
Figures and figure supplements

Nucleosome wrapping energy in CpG islands and the role of epigenetic base modifications

Rasa Giniūnaitė et al.

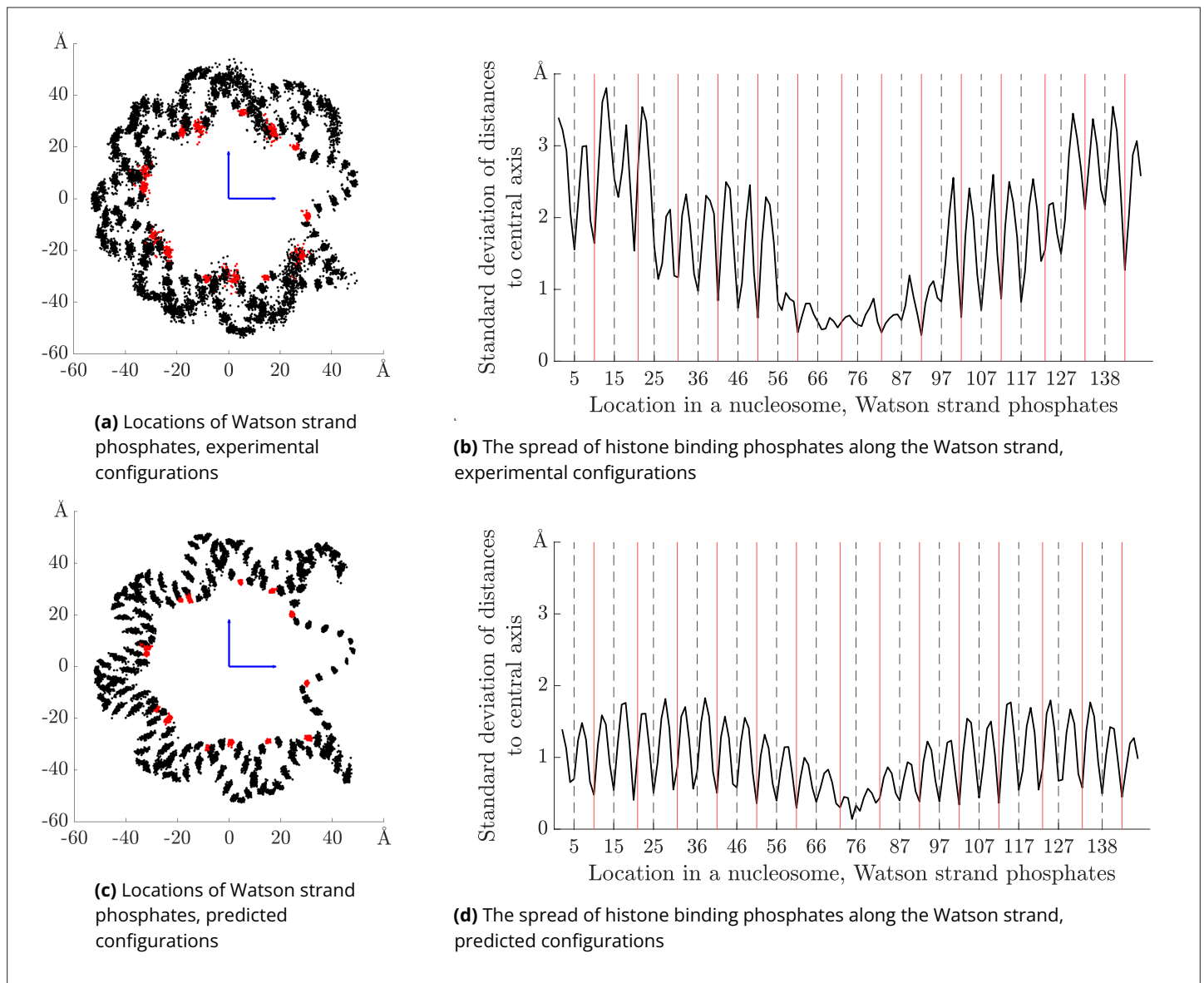


Figure 1. Locations and spread of phosphates in experimental versus predicted nucleosome structures. Left column: locations of the Watson strand phosphates for 100 aligned nucleosome structures, projected to a plane perpendicular to the nucleosome central axis. Top row corresponds to 100 experimental PDB nucleosome structures (not all with independent sequences). Red points are phosphates with local minima of radial distance used to identify bound indices. Bottom row analogous data over 100 predicted minimal energy nucleosomal configurations for sequences drawn from human genome CpG islands. The phosphates with bound indices that are constrained during the optimisation are coloured in red. Right panels: standard deviations over sequence of radial distance of all phosphates against index along the Watson strand. Top PDB structures, bottom model computations. Bound indices are marked with solid red vertical lines. Dashed black vertical lines mark indices of bound complementary (Crick) strand phosphates.

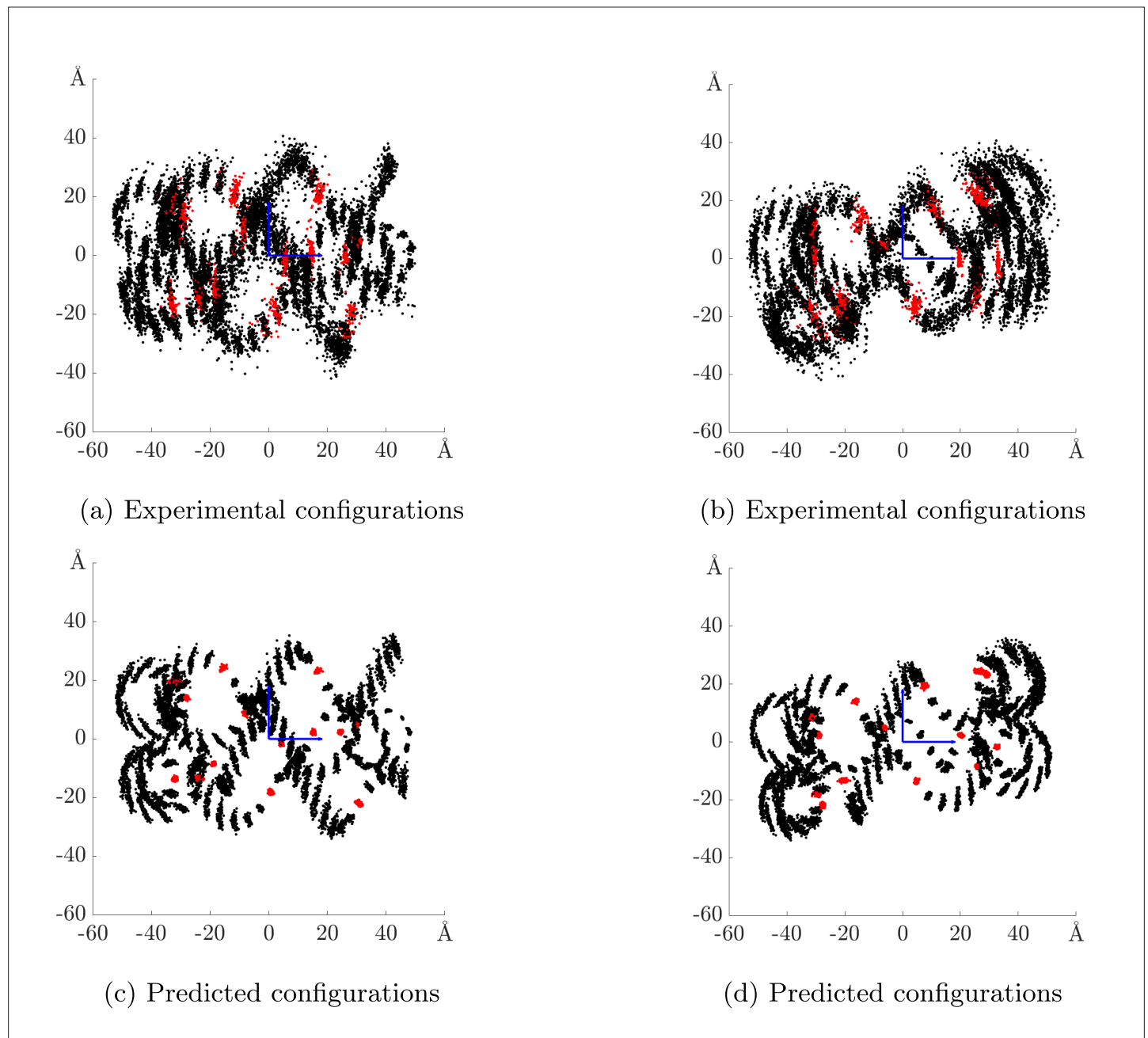


Figure 1—figure supplement 1. Locations of the Watson strand phosphates for 100 aligned nucleosome structures, projected to planes parallel to the nucleosome central axis (side views of the nucleosomes). Top row corresponds to 100 experimental PDB nucleosome structures (not all with independent sequences). Red points are phosphates with local minima of radial distance used to identify bound indices. Bottom row analogous data over 100 predicted minimal energy nucleosomal configurations for sequences drawn from human genome CpG islands. The phosphates with bound indices that are constrained during the optimisation are coloured in red. Left panels: horizontal axis is pointing to the nucleosome dyad, right panels: horizontal axis is perpendicular to the dyad axis and to the nucleosome central axis.

