-known BI/BII polymorphism, responsible for the high twist–low twist bimodality. We found little variation in BI/BII populations with the position along the oligomer (see Supplementary Figure S19), and a very remarkable dependence on the sequence. Thus, while on average 20% of the bps are in the BII conformation (very close to average experimental estimates⁸⁰) there are dramatic sequence differences. For example, 48% of BII is expected for the d(GpA) step, very close to 40% from NMR estimates at the same temperature⁸¹, while no significant BII is expected in d(YpY) steps or some d(RpY) steps such as d(ApT) or d(GpT), again in perfect agreement with accurate NMR experiments⁸².

Hidden physical rules in genome evolution. While the rules governing sequence prevalence, conservation and evolution of DNA in coding regions are clear and related to maintenance of the sequence of proteins and/or efficiency of translation, those governing non-coding regions are unknown. In other words, it is easy to determine the evolutionary pressure against a single nucleotide polymorphism (SNP) in a coding region simply by translating the SNP into an amino acid substitution. However, it is unclear in the case of non-coding regions (extending these to the associated regulatory segments) where the “evolutionary pressure” should be related exclusively to the intrinsic properties of DNA and not their genetic information. Our massive analysis allows

us to compare all the hexamers from a physical point of view, either considering strand independent (2080) or dependent (4160) hexamers and decipher how the need to maintain physical properties have modulated the sequence of non-coding regions, and the evolutionary constraints that they enforced.

Firstly, we analysed the hexamer similarities using different descriptors, including dihedral angles, principal component analysis (PCA), and ion distributions (see Methods), each capturing distinct aspects of hexamer structure, flexibility, and interaction properties. These descriptors enable more meaningful comparisons among hexamers than their base sequence (see Figure 6 and Supplementary Figure S20). The most informative of these descriptors were derived from the analysis of the latent coordinate derived from autoencoders and the density of cations (K^+) in the major groove (MG). Autoencoders based on dihedral angles provide simultaneously information on structure and flexibility of the hexamers, while major groove ion densities (maj_ions) contain both information on direct (H-bond donors and acceptors) and indirect (structure based) recognition of the DNA by proteins. Interestingly, the similarity computed from the different descriptors provides a different landscape of hexamer similarities, as the global correlation between autoencoder latent variables and MG ion distributions is negligible ($r < 0.1$, Supplementary Figure S21). Only a few hexamer pairs appear as consistently similar or different (see Figure 6 and Supplementary Figure S22) in all metrics. However, the similarity/difference among hexamers depends on the comparison metrics used (see Figure 6B and Supplementary Figure S23), revealing different classifications and clustering patterns that reflect fundamentally different underlying properties of the data.

Clustering by structural/dynamic descriptors (dihedral-based autoencoder classification, Figure 6B bottom right) leads to a dendrogram characterized by localized clusters with similar populations. Two clear branches appear in the similarity tree: one with hexamers displaying no d(YpR) or with d(YpR) localized at the ends of the hexamers and another with hexamers with d(YpR) in central positions, with well-defined and differentiated clusters in the context of repeated d(YpR) segments. This similarity landscape agrees with the prevalence of non-harmonic distortions observed in these dinucleotide steps in the atomistic simulations and hexamers where the d(YpR) step is located in the centre are very different from those where it is located in a terminal position (either 5' or 3'). Hexamers with a central d(YpR) step are also the most dissimilar to those with central d(RpY), with ca. 90% of the top 100 dissimilar pairs featuring a central d(YpR) and a central d(RpY) step (see column 3, row 2 of the matrix when divided into a 4×4 grid). Overall, global patterns are mostly determined by the purine/pyrimidine character of bases, as evidenced by the fact that in the consensus patterns, 56% of the sequence codes corresponds to a R or Y, with another 28% base-agnostic (N) and not a single strict preference for a specific base (A/C/G/T).

The presence of the d(GpC) step seems the major determinant of the recognition properties along the major groove (cation densities in the major, Figure 6B at the bottom left), in agreement with previous studies finding that the cation populations are particular high for the major groove on this

step⁸³. In fact, we observe a very large cluster characterized by the lack of d(GpC) steps, where a flat ion distribution is found along the groove, while a smaller cluster is defined by hexamers rich in d(GpC) steps (ex GGGCGC or GCNGCY), where sharp bimodal distribution of ions is found (see Figure 6B).

We then analysed the impact of physical properties in guiding evolutionary variability in non-coding regions. To this end, we computed the sensitivity to mutation of different hexamers (defined as the maximum dissimilarity that can be achieved through a single nucleotide mutation; see Methods) considering both structural/dynamics (dihedral-based autoencoder) and recognition (ion densities) metrics. Figure 6C plots the sensitivities for all hexamers compared with their prevalence in a human chromosome (using as example chromosome 10 as a representative of intermediate length) excluding coding and regulatory regions. Clearly, the most frequently occurring hexamers are among the most mutation-resilient, exhibiting lower sensitivity across both comparison metrics. This relationship is further supported by population-level statistical analyses, which demonstrate a significant enrichment of highly abundant hexamers within the low-sensitivity regime (Supplementary Figure S24). Notably, an increase in A/T content appears to correlate with reduced sensitivity. In other words, our genome seems to be enriched in sequences that function as buffers against mutations, preserving the physical (structural/dynamic and recognition-related) properties of DNA.

In order to get an additional view on the relationship between physical properties and evolution rate we computed the conservation score (human vs 100 related species, <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg38&g=cons100way>). Those data taken from the phylo100 comparison analysis gave us a conservation score at base pair level along the entire human genome. Firstly, we excluded coding regions, then summed up the base pair level conservation rate for each hexamer and compared their average conservation rate with hexamer mutation sensitivity. The data is obviously noisy (see Figure 6D and Supplementary Figure S25-26), however, when comparing with major-groove ion sensitivity, a general trend emerges: higher sensitivity is associated with lower conservation. Although the correlation is weak, it is statistically significant, indicating that maintaining the physical properties of DNA has been a selective constraint during evolution.

We used our physical descriptors to try to understand the promiscuity of transcription factors (TF) to recognize DNA sequences, something that contrasts with the well-defined pattern of interactions found in experimental DNA-TF complexes that suggest a kind of binary recognition of DNA. Results summarized in Figure 7 (and other examples in Supplementary Material) show that using as discriminator the average density of ions in the major groove along a DNA segment, we can distinguish reasonably well good DNA binding sequences (top 50 best binding sequences) from bad binding sites (the 50 sequences showing the poorest binding) for a variety of TF families. It is worth noting that such predictive power stands for TFs that read exclusively the major groove, as well as those that have accessory recognition through the minor groove (like the Sox family). This

suggests that evolution has benefited from the possibility to change the specific sequence along the TF binding region but keeping the general recognition properties of the grooves.

Finally, we quantified the intrinsic affinity of DNA for nucleosome formation by computing the elastic energy required to deform DNA from its naked to its nucleosomal conformation (see Methods)⁸⁴. This deformation energy was evaluated using sequence-dependent equilibrium geometries and stiffness constants for each base-pair step in its hexameric context, and subsequently converted into a nucleosome formation probability (affinity), providing a direct measure of DNA-histone affinity. We applied this framework to analyze nucleosome formation around transcription start sites across the yeast genome. Comparison of the normalised nucleosome affinity values with experimental MNase-seq nucleosome coverage (GEO: GSE255857) shows good agreement, with low nucleosome formation probabilities in nucleosome-free regions and a marked increase at well-positioned nucleosomes (see Supplementary Figure S27).

Discussion

The central dogma of biology presents DNA as a regular helical structure acting as a blind carrier of genetic information, where the role of DNA sequence is limited to coding proteins. However, research over the last decade has shown that the DNA sequence does much more than just code for proteins, it also defines the physical properties of DNA and influences its interactions with effector proteins, as well as the structure of chromatin, which is a key element in gene regulation. In other words, it is now clear that a secondary code exists in DNA⁸⁵, a code that is much more complex to decipher than the primary genetic code, as it is encoded in pattern longer than the three-letter codons of the genetic alphabet. Determining the physical properties of DNA involves in principle solving a 4^n problem (with n is the size of the chromosome), which is far beyond the capabilities of current experimental approaches. In this paper we present an unprecedented community effort in which 14 of the most active groups in DNA simulation joined efforts to determine the minimum length of DNA required to accurately extrapolate the properties of long DNA segments. The resulting 4096 hexamers (reduced to 2080 through symmetry rules) were simulated using accepted community protocols, with individual trajectories at least 20 times longer than those used in current *state-of-the-art* simulations, and all resulting data were deposited in FAIR format and made openly available to the community. The analysis of the 0.25 petabytes of data reveals the presence of non-canonical distortions, characterized in detail here (and validated experimentally), deciphers the canonical physical properties of DNA, and uncovers previously unknown conservation rules beyond the coding regions. Perhaps more importantly, our large-scale simulation effort, conducted over three years and involving 14 experienced groups, provides reference data for developing a generation of mesoscopic and coarse-grained models as well as AI methods designed to bridge the gap between DNA simulation and chromatin structure.