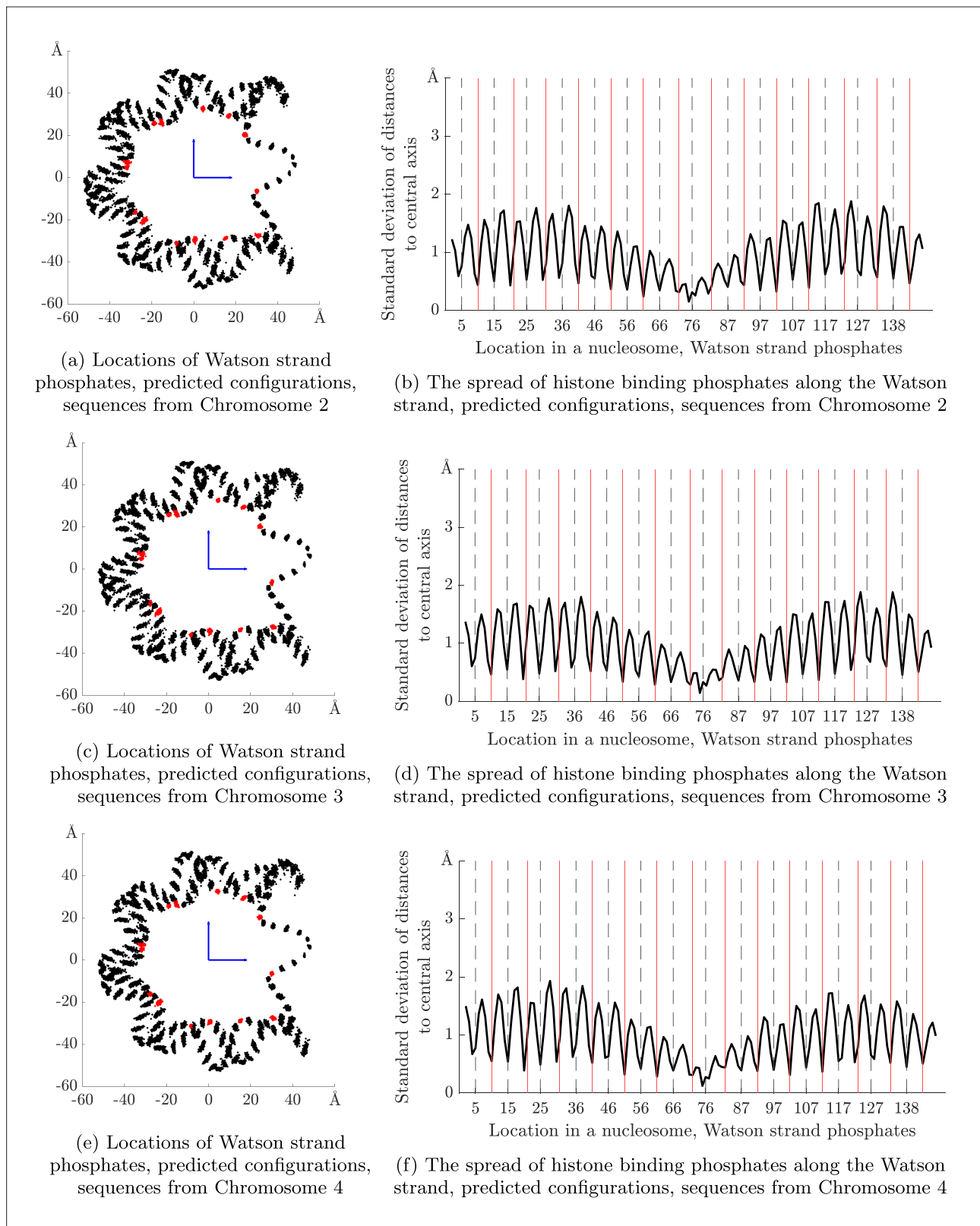


Figure 1—figure supplement 2. Locations and spread of phosphates in nucleosome structures: a comparison of predictions for sequences from different human chromosomes. Left column: locations of the Watson strand phosphates for 100 predicted minimal energy nucleosomal configurations, projected to a plane perpendicular to the nucleosome central axis, for randomly selected sequences from the CpG island (CGI) and non-methylated island (NMI) intersection of Chromosomes 2, 3, and 4 (different rows correspond to different chromosomes). The phosphates with bound indices that are

Figure 1—figure supplement 2 continued on next page

Figure 1—figure supplement 2 continued

constrained during the optimisation are coloured in red. Right panels: standard deviations over sequence of radial distance of all phosphates against index along the Watson strand, computed for the same predicted configurations as shown in plots on the left. Bound indices are marked with solid red vertical lines. Dashed black vertical lines mark indices of bound complementary (Crick) strand phosphates.

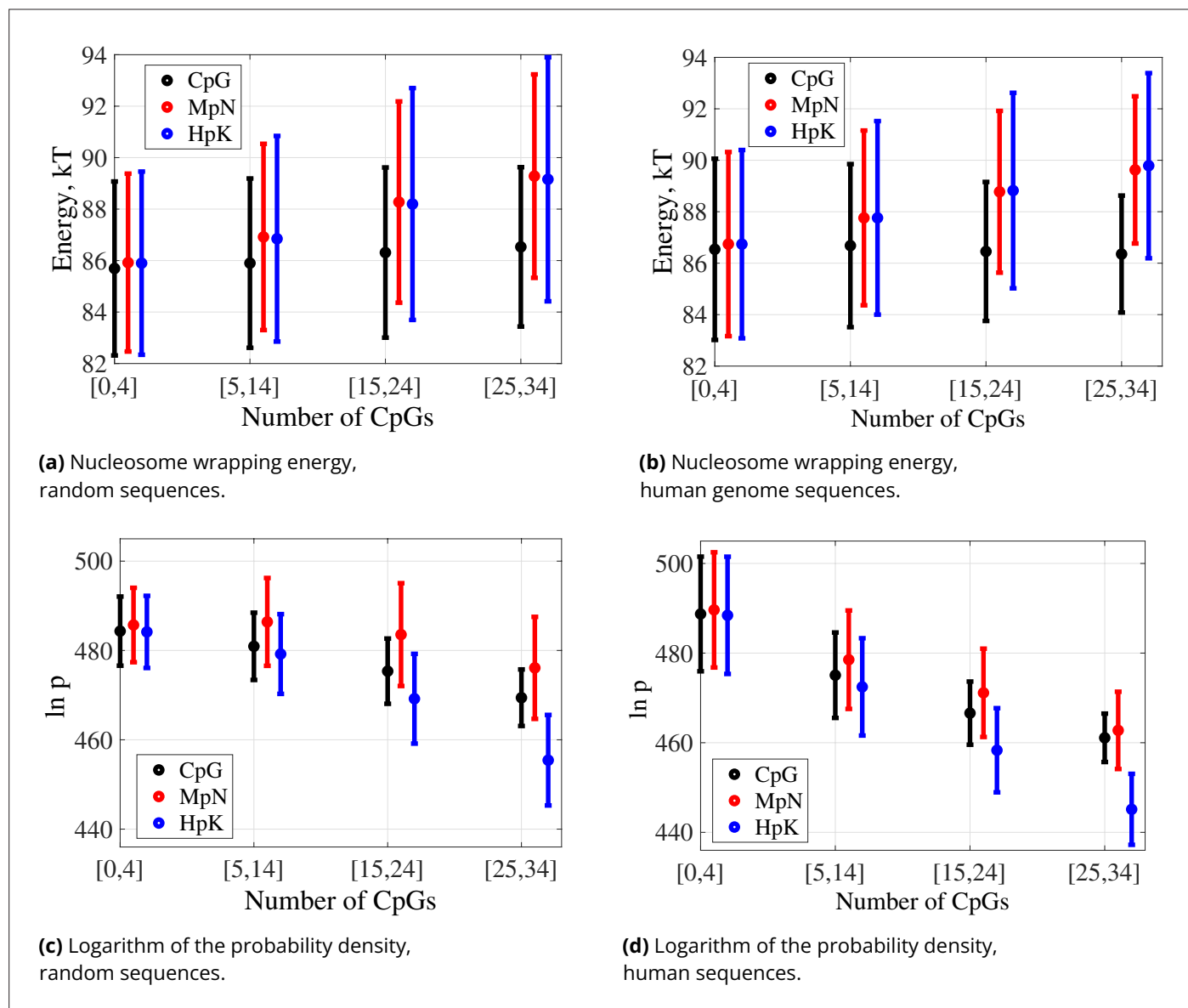


Figure 2. Spectra of nucleosome wrapping energies and logarithms of probability densities for the optimal nucleosomal configurations for 147 bp sequences **(a, c)** generated randomly and **(b, d)** drawn from the human genome, grouped by the indicated ranges of numbers of CpG dinucleotide steps: dots averages, bars standard deviation in sequence. For methylated and hydroxymethylated data, all CpG steps are symmetrically modified. Numbers of sequences falling into each CpG range are given in **Table 2**.

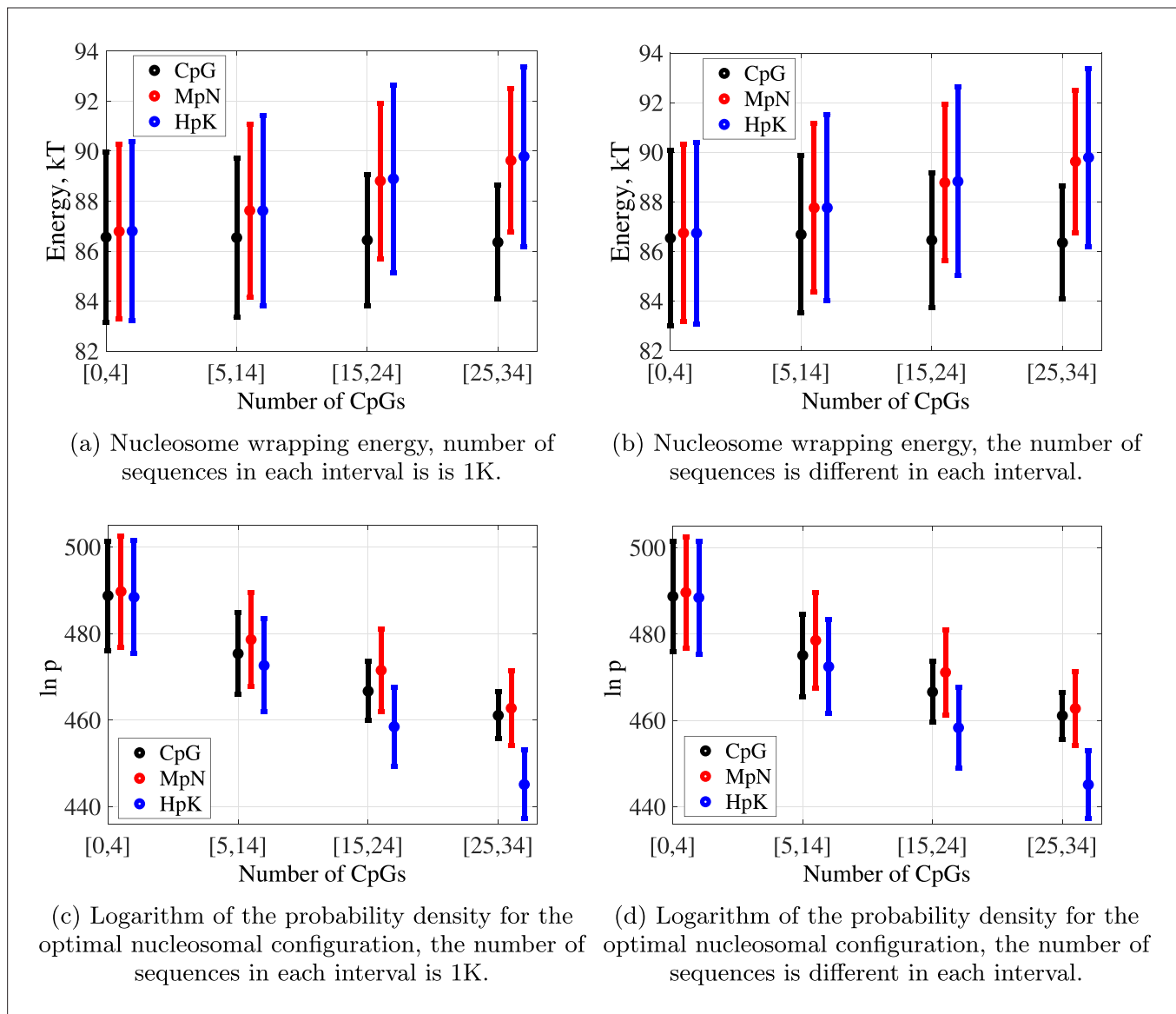


Figure 2—figure supplement 1. Spectra of nucleosome wrapping energies and logarithms of probability densities for the optimal nucleosomal configurations for 147 bp human genome sequences, grouped by the indicated ranges of numbers of CpG dinucleotide steps: dots averages, error bars standard deviation in sequence. For the two plots on the right, we used all the sequences in our ensemble; numbers of sequences falling into each CpG range are given in **Table 2** of the main article. For the plots on the left, we took a random 1K sub-sample of sequences for each CpG range. There are no visible differences between plots on the left and on the right.

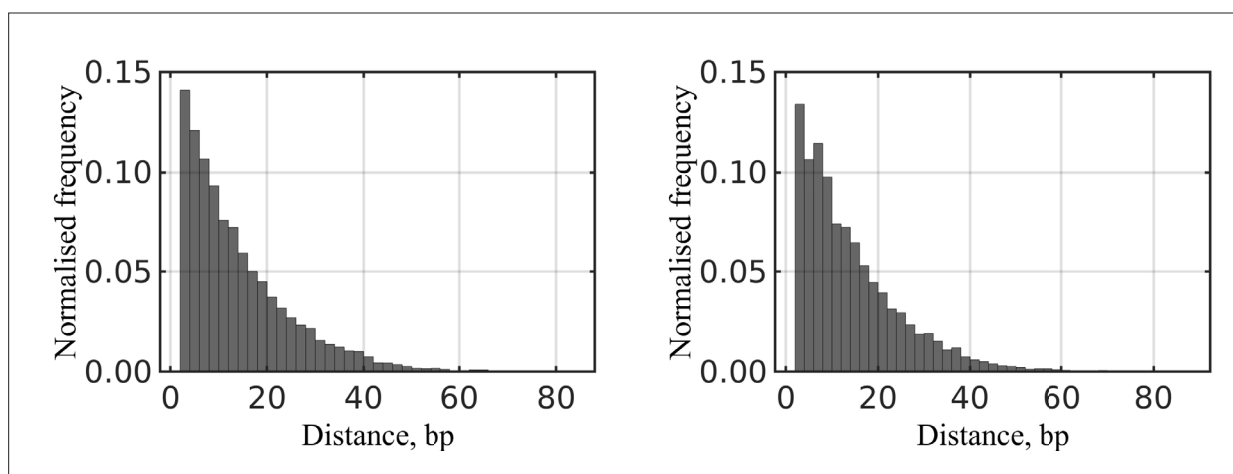


Figure 2—figure supplement 2. Distances between CpG dinucleotides when there are 10 CpG dinucleotides in sequences of length 147. Left – randomly generated DNA sequences (1000). Right – sequences from human genome (2341).