Methods

Systems setup. Applying an adapted version of the Orenstein and Shamir algorithm⁸⁶, we generated an optimal sequence covering all of the 2080 unique hexamers of DNA (of length 2145 bp). This optimal sequence was subsequently split into 190 oligomers of 16 bp each, and GpC terminal pairs were added on both ends to facilitate convergence. This choice of oligomer length provides an optimal balance between number of simulations and total computational effort, and is particularly compact, with only 10 of the 2080 hexamers appearing exactly twice in the dataset (the other 2070 appear exactly once). We generated 100 sets of 190 20 bp oligomers in this way, and chose the one featuring the least extreme GpC content per oligomer, resulting in %G/C values ranging from 25% to 80% (excluding terminal GpCs) (see Supplementary Table 1 for the full list of sequences). Duplexes were built in the canonical B-like conformations using Arnott's parameters as implemented in the Nucleic Acids Builder program in AMBER suite of programs⁸⁷. All systems were surrounded by an octahedral box of water (at least 15Å distance between any DNA residue and the faces of the octahedron), each system was neutralized by adding ions and 150 mM additional KCl to mimic physiological conditions. Simulation boxes containing around 6×10^5 atoms were optimized, thermalized and equilibrated using Roe & Brooks procedure⁸⁸. After a careful pre-analysis and benchmark (to be published) considering a variety of recent force fields (FFs) for DNA, ion and water parameters we decided to use parmbsc1 force field for DNA⁴⁴, OPC water model⁸⁹ and Li-Merz⁹⁰ and in particular Joung-Cheatham for TIP4P-EW for monovalent ion parameters⁹¹. We should note that similar results are expected from OL15 and OL21 FFs, while the use of three-point water models such as TIP3P or SPC/E leads to slightly more rigid ensembles.

Molecular Dynamics Protocol. DNA structures were prepared for explicit solvent molecular dynamics simulations using AMBER tools. Hydrogen Mass Repartitioning (HMR) was applied to the topology files using `parmed_hmassrepartition` from the AMBER suite, tripling hydrogen atom masses and scaling heavy atoms accordingly to enable a 4fs integration timestep. Simulations were run using the AMBER20 suite of programs. The systems, comprising DNA, water, and counterions, underwent a 10-step equilibration protocol adapted from Roe & Brooks, standardized by the ABC consortium. This protocol included progressive energy minimizations and NVT/NPT equilibration stages with gradually reduced positional restraints on the DNA. Initial minimizations were performed with force constants up to 5 kcal/mol·Å², followed by NVT and NPT dynamics with decreasing restraints and increasing timesteps. The final step consisted of a 1 ns unrestrained NPT simulation with a 2 fs timestep. Equilibration ensured structural relaxation and system stability before production simulations. Constant-pressure equilibration (10 ns) and production (10 microseconds) were performed under periodic boundary conditions at 1 bar using the Monte Carlo barostat⁹² and Langevin dynamics to maintain the temperature at 310 K. SHAKE⁹³ constraints were applied on bonds involving hydrogen atoms and a 10 Å cutoff for non-bonded interactions. Trajectories and restart files were written in NetCDF format at regular intervals (20ps).

Production runs. Equilibrated systems were used as the starting point for each production run. Based on a prior pilot study with palindromic sequences, showing general convergence within 2–3 microseconds, each simulation was extended to 10 microseconds and sampled every 20 ps (following again conclusions from a careful benchmarking), the 10 microseconds were used for analyses of the simulation if not differently stated for some analyses. Simulations were performed in two independent replicas at 310 K and 1 atm. When significant distortions were observed in one of the replicas, the trajectory was extended to 20 microseconds and/or 2–3 additional replicas were launched. Overall, c.a. 4 ms of trajectories were collected, accounting for 10^{-1} seconds at the base pair level dynamics. Following FAIR principles, all the set-up, simulation and analyses can be reproduced using BioExcel Building Blocks technology (<https://doi.org/10.48546/workflowhub.workflow.196.6>)^{94,95}.

Molecular dynamics analyses. Canonical analyses: Trajectories were analysed at three levels: i) the entire 20-mer duplexes, ii) the central 16-mer duplexes and iii) each one of the 2080 unique hexamers. Analyses include standard metrics such as the RMSD, helical coordinates⁹⁶, principal component analysis (PCA), groove dimensions and persistence length (PL)⁹⁷. All analyses were performed considering the entire trajectories. The hydrogen bonds were calculated using the program cpptraj of Amber. The same method was used for both terminal and internal base pairs. This program considers Watson and Crick hydrogen bonds disrupted whenever the donor-acceptor distance exceeded 3.0 Å or the donor-hydrogen-acceptor angle dropped below 135°. In the study, to avoid terminal effect we did not consider the terminal bases (GpC) for the selection of the hexamers, they only consider the central 16mers, as well as analysis at helical base-pair level were obtained by removing those snapshots where the canonical hydrogen bonding (HB) scheme was lost.

Persistence lengths (PL) were calculated following the Noy-Golestanian formalism⁶², using its implementation in the SerraNA program⁹⁷. In this approach, the worm-like chain (WLC) model is adapted to atomic-resolution DNA structures, and directional decay is evaluated between the directional vectors defined at each base pair step. Previous studies have demonstrated that the transition from local to global bending regimes in PL occurs within approximately one helical turn as DNA length increases^{62,97}. Beyond this length, PL values converge, indicating that a 16-mer fragment is adequate for estimating the global persistence length.

Helical parameters analyses. The helical distributions and stiffness force constants of all bps (considering only snapshots where at least two of the Watson and Crick hydrogen bonds were preserved) in all hexamer environments were collected in the normalized six-dimensional helical space (rise, slide, shift, roll, twist and tilt). The deviation of the distributions from normality were assessed using the Bayesian Information Criterion (BIC)-Helguero analysis⁹⁸, as described elsewhere^{3,99}. Bimodality, which refers to the existence of two sub-states for a given degree of

freedom, was determined both at tetramer and hexamer level. In the first case all the data on duplicated tetramers were collected and the whole ensemble was analysed. To evaluate structural diversity across hexamers in a 6-dimensional helical space, k-means clustering was applied to each hexamer's conformational ensemble. The optimal number of clusters was determined using Gap Statistic¹⁰⁰; pooled normalized inertia was used as the dispersion measure within this calculation, normalizing each cluster's variance by its size to assess overall compactness. Correlation between bps helical movements were explored using linear or circular correlation coefficients as a function of the sequence distance⁵⁵.

Deformation/Frustration energy. Frustration in the helical space at the base pair step (bps) level refers to the energetic cost required to deform a DNA base pair step from its preferred conformation in one sequence context to that in another. To quantify this effect, we calculated the deformation (or “frustration”) energy, defined as the energy needed to shift a given base pair step (XY) from its most populated (equilibrium) structural state in one hexamer (*i*) to the most populated (equilibrium) state of the same step in another hexamer (*j*) within the helical parameter space. This approach allows us to determine how the conformational preferences of a base pair step vary as a function of sequence context, revealing which steps are more constrained or more adaptable to local structural changes. The frustration energy is given by the following expression (equation (1)):

$$E_{frust}^{i,j}(XY) = \frac{1}{2} K_i (\Delta \phi_i^j)^2 \quad (1)$$

where K_i is the stiffness matrix associated to the inter base pair helical parameters of central bps XY in the hexamer *i* and $\Delta \phi_i^j$ is the distortion vector, representing the difference between the equilibrium conformation (defined by the six inter base pair helical parameters) of the central bps (XY) in hexamer *j* with respect to equilibrium vector for the same bps in hexamer *i*. For instance, if the central base pair step CpG is compared between two different hexamers, for example AGCGTT and TGCGAA, the frustration energy quantifies how much elastic deformation energy is required for the step CpG in the hexamer AGCGTT to adopt the equilibrium conformation of the CpG step in the hexamer TGCGAA. In this example keeping the same central base pair but varying the neighbouring base pairs we could evaluate the tetrameric effect. A larger frustration energy indicates stronger context-dependent deformation or greater incompatibility between the local structural preferences. We note that frustration plays a central role in the harmonic rigid-base coarse-grain model cgDNA, as developed in^{49,50,52,53}.

The energetic cost (deformation energy) required to wrap a 147-bp DNA fragment into the nucleosomal conformation was quantified using the same an elastic energy formalism (equation 1). DNA conformations were described in terms of local helical parameters. The deformation energy was evaluated by defining the equilibrium reference state using the sequence-dependent base pair step parameters for each base pair in its hexameric sequence environment, together with the corresponding force-constant matrices obtained from molecular the dynamics simulations.^{16,84}

The final deformed state was defined by the base pair parameters of DNA in the nucleosome-wrapped conformation. Substituting these quantities into the deformation energy expression allowed a direct estimation of the energetic penalty associated with adopting the nucleosomal geometry. To relate deformation energy to nucleosome affinity, the elastic energy was converted into a Boltzmann-weighted probability ($P=\exp(-E/k_B T)$, where k_B is the Boltzmann constant, T is the temperature. This probability, here called affinity, provides a quantitative measure of the intrinsic affinity of a DNA sequence for nucleosome formation, with higher probabilities corresponding to lower energetic penalties for nucleosomal wrapping.

Principal component analyses (PCA). PCA were done in Cartesian space using standard procedures, to check the global flexibility of the 20-mer oligos and to determine the importance of hexamer effects on the essential deformation mode of the central tetramer, as well as to compare the dynamics of different hexamers. In this latter case instead of a heavy atom representation we use a sequence-agnostic representation defined by all the backbone plus N1/N9 C6/C1 atoms of the nucleobases (Pur/Pyr). The similarity between two essential dynamics was performed using Hess metrics^{101,102} with eigenvectors required to explain 90% of variance.

Interaction properties. DNA-cation interactions were estimated from the density of cations along the major and minor grooves at each bps, which define a 10-dimensional vector for each unique hexamer (2 grooves times 5 bps per hexamer). Cation distributions are analysed using curvilinear helicoidal coordinates^{83,103}, which provide a precise picture of the proximal environment of DNA by accounting for the coupling of ion dynamics with the thermal fluctuations of DNA.

Variational autoencoders (VAE)¹⁰⁴. VAE were used to compress the dynamics of the different hexamers into a 6-dimensional deep-learned latent space. To this end a VAE was trained with 8 dihedral angles per nucleotide (backbone α , β , γ , δ , ϵ , ζ , glycosidic χ , and puckering pseudorotation angle P), converted to sin/cos pairs to avoid discontinuities. At this point a single hexamer is described by a vector of 2 strands $\times$ 6 residues $\times$ 8 dihedrals $\times$ 2 trigonometric functions = 192 values between -1 and 1. Both the encoder and decoder feature 5 hidden layers of dimension 4096 each, with a ReLU nonlinearity. The training was performed for 93 epochs on the Watson-strand hexamers (replica 1). The encodings of each hexamer trajectory (subsamped so as to yield 50000 frames) define a probability distribution in the 6-dimensional latent space. Thus, we use the Hellinger distance¹⁰⁵ between two such distributions (0 corresponding to identical, and 1 to fully non-overlapping distributions) as a measure of (dis-)similarity between two hexamers. To calculate this distance, probability density is sampled (through kernel density estimation with optimized bandwidth) at 10^5 points drawn from the multivariate normal distribution, and the result is calculated as the scaled Euclidean norm of the difference of the square roots of the vectors of probability density at the sampled points.

Dissimilarities. Dissimilarities between each hexamer were calculated and for technical reasons, different dissimilarity metrics were applied for different datasets. For ion distributions, the chosen metric was the root-mean-square difference between the two density profiles, each composed of 5 points. For PCA, the chosen metric was 1 minus the normalized overlap between the principal axes (20 eigenvectors that explain around 90% of the variance). Similarly, for the dihedral distributions, the chosen metric was 1 minus the normalized overlap between the distributions in the low-dimensional latent space. Note that, for the above reasons, dissimilarity metrics are normalized and constrained for PCA and dihedrals, but unnormalized and unconstrained for ion distributions.

Consistency of dissimilarity patterns. For each hexamer, and for each pair of features, Pearson's correlation coefficient was calculated between the given rows of the corresponding dissimilarity matrices. The resulting product of these 6 correlation coefficients is reported as the "cross-feature consistency score", indicating whether for a given hexamer all features follow the same dissimilarity pattern against all other hexamers.

Sensitivity. The sensitivities of different hexamers were calculated as the maximum dissimilarity that can be achieved through a single nucleotide mutation.

Clustering methods. To identify sequence patterns that yield similar descriptors, we performed clustering based on (dis-)similarity matrices in each descriptor space, and presented the clusters as a dendrogram. The number of clusters was optimized for each dataset (PCA, minor groove ions, major groove ions, dihedrals) by calculating the sums of intra-cluster dissimilarities for cluster numbers between 2 and 20, and choosing the number after the largest drops in this metric, indicating the identification of large meaningful distinctions (as in the "elbow" method). Then, cluster divisions and identities were iteratively reconstructed starting from the smallest to the optimal number of clusters to create the dendrogram. The reported cluster entropy was calculated as the sum of the 6 entropies of the individual bases' probability distributions. The affinity propagation algorithm, using the precomputed dissimilarity matrix, was used as implemented in Python's scikit-learn¹⁰⁶. We note that analogous dendrograms for tetramer sequences have previously been presented for crystal structure data and cgDNA coarse grain energies in¹⁰⁷.

Dendrogram construction. Dendrograms were constructed by following consecutive cluster splitting events in the hierarchical clustering algorithm, using the custom (precomputed) dissimilarity matrix. Vertical position of the split on the dendrogram was determined based on the population-weighted average of mean inter-cluster dissimilarity. To that end, the average dissimilarity between all pairs of samples in each cluster was first calculated, and a population-weighted average of these per-cluster values was then computed for increasing numbers of clusters. These values were then normalized, i.e., mapped onto the $\{0,1\}$ interval via a linear transformation. For each final cluster, the total cluster population and sequence entropy (in bits, with a maximum of $6 \times \log_2(4) = 12$) are reported, along with an IUPAC-style consensus sequence.

The latter was calculated by representing the probability distribution of the 4 DNA bases at each position in a 3-simplex, dividing up the 3-simplex into equal-volume regions corresponding to each combination (A, T, C, G for 1 base; R, Y, K, M, S, W for 2 bases; B, D, H, V for 3 bases; and N for 4 bases) and assigning the code of the region where the probability distribution mapped to. Sequence logos were generated using logomaker¹⁰⁸, and cluster centers (representatives) were chosen as the ones with the lowest mean dissimilarity to all other members of the cluster⁴⁹.

All analyses were done using standard software (AMBER analysis suite, Curves+⁹⁶, MDTraj¹⁰⁹) as well as “in house” scripts. Analyses were integrated in workflows using BioExcel Building Blocks infrastructure (BioBB)^{94,95}.

UV and NMR experiments. Unlabelled DNA oligonucleotides d(CGACTCAGCG)·d(CGCTGAGTGC) and d(CGCATACGCG)·d(CGCGTATGCG) were purchased from IDT with a standard desalting purification. The samples were dissolved in a buffer containing 50 mM NaCl and 25 mM sodium phosphate buffer (pH 7) and hybridized. They were then heated to 90°C and slowly cooled down to allow duplex formation. Melting experiments were performed by heating from low temperature to 95 °C at 0.5°C/min, monitoring absorbance at 260 nm every 0.5°C with samples at 5 µM. NMR spectra were acquired in the described buffer with 10% of D₂O at 5 °C, 15 °C, 25 °C, 35 °C and 45 °C. Spectra were acquired in a Bruker spectrometer operating at 600 MHz, equipped with cryoprobe and processed with the TOPSPIN software. Water suppression was achieved by including an excitation sculpting module in the pulse sequence¹¹⁰.

Data Storage. All trajectories were stored in a non-relational database using MongoDB infrastructure and MDDB protocols¹¹¹. All trajectories are stored in a FAIR format with detailed provenance. This includes the group performing the simulation, validation, and all the individual steps followed to launch the simulation. Overall, we have collected around 0.25 PB of trajectories, which are stored in Barcelona, with a backup located at the Center for High Performance Computing (University of Utah, Salt Lake City, Utah, USA).

Data availability

Trajectories, set up files, metadata and standard trajectory analyses are available in the MDDb database: [<https://mdposit.mddbr.eu/#/collection/abc>] and [<https://mdposit.mddbr.eu/#/id/MD-A003ZZ/overview>].

The Source Data generated and analysed in this study have been deposited in the repository at: [DOI <https://doi.org/10.5281/zenodo.18361274>]; including processed datasets for each figure, analysis outputs, and representative trajectory frames. Due to the large size of the full MD trajectories, the associated raw data are available from the corresponding author upon request.

ARTICLE IN PRESS

References

1. Dans, P. D., Pérez, A., Faustino, I., Lavery, R. & Orozco, M. Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.* **40**, 10668–78 (2012).
2. Dans, P. D. *et al.* Unraveling the sequence-dependent polymorphic behavior of d (CpG) steps in B-DNA. *Nucleic Acids Res.* **42**, 11304–11320 (2015).
3. Dans, P. D. *et al.* The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res.* **47**, 11090 (2019).
4. da Rosa, G. *et al.* Sequence-dependent structural properties of B-DNA: what have we learned in 40 years? *Biophys. Rev.* **13**, 995–1005 (2021).
5. Xu, F. & Olson, W. K. DNA architecture, deformability, and nucleosome positioning. *J. Biomol. Struct. Dyn.* **27**, 725–39 (2010).
6. Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences* **95**, 11163–11168 (1998).
7. Dixit, S. B., Mezei, M. & Beveridge, D. L. Studies of base pair sequence effects on DNA solvation based on all-atom molecular dynamics simulations. *J. Biosci.* **37**, 399–421 (2012).
8. Dixit, S. B. *et al.* Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.* **89**, 3721 (2005).
9. Lankaš, F., Šponer, J., Hobza, P. & Langowski, J. Sequence-dependent elastic properties of DNA. *J. Mol. Biol.* **299**, 695–709 (2000).
10. Pasi, M. *et al.* μ ABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42**, 12272–12283 (2014).
11. Beveridge, D. L. *et al.* Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. I. Research Design and Results on d(CpG) Steps. *Biophys. J.* **87**, 3799–3813 (2004).
12. Lavery, R. *et al.* A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **38**, 299–313 (2010).
13. Balaceanu, A. *et al.* The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.* **8**, 21–28 (2017).
14. Battistini, F. *et al.* How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. *J. Mol. Biol.* **431**, 3845–3859 (2019).
15. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–53 (2009).
16. Portella, G., Battistini, F. & Orozco, M. Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Comput. Biol.* **9**, e1003354 (2013).
17. Battistini, F. *et al.* The Impact of the HydroxyMethylCytosine epigenetic signature on DNA structure and function. *PLoS Comput. Biol.* **17**, e1009547 (2021).
18. Pérez, A. *et al.* Impact of Methylation on the Physical Properties of DNA. *Biophys. J.* **102**, 2140–2148 (2012).
19. Severin, P. M. D., Zou, X., Gaub, H. E. & Schulten, K. Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Res.* gkr578- (2011) doi:10.1093/nar/gkr578.