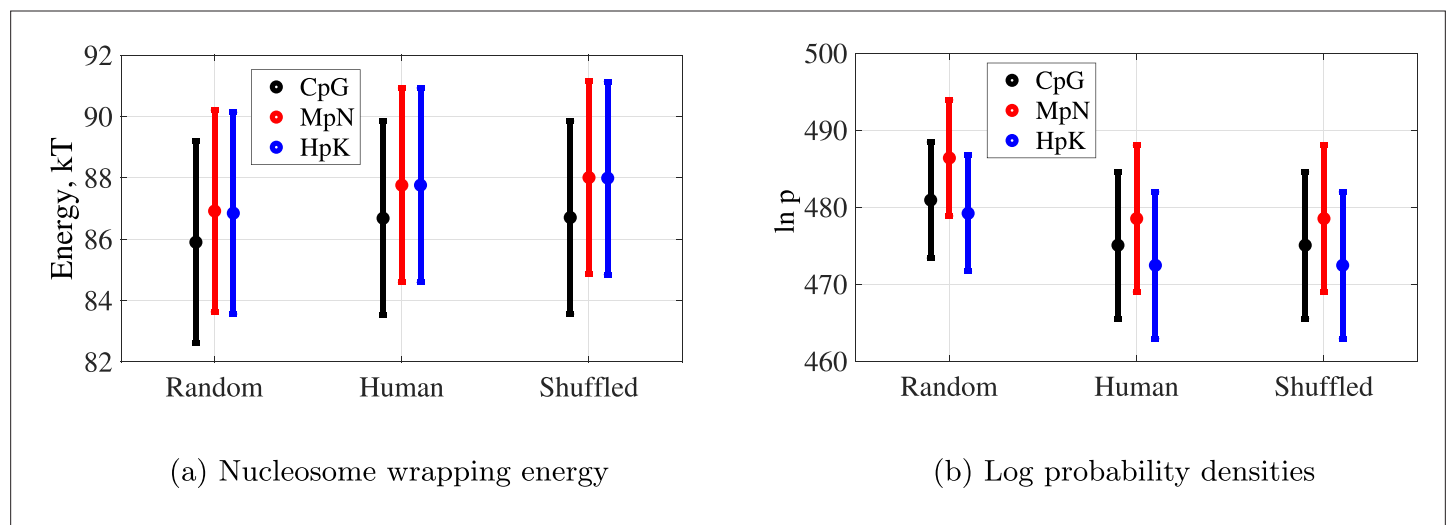


Figure 2—figure supplement 3. Spectra of (a) nucleosome wrapping energy and (b) natural logarithms of probability densities for DNA nucleosomal configurations for unmethylated (CpG), methylated (MpN), and hydroxymethylated (HpK) DNA sequences with CpG dinucleotide count from 5 to 14. Random corresponds to randomly generated sequences and Human to sequences from the human genome. Shuffled corresponds to sequences with the same count of dinucleotides as in the human genome sequences but shuffled. The difference between random and human sequences is significantly larger than the difference between human and shuffled (human) sequences. The results are consistent for different independent shufflings which we verified by performing three independent shufflings for all the sequences.

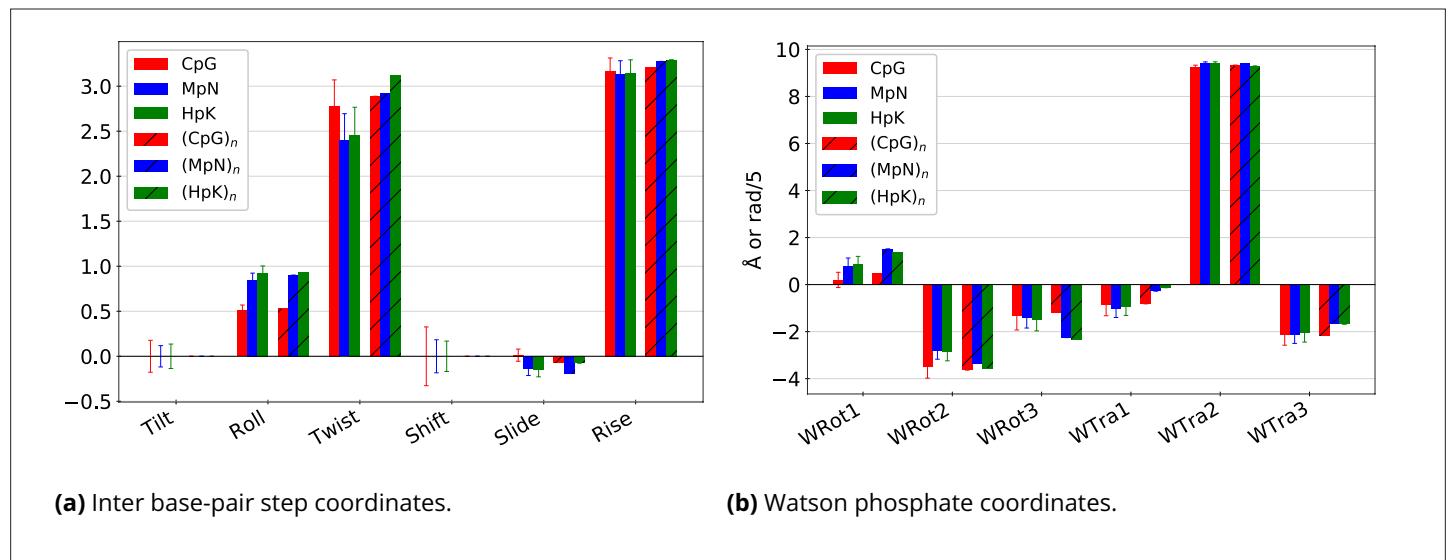


Figure 3. Effects of sequence context and epigenetic base modifications on the cgNA+ model predicted ground state shape of CpG steps. Statistics over $4^8=65,536$ sequences of 22 bp length, constructed around the central CpG step as GCGTCGX4X3X2X1CGY1Y2Y3Y4GTCGGC, with all the possible X_j and $Y_j \in \{A, T, C, G\}$, $\forall j \in \{1, 2, 3, 4\}$. Bar plots show the ground state values of (a) six inter base-pair step and (b) six Watson phosphate coordinates for CpG steps (i) averaged over sequence context with standard deviations in thin lines and (ii) the extreme case of poly(CpG) (in hatch). In each case, three versions corresponding to unmodified, methylated, and hydroxymethylated steps. The standard deviations highlight the crucial role of non-local sequence dependence in the equilibrium structure of CpG/MpN/HpK steps. Analogous plots for the remaining intra base-pair coordinates and Crick phosphate coordinates are shown in **Figure 3—figure supplement 1**.

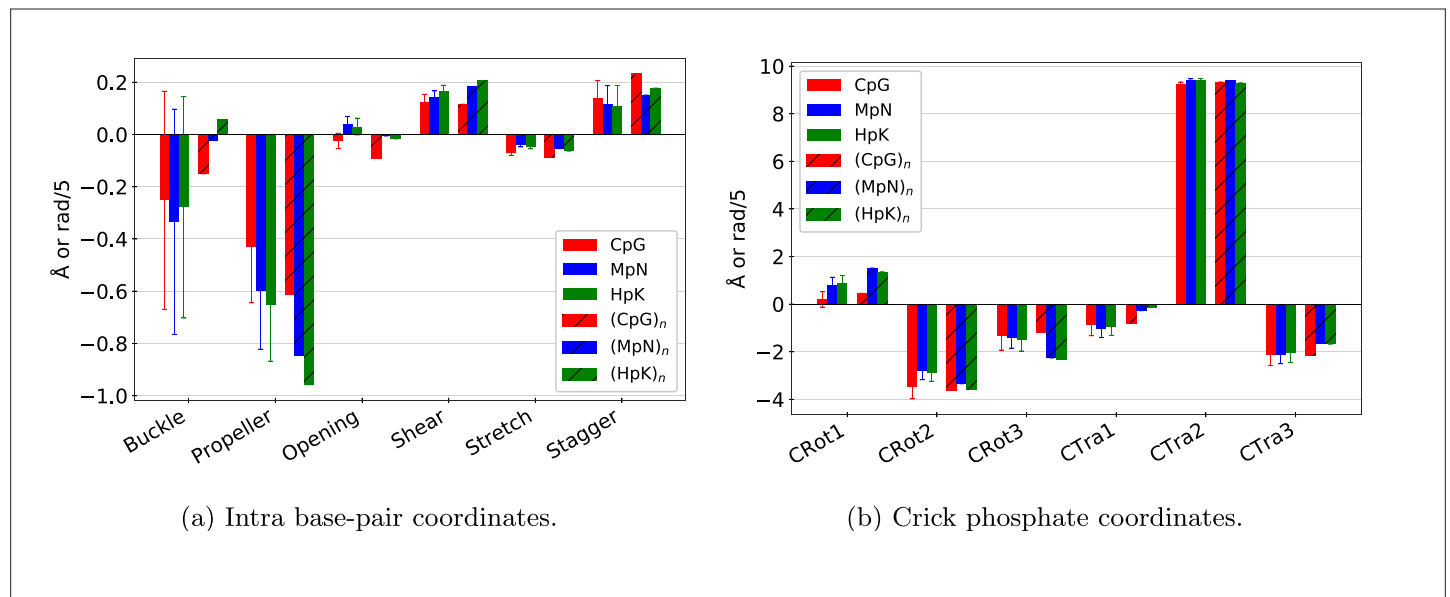
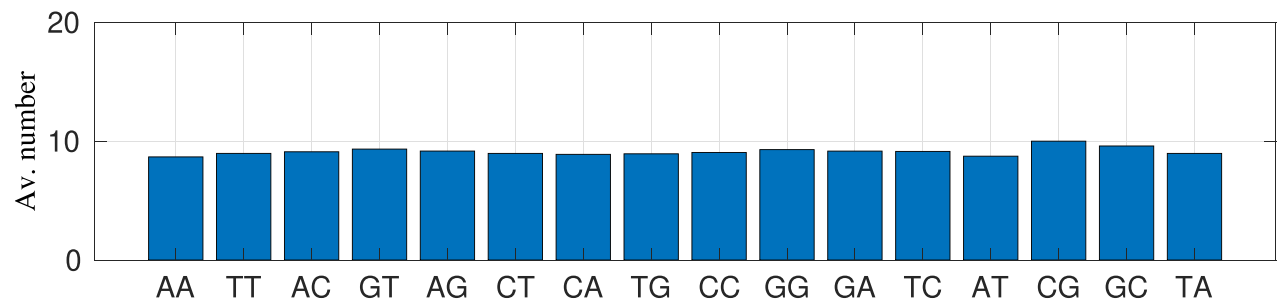
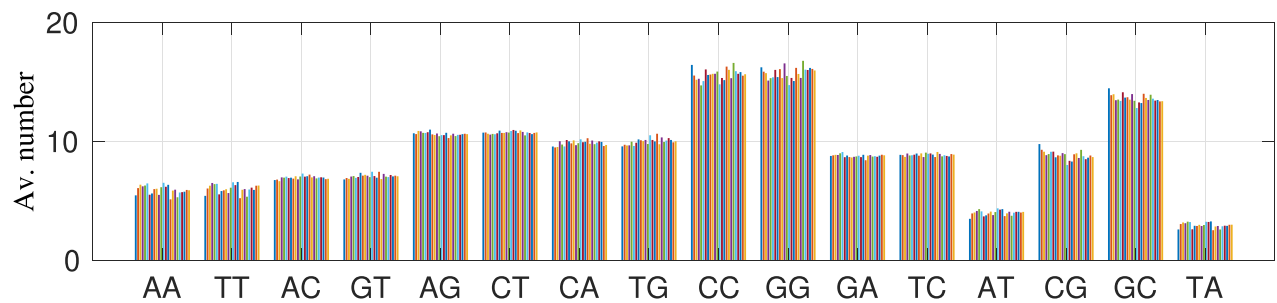


Figure 3—figure supplement 1. Effects of sequence context and epigenetic base modifications on the ground state shape of CpG steps. Bar plots of the ground state values of (a) six intra base-pair and (b) six Crick phosphate coordinates for CpG steps (i) averaged over sequence context with standard deviations in thin lines and (ii) the extreme case of poly(CpG) (in hatch). In each case, three versions corresponding to unmodified, methylated, and hydroxymethylated steps. The standard deviations highlight the crucial role of non-local sequence dependence in the equilibrium structure of CpG/MpN/HpK steps.



(a) Random sequences.



(b) Human genome sequences.

Figure 4. Average number of instances of the 16 different dinucleotide steps for **(a)** 1000 random 147 bp sequences and for **(b)** our 147 bp human genome sequence ensemble, with [5, 14] CpGs. Different colours in **(b)** correspond to fragments taken from different chromosomes. Dinucleotide steps are ordered next to their complements, with self-complementary steps listed on the right.

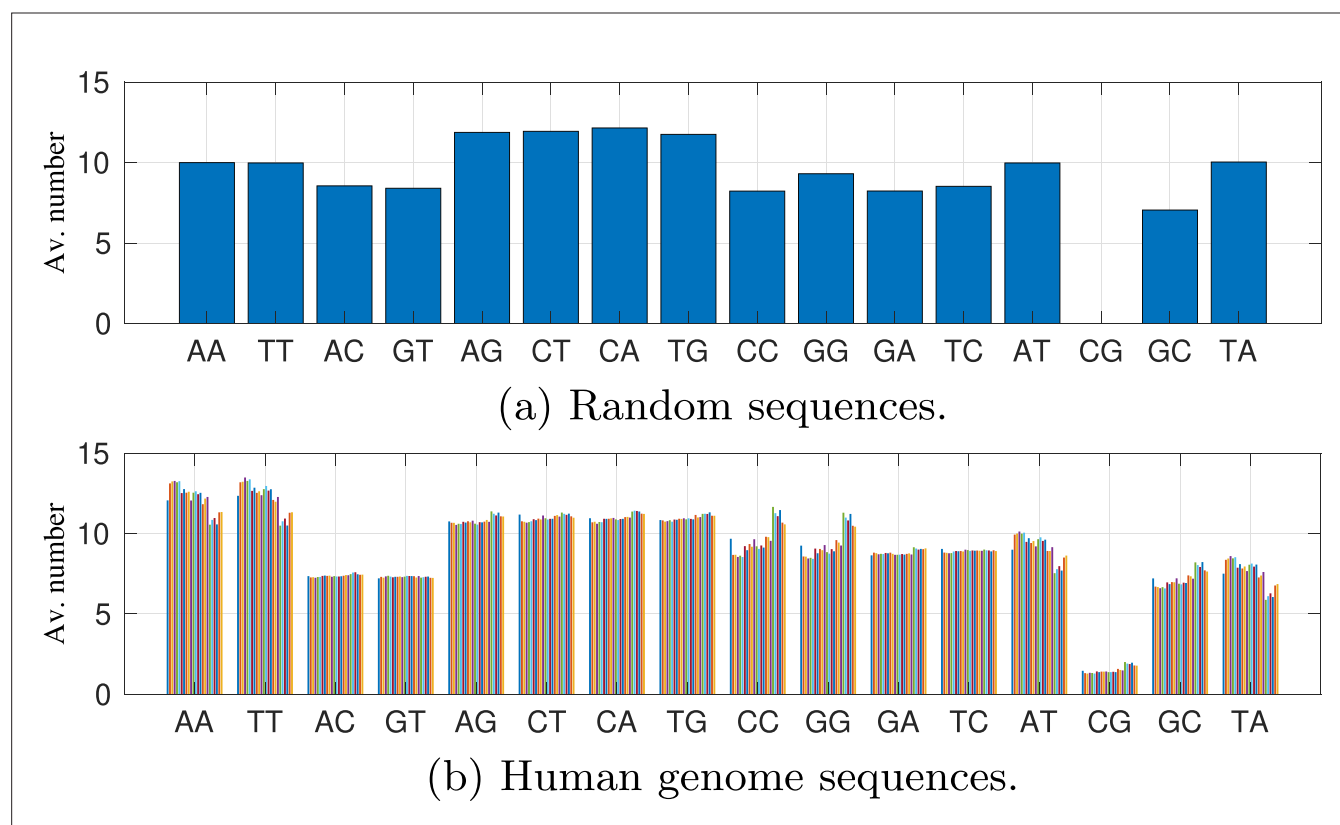


Figure 4—figure supplement 1. Average number of the 16 different dinucleotide steps for (a) 1000 random sequences and for (b) our human genome sequence ensemble, with [0, 4] CpGs. Different colours in (b) correspond to fragments taken from different chromosomes. Dinucleotide steps are ordered next to their complements, with self-complementary steps listed on the right.