20. Choy, J. S. *et al.* DNA methylation increases nucleosome compaction and rigidity. *J. Am. Chem. Soc.* **132**, 1782–3 (2010).
21. Buitrago, D. *et al.* Impact of DNA methylation on 3D genome structure. <https://doi.org/10.21203/rs.3.rs-36311/v1> (2020) doi:10.21203/rs.3.rs-36311/v1.
22. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388–92 (2010).
23. Rossetti, G. *et al.* The structural impact of DNA mismatches. *Nucleic Acids Res.* **43**, 4309–4321 (2015).
24. Sharma, M., Predeus, A. V., Mukherjee, S. & Feig, M. DNA bending propensity in the presence of base mismatches: implications for DNA repair. *J. Phys. Chem. B* **117**, 6194–205 (2013).
25. Wang, Y., Zhang, X., Wu, F., Chen, Z. & Zhou, X. Bisulfite-free, single base-resolution analysis of 5-hydroxymethylcytosine in genomic DNA by chemical-mediated mismatch. *Chem. Sci.* **10**, 447–452 (2019).
26. Kingsland, A. & Maibaum, L. DNA Base Pair Mismatches Induce Structural Changes and Alter the Free-Energy Landscape of Base Flip. *Journal of Physical Chemistry B* **122**, 12251–12259 (2018).
27. Brovarets, O. O. & Hovorun, D. M. Wobble Watson-Crick tautomeric transitions in the homo-purine DNA mismatches: A key to the intimate mechanisms of the spontaneous transversions. *J. Biomol. Struct. Dyn.* **33**, 2710–2715 (2015).
28. Urrutia, K. *et al.* DNA sequence and lesion-dependent mitochondrial transcription factor A (TFAM)-DNA-binding modulates DNA repair activities and products. *Nucleic Acids Res.* **52**, 14093–14111 (2024).
29. Kufner, C. L. *et al.* Sequence dependent UV damage of complete pools of oligonucleotides. *Scientific Reports* 2023 13:1 **13**, 1–10 (2023).
30. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–83 (2008).
31. Chung, H.-R. & Vingron, M. Sequence-dependent nucleosome positioning. *J. Mol. Biol.* **386**, 1411–22 (2009).
32. Bai, L. & Morozov, A. V. Gene regulation by nucleosome positioning. *Trends Genet.* **26**, 476–83 (2010).
33. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–8 (2006).
34. Battistini, F., Hunter, C. A., Moore, I. K. & Widom, J. Structure-based identification of new high-affinity nucleosome binding sequences. *J. Mol. Biol.* **420**, 8–16 (2012).
35. Lowary, P. T. & Widom, J. Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1183–1188 (1997).
36. Deniz, Ö. *et al.* Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics* **12**, 489 (2011).
37. Sereda, Y. V. & Bishop, T. C. Evaluation of elastic rod models with long range interactions for predicting nucleosome stability. *J. Biomol. Struct. Dyn.* **27**, 867–887 (2010).
38. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–69 (2010).

39. Neguembor, M. V. *et al.* MiOS, an integrated imaging and computational strategy to model gene folding with nucleosome resolution. *Nature Structural & Molecular Biology* 2022 29:10 **29**, 1011–1023 (2022).
40. Stolz, R. C. & Bishop, T. C. ICM Web: The interactive chromatin modeling web server. *Nucleic Acids Res.* **38**, (2010).
41. Basu, A. *et al.* Deciphering the mechanical code of the genome and epigenome. *Nature Structural & Molecular Biology* 2022 29:12 **29**, 1178–1187 (2022).
42. Pérez, A., Noy, A., Lankas, F., Luque, F. J. & Orozco, M. The relative flexibility of B-DNA and A-RNA duplexes: Database analysis. *Nucleic Acids Res.* **32**, 6144–6151 (2004).
43. Dans, P. Multiscale simulations of DNA from electrons to nucleosomes: 22 years of the Ascona B-DNA Consortium. *Biophys. Rev.* **16**, 265–267 (2024).
44. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **13**, 55–8 (2015).
45. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–29 (2007).
46. Galindo-Murillo, R. *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.* **12**, 4114–4127 (2016).
47. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **7**, 2886–2902 (2011).
48. Gonzalez, O. & Maddocks, J. H. Extracting parameters for base-pair level models of DNA from molecular dynamics simulations. *Theor. Chem. Acc.* **106**, 76–82 (2001).
49. Sharma, R., Patelli, A. S., De Bruin, L. & Maddocks, J. H. cgNA+web : A Visual Interface to the cgNA+ Sequence-dependent Statistical Mechanics Model of Double-stranded Nucleic Acids. *J. Mol. Biol.* **435**, 167978 (2023).
50. Petkevičiūtė, D., Pasi, M., Gonzalez, O. & Maddocks, J. H. cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.* **42**, e153–e153 (2014).
51. Lankas, F. *et al.* On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Physical Chemistry Chemical Physics* **11**, 10565–10588 (2009).
52. Gonzalez, O., Petkevičiūtė, D. & Maddocks, J. H. A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.* **138**, (2013).
53. De Bruin, L. & Maddocks, J. H. cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.* **46**, W5–W10 (2018).
54. Walther, J. *et al.* A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.* **48**, (2020).
55. Lopez-Güell, K., Battistini, F. & Orozco, M. Correlated motions in DNA: beyond base-pair step models of DNA flexibility. *Nucleic Acids Res.* **51**, 2633–2640 (2023).
56. Liebl, K. & Zacharias, M. Accurate modeling of DNA conformational flexibility by a multivariate Ising model. *Proceedings of the National Academy of Sciences* **118**, (2021).
57. Horenko, I. *et al.* Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to B-DNA. *SIAM J. Appl. Dyn. Syst.* **7**, 532–560 (2008).
58. Balaceanu, A. *et al.* Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.* **47**, 4418–4430 (2019).

59. Hospital, A., Battistini, F., Soliva, R., Gelpí, J. L. & Orozco, M. Surviving the deluge of biosimulation data. *WIREs Computational Molecular Science* **10**, (2020).
60. Arnott, S. & Hukins, D. W. Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* **47**, 1504–9 (1972).
61. Bettotti, P. *et al.* Structure and Properties of DNA Molecules Over The Full Range of Biologically Relevant Supercoiling States. *Sci. Rep.* **8**, 6163 (2018).
62. Noy, A. & Golestanian, R. Length Scale Dependence of DNA Mechanical Properties. *Phys. Rev. Lett.* **109**, 228101 (2012).
63. Mitchell, J. S., Glowacki, J., Grandchamp, A. E., Manning, R. S. & Maddocks, J. H. Sequence-Dependent Persistence Lengths of DNA. *J. Chem. Theory Comput.* **13**, 1539–1555 (2017).
64. Minhas, V. *et al.* Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. *Journal of Physical Chemistry B* **124**, 38–49 (2020).
65. Marin-Gonzalez, A. *et al.* Understanding the paradoxical mechanical response of in-phase A-tracts at different force regimes. *Nucleic Acids Res.* **48**, 5024 (2020).
66. Gardasevic, T. & Noy, A. The impact of sequence periodicity on DNA mechanics: investigating the origin of A-tract's curvature. *Nanoscale* **16**, 18410–18420 (2024).
67. Kruse, H., Banáš, P. & Šponer, J. Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *J. Chem. Theory Comput.* **15**, 95–115 (2018).
68. Priyakumar, U. D. & MacKerell, A. D. NMR imino proton exchange experiments on duplex DNA primarily monitor the opening of purine bases. *J. Am. Chem. Soc.* **128**, 678–679 (2006).
69. Alvey, H. S., Gottardo, F. L., Nikolova, E. N. & Al-Hashimi, H. M. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nature Communications* **2014** 5:1 **5**, 1–8 (2014).
70. Voelz, V. A., Pande, V. S. & Bowman, G. R. Folding@home: Achievements from over 20 years of citizen science herald the exascale era. *Biophys. J.* **122**, 2852–2863 (2023).
71. Cubero, E., Abrescia, N. G. A., Subirana, J. A., Luque, F. J. & Orozco, M. Theoretical Study of a New DNA Structure: The Antiparallel Hoogsteen Duplex. *J. Am. Chem. Soc.* **125**, 14603–14612 (2003).
72. Abrescia, N. G. A., Thompson, A., Huynh-Dinh, T. & Subirana, J. A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2806–2811 (2002).
73. Rueda, M., Cubero, E., Lughton, C. A. & Orozco, M. Exploring the counterion atmosphere around DNA: What can be learned from molecular dynamics simulations? *Biophys. J.* **87**, 800–811 (2004).
74. Cubero, E., Luque, F. J. & Orozco, M. Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys. J.* **90**, 1000–1008 (2006).
75. Shi, H. *et al.* Revealing A-T and G-C Hoogsteen base pairs in stressed protein-bound duplex DNA. *Nucleic Acids Res.* **49**, 12540 (2021).
76. Galindo-Murillo, R. & Cheatham, T. E. Transient Hoogsteen Base Pairs Observed in Unbiased Molecular Dynamics Simulations of DNA. *Journal of Physical Chemistry Letters* **13**, 6283–6287 (2022).
77. Dršata, T. & Lankaš, F. Multiscale modelling of DNA mechanics. *Journal of Physics Condensed Matter* **27**, (2015).

78. Sutthibutpong, T. *et al.* Long-range correlations in the mechanics of small DNA circles under topological stress revealed by multi-scale simulation. *Nucleic Acids Res.* **44**, 9121–9130 (2016).
79. da Rosa, G., Grille, L. & Dans, P. D. Ramachandran-like Conformational Space for DNA. *J. Chem. Inf. Model.* **64**, 8339–8348 (2024).
80. Ben Imeddourene, A. *et al.* Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLoS Comput. Biol.* **11**, e1004631 (2015).
81. Tian, Y. *et al.* ³¹P NMR Investigation of Backbone Dynamics in DNA Binding Sites. *J. Phys. Chem. B* **113**, 2596 (2009).
82. Heddi, B., Oguey, C., Lavelle, C., Foloppe, N. & Hartmann, B. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.* **38**, 1034–1047 (2010).
83. Pasi, M., Maddocks, J. H. & Lavery, R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.* **43**, 2412–2423 (2015).
84. Sala, A. *et al.* An integrated machine-learning model to predict nucleosome architecture. *Nucleic Acids Res.* **52**, 10132–10143 (2024).
85. Travers, A. A., Muskhelishvili, G. & Thompson, J. M. T. DNA information: from digital code to analogue structure. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences* **370**, 2960–2986 (2012).
86. Orenstein, Y. & Shamir, R. Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers. *Bioinformatics* **29**, i71–i79 (2013).
87. Macke, T. J. & Case, D. A. Modeling Unusual Nucleic Acid Structures. in 379–393 (1997). doi:10.1021/bk-1998-0682.ch024.
88. Roe, D. R. & Brooks, B. R. A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations. *Journal of Chemical Physics* **153**, (2020).
89. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: A different approach. *Journal of Physical Chemistry Letters* **5**, 3863–3871 (2014).
90. Li, P., Song, L. F. & Merz, K. M. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. *Journal of Physical Chemistry B* **119**, 883–895 (2015).
91. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *Journal of Physical Chemistry B* **112**, 9020–9041 (2008).
92. Åqvist, J., Wennerström, P., Nervall, M., Bjelic, S. & Brandsdal, B. O. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chem. Phys. Lett.* **384**, 288–294 (2004).
93. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
94. Bayarri, G., Andrio, P., Hospital, A., Orozco, M. & Gelpí, J. L. BioExcel Building Blocks Workflows (BioBB-Wfs), an integrated web-based platform for biomolecular simulations. *Nucleic Acids Res.* **50**, W99–W107 (2022).
95. Andrio, P. *et al.* BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. *Scientific Data* 2019 6:1 **6**, 1–8 (2019).

96. Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37**, 5917–29 (2009).
97. Velasco-Berrelleza, V. *et al.* SerraNA: a program to determine nucleic acids elasticity from simulation data. *Physical Chemistry Chemical Physics* **22**, 19254–19266 (2020).
98. de Helguero, F. Sui Massimi Delle Curve Dimorfiche. *Biometrika* **3**, 84 (1904).
99. Battistini, F., Sala, A., Hospital, A. & Orozco, M. Sequence-Dependent Properties of the RNA Duplex. *J. Chem. Inf. Model.* **63**, 5259–5271 (2023).
100. Tibshirani, R., Walther, G. & Hastie, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *J. R. Stat. Soc. Series B Stat. Methodol.* **63**, 411–423 (2001).
101. Pérez, A. *et al.* Exploring the Essential Dynamics of B-DNA. *J. Chem. Theory Comput.* **1**, 790–800 (2005).
102. Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **62**, 8438 (2000).
103. Lavery, R., Maddocks, J. H., Pasi, M. & Zakrzewska, K. Analyzing ion distributions around DNA. *Nucleic Acids Res.* **42**, 8138–8149 (2014).
104. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* <https://doi.org/10.61603/ceas.v2i1.33> (2013) doi:10.61603/ceas.v2i1.33.
105. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die Reine und Angewandte Mathematik* **1909**, 210–271 (1909).
106. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research* vol. 12 <http://scikit-learn.sourceforge.net>. (2011).
107. Sharma, R. cgNA+: A sequence-dependent coarse-grain model of double-stranded nucleic acids. <https://doi.org/10.5075/EPFL-THESIS-9792> (2023) doi:10.5075/EPFL-THESIS-9792.
108. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
109. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
110. Stott, K., Keeler, J., Hwang, T. L., Shaka, A. J. & Stonehouse, J. Excitation Sculpting in High-Resolution Nuclear Magnetic Resonance Spectroscopy: Application to Selective NOE Experiments. *J. Am. Chem. Soc.* **117**, 4199–4200 (1995).
111. Beltrán, D., Hospital, A., Gelpí, J. L. & Orozco, M. A new paradigm for molecular dynamics databases: the COVID-19 database, the legacy of a titanic community effort. *Nucleic Acids Res.* **52**, D393 (2023).
112. Barissi, S., Sala, A., Wieczór, M., Battistini, F. & Orozco, M. DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors. *Nucleic Acids Res.* **50**, 9105–9114 (2022).

ARTICLE IN PRESS

Acknowledgements

We gratefully acknowledge the pioneering work and lasting scientific legacy of Krystyna Zakrzewska and Richard Lavery, co-founders and driving forces of the ABC consortium, for their fundamental contributions to this field. This work was supported by the Center of Excellence for HPC H2020 European Commission; BioExcel-3: Centre of Excellence for Computational Biomolecular Research [European Union: 101093290; Ministerio de Ciencia e Innovación: PCI2022-134976-2] to MO; the MDDDB project [101094561] to MO; Spanish Ministry of Science [PID2021-122478NB-I00] to MO; Instituto de Salud Carlos III—Instituto Nacional de Bioinformática, Fondo Europeo de Desarrollo Regional [ISCIII PT 17/0009/0007] to MO; and European Regional Development Fund, ERFD Operative Programme for Catalunya, the Catalan Government AGAUR [SGR2021 00863] to MO. The IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO. Portions of this research were conducted with high-performance computational resources provided by the Louisiana Optical Network Infrastructure (<http://www.loni.org>). The authors thankfully acknowledge the NMR resources and the technical support provided by the LRB and LMR of the Spanish ICTS Red de Laboratorios de RMN de Biomoléculas (R-LRB). The authors thank the Barcelona Supercomputing Center for a HPC resources. Modesto Orozco is an ICREA Academia Fellow. Federica Battistini is a Serra Hunter Fellow.

Author contribution Statement

F.B. contributed to study design, performed and coordinated analyses, and contributed to the manuscript preparation. M.W. performed data analyses, including dendrogram construction, correlation analyses, autoencoder development, and contributed to manuscript writing. M.P. carried out ion analyses and contributed to manuscript preparation. A.P. contributed to manuscript preparation. A.H. led the initial simulation setup, implemented transfer-dynamics technologies, performed trajectory and data analyses, and contributed to manuscript preparation. J.P.A. and R.G.-M. contributed to the initial simulation setup and protocol development. C. L. contributed to the initial simulation setup and protocol, and to the simulation checks. I.S.-C. contributed NMR data analysis and its integration with simulations. A.S. performed biological data correlation analyses, including coverage and conservation metrics. S.D. analysed simulation trajectories, including RMSD and hydrogen-bonding patterns. P.D.D. contributed to figures preparation. A.N. contributed to the persistence length analyses. M.O. conceived the central idea of the project, designed the overall study, supervised the work, and contributed to manuscript preparation.

All authors, F.B., M.W., A.H., M.P., J.P.A., I.S.-C., A.S., S.D., A.G., M.B., E.W.C., L.C., G.d.R., J.R.E., G.L.H., K.A.H., M.J., R.P., R.Sh., R.Su., T.C.B., P.C., T.E.C., R.C.-G., J.C., P.D.D., S.A.H., C.L., R.G.-M., J.H.M., A.N., A.P., D.P.-G., M.O., contributed to the setup, design, and execution of molecular dynamics simulations, and reviewed and approved the final manuscript.

Competing interests Statement

The authors declare no competing interests.

ARTICLE IN PRESS

Figure Legends/Captions

Figure 1. Hydrogen bond patterns analyses. Hydrogen bond (Hbond) dynamics for each sequence referred to the max number of Hbond for each sequence. A) Fraction of Watson-Crick hydrogen bond and the average time of conservation (replica 1 and replica 2 merged). Highlighted the sequence number of duplexes affected by fraying along the simulation (interactive plots in Supplementary Material). B) Base opening statistics based on the analysis of the WC hydrogen bonds for replica 1 and replica 2 averaged (% of the time). For the terminal base d(GpC)/d(CpG) it considers each Hbond lost independently, while only in 6% of the time they occur simultaneously. C-D) Examples of the Hbond pattern along the simulations showing fraying and opening with corresponding images of the dsDNA. For each base pair along the y axes, in yellow 3, in green 2, in light blue 1 and in dark blue 0 Watson and Crick hydrogen bonds detected. The blue rectangles identify regions where hydrogen-bond breaking was detected and where the depicted structure was observed. C) Sequence 186_r01 showing fraying at the bottom of the helix, with recover of the helix after extending the dynamics for further 10 microseconds (right panel). D) Sequences 30_r01, 150_r01, 40_r01 (respectively from left to right) showing other examples of fraying and opening along the simulation.

Figure 2. Hoogsteen pairing forming along the simulations. On the left the Watson and Crick hydrogen bond pattern, for each base pair along the y axes, in yellow 3, in green 2, in light blue 1 and in dark blue 0. On the right images of the helix and the detail of the pairing going from Anti to syn conformation from canonical Watson and Crick pairing to Hoogsteen pairing. From top to bottom sequences: Seq152_r01, Seq174_r02, Seq150_r04 and Seq186_r01. Seq152_r01 shows a deep fraying of the helix, followed by a rotation of the Adenine5 and formation of the Hoogsteen pairing with consequential closing of the helix. Seq174_r02 shows a different mechanism, the helix maintains its shape and Adenine26 flips out from it, turns and goes back with the formation of the Hoogsteen pairing. Seq150_r04 shows just a rotation around the chi angle of Adenine38 with the formation of the Hoogsteen pairing. Seq186_r01 shows a G·C potential Hoogsteen pairing consequential of a deep fraying. The orange arrow indicates the possible protonation site, exposed to the solvent and deeply investigated by MD simulation (Supplementary Figure S11).

Figure 3. Multimodality at the hexamer level. A) Percentage of the population of hexamer represented by 1 (15% unimodal distribution, in blue), 2 (48%, bimodal distribution, in orange), 3 (31%, trimodal distribution, in grey), or more than 3 clusters (4, around 6% in yellow; quadrimodal distribution, the largest number of clusters required is 5) Gaussian distributions. B) Bimodality of the base pair parameter twist at the tetramer level of the central base-pair (x-axis) in all possible sequence contexts (y-axis); if the tetramer average value is within the global average $\pm$ SD (34.0 ± 2.4 degrees) are represented as a white cell, the blue ones correspond to values above the average tetramer + SD, while red cells mean an average value below the global average-SD; when bimodality exists, the cell is divided into portions and the value refers to each one of the maxima as found by the BIC analysis. In a green square is highlighted one case with d(CpA) central base pair surrounded by the bases A-A (ACAA), which showed bimodality at the tetramer level C) Representative selection of hexameric contexts from the full dataset (see Supplementary Figure S14), highlighting the bimodality of the base pair parameter twist at the hexamer level of some central

tetramer (y-axis) across all possible sequence contexts (top x-axis). Example of a selected tetramer ACAA in four different hexameric environments is highlighted in a green square. The selected cases show both unimodality and bimodality even if the same case at tetramer level in panel B showed only bimodality. D) Density maps of slide vs. twist of the cases highlighted in panel C showing unimodality (left panel cases) and bimodality (right panel cases).