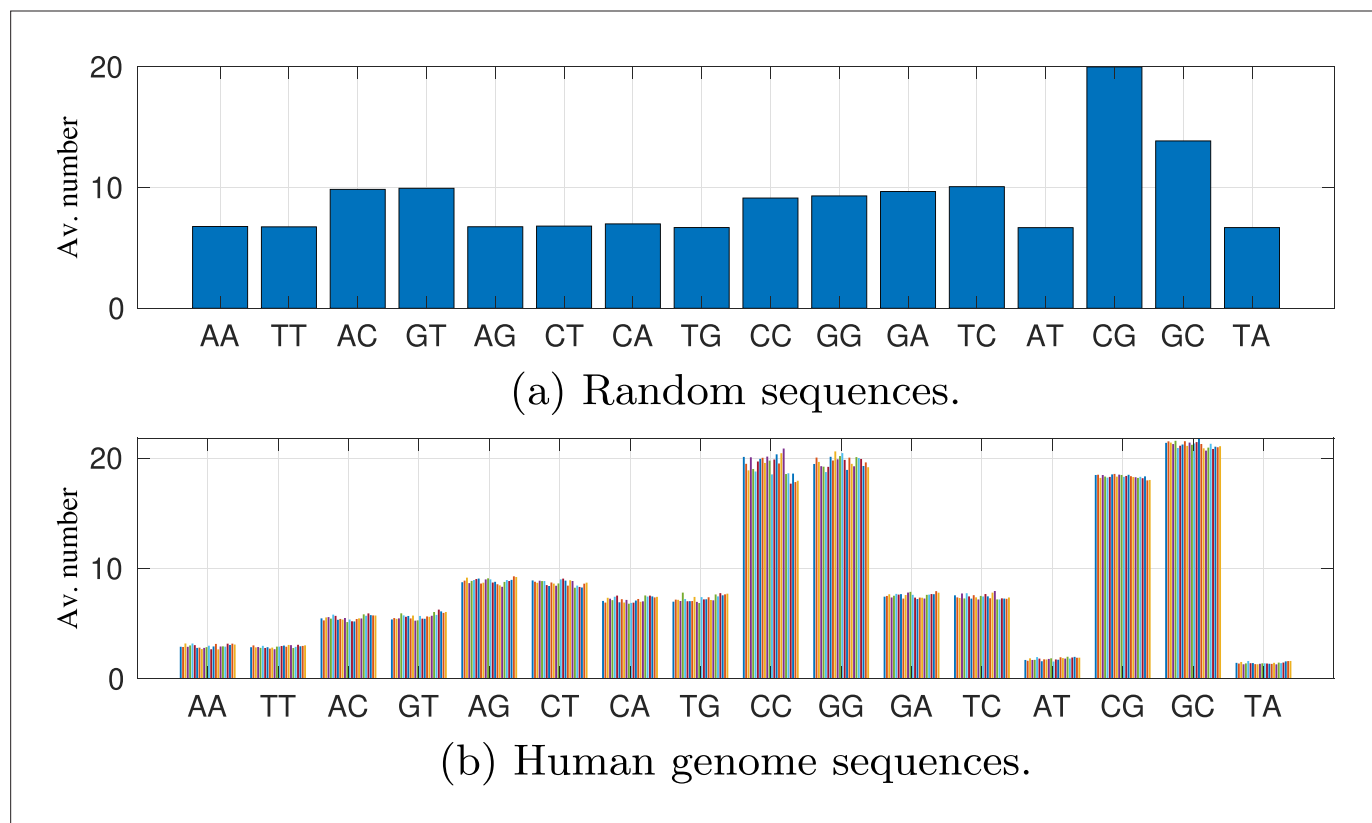


Figure 4—figure supplement 2. Average number of the 16 different dinucleotide steps for (a) 1000 random sequences and for (b) our human genome sequence ensemble, with [15, 24] CpGs. Different colours in (b) correspond to fragments taken from different chromosomes. Dinucleotide steps are ordered next to their complements, with self-complementary steps listed on the right.

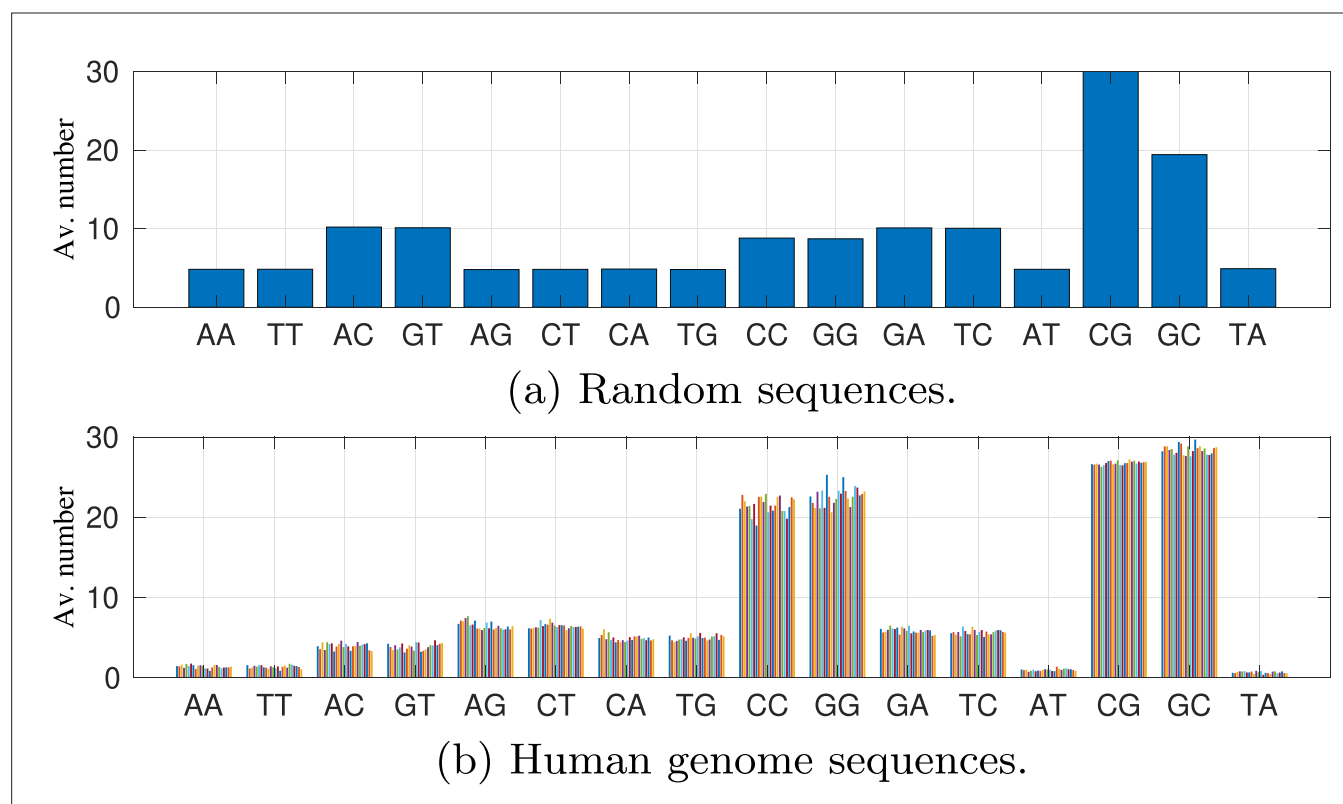


Figure 4—figure supplement 3. Average number of the 16 different dinucleotide steps for (a) 1000 random sequences and for (b) our human genome sequence ensemble, with [25, 34] CpGs. Different colours in (b) correspond to fragments taken from different chromosomes. Dinucleotide steps are ordered next to their complements, with self-complementary steps listed on the right.

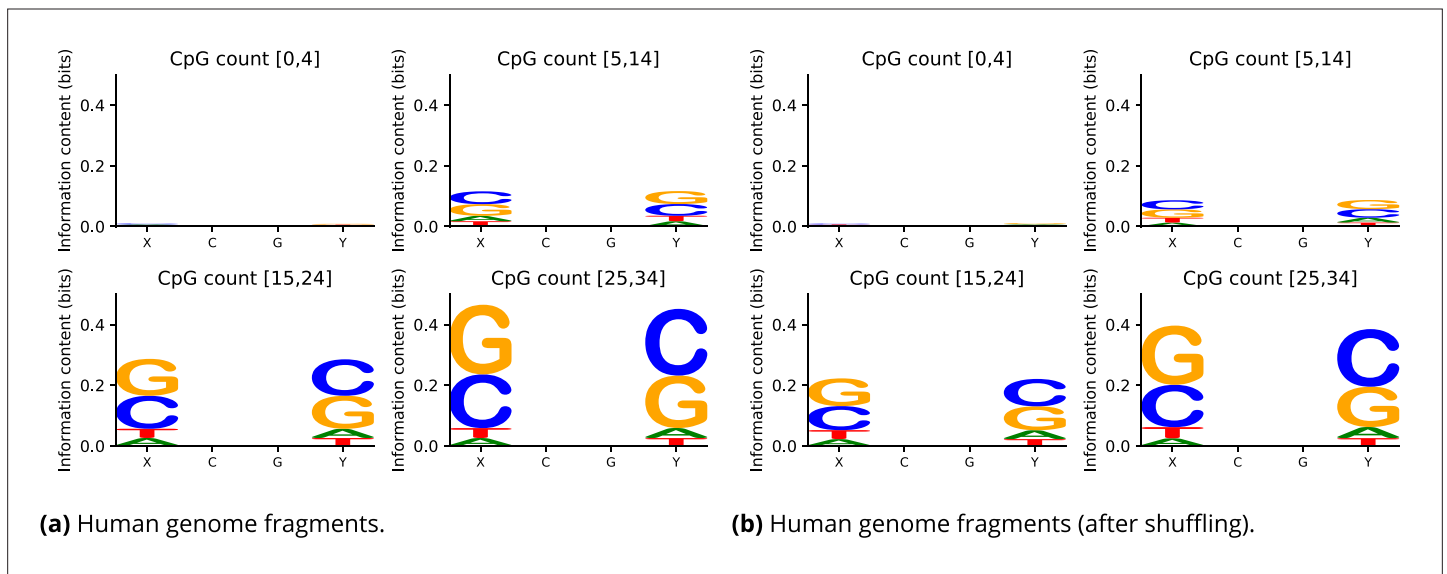


Figure 5. Sequence logos for tetramer flanking context of CpG dinucleotide steps for **(a)** all four sequence ensembles from the human genome with varying numbers of CpG junctions, and **(b)** all four sequence ensembles from the human genome after dinucleotide shuffling (but respecting the numbers of dinucleotide steps). Just specifying the numbers of CpG dinucleotide steps is a strong enough constraint to leave the tetramer sequence context logos largely unchanged after shuffling. The sequence logos in panel **(a)** for the human sequence ensemble before sequence shuffling suggest a slightly stronger C/G flanking enrichment than after shuffling.

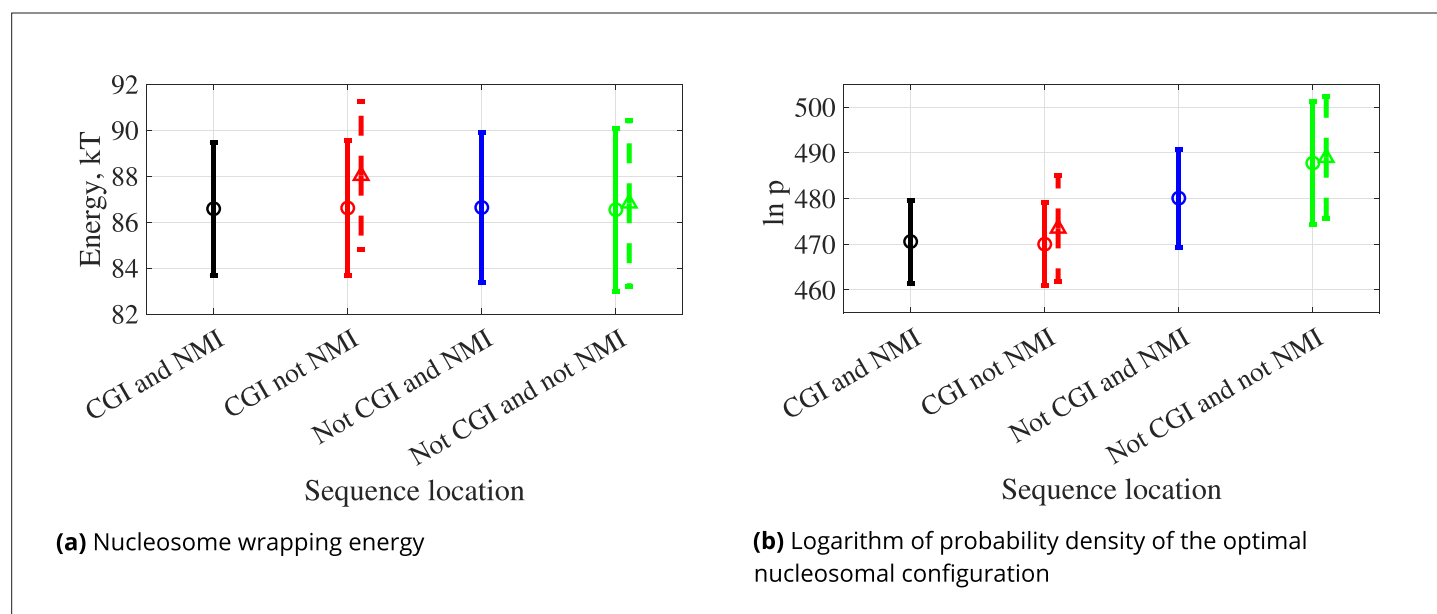


Figure 6. Spectra of (a) nucleosome wrapping energies and (b) log probability densities of the optimal nucleosomal configurations for 147 bp sequences drawn from four different regions of the human genome: (A) intersection of CpG island (CGI) and non-methylated island (NMI), (B) NMI and not CGI, (C) CGI and not NMI, (D) not CGI and not NMI (Table 1). Dots represent averages, error bars represent standard deviation over sequence, solid and circles when CpG dinucleotides are not methylated, dashed and triangles when CpGs are methylated.

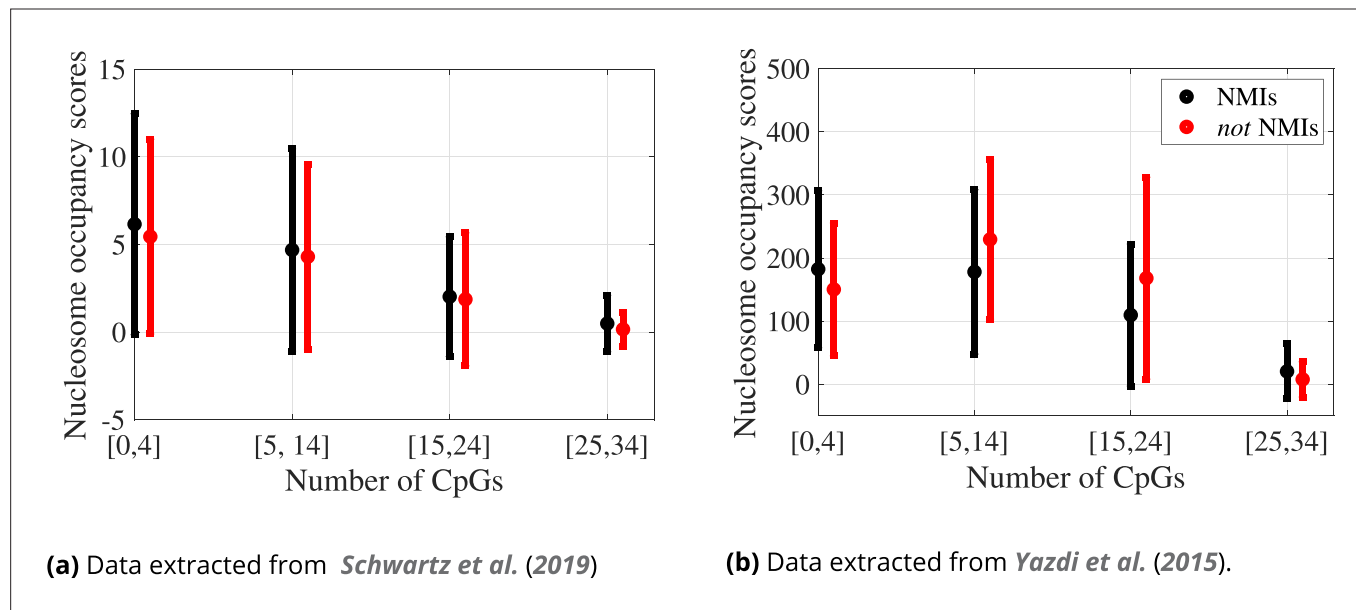


Figure 7. Spectra of nucleosome occupancy scores for our 86,874 selected sequences, grouped by the genomic regions (non-methylated island [NMI] and not NMIs) and by indicated ranges of numbers of CpG dinucleotide steps: dots averages, error bars standard deviation in sequence. The number of sequences in each group is listed in **Table 2**. See also **Figure 2d**.