Figure 4. Base pair parameters correlation effects. A) Average correlation effects for all the sequences simulated with a variation in correlation $|r| > 0.3$ (Sh, Ti, Tw, Sl and Ro stand for the base pair parameters considered that showed correlations: Shift, Tilt, Twist, Slide and Roll respectively). B) Statistical distribution of the Ratio Value (see C_{ij} formula and details in the text) showing clear coupling between states illustrated in the difference in the probability of a combination of states detected along the MD compared (>1) to that expected from uncorrelated-independent movements (equal to 1).

Figure 5. Sequence-dependent frustration energy. A) Frustration at the tetramer level: deformation energy values (kcal/mol) calculated between hexamers sharing the same central dimer (x-axis) (for each dimer, number of points $n=65536$). B) Example of deformation energy variations at the hexamer level (kcal/mol), calculated between hexamers that share the same central tetramer (x-axis); shown here for all tetramers containing a central d(CpG) step (for each tetramer, number of points $n=4096$). In each box plot the centre line indicates the median, the box spans the 25th–75th percentiles, and whiskers extend to $1.5 \times \text{IQR}$; points beyond whiskers are outliers. C) Heatmap showing deformation energy variations at the tetramer/hexamer level for the central base-pair step CpG. The colour scale ranges from 0 kcal/mol (dark blue) to 2.4 kcal/mol (dark red). Highlighted on the x-axis are the tetramers ACGA, ACGC, ACGG, ACGT, CCGA, and CCGG, with the corresponding hexamers shown on the y-axis. Red regions indicate highly sensitive sequence contexts where changes in tetrameric or hexameric environment strongly alter local geometry and stiffness of the d(CpG) step. Dark-blue regions indicate contexts in which neighbouring steps have little influence on d(CpG) geometry or flexibility.

Figure 6. Hexamer dissimilarity patterns. A) Heatmap of the hexamers' dissimilarity using autoencoders-dihedrals (bottom triangle) and major groove ion population (upper triangle) data, highlighting the central base pair for each block of hexamers (see Supplementary Methods for ordering). B) Dendrograms illustrating similarities in major groove ion population (left panel) and dihedral angle distributions (right panel); logo plots indicate the statistical distribution of bases in each cluster, and grey contours at the bottom show the distribution of the ions along the major groove for each sequence in the cluster. Text labels indicate the consensus sequence, the cluster centre (representative), statistical entropy, and cluster population. C) Correlation between frequency (order of 10^6) in the human genome and ion (in red) and dihedrals (in blue) sensitivity mutation. To facilitate reading of the plot poly-d(A) and poly-d(T) at the bottom right of the Figure were removed (see Supplementary Figure S25 for the whole dataset). See Supplementary Information and Supplementary Figure S24 for statistical analyses on how the most populated hexamers are enriched in the low-sensitivity regime at the population level. D). Correlation between total hexamer conservation rate (score) and major groove ion sensitivity; the correlation is statistically significant, the association is quantified using a two-sided Pearson correlation test ($r = -0.376$, $p = 5.396 \times 10^{-138}$, $n=4096$) and two-sided Spearman rank correlation ($r = -0.383$, $p = 3.244 \times 10^{-143}$, $n=4096$).

[&]IUPAC code is used for nucleobases: R (A or G), Y(C or T), S (G or C), W (A or T), K (G or T), M (A or C), B (all but A), D (all but C), H (all but G), V (all but T) and N (any base), see Supplementary Methods.

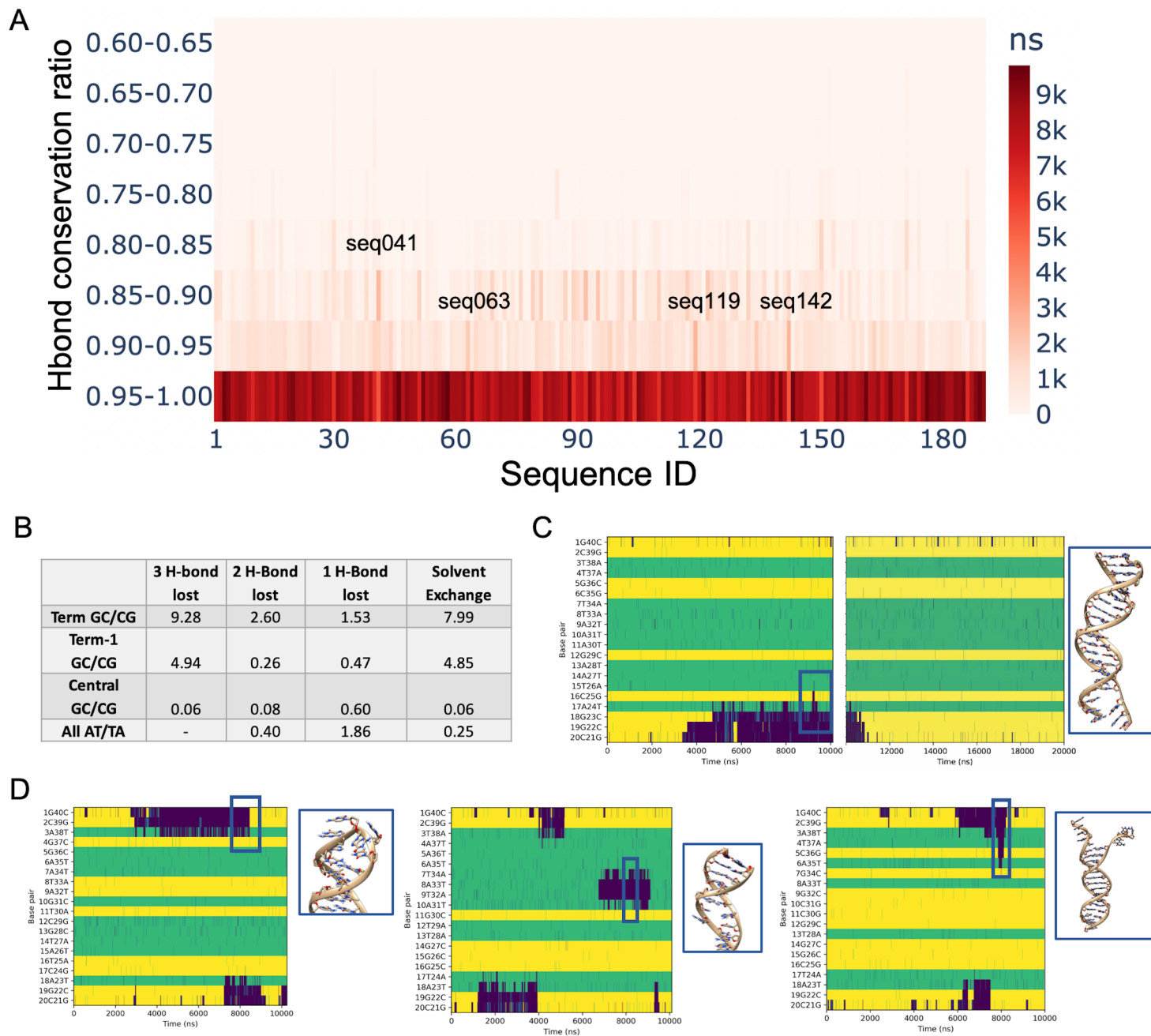
^{&&}Note that here we consider un-symmetrized hexamers, i.e. the sequences in both strands were considered independently.

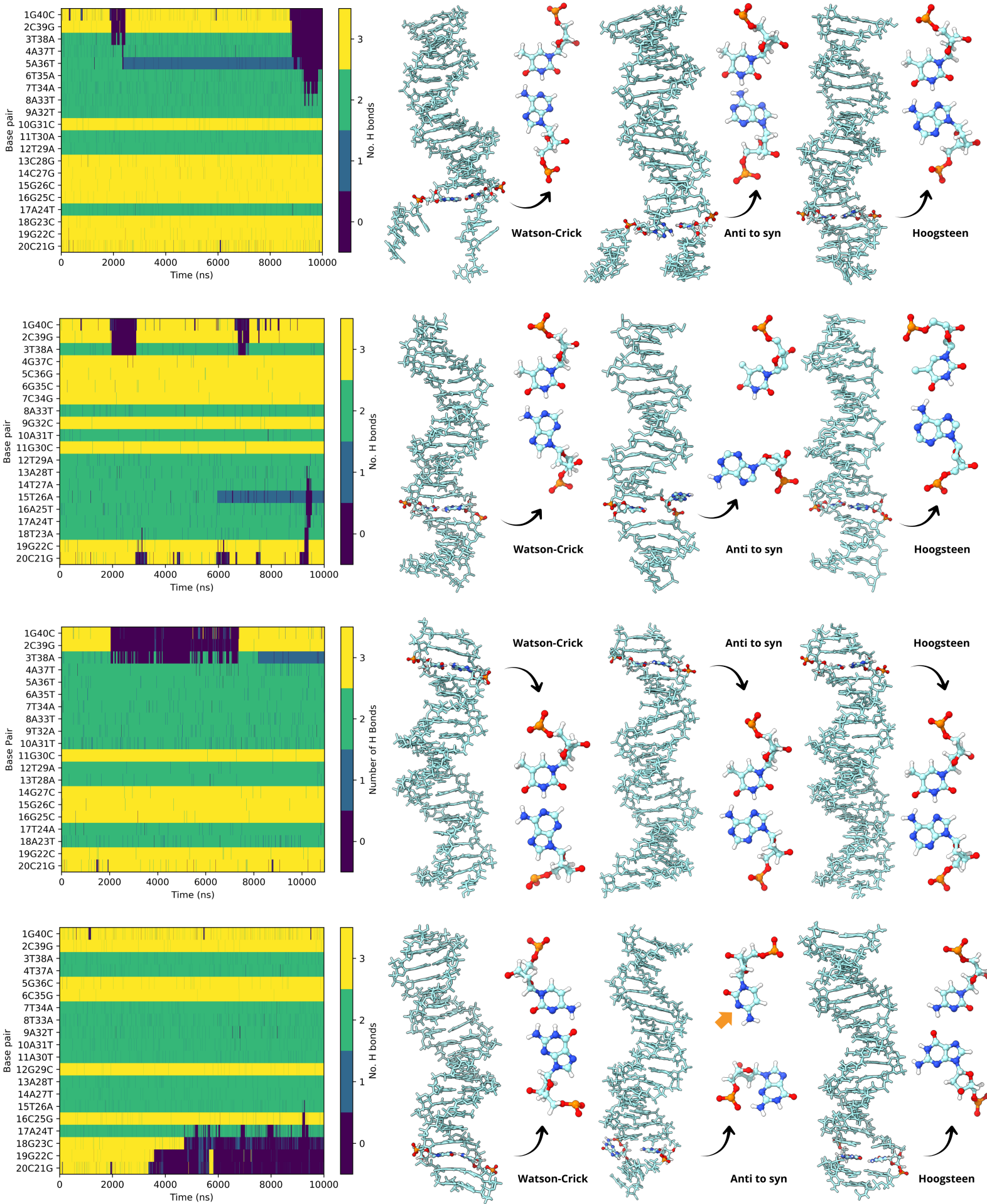
Figure 7. Transcription factors recognition and ion population correlation. Average major groove ion population across 10-mers selected experimentally using HT-SELEX for transcription factors (TFs) binding via the major or minor DNA groove. The bar plot at the top display the distribution of p-values from two-sided Mann-Whitney U tests (test statistic U, exact P value) assessing whether high-affinity sequences (top 50, n=50) are statistically distinct from low-affinity sequences (bottom 50, n=50). P-values below 0.001 indicate significant differences in cation densities, suggesting that enriched 10-mers (top 50, in light blue) have distinct ion interactions in grooves compared to the background and in particular to the low affinity sequences (bottom 50, in light red). Statistics were computed separately for two datasets, corresponding to TFs that preferentially bind via the major or minor groove. Example distributions of average ion occupancy in the major groove (average number of ions along the major groove of each k-mer) for both high- and low-affinity sequences across selected protein families are shown below. The TFs were selected to cover both major groove binding and minor groove binding. Binding affinity data were pre-processed from ENA projects PRJEB14744 and PRJEB29730 using the SELEX R package (<https://bioconductor.org/packages/release/bioc/html/SELEX.html>)¹¹². Transcription factors lacking a clearly enriched 10-mer (≥ 100 counts in the 0th SELEX cycle) were excluded due to the absence of a defined binding motif. More cases across different TFs families available in Supplementary Material.

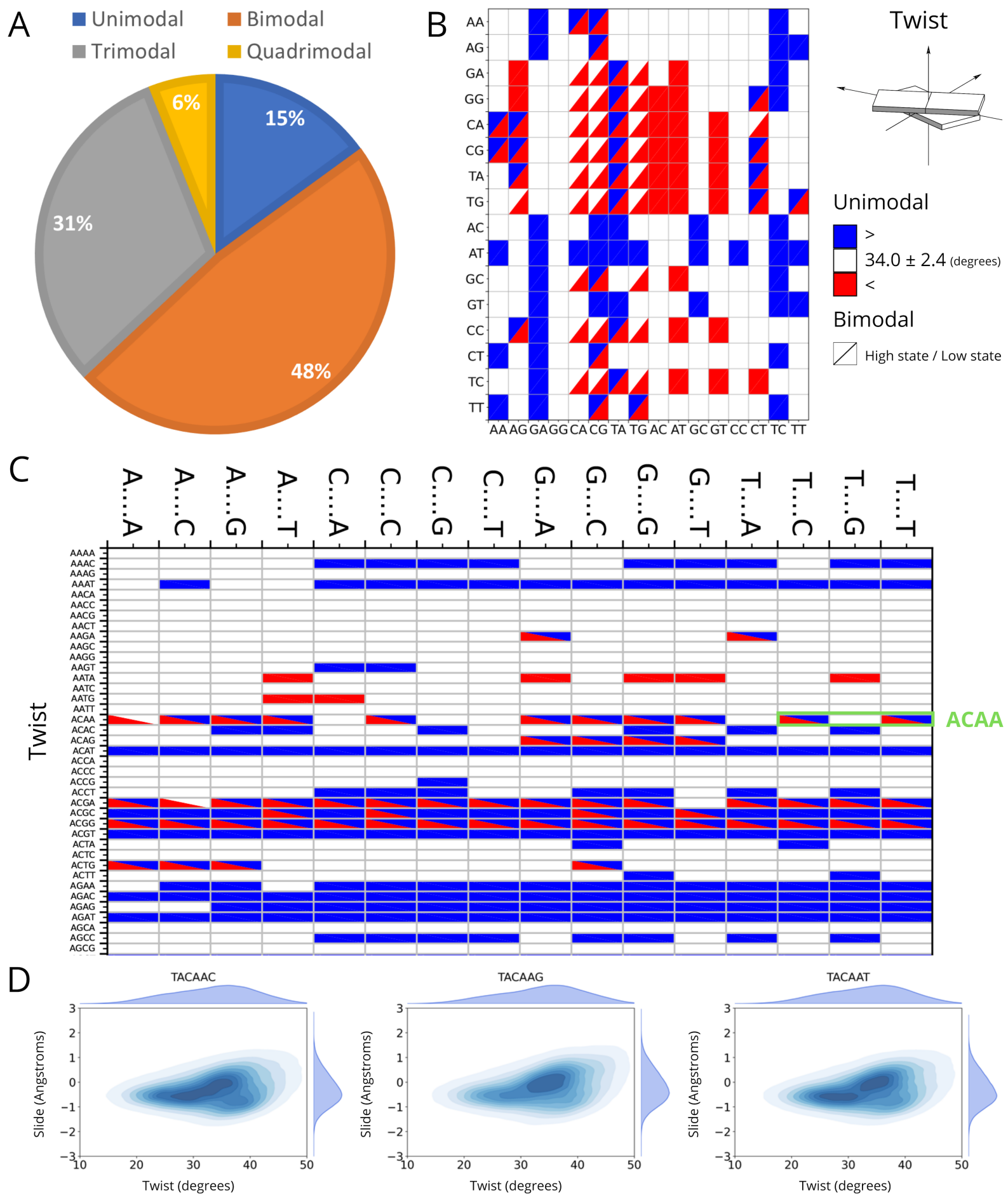
Editorial Summary:

DNA's physical properties are key to gene regulation and protein binding. Here, authors use massive simulations to map all 2,080 unique DNA hexamers, revealing a "physical code" that explains how sequence mechanics shape genome evolution.

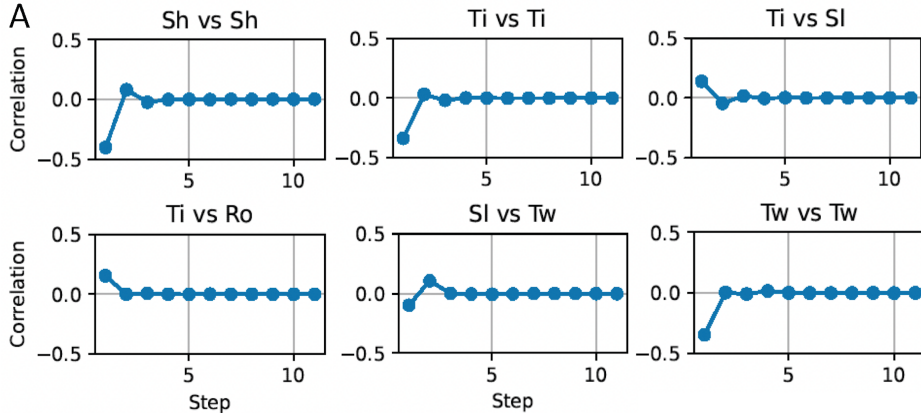
Peer Review Information: *Nature Communications* thanks Alexander P. Lyubartsev, Aleksei Aksimentiev, who co-reviewed with Hemani Chhabra, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.