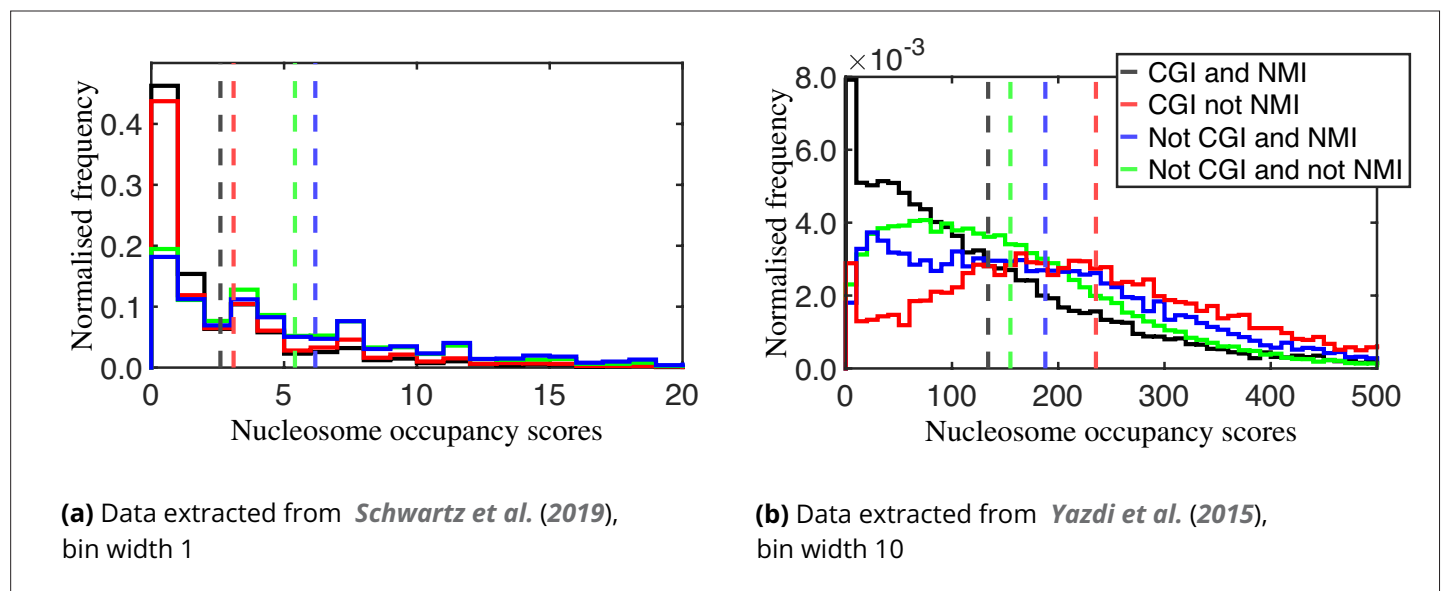


Figure 8. Normalised frequencies (each of the four histograms in each plot normalised independently) of experimental nucleosome occupancy scores for our 86,874 selected sequences grouped by each of the four types of regions in the genome (**Table 1**). Average score for each region is indicated by a vertical dashed line of appropriate colour. The black and red (but not blue or green) histograms have significant spikes reflecting many instances of zero occupancy in the experimental data.

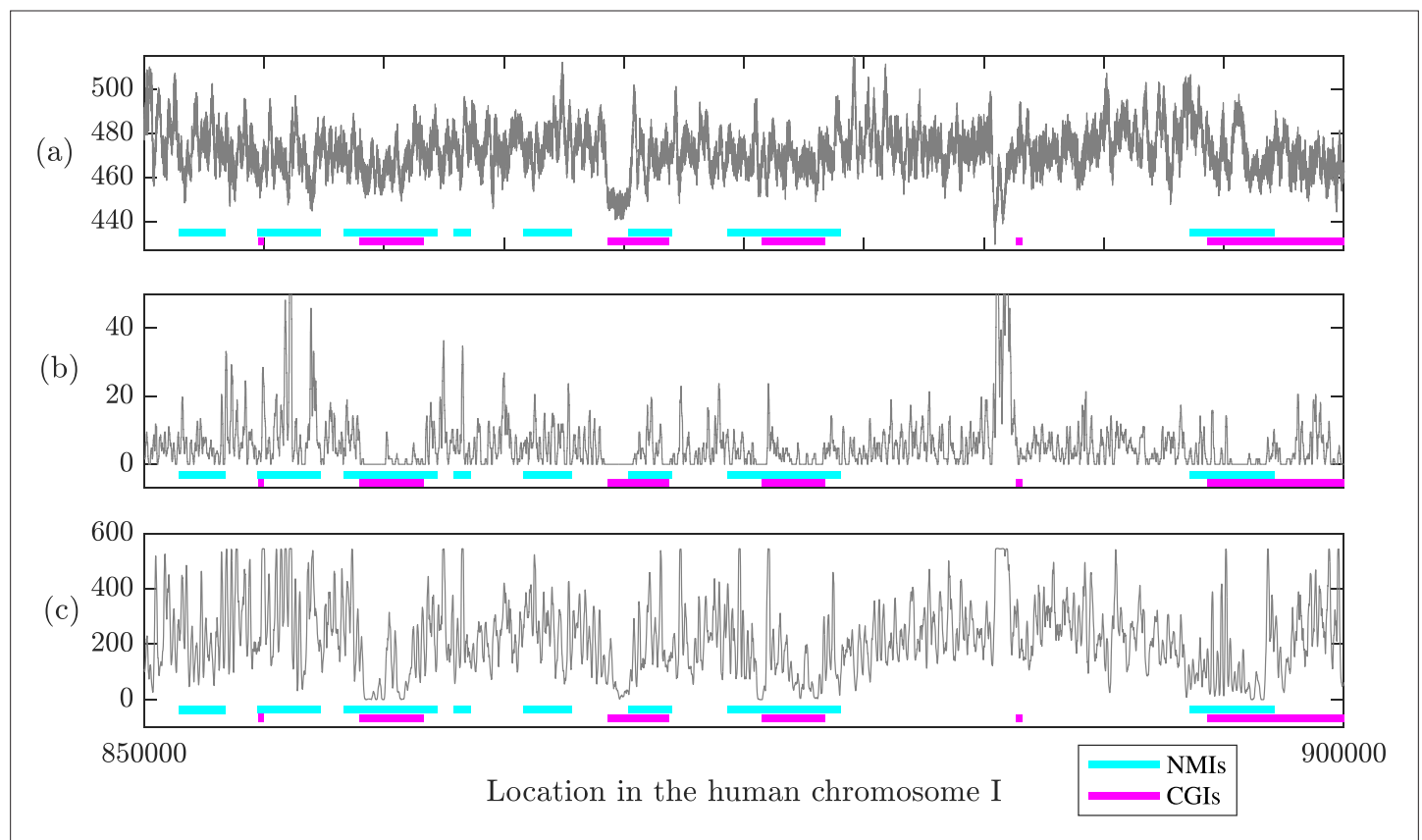


Figure 9. Predicted log probability density for an optimal nucleosomal configuration (a), nucleosome occupancy scores from *Schwartz et al., 2019* (b), and nucleosome occupancy scores from *Yazdi et al., 2015* (c), for sequence positions 850K–900K of human chromosome I. In the regions corresponding to the intersection of CpG islands (CGIs) and non-methylated islands (NMIs), both the mean log probability density (468.61) and mean scores (2.62 and 139.53) are smaller than outside of the intersection regions (476.10, 5.89, and 212.00, respectively).