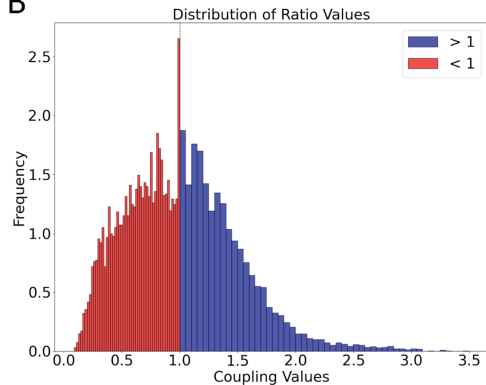


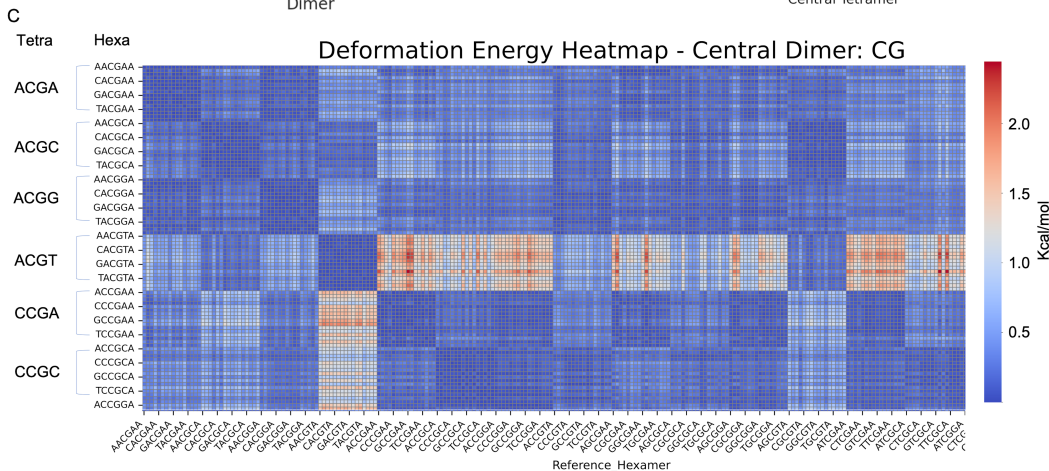
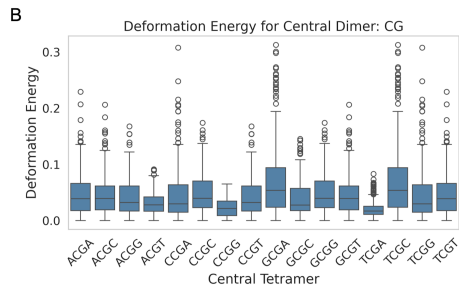
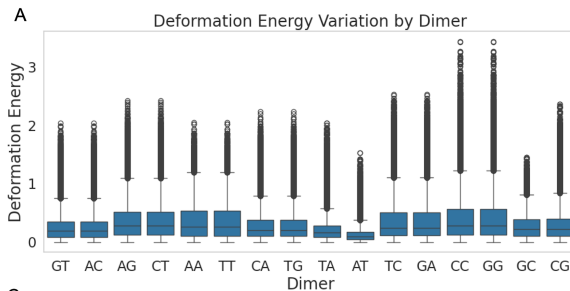


A

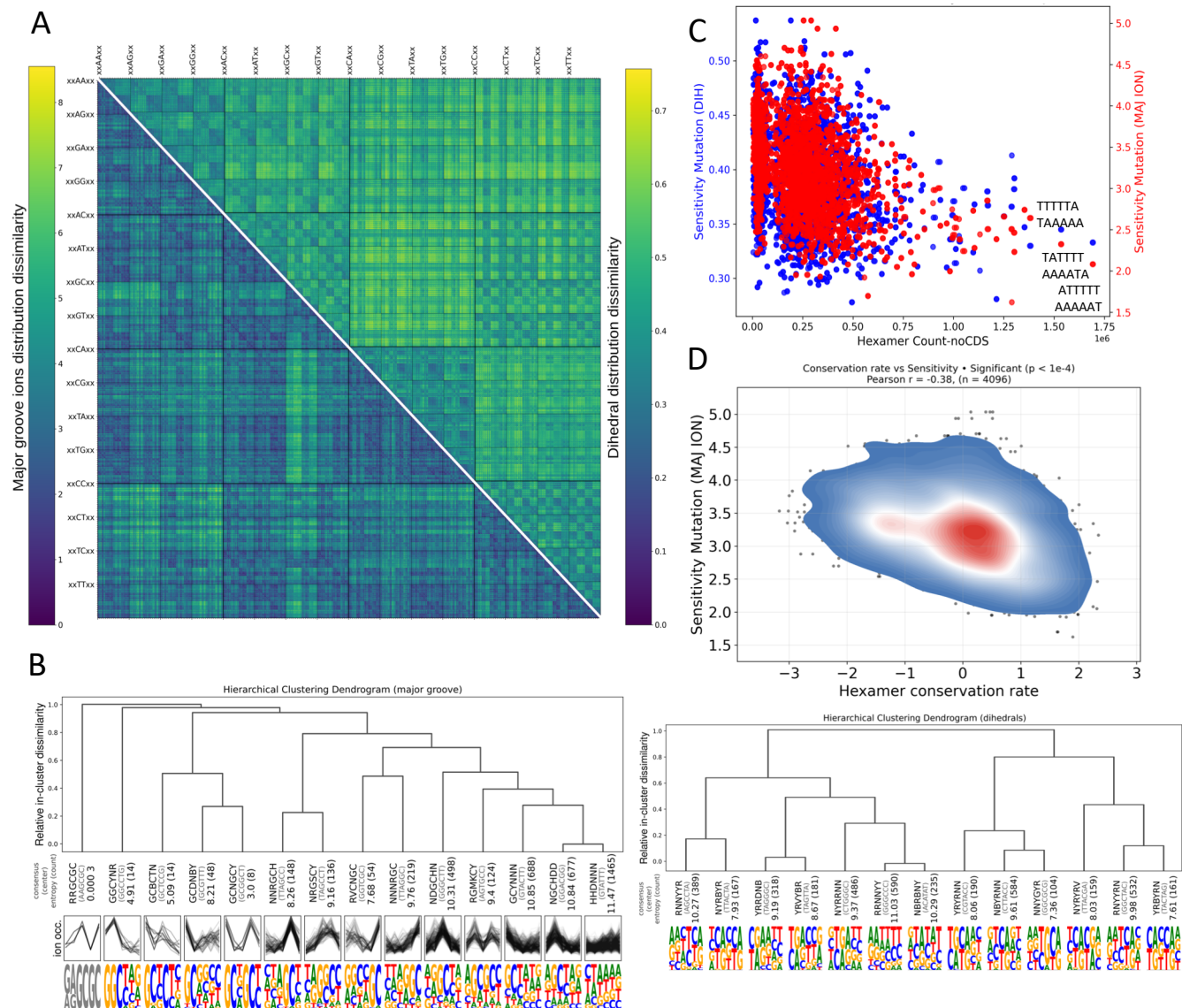


B

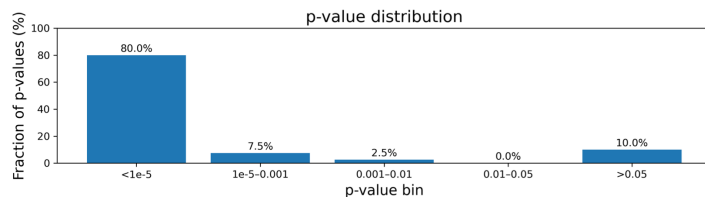




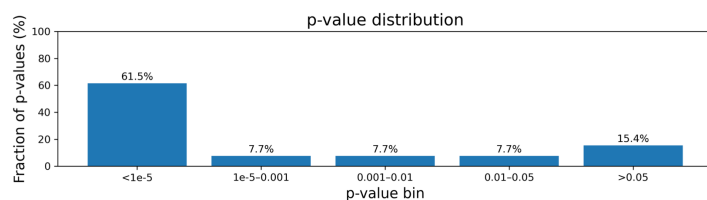
B



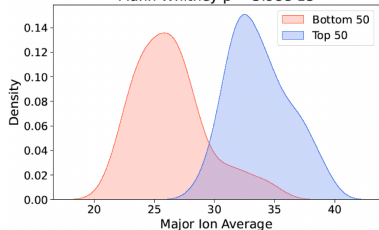
Major groove TFs



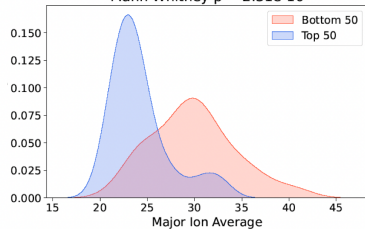
Minor groove TFs



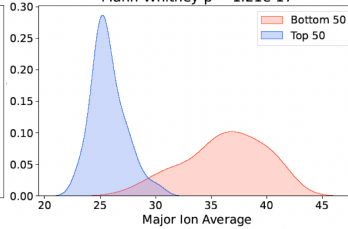
klf15 — Major Groove
Mann-Whitney $p = 3.98\text{e-}15$



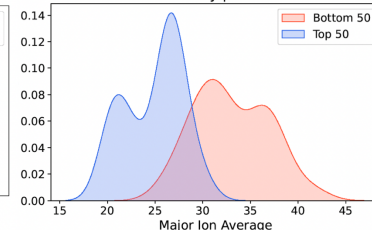
foxl2 — Major Groove
Mann-Whitney $p = 2.31\text{e-}10$



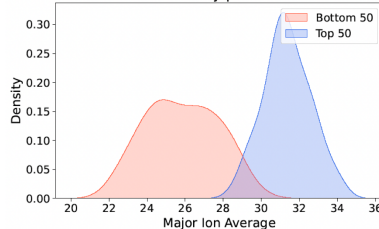
hmbox1 — Major Groove
Mann-Whitney $p = 1.21\text{e-}17$



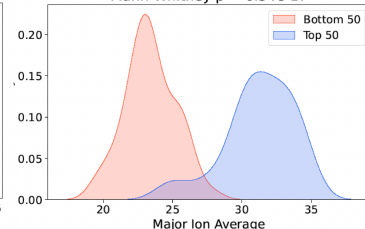
zfc2h2-6 — Major Groove
Mann-Whitney $p = 1.31\text{e-}15$



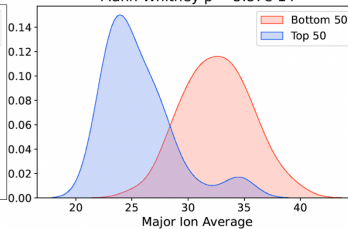
ergb — Major Groove
Mann-Whitney $p = 1.01\text{e-}17$



arid3 — Major Groove
Mann-Whitney $p = 6.34\text{e-}17$



soxc — Major Groove
Mann-Whitney $p = 5.87\text{e-}14$



gata1 — Major Groove
Mann-Whitney $p = 2.99\text{e-}04$

